

# INTRODUCTION TO STATISTICS THROUGH RESAMPLING METHODS AND MICROSOFT<sup>®</sup> OFFICE EXCEL



PHILLIP I. GOOD

---

# INTRODUCTION TO STATISTICS THROUGH RESAMPLING METHODS AND MICROSOFT OFFICE EXCEL<sup>®</sup>

---

Phillip I. Good

 **WILEY-  
INTERSCIENCE**

A JOHN WILEY & SONS, INC., PUBLICATION



---

**INTRODUCTION TO  
STATISTICS THROUGH  
RESAMPLING METHODS  
AND MICROSOFT  
OFFICE EXCEL<sup>®</sup>**

---



---

# INTRODUCTION TO STATISTICS THROUGH RESAMPLING METHODS AND MICROSOFT OFFICE EXCEL<sup>®</sup>

---

Phillip I. Good

 **WILEY-  
INTERSCIENCE**

A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2005 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

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, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-646-8600, or on the web at [www.copyright.com](http://www.copyright.com). Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008.

**Limit of Liability/Disclaimer of Warranty:** While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

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

Good, Phillip L

Introduction to statistics through resampling methods and Microsoft Office Excel /  
Phillip I. Good.

p. cm.

Includes bibliographical references and index.

ISBN-13: 978-0-471-73191-7 (acid-free paper)

ISBN-10: 0-471-73191-9 (pbk : acid-free paper)

1. Resampling (Statistics) 2. Microsoft Excel (Computer file) I. Title.

QA278.8.G62 2005

519.5'4—dc22

2005040801

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

# Contents

<b>Preface</b>	<b>xi</b>
1. Variation (or What Statistics Is All About)	1
1.1. Variation	1
1.2. Collecting Data	2
1.3. Summarizing Your Data	3
1.3.1 Learning to Use Excel	4
1.4. Reporting Your Results: the Classroom Data	7
1.4.1 Picturing Data	10
1.4.2 Displaying Multiple Variables	10
1.4.3 Percentiles of the Distribution	15
1.5. Types of Data	20
1.5.1 Depicting Categorical Data	21
1.5.2 From Observations to Questions	23
1.6. Measures of Location	23
1.6.1 Which Measure of Location?	25
1.6.2 The Bootstrap	27
1.7. Samples and Populations	30
1.7.1 Drawing a Random Sample	32
1.7.2 Ensuring the Sample is Representative	34
1.8. Variation—Within and Between	34
1.9. Summary and Review	36
2. Probability	39
2.1. Probability	39
2.1.1 Events and Outcomes	41
2.1.2 Venn Diagrams	41
2.2. Binomial	43
2.2.1 Permutations and Rearrangements	45
2.2.2 Back to the Binomial	47



2.2.3	The Problem Jury	47
2.2.4	Properties of the Binomial	48
2.2.5	Multinomial	52
2.3.	Conditional Probability	53
2.3.1	Market Basket Analysis	55
2.3.2	Negative Results	56
2.4.	Independence	57
2.5.	Applications to Genetics	59
2.6.	Summary and Review	60
3.	Distributions	63
3.1.	Distribution of Values	63
3.1.1	Cumulative Distribution Function	64
3.1.2	Empirical Distribution Function	66
3.2.	Discrete Distributions	66
3.3.	Poisson: Events Rare in Time and Space	68
3.3.1	Applying the Poisson	69
3.3.2	Comparing Empirical and Theoretical Poisson Distributions	70
3.4.	Continuous Distributions	71
3.4.1	The Exponential Distribution	71
3.4.2	The Normal Distribution	72
3.4.3	Mixtures of Normal Distributions	74
3.5.	Properties of Independent Observations	74
3.6.	Testing a Hypothesis	76
3.6.1	Analyzing the Experiment	77
3.6.2	Two Types of Errors	80
3.7.	Estimating Effect Size	81
3.7.1	Confidence Interval for Difference in Means	82
3.7.2	Are Two Variables Correlated?	84
3.7.3	Using Confidence Intervals to Test Hypotheses	86
3.8.	Summary and Review	87
4.	Testing Hypotheses	89
4.1.	One-Sample Problems	89
4.1.1	Percentile Bootstrap	89
4.1.2	Parametric Bootstrap	90
4.1.3	Student's $t$	91
4.2.	Comparing Two Samples	93
4.2.1	Comparing Two Poisson Distributions	93
4.2.2	What Should We Measure?	94

4.2.3	Permutation Monte Carlo	95
4.2.4	Two-Sample $t$ -Test	97
4.3.	Which Test Should We Use?	97
4.3.1	$p$ Values and Significance Levels	98
4.3.2	Test Assumptions	98
4.3.3	Robustness	99
4.3.4	Power of a Test Procedure	100
4.3.5	Testing for Correlation	101
4.4.	Summary and Review	104
5.	Designing an Experiment or Survey	105
5.1.	The Hawthorne Effect	106
5.1.1	Crafting an Experiment	106
5.2.	Designing an Experiment or Survey	108
5.2.1	Objectives	109
5.2.2	Sample from the Right Population	110
5.2.3	Coping with Variation	112
5.2.4	Matched Pairs	113
5.2.5	The Experimental Unit	114
5.2.6	Formulate Your Hypotheses	114
5.2.7	What Are You Going to Measure?	115
5.2.8	Random Representative Samples	116
5.2.9	Treatment Allocation	117
5.2.10	Choosing a Random Sample	118
5.2.11	Ensuring that Your Observations are Independent	119
5.3.	How Large a Sample?	120
5.3.1	Samples of Fixed Size	121
•	Known Distribution	122
•	Almost Normal Data	125
•	Bootstrap	127
5.3.2	Sequential Sampling	129
•	Stein's Two-Stage Sampling Procedure	129
•	Wald Sequential Sampling	129
•	Adaptive Sampling	133
5.4.	Meta-Analysis	134
5.5.	Summary and Review	135
6.	Analyzing Complex Experiments	137
6.1.	Changes Measured in Percentages	137
6.2.	Comparing More Than Two Samples	138

6.2.1	Programming the Multisample Comparison with Excel	139
6.2.2	What Is the Alternative?	141
6.2.3	Testing for a Dose Response or Other Ordered Alternative	141
6.3.	Equalizing Variances	145
6.4.	Stratified Samples	147
6.5.	Categorical Data	148
6.5.1	One-Sided Fisher's Exact Test	150
6.5.2	The Two-Sided Test	151
6.5.3	Multinomial Tables	152
6.5.4	Ordered Categories	153
6.6.	Summary and Review	154
7.	Developing Models	155
7.1.	Models	155
7.1.1	Why Build Models?	156
7.1.2	Caveats	158
7.2.	Regression	159
7.2.1	Linear Regression	160
7.3.	Fitting a Regression Equation	161
7.3.1	Ordinary Least Squares	162
	• Types of Data	166
7.3.2	Least Absolute Deviation Regression	168
7.3.3	Errors-in-Variables Regression	168
7.3.4	Assumptions	171
7.4.	Problems with Regression	172
7.4.1	Goodness of fit versus prediction	172
7.4.2	Which Model?	173
7.4.3	Measures of Predictive Success	174
7.4.4	Multivariable Regression	175
7.5.	Quantile Regression	182
7.6.	Validation	183
7.6.1	Independent Verification	183
7.6.2	Splitting the Sample	184
7.6.3	Cross-Validation with the Bootstrap	185
7.7.	Classification and Regression Trees	186
7.8.	Data Mining	190
7.9.	Summary and Review	193

8. Reporting Your Findings	195
8.1. What to Report	195
8.2. Text, Table, or Graph?	199
8.3. Summarizing Your Results	200
8.3.1 Center of the Distribution	201
8.3.2 Dispersion	203
8.4. Reporting Analysis Results	204
8.4.1 $p$ Values? Or Confidence Intervals?	205
8.5. Exceptions Are the Real Story	206
8.5.1 Nonresponders	206
8.5.2 The Missing Holes	207
8.5.3 Missing Data	207
8.5.4 Recognize and Report Biases	208
8.6. Summary and Review	209
9. Problem Solving	211
9.1. The Problems	211
9.2. Solving Practical Problems	215
9.2.1 The Data's Provenance	215
9.2.2 Inspect the Data	216
9.2.3 Validate the Data Collection Methods	217
9.2.4 Formulate Hypotheses	217
9.2.5 Choosing a Statistical Methodology	218
9.2.6 Be Aware of What You Don't Know	218
9.2.7 Qualify Your Conclusions	218
<b>Appendix: An Microsoft Office Excel Primer</b>	<b>221</b>
<b>Index to Excel and Excel Add-In Functions</b>	<b>227</b>
<b>Subject Index</b>	<b>229</b>



# Preface

**INTENDED FOR CLASS USE OR SELF-STUDY**, this text aspires to introduce statistical methodology to a wide audience, simply and intuitively, through resampling from the data at hand.

The resampling methods—permutations and the bootstrap—are easy to learn and easy to apply. They require no mathematics beyond introductory high-school algebra, yet are applicable in an exceptionally broad range of subject areas.

Introduced in the 1930s, the numerous, albeit straightforward calculations resampling methods require were beyond the capabilities of the primitive calculators then in use. They were soon displaced by less powerful, less accurate approximations that made use of tables. Today, with a powerful computer on every desktop, resampling methods have resumed their dominant role and table lookup is an anachronism.

Physicians and physicians in training, nurses and nursing students, business persons, business majors, research workers, and students in the biological and social sciences will find here a practical and easily grasped guide to descriptive statistics, estimation, testing hypotheses, and model building.

For advanced students in biology, dentistry, medicine, psychology, sociology, and public health, this text can provide a first course in statistics and quantitative reasoning.

For mathematics majors, this text will form the first course in statistics, to be followed by a second course devoted to distribution theory and asymptotic results.

Hopefully, all readers will find my objectives are the same as theirs: *To use quantitative methods to characterize, review, report on, test, estimate, and classify findings.*

Warning to the autodidact: You can master the material in this text without the aid of an instructor. But you may not be able to grasp even

the more elementary concepts without completing the exercises. Whenever and wherever you encounter an exercise in the text, stop your reading and complete the exercise before going further.

You'll need to download and install several add-ins for Excel to do the exercises, including BoxSampler, Ctree, DDXL, Resampling Statistics for Excel, and XLStat. All are available in no-charge trial versions. Complete instructions for doing the installations are provided in Chapter 1. For those brand new to Excel itself, a primer is included as an Appendix to the text.

For a one-quarter short course, I'd recommend taking students through Chapters 1 and 2 and part of Chapter 3. Chapters 3 and 4 would be completed in the winter quarter along with the start of chapter 5, finishing the year with Chapters 5, 6, and 7. Chapters 8 and 9 on "Reporting Your Findings" and "Problem Solving" convert the text into an invaluable professional resource.

An Instructor's Manual is available to qualified instructors and may be obtained by contacting the Publisher. Please visit [ftp://ftp.wiley.com/public/sci\\_tech\\_med/introduction\\_statistics/](ftp://ftp.wiley.com/public/sci_tech_med/introduction_statistics/) for instructions on how to request a copy of the manual.

Twenty-eight or more exercises included in each chapter plus dozens of thought-provoking questions in Chapter 9 will serve the needs of both classroom and self-study. The discovery method is utilized as often as possible, and the student and conscientious reader are forced to think their way to a solution rather than being able to copy the answer or apply a formula straight out of the text. To reduce the scutwork to a minimum, the data sets for the exercises may be downloaded from [ftp://ftp.wiley.com/public/sci\\_tech\\_med/statistics\\_resampling](ftp://ftp.wiley.com/public/sci_tech_med/statistics_resampling).

If you find this text an easy read, then your gratitude should go to Cliff Lunneborg for his many corrections and clarifications. I am deeply indebted to the students in the Introductory Statistics and Resampling Methods courses that I offer on-line each quarter through the auspices of [statistics.com](http://statistics.com) for their comments and corrections.

**Phillip I. Good**  
Huntington Beach, CA  
[frere\\_until@hotmail.com](mailto:frere_until@hotmail.com)

# Chapter 1

## Variation (or What Statistics Is All About)

*If there were no variation, if every observation were predictable, a mere repetition of what had gone before, there would be no need for statistics.*

### 1.1. VARIATION

We find physics extremely satisfying. In high school, we learned the formula  $S = VT$ , which in symbols relates the distance traveled by an object to its velocity multiplied by the time spent in traveling. If the speedometer says 60 miles an hour, then in half an hour you are certain to travel exactly 30 miles. Except that during our morning commute, the speed we travel is seldom constant.

In college, we had Boyle's law,  $V = KT/P$ , with its tidy relationship between the volume  $V$ , temperature  $T$ , and pressure  $P$  of a perfect gas. This is just one example of the perfection encountered there. The problem was we could never quite duplicate this (or any other) law in the freshman physics laboratory. Maybe it was the measuring instruments, our lack of familiarity with the equipment, or simple measurement error, but we kept getting different values for the constant  $K$ .

By now, we know that variation is the norm. Instead of getting a fixed, reproducible  $V$  to correspond to a specific  $T$  and  $P$ , one ends up with a distribution of values instead as a result of errors in measurement. But we also know that with a large enough sample, the mean and shape of this distribution are reproducible.

That's the good news: Make astronomical, physical, or chemical measurements and the only variation appears to be due to observational error. But try working with people.

Anyone who has spent any time in a schoolroom, whether as a parent or as a child, has become aware of the vast differences among individuals.



Our most distinct memories are of how large the girls were in the third grade (ever been beat up by a girl?) and the trepidation we felt on the playground whenever teams were chosen (not right field again!). Much later, in our college days, we were to discover there were many individuals capable of devouring larger quantities of alcohol than we could without noticeable effect, and a few, mostly of other nationalities, whom we could drink under the table.

Whether or not you imbibe, we're sure you've had the opportunity to observe the effects of alcohol on others. Some individuals take a single drink and their nose turns red. Others can't seem to take just one drink.

The majority of effort in *experimental design*, the focus of Chapter 5 of this text, is devoted to finding ways in which this variation from individual to individual won't swamp or mask the variation that results from differences in treatment or approach. It's probably safe to say that what distinguishes statistics from all other branches of applied mathematics is that it is devoted to characterizing and then accounting for *variation*.

#### SOURCES OF VARIATION

You catch three fish. You heft each one and estimate its weight; you weigh each one on a pan scale when you get back to dock, and you take them to a chemistry laboratory and weigh them there. Your two friends on the boat do exactly the same thing. (All but Mike; the chem professor catches him and calls campus security. This is known as missing data.)

The 26 weights you've recorded ( $3 \times 3 \times 3 - 1$  when they nabbed Mike) differ as result of measurement error, observer error, differences among observers, differences among measuring devices, and differences among fish.

## 1.2. COLLECTING DATA

The best way to observe variation is for you, the reader, to collect some data. But before we make some suggestions, a few words of caution are in order: 80% of the effort in any study goes into data collection and preparation for data collection. Any effort you don't expend goes into cleaning up the resulting mess.

We constantly receive letters and E-mails asking which statistic we would use to rescue a misdirected study. There is no magic formula, no secret procedure known only to PhD statisticians. The operative phrase is GIGO: Garbage In, Garbage Out. So think carefully before you embark on your collection effort. Make a list of possible sources of variation and see whether you can eliminate any that are unrelated to the objectives of

your study. If midway through, you think of a better method—don't use it. Any inconsistency in your procedure will only add to the undesired variation.

Let's get started. Here are three suggestions. Before continuing with your reading, follow through on at least one of them or an equivalent idea of your own, as we will be using the data you collect in the very next section:

1. Measure the height, circumference, and weight of a dozen humans (or dogs, or hamsters, or frogs, or crickets).
2. Time some tasks. Record the times of 5–10 individuals over three track lengths (say 50 meters, 100 meters, and a quarter mile). Because the participants (or trial subjects) are sure to complain they could have done much better if only given the opportunity, record at least two times for each study subject. (Feel free to use frogs, hamsters, or turtles in place of humans as runners to be timed. Or to replace foot races with knot tying, bandaging, or putting on or taking off a uniform.)
3. Take a survey. Include at least three questions and survey at least 10 subjects. All your questions should take the form “Do you prefer A to B? Strongly prefer A, slightly prefer A, indifferent, slightly prefer B, strongly prefer B.” For example, “Do you prefer Britney Spears to Jennifer Lopez?” or “Would you prefer spending money on new classrooms rather than guns?”

#### SOURCES OF VARIATION

- Characteristics of the observer(s)
- Characteristics of the environment in which observations are made
- Characteristics of the measuring device(s)
- Characteristics of the subjects or objects observed

**Exercise 1.1.** Collect data as described above. Before you begin, write down a complete description of exactly what you intend to measure and how you plan to make your measurements. Make a list of all potential sources of variation. When your study is complete, describe what deviations you had to make from your plan and what additional sources of variation you encountered.

### 1.3. SUMMARIZING YOUR DATA

Learning how to adequately summarize one's data can be a major challenge. Can it be explained with a single number like the median? The

*median* is the middle value of the observations you have taken, so that half of the data have a smaller value and half have a greater value. Take the observations 1.2, 2.3, 4.0, 3, and 5.1. The observation 3 is the one in the middle. If we have an even number of observations such as 1.2, 2.3, 3, 3.8, 4.0, and 5.1, then the best one can say is that the median or midpoint is a number (any number) between 3 and 3.8. Now, a question for you: What are the median values of the measurements you made?

Hopefully, you've already collected data as described in Section 1.2; otherwise, face it, you are behind. Get out the tape measure and the scales. If you conducted time trials, use those data instead. Treat the observations for each of the three distances separately.

If you conducted a survey, we have a bit of a problem. How does one translate "I would prefer spending money on new classrooms rather than guns" into a number a computer can add and subtract? There is more one way to do this, as we'll discuss in what follows under the heading, "Types of Data." For the moment, assign the number 1 to "Strongly prefer classrooms," the number 2 to "Slightly prefer classrooms," and so on.

### 1.3.1. Learning to Use Excel

Calculating the value of a statistic is easy enough when we've only 1 or 2 observations, but a major pain when we have 10 or more. And as for drawing graphs—one of the best ways to summarize your data—we're no artists. Let the computer do the work.

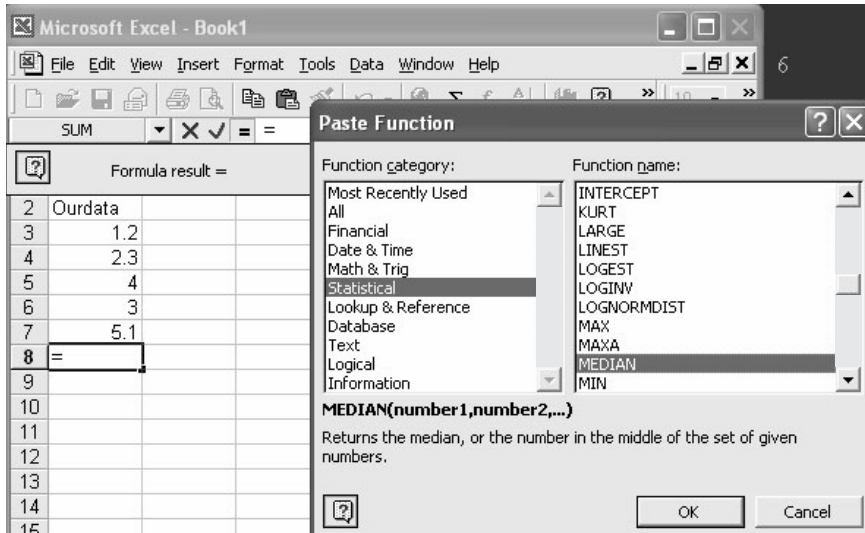
We're going to need the help of Excel, a spreadsheet program with many built-in statistics and graphics functions. We'll assume that you already have Microsoft Office Excel installed and have some familiarity with its use.<sup>1</sup> To enter the observations 1.2, 2.3, 4.0, 3, and 5.1, simply type these values down the first column starting in the third row. Notice in Fig. 1.1 that we've put a description of the column in the second row. The first row is reserved for a more lengthy description of the project should one be required.

In Fig. 1.1, we've begun in Row 8 to start the computation of the median of our data. Here are the steps we went through:

1. Type the first data element (1.2 in this example) in the third row of the first column.
2. Press the "Enter" key to go to the next row.

---

<sup>1</sup> If you're an absolute beginner, we've included an Appendix to the text to help you get started. If you already own and are familiar with some other statistics package or spreadsheet, feel free to use it instead. The objective of this text is to help you understand and make use of basic statistics principles. Excel is merely a convenient tool.



**FIGURE 1.1** Using Excel to compute the median of a data set.

3. Repeat steps 1 and 2 until all the data are entered.
4. Use your mouse to depress the = button in the row.

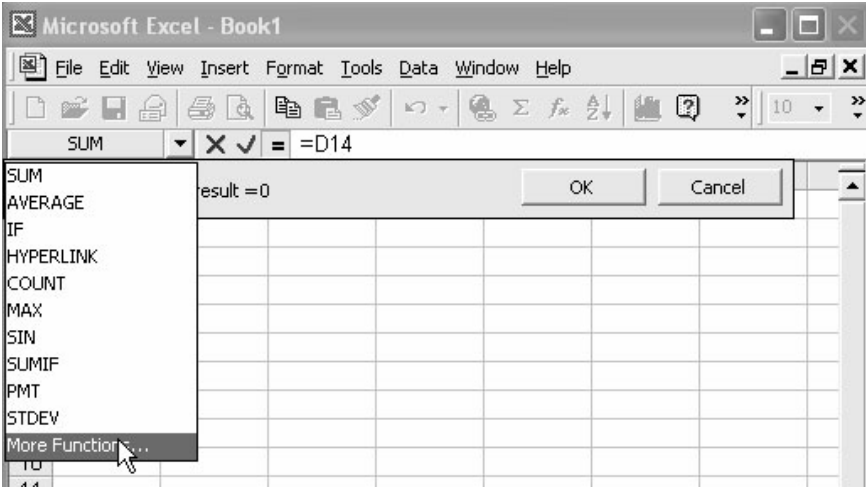


5. Depress the down arrow next to the word SUM and select “More Functions” from the resultant display (Fig. 1.2).
6. Select “Statistical” from the Function category menu and “Median” from the Function name menu.
7. Press “OK” or the “Enter” key to learn that the median of the five numbers we entered is 2.65.

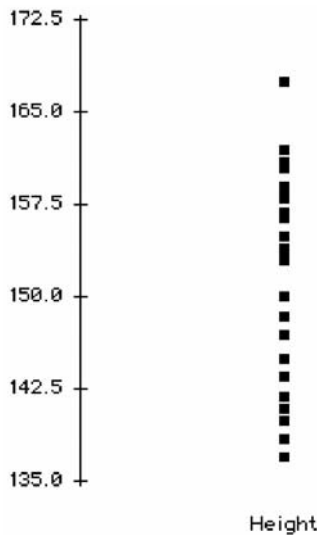
The *median* of a sample tells us where the center of a set of observations is, but it provides no information about the variability of our observations, and variation is what statistics is all about. Pictures tell the story the best.

In Section 1.4, we’ll consider some data on heights I collected while teaching sixth-graders mathematics. The one-way *strip chart* or *dotplot* (Fig. 1.3) created with the aid of Data Desk/XL<sup>2</sup>, an Excel add-in, reveals that the *minimum* of this particular set of data is approximately 137 cm

<sup>2</sup> A trial version may be downloaded from [http://www.datadesk.com/products/data\\_analysis/ddxl/](http://www.datadesk.com/products/data_analysis/ddxl/).



**FIGURE 1.2** A partial list of the functions available in Excel.

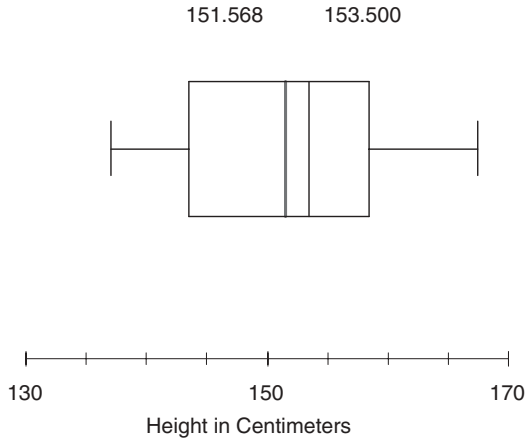


**FIGURE 1.3** One-way strip chart or dotplot.

and the *maximum* approximately 167 cm. Each dot in this strip chart corresponds to an observation. Blotches correspond to multiple observations. The *range* over which these observations extend is 167–137, or 30.

By the way, DataDesk/XL is just one of a hundred or more programs that can add in capabilities to Excel. We'll be using several such add-ins to carry out the necessary calculations to complete this course.

Box plot - Heights of Sixth Graders

**FIGURE 1.4** Box and whiskers plot of classroom data.

A weakness of Fig. 1.3 is that it's hard to tell exactly what the values of the various percentiles are. A glance at the *box and whiskers plot* (Fig. 1.4) created with the aid of *XLStat* (Addinsoft, 2004),<sup>3</sup> a second Excel add-in, tells us that the median of the classroom data described in Section 1.4 is 153.5 cm, the mean is 151.6 cm, and the interquartile range (the “box”) is close to 14 cm. The minimum and maximum of the sample are located at the ends of the “whiskers.”

In Section 1.4, you'll learn how to create these and other graphs.

## 1.4. REPORTING YOUR RESULTS: THE CLASSROOM DATA

Imagine you are in the sixth grade and you have just completed measuring the heights of all your classmates.

Once the pandemonium has subsided, your instructor asks you and your team to prepare a report summarizing your results.

Actually, you have two sets of results. The first set consists of the measurements you made of you and your team members, reported in centimeters, 148.5, 150.0, and 153.0. (Kelly is the shortest, incidentally, and you are the tallest.) The instructor asks you to report the minimum, the

<sup>3</sup> A trial version may be downloaded from <http://www.xlstat.com/download.htm>.

median, and the maximum height in your group. This part is easy, or at least it's easy once you look the terms up in the glossary of your textbook and discover that minimum means smallest, maximum means largest, and median is the one in the middle. Conscientiously, you write these definitions down—they could be on a test.

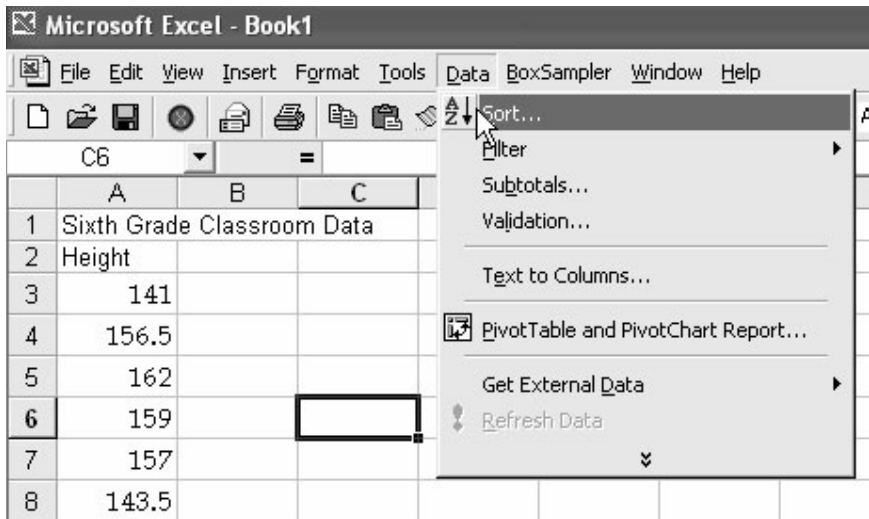
In your group, the minimum height is 148.5 centimeters, the median is 150.0 centimeters, and the maximum is 153.0 centimeters.

Your second assignment is more challenging. The results from all your classmates have been written on the blackboard—all 22 of them.

141, 156.5, 162, 159, 157, 143.5, 154, 158, 140, 142, 150, 148.5,  
138.5, 161, 153, 145, 147, 158.5, 160.5, 167.5, 155, 137

You copy the figures neatly into the first column of an Excel worksheet as described in the previous section. Next, you brainstorm with your teammates. Nothing. Then John speaks up—he's always interrupting in class. Shouldn't we put the heights in order from smallest to largest? "Of course," says the teacher, "you should always begin by ordering your observations."

You go to the Excel menu bar as shown in Fig. 1.5 and access the "sort" command from the "data" menu. As a result, your data are now in sorted in order from smallest to largest:



**FIGURE 1.5** Accessing the sort command.

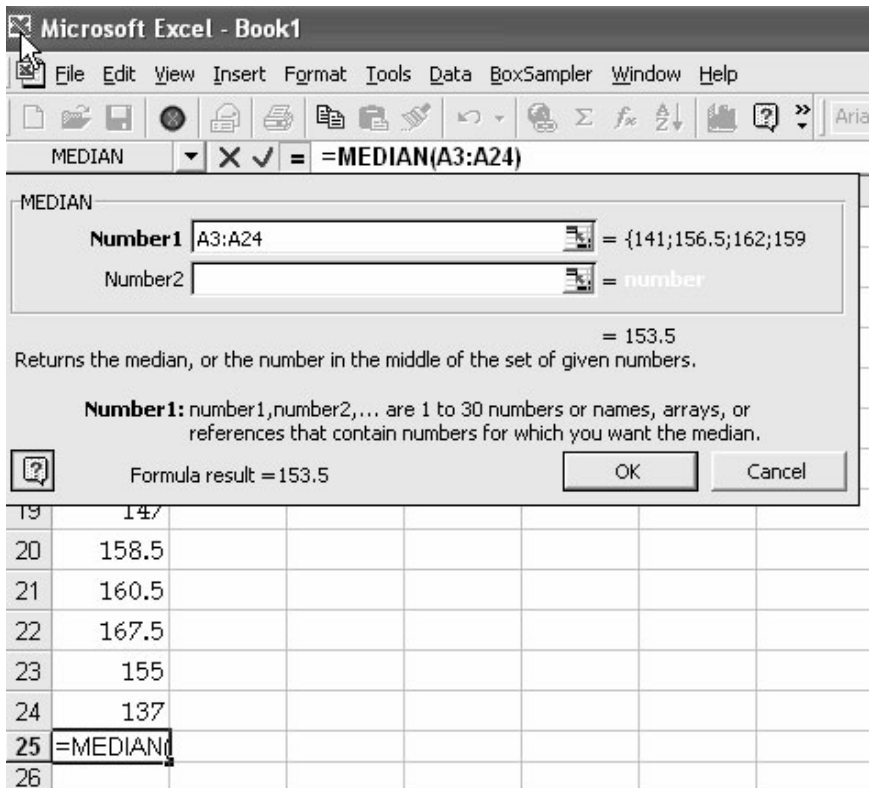
137.0 138.5 140.0 141.0 142.0 143.5 145.0 147.0 148.5 150.0 153.0  
154.0 155.0 156.5 157.0 158.0 158.5 159.0 160.5 161.0 162.0 167.5

“I know what the minimum is,” you say—come to think of it, you are always blurting out in class, too, “137 millimeters, that’s Tony.”

“The maximum, 167.5, that’s Pedro, he’s tall,” hollers someone from the back of the room.

As for the median height, the one in the middle is just 153 centimeters (or is it 154)? What does Excel tell us? As illustrated in Fig. 1.6, we need to do the following to find out:

1. Put our cursor in the first empty cell after the data; A25 in our example.
2. Click the = key on the formula menu bar.
3. Select “median” by using the down arrow ▼ on the formula bar.



**FIGURE 1.6** Computing the median of the classroom data.



4. Use the cursor to select the data range or enter the data range using the form shown in Fig. 1.6 as A3:A24.
5. Press OK.

The result 153.5 will appear in cell A25.

Actually, the median could be any number between 153 and 154, but it is a custom among statisticians, honored by Excel, to report the median as the value midway between the two middle values, when the number of observations is even.

### 1.4.1. Picturing Data

The preceding scenario was a real one. The results reported here, especially the pandemonium, were obtained by my sixth grade homeroom at St. John's Episcopal School in Rancho Santa Margarite, CA. The problem of a metric tape measure was solved by building their own from string and a meter stick.

My students at St. John's weren't through with their assignments. It was important for them to build on and review what they'd learned in the fifth grade, so I had them draw pictures of their data. Not only is drawing a picture fun, but pictures and graphs are an essential first step toward recognizing patterns.

We begin by downloading a trial copy of DataDesk/XL from the website [http://www.datadesk.com/products/data\\_analysis/downloads/ddxl.cfm](http://www.datadesk.com/products/data_analysis/downloads/ddxl.cfm). Note the folder to which you downloaded the program.

To install this add-in, pull down the Excel Tools menu, select "add-ins," and then browse the various folders on the hard disk until you locate the DDXL add-in. Once DDXL is added, a new pull-down menu, labeled DDXL will appear on the menu bar as shown in Fig. 1.7.

After selecting "Charts and Plots" as depicted in Fig. 1.7, we complete the Charts and Plots Dialog shown in Fig. 1.8. Note that among the other possible headings under "Function type" are Box Plot and Histogram.

We click "OK", and Fig. 1.9 reveals the end result. As a by-product, the numeric values of various sample statistics are displayed as well as the dotplot.

**Exercise 1.2.** Generate a dot plot and a box plot for one of the data sets you gathered in your initial assignment. Write down the values of the median, minimum, and maximum that you can infer from the box plot.

### 1.4.2. Displaying Multiple Variables

I'd read, but didn't quite believe, that one's arm span is almost exactly the same as one's height. To test this hypothesis, I had my sixth graders get

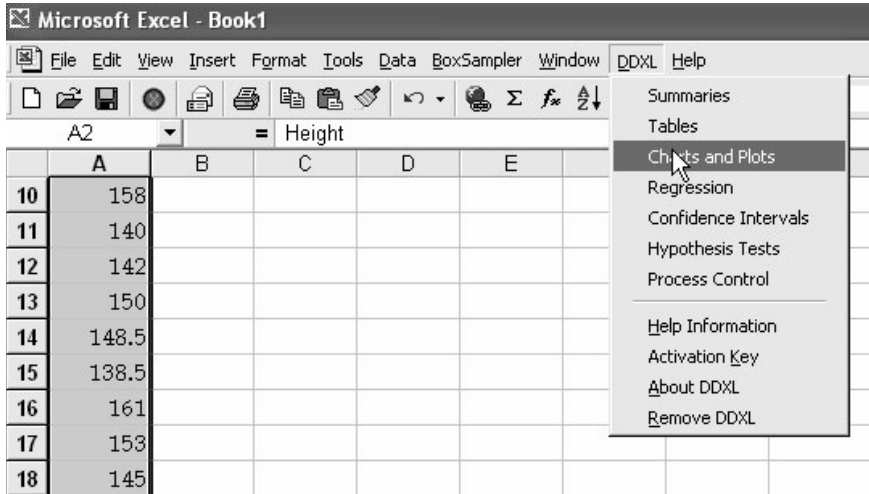


FIGURE 1.7 Selecting charts and plots from the DDXL menu.

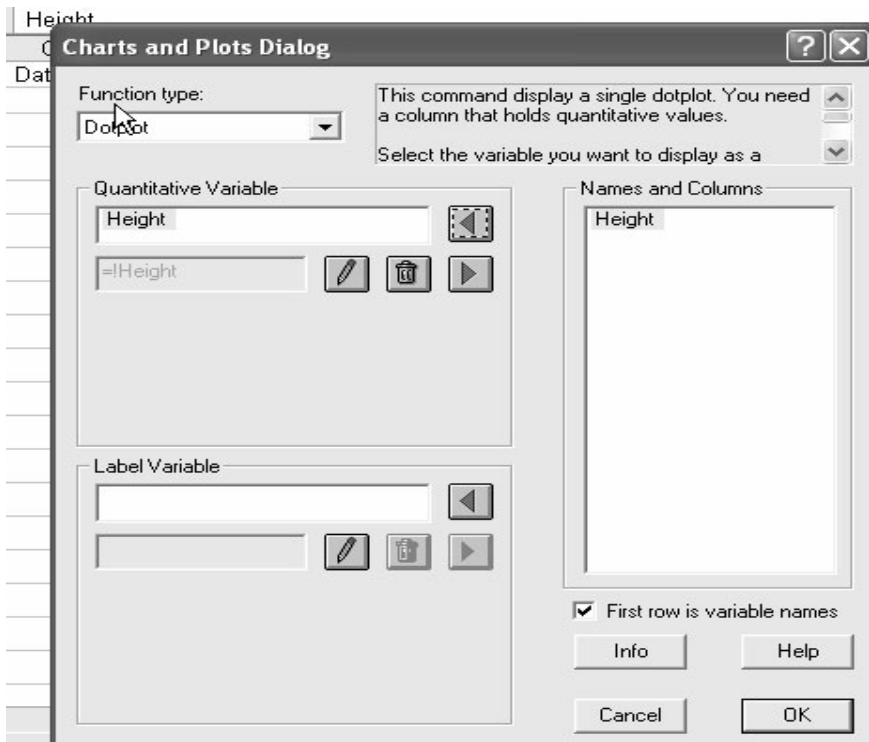
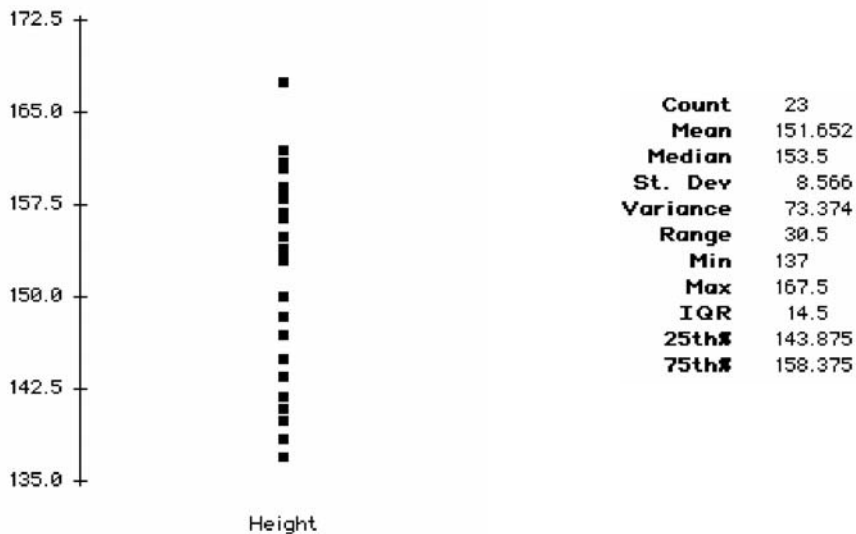


FIGURE 1.8 Selecting the type of graph desired.



**FIGURE 1.9** Dotplot of the classroom height data.

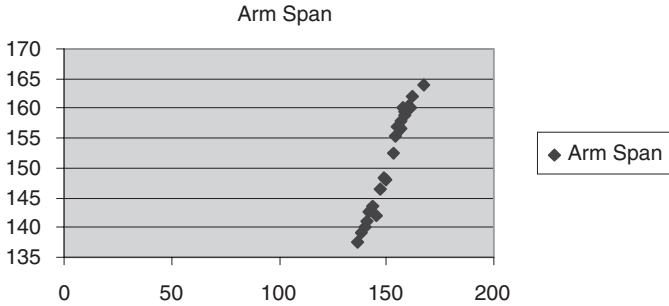
out their tape measures a second time and rule off the distance from the fingertips of the left hand to the fingertips of the right while the student they were measuring stood with arms outstretched like a big bird. After the assistant principal had come and gone (something about how the class was a little noisy, and though we were obviously having a good time, could we just be a little quieter), they recorded their results in the form of a two-dimensional *scatter plot*.

They had to reenter their height data (it had been sorted, remember) and then enter their arm span data :

Height = 141, 156.5, 162, 159, 157, 143.5, 154, 158, 140, 142, 150,  
148.5, 138.5, 161, 153, 145, 147, 158.5, 160.5, 167.5, 155,  
137

Arm span = 141, 156.5, 162, 159, 158, 143.5, 155.5, 160, 140, 142.5,  
148, 148.5, 139, 160, 152.5, 142, 146.5, 159.5, 160.5,  
164, 157, 137.5

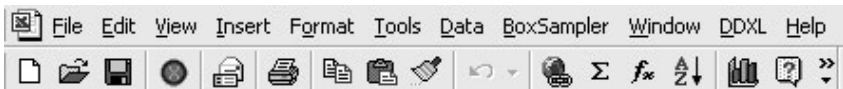
This is trickier than it looks, because unless the data are entered in exactly the same order by student in each data set, the results are meaningless. (We told you that 90% of the problems are in collecting the data and



**FIGURE 1.10** Scatterplot using excel's default settings.

entering it in the computer for analysis. In another text of mine, *A Manager's Guide to The Design and Conduct of Clinical Trials*, I recommend eliminating paper forms completely and entering all data directly into the computer.) Once the two data sets have been read in, creating a scatterplot is easy.

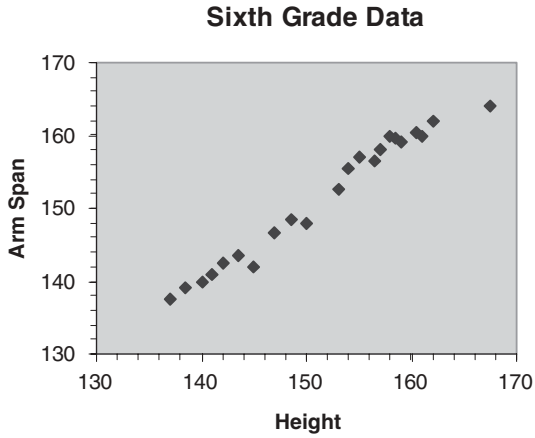
Well, almost easy. The first chart, Fig. 1.10, I created with the Excel Chart menu, next to the question mark, selecting XY(Scatter) and repeatedly pressing Next.



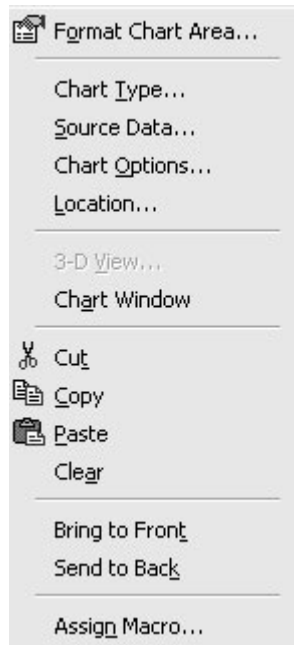
To create Fig. 1.11 from the first scatterplot, I had to complete several steps. Placing my cursor on the chart, and depressing the right mouse button, yielded the menu shown in Fig. 1.12. Clicking on chart options allowed me to enter a title, “Sixth Grade Data” and labels for the  $X$  and  $Y$  axis, “Height” and “Arm Span.”

Escaping from this menu, I put my cursor on the  $X$ -axis and clicked to bring up the menu shown in Fig. 1.13. I changed only one item, setting the Minor tick mark type to “outside.” Then I clicked on the “Scale” tab, removed all the check marks under “Auto,” and put in the values I wanted as shown in Fig. 1.14. I clicked OK to obtain Fig. 1.11.

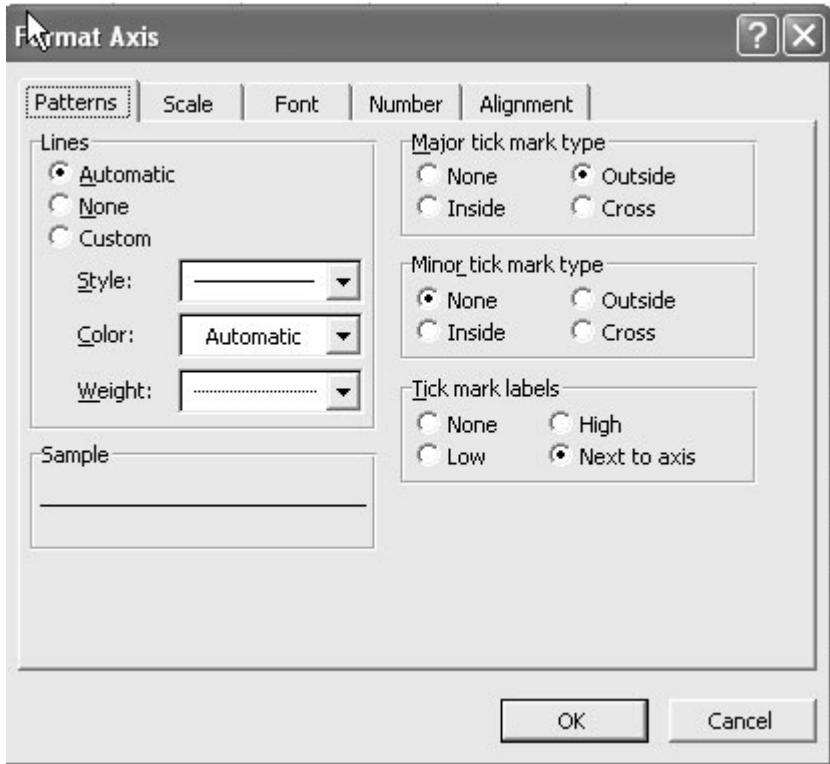
**Exercise 1.3.** Is performance on the LSAT used for law school admission related to one's grade point average? Prepare a scatterplot of the following data drawn from a population of 82 law schools. We'll look at this data again later in this chapter as well as in Chapters 3 and 4.



**FIGURE 1.11** Scatterplot using excel's full capabilities.



**FIGURE 1.12** Chart format menu.



**FIGURE 1.13** Format axis menu.

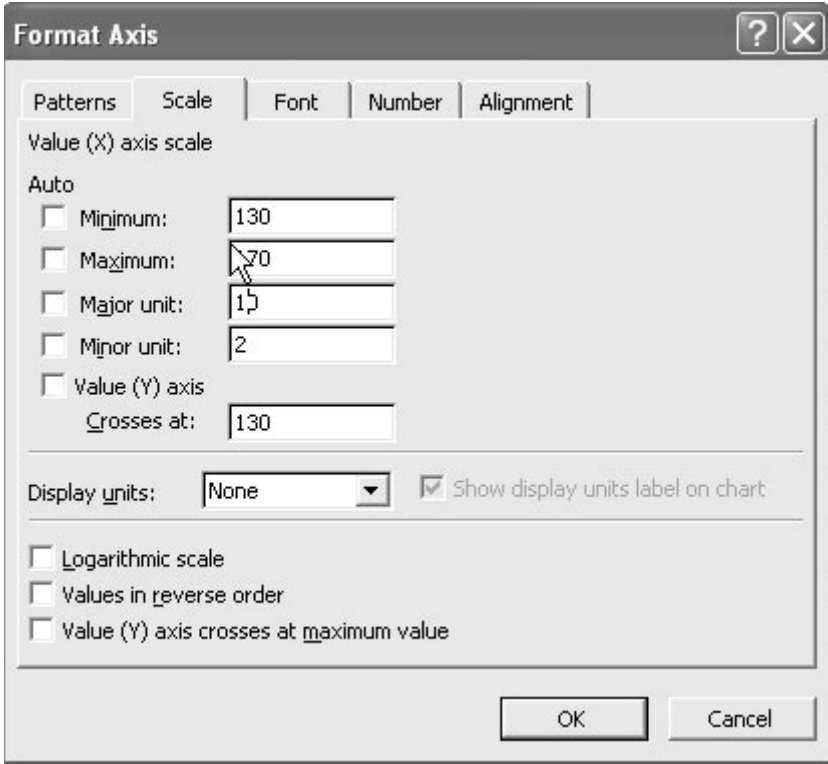
LSAT = 576, 635, 558, 578, 666, 580, 555, 661, 651, 605, 653,  
575, 545, 572, 594

GPA = 3.39, 3.3, 2.81, 3.03, 3.44, 3.07, 3, 3.43, 3.36, 3.13, 3.12,  
2.74, 2.76, 2.88, 2.96

### 1.4.3. Percentiles of the Distribution

The values one reads from a box plot like Fig. 1.4 are approximations. To obtain exact values for the minimum and maximum, you can sort the data as shown in Fig. 1.5. To obtain the values of the median and other percentiles, we would go to Excel's formula bar, choose "Statistical" as our Function category if we have not already done so, and then select "Percentile." The result will be a display similar to Fig. 1.15.

One word of caution: Excel (like most statistics software) yields an excessive number of digits. Because we only measured heights to the nearest centimeter, reporting the 25th percentile as 143.875 would



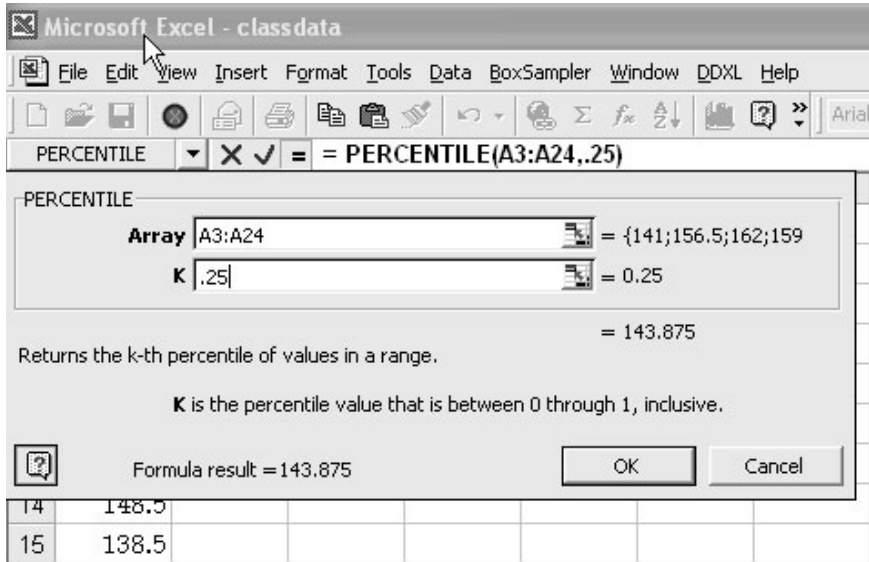
**FIGURE 1.14** Setting up the X-axis for Fig. 1.11.

suggest far more precision in our measurements than actually exists. Report the value 144 centimeters instead.

### PERCENTILES

The 25th percentile of a sample is such that 25% of the observations are smaller in value and 75% are greater. The median or 50th percentile of a sample is such that 50% of the observations are smaller in value and 50% are greater, and so forth. The socially conscious are concerned as much with what the 10th percentile of a population is earning as with what the median income is.

Still another way to display your data is via the *cumulative distribution function*. Begin by sorting the data and then typing the numbers 1, 2, and 3 in Column B opposite the data values as shown in Fig. 1.16. Place your cursor in the first entry in this column (the “1” in B3), hold down your



**FIGURE 1.15** Computing the percentiles of a sample.

	A	B	C
1	Sixth Grade Classroom Data		
2	Height		
3	137	1	
4	138.5	2	
5	140	3	
6	141		

**FIGURE 1.16** The sorted data.

mouse button, and pull the cursor straight down the column, until the numbers 1, 2, and 3 are all highlighted. Release the mouse button. Move your cursor to the lower right corner of B5, until a plus sign appears. Holding down the mouse button, again pull straight down Column B and watch as Excel fills in the numbers 4, 5, . . . , up to 22 (the number of observations) automatically as you pull.

Enter  $= B3/22$  in cell C3, then copy the entry in C3 all the way down the column to C24. The result should look like Fig. 1.17. Note that the entries in Column C are the *cumulative frequencies* of the observations, that is, 0.045 are 137 or less, 0.09 are 138.5 or less, and so forth.



	A	B	C
1	Sixth Grade Classroom Data		
2	Height		
3	137	1	0.045455
4	138.5	2	0.090909
5	140	3	0.136364
6	141	4	0.181818
7	142	5	0.227273
8	143.5	6	0.272727
9	145	7	0.318182
10	147	8	0.363636
11	148.5	9	0.409091
12	150	10	0.454545
13	153	11	0.5

**FIGURE 1.17** Cumulative frequencies.

	A	B	C	D
1	Sixth Grade Classroom Data			
2	Height		Cumulative Frequency	
3	136	0	0	
4	137	1	0.045455	
5	138.5	2	0.090909	
6	140	3	0.136364	

**FIGURE 1.18** Preparing to graph the cumulative frequencies.

The next step in preparing a graph of these cumulative frequencies is to insert an extra row and a column label as shown in Fig. 1.18.

Afterward, highlight the entire region between A2 and C25, select “Charts and Plots” from the DDXL menu, and complete the resulting Charts and Plots Dialog as shown in Fig. 1.19 to obtain the plot of Fig. 1.20.

Note that the  $X$ -axis of the cumulative distribution function extends from the minimum to the maximum value of the class data. The  $Y$ -axis corresponding to the cumulative frequency reveals that the probability that

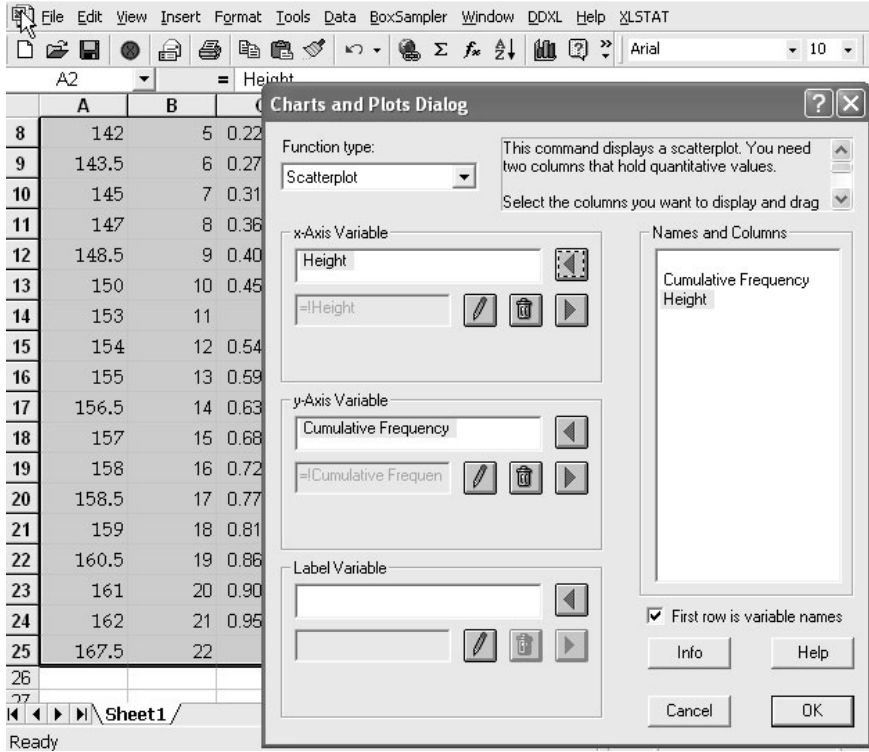


FIGURE 1.19 Plotting the empirical cumulative distribution function.

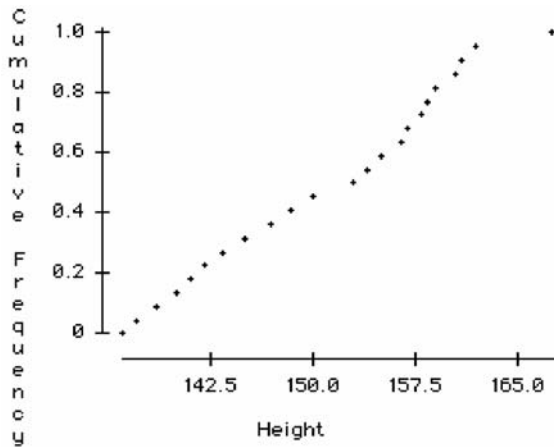


FIGURE 1.20 Cumulative distribution of heights of Dr. Good's sixth-grade class.

a data value is less than the minimum is 0 (you knew that) and the probability that a data value is less than or equal to the maximum is 1. Using a ruler, see what  $X$  value or values correspond to 0.5 on the  $Y$ -scale.

**Exercise 1.4.** What do we call this value(s)?

**Exercise 1.5.** Construct cumulative distribution functions for the data you've collected.

## 1.5. TYPES OF DATA

Statistics such as the minimum, maximum, median, and percentiles make sense only if the data is *ordinal*, that is, if it can be ordered from smallest to largest. Clearly height, weight, number of voters, and blood pressure are ordinal. So are the answers to survey questions such as “How do you feel about President Bush?”

Ordinal data can be subdivided into metric and nonmetric data. *Metric* data like heights and weights can be added and subtracted. We can compute the mean as well as the median of metric data. (We can further subdivide metric data into observations like time that can be measured on a *continuous* scale and counts such as “buses per hour” that are *discrete*.)

But what is the average of “He’s destroying our country” and “He’s no worse than any other politician”? Such preference data is ordinal, in that it may be ordered, but it is *not* metric.

Many times, in order to analyze ordinal data, statisticians will impose a metric on it—assigning, for example, weight 1 to “Bush is destroying our country” and weight 5 to “Bush is no worse than any other politician.” Such analyses are suspect, for another observer using a different set of weights might get quite a different answer.

The answers to other survey questions are not so readily ordered. For example, “What is your favorite color?” Oops, bad example, because we can associate a metric wavelength with each color. Consider instead the answers to “What is your favorite breed of dog?” or “What country do your grandparents come from?” The answers to these questions fall into nonordered categories. Pie charts and bar charts are used to display such categorical data, and contingency tables are used to analyze them. A scatterplot of categorical data would not make sense.

**Exercise 1.6.** For each of the following, state whether the data are metric and ordinal, only ordinal, categorical, or you can’t tell:

- a) Temperature
- b) Concert tickets
- c) Missing data
- d) Postal codes

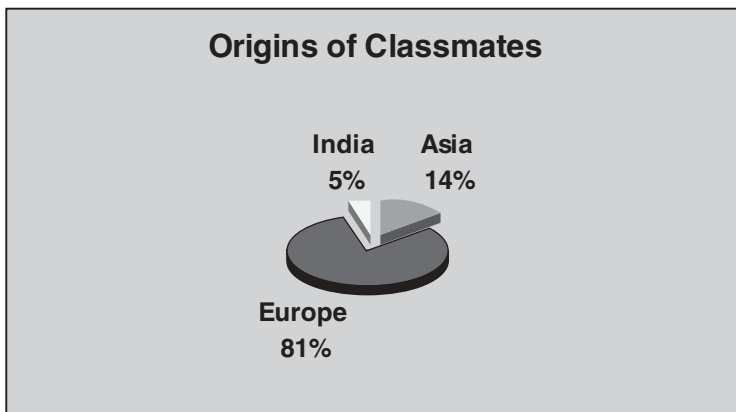
### 1.5.1. Depicting Categorical Data

Three of the students in my class were of Asian origin, 18 were of European origin (if many generations back), and one was part Indian. To depict these categories in the form of a pie chart, I first entered the categorical data Asia, Europe, and India in Column A and the corresponding numbers 3, 18, 1 in Column B.

To obtain the exploded pie chart in Fig. 1.21, I first used my cursor to outline the area on the spreadsheet in which I'd typed my data. I selected the Chart Wizard from Excel's own menu bar, clicked on the Custom Types tab, selected Pie Explosion, and then went step by step through the resulting dialog.

A pie chart also lends itself to the depiction of ordinal data resulting from surveys. If you did a survey as your data collection project, make a pie chart of your results now.

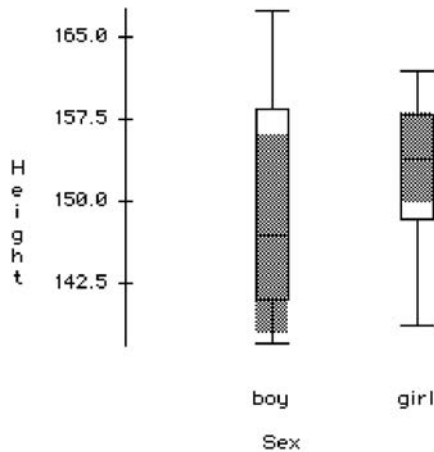
Such plots and charts have several purposes. One is to summarize the data. Another is to compare different samples or different populations (girls versus boys, my class versus your class). For example, we can enter gender data for the students, being careful to enter the gender codes in the same order in which the students' heights and arm spans already have been entered. As shown in Fig. 1.22, the first student on our



**FIGURE 1.21** Region of origin of classmates.

	A	B	C
1	Sixth Grade Classroom Data		
2	Height	Armspan	Sex
3	141	141	boy
4	156.5	156.5	girl
5	162	162	girl
6	159	159	girl
7	157	158	girl
8	143.5	143.5	girl
9	154	155.5	girl
10	158	160	girl
11	140	140	boy
12	142	142.5	girl

**FIGURE 1.22** Classdata by sex of student.



**FIGURE 1.23** Boxplot of class heights by sex.

list is a boy, the next seven are girls, then another boy, six girls, and finally seven boys.

To create the side-by-side boxplots shown in Fig. 1.23, we selected “Boxplot by Groups” from the DDXL Charts and Plots menu.

**Exercise 1.7.** Create a boxplot of arm span by sex for the classdata. Also, create a pie chart by sex for the classdata.

The primary value of charts and graphs is as an aid to critical thinking. The figures in this specific example may make you start wondering about the uneven way in which adolescents go about their growth. The exciting thing, whether you are a parent or a middle-school teacher, is to observe how adolescents get more heterogeneous, more individual with each passing year.

### 1.5.2. From Observations to Questions

You may want to formulate your theories and suspicions in the form of questions: Are girls in the sixth grade taller on the average than sixth-grade boys (not just those in Dr. Good's sixth-grade class, but in all sixth-grade classes)? Are they more homogeneous, that is, less variable, in terms of height? What is the average height of a sixth grader? How reliable is this estimate? Can height be used to predict arm span in sixth grade? Can it be used to predict the arm spans of students of any age?

You'll find straightforward techniques in subsequent chapters for answering these and other questions. First, we suspect, you'd like the answer to one really big question: Is statistics really much more difficult than the sixth-grade exercise we just completed? No, this is about as complicated as it gets.

## 1.6. MEASURES OF LOCATION

Far too often, we find ourselves put on the spot, forced to come up with a one-word description of our results when several pages or, better still, several charts would do. "Take all the time you like," coming from a boss, usually means "Tell me in 10 words or less."

If you were asked to use a single number to describe data you've collected, what number would you use? One answer is "the one in the middle," the *median* that we defined earlier in this chapter.

In the majority of cases, we recommend using the *arithmetic mean* or arithmetic average rather than the median. To calculate the mean of a sample of observations by hand, one adds up the values of the observations, then divides by the number of observations in the sample. If we observe 3.1, 4.5, and 4.4, the arithmetic mean would be  $12/3 = 4$ . In symbols, we write the mean of a sample of  $n$  observations,  $X_i$  with  $i = 1, 2, \dots, n$  as  $(X_1 + X_2 + \dots + X_n)/n = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$ .<sup>4</sup>

$$2, \dots, n \text{ as } (X_1 + X_2 + \dots + X_n)/n = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} .^4$$

<sup>4</sup> The Greek letter  $\Sigma$  is pronounced "sigma".

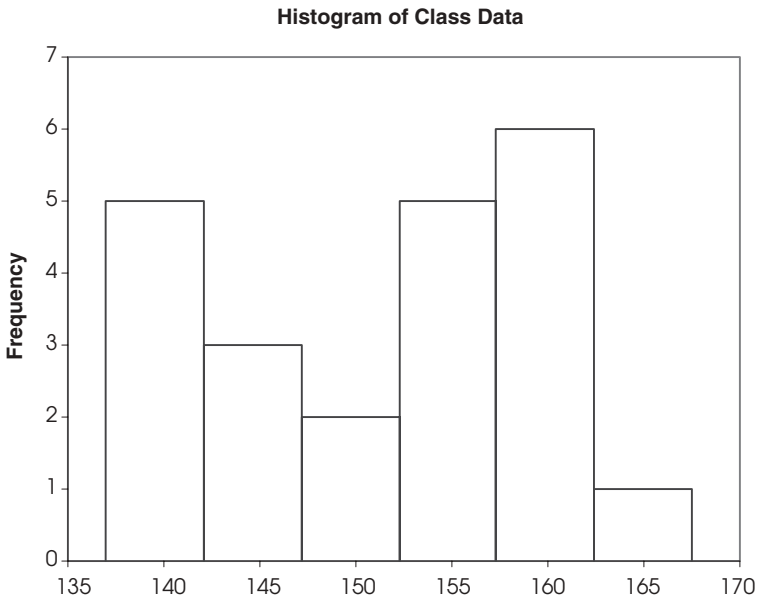
Is adding a set of numbers and then dividing by the number in the set too much work? To find the mean height of the students in my classroom, we would use Excel's average function.

A playground seesaw (or teeter-totter) is symmetric in the absence of kids. Its midpoint or median corresponds to its center of gravity or its mean. If you put a heavy kid at one end and two light kids at the other so that the seesaw balances, the mean will still be at the pivot point, but the median is located at the second kid.

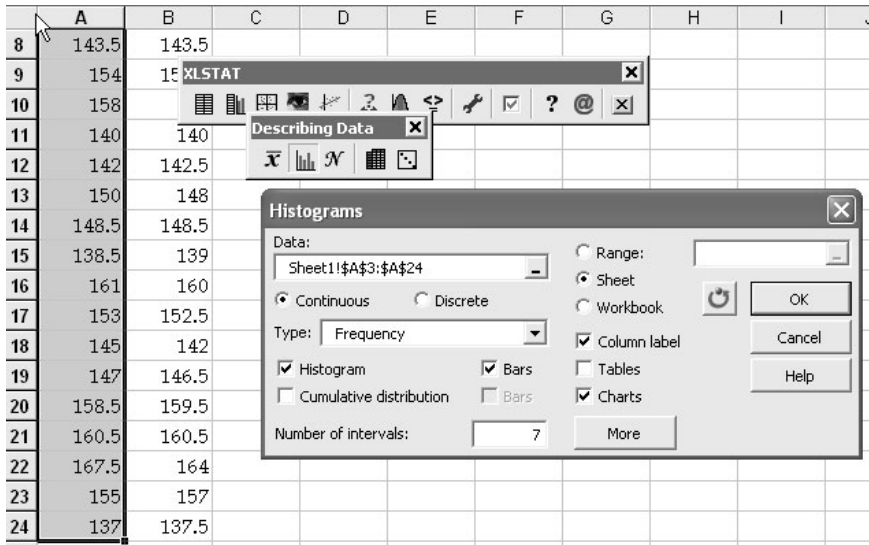
Another population parameter of interest is the most frequent observation or *mode*. In the sample 2, 2, 3, 4 and 5, the mode is 2. Often the mode is the same as the median or close to it. Sometimes it's quite different, and sometimes, particularly when there is a mixture of populations, there may be several modes.

Consider the data on heights collected in my sixth-grade classroom. The mode is at 157.5 cm. But aren't there really two modes, one corresponding to the boys, the other to the girls in the class?

As you can see from Fig. 1.24, a histogram of the heights of my sixth-graders provides evidence of two modes. When we don't know in advance how many subpopulations there are, modes serve a second purpose: to help establish the number of subpopulations.



**FIGURE 1.24** Histogram of class data.



**FIGURE 1.25** Using XLStat to create a histogram from the class heights.

To construct this histogram, I downloaded a trial version of XLStat from <http://www.xlstat.com/index.html> and installed this program after selecting “Add-ins” from Excel’s Tools menu.

As you can see from Fig. 1.25, I selected Describing Data and the Histograms from XLStat’s menu.

**Exercise 1.8.** Compare the mean, median, and mode of the data you’ve collected.

**Exercise 1.9.** A histogram can be of value in locating the modes when there are 20 to several hundred observations, because it groups the data. Draw histograms for the data you’ve collected.

### 1.6.1. Which Measure of Location?

The mean, the median, and the mode are examples of sample statistics. Statistics serve three purposes:

1. Summarizing data
2. Estimating population parameters
3. Aids to decision making

Our choice of one statistic rather than another depends on the use(s) to which it is to be put.



**THE CENTER OF A POPULATION**

**Median:** the value in the middle; the halfway point; that value which has equal numbers of larger and smaller elements around it.

**Arithmetic mean or arithmetic average:** the sum of all the elements divided by their number or, equivalently, that value such that the sum of the deviations of all the elements from it is zero.

**Mode:** the most frequent value. If a population consists of several sub-populations, there may be several modes.

**For summarizing data:** Graphs—boxplots, strip plots, cumulative distribution functions, and histograms—are essential. If you're not going to use a histogram, then for samples of 20 or more be sure to report the number of modes.

We always recommend using the median if the data are ordinal but not metric, as well as when the distribution is highly skewed with a few very large or very small values.

Two good examples of skewness are incomes and house prices. A recent *Los Angeles Times* featured a great house in Beverly Park at \$80 million US. A house like that has a large effect on the mean price of homes in an area. The median house price is far more representative than the mean, even in Beverly Hills.

The weakness of the arithmetic mean is that it is too easily biased by extreme values. If we eliminate Pedro from our sample of sixth graders—he's exceptionally tall for his age at 5'7" or 167 cm—the mean would change from 151.6 to  $3167/21 = 150.8$  cm. The median would change to a much lesser degree, shifting from 153.5 to 153 cm. Because the median is not as readily biased by extreme values, we say that the median is more *robust* than the mean.

**For estimation:** In deciding which *sample statistic* to use in estimating the corresponding *population parameter*, we need to distinguish between precision and accuracy. Let us suppose that Robin Hood and the Sheriff of Nottingham engage in an archery contest. Each is to launch three arrows at a target 50 meters (half a soccer pitch) away. The Sheriff launches first, and his three arrows land one atop the other in a dazzling display of shooting *precision*. Unfortunately, all three arrows penetrate and fatally wound a cow grazing peacefully in the grass nearby. The Sheriff's *accuracy* leaves much to be desired.

We can show mathematically that for very large samples the sample median and the median of the population from which the sample is drawn will almost coincide. The same is true for large samples and the mean. Alas, “large” in this instance may mean larger than we can afford. As you saw in Exercise 1.1, gathering data takes time and money. With small samples, the accuracy of an estimator is always suspect.

With most of the samples we encounter in practice, we can expect the value of the sample median and virtually any other estimator to vary from sample to sample. One way to find out for small samples how *precise* a method of estimation is would be to take a second sample the same size as the first and see how the estimator varies between the two, then a third, and fourth, . . . , say 20 samples. *But a large sample will always yield more precise results than a small one.* So, if we’d been able to afford it, the sensible thing would have been to take 20 times as large a sample to begin with.<sup>5</sup>

Still, there is an alternative. We can treat our sample as if it were the original population and take a series of *bootstrap samples* from it. The variation in the value of the estimator from bootstrap sample to bootstrap sample will be a measure of the variation to be expected in the estimator had we been able to afford to take a series of samples from the population itself. The larger the size of the original sample, the closer it will be in composition to the population from which it was drawn, and the more accurate this measure of precision will be.

### 1.6.2. The Bootstrap

Let’s see how this process, called bootstrapping, would work with a specific set of data. Once again, here are the heights of the 22 students in my sixth-grade class, measured in centimeters and ordered from shortest to tallest:

137.0 138.5 140.0 141.0 142.0 143.5 145.0 147.0 148.5  
150.0 153.0 154.0 155.0 156.6 157.0 158.0 158.5 159.0  
160.5 161.0 162.0 167.5

Let’s assume we record each student’s height on an index card, 22 index cards in all. We put the cards in a big hat, shake them up, pull one out, and make a note of the height recorded on it. *We return the card to the hat* and repeat the procedure for a total of 22 times until I have a second

<sup>5</sup> Of course, there is a point at which each additional observation will cost more than it yields in information. The bootstrap described here will also help us to find the “optimal” sample size.

sample, the same size as the original. Note that we may draw Jane's card several times as a result of using this method of *sampling with replacement*.

Our first bootstrap sample, arranged in increasing order of magnitude for ease in reading, might look like this:

138.5 138.5 140.0 141.0 141.0 143.5 145.0 147.0 148.5 150.0 153.0  
154.0 155.0 156.5 157.0 158.5 159.0 159.0 159.0 160.5 161.0 162.

Several of the values have been repeated; not surprising as we are sampling with replacement, treating the original sample as a stand-in for the much larger population from which the original sample was drawn. The minimum of this bootstrap sample is 138.5, higher than that of the original sample; the maximum at 162.0 is less than the original, whereas the median remains unchanged at 153.5.

137.0 138.5 138.5 141.0 141.0 142.0 143.5 145.0 145.0 147.0  
148.5 148.5 150.0 150.0 153.0 155.0 158.0 158.5 160.5 160.5  
161.0 167.5

In this second bootstrap sample, again we find repeated values; this time the minimum, maximum, and median are 137.0, 167.5, and 148.5, respectively.

Two bootstrap samples cannot tell us very much. But suppose we were to take 50 or 100 such samples. Here is a one-way strip plot of the



medians of 50 bootstrap samples taken from the classroom data:

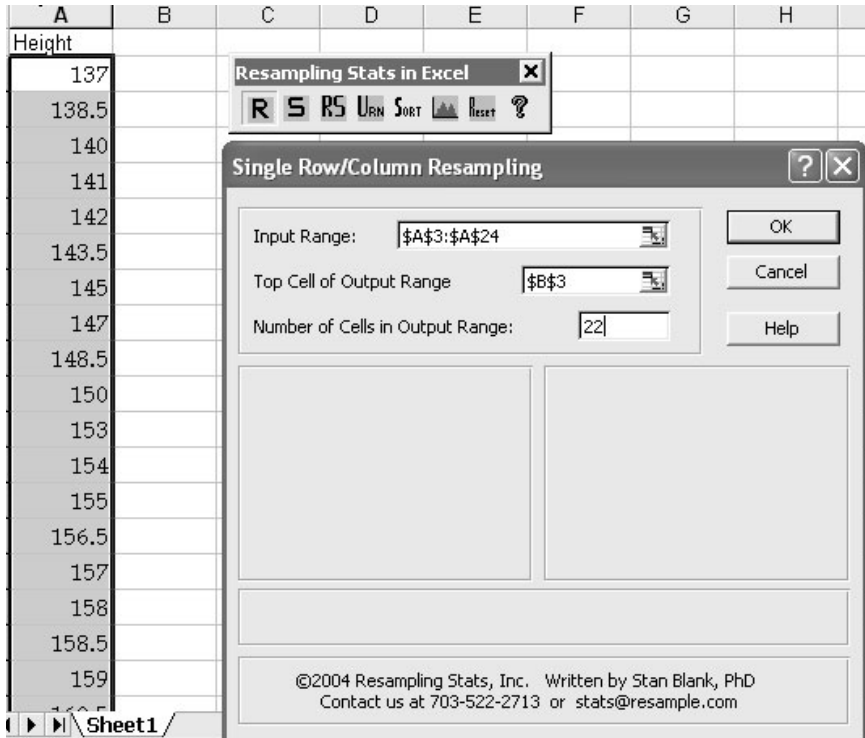
These values provide an insight into what might have been had we sampled repeatedly from the original population.

Quick question: What is that population? Does it consist of all classes at the school where I was teaching? All sixth-grade classes in the district? All sixth-grade classes in the state? The school was Episcopalian, so perhaps the population was all sixth-grade classes in Episcopalian schools.

To apply the bootstrap, you'll need to download and install a trial version of the Resampling Stats in Excel add-in from <http://www.resample.com/content/software/excel/download.shtml>

Before you add it in, make sure that the "Analysis Toolpak" and "Analysis Toolpak VBA" options are checked in Excel's Tools/Add-ins menu.

Clicking on the R on the newly appeared Resampling Stats in Excel menu yields the display of Fig. 1.26. Pressing OK in the dialog box

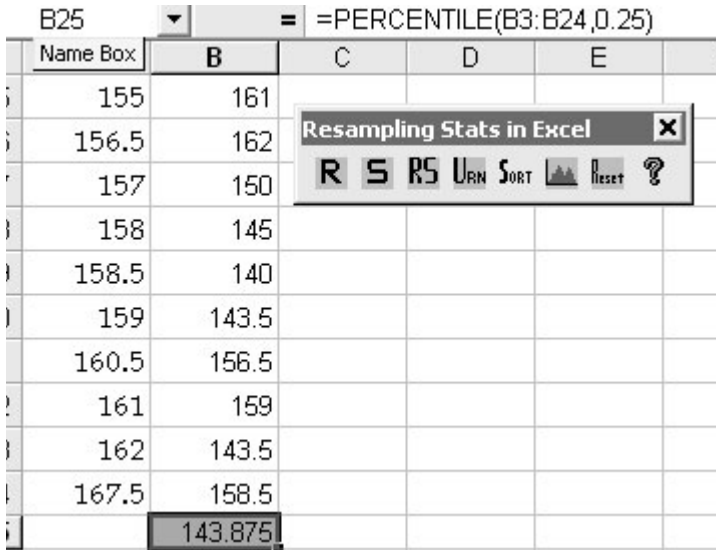


**FIGURE 1.26** Preparing to generate a bootstrap sample.

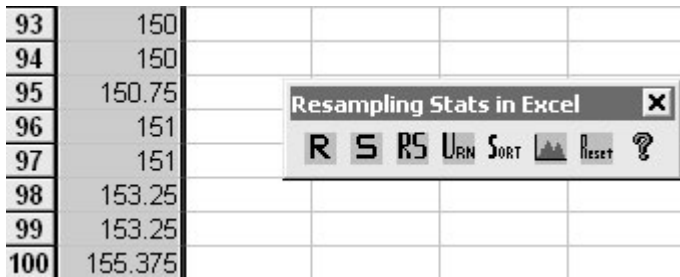
results in a single bootstrap sample (with replacement) in the second column.

To obtain a confidence interval for the 25th percentile of the original sample, I inserted the percentile formula in the first cell immediately beneath the bootstrap sample as in Fig. 1.27. I clicked on the RS on Resampling Stats in Excel menu, and the 25th percentile of each of 100 bootstrap samples was displayed in the first column of a second worksheet, labeled “Results.” To obtain a confidence interval for the original estimate, I sorted the values in the column and then selected the end points of the interval. In Fig. 1.28 we see that in 90 out of 100 instances, the 25th percentile of the bootstrap sample was 150.75 or less.

**Exercise 1.10.** Our original question, you’ll recall, is which is the least variable (most precise) estimate: mean or median? To answer this question, at least for the data on heights I collected in my classroom, apply the bootstrap, then construct side-by-side boxplots for the results.



**FIGURE 1.27** First step in getting a confidence interval for  $P_{25}$ .



**FIGURE 1.28** The eight largest values of the 25th percentile for 100 bootstrap samples.

**Exercise 1.11.** Apply the bootstrap to the data you collected in Exercise 1.1 to see whether the mean or the median is the more precise estimator.

**Exercise 1.12.** Can you tell which is the more accurate estimator in the two previous cases? If not, why not?

## 1.7. SAMPLES AND POPULATIONS

If it weren't for person-to-person variation, it really would be easy to find out what brand of breakfast cereal people prefer or which movie star they

want as their leader. Interrogate the first person you encounter on the street and all will be revealed. As things stand, we must either pay for and take a total census of everyone's view (the cost of the 2003 recall election in California pushed an already near-bankrupt state one step closer to the edge) or take a sample and learn how to extrapolate from that sample to the entire population.

In each of the data collection examples in Section 1.2, our observations were limited to a sample from a population. We measured the height, circumference, and weight of a dozen humans (or dogs, or hamsters, or frogs, or crickets) but not all humans or dogs or hamsters. We timed some individuals (or frogs or turtles) in races but not all. We interviewed some fellow students but not all.

If we had interviewed a different set of students, would we have gotten the same results? Probably not. Would the means, medians, IQRs, and so forth have been similar for the two sets of students? Maybe, if the two samples had been large enough and similar to each other in composition.

If we interviewed a sample of women and a sample of men regarding their views on women's right to choose, would we get similar answers? Probably not, as these samples would be drawn from completely different populations (different, that is, with regard to their views on women's right to choose). If we want to know how the citizenry as a whole feels about an issue, we need to be sure to interview both men and women.

In every statistical study, two questions immediately arise:

1. How large should my sample be?
2. How can I be sure this sample is representative of the population in which my interest lies?

By the end of Chapter 5, we'll have enough statistical knowledge to address the first question, but we can start now to discuss the second.

After I deposited my ballot in a recent election, I walked up to the interviewer from the *Los Angeles Times* who was taking an exit poll and offered to tell her how I'd voted. "Sorry," she said, "I can only interview every ninth person."

What kind of a survey wouldn't want my views? Obviously, a survey that wanted to ensure that shy people were as well represented as boisterous people and that a small group of activists couldn't bias the results.<sup>6</sup>

---

<sup>6</sup> To see how surveys could be biased deliberately, you might enjoy reading Grisham's *The Chamber*.

One sample we would all insist be representative is the jury.<sup>7</sup> The Federal Jury Selection and Service Act of 1968 as revised<sup>8</sup> states that citizens cannot be disqualified from jury duty “on account of race, color, religion, sex, national origin or economic status.”<sup>9</sup> The California Code of Civil Procedure, section 197, tells us *how* to get a representative sample. First, you must be sure your sample is taken from the appropriate population. In the case of California, the “list of registered voters and the Department of Motor Vehicles list of licensed drivers and identification card holders . . . shall be considered inclusive of a representative cross section of the population.” The Code goes on to describe how a table of random numbers or a computer could be used to make the actual selection. The bottom line is that to obtain a random, representative sample:

- Each individual (or item) in the population must have an equal probability of being selected.
- No individual (item) or class of individuals may be discriminated against.

There’s good news and bad news. The bad news is that any individual sample may not be representative. You can flip a coin six times, and every so often it will come up heads six times in a row. A jury may consist entirely of white males. The good news is that as we draw larger and larger samples, samples will resemble the population from which they are drawn more and more closely.

**Exercise 1.13.** For each of the three data collection examples of Section 1.2, describe the populations you would hope to extend your conclusions to and how you would go about ensuring that your samples were representative in each instance.

### 1.7.1. Drawing a Random Sample

Recently, one of our clients asked for help with an audit. Some errors had been discovered in an invoice they’d submitted to the government for reimbursement. Because this client, an HMO, made hundreds of such submissions each month, they wanted to know how prevalent such errors were. Could we help them select a sample for analysis?

<sup>7</sup> Unless, of course, we are the ones on trial.

<sup>8</sup> 28 U.S.C.A. x1861 et. seq (1993).

<sup>9</sup> See 28 U.S.C.A. x1862 (1993).

We could, but we needed to ask the client some questions first. We had to determine what the population was from which the sample would be taken and what constituted a *sampling unit*.

Were we interested in all submissions or just some of them? The client told us that some submissions went to state agencies and some to Federal agencies, but for audit purposes their sole interest was in certain Federal submissions, specifically in submissions for reimbursement for a certain type of equipment. Here, too, a distinction needed to be made between custom equipment (with respect to which there was virtually never an error) and more common off-the-shelf supplies. At this point in the investigation, our client breathed a sigh of relief. We'd earned our fee, it appeared, merely by observing that instead of 10,000 plus potentially erroneous claims, the entire population of interest consisted of only 900 or so items. (When you read earlier that 90% of the effort in statistics was in collecting the data, we meant exactly that.)

Our client's staff, like that of most businesses, was used to working with an electronic spreadsheet. "Can you get us a list of all the files in spreadsheet form?" we asked.

**Files Sorted By Date**

Name	Start Date	
Reed, Agnes	23-Jan-03	0.0055
Ellis, Cynthia	24-Jun-03	0.0991
Wolfe, Carissa	25-Jun-03	0.0173
Rooney, Kevin	9-Jul-03	0.0332
Lane, Lori	18-Jul-03	0.0550
Russo, Will	25-Jul-03	0.1983
Gabel, Steven	28-Jul-03	0.1767
Reed, Oliver	1-Aug-03	0.1913
Huff, Elouise	5-Aug-03	0.0916

They could and did. The first column of the spreadsheet held each claim's ID. The second held the date. We used the spreadsheet's sort function to sort all the claims by date and then deleted all those that fell outside the date range of interest. Next, we inserted a new column and in the top cell (just below the label row) of the new column, we put the command `=rand()`. We copied this command all the way down the column, using Windows' standard cut and paste commands `ctrl-C` and `ctrl-V`.



**Files Included in Initial Audit**

Name	Start Date	rand()
Reed, Agnes	23-Jan-03	0.0055
Hason, Arnold	13-Aug-03	0.0104
Wolfe, Carissa	25-Jun-03	0.0173
Sartre, Jean-Paul	17-Oct-03	0.0222
Brown, James	29-Oct-03	0.0226
Rooney, Kevin	9-Jul-03	0.0332
Mills, Louise	4-Sep-03	0.0412
Smith, Thomas	2-Oct-03	0.0497
Dudley, Morris	8-Aug-03	0.0540

A series of numbers was displayed down the column. To lock these in place, we went to the Tools menu, clicked on “options” and then on the calculation tab. We made sure that Calculation was set to manual and there was no check mark opposite “recalculate before save.”

Now, we resorted the data based on the results of this column. Beforehand, we’d decided there would be exactly 35 claims in the sample, so we simply cut and pasted the top 35 items.

### 1.7.2. Ensuring the Sample is Representative

**Exercise 1.14.** We’ve already noted that a random sample might not be representative. By chance alone, our sample might include men only, or African Americans but no Asians, or no smokers. How would you go about ensuring that a random sample is representative?

## 1.8. VARIATION—WITHIN AND BETWEEN

Our work so far has revealed that the values of our observations vary within a sample as well as between samples taken from the same population. Not surprisingly, we can expect even greater variability when our samples are drawn from different populations. Several different statistics are used to characterize and report on the *within-sample variation*.

The most common statistic is termed the *variance* and is defined as the sum of the squares of the deviations of the individual observations about their mean divided by the sample size minus 1. In symbols, if our observations are labeled  $X_1, X_2,$  up to  $X_n,$  and the mean of these observations is written as  $\bar{X}$ , then the variance  $\sigma^2$  (pronounced sigma squared) is equal to

$$\frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2$$

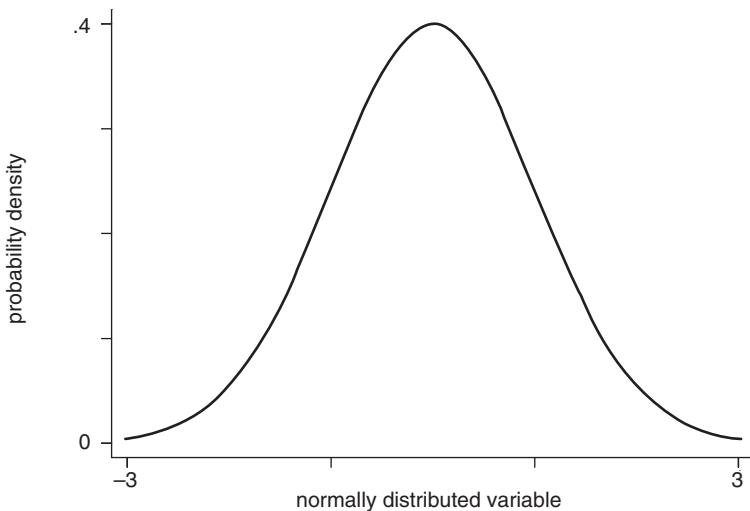
**Exercise 1.15.** What is the sum of the deviations of the observations from their arithmetic mean? That is, what is  $\sum_{i=1}^n (X_i - \bar{X}_i)$ ?

The problem with using the variance is that if our observations, on temperature for example, are in degrees Celsius, then the variance would be expressed in square degrees, whatever these are. More often, we report the *standard deviation*  $\sigma$ , the square root of the variance, as it is in the same units as our observations.

Reporting the standard deviation has the further value that if our observations come from a *normal* distribution like that depicted in Fig. 1.29, then we know that the probability is 68% that an observation taken from such a population lies within plus or minus one standard deviation of the population mean.

If we have two samples and aren't sure whether they come from the same population, one way to check is to express the difference in the sample means, the *between-sample variation*, in terms of the within-sample variation or standard deviation. We'll investigate this approach in Chapter 3.

If the observations do not come from a normal distribution, then the standard deviation is less valuable. In such a case, we might want to report as a measure of dispersion the sample *range*, which is just the maximum minus the minimum, or the *interquartile range*, which is the distance between the 75th and 25th percentiles. From a boxplot of our data, we



**FIGURE 1.29** Bell-shaped symmetric curve of a normally distributed population.

can get eyeball estimates of the range, as the distance from whisker end to whisker end, and the interquartile range, which is the length of the box. Of course, to obtain exact values, we would use R's quantile function.

**Exercise 1.16.** What are the variance, standard deviation, and interquartile range of the classroom data? What are the 90th and 5th percentiles?

This next exercise is only for those familiar with calculus.

**Exercise 1.17.** Show that we can minimize the sum of squares  $\sum_{i=1}^n (X_i - A)^2$  if we let  $A$  be the sample mean.

## 1.9. SUMMARY AND REVIEW

In this chapter, you learned how to do the following:

- Compute mathematical (log, exp, sqrt) and statistical (median, percentile, variance) functions using Excel.
- Create graphs (boxplot, histogram, scatterplot, pie chart, and dotplot).
- Select random samples.

And we showed how to expand Excel's capabilities by downloading and installing add-ins.

The best way to summarize and review the statistical material we've covered so far is with the aid of three additional exercises.

**Exercise 1.18.** Make a list of all the *italicized* terms in this chapter. Provide a definition for each one, along with an example.

**Exercise 1.19.** The following data on the relationship of performance on the LSATs to GPA is drawn from a population of 82 law schools. We'll look at this data again in Chapters 3 and 4.

LSAT = 576, 635, 558, 578, 666, 580, 555, 661, 651, 605, 653,  
575, 545, 574, 594

GPA = 3.39, 3.3, 2.81, 3.03, 3.44, 3.07, 3, 3.43, 3.36, 3.13, 3.12,  
2.74, 2.76, 2.88, 2.96

Make boxplots and histograms for both the LSAT score and GPA. Tabulate the mean, median, interquartile range, standard deviation, and 95th and 5th percentiles for both variables.

**Exercise 1.20.** I have a theory that literally all aspects of our behavior are determined by our birth order (oldest/only, middle, youngest) including clothing, choice of occupation, and sexual behavior. How would you go about collecting data to prove or disprove some aspect of this theory?



# Chapter 2

# Probability

**IN THIS CHAPTER, YOU'LL LEARN THE RULES** of probability and apply them to games of chance, jury selection, surveys, diagnostic tests, and blood types. You'll use R to generate simulated random data and learn how to create your own R functions.

## 2.1. PROBABILITY

Because of the variation inherent in the processes we study, we are forced to speak in probabilistic terms rather than absolutes. We talk about the probability that a sixth-grader is exactly 150 cm tall or, more often, that his height will lie between two values such as 150 cm and 155 cm. The events we study may happen a large proportion of the time, or “almost always,” but seldom “always” or “never.”

Rather arbitrarily, and some time ago, it was decided that probabilities would be assigned a value between 0 and 1, that events that were certain to occur would be assigned probability 1, and that events that would “never” occur would be given probability 0. When talking about a set of *equally likely* events, such as the probability that a fair coin will come up heads, or an unweighted die will display a “6,” this limitation makes a great deal of sense. A coin has two sides; we say the probability it comes up heads is a half and the probability of tails is a half also:  $\frac{1}{2} + \frac{1}{2} = 1$ , the probability that a coin comes up something.<sup>1</sup> Similarly, the probability that a six-sided die displays a “6” is  $1/6$ . The probability it does *not* display a 6 is  $1 - \frac{1}{6} = \frac{5}{6}$ .

---

<sup>1</sup> I had a professor at Berkeley who wrote a great many scholarly articles on the subject of “coins that stand on edge,” but then that is what professors at Berkeley do.

For every dollar you bet, roulette wheels pay off \$36 if you win. This certainly seems fair, until you notice that not only does the wheel have slots for the numbers 1 through 36, but there is a slot for 0, and sometimes for double 0, and for triple 000 as well. Thus the real probabilities of winning and losing are, respectively, 1 chance in 39 and 38/39. In the long run, you lose one dollar thirty-eight times as often as you win \$36. Even when you win, the casino pockets your dollar, so that in the long run the casino pockets \$3 for every \$39 that is bet. (And from whose pockets does that money come?)

Ah, but you have a clever strategy called a *martingale*. Every time you lose, you simply double your bet. So if you lose a dollar the first time, you lose two dollars the next. Hmm. As the casino always has more money than you do, you still end up broke. Tell me again why this is a clever strategy.

**Exercise 2.1.** List the possible ways in which the following can occur:

- a) A person, call him Bill, is born on a specific day of the week.
- b) Bill and Alice are born on the same day of the week.
- c) Bill and Alice are born on different days of the week.
- d) Bill and Alice play a round of a game called “paper, scissor, stone” and simultaneously display an open hand, two fingers, or a closed fist.

**Exercise 2.2.** Match the probabilities with their descriptions. A description may match more than one probability.

- |              |                         |
|--------------|-------------------------|
| a) -1        | 1) infrequent           |
| b) 0         | 2) virtually impossible |
| c) 0.10      | 3) certain to happen    |
| d) 0.25 inch | 4) typographical error  |
| e) 0.50      | 5) more likely than not |
| f) 0.80      | 6) certain              |
| g) 1.0       | 7) highly unlikely      |
| h) 1.5       | 8) even odds            |
|              | 9) highly likely        |

To determine whether a gambling strategy or a statistic is optimal, we need to know a few of the laws of probability. These laws show us how to determine the probabilities of combinations of events. For example, if the probability that an event  $A$  will occur is  $P\{A\}$ , then the probability that  $A$  won't occur  $P\{\text{not}A\} = 1 - P\{A\}$ . This makes sense because either the event  $A$  occurs or it doesn't, and thus  $P\{A\} + P\{\text{not}A\} = 1$ .

We'll also be concerned with the probability that both A and B occur,  $P\{A \text{ and } B\}$ , or with the probability that either A occurs or B occurs or both do,  $P\{A \text{ or } B\}$ . If two events A and B are *mutually exclusive*, that is, if when one occurs the other cannot possibly occur, then the probability that A **or** B will occur,  $P\{A \text{ or } B\}$ , is the sum of their separate probabilities. (Quick, what is the probability that both A **and** B occur.) The probability that a six-sided die will show an odd number is thus  $\frac{3}{6}$  or  $\frac{1}{2}$ . The probability that a six-sided die will *not* show an even number is equal to the probability that a six-sided die will show an odd number.

### 2.1.1. Events and Outcomes

An *outcome* is something we can observe, for example, “the coin lands heads” or “an odd number appears on the die.” Outcomes are made up of *events* that may or may not be completely observable. The referee tosses the coin into the air; it flips over three times before he catches it and places it face upward on his opposite wrist. “Heads,” and Manchester United gets the call. But the coin might also have come up heads had the coin been tossed higher in the air so that it spun three and a half or four times before being caught. A literal infinity of events makes up the single observed outcome, “Heads.”

The outcome “an odd number appears on the six-sided die” is composed of three outcomes, 1, 3, and 5, each of which can be the result of any of an infinity of events. By definition, events are mutually exclusive. Outcomes may or may not be mutually exclusive, depending on how we aggregate events.

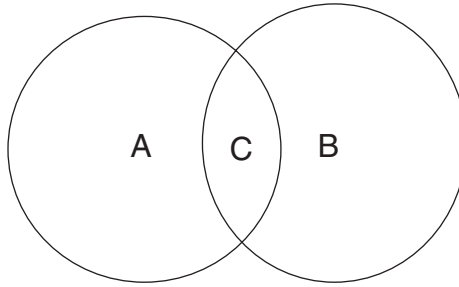
### 2.1.2. Venn Diagrams

An excellent way to gain insight into the distinction between events and outcomes and the laws of probability is via the Venn diagram.<sup>2</sup> Figure 2.1 pictures two overlapping outcomes, A and B. For example, A might consist of all those who respond to a survey that they are nonsmokers, while B corresponds to the outcome that the respondent has lung cancer.

Every point in the figure corresponds to an event. The events within the circle A all lead to the outcome A. Note that many of the events or points in the diagram lie outside both circles. These events correspond to the outcome “neither A nor B” or, in our example, “an individual who does smoke and does not have lung cancer.”

<sup>2</sup> Curiously, not a single Venn diagram is to be found in John Venn's text, *The Logic of Chance*, published by Macmillan and Co, London, 1866, with a third edition in 1888.





**FIGURE 2.1** Venn diagram depicting two overlapping outcomes.

The circles overlap; thus outcomes A and B are not mutually exclusive. Indeed, any point in the region of overlap between the two, marked C, leads to the outcome “A and B.” What can we say about individuals who lie in region C?

**Exercise 2.3.** Construct a Venn diagram corresponding to the possible outcomes of throwing a six-sided die. (I find it easier to use squares than circles to represent the outcomes, but the choice is up to you.) Does every event belong to one of the outcomes? Can an event belong to more than one of these outcomes? Now, shade the area that contains the outcome “the number face up on the die is odd.” Use a different shading to outline the outcome “the number on the die is greater than 3.”

**Exercise 2.4.** Are the outcomes “the number face up on the die is odd” and “the number on the die is greater than 3” mutually exclusive?

You’ll find many excellent Venn diagrams illustrating probability concepts at <http://stat-www.berkeley.edu/~stark/Java/Venn.htm>.

**Exercise 2.5.** According to the *Los Angeles Times*, scientists are pretty sure planetoid Sedna has a moon, although as of April 2004 they’d been unable to see it. The scientists felt at the time there was a 1 in 100 possibility that the moon might have been directly in front of or behind the planetoid when they looked for it, and a 5 in 100 possibility that they’d misinterpreted Sedna’s rotation rate. How do you think they came up with those probabilities?

**IN THE LONG RUN: SOME MISCONCEPTIONS**

When events occur as a result of chance alone, anything can happen and usually will. You roll craps 7 times in a row, or you flip a coin 10 times and 10 times it comes up heads. Both these events are unlikely, but they are not impossible. Before reading the balance of this section, test yourself by seeing if you can answer the following:

You've been studying a certain roulette wheel that is divided into 38 sections for over 4 hours now, and not once during those 4 hours of continuous play has the ball fallen into the number 6 slot. Which of the following do you feel is more likely?

- (1) Number 6 is bound to come up soon.
- (2) The wheel is fixed so that number 6 will never come up.
- (3) The odds are exactly what they've always been, and in the next 4 hours number 6 will probably come up about 1/38th of the time.

If you answered (2) or (3) you're on the right track. If you answered (1), think about the following equivalent question:

You've been studying a series of patients treated with a new experimental drug, all of whom died in excruciating agony despite the treatment. Do you conclude the drug is bound to cure somebody sooner or later and take it yourself when you come down with the symptoms? Or do you decide to abandon this drug and look for an alternative?

**2.2. BINOMIAL**

Many of our observations take a yes/no or dichotomous form: "My headache did/didn't get better." "Chicago beat/was beaten by Los Angeles." "The respondent said he would/wouldn't vote for Dean." The simplest example of a *binomial trial* is that of a coin flip: Heads I win, tails you lose.

If the coin is fair, that is, if the only difference between the two mutually exclusive outcomes lies in their names, then the probability of throwing a head is  $\frac{1}{2}$ , and the probability of throwing a tail is also  $\frac{1}{2}$ . (That's what I like about my bet, either way I win.)

By definition, the probability that something will happen is 1 and the probability that nothing will occur is 0. All other probabilities are somewhere in between.<sup>3</sup>

<sup>3</sup> If you want to be precise, the probability of throwing a head is probably only 0.49999, and the probability of a tail is also only 0.49999. The leftover probability of 0.00002 is the probability of all the other outcomes—the coin stands on edge, a sea gull drops down out of the sky and takes off with it, and so forth.

What about the probability of throwing heads twice in a row? Ten times in a row? If the coin is fair and the throws independent of one another, the answers are easy:  $1/4$ th and  $1/1024$ th or  $(1/2)^{10}$ .

These answers are based on our belief that when the only differences among several possible *mutually exclusive* outcomes are their labels, “heads” and “tails,” for example, the various outcomes will be *equally likely*. If we flip two fair coins or one fair coin twice in a row, there are four possible outcomes: HH, HT, TH, and TT. Each outcome has equal probability of occurring. The probability of observing the one outcome in which we are interested is 1 in 4 or  $1/4$ th. Flip the coin 10 times and there are  $2^{10}$  or a thousand possible outcomes; one such outcome might be described as HTTTTTTTTTH.

Unscrupulous gamblers have weighted coins so that heads comes up more often than tails. In such a case, there is a real difference between the two sides of the coin and the probabilities will be different from those described above. Suppose as a result of weighting the coin, the probability of getting a head is now  $p$ , where  $0 \leq p \leq 1$ , and the complementary probability of getting a tail (or not getting a head) is  $1 - p$ , because  $p + (1 - p) = 1$ . Again, we ask the question, What is the probability of getting two heads in a row? The answer is  $p^2$ . Here is why:

To get two heads in a row, we must throw a head on the first toss, which we expect to do in a proportion  $p$  of attempts. Of this proportion, only a further fraction  $p$  of two successive tosses also end with a head, that is, only  $p \times p$  trials result in HH. Similarly, the probability of throwing 10 heads in a row is  $p^{10}$ .

By the same line of reasoning, we can show that the probability of throwing nine heads in a row followed by a tail when we use the same weighted coin each time is  $p^9(1 - p)$ . What is the probability of throwing 9 heads in 10 trials? Is it also  $p^9(1 - p)$ ? No, for the outcome “nine heads out of ten” includes the case where the first trial is a tail and all the rest are heads, the second trial is a tail and all the rest are heads, the third trial is . . . , and so forth, 10 different ways in all. These different ways are *mutually exclusive*, that is, if one of these events occurs, the others are excluded. The probability of the overall event is the sum of the individual probabilities, or  $10 p^9(1 - p)$ .

#### RULES OF PROBABILITY

- The probability that one of several mutually exclusive events will occur is the sum of the individual probabilities.
- The probability that a series of independent events will occur is the product of the individual probabilities.

**Exercise 2.6.** What is the probability that if you flipped a fair coin you would get heads five times in a row?

**Exercise 2.7.** Suppose the incidence of individuals infected with tuberculosis on an Indian reservation was 10%. Suppose we test 100 individuals on the reservation for TB, using a test that was known to be 100% accurate for infected individuals but also yielded positive and erroneous results for noninfected individuals 10% of the time. How many of these 100 individuals would you expect to test positive for TB?

**Exercise 2.8.** The strength of support for our candidate seems to depend on whether we are interviewing men or women: 50% of male voters support our candidate, but only 30% of women. What percentage of women favor some other candidate? If we select a woman and a man at random and ask which candidate they support, in what percentage of cases do you think both will say they support our candidate?

**Exercise 2.9.** Would your answer to the last question in Exercise 2.8 be the same if the man and the woman were co-workers?

**Exercise 2.10.** Which do you think would be preferable in a customer-satisfaction survey? To ask customers if they were or were not satisfied? Or to ask them to specify their degree of satisfaction on a 5-point scale? Why?

### 2.2.1. Permutations and Rearrangements

What is the probability of throwing exactly 5 heads in 10 tosses of a coin? The answer to this last question requires we understand something about permutations and combinations or rearrangements, a concept that will be extremely important in succeeding chapters.

Suppose we have three horses in a race. Call them A, B, and C. A could come in first, B could come in second, and C would be last. ABC is one possible outcome or permutation. But so are ACB, BAC, BCA, CAB, CBA, six possibilities or *permutations* in all. Now suppose we have a nine-horse race. We could write down all the possibilities, or we could use the following trick: We choose a winner (nine possibilities); we choose a second-place finisher (eight remaining possibilities), and so forth until all positions are assigned. A total of  $9! = 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$  possibilities in all. Had there been  $N$  horses in the race, there would have been  $N!$  possibilities.  $N!$  is read “ $N$  factorial.”

Note that  $N! = N(N-1)!$ .

Normally in a horse race, all our attention is focused on the first three finishers. How many possibilities are there? Using the same reasoning, it is easy to see there are  $9 \times 8 \times 7$  possibilities or  $9!/6!$ . Had there been  $N$  horses in the race, there would have been  $N!/(N-3)!$  possibilities.

Suppose we ask a slightly different question: In how many different ways can we select three horses from nine entries without regard to order (that is, we don't care which comes first, which second, or which third)? In the previous example, we distinguished between first-, second-, and third-place finishers; now we're saying the order of finish doesn't make any difference. We already know there are  $3! = 3 \times 2 \times 1 = 6$  different permutations of the three horses that finish in the first three places. So we take our answer to the preceding question  $9!/6!$  and divide this answer in turn by  $3!$ . We write the result as  $\binom{9}{3}$ , which is usually read as 9 choose 3.

Note that  $\binom{9}{3} = \binom{9}{6}$ .

In how many different ways can we assign nine cell cultures to two unequal experimental groups, one with three cultures and one with six? This would be the case if we had nine labels and three of the labels read "vitamin E" while six read "controls." If we could distinguish the individual labels, we could assign them in  $9!$  different ways. But the order they are assigned in each of the experimental groups,  $3!$  ways in the first instance and  $6!$  in the other, won't affect the results. Thus there are only  $9!/6!3!$  or  $\binom{9}{3} = 84$  distinguishable ways. We can generalize this result to show that the number of distinguishable ways  $N$  items can be assigned to two groups, one of  $k$  items and one of  $N-k$  is  $\frac{N!}{k!(N-k)!} = \binom{N}{k}$ .

What if we were to divide these same nine cultures among three equal-sized experimental groups? Then we would have  $9!/3!3!3!$  distinguishable ways or *rearrangements*, written as  $\binom{9}{3 \ 3 \ 3}$ .

**Exercise 2.11.** What is the value of 4!?

**Exercise 2.12.** In how many different ways can we divide eight subjects into two equal-sized groups? Use the Excel formula =COMBIN(8,4).

**Exercise 2.13.** In how many different ways can we choose 5 from 10 things?

### 2.2.2. Back To the Binomial

We used horses in an example in the previous section, but the same reasoning can be applied to coins or survivors in a clinical trial.<sup>4</sup> What is the probability of five heads in 10 tosses? What is the probability that five of 10 breast cancer patients will still be alive after six months?

We answer this question in two stages. First, what is the number of different ways we can get five heads in 10 tosses? We could have thrown HHHHHTTTTT or HHHHTHTTTT, or some other combination of five heads and five tails for a total of 10 choose 5 or  $10!/(5!5!)$  ways. The probability the first of these events occurring—five heads followed by five tails—is  $(\frac{1}{2})^{10}$ . Combining these results yields

$$\Pr \{5 \text{ heads in } 10 \text{ throws of a fair coin}\} = \binom{10}{5} \left(\frac{1}{2}\right)^{10}$$

We can generalize the preceding to an arbitrary probability of success  $p$ ,  $0 \leq p \leq 1$ . The probability of failure is  $1 - p$ . The probability of  $k$  successes in  $n$  trials is given by the binomial formula

$$\binom{n}{k} (p)^k (1-p)^{n-k} \text{ for } 0 \leq k \leq n.$$

**Exercise 2.14.** What is the probability of getting at least one head in six flips of a fair coin? (Hint: Think negatively.)

### 2.2.3 The Problem Jury

At issue in *Ballew v. Georgia*<sup>5</sup> brought before the Supreme Court in 1978 was whether the all-white jury in Ballew's trial represented a denial of Ballew's rights.<sup>6</sup> In the 1960s and 1970s, United States courts held uniformly that the use of race, gender, religion, or political affiliation to bar citizens from jury service would not be tolerated. In one case in 1963 in which I assisted the defense on appeal, we were able to show that only one black had served on some 163 consecutive jury panels. In this case, we were objecting—successfully—to the methods used to select the jury.

<sup>4</sup> If, that is, the probability of survival is the same for every patient. When there are obvious differences from trial to trial—for example, one subject is an otherwise healthy 35-year old male and the other an elderly 89-year old who has just recovered from pneumonia this simple binomial model would not apply.

<sup>5</sup> 435 U.S. 223, 236–237 (1978)

<sup>6</sup> Strictly speaking, it is not the litigant but the potential juror whose rights might have been interfered with. For more on this issue, see Chapter 2 of *Applying Statistics in the Courtroom*, Phillip Good, Chapman and Hall, 2001.

In *Ballew*, the defendant was not objecting to the methods but to the composition of the specific jury that had judged him at trial.

In the district in which Ballew was tried, blacks comprised 10% of the population, but Ballew's jury was entirely white. Justice Blackmun wanted to know what the probability was that a jury of 12 persons selected from such a population in accordance with the law would fail to include members of the minority.

If the population in question is large enough, say a hundred thousand or so, we can assume that the probability of selecting a nonminority juryperson is a constant 90 out of 100. The probability of selecting two nonminority persons in a row according to the product rule for independent events is  $.9 \times .9$  or  $.81$ . Repeating this calculation 10 more times, once for each of the remaining 10 jurypersons, we get a probability of  $.9 \times .9 \times \dots \times .9 = 0.28243$ , or 28%.

Not incidentally, Justice Blackmun made exactly this same calculation and concluded that Ballew had not been denied his rights.

#### 2.2.4. Properties of the Binomial

Suppose we sent out several hundred individuals to interview our customers and find out whether they are satisfied with our products. Each individual had the responsibility of interviewing exactly 10 customers. Collating the results, we observed several things:

- 740 out of every 1000 customers reported they were satisfied.
- Results varied from interviewer to interviewer.
- About 6% of the samples included no dissatisfied customers.
- A little more than 2% of the samples included 6 or more dissatisfied customers.
- The median number of satisfied customers per sample was 7.
- The modal number of satisfied customers per sample was 8.

When we reported these results to our boss, she only seemed interested in the first of them. "Results always vary from interviewer to interviewer, from sample to sample. And the percentages you reported, apart from the 74% satisfaction rate, are immediate consequences of the binomial distribution."

Clearly, our boss was familiar with the formula for  $k$  successes in  $n$  trials given in Section 2.2.2. From our initial finding, she knew that  $P = 0.74$ . Thus,

$$\Pr\{k \text{ satisfied customers in sample of } 10\} = \binom{10}{k} (.74)^k (.26)^{n-k} \text{ for } 0 \leq k \leq 10.$$

To find the median of this distribution, go to any vacant cell on the spreadsheet and type = BinomDist( to bring up the menu shown in Fig. 2.2.

By entering a series of successively larger guesses 5, 6, and then 7 in the Number\_s space, I was able to determine that the median (the 50th percentile) was 7.

The proportion of samples with *no* dissatisfied customers is the same as the percentage of samples *all* of whose customers are satisfied. To determine the probability of such an outcome, I filled out the BinomDist menu as shown in Fig. 2.3.

BINOMDIST

Number_s	7	= 7
Trials	10	= 10
Probability_s	0.74	= 0.74
Cumulative	True	= TRUE

Returns the individual term binomial distribution probability.

Number\_s is the number of successes in trials.

Formula result =0.504219988

OK Cancel

FIGURE 2.2 Excel's BinomDist menu.

BINOMDIST

Number_s	10	= 10
Trials	10	= 10
Probability_s	0.74	= 0.74
Cumulative	False	= FALSE

Returns the individual term binomial distribution probability.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability mass function, use FALSE.

Formula result =0.049239904

OK Cancel

FIGURE 2.3 Finding the probability of a specific binomial outcome.



To find the proportion of samples with four or less satisfied customers, set Cumulative to True and Number\_s to 4.

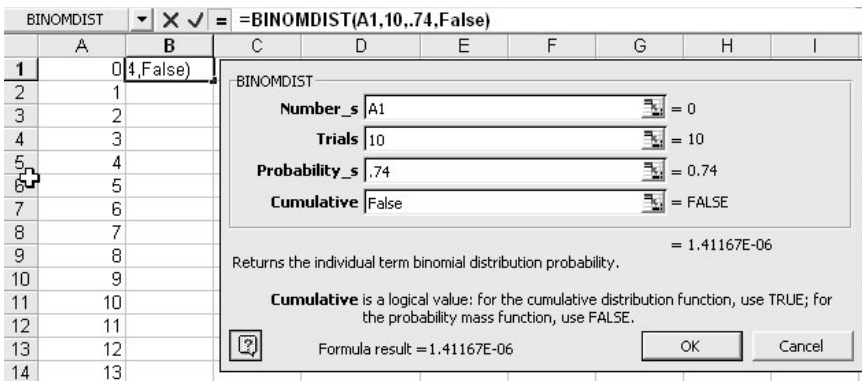
To find the mean or *expected value* of this binomial distribution, let us first note that the computation of the arithmetic mean can be simplified when there are a large number of ties by multiplying each distinct number  $k$  in a sample by the frequency  $f_k$  with which it occurs;  $\bar{X} = \sum_k k f_k$ . We can have only 11 possible outcomes as a result of our interviews: 0, 1, . . . , or 10 satisfied customers. We know from the binomial distribution the frequency  $f_i$  with which each outcome may be expected to occur; the population mean is given by the formula  $\sum_{i=0}^{10} i \binom{10}{i} (p)^i (1-p)^{10-i}$ .

To let Excel make the calculations for us, proceed as follows:

1. Enter the numbers 0 through 10 in the first column.
2. Enter the probability of zero successes in the first cell of the second column as shown in Fig. 2.4.
3. Copy this cell down the second column.
4. In the first cell of the third column, enter the cross product = A1\*B1
5. To find the mean, sum the products in the third column = SUM(C1:C11) or 7.4.

This result, we notice, is equal to 10\*0.74 and, more generally, the expected value of the binomial distribution is equal to the product of the sample size and the probability of success at each trial.

**Warning:** In the preceding example, we assumed that our sample of 1000 customers was large enough that we could use the proportion of successes in that sample, 740 out of 1000, as if it were the true proportion in the entire distribution of customers. Because of the variation



**FIGURE 2.4** Preparing to calculate the mean of a binomial distribution.

inherent in our observations, the true proportion might have been greater or less than our estimate.

**Exercise 2.15.** Which is more likely, observing two or more successes in 8 trials with a probability of one-half of observing a success in each trial, or observing three or more successes in 7 trials with a probability of 0.6 of observing a success. Which set of trials has the greater expected number of successes?

**Exercise 2.16.** Show without using algebra that if  $X$  and  $Y$  are independent identically distributed binomial variables  $B(n, p)$ , then  $X + Y$  is distributed as  $B(2n, p)$ .

Unless we have a large number of samples, the observed or *empirical distribution* may differ radically from the expected or theoretical distribution. To generate random samples from a binomial distribution, we need to download and install BoxSampler, an Excel add-in, from the website <http://www.introductorystatistics.com/escout/tools/boxsampler.htm>.

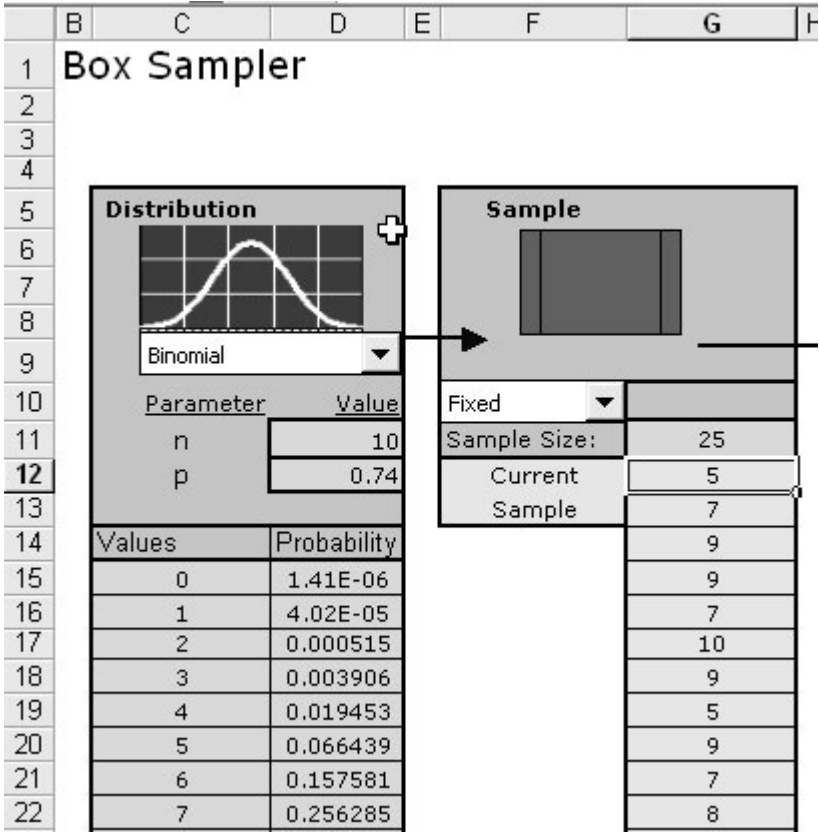
To assist you in using the program, you'll find full documentation at <http://www.introductorystatistics.com/escout/BSHelp/Main.htm>.

Let me walk you through the steps for downloading and installation.

1. Once on the website, click on the appropriate "Click Here."
2. Download to any convenient folder. But be sure to write down the location where you download the file!!
3. Go to this folder when downloading is complete, and click on BoxSamplerInstall. You'll be asked a series of questions, but all are straightforward and in most cases simply clicking on "Next" will be sufficient. Make a note of the folder in which the program and associated files are being installed!!
4. Bring up Excel. Click Tools on the menu bar, then select Add-Ins/Browse. Locate the folder with the BoxSampler program and add it in.
5. The BoxSampler menu should now appear on your Excel menu bar. (If not, you need to go to your Windows program menu, find Box Sampler, and read the BoxSampler Installation Document.)

Once BoxSampler is installed, we can proceed to generate samples from a binomial distribution as follows:

1. Choose "New Model" from the BoxSampler menu and specify "Distribution" as the model type.
2. Once the BoxSampler worksheet is displayed, set Distribution to Binomial,  $n$  to 10, and  $p$  to 0.74 as shown in Fig. 2.5. Set the Sample Size to 25.



**FIGURE 2.5** Sampling from a binomial frequency distribution.

3. Click the solid arrow ► on the simulator menu



to display both the complete frequency distribution (cells C15 to C25) and the results of 25 samples from that distribution (cells G12 through G36).

**Exercise 2.17.** Generate 100 random samples of 10 binomial trials where each trial has a probability 0.6 of success. Construct a column chart of the results, using Excel’s Chart Wizard.

### 2.2.5. Multinomial

Suppose now reporters were to take a survey before an election in which multiple candidates were competing for the same office. The reporters wouldn’t just be interested in whether or not votes were going to be cast for our candidate (a binomial) but which candidate the votes were going

to go to (a *multinomial*). A proportion  $p_i$  of the population intends to vote for the  $i$ th candidate where  $\sum_i p_i = 1$ . The reporter is going to use the frequencies  $\{f_i\}$  he observes in his survey to estimate the unknown population proportions  $\{p_i\}$ .<sup>7</sup>

In another application of the multinomial, we might want to do a survey of consumers and have them try washing with our soap. Afterwards, we would ask them to state their degree of satisfaction on a 5-point scale and, at the same time, state their degree of satisfaction with their present soap. With the comparative data in hand, we could create side-by-side bar charts of the two sets of preferences to use in our advertising.

### 2.3. CONDITIONAL PROBABILITY

Conditional probability is one of the most difficult of statistical concepts, not so much to understand as to accept in all its implications. Recall that mathematicians arbitrarily assign a probability of 1 to the result that something will happen—the coin will come up heads or tails—and 0 to the probability that nothing will occur. But real life is more restricted: A series of past events has preceded our present, and every future outcome is conditioned on this past. Consequently, we need a method whereby the probabilities of just the remaining possibilities sum to 1.

We define the *conditional probability* of an event A given another event B, written  $P(A|B)$ , to be the ratio  $P(A \text{ and } B)/P(B)$ . To show how this would work, suppose we are playing craps, a game in which we throw two six-sided die. Clearly, there are  $6 \times 6 = 36$  possible outcomes. One (and only one) of these 36 outcomes is snake eyes, a 1 and a 1.

Now, suppose we throw one die at a time (a method that is absolutely forbidden in any real game of craps, whether in the Bellagio or a back alley) and a 1 appears on the first die. The probability that we will now roll snake eyes, that is, that the second die will reveal a 1 also, is 1 out of 6 possibilities or  $(\frac{1}{36})/(\frac{1}{6}) = \frac{6}{36} = \frac{1}{6}$ .

The conditional probability of rolling a total of 7 spots on the two dice is  $\frac{1}{6}$ . And the conditional probability of the spots on the two die summing to 11, another winning combination, is 0. Yet before we rolled the two dice, the unconditional probability of rolling snake eyes was 1 out of 36 possibilities and the probability of 11 spots on the two die was  $\frac{2}{36}$ th (a 5 and a 6 or a 6 and a 5).

Now, suppose I walk into the next room where I have two decks of cards. One is an ordinary deck of 52 cards, half red and half black. The

<sup>7</sup> The choice of letter used for the index is unimportant.  $\sum_i p_i$  means the same as  $\sum_k p_k$ .

other is a trick deck in which all the spots on the cards are black. I throw a coin—I'm still in the next room so you don't get to see the result of the coin toss—and if the coin comes up heads I stick the trick deck in my pocket, otherwise I take the normal deck. Now, I come back into the room and offer to show you a card chosen at random from the deck in my pocket. The card has black spots. Would you like to bet on whether or not I'm carrying the trick deck?

[STOP: Think about your answer before reading further.]

Common sense would seem to suggest that the odds are still only 50-50 that it's the trick deck I'm carrying. You didn't really learn anything from seeing a card that could have come from either deck. Or did you?

Let's use our conditional probability relation to find out whether the odds have changed. First, what do we know? As the deck was chosen at random, we know that the probability of the card being drawn from the trick deck is the same as the probability of it being drawn from the normal one:

$$P(T^c) = P(T) = \frac{1}{2}.$$

Here,  $T$  denotes the event that I was carrying a trick deck and  $T^c$  denotes the complementary event that I was carrying the normal deck.

We also know two conditional probabilities. The probability of drawing a black card from the trick deck is, of course, 1 while that of drawing a black card from a deck that has equal numbers of black and red cards is  $\frac{1}{2}$ . In symbols,  $P(B|T) = 1$  and  $P(B|T^c)$  is  $\frac{1}{2}$ .

What we'd like to know is whether the two conditional probabilities  $P(T|B)$  and  $P(T^c|B)$  are equal. We begin by putting the two sets of facts we have together, using our conditional probability relation,  $P(B|T) = P(T \text{ and } B)/P(T)$ .

We know two of the values in the first relation,  $P(B|T)$  and  $P(T)$ , and so we can solve for  $P(B \text{ and } T) = P(B|T) P(T) = 1 \times \frac{1}{2}$ . Similarly,  $P(B \text{ and } T^c) = P(B|T^c) P(T^c) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ .

Take another look at our Venn diagram in Fig. 2.1. All the events in outcome  $B$  are either in  $A$  or in its complement  $A^c$ . Similarly  $P(B) = P(B \text{ and } T) + P(B \text{ and } T^c) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$ .

We now know all we need to know to calculate the conditional probability  $P(T|B)$ , for our conditional probability relation can be rearranged to interchange the roles of the two outcomes, giving  $P(T|B) = P(B \text{ and } T)/P(B) = \frac{1}{2} / \frac{3}{4} = \frac{2}{3}$ . By definition  $P(T^c|B) = 1 - P(T|B) = \frac{1}{3} < P(T|B)$ .

The odds have changed. Before I showed you the card, the probability of my showing you a black card was  $1 \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2}$  or  $\frac{3}{4}$ . When I

showed you a black card, the probability it came from a black deck was  $\frac{1}{2}$  divided by  $\frac{3}{4}$  or  $\frac{2}{3}$ !

**Exercise 2.18.** If R denotes a red card, what would be  $P(T|R)$  and  $P(T^c|R)$ ?

#### A TOO-REAL EXAMPLE

Think the previous example was artificial? That it would never happen in real life? My wife and I just came back from a car trip. On our way up the coast, I discovered that my commuter cup leaked, but, desperate for coffee, I wrapped a towel around the cup and persevered. Not in time, my wife noted, pointing to the stains on my jacket.

On our way back down, I lucked out and drew the cup that didn't leak. My wife congratulated me on my good fortune and then, ignoring all she might have learned had she read this text, proceeded to drink from the remaining cup! So much for her new Monterey Bay Aquarium sweat shirt.

### 2.3.1. Market Basket Analysis<sup>8</sup>

Many supermarkets collect data on purchases with bar code scanners located at the checkout counter. Each transaction record lists all items bought by a customer in a single purchase transaction. Executives want to know whether certain groups of items are consistently purchased together. They use this data for adjusting store layouts (placing items optimally with respect to each other), for cross-selling, for promotions, for catalog design, and to identify customer segments based on buying patterns.

If a supermarket database has 100,000 point-of-sale transactions, out of which 2000 include both items A and B and 800 of these include item C, the *association rule* “If A and B are purchased then C is purchased on the same trip” has a *support* of 800 transactions (alternatively  $0.8\% = 800/100,000$ ) and a *confidence* of 40% ( $=800/2000$ ).

**Exercise 2.19.** Suppose you have the results of a market basket analysis in hand. a) If you wanted an estimate of the probability that a customer will purchase anchovies, would you use the support or the confidence? b) If you wanted an estimate of the probability that a customer carrying

<sup>8</sup> In Section 7.8, we make use of a data mining procedure to do a market basket analysis when there are hundreds of items to choose from.

anchovies will also purchase hot dogs, would you use the support or the confidence?

### 2.3.2. Negative Results

Suppose you were to bet on a six-horse race in which the horses carried varying weights on their saddles. As a result of these handicaps, the probability that a specific horse will win is exactly the same as that of any other horse in the race. What is the probability that your horse will come in first?

Now suppose, to your horror, a horse other than the one you bet on finishes first. No problem; you say, “I bet on my horse to place,” that is, you bet it would come in first *or* second. What is the probability you still can collect on your ticket? That is, what is the conditional probability of your horse coming in second when it did not come in first?

One of the other horses did finish first, which leaves five horses still in the running for second place. Each horse, including the one you bet on, has the same probability to finish second, so the probability you can still collect is one out of five. Agreed?

Just then, you hear the announcer call out that the horses are about to line up for the second race. Again there are six horses and each is equally likely to finish first. What is the probability that if you bet on a horse to place in the second race you will collect on your bet? Is this  $\frac{1}{6} + \frac{1}{5}$ ?

There are three ways we can arrive at the correct answer when all horses are equally fast:

1. We could notice that the probability that your horse will finish second is exactly the same as the probability that it will finish first (or the probability that it will finish dead last, for that matter). As these are mutually exclusive outcomes, their probabilities may be added. The probability is  $\frac{2}{6}$  that your horse finishes first or second.
2. We could list all  $6!$  mutually exclusive outcomes of the race and see how many would lead to our collecting on our bet—but this would be a lot of work.
3. Or we could trace the paths that lead to the desired result. For example, either your horse comes in first, with probability  $\frac{1}{6}$ , or it does not, with probability  $\frac{5}{6}$ . If it doesn't come in first, it might still come in second, with probability  $\frac{1}{5}$ . The overall probability of your collecting on your bet is  $\Pr\{\text{your horse wins}\} + \Pr\{\text{your horse doesn't win}\} * \Pr\{\text{your horse is first among the nonwinning horses}\} = \frac{1}{6} + \frac{5}{6} * \frac{1}{5} = \frac{2}{6}$ .

**Exercise 2.20.** Suppose 10 people are in a class. What is the probability that no two of them were born on the same day of the week? What is the

probability that all of them were born in different nonoverlapping four-week periods? Hint: Write down some of the possibilities, Sam—Monday, Bill—Tuesday, and so forth.

**Exercise 2.21\*.** A spacecraft depends on five different mission-critical systems. If any of these systems fail, the flight will end in catastrophe. Taken on an individual basis, the probability that a mission-critical system will fail during the flight is  $\frac{1}{10}$ . a) What is the probability that the flight will be successful?

NASA decides to build in redundancies. Every mission-critical system has exactly one back-up system that will take over in the event that the primary system fails. The back-up systems have the same probability of failure as the primaries. b) What is the probability that the flight will be successful?

**Exercise 2.22.** A woman sued a Las Vegas casino alleging the following: She asked a security guard to hold her slot machine while she hit the buffet; he let somebody else use “her” machine while she was gone; that “somebody else” hit the jackpot; that jackpot was rightfully hers. The casino countered that jackpots were triggered by a random clock keyed to the  $\frac{1}{1000}$ th of a second; thus, even had the woman been playing the machine continuously, she might not have hit the jackpot. How would you rule if you were a judge?

**Exercise 2.23.** In the U.S. in 1985, there were 2.1 million deaths from all causes, compared to 1.7 million in 1960. Does this mean it was safer to live in the U.S. in the '60s than in the '80s?

**Exercise 2.24\*.** You are a contestant on “Let’s Make a Deal.” Monty offers you a choice of three different curtains and tells you there is a brand new automobile behind one of them plus enough money to pay the taxes in case you win the car. You tell Monty you want to look behind curtain number 1. Instead, he throws back curtain number 2 to reveal . . . a child’s toy. “Would you like to choose curtain number 3 instead of number 1?” Monty asks. Well, would you?

## 2.4. INDEPENDENCE

A key element in virtually all the statistical procedures we will consider in this text is that the selection of one member of a sample takes place



*independently* of the selection of another. In discussing the game of craps, we assumed that the spots displayed on the first die were *independent* of the spots displayed on the second. When statistics are used, we:

1. Assume observations are independent.
2. Test for independence.
3. Try to characterize the nature of the dependence (Chapter 7).

Two events or observations are said to be independent of one another providing that knowledge of the outcome or value of the one gives you no information regarding the outcome or value of the other.

In terms of conditional probabilities, two events A and B are independent of one another providing that  $P(A|B) = P(A)$ , that is, our knowledge that B occurred does not alter the likelihood of A. We can use this relation to show that if A and B are independent, then the probability they will *both* occur is the product of their separate probabilities,  $P(A \text{ and } B) = P(A) \cdot P(B)$ , for from the definition of conditional probability,  $P(A \text{ and } B) = P(A) \cdot P(A \text{ and } B|A) = P(A) \cdot P(B|A) = P(A) \cdot P(B)$ .

**Warning:** Whether events are independent of one another will depend upon the context. Imagine that three psychiatrists interview the same individual, who we shall suppose is a paranoid schizophrenic. The interviews take place at different times, and the psychiatrists are not given the opportunity to confer with each other either before or after the interviews take place.

Suppose now that these psychiatrists are asked for their opinions on i) the individual's sanity, and, having been informed of the patient's true condition, ii) their views on paranoid schizophrenia. In the first case, their opinions will be independent of one another; in the second case, they will not.

**Exercise 2.25.** Can two independent events be mutually exclusive?

**Exercise 2.26.** Draw a Venn diagram depicting two independent events one of which is twice as likely to occur as the other.

**Exercise 2.27.** Do the following constitute independent observations?

- A. Several students sitting together at a table asked who their favorite movie actress is
- B. The number of abnormalities in each of several tissue sections taken from the same individual

- C. Opinions of several individuals whose names you obtained by sticking a pin through a phone book, and calling the “pinned” name on each page
- D. Opinions of an ardent Democrat and an ardent Republican
- E. Today’s price in Australian dollars of the Euro and the Japanese yen.

**Exercise 2.28.** On the basis of the results in the following *contingency tables*, would you say that sex and survival are independent of one another in Table A? In Table B?

**Table A**

	Alive	Dead
Men	15	5
Women	15	10

**Table B**

	Alive	Dead
Men	15	10
Women	15	8

**Exercise 2.29.** Provide an example in which an observation  $X$  is independent of the value taken by an observation  $Y$ ,  $X$  is independent of a third observation  $Z$ , and  $Y$  is independent of  $Z$ , but  $X$ ,  $Y$ , and  $Z$  are dependent.

## 2.5. APPLICATIONS TO GENETICS

All the information needed to construct an organism, whether a pea plant, a jellyfish, or a person, is encoded in its genes. Each gene contains the information needed to construct a single protein. Each of our cells has two copies of each gene, one obtained from our mother and one from our father. We will contribute just one of these copies to each of our offspring. Whether it is the copy we got from our father or the one from our mother is determined entirely by chance.

You could think of this as flipping a coin: One side says “mother’s gene,” the other side says “father’s gene.” Each time a sperm is created in our testis or an ovum in our ovary, the coin is flipped.

There may be many forms of a single gene; each such form is called an allele. Some alleles are defective, incapable of constructing the necessary protein. For example, my mother was  $rh^-$ , meaning that both her copies of the  $rh$  gene were incapable of manufacturing the  $rh$  protein that is found in red blood cells. This also means that the copy of the  $rh$  gene I obtained from my mother was  $rh^-$ . But my blood tests positive for the  $rh$  protein, which means that the  $rh$  gene I got from my father was  $rh^+$ .

**Exercise 2.30.** The mother of my children was also  $rh^-$ . What proportion of our children would you expect to be  $rh^-$ ?

**Exercise 2.31.** Sixteen percent of the population of the United States are rh<sup>-</sup>. What percentage do you expect to have at least one rh<sup>-</sup> gene? (Remember, as long as a person has even one rh<sup>+</sup> gene, they can manufacture the rh protein.)

The gene responsible for making the A and B blood proteins has three alleles, A, B, and O. A person with two type O alleles will have blood type O. A person with one A allele and one B allele will have blood type AB. Only 4% of the population of the United States have this latter blood type.

Our genes are located on chromosomes. The chromosomes come in pairs, one member of each pair being inherited from the father and one from the mother. Your chromosomes are passed onto the offspring independently of one another.

**Exercise 2. 32.** The ABO and rh genes are located on different chromosomes. What percentage of the population of the United States would you expect to have the AB rh<sup>+</sup> blood type?

**Exercise 2. 33.** Forty-five percent of the population of the United States have type O blood. That is, they do not test positive for either the A or the B protein. What percentage of the population do you expect to have at least one O allele?

## 2.6. SUMMARY AND REVIEW

In this chapter, we introduced the basics of probability theory and independence, considered the properties of a *discrete* probability distribution, the binomial, and applied the elements of probability to genetics. We learned how to use the BoxSampler add-in to generate random samples from various distributions.

**Exercise 2.34.** Make a list of all the italicized terms in this chapter. Provide a definition for each one, along with an example.

**Exercise 2.35.** (Read and reread carefully before even attempting an answer.) A magician has three decks of cards, one with only red cards, one that is a normal deck, and one with only black cards. He walks into an adjoining room and returns with only a single deck. He removes the top card from the deck and shows it to you. The card is black. What is the probability that the deck from which the card came consists only of black cards?

**Exercise 2.36.** An integer number is chosen at random. What is the probability that it is divisible by 2? What is the probability that it is divisible by 17? What is the probability that it is divisible by 2 and 17? What is the probability that it is divisible by 2 or 17? (Hint: A Venn diagram would be a big aid in solving this last part.)

**Exercise 2.37.** Pete, Phil, and Myron are locked in a squash court after hours with only a Twinkie and a coin between them. The only thing all three can agree on is that they want a whole Twinkie or nothing. Myron suggests that Pete and Phil flip the coin, and that the winner flips a coin with him to see who gets the Twinkie. Phil, who is a graduate student in statistics, says this is unfair. Is it unfair and why? How would you decide who gets the Twinkie?



# Chapter 3

## Distributions

**IN THIS CHAPTER, YOU'LL LEARN TO RECOGNIZE** and describe the probability distributions of numerical observations made on random selections from a population. You'll learn methods for estimating the parameters of these distributions and for testing hypotheses.

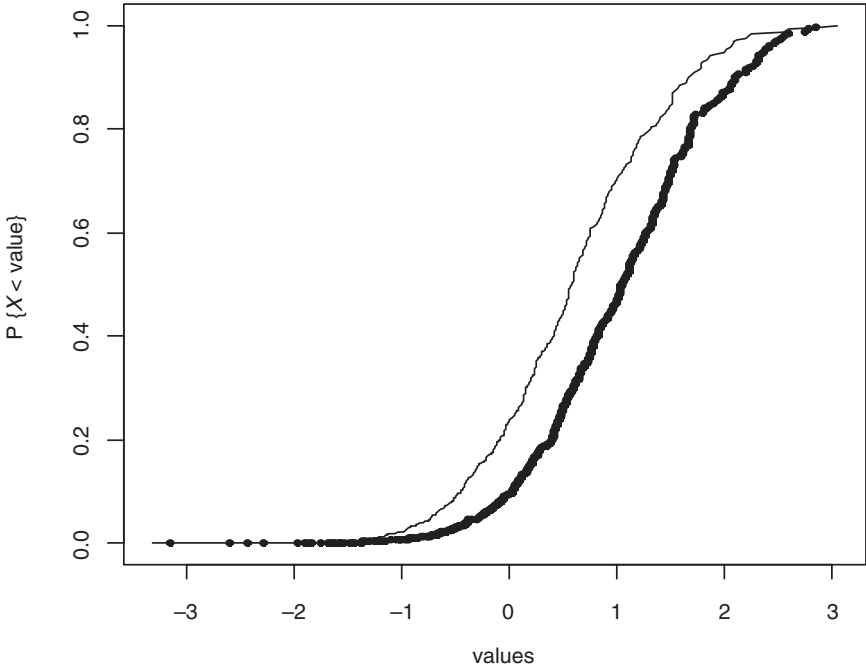
### 3.1. DISTRIBUTION OF VALUES

Life constantly calls upon us to make decisions. Should penicillin or erythromycin be prescribed for an infection? Which fertilizer should be used to get larger tomatoes? Which style of dress should our company manufacture (that is, which style will lead to greater sales)?

My wife often asks me what appears to be a very similar question, “Which dress do you think I should wear to the party?” But this question is really quite different from the others as it asks what a specific individual, me, thinks about dresses that are to be worn by another specific individual, my wife. All the other questions reference the behavior of a yet-to-be-determined individual selected *at random* from a population.

Is Alice taller than Peter? It won't take us long to find out: We just need to put the two back to back or measure them separately. The somewhat different question, “Are girls taller than boys?” is not answered quite so readily. How old are the boys and girls? Generally, but not always, girls are taller than boys until they enter adolescence. Even then, what may be true in general for boys and girls of a specified age group may not be true for a particular girl and a particular boy.

Put in its most general and abstract form, what we are asking is whether a numerical observation  $X$  made on an individual drawn at random from



**FIGURE 3.1** Two cumulative distributions that differ by a shift in the median value.

one population will be larger than a similar numerical observation  $Y$  made on an individual drawn at random from a second population.

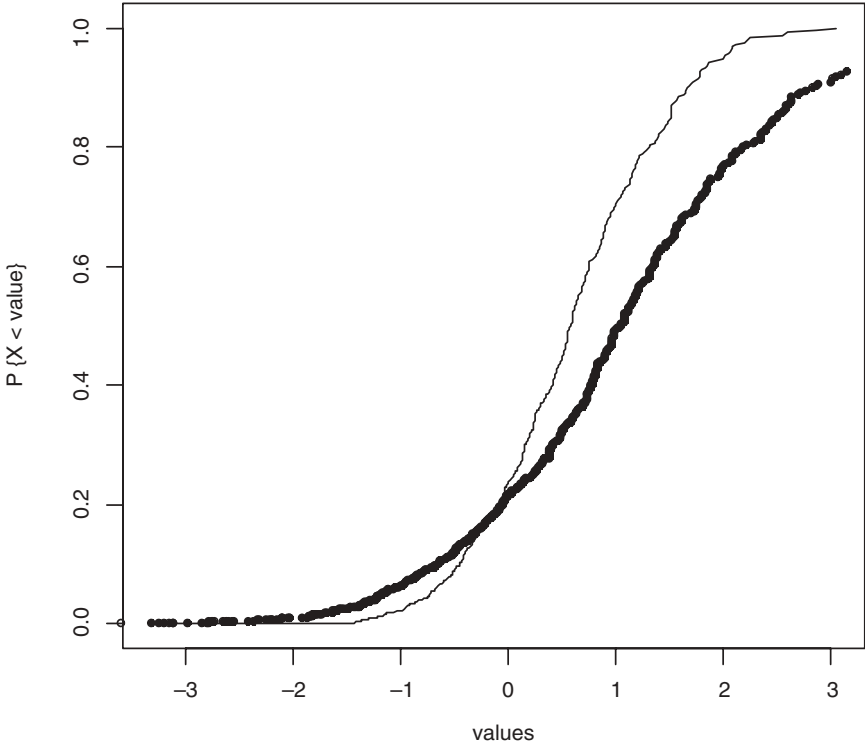
**3.1.1. Cumulative Distribution Function**

Let  $F_W[w]$  denote the probability that the numerical value of an observation from the distribution  $W$  will be less than or equal to  $w$ .  $F_W[w]$  is a monotone nondecreasing function. In symbols, if  $w < z$ , then  $0 = F_W[-\infty] \leq F_W[w] = \Pr\{W \leq w\} \leq \Pr\{W \leq z\} = F_W[z] \leq F_W[\infty] = 1$ .

Two such *cumulative distribution functions*  $F_X$  and  $G_X$  are depicted in Fig. 3.1.

Note the following in Fig. 3.1:

1.  $F_X$  is to the left of  $G_X$ . As can be seen by drawing lines perpendicular to the value axis,  $F_X[x] > G_X[x]$  for all values of  $X$ . As can be seen by drawing lines perpendicular to the percentile axis, all the percentiles of the cumulative distribution  $G$  are smaller than the percentiles of the cumulative distribution  $F$ .
2. Most of the time an observation taken at random from the distribution of  $X$  will be smaller if that distribution has cumulative distribution  $F$  than if it has cumulative distribution  $G$ .



**FIGURE 3.2** Two cumulative distributions whose mean and variance differ.

3. Still, there is a nonzero probability that an observation from  $F_X$  will be larger than one from  $G_X$ .

Many treatments act by shifting the distribution of values, as shown in Fig. 3.1. The balance of this chapter and Chapter 4 are concerned with the detection of such treatment effects. The possibility exists that the actual effects of treatment are more complex than is depicted in Fig. 3.1. In such cases, (see, for example, Fig. 3.2) the introductory methods described in this text may not be immediately applicable.

**Exercise 3.0.** Is it possible that an observation drawn at random from the distribution  $F_X$  depicted in Fig. 3.1 could have a value larger than an observation drawn at random from the distribution  $G_X$ ?<sup>1</sup>

<sup>1</sup> If the answer to this exercise is not immediately obvious—you'll find the correct answer at the end of this chapter—you should reread Chapters 1 and 2 before proceeding further.



### 3.1.2 Empirical Distribution Function

Suppose we have collected a sample of  $n$  observations  $x_1, x_2, \dots, x_n$ . The *empirical cumulative distribution function*  $F_n[x]$  is equal to the number of observations that are less than or equal to  $x$  divided by the size of our sample,  $n$ . If we've sorted the sample so that  $x_1 \leq x_2 \leq \dots \leq x_n$ , then

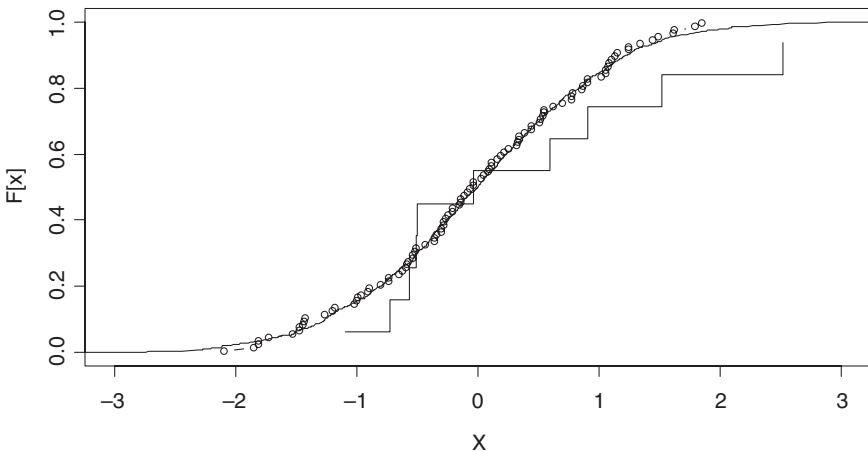
$$\begin{aligned}
 &0 \quad \text{if } x < x_i \\
 F_n[x] &= k/n \quad \text{if } x_k \leq x < x_{k+1}, \text{ for } 1 \leq k \leq (n-1) \\
 &1 \quad \text{if } x > x_n
 \end{aligned}$$

If these observations all come from the same population distribution  $F$  and are independent of one another, then as the sample size  $n$  gets larger,  $F_n$  will begin to resemble  $F$  more and more closely. We illustrate this point in Fig. 3.3 with samples of size 10, 100, and 1000, all taken from the same distribution.

Figure 3.3 reveals what you will find in your own samples in practice: The distance (or fit) between the empirical and theoretical distributions is best in the middle of the distribution near the median and worst in the tails.

### 3.2. DISCRETE DISTRIBUTIONS

We need to distinguish between *discrete* random observations like the binomial and the Poisson (see Section 3.3) made when recording numbers



**FIGURE 3.3** Three empirical distributions based on samples of size 10, 100, and 1000 independent observations from the same population.

of events and the *continuous* random observations that are made when taking measurements.

Discrete random observations usually take only integer values (positive or negative) with nonzero probability. That is,

$$\begin{aligned} \Pr\{X = k\} &= f_k \quad \text{if } k = \{\dots, -1, 0, 1, 2, \dots\} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

The cumulative distribution function  $F[x]$  is  $\sum_{k \leq x} f_k$ .

Recall from the preceding chapter (Section 2.2.2) that binomial variables had probabilities

$$\begin{aligned} \Pr\{X = k | n, p\} &= \binom{n}{k} (p)^k (1-p)^{n-k} \quad \text{for } k \in \{0, 1, 2, \dots, n\} \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

where  $n$  denoted the number of independent trials and  $p$  was the probability of success in each trial. The cumulative distribution function of a binomial is a step function, equal to zero for all values less than 0 and to one for all values greater than or equal to  $n$ .

Although it seems obvious that the mean of a sufficiently large number of sets of  $n$  binomial trials each with a probability  $p$  of success will be equal to  $np$ , many things that seem obvious in mathematics aren't. Besides, if it's that obvious, we should be able to prove it.

If a variable  $X$  takes a discrete set of values  $\{\dots, 0, 1, \dots, k, \dots\}$  with corresponding probabilities  $\{\dots, f_0, f_1, \dots, f_k, \dots\}$ , its mean or expected value, written  $EX$ , is given by the summation:  $\dots + 0f_0 + 1f_1 + \dots + kf_k + \dots$ , which we may also write as  $\sum_k kf_k$ . For a binomial variable, this sum is

$$\begin{aligned} EX &= \sum_{k=0}^n k \binom{n}{k} (p)^k (1-p)^{n-k} \\ &= \sum_{k=0}^n k \frac{n(n-1)!}{k!(n-k)!} p^k (1-p)^{n-1-(k-1)} \end{aligned}$$

Note that the term on the right is equal to zero when  $k = 0$ , so we can start the summation at  $k = 1$ . Factoring  $n$  and  $p$  outside the summation and using the  $k$  in the numerator to reduce  $k!$  to  $(k-1)!$  we have

$$EX = np \sum_{k=1}^n \frac{(n-1)!}{(k-1)!(n-1-(k-1))!} (p)^{k-1} (1-p)^{(n-1)-(k-1)}$$

If we change the notation a bit, letting  $j = (k - 1)$  and  $m = (n - 1)$ , this can be expressed as

$$EX = np \sum_{j=0}^m \binom{m}{j} (p)^j (1-p)^{m-j}$$

The summation on the right side of the equals sign is of the probabilities associated with *all* possible outcomes for the binomial random variable  $B(n, p)$ , so it must be equal to 1. Thus  $EX = np$ , a result that agrees with our intuitive feeling that, in the long run, the number of successes should be proportional to the probability of success.

Suppose  $EX = \theta$  (pronounced theta). The *variance* of  $X$  is defined as  $\text{Var}(X) = E(X - \theta)^2$ . In contrast to the *sample variance* defined in Chapter 1,  $\text{Var}(X)$  stands for a purely hypothetical value. For observations from a discrete distribution,  $\text{Var}(X) = \sum_k (k - \theta)^2 f_k$ .

**Exercise 3.1.** (For math and statistics majors and the intensely curious only.) Show that the variance of a binomial variable is  $np(1 - p)$ .

**Exercise 3.2.** Is the binomial distribution *symmetric* about its mean? Do its mean and median coincide? Does it have more than one mode? [Hint: Use the instructions provided in Section 2.2.4 to display the binomial distribution for various probabilities of success and numbers of trials.]

**Exercise 3.3.** Recently, we interviewed 10 people and found that the majority favored our candidate. Should we conclude that our candidate is sure to win a majority? Support your opinion with numerical values.

**Exercise 3.4.** Create plots of the cumulative distribution functions of the binomial random variables  $B(20, 0.5)$  and  $B(20, 0.7)$ .

### 3.3. POISSON: EVENTS RARE IN TIME AND SPACE

The decay of a radioactive element, an appointment to the United States Supreme Court, and a cavalry officer trampled by his horse have in common that they are relatively rare but inevitable events. They are inevitable, that is, if there are enough atoms, enough seconds or years in the observation period, and enough horses and momentarily careless riders. Their frequency of occurrence has a Poisson distribution.

The number of events in a given interval has the Poisson distribution if

- a) It is the cumulative result of a large number of independent opportunities each of which has only a small chance of occurring and
- b) Events in nonoverlapping intervals are independent.

The intervals can be in space or time. For example, if we seed a small number of cells into a Petri dish that is divided into a large number of squares, the distribution of cells per square follows the Poisson. The same appears to be true in the way a large number of masses in the form of galaxies are distributed across a very large universe.

Like the binomial variable, a Poisson variable only takes nonnegative integer values. If the number of events  $X$  has a Poisson distribution such that we may expect an average of  $\lambda$  events per unit interval, then  $\Pr\{X = k\} = \lambda^k e^{-\lambda}/k!$  for  $k = 0, 1, 2, \dots$ . For the purpose of testing hypotheses concerning  $\lambda$  as discussed in the chapter following, we needn't keep track of the times or locations at which the various events occur; the number of events  $k$  is *sufficient*.

**Exercise 3.5.** Show without using algebra that the sum of a Poisson with expected value  $\lambda_1$  and a second independent Poisson with expected value  $\lambda_2$  is also a Poisson with expected value  $\lambda_1 + \lambda_2$ .

### 3.3.1. Applying the Poisson

John Ross of the Wistar Institute held that there were two approaches to biology: the analog and the digital. The analog was served by the scintillation counter: One ground up millions of cells, then measured whatever radioactivity was left behind in the stew after centrifugation. The digital was to be found in cloning experiments where any necessary measurements would be done on a cell-by-cell basis.

John was a cloner and, later, as his student, so was I. We'd start out with 10 million or more cells in a 10-milliliter flask and try to dilute them down to one cell per milliliter. We were usually successful in cutting down the numbers to ten thousand or so. Then came the hard part. We'd dilute the cells down a second time by a factor of 1:100 and hope we'd end up with 100 cells in the flask. Sometimes we did. Ninety percent of the time, we'd end up with between 90 and 110 cells, just as the binomial distribution predicted. But just because you cut a mixture in half (or a dozen, or a 100 parts) doesn't mean you're going to get equal numbers in each part. It means the probability of getting a particular cell is the same for all the parts. With large numbers of cells, things seem to even out. With small numbers, chance seems to predominate.

Things got worse, when I went to seed the cells into culture dishes. These dishes, made of plastic, had a rectangular grid cut into their

bottoms, so they were divided into approximately 100 equal-sized squares. Dropping 100 cells into the dish meant an average of 1 cell per square. Unfortunately for cloning purposes, this average didn't mean much. Sometimes, 40% or more of the squares would contain two or more cells. It didn't take long to figure out why. Planted at random, the cells obey the Poisson distribution in space. An average of one cell per square means

$$\Pr\{\text{No cells in a square}\} = 1 * e^{-1}/1 = 0.32$$

$$\Pr\{\text{Exactly one cell in a square}\} = 1 * e^{-1}/1 = 0.32$$

$$\Pr\{\text{Two or more cells in a square}\} = 1 - 0.32 - 0.32 = 0.36.$$

Two cells was one too many. A clone or colony must begin with a single cell. I had to dilute the mixture a third time to ensure that the percentage of squares that included two or more cells was vanishingly small. Alas, the vast majority of squares were now empty; I was forced to spend hundreds of additional hours peering through the microscope looking for the few squares that did include a clone.

### 3.3.2. Comparing Empirical and Theoretical Poisson Distributions

BoxSampler includes a Poisson Distribution.

**Exercise 3.6.** Generate the results of 100 samples from a Poisson distribution with an expected number of 2 events per interval. Compare the graph of the resulting empirical frequency distribution with that of the corresponding theoretical frequency distribution. Determine the 10th, 50th, and 90th percentiles of the theoretical Poisson distribution.

**Exercise 3.7.** Show that if  $\Pr\{X = k\} = \lambda^k e^{-\lambda}/k!$  for  $k = 0, 1, 2, \dots$  that is, if  $X$  is a Poisson variable, then the expected value of  $X = \sum_k k \Pr\{X = k\} = \lambda$ .

**Exercise 3.8.** In subsequent chapters, we will learn how the statistical analysis of trials of a new vaccine is often simplified by assuming that the number of infected individuals follows a Poisson rather than a binomial distribution. To see how accurate an approximation this might be, compare the cumulative distribution functions of a binomial variable,  $B(100, 0.01)$  and a Poisson variable,  $P(1)$  over the range 0 to 100.

### 3.4. CONTINUOUS DISTRIBUTIONS

The vast majority of the observations we make are on a continuous scale even if, in practice, we only can make them in discrete increments. For example, a man's height might actually be 1.835421117 meters, but we are not likely to record a value with that degree of accuracy (or want to). If one's measuring stick is accurate to the nearest millimeter, then the probability that an individual selected at random will be exactly 2 meters tall is really the probability that his or her height will lie between 1.9995 and 2.0004 meters. In such a situation, it is more convenient to replace the sum of an arbitrarily small number of quite small probabilities with an integral  $\int_{1.9995}^{2.0004} dF[x] = \int_{1.9995}^{2.0004} f[x]dx$  where  $F[x]$  is the cumulative distribution function of the continuous variable representing height and  $f[x]$  is its probability density. Note that  $F[x]$  is now defined as  $\int_{-\infty}^x f[y]dy$ .<sup>2</sup> As with discrete variables, the cumulative distribution function is monotone non-decreasing from 0 to 1, the distinction being that it is a smooth curve rather than a step function.

The mathematical expectation of a continuous variable is  $\int_{-\infty}^{\infty} yf[y]dy$ , and its variance is  $\int_{-\infty}^{\infty} (y - EY)^2 f[y]dy$ .

#### 3.4.1. The Exponential Distribution

The simplest way to obtain continuously distributed random observations is via the same process that gave rise to the Poisson. Recall that a Poisson process is such that events in nonoverlapping intervals are independent and identically distributed. The times<sup>3</sup> between Poisson events follow an exponential distribution:

$$F[t|\lambda] = \Pr\{T \leq t|\lambda\} = 1 - \exp[-\lambda t] \text{ for } t \geq 0, \lambda > 0.$$

When  $t$  is zero,  $\exp[-\lambda t]$  is 1 and  $F[t|\lambda]$  is 0. As  $t$  increases,  $\exp[-\lambda t]$  decreases rapidly toward zero and  $F[t|\lambda]$  increases rapidly to 1. The rate of increase is proportional to the magnitude of the parameter  $\lambda$ . In fact,  $\log(1 - F[t|\lambda]) = -\lambda t$ . Exercise 3.9 allows you to demonstrate this for yourself.

<sup>2</sup> If it's been a long while or never since you had calculus, note that the differential  $dx$  or  $dy$  is a meaningless index, so any letter will do, just as  $\sum_i f_i$  means exactly the same thing as  $\sum_j f_j$ .

<sup>3</sup> Time is almost but not quite continuous. Modern cosmologists now believe that both time and space are discrete, with time determined only to the nearest  $10^{-23}$  of a second.

**Exercise 3.9.** Draw the cumulative distribution function of an exponentially distributed observation with parameter  $\lambda$ . Is the median the same as the mean?

**Exercise 3.10.** (requires calculus) What is the expected value of an exponentially distributed observation with parameter  $\lambda$ ?

The times between counts on a Geiger counter follow an exponential distribution. So do the times between failures of manufactured items like light bulbs that rely on a single crucial component.

**Exercise 3.11.** When you walk into a room, you discover the light in a lamp is burning. Assuming the life of its bulb is exponentially distributed with an expectation of one year, how long do you expect it to be before the bulb burns out? [Many people find they get two contradictory answers to this question. If you are one of them, see Feller, 1966, p.11–12.]

Most real-life systems (including that complex system known as a human being) have built-in redundancies. Failure can only occur after a series of  $n$  breakdowns. If these breakdowns are independent and exponentially distributed, all with the same parameter  $\lambda$ , the probability of failure of the total system at time  $t > 0$  is

$$f(t) = \lambda \exp(-\lambda t)(\lambda t)^n / n!$$

### 3.4.2. The Normal Distribution

Figure 1.24 depicts the bell-shaped symmetric frequency curve of a normally distributed population. Its probability density  $f(x)$  may be written as

$$f[x|\theta, \sigma^2] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2\right) \quad (3.1)$$

In contrast to the exponential distribution, the normal distribution depends on two parameters: its expected value  $\theta$  (theta) and its variance  $\sigma^2$  (sigma-squared).

**Exercise 3.12.** How do changes in the values of these parameters affect the shape of the normal distribution? Hint: Let  $w = (y - \theta)/\sigma$ , so the probability density function can be written as

$$f[x|\theta, \sigma^2] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2\right)$$

**Exercise 3.13.** (requires calculus) Show that the expected value of a normal distribution whose density is given by Equation 3.1 is  $\theta$  and its variance is  $\sigma^2$ .

#### STATISTICS AND PARAMETERS

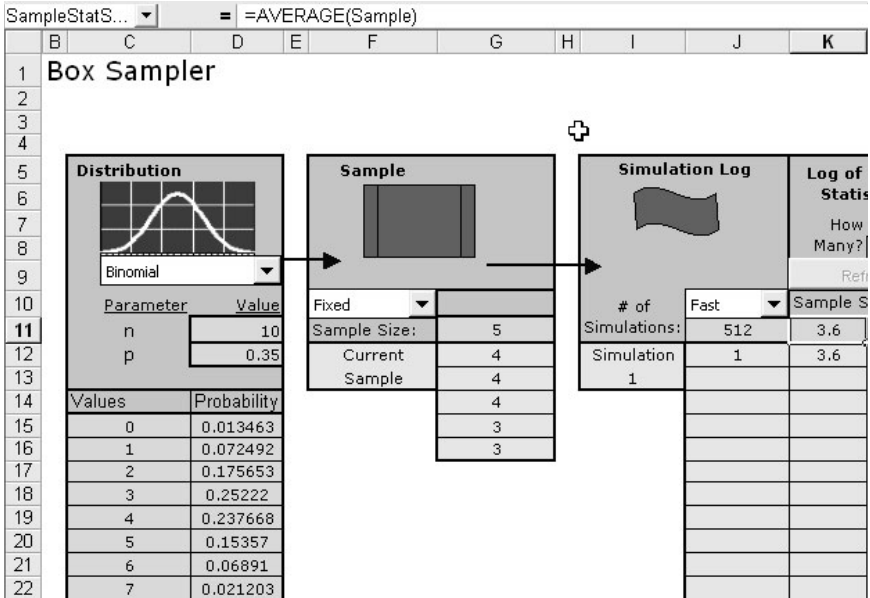
A *statistic* is any single value such as the sample mean  $\frac{1}{n}\sum_{k=1}^n X_k$  that summarizes some aspect of a sample. A *parameter* is any single value such as the mean  $\theta$  of a normal distribution that summarizes some aspect of an entire population. Examples of sample statistics include measures of location and central tendency such as the sample mode, sample median, and sample mean, extrema such as the sample minimum and maximum, and measures of variation and dispersion such as the sample standard deviation. These same measures are considered **parameters** when they refer to an entire population, e.g., population mean  $\theta$ , population range, population variance  $\sigma^2$ .

In subsequent chapters, we will use sample statistics to estimate the values of population parameters and to test hypotheses about them.

To see why the normal distribution plays such an important role in statistics, please complete Exercise 3.14, which requires you to compute the distribution of the mean of a number of binomial observations. As you increase the number of observations used to compute the mean from 5 to 12 so that each individual observation makes a smaller relative contribution to the total, you will see that the distribution of means looks less and less like the binomial distribution from which they were taken and more and more like a normal distribution. This result will prove to be true regardless of the distribution from which the observations used to compute the mean are taken, providing that this distribution has a finite mean and variance.

**Exercise 3.14.** Generate five binomial observations based on 10 trials with probability of success  $p = 0.35$  per trial. Compute the mean value of these five. Repeat this procedure 512 times, computing the mean value each time. Plot the histogram of these means. Compare with the histograms of a) a sample of 512 normally distributed observations with expected value 3.5 and variance 2.3, b) a sample of 512 binomial observations each consisting of 10 trials with probability of success  $p = 0.4$  per





**FIGURE 3.4** Preparing to generate the mean values of binomial samples.

trial. Repeat the entire exercise, computing the mean of 12 rather than 5 binomial observations.

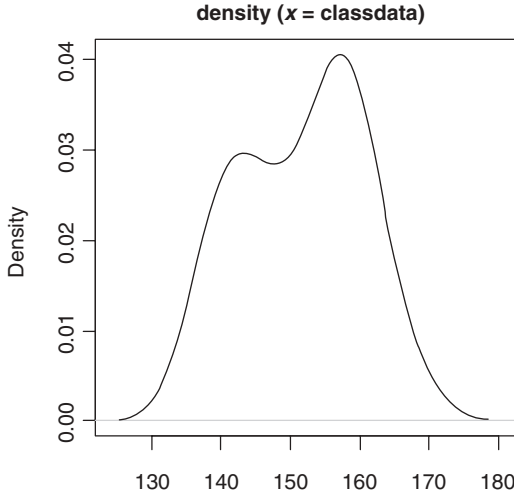
To do this exercise, you'll need to set up BoxSampler as shown in Fig. 3.4, inserting the formula `= Average(Sample)` in cell K11.

**3.4.3. Mixtures of Normal Distributions**

Many real-life distributions strongly resemble mixtures of normal distributions. Figure 3.5 depicts just such an example in the heights of California sixth-graders. Although the heights of boys and girls overlap, it is easy to see that the population of sixth graders is composed of a mixture of the two sexes.

**3.5. PROPERTIES OF INDEPENDENT OBSERVATIONS**

As virtually all the statistical procedures in this text require that our observations be independent of one another, we need to study the properties of independent observations. Recall from the previous chapter that if  $X$  and  $Y$  are independent observations, then



**FIGURE 3.5** Distribution of the heights of California sixth-graders.

$$\Pr\{X = k \text{ and } Y = j\} = \Pr\{X = k\} \Pr\{Y = j\}$$

and

$$\Pr\{X = k | Y = j\} = \Pr\{X = k\}.$$

If  $X$  and  $Y$  are independent discrete random variables and their expectations exist and are finite<sup>4</sup>, then  $E\{X + Y\} = EX + EY$ . To see this, suppose that  $Y = y$ . The conditional expectation of  $X + Y$  given  $Y = y$  is

$$\begin{aligned} E\{X + Y | Y = y\} &= \sum_k (k + y) \Pr\{X + Y = k + y | Y = y\} \\ &= \sum_k (k + y) \Pr\{X = k\} \\ &= EX + y \sum_k \Pr\{X = k\} = EX + y. \end{aligned}$$

Taking the average of this conditional expectation over all possible values of  $Y$  yields

$$E\{X + Y\} = \sum_j (EX + j) \Pr\{Y = j\} = EX * 1 + EY.$$

<sup>4</sup> In real life, expectations almost always exist and are finite—the expectations of ratios are a notable exception.

A similar result holds if  $X$  and  $Y$  have continuous distributions, providing that their individual expectations exist and are finite.

**Exercise 3.15.** Show that for any variable  $X$  with a finite expectation,  $E(aX) = aEX$ , where  $a$  is a constant.

**Exercise 3.16.** Show that the expectation of the mean of  $n$  independent identically distributed random variables with finite expectation  $\theta$  is also  $\theta$ .

### THE CAUCHY DISTRIBUTION: AN EXCEPTION TO THE RULE

It seems obvious that every distribution should have a mean and a variance. But telling a mathematician something is “obvious” only encourages him/her to find an exception. The Cauchy distribution is just one example: The expression  $f(x) = 1/\Pi(1+x)$  for  $-\infty < x < \infty$  is a probability density because  $\int f(x)dx = 1$ , but neither its mean nor its variance exists. Does the Cauchy distribution arise in practice? It might if you study the ratio  $X/Y$  of two independent random variables each distributed as  $N(0,1)$ .

The variance of the sum of two independent variables  $X$  and  $Y$  is the sum of their variances, providing each of these variances exists and is finite.

**Exercise 3.17.** Given the preceding result, show that the variance of the mean of  $n$  independent identically distributed observations is  $1/n$ th of the variance of just one of them. Does this mean that the arithmetic mean is more precise than any individual observation? Does this mean that the sample mean will be closer to the mean of the population from which it is drawn than any individual observation would be, that is, that it would be more accurate?

## 3.6. TESTING A HYPOTHESIS

Suppose we were to pot a half-dozen tomato plants in ordinary soil and a second half-dozen plants in soil enriched with fertilizer. If we wait a few months, we can determine whether the addition of fertilizer increases the resulting yield of tomatoes, at least as far as these dozen plants are concerned. But can we extend our findings to all tomatoes?

To ensure that we can extend our findings we need to proceed as follows: First, the 12 tomato plants used in our study have to be a *random sample* from a nursery. If we choose only plants with especially

green leaves for our sample, then our results can be extended only to plants with especially green leaves. Second, we have to divide the 12 plants into two treatment groups at random. If we subdivide by any other method, such as tall plants in one group and short plants in another, then the experiment would not be about fertilizer but about our choices.

I performed just such an experiment a decade or so ago, only I was interested in the effects vitamin E might have on the aging of human cells in culture. After several months of abject failure—contaminated cultures, spilled containers—I succeeded in cloning human diploid fibroblasts in eight culture dishes. Four of these dishes were filled with a conventional nutrient solution and four held an experimental “life-extending” solution to which vitamin E had been added. All the cells in the dishes came from the same culture so that the initial distribution of cells was completely random.

I waited three weeks with my fingers crossed—there is always a risk of contamination with cell cultures—but at the end of this test period three dishes of each type had survived. I transplanted the cells, let them grow for 24 hours in contact with a radioactive label, and then fixed and stained them before covering them with a photographic emulsion.

Ten days passed, and we were ready to examine the autoradiographs. “121, 118, 110, 34, 12, 22.” I read and reread these six numbers over and over again. The larger numbers were indicative of more cell generations and an extended life span. If the first three generation counts were from treated colonies and the last three were from untreated, then I had found the fountain of youth. Otherwise, I really had nothing to report.

### 3.6.1. Analyzing the Experiment

How had I reached this conclusion? Let’s take a second, more searching look. First, we identify the primary hypothesis and the alternative hypothesis of interest.

I wanted to assess the life-extending properties of a new experimental treatment with vitamin E. To do this, I had divided my cell cultures into two groups: one grown in a standard medium and one grown in a medium containing vitamin E. At the conclusion of the experiment and after the elimination of several contaminated cultures, both groups consisted of three independently treated dishes.

My primary hypothesis was a *null hypothesis*, that the growth potential of a culture would not be affected by the presence of vitamin E in the media: All the cultures would have equal growth potential. The *alterna-*

*tive* hypothesis of interest was that cells grown in the presence of vitamin E would be capable of many more cell divisions.

Under the null hypothesis, the labels “treated” and “untreated” provide no information about the outcomes: The observations are expected to have more or less the same values in each of the two experimental groups. If they do differ, it should only be as a result of some uncontrollable random fluctuation. Thus if this null or no-difference hypothesis were true, I was free to exchange the labels.

The alternative is a distributional shift like that depicted in Fig. 3.1, in which greater numbers of cell generations are to be expected as the result of treatment with vitamin E (though the occasional smaller value cannot be ruled out completely).

The next step is to choose a test statistic that discriminates between the hypothesis and the alternative. The statistic I chose was the sum of the counts in the group treated with vitamin E. *If* the alternative hypothesis is true, most of the time this sum ought to be larger than the sum of the counts in the untreated group. *If* the *null hypothesis* is true, that is, if it doesn’t make any difference which treatment the cells receive, then the sums of the two groups of observations should be approximately the same. One sum might be smaller or larger than the other by chance, but most of the time the two shouldn’t be all that different.

The third step is to compute the test statistic for each of the possible relabelings and compare these values with the value of the test statistic as the data was labeled originally. As it happened, the first three observations—121, 118, and 110—were those belonging to the cultures that received vitamin E. The value of the test statistic for the observations as originally labeled is  $349 = 121 + 118 + 110$ .

I began to rearrange (permute) the observations, randomly reassigning the six labels, three “treated” and three “untreated,” to the six observations, for example, treated, 121 118 34, and untreated, 110 12 22. In this particular rearrangement, the sum of the observations in the first (treated) group is 273. I repeated this step until all  $\binom{6}{3} = 20$  distinct rearrangements had been examined.<sup>5</sup>

<sup>5</sup> Determination of the number of relabelings, “6 choose 3” in the present case, is considered in Section 2.2.1.

	First Group	Sum of Second Group	First Group
1.	121 118 110	34 22 12	349
2.	121 118 34	110 22 12	273
3.	121 110 34	118 22 12	265
4.	118 110 34	121 22 12	262
5.	121 118 22	110 34 12	261
6.	121 110 22	118 34 12	253
7.	121 118 12	110 34 22	251
8.	118 110 22	121 34 12	250
9.	121 110 12	118 34 22	243
10.	118 110 12	121 34 22	240
11.	121 34 22	118 110 12	177
12.	118 34 22	121 110 12	174
13.	121 34 12	118 110 22	167
14.	110 34 22	121 118 12	166
15.	118 34 12	121 110 22	164
16.	110 34 12	121 118 22	156
17.	121 22 12	118 110 34	155
18.	118 22 12	121 110 34	152
19.	110 22 12	121 118 34	144
20.	34 22 12	121 118 110	68

The sum of the observations in the original vitamin E-treated group, 349, is equaled only once and never exceeded in the 20 distinct random re-labelings. If chance alone is operating, then such an extreme value is a rare, only 1-time-in-20 event. If I reject the null hypothesis and embrace the alternative that the treatment is effective and responsible for the observed difference, I only risk making an error and rejecting a true hypothesis once in every 20 times.

In this instance, I did make just such an error. I was never able to replicate the observed life-promoting properties of vitamin E in other repetitions of this experiment. Good statistical methods can reduce and contain the probability of making a bad decision, but they cannot eliminate the possibility.

**Exercise 3.18.** How was the analysis of the cell culture experiment affected by the loss of two of the cultures because of contamination? Suppose these cultures had escaped contamination and given rise to the observations 90 and 95; what would be the results of a permutation

analysis applied to the new, enlarged data set consisting of the following cell counts

Treated 121 118 110 90      Untreated 95 34 22 12?

Hint: To determine how probable an outcome like this is by chance alone, first determine how many possible rearrangements there are. Then list all the rearrangements that are as or more extreme than this one.

### 3.6.2. Two Types of Errors

In the preceding example, I risked rejecting the null hypothesis in error 5% of the time. Statisticians call this making a *Type I error*, and they call the 5%, the *significance level*. In fact, I did make such an error, as in future experiments vitamin E proved to be valueless in extending the life span of human cells in culture.

On the other hand, suppose the null hypothesis had been false, that treatment with vitamin E really did extend life span, and I had failed to reject the null hypothesis. Statisticians call this making a *Type II error*.

The consequences of each type of error are quite different and depend upon the context of the investigation. Consider the table of possibilities (Table 3.1) arising from an investigation of the possible carcinogenicity of a new headache cure.

We may luck out in that our sample supports the correct hypothesis, but we always run the risk of making either a Type I or a Type II error. We can't avoid it. If we use a smaller significance level, say 1%, then if the null hypothesis is false, we are more likely to make a Type II error. If we always accept the null hypothesis, a significance level of 0%, then we guarantee you'll make a Type II error if the null hypothesis is false. This seems kind of stupid: Why bother to gather data if you're not going to use it? But if you read or live Dilbert, then you know this happens all the time.

**TABLE 3.1 Decision Making Under Uncertainty**

The Facts		Investigator's Decision
Not a Carcinogen	Not a Carcinogen	Compound a Carcinogen <i>Type I error.</i> Manufacturer misses opportunity for profit. Public denied access to effective treatment.
Carcinogen	<i>Type II error.</i> Manufacturer sued. Patients die; families suffer.	

**Exercise 3.19.** The nurses have petitioned the CEO of a hospital to allow them to work 12-hour shifts. He wants to please them but is afraid that the frequency of errors may increase as a result of the longer shifts. He decides to conduct a study and to test the null hypothesis that there is no increase in error rate as a result of working longer shifts against the alternative that the frequency of errors increases by at least 30%. Describe the losses associated with Type I and Type II errors.

**Exercise 3.20.** Design a study. Describe a primary hypothesis of your own along with one or more likely alternatives. The truth or falsity of your chosen hypothesis should have measurable monetary consequences. If you were to test your hypothesis, what would be the consequences of making a Type I error? A Type II error?

**Exercise 3.21.** Suppose I'm (almost) confident that my candidate will get 60% or more of the votes in the next primary. The alternative that scares me is that she will get 40% or less. To test my confident hypothesis, I decide to interview 20 people selected at random in a shopping mall and reject my hypothesis if 7 or fewer say they will vote for her. What is the probability of my making a Type I error? What is the probability of my retaining confidence in my candidate if only 40% of the general population favor her, i.e., committing a Type II error? How can I reduce the probability of making a Type II error while keeping the probability of making a Type I error the same?

**Exercise 3.22.** Individuals were asked to complete an extensive questionnaire concerning their political views and eating preferences. Analyzing the results, a sociologist performed 20 different tests of hypotheses. Unknown to the sociologist, the null hypothesis was true in all 20 cases. What is the probability that the sociologist rejected at least one of the hypotheses at the 5% significance level?

### 3.7. ESTIMATING EFFECT SIZE

In the previous example, we developed a test of the null hypothesis of no treatment effect against the alternative hypothesis that a positive effect existed. But in many situations, we would also want to know the magnitude of the effect. Does vitamin E extend cell life span by 3 cell generations? By 10? By 15?

In Section 1.6.2 we showed how to use the bootstrap to estimate the precision of the sample mean or median (or, indeed, almost any sample



statistic) as an estimator of a population parameter. As a by-product, we obtain an *interval estimate* of the corresponding population parameter.

For example, if  $P_{05}$  and  $P_{95}$  are the 5th and 95th percentiles of the bootstrap distribution of the median of the law school LSAT data you used for Exercise 1.16, then the set of values between  $P_{05}$  and  $P_{95}$  provides a 90% *confidence interval* for the median of the population from which the data were taken.

**Exercise 3.23.** Obtain an 85% confidence interval for the median of the population from which the LSAT data were taken.

**Exercise 3.24.** Can this same bootstrap technique be used to obtain a confidence interval for the 90th percentile of the population? For the maximum value in the population?

### 3.7.1. Confidence Interval for Difference in Means

Suppose we have independently collected samples from two populations and want to know the following:

- Do the populations from which they are drawn have the same means?
- If the means are not the same, then what is the difference between them?

To find out, we would let each sample stand in place of the population from which it is drawn, take a series of bootstrap samples separately from each sample, and compute the difference in means each time.

Suppose our data are stored in two vectors called “control” and “treated” as shown in Fig. 3.6. We begin by creating a new BoxSampler model.

Next, we enter the formula = AVERAGE(Sample1)–AVERAGE(Sample2) in cell R11 of the BoxSampler worksheet (Fig. 3.7). We click the double arrow ►► on the simulation bar to generate a set of bootstrap results for the difference in means in column R. Finally, we use Excel’s sort command to sort these differences in descending order.

In the simulation I ran, the largest differences were 7.02 6.11 4.70 1.74 1.43 1.08 and –0.01. I discarded the top 5 as well as the bottom 5 in value to obtain a 90% (90 out of a 100) confidence interval of [–13.01, 1.08] for the difference in population means. As this interval contains zero, I wasn’t able to reject the possibility at the 10% significance level that the difference in population means was zero.

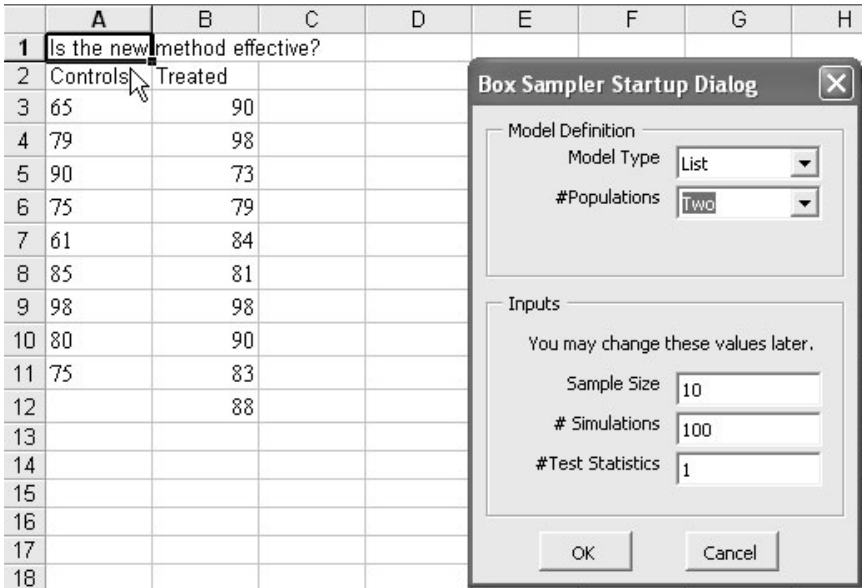


FIGURE 3.6 Preparing to estimate difference in population means.

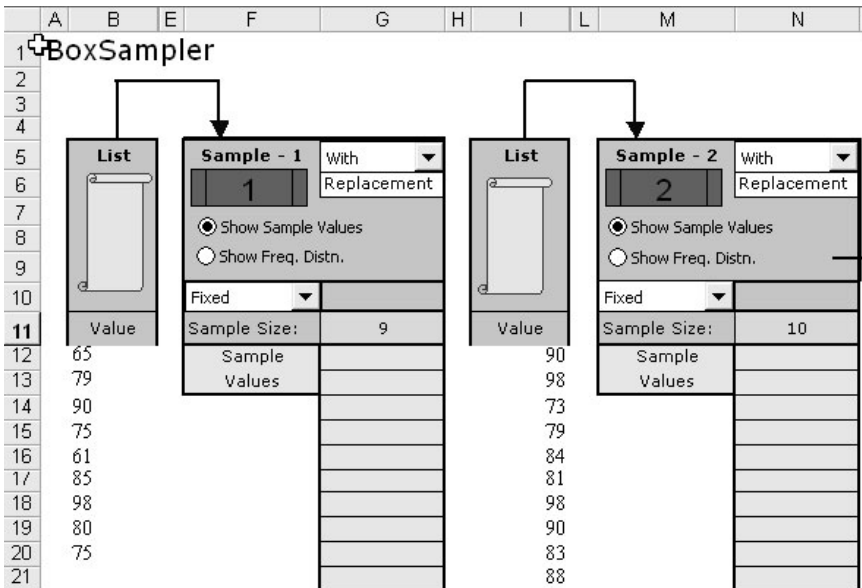


FIGURE 3.7 Entering data and sample sizes in the BoxSampler worksheet.

### 3.7.2. Are Two Variables Correlated?

Yet another example of the bootstrap's application lies in the measurement of the *correlation* or degree of agreement between two variables. The Pearson correlation of two variables  $X$  and  $Y$  is defined as the ratio of the covariance between  $X$  and  $Y$  and the product of the standard deviations of  $X$  and  $Y$ . The covariance of  $X$  and  $Y$  is given by the formula

$$\sum_{k=1}^n (X_k - \bar{X})(Y_k - \bar{Y}) / (n - 1) .$$

Recall that if  $X$  and  $Y$  are independent, the  $E(XY) = (EX)(EY)$ , so that the expected value of the covariance and hence the correlation of  $X$  and  $Y$  is zero. If  $X$  and  $Y$  increase more or less together as do, for example, the height and weight of individuals, their covariance and their correlation will be positive so that we say that height and weight are positively correlated. I had a boss, more than once, who believed that the more abuse and criticism he heaped on an individual the more work he could get out of them. Not. Abuse and productivity are negatively correlated; heap on the abuse and work output declines.

The reason we divide by the product of the standard deviations in assessing the degree of agreement between two variables is that it renders the correlation coefficient free of the units of measurement.

If  $X = -Y$ , so that the two variables are totally dependent, the correlation coefficient, usually represented in symbols by the Greek letter  $\rho$  (rho) will be  $-1$ . In all cases,  $-1 \leq \rho \leq 1$ .

Is systolic blood pressure an increasing function of age? To find out, I entered the data from 15 subjects in an Excel worksheet as shown in Fig. 3.8. Each row of the worksheet corresponds to a single subject. As described in Section 1.4.2, Resampling Stats was used to select a single bootstrap sample of subjects. That is, each row in the bootstrap sample corresponded to one of the rows of observations in the original sample.

Making use of the data from the bootstrap samples, I entered the formula for the correlation of Systolic Blood Pressure and Age in a convenient empty cell of the worksheet as shown in Fig. 3.9 and then used the RS button to generate 100 values of the correlation coefficient.

**Exercise 3.25.** Using the LSAT data from Exercise 1.16 and the bootstrap, obtain an interval estimate for the correlation between the LSAT score and the student's subsequent GPA.

**Exercise 3.26.** Trying to decide whether to take a trip to Paris or Tokyo, a student kept track of how many euros and yen his dollars would buy. Month by month he found that the values of both currencies were rising.

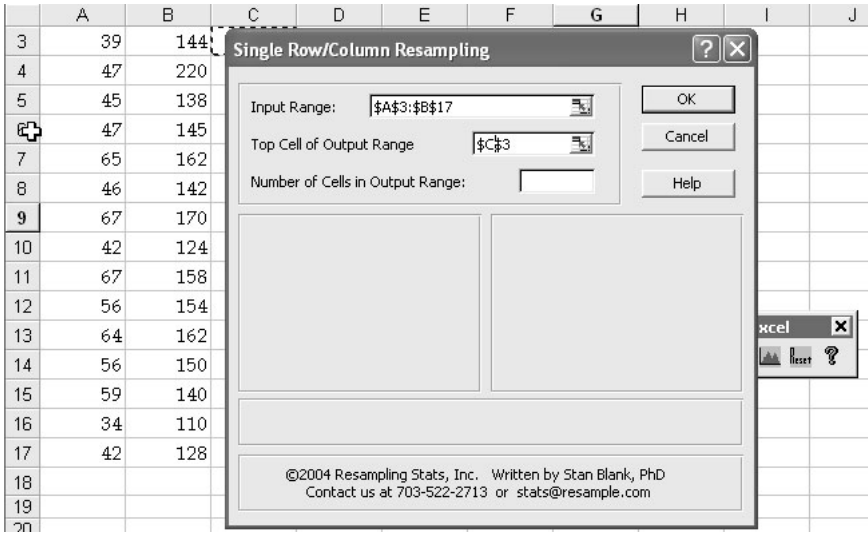


FIGURE 3.8 Preparing to generate a bootstrap sample of subjects.

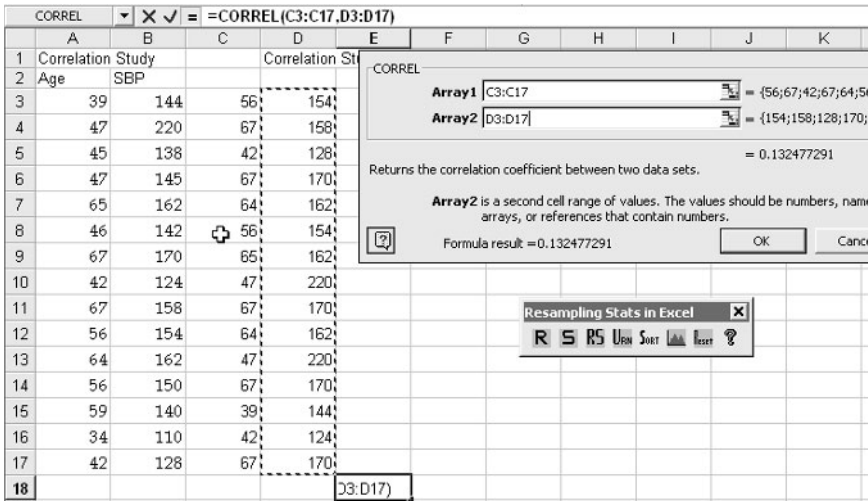


FIGURE 3.9 Calculating the correlation between systolic blood pressure and age.

Does this mean that improvements in the European economy are reflected by improvements in the Japanese economy?

### 3.7.3. Using Confidence Intervals to Test Hypotheses

Suppose we have derived a 90% confidence interval for some parameter, for example, a confidence interval for the difference in means between two populations, one of which was treated and one that was not. We can use this interval to test the hypothesis that the difference in means is 4 units, by accepting this hypothesis if 4 is included in the confidence interval and rejecting it otherwise. If our alternative hypothesis is nondirectional and two-sided,  $\theta_A \neq \theta_B$ , the test will have a Type I error of  $100\% - 90\% = 10\%$ .

Clearly, hypothesis tests and confidence intervals are intimately related. Suppose we test a series of hypotheses concerning a parameter  $\theta$ . For example, in the vitamin E experiment, we could test the hypothesis that vitamin E has no effect,  $\theta = 0$ , or that vitamin E increases life span by 25 generations,  $\theta = 25$ , or that it increases it by 50 generations,  $\theta = 50$ . In each case, whenever we accept the hypothesis, the corresponding value of the parameter should be included in the confidence interval.

In this example, we are really performing a series of one-sided tests. Our hypotheses are that  $\theta = 0$  against the one-sided alternative that  $\theta > 0$ , that  $\theta \leq 25$  against the alternative that  $\theta > 25$  and so forth. Our corresponding confidence interval will be one-sided also; we will conclude  $\theta < \theta_U$  if we accept the hypothesis  $\theta = \theta_0$  for all values of  $\theta_0 < \theta_U$  and reject it for all values of  $\theta_0 \geq \theta_U$ . One-sided tests lead to one-sided confidence intervals and two-sided tests to two-sided confidence intervals.

**Exercise 3.27.** What is the relationship between the significance level of a test and the confidence level of the corresponding interval estimate?

**Exercise 3.28.** In each of the following instances would you use a one-sided or a two-sided test?

- i. Determine whether men or women do better on math tests.
- ii. Test the hypothesis that women can do as well as men on math tests.
- iii. In *Commonwealth v. Rizzo et al.*, 466 F. Supp 1219 (E.D. Pa 1979), help the judge decide whether certain races were discriminated against by the Philadelphia Fire Department by means of an unfair test.
- iv. Test whether increasing a dose of a drug will increase the number of cures.

**Exercise 3.29.** Use the data of Exercise 3.18 to derive an 80% upper confidence bound for the effect of vitamin E to the nearest 5 cell generations.

### 3.8. SUMMARY AND REVIEW

In this chapter, we considered the form of four common distributions, two discrete—the binomial and the Poisson—and two continuous—the normal and the exponential. We provided the R functions necessary to generate random samples from the various distributions and to display plots side by side on the same graph.

We noted that, as sample size increases, the observed or empirical distribution of values more closely resembles the theoretical. The distributions of sample statistics such as the sample mean and sample variance are different from the distribution of individual values. In particular, under very general conditions with moderate-size samples, the distribution of the sample mean will take on the form of a normal distribution. We considered two nonparametric methods—the bootstrap and the permutation test—for estimating the values of distribution parameters and for testing hypotheses about them. We found that because of the variation from sample to sample, we run the risk of making one of two types of error when testing a hypothesis, each with quite different consequences. Normally when testing hypotheses, we set a bound called the significance level on the probability of making a Type I error and devise our tests accordingly.

Finally, we noted the relationship between our interval estimates and our hypothesis tests.

**Exercise 3.30.** Make a list of all the italicized terms in this chapter. Provide a definition for each one, along with an example.

**Exercise 3.31.** A farmer was scattering seeds in a field so they would be at least a foot apart 90% of the time. On the average, how many seeds should he sow per square foot?

The answer to Exercise 3.0 is yes, of course; an observation or even a sample of observations from one population may be larger than observations from another population even if the vast majority of observations are quite the reverse. This variation from observation to observation is why before a drug is approved for marketing its effects must be demonstrated in a large number of individuals and not just in one or two.



# Chapter 4

## Testing Hypotheses

**IN THIS CHAPTER, WE DEVELOP IMPROVED METHODS** for testing hypotheses by means of the bootstrap, introduce parametric hypothesis testing methods, and apply these and other methods to problems involving one sample, two samples, and many samples. We then address the obvious but essential question: How do we choose the method and the statistic that is best for the problem at hand?

### 4.1. ONE-SAMPLE PROBLEMS

A fast-food restaurant claims that 75% of its revenue is from the “drive-thru.” The owner collected two weeks’ worth of receipts from the restaurant and turned them over to you. Each day’s receipt shows the total revenue and the “drive-thru” revenue for that day.

The owner does not claim that their drive-thru produces 75% of their revenue, day in and day out, only that their overall average is 75%. In this section, we consider four methods for testing the restaurant owner’s hypothesis.

#### 4.1.1. Percentile Bootstrap

We’ve already made use of the percentile or uncorrected bootstrap on several occasions, first to estimate precision and then to obtain interval estimates for population parameters. Readily computed, the bootstrap seems ideal for use with the drive-thru problem. Still, if something seems too good to be true, it probably is. Unless corrected, bootstrap interval estimates are *inaccurate* (that is, they will include the true value of the unknown parameter less often than the stated confidence probability) and



*imprecise* (that is, they will include more erroneous values of the unknown parameter than is desirable). When the original samples contain less than a hundred observations, the confidence bounds based on the primitive bootstrap may vary widely from simulation to simulation.

#### 4.1.2. Parametric Bootstrap

If we know something about the population from which the sample is taken, we can improve our bootstrap confidence intervals, making them both more accurate (more likely to cover the true value of the population parameter) and more precise (narrower and thus less likely to include false values of the population parameter). For example, if we know that this population has an exponential distribution, we would use the sample mean to estimate the population mean. Then we would draw a series of random samples of the same size as our original sample from an exponential distribution whose mathematical expectation was equal to the sample mean to obtain a confidence interval for the population parameter of interest.

This parametric approach is of particular value when we are trying to estimate one of the tail percentiles such as  $P_{10}$  or  $P_{90}$ , for the sample alone seldom has sufficient information.

Here are the steps to deriving a parametric bootstrap:

1. Establish the appropriate distribution, let us say, the exponential.
2. Use Excel to calculate the sample average.
3. Use the sample average as an estimate of the population average in the following steps.
4. Select “NewModel” from the BoxSampler menu. Set ModelType to “Distribution.”
5. Set Distribution to “Exponential” on the BoxSampler worksheet. Set the value of the parameter  $\lambda$  to the sample average.
6. Set the sample size G11 to the size of your original sample. Set the number of simulations J11 to 400. Set the statistic K11 to = Function(Sample) where “Function” is the statistic for which you are attempting to derive a confidence interval. For example = Percentile(Sample, .25)

**Exercise 4.1.** Obtain a 90% confidence interval for the mean time to failure of a new component based on the following observations: 46 97 27 32 39 23 53 60 145 11 100 47 39 1 150 5 82 115 11 39 36 109 52 6 22 193 10 34 3 97 45 23 67 0 37

**Exercise 4.2.** Would you accept or reject the hypothesis at the 10% significance level that the mean time to failure in the population from which the sample depicted in Exercise 4.01 was drawn is 97?

**Exercise 4.3.** Obtain an 80% confidence interval with the parametric bootstrap for the IQR of the LSAT data. Careful: What would be the most appropriate continuous distribution to use?

### 4.1.3. Student's $t$

One of the first hypothesis tests to be developed was that of Student's  $t$ . This test, which dates back to 1908, takes advantage of our knowledge that the distribution of the mean of a sample is usually close to that of a normal distribution. When our observations are normally distributed, then the statistic

$$t = \frac{\bar{X} - \theta}{s/\sqrt{n}}$$

has a  $t$  distribution with  $n - 1$  degrees of freedom where  $n$  is the sample size,  $\theta$  is the population mean, and  $s$  is the standard deviation of the sample. Two things should be noted about this statistic:

1. Its distribution is independent of the unknown population variance.
2. If we guess wrong about the value of the unknown population mean and subtract a guesstimate of  $\theta$  smaller than the correct value, then the observed values of the  $t$  statistic will tend to be larger than the values predicted from a comparison with the Student's  $t$  distribution.

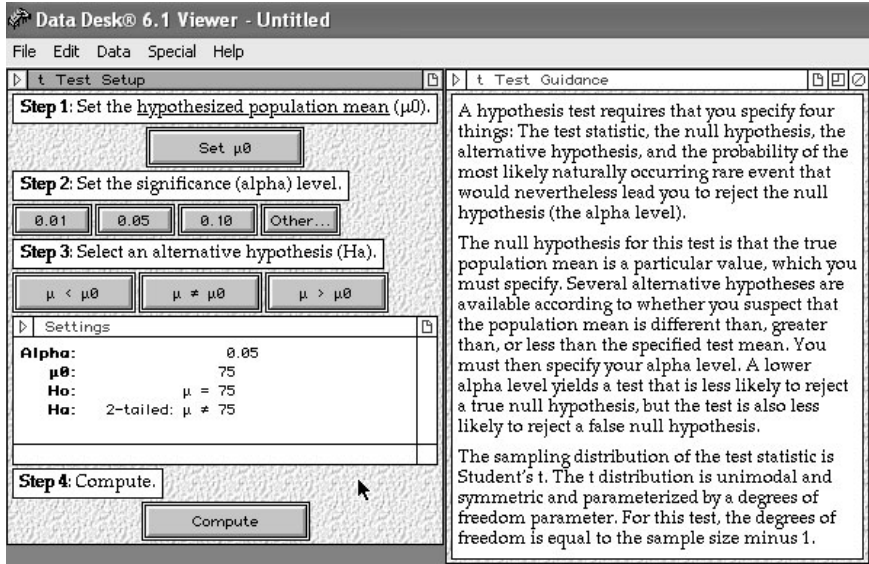
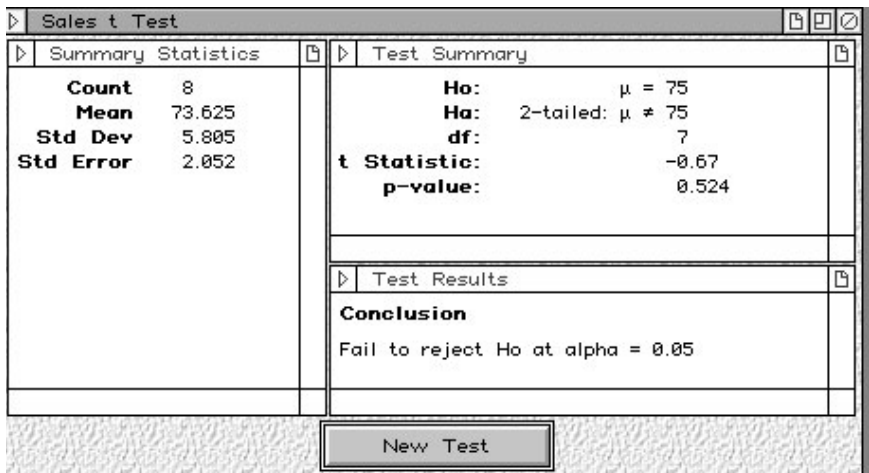
We can make use of this latter property to obtain a test of the hypothesis that the percentage of drive-in sales averages 75%, not just for our sample of sales data, but also for past and near-future sales. (Quick: Would this be a one-sided or a two-sided test?)

To perform the test, we pull down the DDXL menu, select first "Hypothesis Tests" and then "1 Var t Test." Completing the t Test Setup as shown in Fig. 4.1 yields the results in Fig. 4.2.

The sample estimate of \$73.62 is not significantly different from our hypothesis of \$75, the p value is close to 50%, and we accept the claim of the restaurant's owner.

**Exercise 4.4.** Would you accept or reject the restaurant owner's hypothesis at the 5% significance level after examining the entire two weeks' worth of data: 80, 81, 65, 72, 73, 69, 70, 79, 78, 62, 65, 66, 67, 75?

**Exercise 4.5.** In describing the extent to which we might extrapolate from our present sample of drive-in data, we used the qualifying phrase "near-future." Is this qualification necessary, or would you feel confident

FIGURE 4.1 Setting up a one-sample *t*-test using DDXL.FIGURE 4.2 Results of a one-sample *t*-test.

in extrapolating from our sample to all future sales at this particular drive-in? If not, why not?

**Exercise 4.6.** Although some variation is expected in the width of screws coming off an assembly line, the ideal width of this particular type of screw is 10.00 and the line should be halted if it looks as if the mean width of the screws produced will exceed 10.01 or fall below 9.99. On the basis of the following 10 observations, would you call for the line to halt so they can adjust the milling machine: 9.983, 10.020, 10.001, 9.981, 10.016, 9.992, 10.023, 9.985, 10.035, 9.960?

**Exercise 4.7.** In Exercise 4.6, what kind of economic losses do you feel would be associated with Type I and Type II errors?

## 4.2. COMPARING TWO SAMPLES

In this section, we'll examine the use of the binomial, Student's  $t$ , permutation methods, and the bootstrap for comparing two samples and then address the question of which is the best test to use.

### 4.2.1. Comparing Two Poisson Distributions

Suppose in designing a new nuclear submarine you become concerned about the amount of radioactive exposure that will be received by the crew. You conduct a test of two possible shielding materials. During 10 minutes of exposure to a power plant using each material in turn as a shield, you record 14 counts with material A and only four with experimental material B. Can you conclude that B is safer than A?

The answer lies not with the Poisson but the binomial. If the materials are equal in their shielding capabilities, then each of the 18 recorded counts is as likely to be obtained through the first material as through the second. In other words, under the null hypothesis you would be observing a binomial distribution with 18 trials, each with probability  $\frac{1}{2}$  of success or  $B(18, \frac{1}{2})$ .

I used just such a procedure in analyzing the results of a large-scale clinical trial involving some 100,000 service men and women who had been injected with either a new experimental vaccine or a saline control. Epidemics among service personnel can be particularly serious as they live in such close quarters. Fortunately, there were few outbreaks of the disease we were inoculating against during our testing period. Fortunate for the men and women of our armed services, that is.

When the year of our trial was completed, only 150 individuals had contracted the disease, which meant an effective sample size of 150. The differences in numbers of diseased individuals between the control and treated groups were not statistically significant.

**Exercise 4.8.** Can you conclude that material B is safer than A?

### 4.2.2. What Should We Measure?

Suppose you've got this strange notion that your college's hockey team is better than mine. We compare win/lost records for last season and see that while McGill won 11 of its 15 games, your team only won 8 of 14. But is this difference statistically significant? With the outcome of each game being success or failure, and successive games being independent of one another, it looks at first glance as if we have two series of binomial trials (as we'll see in a moment, this is highly questionable). We could derive confidence intervals for each of the two binomial parameters. If these intervals do not overlap, then the difference in win/loss records is statistically significant. But do win/loss records really tell the story?

Let's make the comparison another way by comparing total goals. McGill scored a total of 28 goals last season and your team 32. Using the approach described in the preceding section, we could look at this set of observations as a binomial with  $28 + 32 = 60$  trials, and test the hypothesis that  $p \leq \frac{1}{2}$  (that is, McGill is no more likely to have scored the goal than your team) against the alternative that  $p > \frac{1}{2}$ .

This latter approach has several problems. For one, your team played fewer games than McGill. But more telling, and the principal objection to all the methods we've discussed so far, the schedules of our two teams may not be comparable.

With binomial trials, the probability of success must be the same for each trial. Clearly, this is not the case here. We need to correct for the differences among opponents. After much discussion—what else is the off-season for?—you and I decide to award points for each game using the formula  $S = O + GF - GA$ , where GF stands for goals for, GA for goals against, and O is the value awarded for playing a specific opponent. In coming up with this formula and with the various values for O, we relied not on our knowledge of statistics but on our hockey expertise. This reliance on domain expertise is typical of most real-world applications of statistics.

The point totals we came up with read like this

McGill            4, -2, 1, 3, 5, 5, 0, -1, 6, 2, 2, 3, -2, -1, 4  
 Your School    3, 4, 4, -3, 3, 2, 2, 2, 4, 5, 1, -2, 2, 1

Curiously, your school's first four point totals, all involving games against teams from other leagues, were actually losses, their high point value being the result of the high caliber of the opponents. I'll give you guys credit for trying.

### 4.2.3. Permutation Monte Carlo

Straightforward application of the permutation methods discussed in Section 3.6.1 to the hockey data is next to impossible. Imagine how many years it would take us to look at all  $\binom{14+15}{15}$  possible rearrangements!

What we can do today—something not possible with the primitive calculators that were all that was available in the 1930s when permutation methods were first introduced—is to look at a large random sample of rearrangements.

We prepare to reshuffle the data as shown in Fig. 4.3 with the following steps:

	A	B	C	D	E	F	G	H	I
1	Which is The Best Team?								
2	McGill	Your School							
3	4	3							
4	-2	4							
5	1	4							
6	3	-3							
7	5	3							
8	5	2							
9	0	2							
10	-1	2							
11	6	4							
12	2	5							
13	2	1							
14	3	-2							
15	-2	2							
16	-1	1							
17	4								
18									

**Matrix Shuffle**

Input Range:

Top Left Cell of Output Range:

Normal Shuffle  
 Shuffle Rows as Units  
 Shuffle Within Rows  
 Shuffle Columns as Units  
 Shuffle Within Columns  
 Shuffle Single Column

Stratified Sample

Shuffle blank cells in data

Buttons: OK, Cancel, Help

FIGURE 4.3 Preparing to shuffle the data.

1. Use the cursor to outline the two columns that we wish to shuffle, that is, to rearrange again in two columns, one with 15 observations and one with 14.
2. Press the S on the Resampling Stats in Excel menu.
3. Note the location of the top left cell where you wish to position the reshuffled data.
4. Click OK.

Our objective is to see in what proportion of randomly generated rearrangements the sum of the observations in the first of the sample equals or exceeds the original sum of the observations in the first sample. Once a single rearrangement has been generated, we enter the following formula in any convenient empty cell:

$$=IF(SUM(C3:C17)>=SUM(A3:A17),1,0)$$

Select RS (repeat and score) and set the number of trials to 400. When we click OK, 400 random rearrangements are generated and the preceding formula is evaluated for each rearrangement and the result placed in the first column of a separate “Results” worksheet.

Our  $p$  value is the proportion of 1s among the 1s and 0s in this column. We calculate it with the following formula:

$$=SUM(A1:A400)/400$$

**Exercise 4.9.** Show that we would have gotten exactly the same  $p$  value had we used the difference in means between the samples instead of the sum of the observations in the first sample as our permutation test statistic.

**Exercise 4.10.** (for mathematics and statistics majors only) Show that we would have gotten exactly the same  $p$  value had we used the  $t$  statistic as our test statistic.

**Exercise 4.11.** Use the Monte Carlo approach to rearrangements to test the hypothesis that McGill’s hockey team is superior to your school’s.

**Exercise 4.12.** Compare the 90% confidence intervals for the variance of the population from which the following sample of billing data was taken for a) the original primitive bootstrap, b) the parametric bootstrap, assuming the billing data are normally distributed.

**Hospital Billing Data**

4181, 2880, 5670, 11620, 8660, 6010, 11620, 8600, 12860, 21420, 5510, 12270, 6500, 16500, 4930, 10650, 16310, 15730, 4610, 86260, 65220, 3820, 34040, 91270, 51450, 16010, 6010, 15640, 49170, 62200, 62640, 5880, 2700, 4900, 55820, 9960, 28130, 34350, 4120, 61340, 24220, 31530, 3890, 49410, 2820, 58850, 4100, 3020, 5280, 3160, 64710, 25070

**4.2.4. Two-Sample t-Test**

For the same reasons that Student's  $t$  was an excellent choice in the one-sample case, it is recommended for comparing samples of continuous data from two populations, providing that the only difference between the two is in their mean value, that is, the distribution of one is merely shifted with respect to the other so that  $F_1[x] = F_2[x - \Delta]$ . The test statistic is  $\frac{\bar{X}_1 - \bar{X}_2}{\hat{s}}$ , where  $\hat{s}$  is an estimate of the *standard error* of the numerator:

$$\hat{s} = \sqrt{\frac{\sum (X_{1j} - \bar{X}_1)^2 / (n_1 - 1) + \sum (X_{2j} - \bar{X}_2)^2 / (n_2 - 1)}{n_1 + n_2 - 2}}$$

Note that the square of the  $t$  statistic is the ratio of the variance between the samples from your school and McGill to the variance within these samples.

**Exercise 4.13.** Basing your decision on the point totals, use Student's  $t$  to test the hypothesis that McGill's hockey team is superior to your school's. Is this a one-sided or a two-sided test? (This time when you use DDXL Hypothesis Tests, select "2 Var t Test.")

**4.3. WHICH TEST SHOULD WE USE?**

Four different tests were used for our two-population comparisons. Two of these were *parametric* tests that obtained their  $p$  values by referring to parametric distributions such as the binomial and Student's  $t$ . Two were *resampling methods*—bootstrap and permutation test—that obtained their  $p$  values by sampling repeatedly from the data at hand.

In some cases, the choice of test is predetermined, for example, when the observations take or can be reduced to those of a binomial distribution. In other instances, we need to look more deeply into the consequences of our choice. In particular, we need to consider the assumptions



under which the test is valid, the effect of violations of these assumptions, and the Type I and Type II errors associated with each test.

#### 4.3.1. $p$ Values and Significance Levels

In the preceding sections we have referred several times to  $p$  values and significance levels. We have used both in helping us to make a decision whether to accept or reject a hypothesis and, in consequence, to take a course of action that might result in gains or losses.

To see the distinction between the two concepts, please go through the following steps:

1. Use BoxSampler to generate a sample of size 10 from a Normal Distribution with mean 0.5 and variance 1.
2. Use this sample and the  $t$ -test to test the hypothesis that the mean of the population from which this sample was drawn was 0 (not 0.5). Write down the value of the  $t$  statistic and of the  $p$  value.
3. Repeat Step 1.
4. Repeat Step 2.

The composition of the two samples varies, the value of the  $t$  statistic varies, the  $p$  values vary, and the boundaries of the confidence interval vary. What remains unchanged is the *significance level* of  $100\% - 95\% = 5\%$  that is used to make decisions.

You aren't confined to a 5% significance level. In clinical trials of drug effectiveness, one might use a significance level of 10% in pilot studies but would probably insist on a significance level of 1% before investing large amounts of money in further development.

In summary,  $p$  values vary from sample to sample, whereas significance levels are fixed.

Significance levels establish limits on the overall frequency of Type I errors. The significance levels and confidence bounds of parametric and permutation tests are exact only if all the assumptions that underlie these tests are satisfied. Even when the assumptions that underlie the bootstrap are satisfied, the claimed significance levels and confidence bounds of the bootstrap are only approximations. The greater the number of observations in the original sample, the better this approximation will be.

#### 4.3.2. Test Assumptions

Virtually all statistical procedures rely on the assumption that our observations are independent of one another. When this assumption fails, the computed  $p$  values may be far from accurate, and a specific significance level cannot be guaranteed.

All statistical procedures require that at least one of the following successively stronger assumptions be satisfied under the hypothesis of no differences among the populations from which the samples are drawn:

1. The observations all come from distributions that have the same value of the parameter of interest.
2. The observations are *exchangeable*, that is, each rearrangement of labels is equally likely.
3. The observations are *identically distributed* and come from a distribution of known form.

The first assumption is the weakest. If this assumption is true, a non-parametric bootstrap test<sup>1</sup> will provide an exact significance level with very large samples. The observations may come from different distributions, providing that they all have the same parameter of interest. In particular, the nonparametric bootstrap can be used to test whether the expected results are the same for two groups even if the observations in one of the groups are more variable than they are in the other.<sup>2</sup>

If the second assumption is true, the first assumption is also true. If the second assumption is true, a permutation test will provide exact significance levels even for very small samples.

The third assumption is the strongest assumption. If it is true, the first two assumptions are also true. This assumption must be true for a parametric test to provide an exact significance level.

An immediate consequence is that if observations come from a multi-parameter distribution such as the normal, then all parameters, not just the one under test, must be the same for all observations under the null hypothesis. For example, a *t*-test comparing the *means* of two populations requires that the *variances* of the two populations be the same.

### 4.3.3. Robustness

When a test provides almost exact significance levels despite a violation of the underlying assumptions, we say that it is *robust*. Clearly, the nonparametric bootstrap is more robust than the parametric because it has fewer assumptions. Still, when the number of observations is small, the parametric bootstrap, which makes more effective use of the data, will be preferable, providing enough is known about the shape of the distribution from which the observations are taken.

<sup>1</sup> Any bootstrap but the parametric bootstrap.

<sup>2</sup> We need to modify our testing procedure if we suspect this to be the case; see Chapter 8.

When the variances of the populations from which the observations are drawn are not the same, the significance level of the bootstrap is not affected. Bootstrap samples are drawn separately from each population. Small differences in the variances of two populations will leave the significance levels of permutation tests relatively unaffected, but they will no longer be exact. Student's  $t$  should not be used when there are clear differences in the variances of the two groups.

On the other hand, Student's  $t$  is the exception to the rule that parametric tests should only be used when the distribution of the underlying observations is known. Student's  $t$  tests for differences between means, and means, as we've already noted, tend to be normally distributed even when the observations they summarize are not.

#### 4.3.4. Power of a Test Procedure

Statisticians call the probability of rejecting the null hypothesis when an alternative hypothesis is true the *power* of the test. If we were testing a food additive for possible carcinogenic (cancer producing) effects, this would be the probability of detecting a carcinogenic effect. The power of a test equals one minus the probability of making a Type II error. The greater the power, the smaller the Type II error, the better off we are.

Power depends on all of the following:

1. The true value of the parameter being tested—the greater the gap between our primary hypothesis and the true value, the greater the power will be. In our example of a carcinogenic substance, the power of the test would depend on, whether the substance was a strong or a weak carcinogen and whether its effects were readily detectable.
2. The significance level—the higher the significance level (10% rather than 5%), the larger the probability of making a Type I error we are willing to accept, and the greater the power will be. In our example, we would probably insist on a significance level of 1%.
3. The sample size—the larger the sample, the greater the power will be. In our example of a carcinogenic substance, the regulatory commission (the FDA in the United States) would probably insist on a power of 80%. We would then have to increase our sample size in order to meet their specifications.
4. The method used for testing. Obviously, we want to use the most powerful possible method.

**Exercise 4.14.** To test the hypothesis that consumers can't tell your cola from Coke, you administer both drinks in a blind tasting to 10 people selected at random. A) To ensure that the probability of a Type I error is just slightly more than 5%, how many people should correctly identify the

glass of Coke before you reject this hypothesis? B) What is the power of this test if the probability of an individual correctly identifying Coke is 75%?

**Exercise 4.15.** What is the power of the test in Exercise 4.14 if the probability of an individual correctly identifying Coke is 90%?

**Exercise 4.16.** If you test 20 people rather than 10, what will be the power of a test at the 5% significance level if the probability of correctly identifying Coke is 75%?

**Exercise 4.17.** Physicians evaluate diagnostic procedures on the basis of their “sensitivity” and “selectivity.”

Sensitivity is defined as the percentage of diseased individuals that are correctly diagnosed as such. Is sensitivity related to significance level and power? How?

Selectivity is defined as the percentage of those diagnosed as suffering from a given disease that actually have the disease. Can selectivity be related to the concepts of significance level and power? If so, how?

**Exercise 4.18.** Suppose we wish to test the hypothesis that a new vaccine will be more effective than the old vaccine in preventing infectious pneumonia. We decide to inject some 1000 patients with the old vaccine and 1000 with the new and follow them for one year. Can we guarantee the power of the resulting hypothesis test?

**Exercise 4.19.** Show that the power of a test can be compared to the power of an optical lens in at least one respect.

### 4.3.5. Testing for Correlation

To see how we would go about finding the most powerful test in a specific case, consider the problem of deciding whether two variables are correlated. Let’s take another look at the data from my sixth-grade classroom. The arm span and height of the five shortest students in my sixth grade class are (139, 137), (140, 138.5), (141, 140), (142.5, 141), (143.5, 142). Both arm spans and heights are in increasing order. Is this just coincidence? Or is there a causal relationship between them or between them and a third hidden variable? What is the probability that an event like this could happen by chance alone?

The test statistic of choice is the Pitman correlation,  $S = \sum_{i=1}^n a_i b_i$ , where  $(a_k, b_k)$  denotes the pair of observations made on the  $k$ th individual. To prove to your own satisfaction that  $S$  will have its maximum when both arm spans and heights are in increasing order, imagine that the set of arm spans  $\{a_k\}$  denotes the widths and  $\{b_k\}$  the heights of a set of rectangles. The area inside the rectangles,  $S$ , will be at its maximum when the smallest width is paired with the smallest height, and so forth. If your intuition is more geometric than algebraic, prove this result by sketching the rectangles on a piece of graph paper.

We could list all possible permutations of both arm span and height along with the value of  $S$ , but this won't be necessary. We can get exactly the same result if we fix the order of one of the variables, the height, for example, and look at the  $5! = 120$  ways in which we could rearrange the arm span readings:

(140, 137) (139, 138.5) (141, 140) (142.5, 141) (143.5, 142)  
 (141, 137) (140, 138.5) (139, 140) (142.5, 141) (143.5, 142)

and so forth.

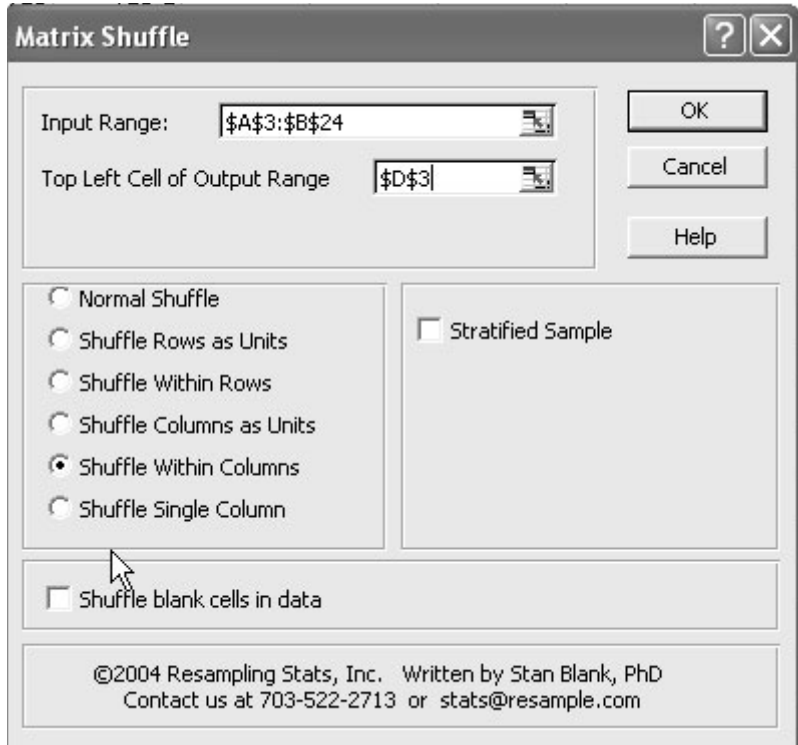
Obviously, the arrangement we started with is the most extreme, occurring exactly one time in 120 by chance alone. Applying this same test to all 22 pairs of observations, we find the odds are less than 1 in a million that what we observed occurred by chance alone and conclude that arm span and height are directly related.

To perform a Monte Carlo estimate of the  $p$  values, we proceed as in Section 4.2.3 with two modifications. We begin by outlining the columns that we wish to shuffle. But when we complete the Matrix Shuffle form, we specify "Shuffle within Columns" as shown in Fig. 4.4. And we compute Excel's **Correl()** function repeatedly.

Note that we would get exactly the same  $p$  value if we used as our test statistic the Pearson correlation  $\rho = \sum_{i=1}^n a_i b_i / \sqrt{\text{Var}[a] * \text{Var}[b]}$ . This is

because the variances of  $a$  and  $b$  are left unchanged by rearrangements. A rearrangement that has a large value of  $S$  will have a large value of  $\rho$  and vice versa.

**Exercise 4.20.** The correlation between the daily temperatures in Cairns and Brisbane is 0.29 and between Cairns and Sydney is 0.52. Or should that be the other way around?



**FIGURE 4.4** Setting up test for correlation between columns.

**Exercise 4.21.** Do DDT residues have a deleterious effect on the thickness of a cormorant's eggshell? (Is this a one-sided or a two-sided test?)

DDT residue in yolk (ppm)	65	98	117	122	393
Thickness of shell (mm)	.52	.53	.49	.49	.37

**Exercise 4.22.** Is there a statistically significant correlation between the LSAT score and the subsequent GPA in law school?

**Exercise 4.23.** If we find that there is a statistically significant correlation between the LSAT score and the subsequent GPA, does this mean the LSAT score of a prospective law student will be a good predictor of that student's subsequent GPA?

#### 4.4. SUMMARY AND REVIEW

In this chapter, we derived permutation, parametric, and bootstrap tests of hypothesis for a single sample, for comparing two samples, and for bivariate correlation. We showed how to improve the accuracy and precision of bootstrap confidence intervals. We explored the relationships and distinctions among  $p$  values, significance levels, alternative hypotheses, and sample sizes. And we provided some initial guidelines to use in the selection of the appropriate test.

**Exercise 4.24.** Make a list of all the italicized terms in this chapter. Provide a definition for each one, along with an example.

**Exercise 4.25.** Some authorities have suggested that when we estimate a  $p$  value via a Monte Carlo as in Section 4.2.3 we should include the original observations as one of the rearrangements. Instead of reporting the  $p$  value as  $\text{cnt}/N$ , we would report it as  $(\text{cnt} + 1)/(N + 1)$ . Explain why this would give a false impression. (Hint: Reread Chapter 2 if necessary.)

**Exercise 4.26.** Efron and Tibshirani (1993) report the survival times in days for a sample of 16 mice undergoing a surgical procedure. The mice were randomly divided into two groups. The following survival times in days were recorded for a group of seven mice that received a treatment expected to prolong their survival:

● 94,197,16,38,99,141,23

The second group of nine mice underwent surgery without the treatment and had these survival times in days:

● 52,104,146,10,51,30,40,27,46

Provide a 75% confidence interval for the difference in mean survival days for the sampled population based on 1000 bootstrap samples.

**Exercise 4.27.** Which test would you use for a comparison of the following treated and control samples?

control = 4,6,3,4,7,6

treated = 14,6,3,12,7,15.

## Chapter 5

# Designing an Experiment or Survey

**SUPPOSE YOU WERE A CONSULTING STATISTICIAN<sup>1</sup>** and were given a data set to analyze. What is the first question you would ask? “What statistic should I use?” No, your first question always should be, “How were these data collected?”

Experience teaches us that garbage in, garbage out or GIGO. To apply statistical methods, you need to be sure that samples have been drawn at random from the population(s) you want represented and are representative of that population. You need to be sure that observations are independent of one another and that outcomes have not been influenced by the actions of the investigator or survey taker.

Many times people who consult statisticians don't know the details of the data collection process, or they do know and look guilty and embarrassed when asked. All too often, you'll find yourself throwing your hands in the air and saying, “If only you'd come to me to design your experiment in the first place.”

The purpose of this chapter is to take you step by step through the design of an experiment and a survey. You'll learn the many ways in which an experiment can go wrong. And you'll learn the right things to do to ensure that your own efforts are successful.

---

<sup>1</sup> The idea of having a career as a consulting statistician may strike you as laughable or even distasteful. I once had a student who said he'd rather eat worms and die. Suppose then that you've eaten worms and died, only to wake to discover that reincarnation is real and that to expiate your sins in the previous life you've been reborn as a consulting statistician. I'm sure that's what must have happened in my case.



## 5.1. THE HAWTHORNE EFFECT

The original objective of the industrial engineers at the Hawthorne plant of Western Electric was to see whether a few relatively inexpensive improvements would increase workers' productivity. They painted the walls green, and productivity went up. They hung posters, and productivity went up. Then, just to prove how important bright paint and posters were to productivity, they removed the posters and repainted the walls a dull gray, only to find that, once again, *productivity went up!*

Simply put, these industrial engineers had discovered that the mere act of paying attention to a person modifies his behavior. (Note: The same is true for animals.)

You've probably noticed that you respond similarly to attention from others, though not always positively. Taking a test under the watchful eye of an instructor is quite different from working out a problem set in the privacy of your room.

Physicians and witch doctors soon learn that merely giving a person a pill (any pill) or dancing a dance often results in a cure. This is called the *placebo* effect. If patients think they are going to get better, they do get better. Thus regulatory agencies insist that, before they approve a new drug, it be tested side by side with a similar looking, similar tasting *placebo*. If the new drug is to be taken twice a day in tablet form, then the placebo must also be given twice a day, also as a tablet, and not as a liquid or an injection. And, most important, the experimental subject should not be aware of which treatment she is receiving. Studies in which the treatment is concealed from the subject are known as *single-blind* studies.

The doctor's attitude is as important as the treatment. If part of the dance is omitted—a failure to shake a rattle, why bother if the patient is going to die anyway—the patient may react differently. Thus the agencies responsible for regulating drugs and medical devices (in the United States this would be the FDA) now also insist that experiments be *double blind*. Neither the patient nor the doctor (or whoever administers the pill to the patient) should know whether the pill that is given the patient is an active drug or a placebo. If the patient searches the doctor's face for clues—Will this experimental pill really help me?—she'll get the same response whether she is in the treatment group or is one of the *controls*.

Note: The double-blind principle also applies to experimental animals. Dogs and primates are particularly sensitive to their handlers' attitudes.

### 5.1.1. Crafting an Experiment

In the very first set of clinical data that was brought to me for statistical analysis, a young surgeon described the problems he was having with his

chief of surgery. “I’ve developed a new method for giving arteriograms that I feel can cut down on the necessity for repeated amputations. But my chief will only let me try out the technique on patients who he feels are hopeless. Will this affect my results?” It would, and it did. Patients examined by the new method had a very poor recovery rate. But, of course, the only patients who’d been examined by the new method were those with a poor prognosis. The young surgeon realized that he would not be able to test his theory until he was able to assign patients to treatment at random.

Not incidentally, it took us three more tries until we got this particular experiment right. In our next attempt, the chief of surgery—Mark Craig of St. Eligius in Boston—announced that he would do the “random” assignments. He finally was persuaded to let me make the assignment by using a table of random numbers. But then he announced that he, and not the younger surgeon, would perform the operations on the patients examined by the traditional method to make sure “they were done right.” Of course, this turned a comparison of methods into a comparison of surgeons and intent.

In the end, we were able to create the ideal “double-blind” study: The young surgeon performed all the operations, but the incision points were determined by his chief after examining one or the other of the two types of arteriogram.

**Exercise 5.1.** Each of the following studies is fatally flawed. Can you tell what the problem is in each instance and, as important, why it is a problem?

1. Class action. Larry the Lawyer was barely paying his rent when he got the bright idea of looking through the county-by-county leukemia rates for our state. He called me a week later and asked what I thought of the leukemia rate in Kay County. I gave a low whistle. “Awfully high,” I said.

The next time I talked to Larry, he seemed happy and prosperous. He explained that he’d gone to Kay County once he’d learned that the principal employer in that area was a multinational chemical company. He’d visited all the families whose kids had been diagnosed with leukemia and signed them up for a class action suit. The company had quickly settled out of court when they looked at the figures.

“How’d you find out about Kay County?” I asked.

“Easy, I just ordered all the counties in the state by their leukemia rates, and Kay came out on top.”

2. Controls. Donald routinely tested new drugs for toxicity by injecting them in mice. In each case, he’d take five animals from a cage and

inject them with the drug. To be on the safe side, he'd take the next five animals from the cage, inject them with a saline solution, and use them for comparison purposes.

3. Survey. Reasoning, correctly, that he'd find more students home at dinnertime, Tom brought a set of survey forms back to his fraternity house and interviewed his frat brothers one by one at the dinner table.
4. Treatment Allocation. Fully aware of the influence that a physician's attitude could have on a patient's recovery, Betty, a biostatistician, provided the investigators in a recent clinical trial with bottles of tablets that were labeled only A or B.
5. Clinical Trials. Before a new drug can be marketed, it must go through a succession of clinical trials. The first set of trials (phase I) is used to establish the maximum tolerated dose. They are usually limited to 25 or so test subjects who will be observed for periods of several hours to several weeks. The second set of trials (phase II) is used to establish the minimum effective dose; they also are limited in duration and in the number of subjects involved. Only in phase III are the trials expanded to several hundred test subjects who will be followed over a period of months or even years. Up until the 1990s, only males were used as test subjects, in order to spare women the possibility of unnecessary suffering.
6. Comparison. Contrary to what one would expect from the advances in medical care, there were 2.1 million deaths from all causes in the U.S. in 1985, compared to 1.7 million in 1960.
7. Survey. The Federal Trade Commission surveyed former correspondence school students to see how they felt about the courses they had taken some two to five years earlier.<sup>2</sup> The survey was accompanied by a form letter signed by an FTC attorney that began, "The Bureau of Consumer Protection is gathering information from those who enrolled in . . . to determine if any action is warranted." Questions were multiple choice and did not include options for "I don't know" or "I don't recall."

## 5.2. DESIGNING AN EXPERIMENT OR SURVEY

Before you complete a single data collection form:

1. Set forth your objectives and the use you plan to make of your research.
2. Define the population(s) to which you will apply the results of your analysis.
3. List all possible sources of variation.

---

<sup>2</sup> Macmillan, Inc. 96 F.T.C. 208 (1980).

4. Decide how you will cope with each source. Describe what you will measure and how you will measure it. Define the experimental unit and all end points.
5. Formulate your hypothesis and all of the associated alternatives. Define your end points. List possible experimental findings, along with the conclusions you would draw and the actions you would take for each of the possible results.
6. Describe in detail how you intend to draw a representative random sample from the population.
7. Describe how you will ensure the independence of your observations.

### 5.2.1. Objectives

In my experience as a statistician, the people who come to consult me before they do an experiment (an all-too-small minority of my clients) aren't always clear about their objectives. I advise them to start with their reports, to write down what they would most like to see in print. For example,

Fifteen thousand of 17,500 surveys were completed and returned. Over half of the respondents were between the ages of 47 and 56. Thirty-six percent (36%) indicated that they were currently eligible or would be eligible for retirement in the next three years. However, only 25% indicated they intended to retire in that time. Texas can anticipate some 5000 retirees in the next three years.

or

743 patients self-administered our psyllium preparation twice a day over a three-month period. Changes in the Klozner–Murphy self-satisfaction scale over the course of treatment were compared with those of 722 patients who self-administered an equally foul-tasting but harmless preparation over the same time period.

All patients in the study reported an increase in self-satisfaction, but the scores of those taking our preparation increased an average of  $2.3 \pm 0.5$  points more than those in the control group.

Adverse effects included. . . .

If taken as directed by a physician, we can expect those diagnosed with. . . .

I have my clients write in exact numerical values for the anticipated outcomes—their best guesses, as these will be needed when determining sample size. My clients go over their reports several times to ensure they’ve included all end points and as many potential discoveries as they can—“Only 25% indicated an intent to retire in that time.” Once the report is fleshed out completely, they know what data need to be collected and do not waste their time and their company’s time on unnecessary or redundant effort.

**Exercise 5.2.** Throughout this chapter, you’ll work on the design of a hypothetical experiment or survey. If you are already well along in your studies, it could be an actual one! Start now by writing the results section.

### 5.2.2. Sample From the Right Population

Be sure you will be sampling from the population of interest as a whole rather than from an unrepresentative subset of that population. The most famous blunder along these lines was basing the forecast of Dewey over Truman in the 1948 U.S. presidential election on a telephone survey: Those who owned a telephone and responded to the survey favored Dewey; those who voted did not.

An economic study may be flawed because we have overlooked the homeless. This was among the principal arguments the cities of New York and Los Angeles advanced against the use of the 1990 and 2000 census to determine the basis for awarding monies to cities. See *City of New York v. Dept of Commerce*.<sup>3</sup>

An astrophysical study was flawed because of overlooking galaxies whose central surface brightness was very low. And the FDA’s former policy of permitting clinical trials to be limited to men (see Exercise 5.1, examples) was just plain foolish.

Plaguing many surveys are the uncooperative and the nonresponder. Invariably, follow-up surveys of these groups show substantial differences from those who responded readily the first time around. These follow-up surveys aren’t inexpensive—compare the cost of mailing out a survey to telephoning or making face-to-face contact with a nonresponder. But if one doesn’t make these calls, one may get a completely unrealistic picture of how the population as a whole would respond.

**Exercise 5.3.** You be the judge. In each of the following cases, how would you rule?

<sup>3</sup> 822 F. Supp. 906 (E.D.N.Y., 1993).

- A. The trial of *People v. Sirhan*<sup>4</sup> followed the assassination of presidential candidate Robert Kennedy. The defense appealed the guilty verdict, alleging that the jury was a nonrepresentative sample and offering anecdotal evidence based on the population of the northern United States. The prosecution said, so what, our jury was representative of Los Angeles where the trial was held. How would you rule? Note that the Sixth Amendment to the Constitution of the United States provides that

A criminal defendant is entitled to a jury drawn from a jury panel which includes jurors residing in the geographic area where the alleged crime occurred.

- B. In *People v. Harris*,<sup>5</sup> a survey of trial court jury panels provided by the defense showed a significant disparity from census figures. The prosecution contended that the survey was too limited, being restricted to the Superior Courts in a single district, rather than being county wide. How would you rule?
- C. Amstar Corporation claimed that “Domino’s Pizza” was too easily confused with its own use of the trademark “Domino” for sugar.<sup>6</sup> Amstar conducted and offered in evidence a survey of heads of households in ten cities. Domino objected to this survey, pointing out that it had no stores or restaurants in eight of these cities and in the remaining two their outlets had been open less than three months. Domino provided a survey it had conducted in its pizza parlors, and Amstar objected. How would you rule?

**Exercise 5.4.** Describe the population from which you plan to draw a sample in your hypothetical experiment. Is this the same population you would extend the conclusions to in your report?

### The Drunk and The Lamppost

There’s an old joke dating back to at least the turn of the previous century about the drunk whom the police officer found searching for his wallet under the lamppost. The policeman offers to help and after searching on hands and knees for fifteen minutes without success asks the inebriated gentleman

<sup>4</sup> 7 Cal.3d 710, 102 Cal. Rptr.385 (1972), *cert. denied*, 410 U.S. 947.

<sup>5</sup> 36 Cal.3d 36, 201 Cal. Rptr 782 (1984), *cert. denied* 469 U.S. 965, *appeal to remand* 236 Cal. Rptr 680, 191 Cal. App. 3d 819, *appeal after remand*, 236 Cal. Rptr 563, 217 Cal. App. 3d 1332.

<sup>6</sup> *Amstar Corp. v. Domino’s Pizza, Inc.*, 205 U.S.P.Q 128 (N.D. Ga. 1979), *rev’d*, 615 F. 2d 252 (5th Cir. 1980).

just exactly where he lost his wallet. The drunk points to the opposite end of the block. “Then why were you searching over here?!” the policeman asks.

“The light’s better.”

It’s amazing how often measurements are made because they are convenient (inexpensive and/or quick to make) rather than because they are directly related to the object of the investigation. Your decisions as to what to measure and how to measure it require as much or more thought as any other aspect of your investigation.

### 5.2.3. Coping with Variation

As noted in the very first chapter of this text, you should begin any investigation where variation may play a role by listing all possible sources of variation—in the environment, in the observer, in the observed, and in the measuring device. Consequently, you need to have a thorough understanding of the domain—biological, psychological, or seismological—in which the inquiry is set.

Will something as simple as the time of day affect results? Body temperature and the incidence of mitosis both depend on the time of day. Retail sales and the volume of mail both depend on the day of the week. In studies of primates (including you) and hunters (tigers, mountain lions, domestic cats, dogs, wolves, and so on) the sex of the observer will make a difference.

Statisticians have found four ways for coping with individual-to-individual and observer-to-observer variation:

1. *Controlling.* Making the environment for the study—the subjects, the manner in which the treatment is administered, the manner in which the observations are obtained, the apparatus used to make the measurements, and the criteria for interpretation—as uniform and homogeneous as possible.
2. *Blocking.* A clinician might stratify the population into subgroups based on such factors as age, sex, race, and the severity of the condition and to restrict subsequent comparisons to individuals who belong to the same subgroup. An agronomist would want to stratify on the basis of soil composition and environment.
3. *Measuring.* Some variables such as cholesterol level or the percentage of CO<sub>2</sub> in the atmosphere can take any of a broad range of values and don’t lend themselves to blocking. As we show in Chapter 6, statisticians have methods for correcting for the values taken by these *covariates*.

4. *Randomizing.* Randomly assign patients to treatment within each block or subgroup so that the innumerable factors that can be neither controlled nor observed directly are as likely to influence the outcome of one treatment as another.

**Exercise 5.5.** List all possible sources of variation for your hypothetical experiment and describe how you will cope with each one.

#### 5.2.4. Matched Pairs

One of the best ways to eliminate a source of variation and the errors in interpretation associated with it is through the use of matched pairs. Each subject in one group is matched as closely as possible by a subject in the treatment group. If a 45-year-old black male hypertensive is given a blood-pressure lowering pill, then we give a second similarly built 45-year-old black male hypertensive a placebo.

Consider the case of a fast-food chain that is interested in assessing the effect of the introduction of a new sandwich on overall sales. To do this experiment, they designate a set of outlets in different areas—two in the inner city, two in the suburbs, two in small towns, and two located along major highways. A further matching criterion is that the overall sales for the members of each pair before the start of the experiment were approximately the same for the months of January through March. During the month of April, the new sandwich is put on sale at one of each pair of outlets. At the end of the month, the results are recorded for each matched pair of outlets.

To analyze this data, we consider the 28 possible rearrangements that result from the possible exchanges of labels within each matched pair of observations. We proceed as in Section 4.3.5, only we select “Shuffle within Rows” from the Matrix Shuffle form.

**Exercise 5.6.** In Exercise 5.5, is the correct p value 0.98, 0.02, or 0.04?

**Exercise 5.7.** Did the increased sales for the new menu justify the increased cost of \$1200 per location?

**TABLE 5.1**

	1	2	3	4	5	6	7	8
New Menu	48722	28965	36581	40543	55423	38555	31778	45643
Standard	46555	28293	37453	38324	54989	35687	32000	43289



### 5.2.5. The Experimental Unit

A scientist repeatedly subjected a mouse named Harold to severe stress. She made a series of physiological measurements on Harold, recording blood pressure, cholesterol levels, and white blood cell counts both before and after stress was applied for a total of 24 observations. What was the sample size?

Another experimenter administered a known mutagen—a substance that induces mutations—into the diet of a pregnant rat. When the rat gave birth, the experimenter took a series of tissue samples from each of the seven offspring, two from each of eight body regions. What was the sample size?

In each of the preceding examples, the sample size was one. In the first example, the sole *experimental unit* was Harold. In the second example, the experimental unit was a single pregnant rat. Would stress have affected a second mouse the same way? We don't know. Would the mutagen have caused similar damage to the offspring of a different rat? We don't know. We do know there is wide variation from individual to individual in their responses to changes in the environment. With data from only a single individual in hand, I'd be reluctant to draw any conclusions about the population as a whole.

**Exercise 5.8.** Suppose we are testing the effect of a topical ointment on pink eye. Is each eye a separate experimental unit, or each patient?

### 5.2.6. Formulate Your Hypotheses

In translating your study's objectives into hypotheses that are testable by statistical means, you need to satisfy all of the following:

- The hypothesis must be numeric in form and must concern the value of some population parameter. Examples: More than 50% of those registered to vote in the state of California prefer my candidate. The arithmetic average of errors in tax owed that are made by U.S. taxpayers reporting \$30,000 to \$50,000 income is less than \$50. The addition of vitamin E to standard cell growth medium will increase the life span of human diploid fibroblasts by no less than 30 generations. Note in these examples that we've also tried to specify the population from which samples are taken as precisely as possible.
- There must be at least one meaningful numeric alternative to your hypothesis.
- It must be possible to gather data to test your hypothesis.

The statement "Redheads are sexy" is not a testable hypothesis. Nor is the statement "Everyone thinks redheads are sexy." Can you explain why?

The statement “At least 80% of Reed College students think redheads are sexy” is a testable hypothesis.

You should also decide at the same time as you formulate your hypotheses whether the alternatives of interest are one-sided or two-sided, ordered or unordered.

**Exercise 5.9.** Are the following testable hypotheses? Why or why not?

- a. A large meteor hitting the Earth would dramatically increase the percentage of hydrocarbons in the atmosphere.
- b. Our candidate can be expected to receive votes in the coming election.
- c. Intelligence depends more on one’s genes than on one’s environment.

### 5.2.7. What Are You Going to Measure?

To formulate a hypothesis that is testable by statistical means, you need decide on the variables you plan to measure. Perhaps your original hypothesis was that men are more intelligent than women. To put this in numerical terms requires a scale by which intelligence may be measured. Which of the many scales do you plan to use, and is it really relevant to the form of intelligence you had in mind?

Be direct. To find out which drugs individuals use and in what combinations, which method would yield more accurate data: a) a mail survey of households, b) surveying customers as they step away from a pharmacy counter, or c) accessing pharmacy records?

Clinical trials often make use of *surrogate response variables* that are less costly or less time-consuming to measure than the actual variable of interest. One of the earliest examples of the use of a surrogate variable was when coal miners would take a canary with them into the mine to detect a lack of oxygen well before the miners themselves fell unconscious. Today, with improved technology, they would be able to measure the concentration of oxygen directly.

The presence of HIV often serves as a surrogate for the presence of AIDS. But is HIV an appropriate surrogate? Many individuals have tested positive for HIV who do not go on to develop AIDS.<sup>7</sup> How shall we measure the progress of arteriosclerosis? By cholesterol levels? Angiography? Electrocardiogram? Or by cardiovascular mortality?

**Exercise 5.10.** Formulate your hypothesis and all of the associated alternatives for your hypothetical experiment. Decide on the variables you will

<sup>7</sup> A characteristic of most surrogates is that they are not one-to-one with the gold standard.

measure. List possible experimental findings, along with the conclusions you would draw and the actions you would take for each possible outcome. (A spreadsheet is helpful for this last.)

### 5.2.8. Random Representative Samples

Is randomization really necessary? Would it matter if you simply used the first few animals you grabbed out of the cage as controls? Or if you did all your control experiments in the morning and your innovative procedures in the afternoon? Or let one of your assistants perform the standard procedure while you performed and perfected the new technique?

A sample consisting of the first few animals to be removed from a cage will not be random because, depending on how we grab, we are more likely to select more active or more passive animals. Activity tends to be associated with higher levels of corticosteroids, and corticosteroids are associated with virtually every body function.

We've already discussed in Section 5.1.1 why a simple experiment can go astray when we *confound* competing sources of variation such as the time of day and the observer with the phenomenon that is our primary interest. As we saw in the preceding section, we can block our experiment and do the control and the innovative procedure both in the afternoon and in the morning, but we should not do one at one time and one at the other. Recommended in the present example would be to establish four different blocks (you observe in the morning, you observe in the afternoon, your assistant observes in the morning, your assistant observes in the afternoon) and to replicate the experiment separately in each block.

Samples also are taken whenever records are audited. Periodically, federal and state governments review the monetary claims made by physicians and health maintenance organizations (HMOs) for accuracy. Examining each and every claim would be prohibitively expensive, so governments limit their audits to a sample of claims. Any systematic method of sampling, examining every 10th claim say, would fail to achieve the desired objective. The HMO would soon learn to maintain its files in an equally systematic manner, making sure that every 10th record was error- and fraud free. The only way to ensure honesty by all parties submitting claims is to let a sequence of random numbers determine which claims will be examined.

The same reasoning applies when we perform a survey. Let us suppose we've decided to subdivide (block) the population whose properties we are investigating into strata—males, females, city dwellers, farmers—and to draw separate samples from each stratum. Ideally, we would assign a random number to each member of the stratum and let a computer's

random number generator determine which members are to be included in the sample.

By the way, if we don't *block* our population, we run the risk of obtaining a sample in which members of an important subgroup are absent or underrepresented. Recall from Section 2.2.3, that a single jury (or sample) may not be representative of the population as a whole. We can forestall this happening by deliberately drawing samples from each important subgroup.<sup>8</sup>

**Exercise 5.11.** Suppose you were to conduct a long-term health survey of our armed services personnel. What subgroups would you want to consider? Why?

**Exercise 5.12.** Show that the size of each subsample need not be proportional to its size in the population at large. For example, suppose your objective was to estimate the median annual household income for a specific geographic area and you were to take separate samples of households whose heads were male and female, respectively. Would it make sense to take samples of the same size from each group?

### 5.2.9. Treatment Allocation

If the members of a sample taken from a stratum are to be exposed to differing test conditions or treatments, then we must make sure that treatment allocation is random and that the allocation is concealed from both the investigator and the experimental subjects insofar as this is possible.

Treatment allocation cannot be left up to the investigator because of the obvious bias that would result. Having a third party label the treatments (or the treated patients) with seemingly meaningless labels such as A or B won't work either. The investigator will soon start drawing conclusions—not necessarily the correct ones—about which treatment the As received. In clinical trials, sooner or later the code will need to be broken for a patient who is exhibiting severe symptoms that require immediate treatment. Breaking the code for one patient when the A/B method of treatment allocation is used will mean the code has been broken for all patients.

Similar objections can be made to any system of treatment allocation in which subjects are assigned on a systematic basis to one treatment regimen or the other. For example, injecting the first subject with the experimental

---

<sup>8</sup> There has been much argument among legal scholars as to whether such an approach would be an appropriate or constitutional way to select juries.

vaccine, the next with saline, and so forth. The only safe system is one in which the assignment is made on the basis of random numbers.

### 5.2.10. Choosing a Random Sample

My clients often provide me with a spreadsheet containing a list of claims to be audited. Using Excel, I'll insert a new column and type =RAND() in the top cell. I'll copy this cell down the column and then SORT the entire worksheet on the basis of this column. (You'll find the SORT command in Excel's DATA menu.) The final step is to use the top 10 entries or the top 100 or whatever sample size I've specified for my audit.

**Exercise 5.13.** Describe the method of sampling you will use in your hypothetical experiment. If you already have the data, select the sample.

**Exercise 5.14.** Once again, you be the judge. The California Trial Jury Selection and Management Act<sup>9</sup> states that

It is the policy of the State of California that all persons selected for jury service shall be selected *at random* from the population of the area served by the court; that all qualified persons have an equal opportunity, in accordance with this chapter, to be considered for jury service in the state and an obligation to serve as jurors when summoned for that purpose; and that it is the responsibility of jury commissioners to manage all jury systems in an efficient, equitable, and cost-effective manner in accordance with this chapter.

In each of the following cases, decide whether this act has been complied with.

1. A trial judge routinely excuses physicians from jury duty because of their importance to the community.
2. Jury panels are selected from lists of drivers compiled by the department of motor vehicles.<sup>10</sup>
3. A trial judge routinely excuses jurors not possessing sufficient knowledge of English.<sup>11</sup>

<sup>9</sup> Title 3, C.C.P. Section 191.

<sup>10</sup> *U.S. v. Bailey*, 862 F. Supp. 277 (D. Colo. 1994) *aff'd in part, rev'd in part*, 76 F.3d 320, *cert. denied* 116 S. Ct. 1889.

<sup>11</sup> *People v. Lesara*, 206 Cal. App. 3d 1305, 254 Cal. Rptr. 417 (1988).

4. A trial judge routinely excuses the “less educated” (12 or fewer years of formal education) or “blue-collar workers.”<sup>12</sup>
5. A trial judge routinely excuses anyone who requests to be excused.
6. Jury selection is usually a two- or three-stage process. At the first stage, a panel is selected at random from the population. At the second stage, jurors are selected from the panel and assigned to a courtroom. In *People v. Viscotti*,<sup>13</sup> the issue was whether the trial court erred in taking the first 12 jurors from the panel rather than selecting 12 at random. How would you rule?
7. A jury of 12 black males was empaneled in an area where blacks and whites were present in equal numbers.

### 5.2.11. Ensuring that Your Observations Are Independent

Independence of the observations is essential to most statistical procedures. When observations are related as in the analysis of multifactor designs described in Chapter 6, it is essential that the residuals be independent. Any kind of dependence, even if only partial, can make the analysis suspect.

Too often, surveys take advantage of the cost savings that result from naturally occurring groups such as work sites, schools, clinics, neighborhoods, even entire towns or states. Not surprisingly, the observations within such a group are correlated. Any group or cluster of individuals who live, work, study, or pray together may fail to be representative for any or all of the following reasons:

- Shared exposure to the same physical or social environment
- Self-selection in belonging to the group
- Sharing of behaviors, ideas, or diseases among members of the group

Two events A and B are independent only if knowledge of B is NEVER of value in predicting A. In statistics, two events or two observations are said to be independent if knowledge of the outcome of the one tells you nothing about the likelihood of the other. My knowledge of who won the first race will not help me predict the winner of the second. On the other hand, knowledge of the past performance of the horses in the second race would be quite helpful.

The UCLA statistics professor who serves as consultant to the California state lottery assures me that the winning numbers on successive days are completely independent of one another. Despite the obvious, the second

<sup>12</sup> *People v. Estrada*, 93 Cal. App. 3d 76, 155 Cal. Rptr. 731 (1979).

<sup>13</sup> 2 Cal. 4th 1, 5 Cal. Rptr.2d 495 (1992).

most common pick in the California lottery are the numbers that won the previous day!

Pick two voters at random, and the knowledge of one person's vote won't help me forecast the others. But if you tell me that the second person is the spouse of the first, then there is at least a partial dependence. (The two spouses may vote differently on specific occasions, but if one generalizes to all spouses on all occasions, the dependence is obvious.) The effect of such correlation must be accounted for by the use of the appropriate statistical procedures.<sup>14</sup>

The price of Coca Cola stock tomorrow does depend upon the closing price today. But the change in price between today and tomorrow's closing may well be independent of the change in price between yesterday's closing and today's. When monitoring an assembly line or a measuring instrument, it is the changes from hour to hour and day to day that concern us. Change is expected and normal. It is the trends in these changes that concern us as these may indicate an underlying mutual dependence on some other hidden factors.

**Exercise 5.15.** Review Exercise 2.26.

### 5.3. HOW LARGE A SAMPLE?

Once we have mastered the technical details associated with making our observations, we are ready to launch our experiment or survey, but for one unanswered question: How large a sample should we take?

The effect of increasing sample size is best illustrated in the following series of photographs copied from <http://www.oztam.com.au/faq/#erwin>. The picture (below) is comprised of several hundred thousand tiny dots (the population).



---

<sup>14</sup> See, for example, Feng et al. [2001].

Now suppose we were to take successive representative samples from this population consisting of 250, 1000, and 2000 dots, respectively. They are “area probability” samples of the original picture, because the dots are distributed in proportion to their distribution in the picture. If we think of homes instead of dots, this is the sampling method used for most door-to-door surveys.



Having trouble recognizing the photo? Move back 30 inches or so from the page. When your eye stops trying to read the dots, even the smallest sample provides a recognizable picture. You would have trouble picking this woman out of a group based on the 250-dot sample. But at 1000 dots, if you squint to read the pattern of light and dark, you might recognize her. At 2000 dots, you see her more clearly—but the real improvement is between 250 and 1000—an important point. In sampling, the ability to see greater detail is a “squared function”—it takes four times as large a sample to see twice the detail. This is the strength and weakness of sample-based research. You can get the general picture cheap, but precision costs a bundle.

In our hypothetical experiment, we have a choice either of using a sample of fixed size or of a sequential sampling method in which we proceed in stages, deciding at each stage whether to terminate the experiment and make a decision. For the balance of this chapter, we shall focus on methods for determining a fixed sample size, merely indicating some of the possibilities associated with sequential sampling.

### 5.3.1. Samples of Fixed Size

Not surprisingly, many of the factors that go into determining optimal sample size are identical with those needed to determine the power of a test (Section 4.3.4):

1. The true value of the parameter being tested. The greater the gap between our primary hypothesis and the true value, the smaller the sample needed to detect the gap.



2. The variation of the observations. The more variable the observations, the more observations we will need to detect an effect of fixed size.
3. The significance level and the power. If we fix the power against a specific alternative, then working with a higher significance level (10% rather than 5%) will require fewer observations.
4. The relative costs of the observations and of the losses associated with making Type I and Type II errors. If our measurements are expensive, then to keep the overall cost of sampling under control, we may have to accept the possibility of making Type I and Type II errors more frequently.
5. The method used for testing. Obviously, we want to use the most powerful possible method to reduce the number of observations.

The sample size that we determine by consideration of these factors is the sample size we need to end the study with. We may need to take a much larger sample to begin with in order to account for drop-outs and withdrawals, animals that escape from their cages, get mislabeled or misclassified, and so forth. Retention is a particular problem in long-term studies. In a follow-up survey conducted five years after the original, one may be able to locate as few as 20% of the original participants.

We have a choice of methods for determining the appropriate sample size. If we know how the observations are distributed, we should always take advantage of our knowledge. If we don't know the distribution exactly, but the sample is large enough that we feel confident that the statistic we are interested in has almost a normal distribution, then we should take advantage of this fact. As a method of last resort, we can run a simulation or bootstrap.

In what follows, we consider each of these methods in turn.

**Known Distribution.** One instance in which the distribution of the observations is always known is when there are only two possible outcomes—success or failure, a vote for our candidate or a vote against. Suppose we desire to cut down a stand of 500-year-old redwoods in order to build and sell an expensive line of patio furniture. Unfortunately, the stand is located on state property, but we know a politician who we feel could be persuaded to help facilitate our purchase of the land. In fact, his agent has provided us with a rate card.

The politician is up for reelection and believes that a series of TV and radio advertisements purchased with our money could guarantee him a victory. He claims that those advertisements would guarantee him 55% of the vote. Our own advisors say he'd be lucky to get 40% of the vote without our ads and the best the TV exposure could do is give him another 5%.

We decide to take a poll. If it looks like only 40% of the voters favor the candidate, we won't give him a dime. If 46% or more of the voters already favor him, we'll pay to saturate the airwaves with his promises. We decide we can risk making a Type I error 5% of the time and a Type II error at most 10% of the time. That is, if  $p = 0.40$ , then the probability of rejecting the hypothesis that  $p = 0.40$  should be no greater than 5%. And if  $p = 0.46$ , then the probability of rejecting the hypothesis that  $p = 0.40$  should be at least 90%.

- To calculate the 95% percentile of the binomial distribution with 10 trials and  $p = 0.4$ , enter = BINOMDIST(k,10,0.4,1) in a vacant cell and experiment with various values for k. What would be a good starting value?
- 7 is the answer. Our rejection region will include not only 7 but all more extreme values, 8, 9, and 10. To calculate the probability of observing 7 or more successes for a binomial distribution with 10 trials and  $p = 0.4$ , enter = 1 - BINOMDIST(6,10,0.4,1). 0.05476; close enough to 5% given the small sample size.

Now let's see what the Type II error would be:

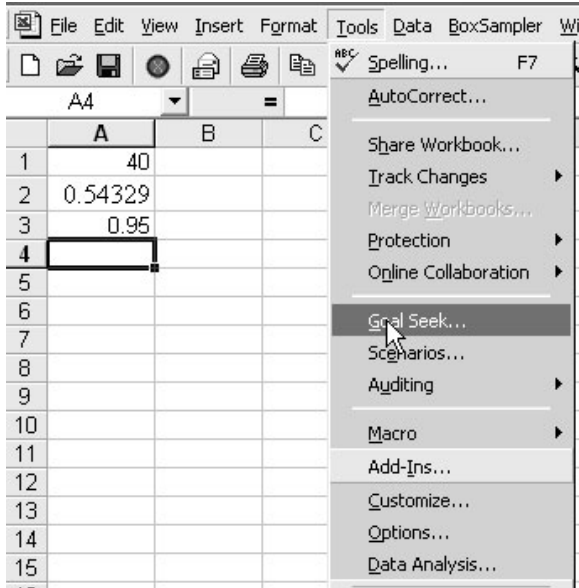
- To calculate the probability of observing 6 or fewer successes for a binomial distribution with 10 trials and  $p = 0.46$ , enter BINOMDIST(6,10,0.46,1) or 0.8859388.

Much too large; let's try a larger sample size to begin with, say 100. Repeating the previous calculations, we find by trial and error that . . . wait one minute . . . that could be as many as 50 or more calculations if I make a dumb series of guesses. Is there a better way? Absolutely.

You will first need to install the Solver add-in. Although Solver is supplied with Excel, it often needs to be installed separately. Pull down your Tools menu and see whether "Goal Seeking" is one of the options. If not, you'll need to complete the following steps.

1. Click "start" and then select either "Settings" or "Control Panel" depending on your version of Windows. In either case, from the Control Panel select "Add or Remove Programs."
2. Select "Microsoft Office Professional" from the resulting menu and click on "Change."
3. Select "Add or Remove Features."
4. Select "Microsoft Excel for Windows," "Add-ins," and "Solver."

Once Solver is installed, select "Goal Seeking" from the tools menu as shown in Fig. 5.1.



**FIGURE 5.1** Preparing to let Excel do the work.

Next, complete the Goal Seek menu as shown in Fig. 5.2. Press OK. Excel reports that it cannot find a solution but suggests 47.47 as a possibility. To be on the conservative side, let us use a sample size of 48.

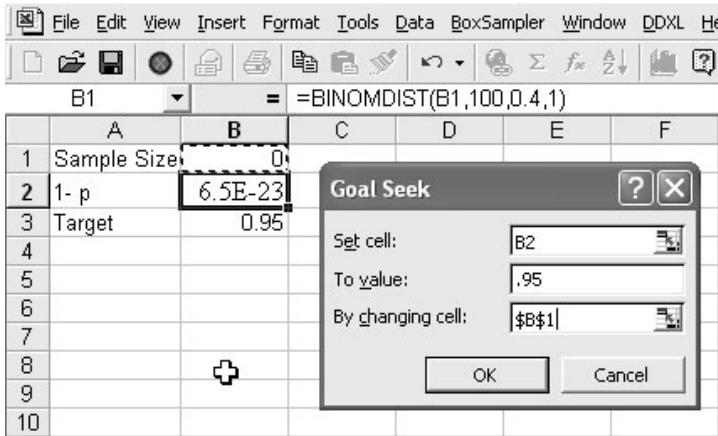
**Warning:** Excel's Goal Seek procedure is not always successful. For, if one starts with a guess of 40 in cell B1 instead of 0, Goal Seek fails to find the answer.

Our next step is to determine the power of the test associated with the new sample size of 100.  $1 - \text{BINOMDIST}(48, 100, 0.46, 1) = 0.6191224$ . Not large enough.

Let's try a sample size of 400. Trial and error and a occasional help from Goal Seek yields a cutoff value of 175 successes, with a Type I error of 0.057 and a Type II error of 0.19.

We're getting close. Let's try a sample size of 800. Trial and error yields a cutoff value of 342 successes, with a Type I error of 0.052 and a Type II error of 0.04, which is less than 10%.

A poll of 800 people will give me the results I need. But why pay for that large a sample, when fewer observations will still result in the desired Type I and Type II errors?



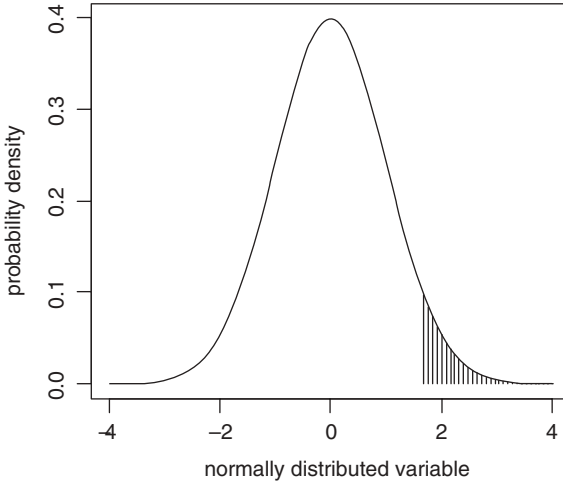
**FIGURE 5.2** Completing the Goal Seek menu.

**Exercise 5.16.** Find, to the nearest 20 observations, the smallest sample size needed to yield a Type I error of 5% when  $p = 0.40$  and a Type II error of 10% when  $p = 0.46$ .

**Exercise 5.17.** A friend of ours has a “lucky” coin that seems to come up heads every time he flips it. We examine the coin and verify that one side is marked tails. How many times should we flip the coin to test the hypothesis that it is fair so that a) the probability of making a Type I error is no greater than 10% and b) we have a probability of 80% of detecting a weighted coin that will come up heads 70 times out of one hundred on the average?

**Almost Normal Data.** As noted in Chapter 3, the mean of a sample will often have an almost normal distribution similar to that depicted in Fig. 1.23 even when the individual observations come from some quite different distribution. See, for example, Exercise 3.14.

In the previous section, we derived the ideal sample size more or less by trial and error. We could proceed in much the same way with normally distributed data, but there is a much better way. Recall that in Exercise 3.17 we showed that the variance of the mean of  $n$  observations each with variance  $\sigma^2$  was  $\sigma^2/n$ . If we know the cutoff value for testing the mean of a normal distribution with variance 1, we can find the cutoff value and subsequently the power of a test for the mean of a normal distribution with any variance whatever.



**FIGURE 5.3** A cutoff value of 1.64 excludes 5% of  $N(0,1)$  observations.

Let's see how. Typing

$$\bullet = \text{NORMSINV}(0.95)$$

we find that the 95th percentage point of an  $N(0,1)$  distribution is 1.644854

We illustrate this result in Fig. 5.3.

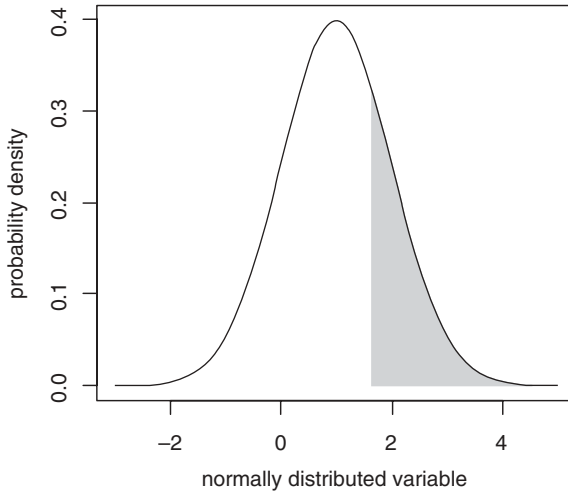
If the true mean is actually one standard deviation larger than 0, the probability of observing a value less than or equal to 1.644853 is given by  $1 - \text{NORMS}(1.644853 - 1) = 0.74$ . We illustrate this result in Fig. 5.4.

We can use Exercise 3.12 to show that if each of  $n$  independent observations is normally distributed as  $N(0,1)$  then their mean is distributed as  $N(0,1/n)$ . Detecting a difference of 1 now becomes a task of detecting a difference of  $\sqrt{n}$  standard deviation units.

$$\bullet \text{ NORMSINV}(0.10) = -1.2816$$

Thus we require a sample size  $n$  such that  $1.644 - \sqrt{n} = -1.281552$ , or  $n = 9$ .

More often, we need to test against an alternative of fixed size. For, the population mean is really equal to 10 units. Thus, to determine the required sample size for testing, we would also need to know the population variance or at least to have some estimate of what it is. If we have taken a preliminary sample, then the variance of this sample  $s^2$  could serve as an estimate of the unknown population variance  $\sigma^2$ .



**FIGURE 5.4** A cutoff value of 1.64 detects 36% of  $N(1,1)$  observations.

Let us suppose the sample variance is 25. The sample standard deviation is 5, and we are testing for an estimated difference of 2 standard deviations. At a significance level of 5%, we require a sample size  $n$  such that  $1.644854 - \sqrt{n} \cdot 10/5 = -1.281552$ , or no more than 3 observations.

To generalize the preceding results, suppose that  $C_\alpha$  is the cutoff value for a test of significance at level  $\alpha$ , and we want to have power  $\beta$  to detect a difference of size  $\delta$ .  $C_\beta$  is the value of a  $N(0,1)$  distributed variable  $Z$  for which  $P\{Z > C_\beta\} = \beta$ , and  $\sigma$  is the standard deviation of the variable we are measuring. Then  $\sqrt{n} = (C_\alpha - C_\beta) \sigma / \delta$ .

**Exercise 5.18.** How big a sample would you need to test the hypothesis that the average sixth-grader is 150 mm in height at a 5% significance level so that the probability of detecting a true mean height of 160 mm is 90%?

**Exercise 5.19.** When his results were not statistically significant at the 10% level, an experimenter reported that a “new experimental treatment was ineffective.” What other possibility is there?

**Bootstrap.** If you have reason to believe that the distribution of the sample statistic is not normal, for example, if you are testing hypotheses regarding variances or ratios, the best approach to both power and sample size determination is to bootstrap from the empirical distribution under both the primary and the alternative hypothesis.

Recently, one of us was helping a medical device company design a comparison trial of their equipment with that of several other companies. They had plenty of information on their own equipment but could only guess at the performance characteristics of their competitors. As they were going to have to buy, and then destroy, their competitors' equipment to perform the tests, they wanted to keep the sample size as small as possible.

Stress test scores took values from 0 to 5, with 5 being the best. The idea was to take a sample of  $k$  units from each lot and reject if the mean score was too small. To determine the appropriate cutoff value for each prospective sample size, we took a series of simulated samples from the empirical distribution for our client's equipment. The individual frequencies for this distribution were  $f_0, f_1, f_2, f_3, f_4,$  and  $f_5$ . We let the computer choose a random number from 0 to 1. If this number was less than  $f_0$ , we set the simulated test score to 0. If the random number was greater than  $f_0$  but less than  $f_0 + f_1$ , we set it to 1, and so forth. We did this  $k$  times, recorded the mean, and then repeated the entire process. For  $k = 4$ , 95% of the time this mean was greater than 3. So 3 was our cutoff point for  $k = 4$ .

Next, we guesstimated an empirical distribution for the competitor's product. We repeated the entire simulation using this guesstimated distribution. (Sometimes, you just have to reply on your best judgment.) For  $k = 4$ , 40% of the time the mean of our simulated samples of the competitors' products was less than 3. Not good enough. We wanted a test our product could pass and their products wouldn't.

By trial and error, we finally came up with a sample size of 6 and a test our product could pass 95% of the time and the competitors' products would fail at least 70% of the time. We were happy.

How many simulated samples  $n$  did we have to take each time? The proportion of values that fall into the rejection region is a binomial random variable with  $n$  trials and a probability  $\beta$  of success in each trial, where  $\beta$  is the desired power of the test.

We used  $n = 100$  until we were ready to fine-tune the sample size, when we switched to  $n = 400$ .

**Exercise 5.20.** In preliminary trials of a new medical device, test results of 7.0 were observed in 11 out of 12 cases and 3.3 in 1 out of 12 cases. Industry guidelines specified that any population with a mean test result greater than 5 would be acceptable. A worst-case or boundary-value scenario would include one in which the test result was 7.0, 3/7th of the time, 3.3, 3/7th of the time, and 4.1, 1/7th of the time.

The statistical procedure with significance level 6% requires us to reject if the sample mean of subsequent test results is less than 6. What sample size is required to obtain a power of at least 80% for the worst-case scenario?

### 5.3.2. Sequential Sampling

The computational details of sequential sampling procedures are beyond the scope of the present text. Still, realizing that many readers will go on to design their own experiments and surveys, we devote the balance of this chapter to outlining some of the possibilities.

**Stein's Two-Stage Sampling Procedure.** Charles Stein's two-stage sampling procedure makes formal recognition of the need for some estimate of variation before we can decide on an optimal sample size. The procedure assumes that the test statistic will have an almost normal distribution. We begin by taking a relatively small sample and use it and the procedures of the preceding sections to estimate the optimal sample size.

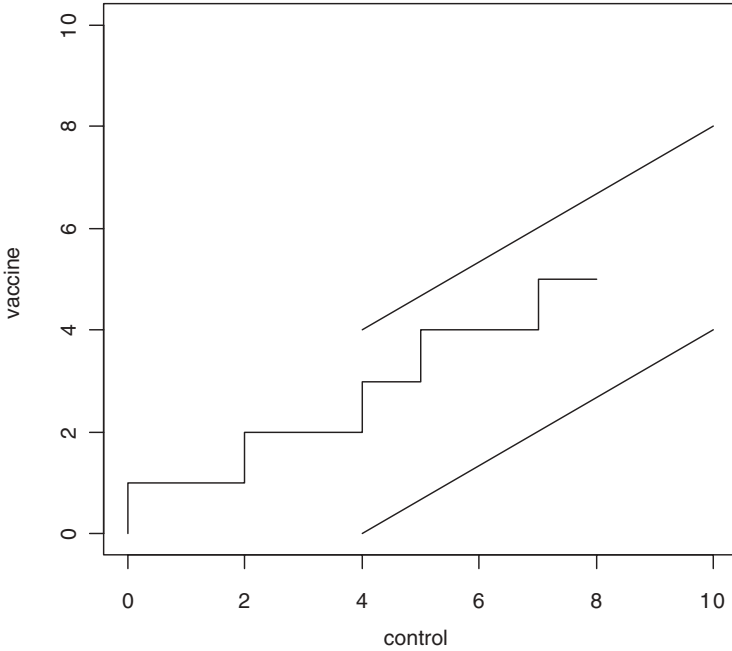
If the estimated optimal sample size is less than or equal to the size of the sample we've already taken, we stop; otherwise we will take the suggested number of observations plus one.

**Exercise 5.21.** Apply Stein's two-stage sampling procedure to the data of Exercise 5.17. How many additional observations would we need to detect an improvement in scores of 4 units 95% of the time?

**Wald Sequential Sampling.** When our experiments are destructive in nature (as in testing condoms) or may have an adverse effect upon the experimental subject (as in clinical trials), we would prefer not to delay our decisions until some fixed sample size has been reached.

Figure 5.5 depicts a sequential trial of a new vaccine after eight patients who had received either the vaccine or an innocuous saline solution had come down with the disease. Each time a control patient came down with the disease, the jagged line was extended to the right. Each time a patient who had received the experimental vaccine came down with the disease, the jagged line was extended upward one notch. The experiment will continue until either the jagged line crosses the lower boundary—in which case we will stop the experiment, reject the null hypothesis, and immediately put the vaccine into production, or the jagged line crosses the upper boundary—in which case we will stop the experiment, accept the null hypothesis, and abandon further work with this vaccine. What Abraham Wald [1945] showed in his pioneering research was that on the average





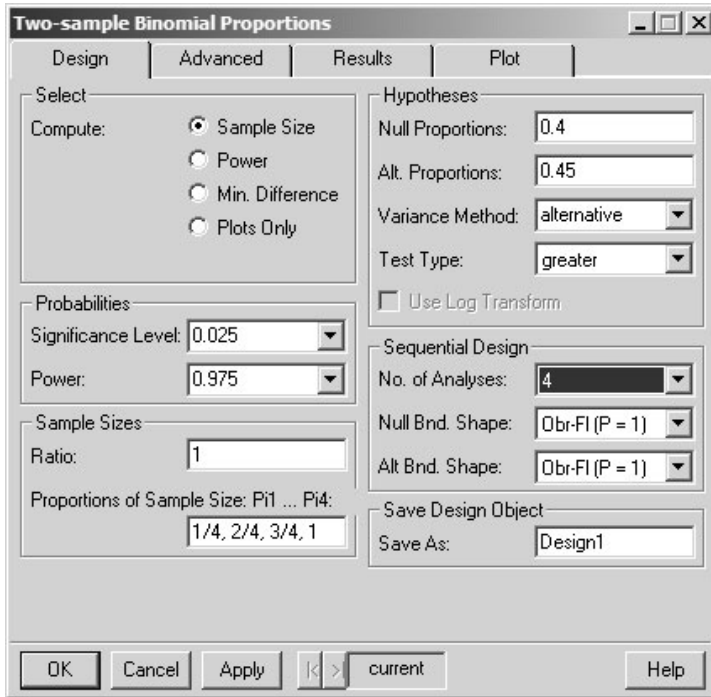
**FIGURE 5.5** Sequential trial in progress.

the resulting sequential experiment would require many fewer observations whether or not the vaccine was effective than would a comparable experiment of fixed sample size.

**Exercise 5.22.** Suppose we were to take a series of observations and, after each one, reject if the test statistic is greater than the 95th percentile of its distribution under the null hypothesis. Show that the Type I error would exceed 5% even if we only took two observations.

As Exercise 5.22 illustrates, simply performing a standard statistical test after each new observation as if the sample size were fixed will lead to inflated values of Type I error. The boundaries depicted in Fig. 5.3 were obtained by using formulas specific to sequential design. Not surprisingly, these formulas require us to know each and every one of the factors required to determine the number of samples when an experiment is of fixed size.

More recent developments include “group-sequential designs,” which involve testing not after every observation, as in a fully sequential design, but rather after groups of observations, e.g., after every 6 months in a



**FIGURE 5.6** Group-sequential design menu in S+SeqTrial.

clinical trial. The design and analysis of such experiments is best done with specialized software such as S+SeqTrial, from <http://www.insightful.com>. For example, Fig. 5.6 is the main menu for designing a trial to compare binomial proportions in a treatment and control group, with the null hypothesis being  $p = 0.4$  in both groups, and the alternative hypothesis that  $p = 0.45$  in the treatment group, using an “O’Brien–Fleming” design, with a total of four analyses (three “interim analyses” and a final analysis).

The resultant output (see sidebar) begins with the call to the “seqDesign” function that you would use if working from the command line rather than using the menu interface. The null hypothesis is that Theta (the difference in proportions, e.g., survival probability, between the two groups) is 0.0, and the alternative hypothesis is that Theta is at least 0.05. The last section indicates the stopping rule, which is also shown in the next plot. After 1565 observations (split roughly equally between the two groups) we should analyze the interim results. At the first analysis, if the treatment group has a survival probability that is 10% greater than the control group, we stop early and reject the null hypothesis; if the treat-

ment group is doing 5% worse, we also stop early, and accept the null hypothesis (at this point it appears that our treatment is actually killing people; there is little point in continuing the trial). Any ambiguous result, in the middle, causes us to collect more data. At the second analysis time the decision boundaries are narrower, with lower and upper boundaries 0% and 5%; stop and declare success if the treatment group is doing 5% better, stop and give up if the treatment group is doing at all worse. The decision boundaries at the third analysis time are even narrower, and at the final time (6260 total observations) they coincide; at this point we make a decision one way or the other. For comparison, the sample size and critical value for a fixed-sample trial is shown; this requires somewhat less than 6000 subjects.

```

*** Two-sample Binomial Proportions Trial ***

Call:
seqDesign(prob.model = "proportions", arms = 2,
  null.hypothesis = 0.4, alt.hypothesis = 0.45, ratio
  = c(1., 1.), nbr.analyses = 4, test.type = "greater",
  power = 0.975, alpha = 0.025, beta = 0.975, epsilon =
  c(0., 1.), display.scale = seqScale(scaleType = "X"))

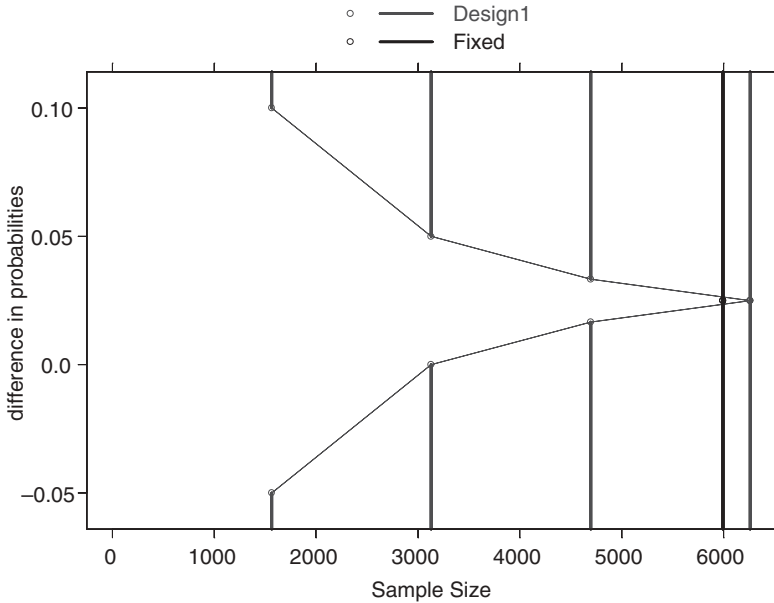
PROBABILITY MODEL and HYPOTHESES:

Two-arm study of binary response variable
Theta is difference in probabilities (Treatment -
  Comparison)
One-sided hypothesis test of a greater alternative:
Null hypothesis : Theta <= 0      (size = 0.025)
Alternative hypothesis : Theta >= 0.05      (power =
  0.975)
[Emerson & Fleming (1989) symmetric test]

STOPPING BOUNDARIES: Sample Mean scale
              a          d
Time 1 (N= 1565.05) -0.0500 0.1000
Time 2 (N= 3130.09)  0.0000 0.0500
Time 3 (N= 4695.14)  0.0167 0.0333
Time 4 (N= 6260.18)  0.0250 0.0250

```

Figure 5.7 depicts the boundaries of a group-sequential trial. At each of four analysis times, at each time a difference in proportions below the lower boundary or above the upper boundary causes the trial to stop; anything in the middle causes it to continue. For comparison, a fixed trial (in

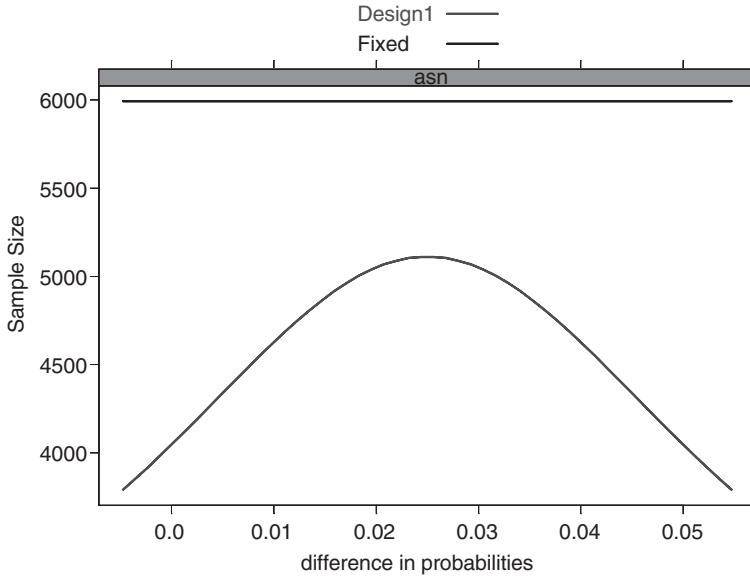


**FIGURE 5.7** Group-sequential decision boundaries.

which one only analyzes the data at the completion of the study) is shown; this would require just under 6000 subjects for the same Type I error and power.

The major benefit of sequential designs is that we may stop early if results clearly favor one or the other hypothesis. For example, if the treatment really is worse than the control, we are likely to hit one of the lower boundaries early. If the treatment is much better than the control, we are likely to hit an upper boundary early. Even if the true difference is right in the middle between our two hypotheses, say that the treatment is 2.5% better (when the alternative hypothesis is that it is 5% better), we may stop early on occasion. Figure 5.8 shows the average sample size as a function of  $\Theta$ , the true difference in means. When  $\Theta$  is less than 0% or greater than 5%, we need about 4000 observations on average before stopping. Even when the true difference is right in the middle, we stop after about 5000 observations, on average. In contrast, the fixed-sample design requires nearly 6000 observations for the same Type I error and power.

**Adaptive Sampling.** The adaptive method of sequential sampling is used primarily in clinical trials where the treatment or the condition being treated presents substantial risks to the experimental subjects. Suppose, for



**FIGURE 5.8** Average sample sizes, for group-sequential design.

example, 100 patients have been treated, 50 with the old drug and 50 with the new. If, on review of the results, it appears that the new experimental treatment offers substantial benefits over the old, we might change the proportions given each treatment, so that in the next group of 100 patients, just 25 randomly chosen patients receive the old drug and 75 receive the new.

## 5.4. META-ANALYSIS

Such is the uncertain nature of funding for scientific investigation that experimenters often lack the means necessary to pursue a promising line of research. A review of the literature in your chosen field is certain to turn up several studies in which the results are inconclusive. An experiment or survey has ended with results that are “almost” significant, say with  $p = 0.075$  but not  $p = 0.049$ . The question arises whether one could combine the results of several such studies, thereby obtaining, in effect, a larger sample size and a greater likelihood of reaching a definitive conclusion. The answer is yes, through a technique called *meta-analysis*.

Unfortunately, a complete description of this method is beyond the scope of this text. There are some restrictions on meta-analysis, for example, that the experiments whose  $p$  values are to be combined should

be comparable in nature. Formulas and a set of Excel worksheets may be downloaded from <http://www.ucalgary.ca/~steel/procrastinus/meta/Meta%20Analysis%20-%20Mark%20IX.xls>

**Exercise 5.23.** List all the respects in which you feel experiments ought to be comparable in order that their p-values should be combined in a meta-analysis.

## 5.5. SUMMARY AND REVIEW

In this chapter, you learned the principles underlying the design and conduct of experiments and surveys. You learned how to cope with variation through controlling, blocking, measuring, or randomizing with respect to all contributing factors. You learned the importance of giving a precise, explicit formulation to your objectives and hypotheses. You learned a variety of techniques to ensure that your samples will be both random and representative of the population of interest. And you learned a variety of methods for determining the appropriate sample size.

You also learned that there is much more to statistics than can be presented within the confines of a single introductory text.

**Exercise 5.24.** A highly virulent disease is known to affect one in 5000 people. A new vaccine promises to cut this rate in half. Suppose we were to do an experiment in which we vaccinated a large number of people, half with an ineffective saline solution and half with the new vaccine. How many people would we need to vaccinate to ensure that the probability was 80% of detecting a vaccine as effective as this one purported to be while the risk of making a Type I error was no more than 5%? (Hint: See Section 4.2.1.)

There was good news and bad news when one of us participated in just such a series of clinical trials recently. The good news was that almost none of the subjects—control or vaccine treated—came down with the disease. The bad news was that with so few diseased individuals the trials were inconclusive.

**Exercise 5.25.** To compare teaching methods, 20 school children were randomly assigned to one of two groups. The following are the test results:

conventional	85	79	80	70	61	85	98	80	86	75
new	90	98	73	74	84	81	98	90	82	88

Are the two teaching methods equivalent in result?

What sample size would be required to detect an improvement in scores of 5 units 90% of the time where our test is carried out at the 5% significance level?

**Exercise 5.26.** To compare teaching methods, 10 school children were first taught by conventional methods, tested, and then taught by an entirely new approach. The following are the test results:

conventional	85	79	80	70	61	85	98	80	86	75
new	90	98	73	74	84	81	98	90	82	88

Are the two teaching methods equivalent in result?

What sample size would be required to detect an improvement in scores of 5 units 90% of the time? Again, the significance level for the hypothesis test is 5%.

**Exercise 5.27.** Make a list of all the italicized terms in this chapter. Provide a definition for each one along with an example.

# Chapter 6

## Analyzing Complex Experiments

**IN THIS CHAPTER, YOU'LL LEARN HOW** to analyze a variety of different types of experimental data including changes measured in percentages, samples drawn from more than two populations, categorical data presented in the form of contingency tables, samples with unequal variances, and multiple end points.

### 6.1. CHANGES MEASURED IN PERCENTAGES

In Chapter 5, we learned how we could eliminate one component of variation by using each subject as its own control. But what if we are measuring weight gain or weight loss, where the changes, typically, are best expressed as percentages rather than absolute values? A 250-pounder might shed 20 pounds without anyone noticing; not so with a 125-pounder.

The obvious solution is to work not with the before-after differences but with the before/after ratios.

But what if the original observations are on growth processes—the size of a tumor or the size of a bacterial colony—and vary by several orders of magnitude? H. E. Renis of the Upjohn Company observed the following vaginal virus titers in mice 144 hours after inoculation with herpesvirus type II:

Saline controls	10,000,	3000,	2600,	2400,	1500
Treated with antibiotic	9000,	1700,	1100,	360,	1

In this experiment the observed values vary from 1, which may be written as  $10^0$ , to 10,000, which may be written as  $10^4$  or 10 times itself 4



times. With such wide variation, how can we possibly detect a treatment effect?

The trick employed by statisticians is to use the *logarithms* of the observations in the calculations rather than their original values. The logarithm or log of 10 is 1, the log of 10,000 written  $\log_{10}(10000)$  is 4.  $\log_{10}(0.1)$  is  $-1$ . (Yes, the trick is simply to count the number of decimal places that follow the leading digit.)

Using logarithms with growth and percentage-change data has a second advantage. In some instances, it equalizes the variances of the observations or their ratios so that they all have the identical distribution up to a shift. Recall that equal variances are necessary if we are to apply any of the methods we learned for detecting differences in the means of populations.

**Exercise 6.1.** Was the antibiotic used by H. E. Renis effective in reducing viral growth? (Hint: First convert all the observations to their logarithms using the function  $\log_{10}(\cdot)$ .)

**Exercise 6.2.** Although crop yield improved considerably this year on many of the plots treated with the new fertilizer, there were some notable exceptions. The recorded after/before ratios of yields on the various plots were as follows: 2, 4, 0.5, 1, 5.7, 7, 1.5, 2.2. Is there a statistically significant improvement?

## 6.2. COMPARING MORE THAN TWO SAMPLES

The comparison of more than two samples is an easy generalization of the method we used for comparing two samples. As in Chapter 4, we want a test statistic that takes more or less random values when there are no differences among the populations from which the samples are taken but tends to be large when there are differences. Suppose we have taken samples of sizes  $n_1, n_2, \dots, n_I$  from  $I$  populations. Consider either of the statistics

$$F_2 = \sum_{i=1}^I n_i (\bar{X}_i - \bar{X}_{..})^2$$

or

$$F_1 = \sum_{i=1}^I n_i |\bar{X}_i - \bar{X}_{..}|$$

where  $\bar{X}_i$  is the mean of the  $i$ th sample and  $\bar{X}_{..}$  is the grand mean of all the observations.

Recall from Chapter 1 that the symbol  $\Sigma$  stands for sum of, so that  $\sum_{i=1}^I n_i(\bar{X}_i - \bar{X}_{..})^2 = n_1(\bar{X}_1 - \bar{X}_{..})^2 + n_2(\bar{X}_2 - \bar{X}_{..})^2 + \dots + n_I(\bar{X}_I - \bar{X}_{..})^2$ . If the means of the  $I$  populations are approximately the same, then changing the labels on the various observations will not make any difference as to the expected value of  $F_2$  or  $F_1$ , as all the sample means will still have more or less the same magnitude. On the other hand, if the values in the first population are much larger than the values in the other populations, then our test statistic can only get smaller if we start rearranging the observations among the samples. We can show this by drawing a series of figures as we did in Section 4.3.4 when we developed a test for correlation.

Because the grand mean remains the same for all possible rearrangements of labels, we can use a simplified form of the  $F_2$  statistic,

$$F_2 = \sum_{i=1}^I n_i \bar{X}_i^2.$$

Our permutation test consists of rejecting the hypothesis of no difference among the populations when the original value of  $F_2$  (or of  $F_1$  should we decide to use it as our test statistic) is larger than all but a small fraction, say 5%, of the possible values obtained by rearranging labels.

### 6.2.1. Programming the Multisample Comparison with Excel

To minimize the work involved, the worksheet depicted in Fig. 6.1 was assembled in the following order:

1. The original data were placed in cells A3 through D8, with each sample in a separate column.
2. The sample sizes were placed in cells A9 through D9.

	A	B	C	D	E	F	G	H	I	J	K
1	k-sample comparison with unordered categories										
2	Original Data						Rearranged Data				
3	28	33	18	11			14	34	11	34	
4	23	36	21	14			33	14	18	21	
5	14	34	20	11			22	29	16	11	
6	27	29	22	16			20	23	31	28	
7		31	24					24	36		
8		34						27			
9	4.00	6.00	5.00	4.00	19.00		4.00	6.00	5.00	4.00	
10	92.00	197.00	105.00	52.00	446.00		89.00	151.00	112.00	94.00	
11					23.47						
12	2116.00	6468.17	2205.00	676.00	11465.17	F2	1980.25	3800.17	2508.80	2209.00	10498.22
13	1.89	56.16	12.37	41.89	112.32	F1	4.89	10.16	5.37	0.11	20.53

**FIGURE 6.1** Preparing to make a  $k$ -sample comparison by permutation means.

3. The sum of the observations in the first sample =SUM(A3:A8) was placed in cell A10.
4. The square of the sum of the observations in the first sample divided by the sample size =A10 \* A10/A9 was placed in cell A12.
5. The S command of the Resampling Stats add-in was used to generate the rearranged data in Cells G3 through J8 as described in Section 4.2.2.
6. Cells A10 through A11 were copied, first to cells B10 through B12 and then to cells G10 through G12. Note that Excel modifies the formula automatically.
7. The total sample size =Sum(A9:D9) was placed in cell E9.
8. Cell E9 was copied to cells E10 through E13.
9. Cell E11 was overwritten with the grand mean =E10/E9.
10. The formula =ABS(A10-A9 \* \$E\$11) was put in cell A13.
11. The contents of cell A13 were copied and pasted first into cells B13 through D13 and then into cells G13 to J13. Note that Excel does not modify row and column headings that are preceded by a dollar sign. Thus the contents of cell J13 are now =ABS(J10-J9 \* \$E\$11).
12. Cell E12 was copied and pasted first into cell E13 and then into cells K12 through K13.

The next step is to run the Resampling Stats RS command for either F2 in cell K12 or F1 in cell K13. Finish by sorting the first column on the Results worksheet to determine the  $p$  value, that is, what proportion of the rearrangements yield values of F2 greater than 11465? Or of F1 greater than 112?

**Exercise 6.3.** Use BoxSampler to generate four samples from a  $N(0,1)$  distribution. Use sample sizes of 4, 4, 3, and 5, respectively. Repeat the preceding steps using the F2 statistic to see whether this procedure will detect differences in these four samples despite their all being drawn from the same population. (If you've set up the worksheet correctly, the answer should be "no.")

**Exercise 6.4.** Modify your data by adding the value 2 to each member of the first sample. Now test for differences among the populations.

**Exercise 6.5.** We saw in Exercise 6.4 that if the expected value of the first population was much larger than the expected values of the other populations we would have a high probability of detecting the difference. Would the same be true if the mean of the second population was much higher than that of the first? Why?

**Exercise 6.6.** Modify your data by adding 1 to all the members of the first sample and subtracting 1.2 from each of the three members of the third sample. Now test for differences among the populations.

### 6.2.2. What Is the Alternative?

We saw in the preceding exercises that we can detect differences among several populations if the expected value of one population is much larger than the others or if the mean of one of the populations is a little higher and the mean of a second population is a little lower than the grand mean.

Suppose we represent the expectations of the various populations as follows:  $EX_i = \mu + \delta_i$  where  $\mu$  (pronounced mu) is the grand mean of all the populations and  $\delta_i$  represents the deviation of the expected value of the  $i$ th population from this grand mean. The sum of these deviations  $\Sigma \delta_i = \delta_1 + \delta_2 + \dots + \delta_t = 0$ . We will sometimes represent the individual observations in the form  $X_{ij} = \mu + \delta_i + z_{ij}$ , where  $z_{ij}$  is a random deviation with expected value 0 at each level of  $i$ . The permutation tests we describe in this section are applicable only if all the  $z_{ij}$  have the same distribution at each level of  $i$ .

One can show, although the mathematics is tedious, that the power of a test using the statistic  $F_2$  is an increasing function of  $\Sigma \delta_i^2$ . The power of a test using the statistic  $F_1$  is an increasing function of  $\Sigma |\delta_i|$ . The problem with these omnibus tests is that although they allow us to detect any of a large number of alternatives, they are not especially powerful for detecting any specific alternative. As we shall see in the next section, if we have some advance information that the alternative is, for example, an ordered dose response, then we can develop a much more powerful statistical test specific to that alternative.

**Exercise 6.7.** Suppose a car manufacturer receives four sets of screws, each from a different supplier. Each set is a population. The mean of the first set is 4 mm, the second set 3.8 mm, the third set 4.1 mm, and the fourth set 4.1 mm, also. What would the values of  $\mu$ ,  $\delta_1$ ,  $\delta_2$ ,  $\delta_3$ , and  $\delta_4$  be? What would be the value of  $\Sigma |\delta_i|$ ?

### 6.2.3. Testing for a Dose Response or Other Ordered Alternative

Frank, Trzos, and Good studied the increase in chromosome abnormalities and micronuclei as the dose of various compounds known to cause mutations was increased. Their object was to develop an inexpensive but sensitive biochemical test for mutagenicity that would be able to detect even

**TABLE 6.1 Micronuclei in Polychromatophilic Erythrocytes and Chromosome Alterations in the Bone Marrow of Mice Treated with CY**

Dose (mg/kg)	Number of Animals	Micronucelii per 200 cells	Breaks per 25 cells
0	4	0 0 0 0	0 1 1 2
5	5	1 1 1 4 5	0 1 2 3 5
20	4	0 0 0 4	3 5 7 7
80	5	2 3 5 11 20	6 7 8 9 9

marginal effects. The results of their experiment are reproduced in Table 6.1.

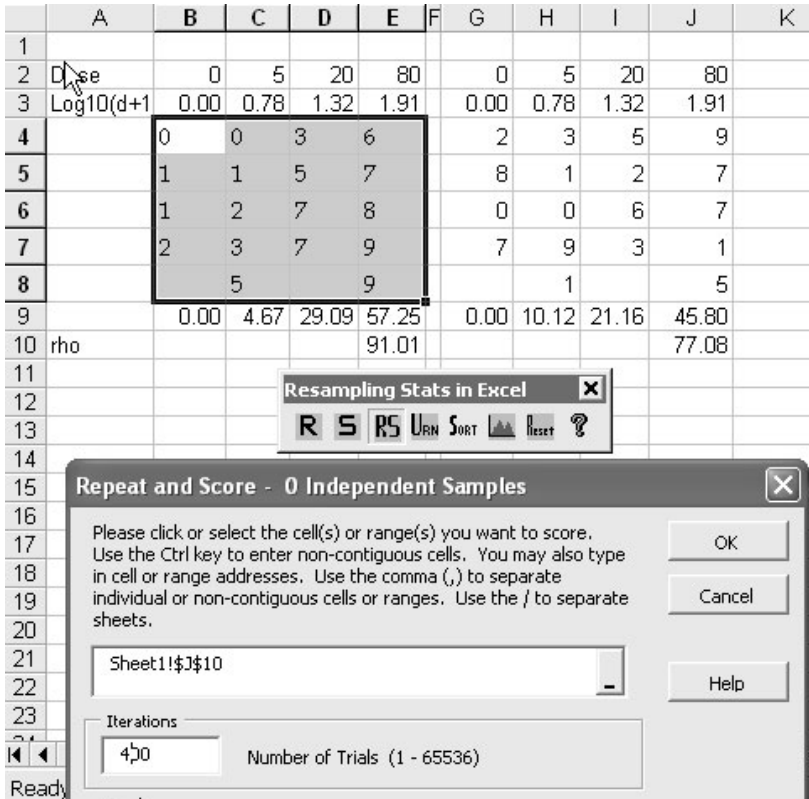
To analyze such data, Pitman proposes a test for linear correlation with three or more *ordered samples* using as test statistic  $S = \sum g[i]s_i$ , where  $s_i$  is the sum of the observations in the  $i$ th dose group, and  $g[i]$  is any monotone increasing function of  $i$ . The simplest example of such a function is  $g[i] = i$ , with test statistic  $S = \sum g[i]s_i$ . In this instance, based on the recommendation of experts in toxicology, we take  $g[\text{dose}] = \log[\text{dose} + 1]$ , as the anticipated effect is proportional to the logarithm of the dose. Our test statistic is  $S = \sum \log[\text{dose}_i + 1]s_i$ .

The original data for breaks may be written in the form

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

As  $\log [0 + 1] = 0$ , the value of the Pitman statistic for the original data is  $0 + 11 * \log[6] + 22 * \log[21] + 39 * \log[81] = 112.1$ . The only larger values are associated with the small handful of rearrangements of the form

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



**FIGURE 6.2** Preparing to compute the permutation distribution of the Pitman correlation.

As there are  $\binom{18}{4\ 5\ 4}$  771,891,120 rearrangements in all,<sup>1</sup> a statistically significant ordered dose response of  $p < 0.001$  has been detected. The micronuclei also exhibit a statistically significant dose response when we calculate the permutation distribution of  $S = \sum \log[\text{dose}_i + 1]n_i$ .

To make the calculations for this second test, we took advantage once again of the Resampling Statistics add-in as shown in Fig. 6.2. The doses were entered in row 2 and converted to log doses in row 3. The original data were entered in B4:E8. Row 9 contains the cross products. As in previous sections, the Shuffle command was used to generate a single

<sup>1</sup> See Section 2.2.1.

rearrangement and then the Repeat and Shuffle command to generate the permutation distribution of the test statistic in cell J10.

A word of caution: If we use as the weights some function of the dose other than  $g[\text{dose}] = \log[\text{dose} + 1]$ , we might observe a different result. Our choice of a test statistic must always make practical as well as statistical sense.

**Exercise 6.8.** Using the data for micronuclei, see if you can detect a significant dose effect. (Hint: I usually use  $N = 400$  repetitions to begin with. Try with both  $N = 400$  and  $N = 1600$ .)

### **k-SAMPLE TEST FOR ORDERED SAMPLES**

Hypothesis H: All distributions and all population means are the same.

Alternative K: The population means are ordered.

Assumptions under the null hypothesis:

- 1) Labels on the observations can be exchanged if the hypothesis is true.
- 2) All the observations in the  $i$ th sample come from the same distribution  $G_i$ ,

where  $G[x] = \Pr\{X \leq x\} = F[x - \delta]$ .

Test statistic:

$S = \sum g[i]x_i$  where  $x_i$  is the sum of the observations in the  $i$ th sample.

**Exercise 6.9.** Aflatoxin is a common and undesirable contaminant of peanut butter. Are there significant differences in aflatoxin levels among the following brands?

Snoopy	0.5	7.3	1.1	2.7	5.5	4.3
Quick	2.5	1.8	3.6	5.2	1.2	0.7
Mrs. Good's	3.3	1.5	0.4	4.8	2.2	1.0

(Hint: What is the null hypothesis? What alternative or alternatives are of interest?)

**Exercise 6.10.** Does the amount of potash in the soil affect the strength of fibers made of cotton grown in that soil? Consider the data in the following table:

	Potash Level (lb/acre)				
	144	108	72	54	36
Breaking	7.46	7.17	7.76	8.14	7.63
Strength	7.68	7.57	7.73	8.15	8.00
	7.21	7.80	7.74	7.87	7.93

### 6.3. EQUALIZING VARIANCES

Suppose that to cut costs on our weight loss experiment, we have each participant weigh him or herself. Some individuals will make very *precise* measurements, perhaps repeating the procedure three or four times to make sure they've performed the measurement correctly. Others, will say "close enough," and get the task done as quickly as possible. The problem with our present statistical methods is they treat each observation as if it were equally important. Ideally, we should give the least consideration to the most variable measurements and the greatest consideration to those that are least variable. The problem is that we seldom have any precise notion of what these variances are.

One possible solution is to put all results on a pass-fail or success-failure basis. This way, if the goal is to lose at least 7% of body weight, losses of 5% and 20% would offset each other, rather than a single 20% loss being treated as if it were equivalent to four losses of 5%. These pass-fail results would follow a binomial distribution, and the appropriate method for testing binomial hypotheses could be applied.

The downside is the loss of information, but if our measurements are not equally precise, perhaps it is noise rather than useful information that we are discarding.

Here is a second example: An experimenter administered three drugs in random order to each of five recipients. He recorded their responses and now wishes to decide whether there are significant differences among treatments. The problem is that the five subjects have quite different baselines. A partial solution would be to subtract an individual's baseline value from all subsequent observations made on that individual. But who is to say that an individual with a high baseline value will respond in the same way as an individual with a low baseline reading?

An alternate solution would be to treat each individual's readings as a *block* (see Section 5.2.3) and then combine the results. But then we run the risk that the results from an individual with unusually large responses might mask the responses of the others. Or suppose the measurements actually had been made by five different experimenters using five different



**TABLE 6.2a Original Observations**

	A	B	C	D	E
Control	89.7	75	105	94	80
Treatment 1	86.2	74	95	98	79
Treatment 2	76.5	68	94	93	84

**TABLE 6.2b Ranks**

	A	B	C	D	E
Control	1	1	1	2	2
Treatment 1	2	2	2	1	3
Treatment 2	3	3	3	3	1

measuring devices in five different laboratories. Would it really be appropriate to combine them?

No, for sets of observations measured on different scales are not *exchangeable*. By converting the data to *ranks*, separately for each case, we are able to put all the observations on a common scale, and then combine the results.

When we replace observations by their ranks, we generally give the smallest value rank 1, the next smallest rank 2, and so forth. If there are  $N$  observations, the largest will have rank  $N$ . In Tables 6.2a and 6.2b, we've made just such a transformation using the following procedure:

1. As the number 89.7 was in cell B3, in cell B7 we inserted the formula = RANK(B3,B\$3:B\$5).
2. We then copied this formula into the region B7:F9.

When three readings are made on each of five subjects, there are a total of  $(3!)^5 = 7776$  possible rearrangements of labels within blocks (subjects). For a test of the null hypothesis against any and all alternatives using  $F_2$  as our test statistic, as can be seen in Table 6.2b, only  $2 \times 5 = 10$  of them, a handful of the total number of rearrangements, are as or more extreme than our original set of ranks.

**Exercise 6.11.** Suppose we discover that the measurement for Subject D for Treatment 1 was recorded incorrectly and should actually be 90. How would this affect the significance of the results depicted in Table 6.2?

**Exercise 6.12.** Does fertilizer help improve crop yield when there is minimal sunlight? Here are the results of a recent experiment exactly as

they were recorded. (Numbers in bushels.) No fertilizer: 5, 10, 8, 6, 9, 122. With fertilizer: 11, 18, 15, 14, 21, 14.

### WHEN TO USE RANKS

1. When one or more extreme-valued observations are suspect.
2. When the methods used to make the measurements were not the same for each observation.

Many older textbooks advocate the use of tests based on ranks for a broad variety of applications. But *rank tests* are simply permutation tests applied to the ranks of observations rather than to their original values. Their value has diminished as a result of improvements in computer technology, and they should not be used except in the two instances outlined above.

Ranks are readily obtained in R by use of the function `rank()`.

## 6.4. STRATIFIED SAMPLES

In Section 5.2.3, we discussed how stratification or blocking of our experiments could be used to reduce unwanted variation. Table 6.3 contains the results of just such an experiment in which plants were grown separately in shade and in full sunlight.

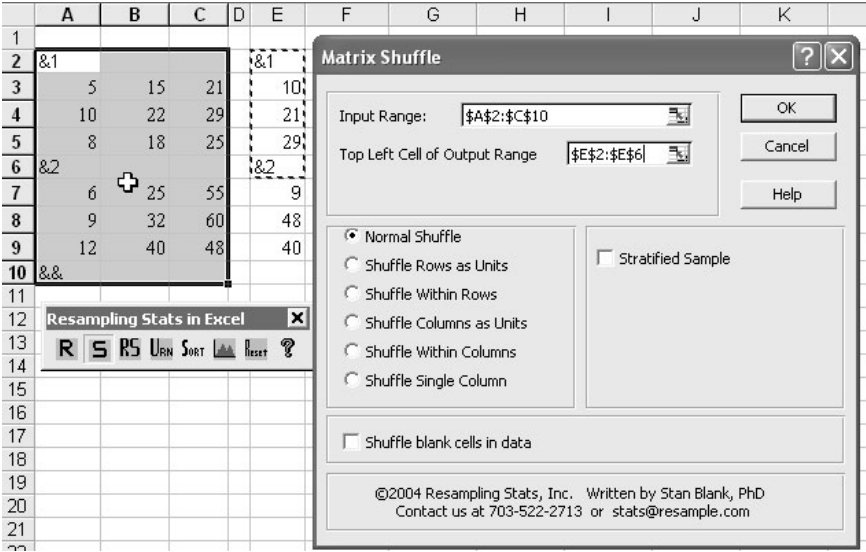
As we would expect yield to increase with increased amounts of fertilizer, Pitman's correlation would appear to be the statistic of choice. We should use the method of Section 6.2.3 to analyze the data with one exception: We have to reshuffle the labels separately in each stratum.

To accomplish this as shown in Fig. 6.3, we need to mark off each stratum with the use of ampersands (&). Note that &1 is placed both before the start of the first stratum of the original observations and to mark the place where the shuffled first stratum is to be located. The marker && is placed in the cell immediately below the first observation in the last stratum.

**Exercise 6.13.** Show that increasing the amount of fertilizer will increase the crop yield.

**TABLE 6.3** Level of Added Fertilizer

Level of Sunlight	Low	Med	High
Low	5,10,8	15,22,18	21,29,25
High	6,9,12	25,32,40	55,60,48



**FIGURE 6.3** Preparing to shuffle data independently within strata.

**TABLE 6.4** Tire Comparison

Vehicle/ Track	Tire Type			
	A	B	C	D
1	15.6	24.6	23.7	16.2
2	9.1	17.1	20.8	11.8
3	13.4	20.3	28.3	16.0
4	12.7	19.8	25.1	15.8
5	11.0	18.2	21.4	14.1

**Exercise 6.14.** Using the data in Table 6.4, determine whether there are significant differences among the various types of tires. Note that we have blocked the data by vehicle to correct for the diverse driving habits of their drivers.

### 6.5. CATEGORICAL DATA

We have shown in two examples (Sections 4.3.4 and 6.2.3) how one may test for the independence of metric variables using a correlation coefficient. But what if our observations are categorical, involving race or gender or some other categorical attribute?

**TABLE 6.5 Cancer Survival as a Function of Gender**

	Survived	Died	Total
Men	9	1	10
Women	4	10	14
Total	13	11	24

Suppose on examining the cancer registry in a hospital, we uncover data that we put in the form of a  $2 \times 2$  contingency table (Table 6.5).

The 9 denotes the number of males who survived, the 1 denotes the number of males who died, and so forth. The four marginal totals or *marginals* are 10, 14, 13, and 11. The total number of men in the study is 10, whereas 14 denotes the total number of women, and so forth.

The marginals in Table 6.5 are fixed because, indisputably, there are 11 dead bodies among the 24 persons in the study and 14 women. Suppose that before completing the table, we lost the subject ids so that we could no longer identify which subject belonged in which category. Imagine you are given two sets of 24 labels. The first set has 14 labels with the word “woman” and 10 labels with the word “man.” The second set of labels has 11 labels with the word “dead” and 12 labels with the word “alive.” Under the null hypothesis that survival is independent of sex, you are allowed to distribute the two sets of labels to the subjects independently of one another. One label from each of the two sets per subject, please.

There are a total of  $\binom{24}{10}$  ways you could hand out the gender labels “man” and “woman.”  $\binom{14}{10}\binom{10}{1}$  of these assignments result in tables that are as extreme as our original table (that is, in which 90% of the men survive) and  $\binom{14}{11}\binom{10}{0}$  in tables that are more extreme (100% of the men survive)—see Tables 6.6a,b.<sup>2</sup>

This is a very small fraction of the total, less than 1%. Consequently, we feel safe in concluding that a difference in survival rates of the two sexes at least as extreme as the difference we observed in our original table is very unlikely to have occurred by chance alone. We reject the hypothesis that the survival rates for the two sexes are the same and accept the alternative

<sup>2</sup> Note that in terms of the relative survival rates of the two sexes, the first of these tables is more extreme than our original Table 6.2. The second is less extreme.

**TABLE 6.6a Cancer Survival as a Function of Gender**

	Survived	Died	Total
Men	10	0	10
Women	3	11	14
Total	13	11	24

**TABLE 6.6b Cancer Survival as a Function of Gender**

	Survived	Died	Total
Men	8	2	10
Women	5	9	14
Total	13	11	24

**TABLE 6.7**

	Category 1	Category 2	Total
Category A	$t - x$	$x$	$t$
Category B	$n - (t - x)$	$M - x$	$M + n - t$
Total	$n$	$m$	$m + n$

hypothesis that, in this instance at least, males are more likely to profit from treatment.

### 6.5.1. One-Sided Fisher’s Exact Test

The preceding test is known as Fisher’s Exact Test as it was first described by R. A. Fisher in 1935. Before we can perform this test, we need to consider the general case depicted in Table 6.7.

If the two attributes represented by the four categories are independent of one another, then each of the tables with the marginals  $n$ ,  $m$ , and  $t$  is equally likely. If  $t$  is the smallest marginal, there are a total of  $\binom{m+n}{t}$  possible tables. If  $t - x$  is the value in the cell with the fewest observations, then  $\sum_{k=0}^{t-x} \binom{m}{t-k} \binom{n}{k}$  tables are as or more extreme than the one we observed.

In Section 2.2.1, we learned to use Excel to compute combinatorials. By making repeated use of the **Combin()** function, we can solve the following exercises.

**Exercise 6.15.** What is the probability of observing Table 6.5 or one more extreme by chance alone?

**Exercise 6.16.** A physician has noticed that half his patients who suffer from sore throats get better within a week if they get plenty of bed rest. (At least they don't call him back to complain that they aren't better.) He decides to do a more formal study and contacts each of 20 such patients during the first week after they've come to see him. What he learns surprises him. Twelve of his patients didn't get much of any bed rest, or if they did go to bed on a Monday, they were back at work on a Tuesday. Of these noncompliant patients, six had no symptoms by the end of the week. The remaining eight patients all managed to get at least three days of bed rest (some by only going to work half-days) and of these, six also had no symptoms by the end of the week. Does bed rest really make a difference?

### 6.5.2. The Two-Sided Test

In the example of the cancer registry, we tested the hypothesis that survival rates do not depend on sex against the alternative that men diagnosed with cancer are likely to live longer than women similarly diagnosed. We rejected the null hypothesis because only a small fraction of the possible tables were as extreme as the one we observed initially. This is an example of a one-tailed test. But is it the correct test? Is this really the alternative hypothesis we would have proposed if we had not already seen the data? Wouldn't we have been just as likely to reject the null hypothesis that men and women profit the same from treatment if we had observed a table like Table 6.8?

Of course we would! In determining the significance level in this example, we must perform a two-sided test and add together the total number of tables that lie in either of the two extremes or tails of the permutation distribution.

Unfortunately, it is not as obvious which tables should be included in the second tail. Is Table 6.8 as extreme as Table 6.5 in the sense that it favors an alternative more than the null hypothesis? One solution is simply to double the  $p$  value we obtained for a one-tailed test. Alternately, we can define and use a test statistic as a basis of comparison. One commonly

**TABLE 6.8**

	Survived	Died	Total
Men	0	10	10
Women	13	1	14
Total	13	11	24

used measure is the Pearson  $\chi^2$  (chi-square) statistic defined for the  $2 \times 2$  contingency table after eliminating terms that are invariant under permutations as  $[\sum - tm/(m+n)]^2$ . For Table 6.5, this statistic is 12.84; for Table 6.8, it is 29.34.

**Exercise 6.17.** Show that Table 6.9a is more extreme (in the sense of having a larger value of the chi-square statistic) than Table 6.5, but Table 6.9b is not.

### 6.5.3. Multinomial Tables

It is possible to extend the approach described in the previous sections to tables with multiple categories such as Table 6.10 and Table 6.11.

**TABLE 6.9a**

	Survived	Died	Total
Men	1	9	10
Women	12	2	14
Total	13	11	24

**TABLE 6.9b**

	Survived	Died	Total
Men	2	8	10
Women	11	3	14
Total	13	11	24

**TABLE 6.10**

	Full Recovery	Partial Recovery	No Improvement
Untreated			
Low Dose			
High Dose			

**TABLE 6.11**

	Urban	Suburban	Rural
Republican			
Democrat			
Independent			

As in the preceding sections, we need only to determine the total number of tables with the same marginals as well as the number of tables that are as or more extreme than the table at hand. Two problems arise. First, what do we mean by more extreme? In Table 6.10, would a row that held one more case of “Full Recovery” be more extreme than a table that held two more cases of “Partial Recovery?” At least a half-dozen different statistics including the Pearson  $\chi^2$  statistic have been suggested for use with tables like Table 6.11 in which neither category is ordered.

The second problem that arises lies in the computations, which are not a simple generalization of the program for the  $2 \times 2$  case. The sole exception, as we shall see in the next section, is if the categories of either the rows or columns can be ordered.

**Exercise 6.18.** In a two-by-two contingency table, once we fix the marginals, we are only free to modify a single entry. In a three-by-three table, how many different entries are we free to change without changing the marginals? Suppose the table has  $R$  rows and  $C$  columns, how many different entries are we free to change without changing the marginals?

#### 6.5.4. Ordered Categories

When either the rows or columns of a contingency table represent ordered categories, we can analyze the data by any of the methods we used for continuous observations, providing we can assign a numeric value to the categories. The leading choices for a scoring method are the following:

1. The category number: 1 for the first ordered category, 2 for the second and so forth
2. The midrank scores
3. Scores determined by the domain expert—a biologist, a physician, a physiologist

To show how such scores might be computed, consider Table 6.12.

The category or equidistant scores are 0, 1, and 2. The ranks of the 44 observations are 1 through 15, 16 through 25, and 26 through 44, so

**TABLE 6.12 Antiemetic Response Data After 2 Days**

	Level of Response			Total
	None	Partial	Complete	
Control	$X_1 = 12$	$X_2 = 3$	$X_3 = 7$	$N_1 = 22$
Treatment	$Y_1 = 3$	$Y_2 = 7$	$Y_3 = 12$	$N_2 = 22$
Total	$T_1 = 15$	$T_2 = 10$	$T_3 = 19$	$N = 44$

Fox et al., 1993.



**TABLE 6.13 "Stem ± Cell Research is Essential"**

	1	2	3	4	5
Republican	25	18	8	20	15
Democrat	8	11	15	20	30
Independent	3	1	8	2	4

that the midrank score of those in the first category is 8, the second 20.5, and the third 35, while physician-chosen scores might be  $-5$ ,  $-2$ , and  $+1$ .

**Exercise 6.19.** Does treatment have an effect on antimetic response after two days?

**Exercise 6.20.** Many surveys use a five-point Likert scale to measure respondents' attitudes, where a value of "1" means the respondent definitely disagrees with a statement and a value of "5" means he definitely agrees. Suppose you have collected the views of Republicans, Democrats, and Independents as in Table 6.13. Analyze the results.

## 6.6. SUMMARY AND REVIEW

In this chapter, you learned to analyze a variety of different types of experimental data. You learned to convert your data to logarithms when changes would be measured in percentages or when analyzing data from dividing populations. You learned to convert your data to ranks when observations were measured on different scales or when you wanted to minimize the importance of extreme observations.

You learned to specify in advance of examining your data whether your alternative hypotheses of interest were one-sided or two-sided, ordered or unordered and to make use of stratification in design and analysis.

You learned how to compare binomial populations, to analyze  $2 \times 2$  contingency tables, and to analyze ordinal data.

**Exercise 6.21.** Write definitions for all italicized words in this chapter.

# Chapter 7

## Developing Models

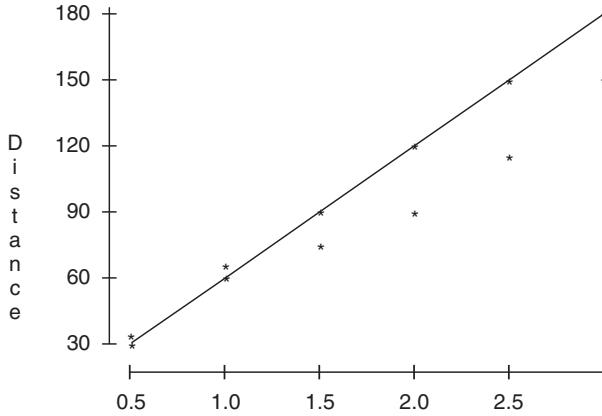
**IN THIS CHAPTER YOU WILL LEARN VALUABLE TECHNIQUES** with which to develop forecasts and classification schemes. These techniques have been used to forecast parts sales by the Honda Motors Company and epidemics at naval training centers, to develop criteria for retention of marine recruits, optimal tariffs for Federal Express, and multitiered pricing plans for Delta Airlines. And these are just examples in which I've been personally involved!

### 7.1. MODELS

A model in statistics is simply a way of expressing a quantitative relationship between one variable, usually referred to as the *dependent variable*, and one or more other variables, often referred to as the *predictors*. We began our text with a reference to Boyle's law for the behavior of perfect gases,  $V = KT/P$ . In this version of Boyle's law,  $V$  (the volume of the gas) is the dependent variable;  $T$  (the temperature of the gas) and  $P$  (the pressure exerted on and by the gas) are the predictors; and  $K$  (known as Boyle's constant) is the *coefficient* of the ratio  $T/P$ .

An even more familiar relationship is that between the distance  $S$  traveled in  $t$  hours and the velocity  $V$  of the vehicle in which we are traveling:  $S = Vt$ . Here  $S$  is the dependent variable and  $V$  and  $t$  are predictors. If we travel at a velocity of 60 mph for 3 hours we can plot the distance we travel over time with Excel as follows:

1. Put the labels Time and Distance at the head of the first two columns.
2. Put the values 0.5, 1, 1.5, 2, 2.5, and 3 in the first column.



**FIGURE 7.1** Distance expected at 60 mph (straight line) vs. distance observed.

- Put the formula = 60 \* A3 in cell B3 and copy it down the column.
- Create a scatterplot, using Excel's Chart Wizard. Select "XY(Scatter)" but use the option "Scatter with data points connected by smoothed lines without markers."

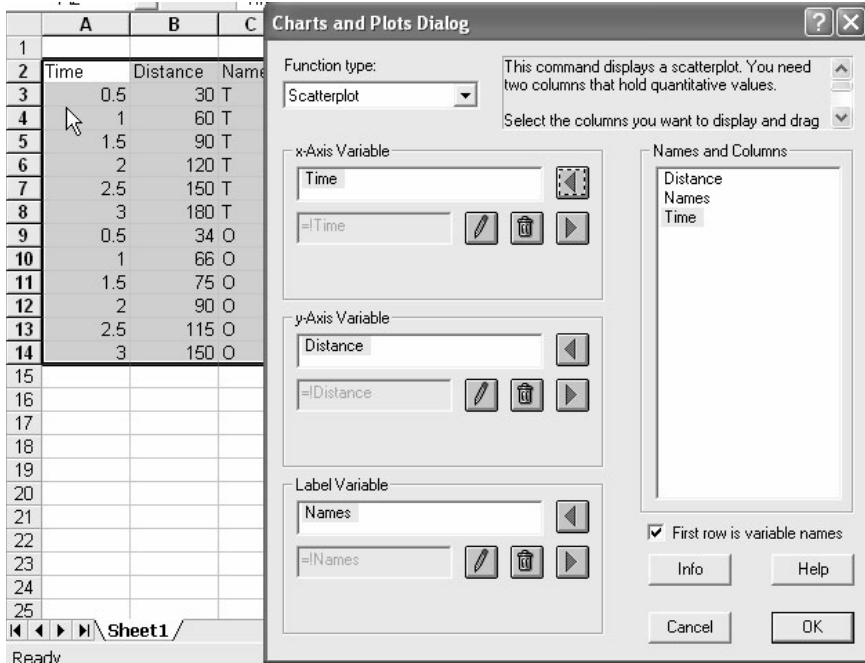
I attempted to drive at 60 mph on a nearby highway past where a truck had recently overturned. Recording the distances at half-hour intervals, I found I'd traveled 32, 66, 75, 90, 115, and 150 miles.

As you can see from Fig. 7.1, the reality on a busy highway was quite different from what theory would predict. Incidentally, I created this figure with the aid of DDXL. The setup is depicted in Fig. 7.2.

**Exercise 7.1.** My average velocity over the three-hour period was equal to distance traveled/time =  $150/3 = 50$  miles per hour, or  $\text{Distance}_i = 50 * \text{Time}_i + z_i$ , where the  $\{z_i\}$  are random deviations from the expected distance. Construct a graph to show that this new model is a much better fit than the old.

### 7.1.1. Why Build Models?

We develop models for at least three different purposes. First, as the term "predictors" suggests, models can be used for *prediction*. A manufacturer of automobile parts will want to predict part sales several months in advance to ensure that its dealers have the necessary parts on hand. Too few parts in stock will reduce profits; too many may necessitate interim borrowing. So entire departments are hard at work trying to come up with the needed formula.



**FIGURE 7.2** Preparing a scatterplot that will depict multiple lines.

At one time, I was part of just such a study team. We soon realized that the primary predictor of part sales was the weather. Snow, sleet, and freezing rain sent sales skyrocketing. Unfortunately, predicting the weather is as or more difficult than predicting part sales.

Models can be used to develop additional insight into cause-and-effect relationships. At one time, it was assumed that the growth of the welfare caseload  $L$  was a simple function of time  $t$ , so that  $L = ct$ , where the growth rate  $c$  was a function of population size. Throughout the 1960s, in state after state, the constant  $c$  constantly had to be adjusted upward if this model were to fit the data. An alternative and better-fitting model proved to be  $L = ct + dt^2$ , an equation often used in modeling the growth of an epidemic. As it proved, the basis for the new second-order model was the same as it was for an epidemic: Welfare recipients were spreading the news of welfare availability to others who had not yet taken advantage of the program much as diseased individuals might spread an infection.

Boyle's law seems to fit the data in the sense that if we measure both the pressure and volume of gases at various temperatures, we find that a plot of pressure times volume versus temperature yields a straight line. Or

if we fix the volume, say by confining all the gas in a chamber of fixed size with a piston on top to keep the gas from escaping, a plot of the pressure exerted on the piston against the temperature of the gas yields a straight line.

Observations such as these both suggested and confirmed what is known today as kinetic molecular theory.

A third use for models is in *classification*. At first glance, the problem of classification might seem quite similar to that of prediction. For example, instead of predicting that  $Y$  would be 5 or 6 or even 6.5, we need only predict that  $Y$  will be greater or less than 6. But the loss functions for the two problems are quite different. The loss connected with predicting  $y_p$  when the observed value is  $y_o$  is usually a monotone increasing function of the difference between the two. By contrast, the loss function connected with a classification problem has jumps, being zero if the classification is correct, and taking one of several possible values otherwise, depending on the nature of the misclassification.

Not surprisingly, different modeling methods have developed to meet the different purposes. For the balance of this chapter, we shall consider two primary modeling methods: linear regression, whose objective is to predict the expected value of a given dependent variable, and decision trees, which are used for classification. We shall briefly discuss some other alternatives.

### 7.1.2. Caveats

The modeling techniques that you learn in this chapter may seem impressive—they require extensive calculations that only a computer can do—so I feel it necessary to issue three warnings.

- You cannot use the same data both to formulate a model and to test it. It must be independently validated.
- A cause-and-effect basis is required for every model, just as molecular theory serves as the causal basis for Boyle's law.
- Don't let your software do your thinking for you. Just because a model fits the data does not mean that it is appropriate or correct. It must be independently validated and have a cause-and-effect basis.

You may have heard that having a black cat cross your path will bring bad luck. Don't step in front of a moving vehicle to avoid that black cat unless you have some causal basis for believing that black cats can affect your luck. (And why not white cats or tortoiseshell?) I avoid cats myself because cats lick themselves and shed their fur; when I breathe cat hairs,

the traces of saliva on the cat fur trigger an allergic reaction that results in the blood vessels in my nose dilating. Now that is a causal connection.

## 7.2. REGRESSION

Regression combines two ideas with which we gained familiarity in previous chapters:

1. Correlation or dependence among variables
2. Additive model

Here is an example: Anyone familiar with the restaurant business (or indeed, with any number of businesses that provide direct service to the public, including the post office) knows that the volume of business is a function of the day of the week. Using an *additive model*, we can represent business volume via the formula

$$V_{ij} = \mu + \delta_i + z_{ij}$$

where  $V_{ij}$  is the volume of business on the  $i$ th day of the  $j$ th week,  $\mu$  is the average volume,  $\delta_i$  is the deviation from the average volume observed on the  $i$ th day of the week,  $i = 1, \dots, 7$ , and the  $z_{ij}$  are independent, identically distributed random fluctuations.

Many physiological processes such as body temperature have a circadian rhythm, rising and falling each 24 hours. We could represent body temperature by the formula

$$T_{ij} = \mu + \delta_i + z_{ij},$$

where  $i$  (in minutes) takes values from 1 to  $24 * 60$ , but this would force us to keep track of 1441 different parameters. Besides, we can get almost as good a fit to the data by using the formula

$$E(T_{ij}) = \mu + \beta \cos(2\pi * (t + 300)/1440) \quad (7.1)$$

If you are not familiar with the  $\cos()$  function, you can use Excel to gain familiarity as follows:

1. Put the hours from 1 to 24 in the first column.
2. In the third cell of the second column, put  $= \cos(2 * 3.1412 * (A3 + 6)/24)$ .
3. Copy the formula down the column; then construct a scatterplot.

Note how the  $\cos()$  function first falls then rises, undergoing a complete cycle in a 24-hour period.

Why use a formula as complicated as Equation 7.1? Because now we have only two parameters we need to estimate,  $\mu$  and  $\beta$ . For predicting body temperature,  $\mu = 98.6$  and  $\beta = 0.4$  might be reasonable choices. Of course, the values of these parameters will vary from individual to individual. For me,  $\mu = 97.6$ .

**Exercise 7.2.** If  $E(Y) = 3X + 2$ , can  $X$  and  $Y$  be independent?

**Exercise 7.3.** According to the inside of the cap on a bottle of Snapple's Mango Madness, "the number of times a cricket chirps in 15 seconds plus 37 will give you the current air temperature." How many times would you expect to hear a cricket chirp in 15 seconds when the temperature is 39 degrees? 124 degrees?

**Exercise 7.4.** If we constantly observe large values of one variable, call it  $Y$ , whenever we observe large values of another variable, call it  $X$ , does this mean  $X$  is part of the mechanism responsible for increases in the value of  $Y$ ? If not, what are the other possibilities? To illustrate the several possibilities, give at least three real-world examples in which this statement would be false. (You'll do better at this exercise if you work on it with one or two others.)

### 7.2.1. Linear Regression

Equation 7.1 is an example of linear regression. The general form of linear regression is

$$Y = \mu + \beta f[X] + Z \quad (7.2)$$

Where  $Y$  is known as the *dependent* or *response variable*,  $X$  is known as the *independent variable* or *predictor*,  $f[X]$  is a function of known form,  $\mu$  and  $\beta$  are unknown *parameters*, and  $Z$  is a random variable whose expected value is zero. If it weren't for this last random component  $Z$ , then if we knew the parameters  $\mu$  and  $\beta$ , we could plot the values of the dependent variable  $Y$  and the function  $f[X]$  as a straight line on a graph; hence the name: *linear regression*.

For the past year, the price of homes in my neighborhood could be represented as a straight line on a graph relating house prices to time,  $P = \mu + \beta t$ , where  $\mu$  was the price of the house on the first of the year and  $t$  is the day of the year. Of course, as far as the price of any individual house

was concerned, there was a lot of fluctuation around this line depending on how good a salesman the realtor was and how desperate the owner was to sell.

If the price of my house ever reaches \$700K, I might just sell and move to Australia. Of course, a straight line might not be realistic. Prices have a way of coming down as well as going up. A better prediction formula might be  $P = \mu + \beta t - \gamma t^2$ , in which prices continue to rise until  $\beta - \gamma t = 0$ , after which they start to drop. If I knew what  $\beta$  and  $\gamma$  were or could at least get some good estimates of their value, then I could sell my house at the top of the market!

The trick is to look at a graph such as Fig. 7.1 and somehow extract that information.

Note that  $P = \mu + \beta t - \gamma t^2$  is another example of *linear regression*, only with three parameters rather than two. So is the formula  $W = \mu + \beta H + \gamma A + Z$  where  $W$  denotes the weight of a child,  $H$  is its height,  $A$  its age, and  $Z$ , as always, is a purely random component.  $W = \mu + \beta H + \gamma A + \delta AH + Z$  is still another example. The parameters  $\mu$ ,  $\beta$ ,  $\gamma$ , and so forth are sometimes referred to as the *coefficients* of the model.

What then is a *nonlinear* regression? Here are two examples:

$$Y = \beta \log(\gamma X), \text{ which is linear in } \beta \text{ but nonlinear in the unknown parameter } \gamma$$

and

$$T = \beta \cos(t + \gamma), \text{ which also is linear in } \beta \text{ but nonlinear in } \gamma.$$

Regression models that are nonlinear in their parameters are beyond the scope of this text. The important lesson to be learned from their existence is that we need to have some idea of the functional relationship between a response variable and its predictors before we start to fit a linear regression model.

**Exercise 7.5.** Generate a plot of the function  $P = 100 + 10t - 1.5t^2$  for values of  $t = 0, 1, \dots, 10$ . Does the curve reach a maximum and then turn over?

### 7.3. FITTING A REGRESSION EQUATION

Suppose we have determined that the *response variable*  $Y$  whose value we wish to predict is related to the value of a *predictor variable*  $X$  by the



equation,  $E(Y) = a + bX$  and on the basis of a sample of  $n$  paired observations  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  we wish to estimate the unknown coefficients  $a$  and  $b$ . Three methods of estimation are in common use: ordinary least squares, least absolute deviation, and error-in-variable, also known as Deming regression. We will study all three in the next few sections.

### 7.3.1. Ordinary Least Squares

The ordinary least squares (OLS) technique of estimation is the most commonly used, primarily for historical reasons, as its computations can be done (with some effort) by hand or with a primitive calculator. The objective of the method is to determine the parameter values that will minimize the sum of squares  $\Sigma(y_i - EY)^2$  where  $EY$ , the expected or mean value of  $Y$ , is modeled by the right-hand side of our regression equation.

In our example,  $EY = a + bx_i$ , and so we want to find the values of  $a$  and  $b$  that will minimize  $\Sigma(y_i - a - bx_i)^2$ . We can readily obtain the desired estimates with the aid of the XLStat add-in.

Suppose we have the following data relating age and systolic blood pressure (SBP):

- Age 39,47,45,47,65,46,67,42,67,56,64,56,59,34,42
- SBP 144,220,138,145,162,142,170,124,158,154,162,150,140,110,128

From the main XLStat menu



select the

scatterplot (fifth from left). Select the straight-line scatter plot (second

from left)



from the modeling data

menu that pops up. Enter the observations in the first two columns and complete the Linear Regression menu as shown in Fig. 7.3.

A plethora of results appear on a second worksheet. Let's focus on what is important. In Table 7.1, extracted from the worksheet, we see that the best-fitting model by least squares methods is that the expected SBP of an individual is  $95.6125119693584 + 1.04743855729333$  times that person's Age. Note that when we report our results, we write this as  $\hat{E}(\text{SBP}) = \hat{a} + \hat{b}\text{Age} = 95.6 + 1.04\text{Age}$ , dropping decimal places that convey a false impression of precision.

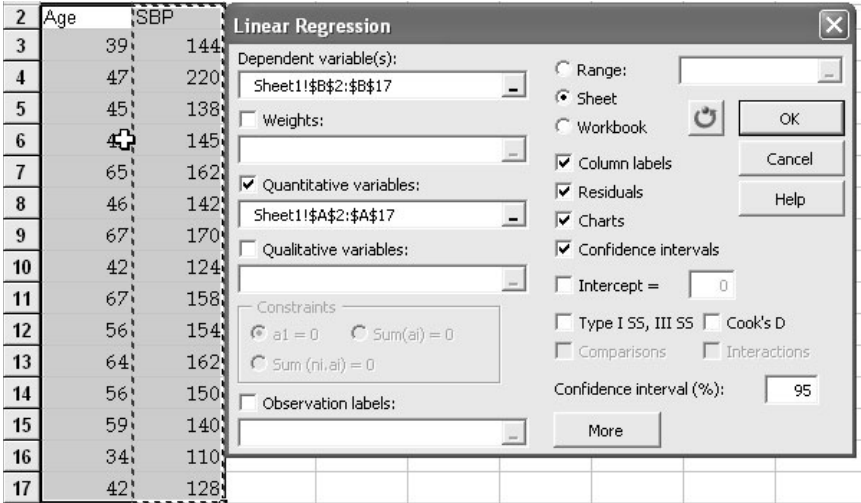


FIGURE 7.3 Preparing to fit a regression line.

TABLE 7.1 Model Parameters

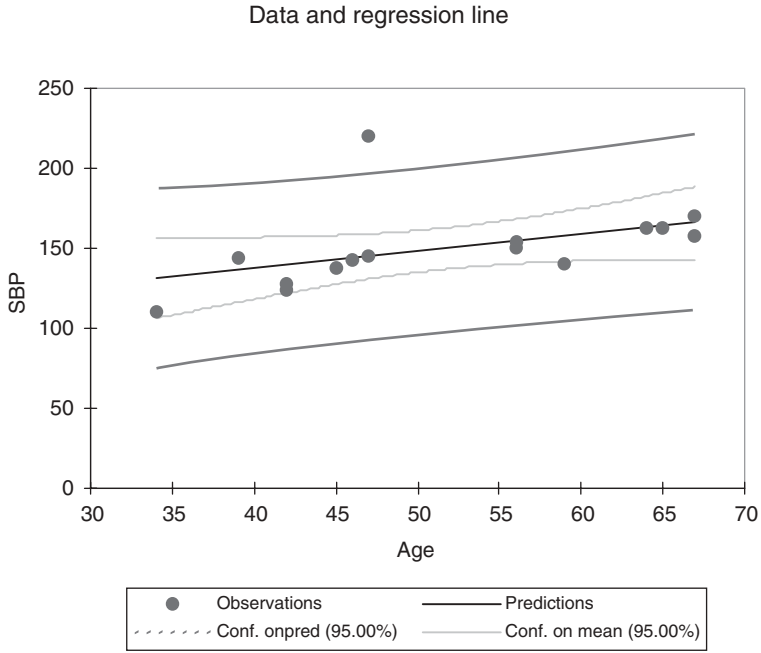
Parameter	Value	Standard Deviation	Student's t	Pr > t	Lower bound 95%	Upper bound 95%
Intercept	95.613	29.894	3.198	0.007	31.031	160.194
Age	1.047	0.566	1.850	0.087	-0.176	2.271

The equation of the model writes:  $SBP = 95.6125119693584 + 1.04743855729333 * Age$

We also see from Table 7.1 that the coefficient of Age, that is, the *slope* of the regression line depicted in Fig. 7.4, is not significantly different from zero at the 5% level. The associated  $p$  value is  $0.087 > 0.05$ . Whether this  $p$  value is meaningful is the topic of Section 7.4.1.

What can be the explanation for the poor fit? Our attention is immediately drawn to the point in Fig. 7.4 that stands out from the rest. It is that of a 47-year old whose systolic blood pressure is 220. Part of our output, reproduced in Table 7.2, includes a printout of all the *residuals*, that is, of the differences between the values our regression equation would predict and the SBPs that were actually observed.

Consider the fourth residual in the series, 0.158. This is the difference between what was observed,  $SBP = 145$ , and what the regression equation estimates as the expected SBP for a 47-year-old individual  $E(SBP) = 95.6 + 1.04 * 47 = 144.8$ . The largest residual is 75, which corresponds to the outlying value we've already alluded to.



**FIGURE 7.4** Data and regression line of SBP vs. Age.

**TABLE 7.2** Deviations from Regression Line

Age	SBP	SBP (Model)	Residuals
39.000	144.000	136.463	7.537
47.000	220.000	144.842	75.158
45.000	138.000	142.747	-4.747
47.000	145.000	144.842	0.158
65.000	162.000	163.696	-1.696
46.000	142.000	143.795	-1.795
67.000	170.000	165.791	4.209
42.000	124.000	139.605	-15.605
67.000	158.000	165.791	-7.791
56.000	154.000	154.269	-0.269
64.000	162.000	162.649	-0.649
56.000	150.000	154.269	-4.269
59.000	140.000	157.411	-17.411
34.000	110.000	131.225	-21.225
42.000	128.000	139.605	-11.605

**Economic Report of the President, 1988,  
Table B-27**

Year	Income 1982 \$s	Expenditures 1982 \$s
1960	6036	5561
1962	6271	5729
1964	6727	6099
1966	7280	6607
1968	7728	7003
1970	8134	7275
1972	8562	7726
1974	8867	7826
1976	9175	8272
1978	9735	8808
1980	9722	8783
1982	9725	8818

**Exercise 7.6.** Do U.S. residents do their best to spend what they earn? Fit a regression line, using OLS, to the data in the accompanying table relating disposable income to expenditures in the United States from 1960 to 1982.

**Exercise 7.7.** Suppose we've measured the dry weights of chicken embryos at various intervals at gestation and recorded our findings in the following table:

Age (days)	6	7	8	9	10	11	12	13	14	15	16
Weight (g)	0.029	0.052	0.079	0.125	0.181	0.261	0.425	0.738	1.130	1.882	2.812

Obtain a plot of the regression line of weight with respect to age on which the actual observations are superimposed. Recall from Section 6.1 that the preferable way to analyze growth data is by using the logarithms of the exponentially increasing values. Obtain a plot of the new regression line of  $\log(\text{weight})$  as a function of age. Which line (or model) appears to provide the better fit to the data?

**Exercise 7.8.** Obtain and plot the OLS regression of systolic blood pressure with respect to age after discarding the outlying value of 220 recorded for a 47-year-old individual. Is the slope of this regression line significant at the 5% level?

Of course, we just can't go around discarding observations because they don't quite fit our preconceptions. There are two possible reasons why we may have had an outlier in this example:

1. We made mistakes when we recorded this particular individual's age and blood pressure.
2. Other factors such as each individual's weight-to-height ratio might be as or more important than age in determining blood pressure. Or the 47-year-old individual whose readings we question might suffer from diabetes, unlike the others in our study.

If we had data on weight and height as well as age and systolic blood pressure, we might write

$$\bullet \text{ SBP} = a + b * \text{Age} + c * \text{weight}/(\text{height} * \text{height}).$$

**Exercise 7.9.** In a further study of systolic blood pressure as a function of age, the height and weight of each individual were recorded. The latter were converted to a Quetlet index using the formula  $\text{QUI} = 100 * \text{weight}/\text{height}^2$ . Fit a multivariate regression line of systolic blood pressure with respect to age and the Quetlet index, using the following information:

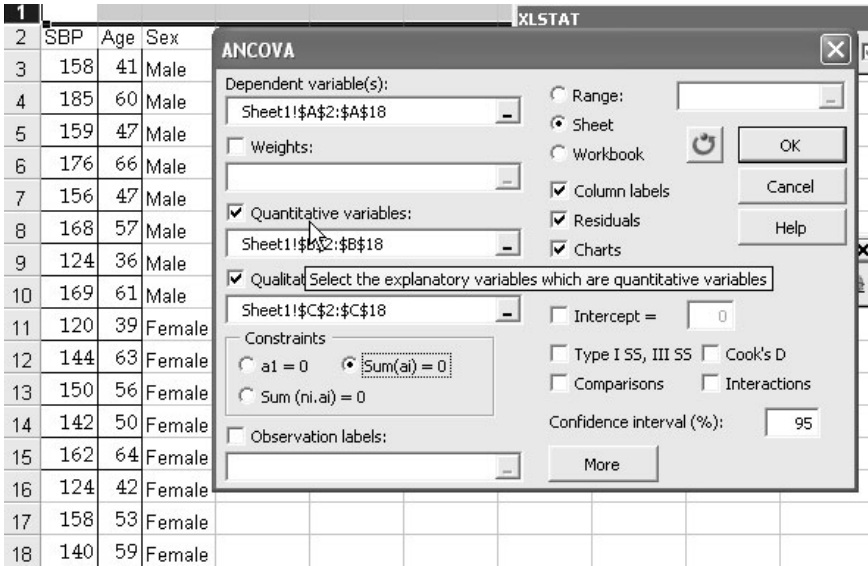
Age	41,	43,	45,	48,	49,	52,	54,	56,	57,	59,	62,	63,	65
SBP	122,	120,	135,	132,	130,	148,	146,	138,	135,	166,	152,	170,	164
QUI	3.25,	2.79,	2.88,	3.02,	3.10,	3.77,	2.98,	3.67,	3.17,	3.88,	3.96,	4.13,	4.01

**Types of Data.** The linear regression model is a *quantitative* one. When we write  $Y = 3 + 2X$ , we imply that the product  $2X$  will be meaningful. This will be the case if  $X$  is a metric variable. In many surveys, respondents use a nine-point *Likert scale*, where a value of “1” means they definitely disagree with a statement and “9” means they definitely agree. Although such data are ordinal and not metric, the regression equation is still meaningful.

When one or more predictor variables are categorical, we must use a different approach. The regression model will include a different additive component for each level of the categorical or *qualitative* variable. Thus we can include sex or race as predictors in a regression model.

Figure 7.5 illustrates the setup of a linear regression model with both quantitative (continuous) and qualitative predictors. Note that if you have multiple predictors of the same data type, they should be placed in adjacent columns.

As can be seen in Table 7.3, providing for differences in the sexes appears to lead to a better-fitting model. One caveat: By including sex as a factor in the model, we have tacitly assumed that the slope of the regres-



**FIGURE 7.5** Setting up a regression model to make use of both continuous and categorical predictors.

**TABLE 7.3** Parameters of a model with both quantitative and qualitative predictors

Parameter	Value	Standard Deviation	Student's t	Pr > t	Lower Bound 95%	Upper Bound 95%
Intercept	79.027	13.592	5.814	< 0.0001	49.663	108.391
Age	1.392	0.255	5.466	0.000	0.842	1.942
Sex—Male	10.644	—	—	—	—	—
Sex—Female	-10.644	2.375	-4.482	0.001	-15.775	-5.513

sion line is the same for both sexes. If this is not the case, we would be better to fit separate regression lines to the data for each sex.

**Exercise 7.10.** Use the data displayed in Fig. 7.5 to fit separate regression lines of systolic blood pressure as a function of age for each sex. Are the slopes of the two regression lines approximately the same?

**Exercise 7.11.** Make use of the following data regarding smoking habits in fitting a model to the systolic blood pressure data: Smoke 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0

**Exercise 7.12.** The following data are based on samples taken from swimming areas off the coast of Milazzo (Italy) from April through

September 1998. Included in this data set are levels of fecal coliform, dissolved oxygen, and temperature.

- Are there significant differences in each of these variables from month to month?
- Develop a model for fecal coliform levels in terms of month, temperature, and dissolved oxygen.

Month 4, 5, 6, 7, 8, 9, 4, 5, 6, 7, 8, 9, 4, 5, 6, 7, 8, 9, 4, 5, 6, 7, 8, 9, 4, 5, 6, 7, 8, 9, 4, 5, 6, 7, 8, 9, 4, 5, 6, 7, 8, 9, 4, 5, 6, 7, 8, 9, 4, 5, 6, 7, 8, 9, 4, 5, 6, 7, 8, 9, 4, 5, 6, 7, 8, 9

Temp 14, 17, 24, 21, 22, 20, 14, 17, 24, 21, 23, 22, 14, 17, 25, 21, 21, 22, 14, 17, 25, 21, 25, 20, 14, 17, 25, 21, 21, 19, 14, 17, 25, 21, 25, 19, 14, 16, 25, 21, 25, 19, 15, 19, 18, 21, 25, 19, 15, 19, 18, 20, 22, 17, 15, 19, 18, 20, 23, 18, 15, 17, 18, 20, 25, 17, 15, 17, 18, 20, 25, 19, 15, 17, 19, 20, 25, 18, 15, 18, 19, 21, 24, 19

FecalColiform 16, 8, 8, 11, 11, 21, 34, 11, 11, 7, 11, 6, 8, 6, 35, 18, 18, 21, 13, 9, 32, 11, 29, 11, 28, 7, 12, 7, 12, 9, 10, 3, 43, 5, 12, 14, 4, 9, 8, 10, 4, 12, 0, 4, 7, 5, 12, 26, 0, 3, 32, 0, 8, 12, 0, 0, 21, 0, 7, 8, 0, 0, 17, 4, 0, 14, 0, 0, 11, 7, 6, 0, 8, 0, 6, 4, 5, 10, 14, 3, 8, 12, 11, 27

Oxygen 95.64, 102.09, 104.76, 106.98, 102.6, 109.15, 96.12, 111.98, 100.67, 103.87, 107.57, 106.55, 89.21, 100.65, 100.54, 102.98, 98, 106.86, 98.17, 100.98, 99.78, 100.87, 97.25, 97.78, 99.24, 104.32, 101.21, 102.73, 99.17, 104.88, 97.13, 102.43, 99.87, 100.89, 99.43, 99.5, 99.07, 105.32, 102.89, 102.67, 106.04, 106.67, 98.14, 100.65, 103.98, 100.34, 98.27, 105.69, 96.22, 102.87, 103.98, 102.76, 107.54, 104.13, 98.74, 101.12, 104.98, 101.43, 106.42, 107.99, 95.89, 104.87, 104.98, 100.89, 109.39, 98.17, 99.14, 103.87, 103.87, 102.89, 108.78, 107.73, 97.34, 105.32, 101.87, 100.78, 98.21, 97.66, 96.22, 22, 99.78, 101.54, 100.53, 109.86

**Exercise 7.13.** The slope of a regression line is zero if and only if the correlation between the predictor and the predicted variable is zero. Use what you learned in previous chapters to test whether the slope of the regression line of systolic blood pressure versus age is zero.

### 7.3.2. Least Absolute Deviation Regression

Least absolute deviation regression (LAD) attempts to correct one of the major flaws of OLS, that of giving sometimes excessive weight to extreme values. The LAD method solves for those values of the coefficients in the regression equation for which the sum of the absolute deviations  $\sum |y_i - R[x_i]|$  is a minimum. Unfortunately, no add-in for Excel is available to do LAD regression at the time of this writing.

### 7.3.3. Errors-in-Variables Regression

The need for errors-in-variables (EIV) or Deming regression is best illustrated by the struggles of a small medical device firm to bring its product

to market. Their first challenge was to convince regulators that their long-lasting device provided results equivalent to those of a less-efficient device already on the market. In other words, they needed to show that the values  $V$  recorded by their device bore a linear relation to the values  $W$  recorded by their competitor, that is, that  $E(V) = a + bW$ .

In contrast to the examples of regression we looked at earlier, the errors inherent in measuring  $W$  (the so-called predictor) were as large if not larger than the variation inherent in the output  $V$  of the new device.

The EIV regression method they used to demonstrate equivalence differs in two respects from that of OLS:

1. With OLS, we are trying to minimize the sum of squares  $\sum(y_{oi} - y_{pi})^2$  where  $y_{oi}$  is the  $i$ th observed value of  $Y$  and  $y_{pi}$  is the  $i$ th predicted value. With EIV, we are trying to minimize the sums of squares of errors, going both ways:  $\sum(y_{oi} - y_{pi})^2/\text{Var}Y + \sum(x_{oi} - x_{pi})^2/\text{Var}X$ .
2. The coefficients of the EIV regression line depend on  $\lambda = \text{Var}X/\text{Var}Y$ . ( $\lambda$  is pronounced lambda.)

Minimizing the sum of squares yields the formulas:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \frac{-(S_{xx} - \lambda S_{yy}) + \sqrt{(S_{xx} - \lambda S_{yy})^2 + 4S_{xy}^2}}{2\lambda S_{xy}}$$

where

$$S_{xx} = \sum (x_i - \text{Mean}[x])^2 \quad S_{yy} = \sum (y_i - \text{Mean}[y])^2$$

$$S_{xy} = \sum (x_i - \text{Mean}[x])(y_i - \text{Mean}[y]).$$

To prepare the worksheet depicted in part in Fig. 7.6, I went through the following steps:

1. Placed the observed values in columns 1 and 2. In this example, the observations occupy cells A3:B26.
2. Placed the average of the first column in A29.
3. Placed the formula  $(X - MX)^2 = (B3 - B\$26) * (B3 - B\$26)$  in cell C3. The \$ is essential.
4. Copied these formulas to C4:C29 and E3:E29.
5. Placed the formula for  $S_{xx} = \text{SUM}(C3:C26)$  in B30.
6. Copied cells B29 and B30 to D29 and D30.
7. Placed the formula for the cross product of  $xy = (B3 - B\$29) * (D3 - D\$29)$  in F3.



	A	B	C	D	E	F	G	H	
1									
2		Old	Devbold	New	DevNew	Xprod	Predicted	Residual	
18		0.95	0.9933	2.28	0.8055	0.8945	1.31	-0.97	
19		1.05	0.8040	3.60	0.1785	-0.3788	1.49	-2.11	
20		2.92	0.9474	5.39	4.8952	2.1535	5.00	-0.39	
21		1.76	0.0348	4.12	0.8883	-0.1759	2.83	-1.29	
22		0.51	2.0640	3.16	0.0003	0.0251	0.48	-2.68	
23		2.17	0.0499	4.40	1.4945	0.2730	3.60	-0.80	
24		⊕	1.99	0.0019	1.18	3.9900	-0.0866	3.26	2.08
25		1.53	0.1736	2.54	0.4064	0.2656	2.40	-0.14	
26		2.60	0.4268	4.89	2.9327	1.1188	4.40	-0.49	
27									
28									
29	Mx	1.95	My	3.18					
30	Sxx	22.82	Syy	43.62					
31	Sxy	15.48	diff=Syy-Sxx	20.80					
32	sqrt	37.30	b=(diff+sqrt)	1.88					
33			a=My-bMX	-0.48					

**FIGURE 7.6** Worksheet for calculating error-in-variable regression coefficients.

8. Placed the formula for  $S_{xy} = \text{SUM}(F3:F26)$  in B31.
9. Placed the formula for the difference  $S_{yy} - S_{xx}$  in D31.
10. Placed the formula  $= \text{SQRT}(D31 * D31 + 4 * B31 * B31)$  in B32.
11. Placed the formula for the slope  $b = (D31 + B32) / (2 * B31)$  in D32.
12. Placed the formula for the intercept  $a$  of the regression line with the X-axis in D33.
13. These coefficients were used to place the first predicted value  $= D\$33 + D\$32 * B3$  in G3.
14. Placed the first residual  $\text{New}(\text{Predicted}) - \text{New}(\text{observed}) = G3 - D3$  in H3.
15. Finished by copying the formula in G3 and H3 down their respective columns.

To check your results against mine, here is the complete set of values that I used:

Old 3.74, 3.66, 0.78, 2.40, 2.18, 1.93, 0.20, 2.50, 3.50, 1.35, 2.36, 3.13, 1.22, 1.00, 1.29, 0.95, 1.05, 2.92, 1.76, 0.51, 2.17, 1.99, 1.53, 2.60

New 3.22, 4.87, 0.12, 2.31, 4.25, 2.24, 2.81, 3.71, 3.11, 0.90, 4.39, 4.36, 1.23, 3.13, 4.05, 2.28, 3.60, 5.39, 4.12, 3.16, 4.40, 1.18, 2.54, 4.89

**Exercise 7.14.** Are the following two sets of measurements comparable, that is, does the slope of the EIV regression line differ significantly from unity?

OLD 2.521, 3.341, 4.388, 5.252, 6.422, 7.443, 8.285, 9.253, 10.621, 10.405, 11.874, 13.444, 13.343, 16.402, 19.108, 19.25, 20.917, 23.409, 5.583, 5.063, 6.272, 7.469, 10.176, 6.581, 7.63

NEW 2.362, 3.548, 4.528, 4.923, 6.443, 6.494, 8.275, 9.623, 9.646, 11.542, 10.251, 11.866, 13.388, 17.666, 17.379, 21.089, 21.296, 23.983, 5.42, 6.369, 7.899, 8.619, 11.247, 7.526, 7.653

**Exercise 7.15.** Which method should be used to regress  $U$  as a function of  $W$  in the following cases, OLS, LAD, or EIV?

- Some of the  $U$  values are suspect.
- It's not clear whether  $U$  or  $W$  is the true independent variable or whether both depend on the value of a third hidden variable.
- Minor errors in your predictions aren't important; large ones could be serious.

### 7.3.4. Assumptions

To use any of the preceding linear regression methods the following as-yet-unstated assumptions must be satisfied:

- Independent random components.* In the model  $y_i = \mu + \beta x_i + z_i$ , the random fluctuations  $z_i$  must be independent of one another. If the  $z_i$  are not, it may be that a third variable  $W$  is influencing their values. In such a case, we would be advised to try the model  $y_i = \mu + \beta x_i + \gamma w_i + \varepsilon_i$ .  
When observations are made one after the other in time, it often is the case that successive errors are dependent on one another, but we can remove this dependence if we work with the increments  $y_2 - y_1$ ,  $y_3 - y_2$ , and so forth. Before we start fitting our model, we would convert to these differences.
- Identically distributed random components.* Often, it is the case that large random fluctuations are associated with larger values of the observations. In such cases, techniques available in more advanced textbooks provide for weighting the observations when estimating model coefficients so that the least variable observations receive the highest weight.
- Random fluctuations come from a specific distribution.* Most statistics packages provide tests of the hypotheses that the model parameters are significantly different from zero given that the random fluctuations come from a normal distribution. This is true of the  $p$  values and confident limits displayed in Table 7.1. If the fluctuations come from some other distribution, the displayed values may be quite misleading.

An alternate approach is to obtain confidence interval for the coefficients by taking a series of bootstrap samples from the collection of pairs of observations (39,144), (47,220), . . . , (42,128). To accomplish this with the Resampling Stats add-in, perform a “Matrix Shuffle” using the “Shuffle Rows Within Units” option.

**Exercise 7.16.** Using the data from Exercise 7.14, obtain bootstrap confidence intervals for the EIV regression coefficients.

## 7.4. PROBLEMS WITH REGRESSION

At first glance, regression seems to be a panacea for all modeling concerns. But it has a number of major limitations, just a few of which we will discuss in this section.

- **The model that best fits the data we have in hand may not provide the best fit to the data we gather in the future.**
- **More than one linear regression model may provide a statistically significant fit to our data.**

### 7.4.1. Goodness of Fit Versus Prediction

Two assumptions we make whenever we use a regression equation to make predictions are:

1. Relationships among the variables and, thus, the true regression line remain unchanged over time.
2. The sources of variation are the same as when we first estimated the coefficients.

We are seldom on safe grounds when we attempt to use our model outside the range of predictor values for which it was developed originally. For one thing, literally every phenomenon seems to have nonlinear behavior for very small and very large values of the predictors. Treat every predicted value outside the original data range as suspect.

In my lifetime, regression curves failed to predict a drop in the sales of '78 records as a result of increasing sales of '45s, a drop in the sales of '45s as a result of increasing sales of 8-track tapes, a drop in the sales of 8-track tapes as a result of increasing sales of cassettes, nor a drop in the sales of cassettes as a result of increasing sales of CDs. It is always advisable to revalidate any model that has been in use for an extended period (see Section 7.5.2).

Finally, let us not forget that our models are based on samples, and that sampling error must always be inherent in the modeling process.

**Exercise 7.17.** Redo Exercise 7.2.

**Exercise 7.18.** The state of Washington uses an audit recovery formula in which the percentage to be recovered (the overpayment) is expressed as a linear function of the amount of the claim. The slope of this line is close to zero. Sometimes, depending on the audit sample, the slope is positive, and sometimes it is negative. Can you explain why?

### 7.4.2. Which Model?

The exact nature of the formula connecting two variables cannot be determined by statistical methods alone. If a linear relationship exists between two variables  $X$  and  $Y$ , then a linear relationship also exists between  $Y$  and any monotone (nondecreasing or nonincreasing) function of  $X$ . Assume  $X$  can only take positive values. If we can fit Model I:  $Y = \alpha + \beta X + \varepsilon$  to the data, we also can fit Model II:  $Y = \alpha' + \beta' \log[X] + \varepsilon$ , and Model III:  $Y = \alpha'' + \beta'' X + \gamma X^2 + \varepsilon$ . It can be very difficult to determine which model if any is the “correct” one.

Five principles should guide you in choosing among models:

1. *Prevention.* The data you collect should span the entire range of interest. For example, when employing EIV regression to compare two methods of measuring glucose, it is essential to observe many pairs of observed abnormal values (characteristic of a disease process) along with the more readily available pairs of normal values. Don't allow your model to be influenced by one or two extreme values—whenever this is a possibility, use LAD regression rather than OLS. Strive to obtain response observations at intervals throughout the relevant range of the predictor. Only when we have observations spanning the range of interest can we begin to evaluate competitive models.
2. *Think why rather than what.* In Exercise 7.7, we let our knowledge of the underlying growth process dictate the use of  $\log(X)$  rather than  $X$ . As a second example, consider that had we wanted to find a relationship between the volume  $V$  and temperature  $T$  of a gas, any of the preceding three models might have been used to fit the relationship. But only one, the model  $V = a + KT$ , is consistent with kinetic molecular theory.
3. *Plot the residuals.* That is, plot the error or difference between the values predicted by the model and the values that were actually observed. If a pattern emerges from the plot, then modify the model to correct for the pattern. The Exercises 7.19 and 7.20 illustrate this approach.

**Exercise 7.19.** Apply Model III to the blood pressure data at the beginning of Section 7.3.1. Examine a plot of the residuals. What does this plot

suggest about the use of Model III in this context versus the simpler model that was used originally?

**Exercise 7.20.** Plot the residuals for the models and data of Exercise 7.6. What do you observe?

The final two guidelines are contradictory in nature:

4. *The more parameters the better the fit.* Thus Model III is to be preferred to the two simpler models.
5. *The simpler, more straightforward model* is more likely to be correct when we come to apply it to data other than the observations in hand; thus Models I and II are to be preferred to Model III.

### 7.4.3. Measures of Predictive Success

The values we observe will be normally distributed about their expected values only if the deviations about the expected values are the sum of a large number of factors each of which only makes a small contribution to the total.

At the beginning of this section, we tried to build a model of systolic blood pressure purely as a function of age. The Quetlet index was lumped in with the “random” factors and, not surprisingly, made a disproportionate contribution to the total. Consequently, the  $p$  values and confidence limits of Table 7.1 are suspect.

In such a case, we can obtain more accurate confidence limits in two ways:

1. By blocking—deriving separate regression lines for males and females, smokers and nonsmokers, diabetics and nondiabetics.
2. By including additional factors such as the Quetlet index in our models.

Unfortunately, the latter approach is subject to diminishing returns. As we add more factors, we appear to be getting a better fit, but the result may be spurious, a purely chance effect. The fraudulent nature of our model will be revealed only later when we attempt to use it for prediction.

Two preventive measures can help you avoid this situation. First, you should always validate your model. Section 7.6 is devoted to this important topic.

Second, you should take advantage of one of the indexes that have been developed to let you know when you’ve begun to overfit a model.

Recall that in selecting among models, we used as one measure of goodness of fit  $SSE = \Sigma(y_i - y_i^*)^2$ , where  $y_i$  and  $y_i^*$  denote the  $i$ th observed

value and the corresponding value obtained from the model. The smaller this sum of squares, the better the fit.

If the observations are independent, then

$$\sum (y_i - y_i^*)^2 = \sum (y_i - \bar{y})^2 - \sum (\bar{y} - y_i^*)^2.$$

The first sum on the right-hand side of the equation is the total sum of squares (SST). Most statistics software uses as a measure of fit  $R^2 = 1 - \text{SSE}/\text{SST}$ . The closer the value of  $R^2$  is to 1, the better.

The automated entry of predictors into the regression equation using  $R^2$  runs the risk of overfitting, as  $R^2$  is guaranteed to increase with each predictor entering the model. To compensate, one may use the adjusted  $R^2$

$$1 - [((n - i)(1 - R^2))/(n - p)]$$

where  $n$  is the number of observations used in fitting the model,  $p$  is the number of regression coefficients in the model, and  $i$  is an indicator variable that is 1 if the model includes an intercept and 0 otherwise.

One rule of thumb is to continue to add variables as long as the value of  $R^2$ adj (the adjusted coefficient of determination) continues to increase.

Using the data of Exercise 7.9, we fit two models, one with systolic blood pressure solely a function of age, the other incorporating both age and the Quetlet index as predictors. Table 7.4 summarizes the results as abstracted from the Results worksheet of the XLStat regression procedure. Reading the bottom line of this table, we learn that the addition of a second predictor has increased  $R^2$ adj from 0.74 to 0.77.

**Exercise 7.21.** Does including the data on smoking habits from Exercise 7.9 increase the value of  $R^2$ adj?

#### 7.4.4. Multivariable Regression

We've already studied several examples in which we utilized multiple predictors in order to obtain improved models. Sometimes, as noted in

**TABLE 7.4 Goodness of fit coefficients**

	Model A	Model AQ
$R$ (coefficient of correlation)	0.874	0.900
$R^2$ (coefficient of determination)	0.763	0.811
$R^2$ adj (adjusted coefficient of determination)	0.742	0.773

Section 7.3.4, dependence among the random errors (as seen from a plot of the residuals) may force us to use additional variables. This result is in line with our discussion of experimental design in Chapter 5. We must control all sources of variation, must measure them, or must tolerate the “noise.”

But adding more variables to the model equation creates its own set of problems. Do predictors  $U$  and  $V$  influence  $Y$  in a strictly additive fashion so that we may write  $Y = \mu + \alpha U + \beta V + Z$ ? What if  $U$  represents the amount of fertilizer,  $V$  the total hours of sunlight, and  $Y$  the crop yield? If there are too many cloudy days, then adding fertilizer won't help a bit. The effects of fertilizer and sunlight are superadditive (or *synergistic*). A better model would be  $Y = \mu + \alpha U + \beta V + \gamma UV + Z$ .

To achieve predictive success, our observations should span the range over which we wish to make predictions. With only a single predictor, we might make just 10 observations spread across the predictor's range. With two synergistic or antagonist predictors we are forced to make  $10 \times 10$  observations, with three,  $10 \times 10 \times 10 = 1000$  observations, and so forth. We can cheat, scattering our observations at random or in some optimal systematic fashion across the grid of possibilities, but there will always be a doubt as to our model's behavior in the unexplored areas.

The vast majority of predictors are interdependent. Changes in the value of one will be accompanied by changes in the other. (Note that we do *not* write, “Changes in the value of one will cause or result in changes in the other.” There may be yet a third, hidden variable responsible for all the changes.) What this means is that more than one set of coefficients may provide an equally good fit to our data. And more than one set of predictors!

Exercise 7.22 illustrates that whether or not a given predictor will be found to make a statistically significant contribution to a model will depend upon what other predictors are present.

**Exercise 7.22.** To optimize an advertising campaign for a new model of automobile by directing the campaign toward the best potential customers, a study of consumers' attitudes, interests, and opinions was commissioned. The questionnaire consisted of a number of statements covering a variety of dimensions, including consumers' attitudes towards risk, foreign-made products, product styling, spending habits, emissions, pollution, self-image, and family. The final question concerned the potential customer's attitude toward purchasing the product itself. All responses were tabulated on a nine-point Likert scale. Utilize the data below to construct a series of models as follows:

Express Purchase as a function of Fashion and Gamble

Express Purchase as a function of Fashion, Gamble, and Ozone

Express Purchase as a function of Fashion, Gamble, Ozone, and  
Pollution

In each instance, determine the values of the coefficients, the associated  $p$ -values, and the value of Multiple  $R^2$  and Adjusted  $R^2$ . (As noted in the preface, for your convenience the following datasets may be downloaded from [ftp://ftp.wiley.com/public/sci\\_tech\\_med/statistics\\_resampling/](ftp://ftp.wiley.com/public/sci_tech_med/statistics_resampling/).)

Purchase 6, 9, 8, 3, 5, 1, 3, 3, 7, 4, 2, 8, 6, 1, 3, 6, 1, 9, 9, 7, 9, 2, 2, 8, 8, 5, 1, 3, 7, 9, 3, 6, 9, 8, 5, 4, 8, 9, 6, 2, 8, 5, 6, 5, 5, 3, 7, 6, 4, 5, 9, 2, 8, 2, 8, 7, 9, 4, 3, 3, 4, 1, 3, 6, 6, 5, 2, 4, 2, 8, 7, 7, 6, 1, 1, 9, 4, 4, 6, 9, 1, 6, 9, 6, 2, 8, 6, 3, 5, 3, 6, 8, 2, 5, 6, 7, 7, 5, 7, 6, 3, 5, 8, 8, 1, 9, 8, 8, 7, 5, 2, 2, 3, 8, 2, 2, 8, 9, 5, 6, 7, 4, 6, 5, 8, 4, 7, 8, 2, 1, 7, 9, 7, 5, 5, 9, 9, 9, 7, 3, 8, 9, 8, 4, 8, 5, 5, 8, 4, 3, 7, 1, 2, 1, 1, 7, 5, 5, 1, 4, 1, 2, 9, 7, 6, 9, 9, 6, 5, 4, 3, 6, 6, 4, 5, 7, 2, 6, 5, 6, 3, 8, 2, 5, 3, 4, 2, 3, 8, 3, 9, 1, 3, 1, 2, 5, 1, 5, 6, 7, 1, 1, 1, 4, 4, 8, 4, 7, 4, 4, 2, 6, 6, 6, 7, 2, 9, 4, 1, 9, 3, 5, 7, 2, 2, 8, 9, 2, 4, 1, 7, 1, 3, 6, 2, 6, 2, 8, 4, 4, 1, 1, 2, 2, 8, 3, 3, 3, 1, 1, 6, 8, 3, 7, 5, 9, 8, 3, 5, 6, 3, 4, 6, 1, 1, 5, 6, 6, 9, 6, 9, 9, 6, 7, 3, 8, 4, 2, 6, 4, 8, 3, 3, 6, 4, 4, 9, 5, 6, 4, 5, 3, 3, 2, 5, 9, 5, 1, 3, 4, 3, 6, 8, 1, 5, 3, 4, 8, 2, 5, 3, 2, 3, 2, 5, 4, 8, 3, 1, 6, 3, 7, 8, 9, 2, 3, 5, 7, 7, 3, 7, 3, 9, 2, 9, 3, 9, 2, 8, 9, 5, 1, 9, 9, 1, 8, 7, 1, 4, 9, 3, 4, 9, 1, 3, 9, 1, 5, 2, 7, 9, 6, 5, 7, 4, 6, 1, 4, 2, 7, 5, 4, 5, 9, 5, 5, 5, 2, 4, 1, 8, 7, 9, 6, 8, 1, 5, 9, 9, 9, 9, 1, 3, 3, 7, 2, 5, 6, 1, 5, 8

Fashion 5, 6, 8, 2, 5, 2, 5, 1, 7, 3, 5, 5, 4, 3, 3, 5, 6, 3, 4, 3, 4, 6, 4, 6, 3, 6, 5, 4, 6, 5, 5, 3, 4, 4, 4, 3, 6, 2, 3, 4, 4, 4, 5, 2, 3, 4, 5, 5, 6, 4, 5, 5, 6, 3, 4, 4, 5, 8, 4, 5, 6, 4, 2, 5, 3, 6, 2, 3, 2, 5, 3, 5, 4, 4, 5, 4, 6, 6, 5, 8, 2, 6, 5, 6, 4, 7, 4, 5, 5, 3, 6, 6, 4, 5, 5, 4, 4, 4, 4, 3, 5, 3, 3, 5, 4, 4, 5, 7, 6, 6, 4, 4, 5, 5, 2, 2, 7, 5, 1, 6, 5, 4, 7, 7, 6, 5, 6, 3, 2, 4, 5, 3, 9, 4, 4, 4, 6, 6, 9, 4, 4, 3, 3, 3, 2, 4, 4, 5, 4, 6, 6, 3, 3, 3, 5, 4, 4, 5, 4, 6, 3, 4, 6, 3, 4, 3, 1, 4, 5, 5, 6, 2, 6, 6, 5, 5, 3, 9, 3, 1, 1, 4, 3, 3, 3, 7, 6, 6, 4, 4, 1, 3, 5, 5, 4, 6, 4, 5, 4, 6, 5, 6, 2, 4, 4, 3, 8, 5, 3, 6, 5, 3, 5, 3, 3, 5, 3, 2, 2, 3, 5, 5, 5, 1, 6, 5, 1, 5, 4, 4, 3, 6, 4, 4, 5, 5, 4, 5, 5, 3, 7, 4, 7, 6, 1, 5, 4, 4, 4, 3, 3, 5, 4, 7, 4, 6, 7, 6, 4, 6, 3, 4, 4, 2, 6, 3, 6, 5, 2, 2, 5, 3, 4, 4, 4, 3, 2, 4, 6, 4, 6, 5, 6, 2, 4, 2, 3, 6, 2, 6, 5, 6, 4, 4, 4, 6, 5, 5, 1, 4, 5, 5, 4, 4, 2, 3, 6, 5, 5, 2, 2, 5, 2, 5, 4, 3, 8, 3, 6, 3, 4, 3, 6, 4, 3, 4, 2, 5, 6, 4, 5, 5, 6, 4, 6, 5, 4, 3, 8, 2, 5, 5, 3, 2, 3, 5, 4, 3, 4, 3, 5, 2, 3, 1, 4, 4, 6, 6, 6, 6, 6, 6, 4, 4, 3, 4, 4, 3, 3, 5, 4, 4, 5, 4, 6, 8, 3, 3, 5, 4, 5, 4, 5, 4, 4, 6, 6

Gamble 5, 4, 7, 4, 5, 4, 3, 3, 3, 6, 2, 6, 5, 4, 5, 5, 2, 7, 6, 6, 6, 4, 2, 8, 4, 4, 3, 3, 4, 5, 4, 3, 4, 6, 5, 4, 8, 9, 7, 3, 6, 4, 6, 6, 5, 3, 4, 6, 5, 4, 5, 3, 7, 3, 8, 5, 7, 5, 3, 4, 7, 4, 4, 5, 4, 6, 1, 4, 4, 9, 5, 4, 6, 4, 4, 5, 5, 5, 6, 6, 4, 4, 8, 7, 4, 5, 3, 3, 5, 3, 4, 5, 3, 5, 6, 6, 6, 5, 7, 4, 2, 3, 7, 6, 6, 4, 8, 4, 6, 3, 4, 4, 5, 8, 3, 3, 4, 5, 5, 5, 4, 5, 1, 6, 8, 5, 6, 4, 4, 6, 5, 7, 6, 5, 6, 7, 7, 6, 6, 4, 7, 6, 6, 5, 7, 6, 6, 5, 2, 5, 5, 4, 3, 3, 4, 6, 4, 4, 4, 3, 5, 2, 6, 4, 4, 6, 7, 6, 5, 4, 4, 7, 4, 7, 8, 5, 4, 5, 5, 3, 2, 4, 3, 6, 2, 3, 6, 5, 2, 4, 6, 2, 2, 1, 4, 3, 5, 5, 4, 6, 2, 3, 3, 5, 3, 6, 5, 6, 5, 3, 4, 6, 5, 5, 5, 3, 7, 7, 2, 6, 4, 2, 7, 2, 6, 3, 6, 2, 5, 3, 7, 3, 4, 2, 3, 7, 3, 6, 3, 7, 2, 4, 4, 4, 8, 3, 4, 4, 3, 1, 4, 7, 5, 4, 5, 8, 4, 6, 4, 6, 4, 4, 5, 4, 2, 6, 5, 5, 7, 2, 7, 4, 5, 6, 5, 3, 3, 2, 5, 3, 6, 1, 5, 6, 6, 5, 8, 6, 6, 5, 6, 4, 4, 6, 4, 7, 4, 4, 5, 4, 3, 7, 8, 1, 4, 4, 7, 4, 5, 4, 5, 1, 3, 4, 4, 4, 5, 3, 5, 5, 4, 7, 6, 3, 6, 4, 6, 5, 3, 4, 5, 7, 4, 5, 4, 5, 3, 7, 6, 4, 6, 4, 8, 3, 4, 2, 5, 5, 5, 4, 5, 6, 3, 5, 8, 4, 5, 2, 5, 4, 5, 6, 3, 3, 1, 5, 3, 4, 7, 4, 4, 6, 4, 3, 5, 3, 4, 4, 8, 6, 7, 4, 6, 4, 5, 4, 6, 8, 7, 2, 5, 4, 7, 4, 5, 6, 4, 6, 6



Ozone 4, 7, 7, 3, 4, 2, 3, 4, 2, 5, 4, 6, 2, 3, 6, 3, 7, 8, 7, 5, 5, 5, 8, 5, 5, 5, 1, 4, 6, 6, 9, 2, 6, 3, 4, 6, 6, 7, 5, 6, 6, 6, 4, 3, 5, 6, 5, 5, 3, 5, 4, 5, 6, 8, 3, 8, 5, 6, 4, 4, 3, 7, 8, 5, 3, 6, 6, 8, 6, 4, 5, 6, 4, 3, 6, 6, 3, 5, 5, 6, 7, 6, 6, 7, 3, 4, 5, 5, 6, 5, 3, 3, 5, 6, 3, 4, 6, 5, 5, 6, 6, 5, 9, 8, 5, 5, 4, 8, 4, 3, 5, 5, 4, 6, 8, 8, 4, 7, 5, 9, 2, 2, 5, 2, 7, 7, 2, 4, 4, 6, 3, 7, 7, 4, 3, 6, 3, 6, 6, 7, 3, 5, 5, 4, 3, 6, 4, 5, 6, 5, 5, 4, 7, 5, 5, 2, 4, 7, 5, 5, 5, 4, 6, 5, 5, 7, 5, 3, 6, 5, 6, 6, 4, 4, 2, 6, 6, 4, 8, 3, 5, 3, 3, 5, 5, 6, 5, 7, 4, 1, 3, 4, 6, 4, 3, 8, 5, 2, 7, 1, 5, 3, 7, 5, 4, 3, 7, 4, 2, 8, 7, 4, 3, 6, 7, 6, 6, 7, 9, 9, 3, 7, 6, 6, 4, 5, 6, 6, 4, 6, 5, 7, 5, 4, 6, 5, 6, 5, 5, 4, 4, 6, 9, 3, 3, 2, 5, 5, 5, 7, 3, 6, 4, 5, 7, 5, 4, 5, 5, 6, 6, 7, 4, 4, 4, 4, 2, 7, 4, 5, 4, 4, 5, 3, 6, 4, 7, 6, 4, 6, 5, 4, 5, 5, 4, 5, 7, 1, 3, 8, 6, 7, 5, 5, 5, 4, 5, 6, 5, 3, 5, 2, 3, 3, 4, 3, 3, 5, 5, 7, 7, 5, 6, 6, 6, 4, 7, 5, 7, 5, 8, 7, 7, 4, 5, 6, 6, 4, 9, 8, 5, 6, 6, 4, 4, 5, 4, 6, 3, 5, 4, 5, 8, 6, 6, 5, 3, 6, 7, 4, 7, 5, 4, 3, 6, 4, 6, 6, 4, 5, 5, 3, 7, 4, 6, 7, 3, 5, 6, 4, 9, 6, 3, 5, 7, 4, 5, 3, 7, 3, 3, 6, 6, 4, 6, 6, 6, 5, 5, 9, 4, 3, 6, 3, 4, 6

Pollution 5, 7, 7, 4, 5, 1, 3, 6, 3, 5, 5, 6, 2, 2, 7, 2, 6, 7, 7, 5, 5, 5, 6, 8, 6, 4, 5, 1, 4, 6, 6, 9, 3, 6, 3, 6, 4, 6, 8, 6, 6, 5, 6, 5, 3, 3, 8, 7, 7, 3, 4, 5, 5, 6, 8, 3, 8, 5, 6, 5, 4, 6, 5, 7, 7, 6, 4, 6, 5, 8, 5, 6, 6, 6, 4, 4, 5, 6, 5, 6, 5, 6, 8, 5, 5, 6, 5, 4, 5, 6, 5, 6, 3, 4, 6, 6, 5, 6, 6, 5, 4, 6, 8, 4, 9, 7, 6, 4, 5, 9, 4, 4, 4, 5, 4, 5, 7, 8, 3, 7, 7, 7, 2, 2, 5, 3, 5, 7, 4, 5, 5, 7, 5, 5, 6, 5, 4, 7, 4, 7, 7, 7, 4, 5, 5, 5, 3, 6, 5, 5, 7, 6, 4, 6, 7, 4, 6, 4, 4, 7, 6, 6, 7, 4, 6, 5, 6, 5, 6, 5, 4, 6, 6, 5, 6, 5, 6, 3, 6, 6, 5, 7, 4, 3, 3, 4, 6, 5, 5, 6, 6, 5, 2, 4, 4, 6, 2, 3, 6, 5, 4, 6, 2, 5, 2, 8, 4, 5, 4, 7, 5, 1, 7, 5, 6, 4, 5, 7, 7, 5, 6, 8, 8, 5, 7, 5, 5, 6, 6, 5, 6, 4, 7, 5, 6, 5, 4, 5, 5, 6, 5, 4, 5, 5, 6, 4, 6, 4, 8, 4, 2, 4, 5, 6, 6, 4, 5, 6, 4, 6, 4, 6, 6, 4, 5, 4, 7, 6, 7, 5, 5, 4, 2, 6, 3, 5, 3, 4, 5, 3, 5, 4, 7, 7, 4, 6, 5, 4, 6, 4, 6, 6, 1, 4, 7, 5, 8, 3, 6, 4, 4, 4, 6, 7, 6, 5, 3, 3, 5, 4, 4, 4, 6, 5, 7, 6, 4, 7, 6, 6, 4, 8, 4, 8, 6, 7, 8, 8, 4, 5, 4, 6, 5, 7, 7, 5, 6, 6, 5, 6, 6, 4, 6, 6, 5, 4, 7, 6, 6, 5, 6, 3, 4, 9, 5, 6, 5, 3, 5, 6, 4, 6, 4, 4, 6, 6, 3, 7, 5, 5, 7, 4, 6, 6, 4, 9, 6, 5, 6, 7, 4, 5, 3, 5, 4, 3, 7, 6, 6, 6, 6, 5, 6, 6, 9, 6, 3, 6, 3, 5, 7

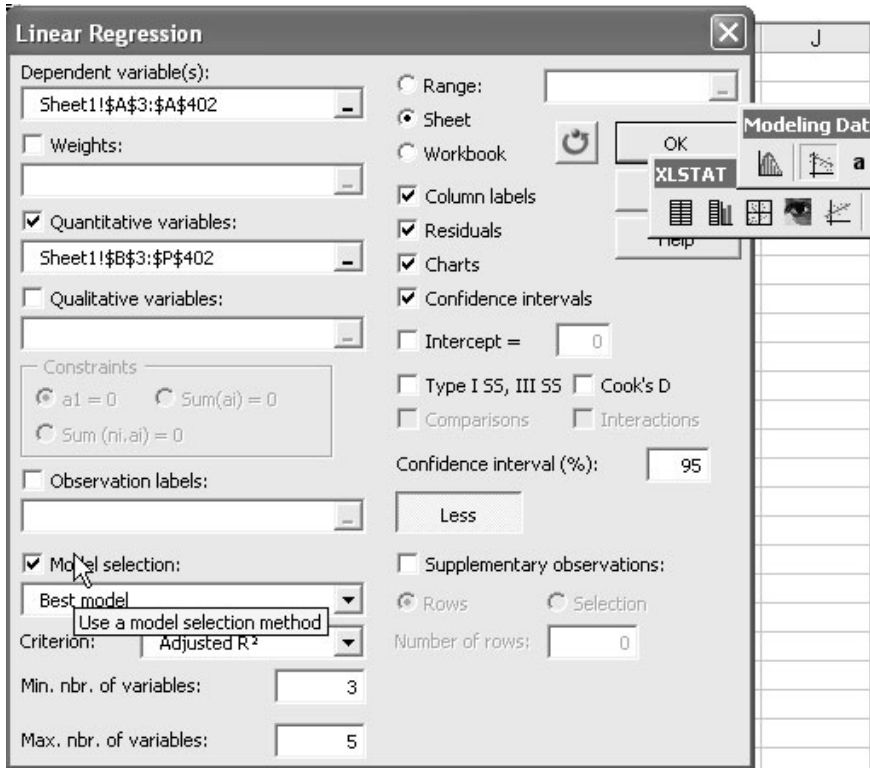
**Exercise 7.23.** The marketing study also included 11 more questions, the responses to which we’ve tabulated below, again using a nine-point Likert scale. Find the best prediction model for Attitude using the answers to all 15 questions.

Stop. This is insane. Fifteen questions leads to tens of thousand of possible models. There has to be a better way than simple trial and error to find the best one. There is. As always, it consists of letting the computer do the work. Figure 7.7 is a display of XLStat’s Linear Regression menu after clicking the “More” button and placing a check beside “Model Selection.”

Note that I’ve chosen “Best Model” as the method, “Adjusted R2” as the criterion, “3” as the minimum number of predictors, and “5” as the maximum. You may wish to experiment with other choices.

**Exercise 7.23 con’t.** Here are the data. Find the best-fitting model for predicting a prospective purchaser.

Today 4, 4, 4, 5, 4, 4, 5, 5, 6, 5, 6, 4, 2, 5, 6, 2, 5, 6, 6, 6, 7, 6, 5, 4, 6, 7, 2, 6, 2, 6, 4, 4, 5, 6, 4, 6, 5, 5, 4, 4, 2, 6, 3, 5, 5, 4, 4, 5, 6, 2, 4, 4, 4, 2, 2, 7, 4, 2, 5, 8, 7, 6, 5,



**FIGURE 7.7** Setting up an automated multiple regression.

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

Coupons 4, 2, 3, 6, 5, 5, 6, 3, 5, 5, 6, 4, 3, 4, 6, 4, 5, 6, 4, 5, 7, 5, 4, 5, 5, 6, 1, 5, 3,  
 7, 5, 5, 4, 5, 3, 6, 6, 5, 3, 4, 4, 7, 4, 6, 5, 5, 3, 4, 7, 1, 4, 5, 6, 3, 3, 6, 5, 3, 6, 5, 7, 7,  
 4, 5, 5, 4, 6, 3, 6, 4, 3, 3, 7, 2, 5, 5, 4, 4, 8, 4, 4, 5, 4, 4, 7, 5, 4, 3, 7, 5, 5, 4, 6, 6,  
 4, 6, 5, 6, 3, 4, 4, 7, 4, 4, 6, 6, 9, 4, 6, 2, 7, 7, 4, 4, 3, 4, 3, 3, 4, 5, 4, 4, 6, 4, 2, 4, 1,  
 2, 5, 2, 3, 5, 3, 3, 5, 5, 5, 6, 3, 3, 3, 4, 3, 3, 4, 6, 3, 5, 4, 3, 4, 5, 4, 4, 3, 5, 9, 3,  
 4, 6, 6, 6, 6, 4, 6, 5, 6, 3, 6, 5, 2, 5, 4, 5, 5, 4, 5, 2, 3, 5, 7, 4, 4, 6, 6, 3, 2, 3, 5, 8, 6,  
 5, 5, 5, 7, 5, 4, 5, 2, 4, 4, 2, 5, 3, 2, 2, 6, 7, 3, 5, 3, 4, 3, 5, 4, 5, 5, 4, 5, 4, 4, 6, 3, 3,  
 5, 7, 4, 6, 5, 3, 5, 5, 6, 4, 6, 3, 5, 2, 5, 3, 2, 4, 5, 5, 4, 4, 5, 4, 5, 5, 4, 3, 4, 5, 5, 4,

5, 5, 4, 4, 4, 4, 6, 6, 4, 5, 2, 4, 2, 3, 6, 8, 1, 5, 6, 5, 3, 5, 3, 6, 5, 5, 6, 4, 7, 7, 6, 3, 3, 3, 3, 5, 5, 5, 5, 6, 7, 4, 5, 3, 4, 3, 3, 6, 4, 3, 5, 8, 6, 5, 4, 8, 5, 7, 3, 5, 6, 5, 1, 7, 5, 6, 5, 1, 2, 5, 5, 3, 3, 5, 5, 6, 4, 5, 5, 7, 4, 5, 5, 3, 4, 4, 3, 4, 6, 5, 3, 6, 7, 5, 4, 2, 8, 2, 6, 5, 2, 5, 6, 5, 4, 4, 5, 5, 4, 2, 4, 3, 6, 6, 7, 5, 4, 5, 3, 4, 5, 4, 4, 6, 5, 7, 4, 7, 4, 5, 5, 7, 2, 6, 7, 5, 5, 3, 5, 4

IntRates 6, 1, 3, 6, 6, 2, 6, 3, 5, 5, 4, 3, 3, 4, 4, 6, 6, 5, 3, 6, 6, 3, 6, 5, 6, 2, 6, 6, 3, 6, 4, 4, 4, 5, 4, 6, 6, 6, 3, 5, 4, 7, 4, 7, 5, 7, 4, 5, 7, 2, 5, 6, 6, 4, 3, 7, 4, 3, 7, 7, 7, 6, 5, 6, 5, 4, 6, 4, 6, 4, 4, 7, 3, 5, 5, 3, 4, 9, 3, 5, 3, 5, 6, 4, 7, 5, 5, 4, 5, 6, 4, 5, 5, 7, 5, 6, 5, 7, 5, 4, 5, 4, 7, 5, 9, 4, 7, 3, 7, 6, 5, 5, 4, 3, 4, 3, 5, 3, 5, 5, 4, 3, 5, 4, 4, 4, 2, 3, 5, 3, 3, 4, 5, 6, 6, 4, 4, 4, 4, 5, 3, 4, 4, 6, 3, 5, 4, 4, 4, 5, 3, 5, 3, 6, 5, 9, 4, 5, 6, 6, 5, 6, 4, 7, 5, 7, 3, 6, 5, 2, 4, 4, 5, 5, 5, 2, 3, 5, 8, 4, 3, 6, 7, 4, 4, 5, 6, 7, 6, 6, 6, 6, 6, 4, 5, 2, 5, 3, 3, 6, 3, 2, 2, 7, 6, 4, 4, 3, 5, 4, 6, 2, 6, 5, 4, 7, 6, 3, 6, 4, 3, 5, 8, 5, 6, 5, 4, 6, 4, 5, 5, 7, 4, 4, 4, 7, 3, 2, 5, 5, 5, 5, 4, 5, 3, 4, 4, 5, 5, 3, 5, 6, 5, 5, 4, 7, 3, 3, 3, 4, 6, 5, 4, 6, 2, 4, 3, 3, 8, 9, 1, 7, 5, 6, 4, 4, 6, 7, 5, 5, 6, 4, 8, 7, 6, 5, 4, 4, 4, 4, 6, 5, 6, 8, 5, 4, 6, 3, 5, 3, 4, 7, 3, 2, 5, 8, 5, 5, 8, 5, 8, 3, 6, 7, 5, 3, 8, 6, 5, 5, 1, 3, 4, 5, 5, 3, 4, 6, 4, 5, 4, 5, 7, 5, 5, 5, 3, 4, 5, 3, 5, 7, 7, 4, 5, 6, 4, 5, 1, 6, 3, 6, 5, 3, 5, 5, 4, 5, 4, 4, 2, 6, 3, 6, 5, 6, 5, 4, 5, 2, 5, 5, 4, 3, 6, 4, 8, 3, 7, 4, 5, 4, 7, 1, 7, 7, 6, 5, 4, 6, 4

Selfconf 6, 7, 4, 6, 6, 6, 6, 3, 5, 6, 4, 1, 5, 5, 7, 3, 4, 6, 7, 5, 5, 4, 6, 4, 8, 5, 6, 5, 3, 5, 4, 4, 5, 8, 3, 5, 6, 3, 6, 5, 4, 7, 5, 2, 6, 5, 5, 6, 3, 5, 5, 4, 8, 6, 4, 7, 4, 5, 3, 4, 5, 3, 5, 7, 8, 6, 6, 3, 6, 6, 5, 3, 4, 5, 4, 6, 4, 6, 5, 5, 4, 5, 6, 7, 5, 6, 4, 6, 4, 3, 4, 2, 4, 1, 5, 5, 5, 6, 5, 5, 4, 4, 3, 7, 5, 5, 5, 5, 4, 3, 6, 3, 5, 3, 4, 7, 6, 5, 5, 4, 3, 5, 3, 6, 5, 4, 5, 2, 5, 5, 3, 6, 7, 5, 4, 6, 5, 6, 5, 3, 6, 5, 3, 6, 6, 5, 6, 6, 1, 4, 4, 9, 7, 5, 8, 7, 4, 3, 3, 6, 6, 8, 3, 7, 6, 4, 7, 4, 6, 6, 5, 6, 3, 5, 5, 3, 5, 6, 6, 6, 4, 4, 5, 4, 5, 5, 3, 6, 3, 6, 4, 4, 5, 3, 6, 6, 5, 1, 4, 6, 4, 6, 4, 3, 6, 6, 5, 3, 5, 6, 6, 5, 4, 6, 5, 3, 5, 5, 7, 4, 7, 3, 5, 2, 3, 2, 3, 4, 4, 5, 7, 3, 6, 6, 6, 7, 4, 3, 4, 4, 5, 2, 8, 5, 2, 5, 6, 7, 1, 4, 5, 2, 7, 6, 6, 3, 4, 2, 4, 6, 6, 3, 2, 6, 3, 4, 5, 5, 5, 4, 3, 6, 3, 4, 4, 5, 7, 7, 5, 3, 2, 6, 4, 1, 5, 4, 5, 5, 4, 5, 7, 3, 6, 6, 8, 2, 5, 6, 4, 5, 7, 3, 5, 6, 6, 4, 4, 6, 4, 4, 4, 7, 6, 4, 6, 3, 4, 5, 3, 4, 6, 6, 4, 5, 7, 6, 6, 4, 4, 4, 6, 5, 7, 4, 7, 4, 8, 6, 6, 6, 7, 4, 3, 7, 4, 3, 4, 4, 6, 6, 5, 5, 3, 3, 5, 4, 4, 5, 6, 4, 8, 2, 3, 3, 2, 6, 7, 2, 7, 7, 4, 6, 6, 5, 4, 3, 3, 7, 6, 5, 5, 1, 4, 5, 8, 5, 6, 3, 5, 8, 4

Leader 5, 6, 6, 6, 7, 7, 6, 5, 6, 7, 6, 3, 5, 6, 7, 4, 4, 7, 6, 6, 6, 5, 6, 5, 9, 5, 8, 6, 3, 5, 5, 6, 5, 9, 4, 4, 7, 5, 7, 6, 5, 6, 7, 3, 7, 6, 7, 7, 4, 8, 7, 6, 8, 6, 5, 7, 5, 5, 5, 5, 6, 4, 6, 8, 8, 5, 6, 6, 6, 7, 5, 4, 5, 6, 4, 6, 6, 6, 6, 6, 4, 6, 6, 8, 5, 6, 6, 5, 6, 5, 4, 3, 5, 3, 7, 5, 6, 6, 6, 6, 4, 5, 5, 7, 7, 6, 6, 6, 6, 5, 3, 8, 5, 6, 3, 5, 7, 6, 4, 6, 4, 4, 6, 4, 7, 5, 4, 7, 4, 6, 5, 5, 5, 8, 7, 6, 6, 6, 5, 6, 5, 3, 7, 4, 4, 6, 7, 5, 5, 6, 3, 5, 5, 9, 8, 8, 7, 6, 5, 3, 4, 8, 6, 9, 5, 6, 5, 6, 6, 9, 5, 6, 5, 3, 7, 7, 5, 5, 6, 7, 7, 5, 4, 6, 6, 7, 5, 6, 4, 8, 4, 5, 6, 3, 6, 7, 5, 1, 6, 5, 6, 7, 5, 4, 6, 7, 5, 3, 5, 6, 7, 6, 5, 7, 7, 5, 6, 5, 5, 7, 5, 7, 5, 7, 2, 6, 3, 5, 5, 6, 6, 7, 3, 9, 7, 6, 8, 5, 4, 5, 4, 5, 4, 9, 6, 4, 6, 5, 8, 3, 5, 4, 6, 6, 7, 4, 5, 4, 5, 7, 6, 3, 1, 7, 3, 5, 5, 6, 6, 5, 3, 7, 3, 5, 4, 5, 7, 8, 4, 3, 3, 8, 5, 2, 6, 5, 7, 5, 3, 6, 8, 3, 8, 5, 8, 2, 6, 6, 5, 6, 9, 3, 6, 7, 6, 5, 3, 5, 5, 6, 6, 5, 4, 7, 7, 4, 5, 4, 5, 5, 3, 7, 6, 6, 4, 7, 7, 6, 6, 6, 5, 4, 7, 5, 8, 4, 7, 6, 8, 8, 7, 6, 8, 5, 5, 7, 6, 4, 6, 6, 7, 8, 7, 5, 3, 4, 6, 7, 6, 4, 4, 7, 6, 8, 3, 4, 4, 4, 7, 9, 4, 8, 7, 4, 5, 7, 5, 5, 4, 5, 7, 8, 7, 6, 1, 4, 6, 7, 5, 7, 4, 7, 7, 6

Trip 4, 5, 6, 4, 5, 2, 3, 5, 4, 2, 3, 6, 4, 3, 2, 7, 1, 6, 9, 6, 7, 4, 1, 6, 3, 3, 4, 3, 4, 5, 5, 3, 6, 4, 4, 3, 4, 6, 7, 2, 4, 2, 3, 7, 4, 3, 4, 3, 4, 3, 5, 6, 2, 5, 1, 7, 2, 5, 4, 4, 2, 3, 3, 3, 5, 3, 4, 4, 4, 7, 5, 6, 5, 5, 2, 2, 7, 4, 2, 5, 6, 3, 4, 4, 4, 4, 6, 5, 3, 5, 3, 4, 4, 2, 3, 5, 4, 4, 2, 4, 3, 5, 5, 6, 4, 3, 4, 4, 5, 4, 6, 4, 3, 2, 5, 4, 5, 4, 6, 4, 5, 6, 5, 3, 4, 6, 4, 5, 5, 3, 3, 5, 6, 4, 2, 5, 5, 3, 6, 5, 5, 7, 3, 4, 4, 4, 4, 5, 5, 4, 5, 2, 3, 2, 4, 6, 4, 4, 2, 2, 2, 5, 6, 3, 4, 7, 6, 3, 5, 5, 3, 3, 4, 3, 5, 4, 4, 4, 4, 3, 7, 4, 4, 3, 3, 2, 4, 6, 4, 6, 1, 2, 1, 2, 4, 5, 5, 5, 3, 2, 1, 4, 4, 4, 3, 3, 5, 4, 7, 4, 8, 5, 4, 5, 4, 7, 6, 2, 6, 4, 6, 5, 4, 3, 6, 4, 3, 4,

4, 5, 4, 3, 4, 3, 3, 6, 6, 5, 5, 3, 4, 1, 3, 4, 2, 4, 4, 1, 2, 4, 5, 3, 4, 4, 5, 5, 4, 3, 5, 6, 5,  
 5, 3, 4, 3, 4, 6, 6, 6, 7, 6, 5, 3, 6, 5, 5, 5, 3, 5, 4, 3, 2, 4, 5, 3, 6, 3, 3, 4, 5, 4, 2, 2, 2,  
 5, 4, 4, 4, 6, 5, 7, 7, 2, 5, 2, 2, 3, 4, 5, 2, 2, 2, 1, 2, 6, 4, 2, 6, 3, 4, 5, 5, 3, 4, 5, 4, 6,  
 4, 1, 5, 6, 4, 2, 3, 2, 2, 6, 5, 4, 2, 7, 7, 4, 7, 7, 3, 3, 5, 5, 3, 4, 2, 4, 4, 4, 2, 3, 4, 5, 4,  
 4, 4, 5, 4, 5, 5, 3, 3, 5, 3, 5, 4, 3, 5, 5, 6, 5, 4, 6, 7, 5, 5, 8, 1, 4, 7, 5, 4, 7, 4, 2, 4, 2,  
 2, 6, 6, 1, 5, 2

Change 3, 5, 5, 4, 4, 2, 5, 6, 5, 3, 3, 6, 4, 4, 3, 8, 2, 7, 9, 5, 8, 3, 2, 6, 4, 3, 4, 4, 4,  
 5, 6, 3, 6, 4, 4, 4, 5, 7, 7, 3, 4, 5, 4, 7, 4, 3, 4, 4, 5, 6, 2, 4, 1, 6, 2, 4, 4, 4, 3, 4, 4,  
 4, 6, 3, 4, 4, 8, 6, 7, 5, 5, 4, 1, 3, 9, 4, 2, 4, 7, 3, 4, 4, 4, 5, 6, 6, 3, 5, 2, 4, 3, 3, 4, 5,  
 5, 4, 3, 5, 3, 4, 5, 6, 4, 4, 5, 5, 6, 4, 5, 4, 3, 2, 6, 5, 5, 5, 7, 5, 4, 6, 4, 4, 4, 6, 4, 5, 4,  
 4, 4, 6, 6, 4, 2, 5, 5, 3, 6, 6, 5, 5, 3, 4, 5, 5, 5, 6, 5, 5, 6, 3, 3, 2, 4, 6, 4, 4, 3, 3, 2,  
 4, 6, 3, 3, 7, 7, 3, 6, 6, 3, 3, 6, 3, 5, 4, 4, 6, 4, 3, 3, 6, 3, 4, 3, 3, 2, 5, 7, 5, 6, 1, 1, 2,  
 2, 4, 5, 5, 5, 3, 2, 1, 4, 5, 5, 4, 4, 6, 5, 7, 4, 8, 6, 3, 5, 3, 6, 6, 3, 7, 5, 7, 5, 5, 3, 7, 3,  
 4, 4, 4, 4, 4, 3, 4, 3, 7, 7, 5, 5, 3, 3, 1, 3, 4, 2, 4, 4, 1, 3, 4, 5, 4, 5, 5, 6, 5, 5, 4, 5,  
 7, 5, 5, 3, 5, 5, 6, 6, 5, 8, 6, 4, 2, 6, 5, 5, 5, 3, 6, 4, 3, 3, 5, 5, 3, 6, 4, 3, 5, 5, 3, 3,  
 2, 2, 5, 4, 4, 5, 7, 6, 7, 2, 5, 3, 3, 3, 4, 5, 2, 2, 3, 1, 1, 6, 4, 2, 6, 3, 4, 4, 6, 4, 4, 5,  
 3, 7, 4, 3, 5, 6, 4, 3, 2, 3, 2, 6, 7, 5, 3, 7, 7, 5, 7, 7, 5, 3, 5, 6, 4, 5, 3, 4, 5, 3, 2, 4, 4,  
 5, 4, 4, 5, 5, 5, 6, 3, 3, 5, 3, 5, 4, 3, 5, 5, 6, 5, 3, 6, 8, 6, 5, 7, 1, 4, 7, 6, 4, 6, 3, 2,  
 4, 2, 3, 6, 6, 1, 6, 4

Pioneer 3, 6, 5, 5, 4, 4, 7, 7, 5, 2, 6, 5, 5, 3, 3, 9, 3, 5, 7, 2, 6, 3, 3, 5, 5, 4, 5, 6, 2,  
 4, 6, 2, 6, 3, 6, 5, 4, 6, 5, 3, 4, 5, 5, 7, 4, 2, 4, 4, 7, 5, 6, 3, 4, 2, 6, 6, 4, 3, 4, 2, 5, 4,  
 6, 5, 5, 4, 4, 7, 5, 7, 6, 6, 5, 1, 4, 7, 4, 2, 7, 6, 5, 7, 5, 3, 4, 5, 5, 2, 3, 4, 6, 4, 4, 5, 5,  
 6, 4, 5, 6, 4, 5, 4, 7, 4, 5, 6, 4, 7, 3, 7, 4, 4, 4, 6, 5, 5, 4, 6, 8, 5, 5, 4, 3, 6, 7, 5, 4, 5,  
 4, 6, 6, 4, 6, 5, 6, 8, 3, 4, 6, 6, 5, 3, 5, 4, 6, 5, 4, 3, 5, 6, 6, 5, 6, 4, 4, 6, 5, 3, 3, 3, 4,  
 6, 4, 4, 4, 6, 6, 4, 4, 4, 5, 3, 7, 2, 3, 5, 6, 4, 4, 6, 4, 5, 4, 5, 3, 6, 3, 7, 5, 7, 7, 5, 3, 3,  
 3, 4, 4, 4, 4, 5, 4, 2, 5, 4, 6, 5, 6, 3, 6, 7, 4, 8, 7, 4, 6, 5, 3, 5, 3, 4, 4, 7, 6, 5, 5, 5, 3,  
 4, 3, 4, 5, 5, 3, 5, 3, 3, 7, 7, 6, 6, 5, 4, 1, 4, 5, 3, 5, 4, 5, 4, 5, 6, 5, 5, 5, 5, 6, 4, 6, 4,  
 6, 7, 7, 4, 6, 3, 5, 5, 5, 7, 6, 3, 5, 4, 5, 5, 5, 4, 5, 6, 6, 2, 4, 5, 6, 4, 6, 4, 4, 8, 4, 6, 3,  
 5, 2, 8, 6, 4, 4, 4, 5, 6, 4, 5, 1, 3, 6, 3, 6, 3, 4, 4, 5, 3, 5, 6, 4, 6, 3, 6, 5, 6, 3, 5, 7,  
 4, 7, 4, 4, 5, 4, 4, 4, 1, 1, 5, 6, 5, 3, 5, 6, 6, 7, 7, 5, 4, 4, 5, 7, 6, 4, 4, 7, 3, 3, 4, 3,  
 5, 4, 4, 5, 4, 5, 4, 6, 2, 3, 4, 4, 5, 3, 4, 5, 4, 4, 4, 4, 6, 7, 5, 6, 5, 5, 5, 5, 5, 6, 6, 5, 5,  
 4, 5, 4, 9, 8, 2, 7, 5

Work 3, 4, 5, 5, 7, 7, 5, 4, 5, 7, 5, 1, 5, 8, 2, 7, 4, 7, 3, 6, 6, 6, 5, 5, 6, 4, 5, 5, 5, 4,  
 5, 3, 8, 7, 4, 4, 7, 6, 8, 4, 4, 5, 5, 6, 3, 6, 6, 5, 7, 9, 4, 5, 6, 4, 3, 6, 8, 3, 5, 8, 5, 5, 7,  
 5, 6, 3, 5, 6, 5, 6, 6, 5, 8, 5, 6, 5, 5, 6, 7, 5, 5, 3, 7, 5, 7, 6, 4, 6, 4, 1, 7, 3, 6, 5, 7, 5,  
 4, 6, 5, 5, 4, 6, 6, 6, 7, 5, 5, 6, 3, 7, 3, 7, 5, 6, 7, 6, 3, 8, 6, 5, 7, 6, 7, 7, 6, 3, 8, 5, 3,  
 8, 7, 6, 7, 6, 8, 9, 4, 6, 4, 7, 7, 3, 7, 6, 4, 5, 5, 4, 7, 7, 9, 6, 7, 5, 5, 6, 7, 6, 7, 7, 5, 5,  
 5, 4, 8, 7, 7, 6, 4, 4, 6, 5, 7, 4, 8, 4, 5, 5, 9, 4, 3, 5, 3, 6, 4, 7, 5, 6, 4, 6, 9, 4, 4, 7, 6,  
 7, 8, 4, 5, 5, 5, 4, 5, 5, 7, 5, 6, 4, 4, 6, 5, 4, 5, 4, 7, 3, 4, 9, 5, 5, 4, 6, 6, 7, 4, 5, 5, 4,  
 5, 4, 5, 7, 3, 6, 5, 5, 7, 7, 3, 4, 3, 6, 5, 3, 1, 3, 6, 5, 7, 4, 7, 6, 6, 5, 8, 7, 6, 4, 7, 5, 6,  
 5, 4, 6, 4, 5, 5, 5, 6, 8, 6, 6, 8, 1, 2, 5, 7, 2, 4, 6, 8, 5, 4, 3, 6, 5, 6, 2, 6, 5, 8, 7, 4, 5,  
 6, 9, 4, 4, 5, 6, 3, 5, 6, 5, 1, 5, 5, 9, 8, 5, 5, 6, 4, 7, 4, 6, 5, 4, 3, 4, 8, 7, 9, 4, 3, 8, 8,  
 6, 7, 3, 5, 5, 5, 4, 6, 3, 3, 4, 5, 7, 8, 8, 4, 6, 7, 5, 4, 5, 5, 3, 5, 6, 6, 4, 6, 1, 6, 5, 4, 6,  
 6, 7, 4, 6, 5, 5, 8, 6, 5, 3, 9, 4, 6, 3, 5, 5, 6, 5, 9, 7, 6, 4, 7, 8, 4, 5, 7, 5, 5, 3, 8, 6, 5,  
 6, 7, 5, 8, 4, 4, 5

Mind 1, 4, 6, 3, 3, 5, 4, 4, 6, 3, 1, 4, 6, 6, 5, 5, 2, 6, 3, 1, 6, 2, 5, 4, 6, 6, 6, 4, 3, 6,  
 2, 6, 6, 4, 4, 5, 9, 6, 6, 2, 4, 6, 7, 5, 3, 3, 6, 7, 7, 8, 6, 4, 5, 2, 3, 5, 5, 2, 4, 5, 3, 4, 6,  
 5, 5, 5, 4, 6, 4, 7, 5, 4, 5, 3, 5, 7, 3, 5, 6, 3, 3, 4, 8, 5, 1, 5, 2, 2, 3, 5, 6, 6, 3, 7, 3, 6,  
 5, 4, 4, 3, 4, 8, 5, 2, 5, 6, 5, 6, 4, 5, 3, 3, 2, 5, 4, 4, 4, 6, 3, 5, 2, 6, 7, 6, 4, 4, 5, 5, 4,  
 5, 5, 6, 2, 4, 7, 7, 5, 5, 6, 5, 6, 6, 3, 4, 6, 4, 4, 6, 3, 5, 3, 3, 4, 6, 4, 2, 3, 2, 4, 3, 3, 6,  
 2, 7, 5, 6, 4, 2, 3, 5, 4, 3, 6, 4, 5, 9, 4, 3, 4, 5, 5, 7, 3, 4, 6, 5, 5, 4, 5, 2, 5, 2, 5, 3, 5,

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

UPM 4, 4, 7, 3, 4, 5, 6, 5, 6, 3, 2, 3, 6, 7, 4, 6, 2, 7, 4, 2, 7, 2, 4, 4, 7, 6, 6, 4, 3, 5, 3, 5, 5, 3, 5, 5, 9, 5, 6, 3, 3, 5, 6, 5, 4, 4, 5, 7, 7, 7, 5, 4, 3, 5, 5, 5, 4, 4, 6, 4, 4, 7, 5, 7, 5, 5, 6, 4, 6, 5, 5, 5, 3, 5, 6, 4, 4, 6, 5, 4, 5, 7, 6, 3, 3, 2, 3, 5, 6, 6, 3, 6, 3, 6, 5, 5, 6, 5, 4, 7, 5, 3, 5, 6, 4, 5, 4, 4, 3, 5, 3, 6, 5, 3, 5, 6, 3, 4, 2, 7, 7, 5, 4, 4, 6, 4, 5, 6, 4, 6, 3, 3, 6, 6, 5, 6, 6, 4, 5, 6, 3, 4, 6, 5, 4, 5, 3, 5, 3, 4, 4, 5, 3, 3, 4, 7, 5, 3, 4, 1, 7, 5, 5, 4, 2, 4, 4, 3, 2, 5, 4, 5, 8, 5, 4, 4, 6, 5, 5, 3, 6, 5, 4, 5, 5, 5, 3, 4, 4, 7, 5, 6, 5, 5, 3, 5, 6, 3, 5, 3, 6, 5, 5, 2, 5, 3, 5, 5, 4, 6, 4, 2, 7, 3, 5, 7, 7, 3, 2, 5, 3, 4, 8, 4, 5, 2, 5, 5, 4, 6, 4, 3, 5, 4, 2, 2, 4, 5, 3, 4, 7, 4, 4, 6, 4, 7, 4, 7, 4, 5, 1, 4, 6, 7, 3, 3, 5, 3, 7, 5, 7, 6, 4, 5, 4, 4, 5, 7, 5, 5, 8, 4, 5, 3, 5, 3, 7, 4, 5, 5, 7, 5, 3, 5, 4, 5, 8, 3, 4, 4, 7, 4, 3, 5, 6, 3, 4, 7, 4, 3, 5, 4, 4, 3, 5, 6, 5, 3, 1, 2, 5, 4, 6, 5, 3, 4, 6, 5, 2, 5, 4, 2, 2, 6, 5, 5, 6, 6, 3, 5, 7, 9, 5, 4, 6, 5, 4, 3, 7, 4, 4, 5, 3, 6, 7, 3, 2, 4, 4, 4, 4, 5, 5, 2, 5, 3, 4, 4, 6, 3, 4, 5, 4, 4, 8, 5, 4, 4, 4, 4, 4, 7, 4, 4, 2, 4, 6, 6, 5, 8, 2, 5, 5, 1, 5, 4, 2, 6, 6, 9

### 7.5. QUANTILE REGRESSION

Linear regression techniques are designed to help us predict expected values, as in  $E(Y) = \mu + \beta X$ . But what if our real interest is in predicting extreme values, *if*, for example, we would like to characterize the observations of  $Y$  that are likely to lie in the upper and lower tails of  $Y$ 's distribution.

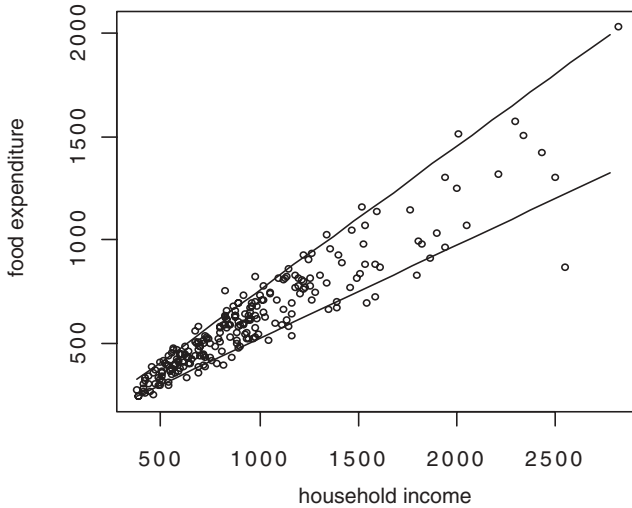
Even when expected values or medians lie along a straight line, other quantiles may follow a curved path. Koenker and Hallock applied the method of quantile regression to data taken from Ernst Engel's study in 1857 of the dependence of households' food expenditure on household income. As Fig. 7.8 reveals, not only was an increase in food expenditures observed as expected when household income was increased, but the dispersion of the expenditures increased also.

In estimating the  $\tau$ th quantile<sup>1</sup>, we try to find that value of  $\beta$  for which  $\sum_i \rho_\tau(y_k - f[x_k, \beta])$  is a minimum, where

$$\begin{aligned} \rho_\tau[x] &= \tau x && \text{if } x > 0 \\ &= (\tau - 1)x && \text{if } x \leq 0 \end{aligned}$$

Unfortunately, as with LAD regression, quantile regression is not readily executed within the Excel framework.

<sup>1</sup>  $\tau$  is pronounced tau.



**FIGURE 7.8** Engel data with quantile regression lines superimposed.

## 7.6. VALIDATION

As noted in the preceding sections, more than one model can provide a satisfactory fit to a given set of observations; even then, goodness of fit is no guarantee of predictive success. Before putting the models we develop to practical use, we need to *validate* them. There are three main approaches to validation:

1. Independent verification (obtained by waiting until the future arrives or through the use of surrogate variables)
2. Splitting the sample (using one part for calibration, the other for verification)
3. Resampling (taking repeated samples from the original sample and refitting the model each time)

In what follows, we examine each of these methods in turn.

### 7.6.1. Independent Verification

Independent verification is appropriate and preferable whatever the objectives of your model. In geologic and economic studies, researchers often return to the original setting and take samples from points that have been bypassed on the original round. In other studies, verification of the model's form and the choice of variables is obtained by attempting to fit the same model in a similar but distinct context.

For example, having successfully predicted an epidemic at one army base, one would then wish to see whether a similar model might be applied at a second and a third almost-but-not-quite identical base.

Independent verification can help discriminate among several models that appear to provide equally good fits to the data. Independent verification can be used in conjunction with either of the two other validation methods. For example, an automobile manufacturer was trying to forecast parts sales. After correcting for seasonal effects and long-term growth within each region, ARIMA techniques were used.<sup>2</sup> A series of best-fitting ARIMA models was derived: one model for each of the nine sales regions into which the sales territory had been divided. The nine models were quite different in nature. As the regional seasonal effects and long-term growth trends had been removed, a single ARIMA model applicable to all regions, albeit with coefficients that depended on the region, was more plausible. The model selected for this purpose was the one that gave the best fit when applied to all regions.

Independent verification also can be obtained through the use of surrogate or proxy variables. For example, we may want to investigate past climates and test a model of the evolution of a regional or worldwide climate over time. We cannot go back directly to a period before direct measurements on temperature and rainfall were made, but we can observe the width of growth rings in long-lived trees or measure the amount of carbon dioxide in ice cores.

### 7.6.2. Splitting the Sample

For validating time series, an obvious extension of the methods described in the preceding section is to hold back the most recent data points, fit the model to the balance of the data, and then attempt to “predict” the values held in reserve.

When time is not a factor, we still would want to split the sample into two parts, one for estimating the model parameters and the other for verification. The split should be made at random. The downside is that when we use only a portion of the sample, the resulting estimates are less precise.

In Exercises 7.24–7.26, we want you to adopt a compromise proposed by Moiser. Begin by splitting the original sample in half; choose your regression variables and coefficients independently for each of the

---

<sup>2</sup> For examples and discussion of AutoRegressive Integrated Moving Average processes used to analyze data whose values change with time.

subsamples. If the results are more or less in agreement, then combine the two samples and recalculate the coefficients with greater precision.

There are several different ways to arrange for the division. Here is one way:

- Suppose we have 100 triples of observations in columns 1 through 4. We start a 4th column as we did in Chapter 1 for an audit, insert the formula = Rand() in the top cell, and copy it down the column. Whenever a value greater than 0.500 appears, the observation will be included in the training set.

**Exercise 7.24.** Apply Moiser's method to the Milazzo data of Exercise 7.12. Can total coliform levels be predicted on the basis of month, oxygen level, and temperature?

TotColi 30, 22, 16, 18, 32, 40, 50, 34, 32, 32, 34, 18, 16, 19, 65, 54, 32, 59, 45, 27, 88, 32, 78, 45, 68, 14, 54, 22, 25, 32, 22, 17, 87, 17, 46, 23, 10, 19, 38, 22, 12, 26, 8, 8, 11, 19, 45, 78, 6, 9, 87, 6, 23, 28, 0, 0, 43, 8, 23, 19, 0, 5, 28, 19, 14, 32, 12, 17, 33, 21, 18, 5, 22, 13, 19, 27, 30, 28, 16, 6, 21, 27, 58, 45

**Exercise 7.25.** Apply Moiser's method to the data provided in Exercises 7.22 and 7.23 to obtain prediction equation(s) for Attitude in terms of some subset of the remaining variables.

**Note:** As conditions and relationships do change over time, any method of prediction should be *revalidated* frequently. For example, suppose we had used observations from January 2000 to January 2004 to construct our original model and held back more recent data from January to June 2004 to validate it. When we reach January 2005, we might refit the model, using the data from 1/2000 to 6/2004 to select the variables and determine the values of the coefficients, then use the data from 6/2004 to 1/2005 to validate the revised model.

**Exercise 7.26.** Some authorities would suggest discarding the earliest observations before refitting the model. In the present example, this would mean discarding all the data from the first half of the year 2000. Discuss the possible advantages and disadvantages of discarding these data.

### 7.6.3. Cross-Validation with the Bootstrap

Recall that the purpose of bootstrapping is to simulate the taking of repeated samples from the original population (and to save money and time by not having to repeat the entire sampling procedure from scratch). By bootstrapping, we are able to judge to a limited extent whether the



models we derive will be useful for predictive purposes or whether they will fail to carry over from sample to sample. As Exercise 7.27 demonstrates, some variables may prove more reliable as predictors than others.

**Exercise 7.27.** Bootstrap repeatedly from the data provided in Exercises 7.22 and 7.23 and use the XLSTAT stepwise function to select the variables to be incorporated in the model each time. Are some variables common to all the models?

## 7.7. CLASSIFICATION AND REGRESSION TREES

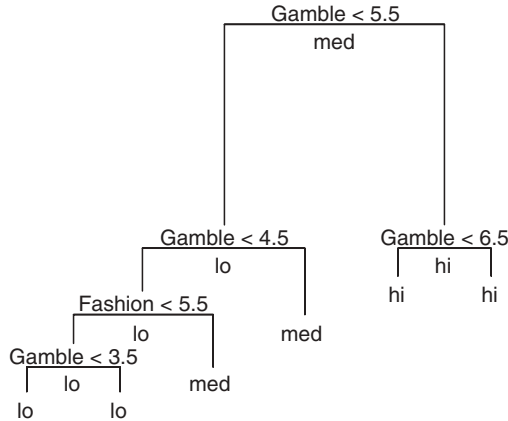
As the number of potential predictors increases, the method of linear regression becomes less and less practical. With three potential predictors, we can have as many as seven coefficients to be estimated: one for the intercept, three for first-order terms in the predictors  $P_i$ , two for second-order terms of the form  $P_i P_j$ , and one third-order term  $P_1 P_2 P_3$ . With  $k$  variables, we have  $k$  first-order terms,  $k(k-1)$  second-order terms, and so forth. Should all these terms be included in our model? Which ones should be neglected? With so many possible combinations, will a single equation be sufficient?

We need to consider alternate approaches. If you're a mycologist, a botanist, a herpetologist, or simply a nature lover you may have made use of some sort of a key. For example,

1. Leaves simple?
  3. Leaves needle-shaped?
    - a. Leaves in clusters of 2 to many?
      - i. Leaves in clusters of 2 to 5, sheathed, persistent for several years?

To derive the decision tree depicted in Fig. 7.9, we began by grouping our prospects' attitudes into categories using the data from Exercise 7.22. Purchase attitudes of 1, 2, or 3 indicate low interest, 4, 5, and 6 indicate medium interest, and 7, 8, and 9 indicate high interest. For example, if the original purchase data were in column L, we might categorize the first entry in an adjacent column via the command = IF(L3 < 4, 1, IF(L3 < 7, 2, 3)), which we then would copy down the column.

As in Exercise 7.22, the object was to express Purchase as a function of Fashion, Gamble, and Ozone. The computer considered each of the variables in turn, looking to find both the variable and the associated value that would be most effective in subdividing the data. Eventually, it settled



**FIGURE 7.9** Labeled classification tree for predicting likelihood of purchase.

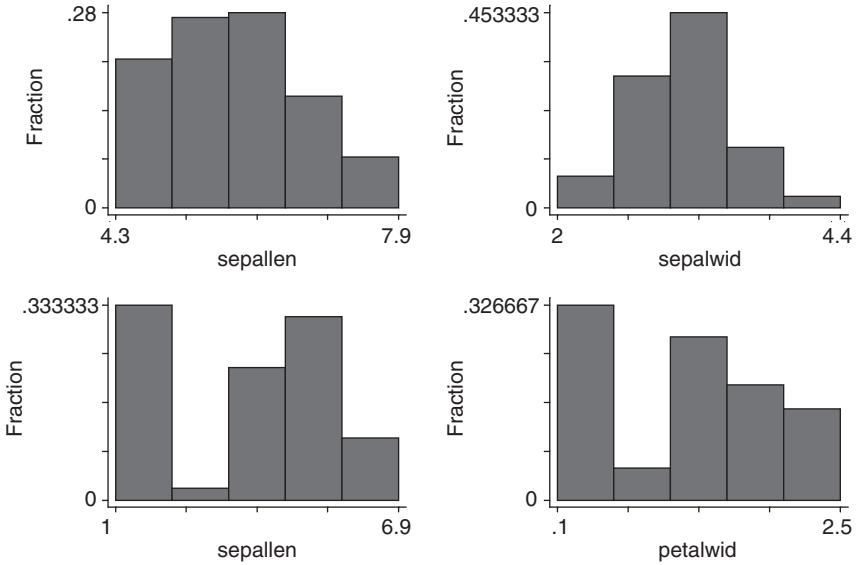
on “Is Gamble < 5.5” as the most effective splitter. This question divides the training data set into two groups, one containing all the most likely prospects.

The computer then proceeded to look for a second splitter that would separate the “lo” prospects from the medium. Again, “Gamble” proved to be the most useful, and so on.

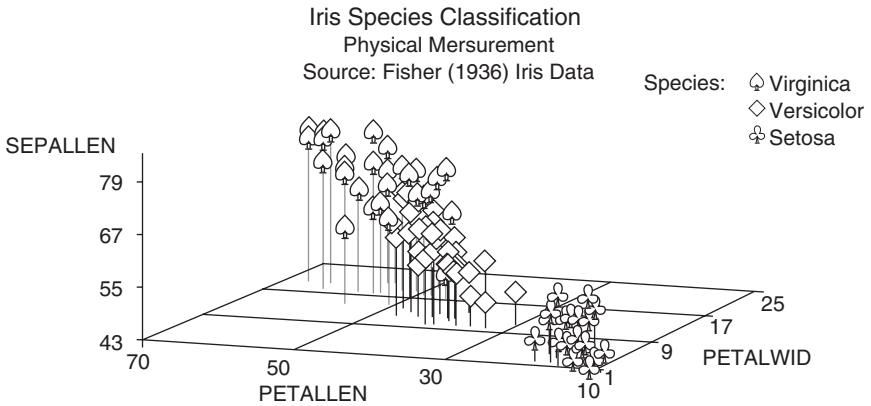
Obviously, building a decision tree is not something you would want to attempt in the absence of a computer and the appropriate software. Fortunately, you can download Ctree, a macro-filled Excel spreadsheet, from <http://www.geocities.com/adotsaha/CTree/CtreeinExcel.html>.

The first of the seven worksheets in the CTree package, labeled “ReadMe,” contains detailed instructions for the use of the remaining worksheets. Initially, the Ctree “Data” worksheet contains the sepal length, sepal width, petal length, and petal width of 150 irises. The attempt at classification of the iris into three separate species on the basis of these measurements dates back to 1935. Our own first clues to the number of subpopulations or categories of iris, as well as to the general shape of the underlying frequency distribution, come from consideration of the histogram in Fig. 7.10. A glance suggests the presence of at least two species, although because of the overlap of the various subpopulations it is difficult to be sure. Three species actually are present as shown in Fig. 7.11.

In constructing the decision tree depicted in part in Fig. 7.12, we made two modifications to the default settings in the Ctree spreadsheet. First, on the Data sheet, we included sepal length and sepal width as



**FIGURE 7.10** Sepal and petal measurements of 150 iris plants.



Petalen: petal length in mm.  
Sepallen: sepal length in mm.

Fetalwid: petal width in mm.  
Sepalwid not shown.

D0335 UC

**FIGURE 7.11** Representing three variables in two dimensions. Iris species. Derived with the help of SAS/Graph®.

Node 1 Petal_width<10						setosa
Node 2 Petal_width>=10						
<b>TRUE</b>						
	Node 3 Petal_width<18					
	<b>TRUE</b>					
		Node 5 Petal_length<50				
		<b>TRUE</b>				
			Node 7 Petal_width<17			versicolor
			<b>TRUE</b>			
				Node 8 Petal_width>=17		virginica
		Node 6 Petal_length>=50				
				Node 9 Petal_length<55		
					Node 11 Petal_width<16	virginica

**FIGURE 7.12** Part of decision tree for iris classification.

explanatory variables, changing the settings in row21 from “omit” to “cont.” Surprisingly, this change did not affect the resulting decision tree, which still made use of only petal length and petal width as classifiers.

Second, on the UserInput sheet, we selected option 1 for partitioning into training and test sets. Option 2 is appropriate only with time series data.

Note in Fig. 7.12 that the *setosa* species is classified on the basis of a single value, while distinguishing *versicolor* and *virginica* subspecies is far more complicated.

How successful is our decision tree? On the Result worksheet, we learn (Table 7.5) that our decision tree correctly classified iris species in 126 out of 128 instances. Amazing? Not really, considering that our tree-building program was given the correct classification of these species to begin with. The true test of the method comes when we attempt to classify without such knowledge.

**TABLE 7.5 Classification of Training Data**

True Class	Predicted Class			
	Setosa	Verginica	Versicolor	
Setosa	41			41
Verginica		43	2	45
Versicolor			42	42
	41	43	44	128

**TABLE 7.6 Classification of Test Data**

True Class	Predicted Class			
	Setosa	Verginica	Versicolor	
Setosa	9			9
Verginica		5		5
Versicolor		1	7	8
	9	6	7	22

As noted earlier, we set aside 20% of the flowers at random to test our classification scheme on. The test results in Table 7.6 abstracted from the Results worksheet are reassuring.

**Exercise 7.28.** Show that the decision tree method only makes use of variables it considers important by constructing a tree for classifying prospective purchasers into hi, med, and lo categories using the model

$$\text{Purchase} \sim \text{Fashion} + \text{Ozone} + \text{Pollution} + \text{Coupons} + \\ \text{Gamble} + \text{Today} + \text{Work} + \text{UPM}$$

**Exercise 7.29.** Apply the CART method to the Milazzo data of Exercise 7.12 to develop a prediction scheme for coliform levels in bathing water based on the month, oxygen level, and temperature.

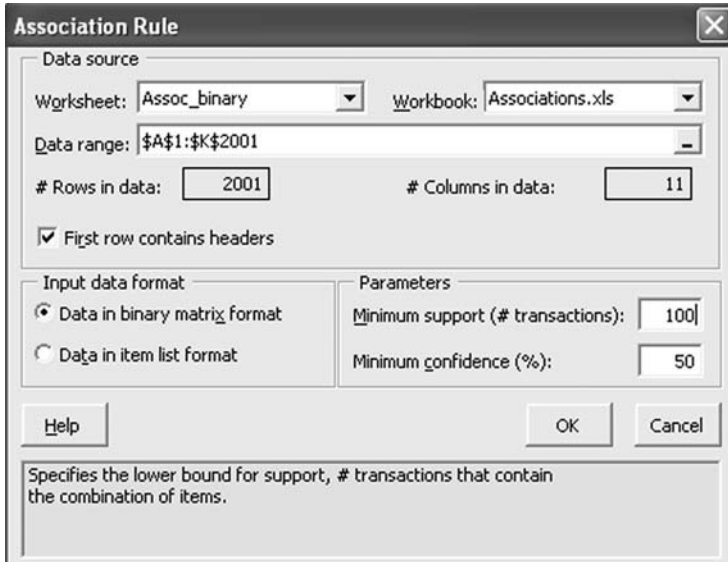
## 7.8. DATA MINING

When data sets are very large with hundreds of rows and dozens of columns, different algorithms come into play. In Section 2.2.1, we considered the possibility of a market basket analysis, when a retail outlet would wish to analyze the pattern of its sales to see what items might be profitably grouped and marketed together.

Table 7.7 depicts part of just such a data set. Each column corresponds to a different type of book and each row corresponds to a single transaction. The complete data set contains 2000 transactions.

TABLE 7.7 Bookstore Data for Use in a Market Basket Analysis

	ChildBks	YouthBks	CookBks	DoltYBks	RefBks	ArtBks	GeogBks	ItalCook	ItalAtlas	ItalArt	Florence
0	1	0	0	1	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	1	0	1	0	0	0	0
0	0	1	1	0	0	0	1	0	0	0	0
1	0	0	0	0	0	1	0	0	0	0	1
0	1	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	1	0	0	0	0	0	0
1	0	0	0	1	0	0	0	0	0	0	0



**FIGURE 7.13** Preparing to do a market basket analysis.

After downloading and installing Xlminer, an Excel add-in, from <http://www.resample.com/xlminer>, select “Affinity” from the XlMiner menu, and then “Association rules.” The Association Rules dialog box will appear as shown in Fig. 7.13.

Completing the dialog as shown, click on the OK button to see the results displayed in Fig. 7.14. Note that each rule is presented along with an estimated confidence level, support and lift ratio.

Rule #1 says that if an Italian cookbook and a Youthbook are bought, a cookbook will also be bought. This particular rule has confidence of 100%, meaning that, of the people who bought an Italian cookbook and a Youthbook, all (100%) bought cookbooks as well. “Support (a)” indicates that it has the support of 118 transactions, meaning that 118 people bought an Italian cookbook and a Youthbook, total. “Support (c)” indicates the number of transactions involving the purchase of cookbooks, total. (This is a piece of side information—it is not involved in calculating the confidence or support for the rule itself.) “Support (a U c)” is the number of transactions where an Italian cookbook and a Youthbook as well as a cookbook were bought.

Lift ratio indicates how much more likely one is to encounter a cookbook transaction if just those transactions where an Italian cookbook and a Youthbook is purchased are considered, as compared to the entire popu-

XLMiner : Association Rules

Data	
Input Data	Assoc_binary\$A\$1:\$K\$2001
Data Format	Binary Matrix
Minimum Support	100
Minimum Confidence %	50
# Rules	260
Overall Time (secs)	5

Rule 1: If item(s) ItalCook= is / are purchased, then this implies item(s) CookBks, YouthBks is / are also purchased. This rule has confidence of 51.98%.
---

Rule #	Conf. %	Antecedent (a)	Consequent (c)	Support(a)	Support(c)
1	51.98	<b>ItalCook=&gt;</b>	<b>CookBks, YouthBks</b>	227	324
2	61.76	ChildBks, ItalCook=>	CookBks, GeogBks	170	385
3	82.03	GeogBks, ItalCook=>	ChildBks, CookBks	128	512
4	51.2	ChildBks, DoltYBks, GeogBks=>	ArtBks, CookBks	209	334
5	50.24	DoltYBks, RefBks=>	ChildBks, YouthBks	211	330
6	54.92	RefBks, YouthBks=>	ChildBks, DoltYBks	193	368
7	75	DoltYBks, GeogBks, YouthBks=>	ChildBks, CookBks	136	512
8	56.39	ItalCook=>	CookBks, GeogBks	227	385
9	56.32	ArtBks, ChildBks, DoltYBks=>	CookBks, GeogBks	190	385
10	74.89	ItalCook=>	ChildBks, CookBks	227	512

**FIGURE 7.14** Results of a market basket analysis of the bookstore data.

lation of transactions—it's the confidence divided by support (c) where the latter is expressed as a percentage. For Rule#1, the confidence is 100%. support (c) (in percentage) =  $(862/2000) * 100 = 43.1$ . So the lift ratio =  $100/43.1 = 2.32$ .

## 7.9. SUMMARY AND REVIEW

In this chapter, you were introduced to two techniques for classifying and predicting outcomes: linear regression and classification trees. Three methods for estimating linear regression coefficients were described along with guidelines for choosing among them. You were provided with a step-wise technique for choosing variables to be included in a regression model. The assumptions underlying the regression technique were discussed, along with the resultant limitations. Overall guidelines for model development were provided.

You learned the importance of validating and revalidating your models before placing any reliance upon them. You were introduced to one of the simplest of pattern recognition methods, the classification tree, to be used whenever there are large numbers of potential predictors or when classification rather than quantitative prediction is your primary goal.

**Exercise 7.30.** Make a list of all the italicized terms in this chapter. Provide a definition for each one, along with an example.



**Exercise 7.31.** It is almost self-evident that levels of toluene, a commonly used solvent, would be observed in the blood after working in a room where the solvent was present in the air. Do the observations recorded below also suggest that blood levels are a function of age and body weight? Construct a model for predicting blood levels of toluene using this data.

Blood Tol	0.494	0.763	0.534	0.552	1.084	0.944	0.955	0.696
Air Tol	50	50	50	50	100	100	100	100
Weight	378	439	302	405	421	370	363	389
Age	95	95	84	85	86	86	83	86

Blood Tol	12.085	9.647	7.524	10.783	38.619	25.402	26.481	28.155
Air Tol	500	500	500	500	1000	1000	1000	1000
Weight	371	347	misg	348	378	433	363	420
Age	83	84	85	85	93	93	85	86

**Exercise 7.32.** Using the data from Exercise 6.18, develop a model for predicting whether an insurance agency will remain solvent.

**Exercise 7.33.** The weights of rat fetuses killed at various intervals after conception are recorded below. Test the hypothesis that the weight of a rat fetus doubles every 2.5 time intervals.

Interval	1	2	2	3	4	5	5	6	7	9	9
Weight	2.44	4.46	4.00	2.21	10.8	10.4	10.13	15.78	15.50	13.2	16.6

# Chapter 8

## Reporting Your Findings

**IN THIS CHAPTER**, we assume you have just completed an analysis of your own or someone else's research and now wish to issue a report on the overall findings. You'll learn what to report and how to go about reporting it, with particular emphasis on the statistical aspects of data collection and analysis.

One of the most common misrepresentations in scientific work is the scientific paper itself. It presents a mythical reconstruction of what actually happened. All of what are in retrospect mistaken ideas, badly designed experiments and incorrect calculations are omitted. The paper presents the research as if it had been carefully thought out, planned and executed according to a neat, rigorous process, for example involving testing of a hypothesis. The misrepresentation of the scientific paper is the most formal aspect of the misrepresentation of science as an orderly process based on a clearly defined method.

Brian Martin

### 8.1. WHAT TO REPORT

Reportable elements include all of the following:

- Study objectives
- Hypotheses
- Power and sample size calculations
- Data collection methods

- Clusters
- Validation methods
- Data summaries
- Details of the statistical analysis
- Sources of missing data
- Exceptions

**Study Objectives.** If you are contributing to the design or analysis of someone else's research efforts, a restatement of the objectives is an essential first step. This ensures that you and the principal investigator are on the same page. This may be necessary in order to formulate quantifiable, testable hypotheses.

Objectives may have shifted or been expanded upon. Often such changes are not documented. You cannot choose or justify the choice of statistical procedures without a thorough understanding of study objectives.

**Hypotheses.** To summarize what was stated in Chapter 5, both your primary and alternate hypotheses must be put in quantified testable form. Your primary hypothesis is used to establish the significance level and your alternative hypothesis to calculate the power of your tests.

Your objective may be to determine whether adding a certain feature to a product would increase sales. (Quick. Will this alternative hypothesis lead to a one-sided or a two-sided test?) Yet for reasons that have to do solely with the limitations of statistical procedures, your primary hypothesis will normally be a null hypothesis of no effect.

Not incidentally, as we saw in Chapter 6, the optimal statistical test for an ordered response is quite different from the statistic one uses for detecting an arbitrary difference among approaches. All the more reason why we need to state our alternative hypotheses explicitly.

**Power and Sample Size Calculations.** Your readers will want to know the details of your power and sample size calculations early on. If you don't let them know, they may assume the worst, for example, that your sample size is too small and your survey is not capable of detecting significant effects. State the alternative hypotheses that your study is intended to detect. Reference your methodology and/or the software you used in making your power calculations. State the source(s) you relied on for your initial estimates of incidence and variability.

Here is one example: "Over a 10-year period in the Himalayas, Dempsey and Peters [1995] observed an incidence of five infected individuals per

100 persons per year. To ensure a probability of at least 90% of detecting a reduction in disease incidence from five persons to one person per 100 persons per year using a one-sided Fisher's exact test at the 2.5% significance level, 400 individuals were assigned to each experimental procedure group. This sample size was determined using the StatXact-5 power calculations for comparing two binomials."

**Data Collection Methods.** Although others may have established the methods of data collection, a comprehensive knowledge of these methods is essential to your choice of statistical procedures and should be made apparent in report preparation. Consider that 90% of all errors occur during data collection as observations are erroneously recorded (GIGO), guessed at, or even faked. Seemingly innocuous work-arounds may have jeopardized the integrity of the study. You need to know and report on exactly how the data were collected, not on how they were supposed to have been collected.

You need to record how study subjects were selected, what was done to them (if appropriate), and when and how this was done. Details of recruitment or selection are essential if you are to convince readers that your work is applicable to a specific population. If incentives were used (phone cards, t-shirts, cash), their use should be noted.

Readers will want to know the nature and extent of any blinding (and of the problems you may have had to overcome to achieve it). They will want to know how each observational subject was selected—random, stratified, or cluster sampling? They will want to know the nature of the controls (and the reasoning underlying your choice of a passive or active control experimental procedure) and of the experimental design. Did each subject act as his own control as in a crossover design? Were case controls used? If they were matched, how were they matched?

You will need the answers to all of the following questions and should incorporate each of the answers in your report:

- What was the survey or experimental unit?
- What were all the potential sources of variation?
- How was each of the individual sources compensated for? In particular, was the sample simple or stratified?
- How were subjects or units grouped into strata?
- Were the units sampled in clusters or individually?
- How were subjects assigned to experimental procedures? If assignment was at random, how was this accomplished?
- How was independence of the observations ensured?

**Clusters.** Surveys often take advantage of the cost savings that result from naturally occurring groups such as work sites, schools, clinics, neighborhoods, even entire towns or states. Not surprisingly, the observations within such a group are correlated. For example, individuals in the same community or work group often have shared views. Married couples, the ones whose marriages last, tend to have shared values. The effect of such correlation must be accounted for by the use of the appropriate statistical procedures. Thus the nature and extent of such cluster sampling must be spelled out in detail in your reports.

**Exercise 8.1.** Examine three recent reports in your field of study or in any field that interests you. (Examine three of your own reports if you have them.) Answer the following in each instance:

1. What was the primary hypothesis?
2. What was the principal alternative hypothesis?
3. Which statistics were used and why?
4. Were all the assumptions underlying this statistic satisfied?
5. Was the power of this statistic reported?
6. Was the statistic the most powerful available?
7. If significant, was the size of the effect estimated?
8. Was a one-tailed or two-tailed test used? Was this correct?
9. What was the total number of tests that were performed?

If the answers to these questions were not present in the reports you reviewed, what was the reason? Had their authors something to hide?

**Validation Methods.** A survey will be compromised if any of the following is true:

- Participants are not representative of the population of interest.
- Responses are not independent among respondents.
- Nonresponders, that is, those who decline to participate, would have responded differently.
- Respondents lie or answer carelessly.
- Forms are incomplete.

Your report should detail the preventive measures used by the investigator and the degree to which they were successful.

You should describe the population(s) of interest in detail—providing demographic information where available—and similarly characterize the samples of participants to see whether they are indeed representative. (Graphs are essential here.)

You should describe the measures taken to ensure that responses were independent, including how participants were selected and where and how the survey took place.

A sample of nonrespondents should be contacted and evaluated. The demographics of the nonrespondents and their responses should be compared with those of the original sample.

How do you know respondents told the truth? You should report the results of any crosschecks such as redundant questions. And you should report the frequency of response omissions on a question-by-question basis.

## 8.2. TEXT, TABLE, OR GRAPH?

*Whatever is the use of a book without pictures?  
Alice in Alice in Wonderland*

A series of major decisions need to be made as to how you will report your results—text, table, or graph? Whatever Alice’s views on the subject, a graph may or may not be more efficient at communicating numeric information than the equivalent prose. This efficiency is in terms of the amount of information successfully communicated and not necessarily any space savings. Resist the temptation to enhance your prose with pictures.

And don’t fail to provide a comprehensive caption for each figure. As Good and Hardin note in their Wiley text, *Common Errors in Statistics* [2003], if the graphic is a summary of numeric information, then the graph caption is a summary of the graphic. The text should be considered part of the graphic design and should be carefully constructed rather than placed as an afterthought. Readers, for their own use, often copy graphics and tables that appear in articles and reports. A failure on your part to completely document the graphic in the caption can result in gross misrepresentation in these cases.

A sentence should be used for displaying 2 to 5 numbers, as in “The blood type of the population of the United States is approximately 45% O, 40% A, 11% B, and 4% AB.” Note that the blood types are ordered by frequency.

Tables with appropriate marginal means are often the best method of presenting results. Consider adding a row (or column, or both) of contrasts; for example, if the table has only two rows we could add a row of differences, row 1 minus row 2.

Tables dealing with two-factor arrays are straightforward, provided confidence limits are clearly associated with the correct set of figures. Tables

involving three or more factors simultaneously are not always clear to the reader and are best avoided.

Make sure the results are expressed in appropriate units. For example, parts per thousand may be more natural than percentages in certain cases.

A table of deviations from row and column means (or tables, if there are several strata) can alert us to the presence of outliers and may also reveal patterns in the data that were not yet considered.

**Exercise 8.2.** To report each of the following, should you use text, a table, or a graph? If a graphic, then what kind?

- Number of goals (each of 5 teams)
- Blood types of Australians
- Comparison treated/control red blood cell counts
- Comparison of blood types in two populations
- Location of leukemia cases by county
- Arm span vs. height (6 persons)

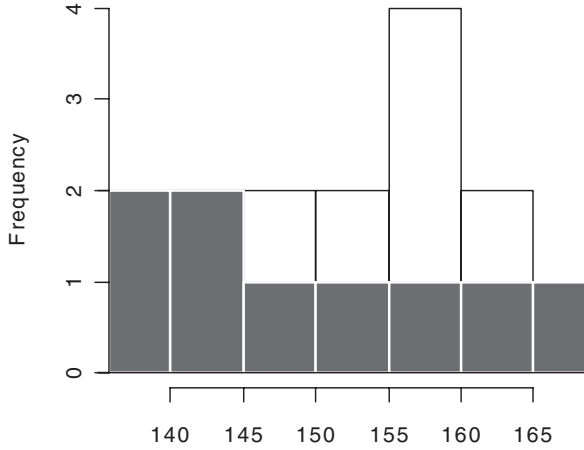
### 8.3. SUMMARIZING YOUR RESULTS

Your objective in summarizing your results should be to communicate some idea of all of the following:

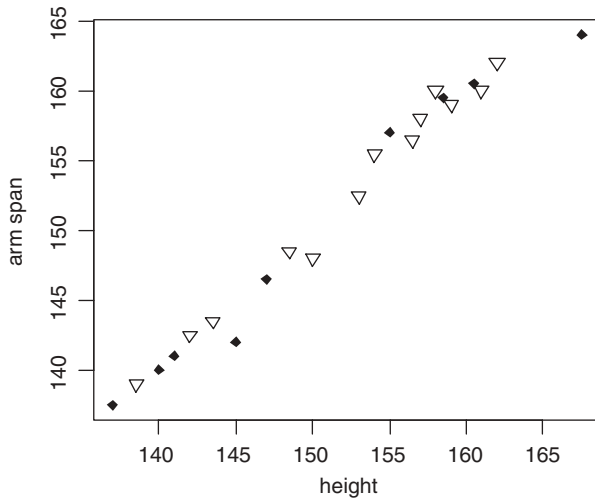
1. The populations from which your samples are drawn
2. Your estimates of population parameters
3. The dispersion of the distribution from which the observations were drawn
4. The precision of your estimates

Proceed in three steps:

First, characterize the populations and subpopulations from which your observations are drawn. Of course, this is the main goal in studies of market segmentation. A histogram or scatterplot can help communicate the existence of such subpopulations to our readers. Few real-life distributions resemble the bell-shaped normal curve depicted in Fig. 1.23. Most are bi- or even trimodal, with each mode or peak corresponding to a distinct subpopulation. We can let the histogram speak for itself, but a better idea, particularly if you already suspect that the basis for market segments is the value of a second variable (such as home ownership or level of education), is to add an additional dimension by dividing each of the histogram's bars into differently shaded segments whose size corresponds to the relative numbers in each subpopulation (Fig. 8.1).



**FIGURE 8.1** Histograms of class data by sex.



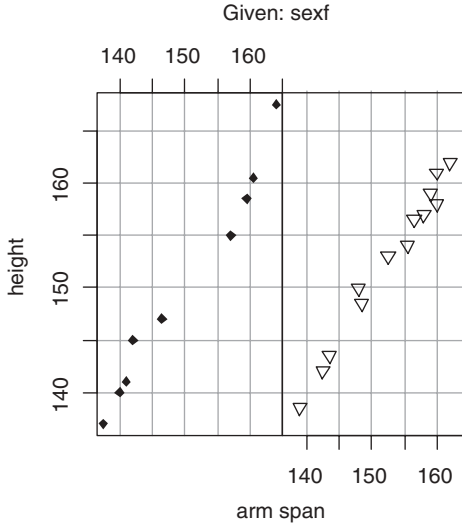
**FIGURE 8.2** Overlying scatterplots of class data by sex.

Similarly, we can provide for different subpopulations on a two-dimensional scatterplot by using different colors or shapes for the points (Figs. 8.2 and 8.3).

### 8.3.1. Center of the Distribution

For small samples of three to five observations, summary statistics are virtually meaningless. Reproduce the actual observations; this is easier to do and more informative.





**FIGURE 8.3** Side-by-side scatterplots of class data by sex.

Although the arithmetic mean or average is in common use, it can be very misleading. For example, the mean income in most countries is far in excess of the *median* income or 50th percentile to which most of us can relate. When the arithmetic mean is meaningful, it is usually equal to or close to the median. Consider reporting the median in the first place.

The *geometric mean* is more appropriate than the arithmetic in three sets of circumstances:

1. When losses or gains can best be expressed as a percentage rather than a fixed value
2. When rapid growth is involved
3. When the data span several orders of magnitude, as with the concentration of pollutants

Because bacterial populations can double in number in only a few hours, many government health regulations utilize the geometric rather than the arithmetic mean. A number of other government regulations also use it, although the sample median would be far more appropriate (and understandable). In any event, as your average reader may be unfamiliar with the geometric mean, be sure to comment on its use and on your reasons for adopting it.

The purpose of your inquiry must be kept in mind. Orders (in \$) from a machinery plant ranked by size may be quite skewed with a few large orders. The median order size might be of interest in describing sales; the mean order size would be of interest in estimating revenues and profits.

Are the results expressed in appropriate units? For example, are parts per thousand more natural in a specific case than percentages? Have we rounded off to the correct degree of precision, taking account of what we know about the variability of the results, and considering whether the reader will use them, perhaps by multiplying by a constant factor or another variable?

Whether you report a mean or a median, be sure to report only a sensible number of decimal places. Most statistical packages including R can give you nine or 10. Don't use them. If your observations were to the nearest integer, your report on the mean should include only a single decimal place. Limit tabulated values to no more than two effective (changing) digits. Readers can distinguish 354691 and 354634 at a glance but will be confused by 354691 and 357634.

### 8.3.2. Dispersion

The standard error of a summary is a useful measure of uncertainty *if* the observations come from a normal or Gaussian distribution. Then in 95% of the samples we would expect the sample mean to lie within two standard errors of the population mean.

But if the observations come from any of the following:

- A nonsymmetric distribution like an exponential or a Poisson
- A truncated distribution like the uniform
- A mixture of populations

we *cannot* draw any such inference. For such a distribution, the probability that a future observation would lie between plus and minus one standard error of the mean might be anywhere from 40% to 100%.

Recall that the standard error of the mean equals the standard deviation of a single observation divided by the square root of the sample size. As the standard error depends on the squares of individual observations, it is particularly sensitive to outliers. A few extra large observations, even a simple typographical error, might have a dramatic impact on its value.

If you can't be sure your observations come from a normal distribution, then for samples from nonsymmetric distributions of size 6 or less, tabulate the minimum, the median, and the maximum. For samples of size 7 and up, consider using a box and whiskers plot. For samples of size 30 and up, the bootstrap may provide the answer you need.

## 8.4. REPORTING ANALYSIS RESULTS

How you conduct and report your analysis will depend upon whether or not

- **Baseline results of the various groups are equivalent**
- **(if multiple observation sites were used) Results of the disparate experimental procedure sites may be combined**
- **(if adjunct or secondary experimental procedures were used) Results of the various adjunct experimental procedure groups may be combined**
- **Missing data, dropouts, and withdrawals are unrelated to experimental procedure**

Thus your report will have to include

1. Demonstrations of similarities and differences for the following:
  - Baseline values of the various experimental procedure groups
  - End points of the various subgroups determined by baseline variables and adjunct therapies
2. Explanations of protocol deviations including:
  - Ineligibles who were accidentally included in the study
  - Missing data
  - Dropouts and withdrawals
  - Modifications to procedures

Further explanations and stratifications will be necessary if the rates of any of the above protocol deviations differ among the groups assigned to the various experimental procedures. For example, if there are differences in the baseline demographics, then subsequent results will need to be stratified accordingly. Moreover, some plausible explanation for the differences must be advanced.

Here is an example: Suppose the vast majority of women in the study were in the control group. To avoid drawing false conclusions about the men, the results for men and women must be presented separately, unless one first can demonstrate that the experimental procedures have similar effects on men and women.

Report the results for each primary end point separately. For each end point:

- a) Report the aggregate results by experimental procedure for all who were examined during the study for whom you have end point or intermediate data.

- b) Report the aggregate results by experimental procedure only for those subjects who were actually eligible, who were treated originally as randomized, or who were not excluded for any other reason. Provide significance levels for comparisons of experimental procedures.
- c) Break down these latter results into subsets based on factors determined before the start of the study as having potential impact on the response to treatment, such as adjunct therapy or gender. Provide significance levels for comparisons of experimental procedures for these subsets of cases.
- d) List all factors uncovered during the trials that appear to have altered the effects of the experimental procedures. Provide a tabular comparison by experimental procedure for these factors, but do *not* include  $p$  values. The probability calculations that are used to generate  $p$  values are not applicable to hypotheses and subgroups that are conceived *after* the data have been examined.

If there are multiple end points, you have the option of providing a further multivariate comparison of the experimental procedures.

Last, but by no means least, you must report the number of tests performed. When we perform multiple tests in a study, there may not be room (or interest) in which to report all the results, but we do need to report the total number of statistical tests performed so that readers can draw their own conclusions as to the significance of the results that are reported. To repeat a finding of previous chapters, when we make 20 tests at the 1 in 20 or 5% significance level, we expect to find at least one or perhaps two results that are “statistically significant” by chance alone.

#### 8.4.1 $p$ Values? Or Confidence Intervals?

As you read the literature of your chosen field, you will soon discover that  $p$  values are more likely to be reported than confidence intervals. We don’t agree with this practice, and here is why:

Before we perform a statistical test, we are concerned with its significance level, that is, the probability that we will mistakenly reject our hypothesis when it is actually true. In contrast to the significance level, the  $p$  value is a random variable that varies from sample to sample. There may be highly significant differences between two populations, and yet the samples taken from those populations and the resulting  $p$  value may not reveal that difference. Consequently, it is not appropriate for us to compare the  $p$  values from two distinct experiments, or from tests on two variables measured in the same experiment, and declare that one is more significant than the other.

If we agree in advance of examining the data that we will reject the hypothesis if the  $p$  value is less than 5%, then our significance level is 5%.

Whether our  $p$  value proves to be 4.9% or 1% or 0.001%, we will come to the same conclusion. One set of results is not more significant than another; it is only that the difference we uncovered was measurably more extreme in one set of samples than in another.

We are less likely to mislead and more likely to communicate all the essential information if we provide confidence intervals about the estimated values. A confidence interval provides us with an estimate of the size of an effect as well as telling us whether an effect is significantly different from zero.

Confidence intervals, you will recall from Chapter 4, can be derived from the rejection regions of our hypothesis tests. Confidence intervals include all values of a parameter for which we would accept the hypothesis that the parameter takes that value.

**Warning:** A common error is to misinterpret the confidence interval as a statement about the unknown parameter. It is not true that the probability that a parameter is included in a 95% confidence interval is 95%. Nor is it at all reasonable to assume that the unknown parameter lies in the middle of the interval rather than toward one of the ends. What is true is that if we derive a large number of 95% confidence intervals, we can expect the true value of the parameter to be included in the computed intervals 95% of the time. Like the  $p$  value, the upper and lower confidence limits of a particular confidence interval are random variables, for they depend upon the sample that is drawn.

The probability that the confidence interval covers the true value of the parameter of interest and the method used to derive the interval must both be reported.

**Exercise 8.3.** Give at least two examples to illustrate why  $p$  values are not applicable to hypotheses and subgroups that are conceived after the data is examined.

## 8.5. EXCEPTIONS ARE THE REAL STORY

Before you draw conclusions, be sure you have accounted for all missing data, interviewed nonresponders, and determined whether the data were missing at random or were specific to one or more subgroups.

Let's look at two examples, the first involving nonresponders and the second airplanes.

### 8.5.1. Nonresponders

A major source of frustration for researchers is when the variances of the various samples are unequal. Alarm bells sound.  $t$ -Tests and the analysis of

variance are no longer applicable; we run to the textbooks in search of some variance-leveling transformation. And completely ignore the phenomena we've just uncovered.

If individuals have been assigned at random to the various study groups, the existence of a significant difference in any parameter suggests that there is a difference in the groups. The primary issue is to understand why the variances are so different, and what the implications are for the subjects of the study. It may well be the case that a new experimental procedure is not appropriate because of higher variance, even if the difference in means is favorable. This issue is important whether or not the difference was anticipated.

In many clinical measurements there are minimum and maximum values that are possible. If one of the experimental procedures is very effective, it will tend to push patient values into one of the extremes. This will produce a change in distribution from a relatively symmetric one to a skewed one, with a corresponding change in variance.

The distribution may not be unimodal. A large variance may occur because an experimental procedure is effective for only a subset of the patients. Then you are comparing mixtures of distributions of responders and nonresponders; specialized statistical techniques may be required.

### 8.5.2. The Missing Holes

During the Second World War, a group was studying planes returning from bombing Germany. They drew a rough diagram showing where the bullet holes were and recommended that those areas be reinforced. Abraham Wald, a statistician, pointed out that essential data were missing. What about the planes that didn't return?

When we think along these lines, we see that the areas of the returning planes that had almost no apparent bullet holes have their own story to tell. Bullet holes in a plane are likely to be at random, occurring over the entire plane. The planes that did not return were those that were hit in the areas where the returning planes had no holes. Do the data missing from your own experiments and surveys also have a story to tell?

### 8.5.3 Missing Data

As noted in an earlier section of this chapter, you need to report the number and source of all missing data. But especially important is to summarize and describe all those instances in which the incidence of missing data varied among the various treatment and procedure groups.

Here are two examples where the missing data was the real finding of the research effort:

To increase participation, respondents to a recent survey were offered a choice of completing a printed form or responding on-line. An unexpected finding was that the proportion of missing answers from the on-line survey was half that from the printed forms.

A minor drop in cholesterol levels was recorded among the small fraction of participants who completed a recent trial of a cholesterol-lowering drug. As it turned out, almost all those who completed the trial were in the control group. The numerous dropouts from the treatment group had only unkind words for the test product's foul taste and undrinkable consistency.

#### 8.5.4. Recognize and Report Biases

Very few studies can avoid bias at some point in sample selection, study conduct, and results interpretation. We focus on the wrong end points; participants and coinvestigators see through our blinding schemes; the effects of neglected and unobserved confounding factors overwhelm and outweigh the effects of our variables of interest. With careful and prolonged planning, we may reduce or eliminate many potential sources of bias, but seldom will we be able to eliminate all of them. Accept bias as inevitable and then endeavor to recognize and report all that do slip through the cracks.

Most biases occur during data collection, often as a result of taking observations from an unrepresentative subset of the population rather than from the population as a whole. An excellent example is the study that failed to include planes that did *not* return from combat.

When analyzing extended seismological and neurological data, investigators typically select specific cuts (a set of consecutive observations in time) for detailed analysis, rather than trying to examine all the data (a near impossibility). Not surprisingly, such “cuts” usually possess one or more intriguing features not to be found in run-of-the-mill samples. Too often theories evolve from these very biased selections.

The same is true of meteorological, geological, astronomical, and epidemiological studies where, with a large amount of available data, investigators naturally focus on the “interesting” patterns.

Limitations in the measuring instrument such as censoring at either end of the scale can result in biased estimates. Current methods of estimating cloud optical depth from satellite measurements produce biased results that depend strongly on satellite viewing geometry. Similar problems arise in high-temperature and high-pressure physics and in radioimmunoassay. In psychological and sociological studies, too often we measure that which is convenient to measure rather than that which is truly relevant.

Close collaboration between the statistician and the domain expert is essential if all sources of bias are to be detected and, if not corrected, accounted for and reported. We read a report recently by economist Otmar Issing in which it was stated that the three principal sources of bias in the measurement of price indices are substitution bias, quality change bias, and new product bias. We've no idea what he was talking about, but we do know that we would never attempt an analysis of pricing data without first consulting an economist.

## 8.6 SUMMARY AND REVIEW

In this chapter, we discussed the necessary contents of your reports, whether on your own work or that of others. We reviewed what to report, the best form in which to report it, and the appropriate statistics to use in summarizing your data and your analysis. We also discussed the need to report sources of missing data and potential biases.





# Chapter 9

## Problem Solving

**IF YOU HAVE MADE YOUR WAY THROUGH THE** first eight chapters of this text, then you may already have found that more and more people, strangers as well as friends, are seeking you out for your newly acquired expertise. (Not as many as if you were stunningly attractive or a film star, but a great many people nonetheless.) Your boss may even have announced that from now on you will be the official statistician of your group.

To prepare you for your new role in life, you will be asked in this chapter to work your way through a wide variety of problems that you may well encounter in practice. A final section will provide you with some overall guidelines. You'll soon learn that deciding which statistic to use is only one of many decisions that need be made.

### 9.1. THE PROBLEMS

1. With your clinical sites all lined up and everyone ready to proceed with a trial of a new experimental vaccine versus a control, the manufacturer tells you that because of problems at the plant, the 10,000 ampoules of vaccine you've received are all he will be able to send you. Explain why you can no longer guarantee the power of the test.
2. After collecting some 50 observations, 25 on members of a control group and 25 who have taken a low dose of a new experimental drug, you decide to add a third high-dose group to your clinical trial, and to take 75 additional observations, 25 on the members of each group. How would you go about analyzing these data?
3. You are given a data sample and asked to provide an interval estimate for the population variance. What two questions ought you to ask about the sample first?

4. John would like to do a survey of the use of controlled substances by teenagers but realizes he is unlikely to get truthful answers. He comes up with the following scheme: Each respondent is provided with a coin, instructions, a question sheet containing two questions, and a sheet on which to write their answer, yes or no. The two questions are:
- A. Is a cola (Coke or Pepsi) your favorite soft drink? Yes or No?
  - B. Have you used marijuana within the past seven days? Yes or No?

The teenaged respondents are instructed to flip the coin so that the interviewer cannot see it. If the coin comes up heads, they are to write their answer to the first question on the answer sheet; otherwise they are to write their answer to question 2.

Show that this approach will be successful, providing John already knows the proportion of teenagers who prefer colas to other types of soft drinks.

5. The town of San Philippe has asked you to provide confidence intervals for the recent census figures for their town. Are you able to do so? Could you do so if you had the some additional information? What might this information be? Just how would you go about calculating the confidence intervals?
6. The town of San Philippe has called on you once more. They have in hand the annual income figures for the past six years for their town and for their traditional rivals at Carfad-sur-la-mer and want you to make a statistical comparison. Are you able to do so? Could you do so if you had the some additional information? What might this information be? Just how would you go about calculating the confidence intervals?
7. You have just completed your analysis of a clinical trial and have found a few minor differences between patients subjected to the standard and revised procedures. The marketing manager has gone over your findings and noted that the differences are much greater if limited to patients who passed their first postprocedure day without complications. She asks you for a  $p$  value. What do you reply?
8. At the time of his death in 1971, psychologist Cyril Burt was viewed as an esteemed and influential member of his profession. Within months, psychologist Leon Kamin reported numerous flaws in Burt's research involving monozygotic twins who were reared apart. Shortly thereafter, a third psychologist, Arthur Jensen, also found fault with Burt's data. Their primary concern was the suspicious consistency of the correlation coefficients for the intelligence test scores of the monozygotic twins in Burt's studies. In each study Burt reported sum totals for the twins he had studied so far. His original results were published in 1943. In 1955 he added 6 pairs of twins and reported results for a total of 21 sets of twins. Likewise in 1966, he reported the results for a total of 53 pairs. In each study Burt reported correlation coefficients indicating the similarity of intelligence scores for monozygotic twins who were reared apart. A high correlation coefficient would make a strong case for Burt's hereditarian views.

Burt reported the following coefficients: 1943:  $r = .770$ ; 1955:  $r = .771$ ; 1966:  $r = .771$ . Why was this suspicious?

9. Which hypothesis testing method would you use to address each of the following? Permutation, parametric, or bootstrap?
  - a. Testing for an ordered dose response.
  - b. Testing whether the mean time to failure of a new light bulb in intermittent operation is one year.
  - c. Comparing two drugs, using the data from the following contingency table.

	Drug A	Drug B
Respond	5	9
No	5	1

- d. Comparing old and new procedures using the data from the following  $2 \times 2$  factorial design.

	Control	Old
Control		1,150 2,520 900 50
Young	5,640 5,120 780 4,430 7,230	7,100 11,020 13,065

### Ethical Standard

Polish-born Jerzy Neyman (1894–1981) is generally viewed as one of the most distinguished statisticians of the twentieth century. Along with Egon Pearson, he is responsible for the method of assigning the outcomes of a set of observations to either an acceptance or a rejection region in such a way that the power is maximized against a given alternative at a specified significance level. He was asked by the United States government to be part of an international committee monitoring the elections held in a newly liberated Greece after World War II. In the oversimplified view of the U.S. State Department, there were two groups running in the election: The Communists and The Good Guys. Professor Neyman’s report that both sides were guilty of extensive fraud pleased no one but set an ethical standard for other statisticians to follow.

10. The government has just audited 200 of your company's submissions over a four-year period and has found that the average claim was in error in the amount of \$135. Multiplying \$135 by the 4000 total submissions during that period, they are asking your company to reimburse them in the amount of \$540,000. List all possible objections to the government's approach.
11. Since I first began serving as a statistical consultant almost 40 years ago, I've made it a practice to begin every analysis by first computing the minimum and maximum of each variable. Can you tell why this practice would be of value to you as well?
12. Your mother has brought your attention to a newspaper article in which it is noted that one school has successfully predicted the outcome of every election of a U.S. president since 1976. Explain to her why this news does not surprise you.
13. A clinical study is well under way when it is noted that the values of critical end points vary far more from subject to subject than was expected originally. It is decided to increase the sample size. Is this an acceptable practice?
14. A clinical study is well under way when an unusual number of side effects is observed. The treatment code is broken, and it is discovered that the majority of the effects are occurring in subjects in the control group. Two cases arise:
  - a. The difference between the two treatment groups is statistically significant. It is decided to terminate the trials and recommend adoption of the new treatment. Is this an acceptable practice?
  - b. The difference between the two treatment groups is *not* statistically significant. It is decided to continue the trials but to assign twice as many subjects to the new treatment as are placed in the control group. Is this an acceptable practice?
15. A jurist has asked for your assistance with a case involving possible racial discrimination. Apparently the passing rate of minorities was 90% compared to 97% for whites. The jurist didn't think this was much of a difference, but then one of the attorneys pointed out that these numbers represented a jump in the failure rate from 3% to 10%. How would you go about helping this jurist to reach a decision?

When you hired on as a statistician at the Bumbling Pharmaceutical Company, they told you they'd been waiting a long time to find a candidate like you. Apparently they had, for your desk is already piled high with studies that are long overdue for analysis. Here is just a sample:

16. The end point values recorded by one physician are easily 10 times those recorded by all other investigators. Trying to track down the discrepancies, you discover that this physician has retired and closed his office. No one knows what became of his records. Your

co-workers instantly begin to offer you advice including all of the following:

- a. Discard all the data from this physician.
- b. Assume this physician left out a decimal point and use the corrected values.
- c. Report the results for this observer separately.
- d. Crack the treatment code and then decide.

What will you do?

17. A different clinical study involved this same physician. This time, he completed the question about side effects that asked whether this effect was “mild, severe, or life threatening” but failed to answer the preceding question that specified the nature of the side effect. Which of the following should you do?
  - a. Discard all the data from this physician.
  - b. Discard all the side effect data from this physician.
  - c. Report the results for this physician separately from the other results.
  - d. Crack the treatment code and then decide.
18. Summarizing recent observations on the planetary systems of stars, the *Monthly Notices* of the Royal Astronomical Society reported that the vast majority of extrasolar planets in our galaxy must be gas giants like Jupiter and Saturn as no Earth-size planet has been observed. What is your opinion?

## 9.2. SOLVING PRACTICAL PROBLEMS

In what follows, we suppose that you have been given a data set to analyze. The data did not come from a research effort that you designed, so there may be problems, many of them. We suggest you proceed as follows:

1. Determine the provenance of the observations.
2. Inspect the data.
3. Validate the data collection methods.
4. Formulate your hypotheses in testable form.
5. Choose methods for testing and estimation.
6. Be aware of what you don't know.
7. Perform the analysis.
8. Qualify your conclusions.

### 9.2.1. The Data's Provenance

Your very first questions should deal with *how* the data were collected. What population(s) were they drawn from? Were the members of the

sample(s) selected at random? Were the observations independent of one another? If treatments were involved, were individuals assigned to these treatments at random? Remember, statistics is applicable only to random samples.<sup>1</sup> You need to find out all the details of the sampling procedure to be sure.

You also need to ascertain that the sample is representative of the population it purports to be drawn from. If not, you'll need to 1) weight the observations, 2) stratify the sample to make it more representative, or 3) redefine the population before drawing conclusions from the sample.

### 9.2.2. Inspect the Data

If satisfied with the data's provenance, you can now begin to inspect the data you've been provided. Your first step should be to compute the minimum and the maximum of each variable in the data set and to compare them with the data ranges you were provided by the client. If any lie outside the acceptable range, you need to determine which specific data items are responsible and have these inspected and, if possible, corrected by the person(s) responsible for their collection.

I once had a long-term client who would not let me look at the data. Instead, he would merely ask me what statistical procedure to use next. I ought to have complained, but this client paid particularly high fees, or at least he did so in theory. The deal was that I would get my money when the firm for which my client worked got its first financing from the venture capitalists. So my thoughts were on the money to come and not on the data.

My client took ill—later I was to learn he had checked into a rehabilitation clinic for a metamphetamine addiction—and his firm asked me to take over. My first act was to ask for my money—they'd gotten their financing. While I waited for my check, I got to work, beginning my analysis as always by computing the minimum and the maximum of each variable. Many of the minimums were zero. I went to verify this finding with one of the technicians, only to discover that zeros were well outside the acceptable range.

The next step was to look at the individual items in the database. There were zeros everywhere. In fact, it looked as if more than half the data were either zeros or repeats of previous entries. Before I could report these discrepancies to my client's boss, he called me in to discuss my fees.

---

<sup>1</sup> The one notable exception is that it is possible to make a comparison between entire populations by permutation means.

“Ridiculous,” he said. We did not part as friends. I almost regret not taking the time to tell him that half the data he was relying on did not exist. *Tant pis*. No, they are not still in business.

Not incidentally, the best cure for bad data is prevention. I strongly urge that all your data be entered directly into a computer so they can be checked and verified immediately upon entry. You don’t want to be spending time tracking down corrections long after whoever entered 19.1 can remember whether the entry was supposed to be 1.91 or 9.1 or even 0.191.

### 9.2.3. Validate the Data Collection Methods

Few studies proceed exactly according to the protocol. Physicians switch treatments before the trial is completed. Sets of observations are missing or incomplete. A measuring instrument may have broken down midway through and been replaced by another, slightly different unit. Scoring methods were modified and observers provided with differing criteria employed. You need to determine the ways in which the protocol was modified and the extent and impact of such modifications.

A number of preventive measures may have been used. For example, a survey may have included redundant questions as crosschecks. You need to determine the extent to which these preventive measures were successful. Was blinding effective? Or did observers crack the treatment code? You need to determine the extent of missing data and whether this was the same for all groups in the study. You may need to ask for additional data derived from follow-up studies of nonresponders and dropouts.

### 9.2.4. Formulate Hypotheses

All hypotheses must be formulated *before* the data are examined. It is all too easy for the human mind to discern patterns in what is actually a sequence of totally random events—think of the faces and animals that always seem to form in the clouds.

As another example, suppose that while just passing the time you deal out a five-card poker hand. It’s a full house! Immediately, someone exclaims “What’s the probability that could happen?” If by “that” a full house is meant, its probability is easily computed. But the same exclamation might have resulted had a flush or a straight been dealt, or even three of a kind. The probability that “an interesting hand” will be dealt is much greater than the probability of a full house. Moreover, this might have been the third or even the fourth poker hand you’ve dealt; it’s just that this one was the first to prove interesting enough to attract attention.



The details of translating objectives into testable hypotheses were given in Chapters 5 and 8.

### 9.2.5. Choosing a Statistical Methodology

For the two-sample comparison, a  $t$ -test should be used. Remember, one-sided hypotheses lead to one-sided tests and two-sided hypotheses to two-sided tests. If the observations were made in pairs, the paired  $t$ -test should be used.

Permutation methods should be used to make  $k$ -sample comparisons. Your choice of statistic will depend upon the alternative hypothesis and the loss function.

Permutation methods should be used to analyze contingency tables.

The bootstrap is of value in obtaining confidence limits for quantiles and in model validation.

### 9.2.6. Be Aware of What You Don't Know

Far more statistical theory exists than can be provided in the confines of an introductory text. Entire books have been written on the topics of survey design, sequential analysis, and survival analysis, and that's just the letter "s." If you are unsure what statistical method is appropriate, don't hesitate to look it up on the Web or in a more advanced text.

### 9.2.7. Qualify Your Conclusions

Your conclusions can only be applicable to the extent that samples were representative of populations and experiments and surveys were free from bias. A report by G.C. Bent and S.A. Archfield is ideal in this regard.<sup>2</sup> This report can be viewed on-line at <http://water.usgs.gov/pubs/wri/wri024043/>.

They devote multiple paragraphs to describing the methods used, the assumptions made, the limitations on their model's range of application, potential sources of bias, and the method of validation. For example: "The logistic regression equation developed is applicable for stream sites with drainage areas between 0.02 and 7.00 mi<sup>2</sup> in the South Coastal Basin and between 0.14 and 8.94 mi<sup>2</sup> in the remainder of Massachusetts, because these were the smallest and largest drainage areas used in equation development for their respective areas.

---

<sup>2</sup> A logistic regression equation for estimating the probability of a stream flowing perennially in Massachusetts USGC. Water-Resources Investigations Report 02-4043.

“The equation may not be reliable for losing reaches of streams, such as for streams that flow off area underlain by till or bedrock onto an area underlain by stratified-drift deposits . . .”

“The logistic regression equation may not be reliable in areas of Massachusetts where ground-water and surface-water drainage areas for a stream site differ.” (Brent and Archfield provide examples of such areas.)

This report also illustrates how data quality, selection and measurement bias can affect results. For example: “The accuracy of the logistic regression equation is a function of the quality of the data used in its development. This data includes the measured perennial or intermittent status of a stream site, the occurrence of unknown regulation above a site, and the measured basin characteristics.

“The measured perennial or intermittent status of stream sites in Massachusetts is based on information in the USGS NWIS database. Stream-flow measured as less than  $0.005 \text{ ft}^3/\text{s}$  is rounded down to zero, so it is possible that several streamflow measurements reported as zero may have had flows less than  $0.005 \text{ ft}^3/\text{s}$  in the stream. This measurement would cause stream sites to be classified as intermittent when they actually are perennial.”

It is essential that your reports be similarly detailed and qualified whether they are to a client or to the general public in the form of a journal article.



# Appendix

## A Microsoft Office Excel Primer



**THIS APPENDIX COVERS WHAT EXCEL IS**, Excel document structure, how to start and quit Excel, and components of the Excel window. An animated HTML version of these guidelines is available online at <http://www.xlminer.com/xlprimer/Primer.htm>. The present version is provided through the courtesy of xlminer.com and statistics.com.

### **WHAT IS EXCEL?**

Microsoft Office Excel is the most commonly used spreadsheet software program. Entering numbers, text, or even a formula into the Excel spreadsheet (or a worksheet, as it is known in Excel) is quick and simple. Excel allows easy ways to calculate, analyze, and format data.

The calculation is instantaneous and allows the user to change data and see the result immediately in a dynamic “what if” scenario. Excel also helps the user to get a quick graphical representation of the worksheet contents. Last but not least, numerous software “add-ins” available from independent vendors allow you to supplement and enhance Excel’s existing capabilities.

### **EXCEL DOCUMENT STRUCTURE**

An Excel document is called a workbook. Workbooks are assigned default names such as Book1, Book2, etc. (You may and should change these names).

Each workbook may contain multiple pages, in the form of worksheets (and also charts). The active worksheet is displayed in the document window of Excel.

The default names of worksheets in a workbook are Sheet1, Sheet2 and so on. The worksheets are easily renamed. The names are displayed in the sheet tab at the bottom of the workbook, with the name of the active sheet shown in bold.

Each worksheet in Excel is made up of rows and columns. The rows are identified by numbers. The columns are identified by letters. The intersection of a row and a column defines a cell. A cell is the smallest unit to store a data element, a formula or a function.

Each cell is identified by a Cell Address (or Cell Reference), which is made up of a column and a row number. (Cell B4 is at the junction of Column B and Row 4).

The cell that is currently in use is called the Current Cell or the Active Cell. Selection of a number of adjacent cells defines a Range.

## HOW TO START AND QUIT EXCEL

Microsoft Excel can be started in many different ways. The two most frequently used methods are:

1. Choose Start ⇒ Programs ⇒ Microsoft Excel

(This notation will be used to mean: From the Windows “Start” menu, click on “Programs” and then click on “Microsoft Excel”)


2. Double-click on Microsoft Excel shortcut if it is available on the Desktop.

When you’re ready to quit Excel, you may Choose File ⇒ Exit, OR Click the “x” (Close) button at the right side of the Title Bar.

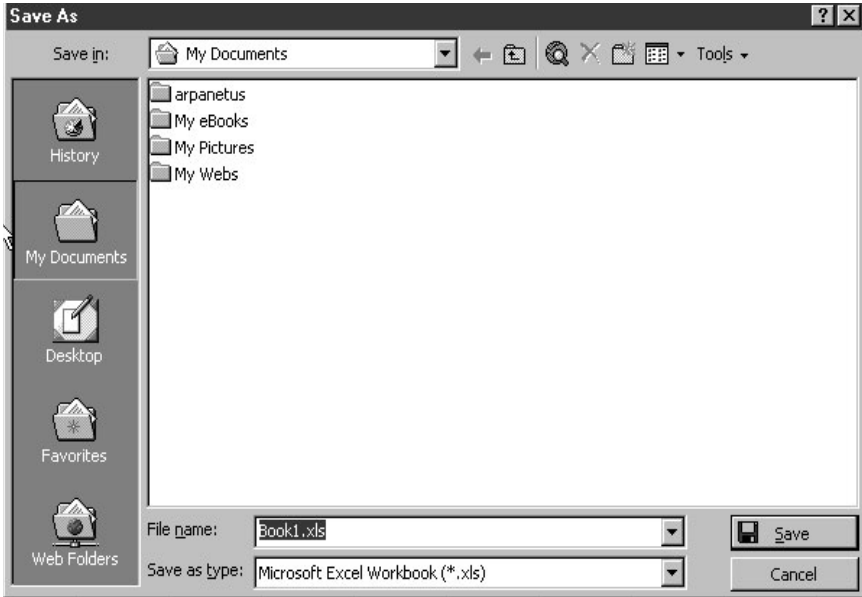
Before you quit and any time you feel apprehensive about losing the work you’ve done so far, you need to save your worksheet.

## Saving An Excel Workbook First Time

To save the workbook first time, do one of the following:


- Choose File → Save
- Choose File → Save As
- Choose the save  button from the Standard Toolbar

Whatever option you choose, Excel brings up the “Save As” dialog box.



The dialog box offers you a few options. You should choose

1. A file name
2. A folder, where you want to save the workbook

After you choose the file name and the folder, click on the  **Save** button to save your workbook.

#### HINTS:


1. If you're using this text as part of a class, create a folder with the name of this class and save all your work there.
2. Use meaningful file names so it will be easy to locate the file later.
3. Save often. But use a different file name each time, for example, class-data01 classdata02 and so forth. If you don't change the file name, the new file will be written on top of the old, destroying its contents.

## Entering Data in Cells

This section covers entering both numeric and text data. To enter data in a cell

1. Select the cell.
2. Type data either **directly** in the cell or in the **Formula Bar**.
3. Press **Enter** to accept the data and move down by a cell.

You may also use the arrow keys on the keyboard to accept the data and move by one cell in the direction of the arrow. The **Tab** key has the same effect as the **Right Arrow** key.

To cancel an entry while typing (i.e., before pressing Enter), press the **Esc** key. If you have already pressed Enter, use **Edit** → **Undo** to cancel the entry. You may also use the **Undo**  button on the Formatting Toolbar.

### IMPORTANT

A cell may not always display all the data it contains. The display of data depends on the cell width and the formatting used for that cell. By contrast, the Formula Bar always shows the entire content of the active cell.

## Entering Text Data

When we enter text data in a cell, the following rules apply

- Alignment:** Texts are automatically left aligned.
- Font:** A 10-point Arial font is used by default.
- Visual Truncation:** If the length of the data exceeds the cell width, the text *appears* to overflow into the next cell. However, if the next cell is not empty, the data *looks* truncated.
- Wrapping:** Text does not wrap, unless explicitly specified.
- Auto Completion:** If the first few characters, entered in the current cell, uniquely match with the text already existing in another cell in the same column, Excel fills the remaining characters for you. This is called the **auto completion** feature of Excel. (You have the option of ignoring this feature and typing your own data).

## Entering Numeric Data

The rules for numeric data are as follows:

- Alignment:** By default, numeric data are right-aligned.
- Precision Limit:** Numbers are stored with a maximum precision of 15 digits. If a number has more than 15 significant digits, the extra digits are converted to zero.

**General Format:****Integers:**

Excel automatically adjusts the column width to accommodate up to 11 digits. If the data is longer than 11 digits, Excel uses scientific (exponential) notation.

For example, if the number is 1234567890123, it will be shown as 1.23457E+12.

**Numbers Containing  
Decimal point:**

For presentation, Excel rounds off these numbers to fit in the cell. The cell width is increased up to 11 digits, depending on the size of the integer part of the number. For a bigger number, Excel uses scientific notation.

**Numbers Containing  
Comma, Dollar Sign,  
and Percent Sign:**

Excel automatically adjusts the column width to fit these numbers.

**Inserting and Deleting Columns and Rows**

To insert a column, use one of the following methods:

**Method 1**

Step 1: Select a cell in the position where you want to insert a column. (To insert a column after Column B, click on any cell in column C, say cell C5).

Step 2: Choose Insert → Columns.

To insert multiple columns, select multiple cells in appropriate positions in Step 1. Selecting cells C5 to E5 in Step 1 will allow you to insert three columns between column B and column C.

**Method 2**

Step 1: Select a column by clicking on the heading of the column.

Step 2: Choose Insert → Columns.

You may select more than one column in Step 1 to insert multiple columns.

To insert a row, use one of the following methods:

**Method 1**

Step 1: Select a cell in the position where you want to insert a row. (To insert a row after row 7, click on any cell in row 8, say cell C8).

Step 2: Choose Insert → Rows.

If you select multiple cells in Step 1, more than one row will get inserted. The positions of these rows will be determined by the cells you choose in Step 1.



## Method 2

Step 1: Select a row by clicking on the heading of the row.

Step 2: Choose Insert → Rows.

To insert multiple rows, select the appropriate rows in Step 1.

Example: If you select row 8 in Step 1, Excel will insert a row after the seventh row. However, if you select row 8 to row 10 in Step 1, Excel will insert three rows between the seventh and the eighth rows.

## Deleting Columns and Rows

To delete Columns and Rows in an Excel worksheet,

Select the Columns or Rows you want to delete.

Chose Edit → Delete

The row and column headings also act as control buttons and can be used to change the sizes of rows and columns. The options available are:

- **Changing Column and Row Sizes Manually:**

Use your mouse to drag the right boundary of a column or the bottom boundary of a row until you get the desired size.

- **Adjust the Sizes Automatically (AutoFit):**

Double-click the right boundary of a column or the bottom boundary of a row. The column/row will resize itself to accommodate the largest entry.

**Note:** If you select multiple rows/columns and use the boundary of one of them for double-clicking, the sizes of the selected rows/columns will be automatically adjusted. If you click on the Select All button at the top left corner of the worksheet (see Animation), and then double-click on a row/column boundary the AutoFit option will adjust the sizes of all the rows/columns in the worksheet.

# Index to Excel Functions and Excel Add-Ins



Absolute value, 140  
Average(), 6, 74  
BinomDist(), 49, 123  
ChartWizard, 21  
Combin(), 46  
Correl(), 102  
Cos(), 159  
Entering data, 224  
Formula Bar, 5  
If(96)

Median(), 5, 9  
Menu Bar, 8  
Normsinv(), 17, 126  
Percentile(), 17  
Rand(), 118  
Rank(), 146  
ScatterPlot, 14  
Sort, 8, 118  
Workbook, 221



# Subject Index

- Accuracy, 26, 89
- Add-Ins
  - Box Sampler, 5, 70, 82, 98, 140
  - Ctree, 186
  - DataDeskXL, 5, 10, 22, 92, 156
  - Resampling Stats, 28, 96, 103, 140–3
  - Solver, 123
  - Xlminer, 192
  - XLStat, 7, 25, 162, 175
- Additive model, 159
- Alternative hypothesis, 141, 198
- ARIMA, 184
- Assumptions, see Tests
- Audit, 116, 173
  
- Baseline, 205
- Bias, 31, 209
- Binomial
  - distribution, 48, 51
  - parameter, 52
  - random variable, 73
  - trial, 43, 94
- Blinding, 106
- Blocking, 12, 117, 145, 174, 197
- bootstrap, 27, 81, 127, 185, 205, 218
  - parametric, 90
  - percentile, 27, 89
- Box and whiskers plot, 7
- Box plot, 10
- Boyle's Law, 1, 155
  
- CART, 186
- Categorical variable, 20, 148, 153
- Cauchy distribution, 71
- Chi-square statistic, 152
- Classification, 158, 186
- Coefficient, 155
- Conditional probability, 53
- Confidence, 55
- Confidence interval, 29, 82, 86, 96, 206
- Contingency table, 59, 153
- Controls, 46
- Correlation, 84, 101
- Cross-validation, 185
- Cumulative distribution, 16, 19, 64
  
- Data collection, 7, 31, 197, 217
- Data
  - categorical, 148
  - continuous, 71, 75
  - discrete, 20, 66, 75
  - metric, 20
  - ordinal, 20
  - types, 20
- Data mining, 190
- Dispersion, see Variance
- Dose response, 141
  
- Empirical distribution, 51, 66, 128
- Equally likely events, 39
- Estimation, 81, 89

- Examples  
   agriculture, 138, 1444  
   astrophysics, 42, 57, 110  
   biology, 69, 77, 84, 100–4, 114–6, 194–6  
   business, 55, 81, 100, 113, 134  
   clinical trials, 86, 98, 106–08, 128, 151, 208  
   economic, 110, 117  
   education, 13, 136  
   epidemiology, 45, 57, 93, 101, 117, 129, 135, 149, 157, 167  
   geologic, 217  
   law, 47, 86, 111, 118  
   political science, 45, 53, 81, 110, 122  
   psychology, 109  
   sociology, 81, 109  
 Exchangeable observations, 99, 146  
 Expected value, 50, 67, 71, 182  
 Experimental unit, 114, 197  
 Exponential distribution, 71
- Factorial, 45  
 Fisher's exact test, 150
- Goodness of fit, 172  
 Graphs, 199, 202, 203  
 Group sequential design, 130  
 Growth processes, 202
- Histogram, 24, 201  
 HIV, 115  
 Hypothesis  
   Alternative, 78, 86, 141  
   formulation, 109, 114, 218  
   null, 77  
   testing, 76, 86, 207
- Identically distributed, 99, 171  
 Independence, 57, 74, 109, 119  
 Independent events, 58  
 Independent variables, 75  
 Interquartile range, 35  
 Interval estimate, see Confidence interval
- Lift ratio, 192  
 Likert scale, 166
- Marginals, 149  
 Martingale, 40  
 Matched pairs, 113  
 Mathematical Expectation, see Expected value.  
 Maximum, 6  
 Mean, arithmetic, 23  
 Mean, geometric, 202  
 Median, 4, 23, 48, 203  
 Meta-analysis, 134  
 Minimum, 5, 8  
 Missing data, 2, 122, 205, 209  
 Modes, 24, 48, 208  
 Model, 155  
 Monte Carlo, 95, 102  
 Multinomial, 53  
 Multisample comparison, 138  
 Mutually exclusive events, 41, 44
- Nonrespondents, 199, 208  
 Nonsymmetrical (see skewed)  
 Normal distribution, 72, 125, 201  
 Null hypothesis, 77
- Objectives, 196  
 Outcomes vs. events, 41
- Parameter, 200  
 Pearson correlation, 102  
 Percentages, 137  
 Percentile, 15, 17, 29, 182  
 Permutation, 45  
 Permutation test, 95, 97, 139, 143  
 Pie chart, 21  
 Pitman correlation, 102  
 Placebo effect, 106  
 Poisson distribution, 68, 93  
 Poll, see survey  
 Pollution, 167  
 Population, 111, 200  
 Power, 100, 122, 125, 141, 196, 198  
 Precision, 26, 200  
 Prediction, 156, 172  
 Predictor, 155  
 Prevention, 173  
 Probability laws, 40, 44  
 p-value, 98, 206

- Qualitative vs quantitative, 166
- Quality control, 72, 93, 141
- Quantile (see percentile)
- Range, 6
- Ranks, 146–7
- Rearrangement, 46, 78, 139
- Regression, 159
  - coefficients, 161, 170
  - Deming (EIV), 168
  - LAD, 168
  - linear, 160
  - multivariable, 175
  - nonlinear, 161
  - OLS, 162
  - quantile, 182
- Regression tree (see CART)
- Rejection region, 151
- Reportable elements, 195
- Resample, 183
- Residual, 163, 173
- Robust, 99
- Sample
  - random, 32, 76, 109, 116, 118
  - representative, 32, 109, 116, 214
  - size, 27, 120, 127, 191
  - stratified, 147
- Sampling, 197
  - adaptive, 133
  - clusters, 119, 198
  - error, 172
  - sequential, 121, 129
  - unit, 33
- Scatter plot, 12, 202
- Selectivity, 101
- Sensitivity, 101
- Shift alternative,, 64
- Significance
  - exact level, 99
  - level, 86, 98–9, 122, 206
  - practical vs statistical, 144
- Simulation, 82, 128
- Skewed, 26
- Slope, 163
- Standard deviation, 35
- Standard error, 97, 204
- Statistic, 26, 73
- Stein's two-stage procedure, 129
- Strata (see blocking)
- Strip chart, 5
- Student's *t*, 91, 97, 100, 218
- Support, 55, 192
- Surrogate variable, 115
- Survey, 3, 31, 45, 108, 116, 119
- Symmetric distribution, 68
- Test
  - assumptions, 98, 171
  - multiple samples, 138
  - one- vs. two-sided, 200
  - one sample, 89
  - parametric, 97
  - resampling, see Permutation test
- Treatment allocation, 117
- Type I, II errors, 80, 98, 100, 124, 130
- Uniform distribution, 204
- Validation, 158, 183, 198
- Variance, 34, 65, 68, 76, 99, 138, 145, 195
- Variation, 2–3, 34–35, 169, 197
- Venn diagram, 41, 64