

*Fifth Edition*

# STATISTICS, DATA ANALYSIS, AND DECISION MODELING

**James R. Evans**

*University of Cincinnati*

*International Edition contributions by*

**Ayanendranath Basu**

*Indian Statistical Institute, Kolkata*

**PEARSON**

© Pearson Education Limited 2013

ISBN 10: 0-273-76822-0

ISBN 13: 978-0-273-76822-7

*Authorized adaptation from the United States edition, entitled Statistics, Data Analysis and Decision Modeling, 5<sup>th</sup> edition,  
ISBN 978-0-13-274428-7 by James R. Evans published by Pearson Education © 2013.*

Typeset in Palatino by Jouve India Pvt Ltd

Printed and bound by Courier Kendalville in The United States of America

# BRIEF CONTENTS

## PART I Statistics and Data Analysis 25

- Chapter 1 Data and Business Decisions 27
- Chapter 2 Descriptive Statistics and Data Analysis 55
- Chapter 3 Probability Concepts and Distributions 89
- Chapter 4 Sampling and Estimation 123
- Chapter 5 Hypothesis Testing and Statistical Inference 162
- Chapter 6 Regression Analysis 196
- Chapter 7 Forecasting 237
- Chapter 8 Introduction to Statistical Quality Control 272

## PART II Decision Modeling and Analysis 293

- Chapter 9 Building and Using Decision Models 295
- Chapter 10 Decision Models with Uncertainty and Risk 324
- Chapter 11 Decisions, Uncertainty, and Risk 367
- Chapter 12 Queues and Process Simulation Modeling 402
- Chapter 13 Linear Optimization 435
- Chapter 14 Integer, Nonlinear, and Advanced Optimization Methods 482

Appendix 533

Index 545

# CONTENTS

Preface 21

## Part I STATISTICS AND DATA ANALYSIS 25

### Chapter 1 DATA AND BUSINESS DECISIONS 27

Introduction	28
Data in the Business Environment	28
Sources and Types of Data	30
Metrics and Data Classification	31
Statistical Thinking	35
Populations and Samples	36
Using Microsoft Excel	37
Basic Excel Skills	38
<i>Skill-Builder Exercise 1.1</i>	38
Copying Formulas and Cell References	38
<i>Skill-Builder Exercise 1.2</i>	39
Functions	40
<i>Skill-Builder Exercise 1.3</i>	42
Other Useful Excel Tips	42
Excel Add-Ins	43
<i>Skill-Builder Exercise 1.4</i>	44
Displaying Data with Excel Charts	45
Column and Bar Charts	45
<i>Skill-Builder Exercise 1.5</i>	46
Line Charts	47
<i>Skill-Builder Exercise 1.6</i>	47
Pie Charts	47
<i>Skill-Builder Exercise 1.7</i>	47
Area Charts	48
Scatter Diagrams	48
<i>Skill-Builder Exercise 1.8</i>	48
Miscellaneous Excel Charts	49
Ethics and Data Presentation	49
<i>Skill-Builder Exercise 1.9</i>	50
Basic Concepts Review Questions	51
Problems and Applications	51
Case: A Data Collection and Analysis Project	52

## **Chapter 2 DESCRIPTIVE STATISTICS AND DATA ANALYSIS 55**

Introduction	56
Descriptive Statistics	56
Frequency Distributions, Histograms, and Data Profiles	57
Categorical Data	58
Numerical Data	58
<i>Skill-Builder Exercise 2.1</i>	62
<i>Skill-Builder Exercise 2.2</i>	62
Data Profiles	62
Descriptive Statistics for Numerical Data	63
Measures of Location	63
Measures of Dispersion	64
<i>Skill-Builder Exercise 2.3</i>	66
Measures of Shape	67
Excel Descriptive Statistics Tool	68
<i>Skill-Builder Exercise 2.4</i>	68
Measures of Association	69
<i>Skill-Builder Exercise 2.5</i>	71
Descriptive Statistics for Categorical Data	71
<i>Skill-Builder Exercise 2.6</i>	72
Visual Display of Statistical Measures	73
Box Plots	73
Dot-Scale Diagrams	73
<i>Skill-Builder Exercise 2.7</i>	73
Outliers	74
Data Analysis Using PivotTables	74
<i>Skill-Builder Exercise 2.8</i>	77
<i>Skill-Builder Exercise 2.9</i>	77
Basic Concepts Review Questions	78
Problems and Applications	78
Case: The Malcolm Baldrige Award	81
<i>Skill-Builder Exercise 2.10</i>	83
<i>Skill-Builder Exercise 2.11</i>	84

## **Chapter 3 PROBABILITY CONCEPTS AND DISTRIBUTIONS 89**

Introduction	90
Basic Concepts of Probability	90
Basic Probability Rules and Formulas	91
Conditional Probability	92
<i>Skill-Builder Exercise 3.1</i>	94
Random Variables and Probability Distributions	94
Discrete Probability Distributions	97
Expected Value and Variance of a Discrete Random Variable	98

<i>Skill-Builder Exercise 3.2</i>	99
Bernoulli Distribution	99
Binomial Distribution	99
Poisson Distribution	100
<i>Skill-Builder Exercise 3.3</i>	102
Continuous Probability Distributions	102
Uniform Distribution	104
Normal Distribution	105
<i>Skill-Builder Exercise 3.4</i>	108
Triangular Distribution	108
Exponential Distribution	109
Probability Distributions in <i>PHStat</i>	110
Other Useful Distributions	110
Joint and Marginal Probability Distributions	113
Basic Concepts Review Questions	114
Problems and Applications	114
Case: Probability Analysis for Quality Measurements	118

## **Chapter 4 SAMPLING AND ESTIMATION 123**

Introduction	124
Statistical Sampling	124
Sample Design	124
Sampling Methods	125
Errors in Sampling	127
Random Sampling From Probability Distributions	127
Sampling From Discrete Probability Distributions	128
<i>Skill-Builder Exercise 4.1</i>	129
Sampling From Common Probability Distributions	129
A Statistical Sampling Experiment in Finance	130
<i>Skill-Builder Exercise 4.2</i>	130
Sampling Distributions and Sampling Error	131
<i>Skill-Builder Exercise 4.3</i>	134
Applying the Sampling Distribution of the Mean	134
Sampling and Estimation	134
Point Estimates	135
Unbiased Estimators	136
<i>Skill-Builder Exercise 4.4</i>	137
Interval Estimates	137
Confidence Intervals: Concepts and Applications	137
Confidence Interval for the Mean with Known Population Standard Deviation	138
<i>Skill-Builder Exercise 4.5</i>	140

Confidence Interval for the Mean with Unknown Population Standard Deviation	140
Confidence Interval for a Proportion	142
Confidence Intervals for the Variance and Standard Deviation	143
Confidence Interval for a Population Total	145
Using Confidence Intervals for Decision Making	146
Confidence Intervals and Sample Size	146
Prediction Intervals	148
Additional Types of Confidence Intervals	149
Differences Between Means, Independent Samples	149
Differences Between Means, Paired Samples	149
Differences Between Proportions	150
Basic Concepts Review Questions	150
Problems and Applications	150
Case: Analyzing a Customer Survey	153
<i>Skill-Builder Exercise 4.6</i>	155
<i>Skill-Builder Exercise 4.7</i>	156
<i>Skill-Builder Exercise 4.8</i>	157
<i>Skill-Builder Exercise 4.9</i>	157
<b>Chapter 5 HYPOTHESIS TESTING AND STATISTICAL INFERENCE</b>	<b>162</b>
Introduction	163
Basic Concepts of Hypothesis Testing	163
Hypothesis Formulation	164
Significance Level	165
Decision Rules	166
Spreadsheet Support for Hypothesis Testing	169
One-Sample Hypothesis Tests	169
One-Sample Tests for Means	169
Using <i>p</i> -Values	171
One-Sample Tests for Proportions	172
One Sample Test for the Variance	174
Type II Errors and the Power of A Test	175
<i>Skill-Builder Exercise 5.1</i>	177
Two-Sample Hypothesis Tests	177
Two-Sample Tests for Means	177
Two-Sample Test for Means with Paired Samples	179
Two-Sample Tests for Proportions	179
Hypothesis Tests and Confidence Intervals	180
Test for Equality of Variances	181
<i>Skill-Builder Exercise 5.2</i>	182
Anova: Testing Differences of Several Means	182
Assumptions of ANOVA	184
Tukey-Kramer Multiple Comparison Procedure	184

Chi-Square Test for Independence	186
<i>Skill-Builder Exercise 5.3</i>	188
Basic Concepts Review Questions	188
Problems and Applications	188
Case: <i>HATCO, Inc.</i>	191
<i>Skill-Builder Exercise 5.4</i>	193
<b>Chapter 6 REGRESSION ANALYSIS</b>	<b>196</b>
Introduction	197
Simple Linear Regression	198
<i>Skill-Builder Exercise 6.1</i>	199
Least-Squares Regression	200
<i>Skill-Builder Exercise 6.2</i>	202
A Practical Application of Simple Regression to Investment Risk	202
Simple Linear Regression in Excel	203
<i>Skill-Builder Exercise 6.3</i>	204
Regression Statistics	204
Regression as Analysis of Variance	205
Testing Hypotheses for Regression Coefficients	205
Confidence Intervals for Regression Coefficients	206
Confidence and Prediction Intervals for X-Values	206
Residual Analysis and Regression Assumptions	206
Standard Residuals	208
<i>Skill-Builder Exercise 6.4</i>	208
Checking Assumptions	208
Multiple Linear Regression	210
<i>Skill-Builder Exercise 6.5</i>	210
Interpreting Results from Multiple Linear Regression	212
Correlation and Multicollinearity	212
Building Good Regression Models	214
Stepwise Regression	217
<i>Skill-Builder Exercise 6.6</i>	217
Best-Subsets Regression	217
The Art of Model Building in Regression	218
Regression with Categorical Independent Variables	220
Categorical Variables with More Than Two Levels	223
<i>Skill-Builder Exercise 6.7</i>	225
Regression Models with Nonlinear Terms	225
<i>Skill-Builder Exercise 6.8</i>	226
Basic Concepts Review Questions	228
Problems and Applications	228
Case: <i>Hatco</i>	231

## **Chapter 7 FORECASTING 237**

Introduction	238
Qualitative and Judgmental Methods	238
Historical Analogy	239
The Delphi Method	239
Indicators and Indexes for Forecasting	239
Statistical Forecasting Models	240
Forecasting Models for Stationary Time Series	242
Moving Average Models	242
Error Metrics and Forecast Accuracy	244
<i>Skill-Builder Exercise 7.1</i>	246
Exponential Smoothing Models	246
<i>Skill-Builder Exercise 7.2</i>	248
Forecasting Models for Time Series with a Linear Trend	248
Regression-Based Forecasting	248
Advanced Forecasting Models	249
Autoregressive Forecasting Models	250
<i>Skill-Builder Exercise 7.3</i>	252
Forecasting Models with Seasonality	252
Incorporating Seasonality in Regression Models	253
<i>Skill-Builder Exercise 7.4</i>	255
Forecasting Models with Trend and Seasonality	255
Regression Forecasting with Causal Variables	255
Choosing and Optimizing Forecasting Models Using <i>CB Predictor</i>	257
<i>Skill-Builder Exercise 7.5</i>	259
The Practice of Forecasting	262
Basic Concepts Review Questions	263
Problems and Applications	264
Case: Energy Forecasting	265

## **Chapter 8 INTRODUCTION TO STATISTICAL QUALITY CONTROL 272**

Introduction	272
The Role of Statistics and Data Analysis in Quality Control	273
Statistical Process Control	274
Control Charts	274
$\bar{x}$ - and R-Charts	275
<i>Skill-Builder Exercise 8.1</i>	280
Analyzing Control Charts	280
Sudden Shift in the Process Average	281
Cycles	281
Trends	281

Hugging the Center Line	281
Hugging the Control Limits	282
<i>Skill-Builder Exercise 8.2</i>	282
<i>Skill-Builder Exercise 8.3</i>	284
Control Charts for Attributes	284
Variable Sample Size	286
<i>Skill-Builder Exercise 8.4</i>	288
Process Capability Analysis	288
<i>Skill-Builder Exercise 8.5</i>	290
Basic Concepts Review Questions	290
Problems and Applications	290
Case: Quality Control Analysis	291

## **Part II Decision Modeling and Analysis 293**

### **Chapter 9 BUILDING AND USING DECISION MODELS 295**

Introduction	295
Decision Models	296
Model Analysis	299
What-If Analysis	299
<i>Skill-Builder Exercise 9.1</i>	301
<i>Skill-Builder Exercise 9.2</i>	302
<i>Skill-Builder Exercise 9.3</i>	302
Model Optimization	302
Tools for Model Building	304
Logic and Business Principles	304
<i>Skill-Builder Exercise 9.4</i>	305
Common Mathematical Functions	305
Data Fitting	306
<i>Skill-Builder Exercise 9.5</i>	308
Spreadsheet Engineering	308
<i>Skill-Builder Exercise 9.6</i>	309
Spreadsheet Modeling Examples	309
New Product Development	309
<i>Skill-Builder Exercise 9.7</i>	311
Single Period Purchase Decisions	311
Overbooking Decisions	312
Project Management	313
Model Assumptions, Complexity, and Realism	315
<i>Skill-Builder Exercise 9.8</i>	317
Basic Concepts Review Questions	317
Problems and Applications	318
Case: An Inventory Management Decision Model	321

## **Chapter 10 DECISION MODELS WITH UNCERTAINTY AND RISK 324**

Introduction	325
Spreadsheet Models with Random Variables	325
Monte Carlo Simulation	326
<i>Skill-Builder Exercise 10.1</i>	327
Monte Carlo Simulation Using <i>Crystal Ball</i>	327
Defining Uncertain Model Inputs	328
Running a Simulation	332
Saving <i>Crystal Ball</i> Runs	334
Analyzing Results	334
<i>Skill-Builder Exercise 10.2</i>	338
<i>Crystal Ball</i> Charts	339
<i>Crystal Ball</i> Reports and Data Extraction	342
<i>Crystal Ball</i> Functions and Tools	342
Applications of Monte Carlo Simulation and <i>Crystal Ball</i> Features	343
Newsvendor Model: Fitting Input Distributions, <i>Decision Table</i> Tool, and Custom Distribution	343
<i>Skill-Builder Exercise 10.3</i>	347
<i>Skill-Builder Exercise 10.4</i>	348
Overbooking Model: <i>Crystal Ball</i> Functions	348
<i>Skill-Builder Exercise 10.5</i>	349
Cash Budgeting: Correlated Assumptions	349
New Product Introduction: <i>Tornado Chart</i> Tool	352
<i>Skill-Builder Exercise 10.6</i>	353
Project Management: Alternate Input Parameters and the <i>Bootstrap</i> Tool	353
<i>Skill-Builder Exercise 10.7</i>	358
Basic Concepts Review Questions	358
Problems and Applications	359
Case: J&G Bank	362

## **Chapter 11 DECISIONS, UNCERTAINTY, AND RISK 367**

Introduction	368
Decision Making Under Certainty	368
Decisions Involving a Single Alternative	369
<i>Skill-Builder Exercise 11.1</i>	369
Decisions Involving Non-mutually Exclusive Alternatives	369
Decisions Involving Mutually Exclusive Alternatives	370
Decisions Involving Uncertainty and Risk	371
Making Decisions with Uncertain Information	371
Decision Strategies for a Minimize Objective	372

<i>Skill-Builder Exercise 11.2</i>	374
Decision Strategies for a Maximize Objective	374
Risk and Variability	375
Expected Value Decision Making	377
Analysis of Portfolio Risk	378
<i>Skill-Builder Exercise 11.3</i>	380
The “Flaw of Averages”	380
<i>Skill-Builder Exercise 11.4</i>	380
Decision Trees	381
A Pharmaceutical R&D Model	381
Decision Trees and Risk	382
Sensitivity Analysis in Decision Trees	384
<i>Skill-Builder Exercise 11.5</i>	384
The Value of Information	384
Decisions with Sample Information	386
Conditional Probabilities and Bayes’s Rule	387
Utility and Decision Making	389
<i>Skill-Builder Exercise 11.6</i>	392
Exponential Utility Functions	393
<i>Skill-Builder Exercise 11.7</i>	394
Basic Concepts Review Questions	394
Problems and Applications	395
Case: The Sandwich Decision	399

## **Chapter 12 QUEUES AND PROCESS SIMULATION MODELING 402**

Introduction	402
Queues and Queuing Systems	403
Basic Concepts of Queuing Systems	403
Customer Characteristics	404
Service Characteristics	405
Queue Characteristics	405
System Configuration	405
Performance Measures	406
Analytical Queuing Models	406
Single-Server Model	407
<i>Skill-Builder Exercise 12.1</i>	408
Little’s Law	408
Process Simulation Concepts	409
<i>Skill-Builder Exercise 12.2</i>	410
Process Simulation with <i>SimQuick</i>	410
Getting Started with <i>SimQuick</i>	411
A Queuing Simulation Model	412

<i>Skill-Builder Exercise 12.3</i>	416
Queues in Series with Blocking	417
Grocery Store Checkout Model with Resources	418
Manufacturing Inspection Model with Decision Points	421
Pull System Supply Chain with Exit Schedules	424
Other <i>SimQuick</i> Features and Commercial Simulation Software	426
Continuous Simulation Modeling	427
Basic Concepts Review Questions	430
Problems and Applications	431
Case: Production/Inventory Planning	434
<b>Chapter 13 LINEAR OPTIMIZATION</b>	<b>435</b>
Introduction	435
Building Linear Optimization Models	436
Characteristics of Linear Optimization Models	439
Implementing Linear Optimization Models on Spreadsheets	440
Excel Functions to Avoid in Modeling Linear Programs	441
Solving Linear Optimization Models	442
Solving the SSC Model Using Standard Solver	442
Solving the SSC Model Using <i>Premium Solver</i>	444
Solver Outcomes and Solution Messages	446
Interpreting Solver Reports	446
<i>Skill-Builder Exercise 13.1</i>	450
How Solver Creates Names in Reports	451
Difficulties with Solver	451
Applications of Linear Optimization	451
Process Selection	453
<i>Skill-Builder Exercise 13.2</i>	454
Blending	454
<i>Skill-Builder Exercise 13.3</i>	456
Portfolio Investment	456
<i>Skill-Builder Exercise 13.4</i>	457
Transportation Problem	457
Interpreting Reduced Costs	461
Multiperiod Production Planning	461
<i>Skill-Builder Exercise 13.5</i>	463
Multiperiod Financial Planning	463
<i>Skill-Builder Exercise 13.6</i>	464
A Model with Bounded Variables	464
A Production/Marketing Allocation Model	469
How Solver Works	473
Basic Concepts Review Questions	474

Problems and Applications 474  
Case: Haller's Pub & Brewery 481

## **Chapter 14 INTEGER, NONLINEAR, AND ADVANCED OPTIMIZATION METHODS 482**

Introduction 482  
Integer Optimization Models 483  
    A Cutting Stock Problem 483  
    Solving Integer Optimization Models 484  
    *Skill-Builder Exercise 14.1* 486  
Integer Optimization Models with Binary Variables 487  
    Project Selection 487  
    Site Location Model 488  
    *Skill-Builder Exercise 14.2* 491  
    Computer Configuration 491  
    *Skill-Builder Exercise 14.3* 494  
    A Supply Chain Facility Location Model 494  
Mixed Integer Optimization Models 495  
    Plant Location Model 495  
    A Model with Fixed Costs 497  
Nonlinear Optimization 499  
    Hotel Pricing 499  
    Solving Nonlinear Optimization Models 501  
    Markowitz Portfolio Model 503  
    *Skill-Builder Exercise 14.4* 506  
    *Evolutionary Solver* for Nonsmooth Optimization 506  
    Rectilinear Location Model 508  
    *Skill-Builder Exercise 14.5* 508  
    Job Sequencing 509  
    *Skill-Builder Exercise 14.6* 512  
Risk Analysis and Optimization 512  
Combining Optimization and Simulation 515  
    A Portfolio Allocation Model 515  
    Using *OptQuest* 516  
    *Skill-Builder Exercise 14.7* 524  
Basic Concepts Review Questions 524  
Problems and Applications 524  
Case: Tindall Bookstores 530

**Appendix 533**

**Index 545**

# PREFACE

## INTENDED AUDIENCE

*Statistics, Data Analysis, and Decision Modeling* was written to meet the need for an introductory text that provides the fundamentals of business statistics and decision models/optimization, focusing on practical applications of data analysis and decision modeling, all presented in a simple and straightforward fashion.

The text consists of 14 chapters in two distinct parts. The first eight chapters deal with statistical and data analysis topics, while the remaining chapters deal with decision models and applications. Thus, the text may be used for:

- MBA or undergraduate business programs that combine topics in business statistics and management science into a single, brief, quantitative methods
- Business programs that teach statistics and management science in short, modular courses
- Executive MBA programs
- Graduate refresher courses for business statistics and management science

## NEW TO THIS EDITION

The fifth edition of this text has been carefully revised to improve clarity and pedagogical features, and incorporate new and revised topics. Many significant changes have been made, which include the following:

1. Spreadsheet-based tools and applications are compatible with *Microsoft Excel 2010*, which is used throughout this edition.
2. Every chapter has been carefully revised to improve clarity. Many explanations of critical concepts have been enhanced using new business examples and data sets. The sequencing of several topics have been reorganized to improve their flow within the book.
3. Excel, *PHStat*, and other software notes have been moved to chapter appendixes so as not to disrupt the flow of the text.
4. “Skill-Builder” exercises, designed to provide experience with applying Excel, have been located in the text to facilitate immediate application of new concepts.
5. Data used in many problems have been changed, and new problems have been added.

## SUBSTANCE

The danger in using quantitative methods does not generally lie in the inability to perform the requisite calculations, but rather in the lack of a fundamental understanding of why to use a procedure, how to use it correctly, and how to properly interpret results. A key focus of this text is conceptual understanding using simple and practical examples rather than a plug-and-chug or point-and-click mentality, as are often done in other texts, supplemented by appropriate theory. On the other hand, the text does not attempt to be an encyclopedia of detailed quantitative procedures, but focuses on useful concepts and tools for today's managers.

To support the presentation of topics in business statistics and decision modeling, this text integrates fundamental theory and practical applications in a spreadsheet environment using *Microsoft Excel 2010* and various spreadsheet add-ins, specifically:

- *PHStat*, a collection of statistical tools that enhance the capabilities of Excel; published by Pearson Education

- *Crystal Ball* (including *CB Predictor* for forecasting and *OptQuest* for optimization), a powerful commercial package for risk analysis
- *TreePlan*, a decision analysis add-in
- *SimQuick*, an Excel-based application for process simulation, published by Pearson Education
- *Risk Solver Platform for Education*, an Excel-based tool for risk analysis, simulation, and optimization

These tools have been integrated throughout the text to simplify the presentations and implement tools and calculations so that more focus can be placed on interpretation and understanding the managerial implications of results.

## TO THE STUDENTS

The Companion Website for this text ([www.pearsoninternationaleditions.com/evans](http://www.pearsoninternationaleditions.com/evans)) contains the following:

- *Data files*—download the data and model files used throughout the text in examples, problems, and exercises
- *PHStat*—download of the software from Pearson
- *TreePlan*—link to a free trial version
- *Risk Solver Platform for Education*—link to a free trial version
- *Crystal Ball*—link to a free trial version
- *SimQuick*—link that will direct you to where you may purchase a standalone version of the software from Pearson
- *Subscription Content*—a Companion Website Access Code accompanies this book. This code gives you access to the following software:
  - *Risk Solver Platform for Education*—link that will direct students to an upgrade version
  - *Crystal Ball*—link that will direct students to an upgrade version
  - *SimQuick*—link that will allow you to download the software from Pearson

To redeem the subscription content:

- Visit [www.pearsoninternationaleditions.com/evans](http://www.pearsoninternationaleditions.com/evans)
- Click on the Companion Website link.
- Click on the Subscription Content link.
- First-time users will need to register, while returning users may log-in.
- Once you are logged in you will be brought to a page which will inform you how to download the software from the corresponding software company's Web site.

## TO THE INSTRUCTORS

To access instructor solutions files, please visit [www.pearsoninternationaleditions.com/evans](http://www.pearsoninternationaleditions.com/evans) and choose the instructor resources option. A variety of instructor resources are available for instructors who register for our secure environment. The Instructor's Solutions Manual files and PowerPoint presentation files for each chapter are available for download.

As a registered faculty member, you can login directly to download resource files, and receive immediate access and instructions for installing Course Management content to your campus server.

Need help? Our dedicated Technical Support team is ready to assist instructors with questions about the media supplements that accompany this text. Visit <http://247.pearsoned.com/> for answers to frequently asked questions and toll-free user support phone numbers.

# PART I

## STATISTICS AND DATA ANALYSIS

## *Chapter 1*

# Data and Business Decisions

- INTRODUCTION 28
- DATA IN THE BUSINESS ENVIRONMENT 28
- SOURCES AND TYPES OF DATA 30
  - Metrics and Data Classification 31
- STATISTICAL THINKING 35
  - Populations and Samples 36
- USING MICROSOFT EXCEL 37
  - Basic Excel Skills 38
  - Copying Formulas and Cell References 38
  - Functions 40
  - Other Useful Excel Tips 42
  - Excel Add-Ins 43
- DISPLAYING DATA WITH EXCEL CHARTS 45
  - Column and Bar Charts 45
  - Line Charts 47
  - Pie Charts 47
  - Area Charts 48
  - Scatter Diagrams 48
  - Miscellaneous Excel Charts 49
  - Ethics and Data Presentation 49
- BASIC CONCEPTS REVIEW QUESTIONS 51
- PROBLEMS AND APPLICATIONS 51
- CASE: A DATA COLLECTION AND ANALYSIS PROJECT 52
- APPENDIX 1.1: EXCEL AND *PHStat* NOTES 52
  - A. Using the *PHStat Stack Data* and *Unstack Data* Tools 52
  - B. Creating Charts in Excel 2010 53

## INTRODUCTION

Since the dawn of the electronic age and the Internet, both individuals and organizations have had access to an enormous wealth of data and information. *Data* are numerical facts and figures that are collected through some type of measurement process. *Information* comes from analyzing data; that is, extracting meaning from data to support evaluation and decision making. Modern organizations—which include for-profit businesses such as retailers, manufacturers, hotels, and airlines, as well as nonprofit organizations like hospitals, educational institutions, and government agencies—need good data to evaluate daily performance and to make critical strategic and operational decisions.

The purpose of this book is to introduce you to statistical methods for analyzing data; ways of using data effectively to make informed decisions; and approaches for developing, analyzing, and solving models of decision problems. Part I of this book (Chapters 1–8) focuses on key issues of statistics and data analysis, and Part II (Chapters 9–14) introduces you to various types of decision models that rely on good data analysis.

In this chapter, we discuss the roles of data analysis in business, discuss how data are used in evaluating business performance, introduce some fundamental issues of statistics and measurement, and introduce spreadsheets as a support tool for data analysis and decision modeling.

## DATA IN THE BUSINESS ENVIRONMENT

Data are used in virtually every major function in business, government, health care, education, and other nonprofit organizations. For example:

- Annual reports summarize data about companies' profitability and market share both in numerical form and in charts and graphs to communicate with shareholders.
- Accountants conduct audits and use statistical methods to determine whether figures reported on a firm's balance sheet fairly represents the actual data by examining samples (that is, subsets) of accounting data, such as accounts receivable.
- Financial analysts collect and analyze a variety of data to understand the contribution that a business provides to its shareholders. These typically include profitability, revenue growth, return on investment, asset utilization, operating margins, earnings per share, economic value added (EVA), shareholder value, and other relevant measures.
- Marketing researchers collect and analyze data to evaluate consumer perceptions of new products.
- Operations managers use data on production performance, manufacturing quality, delivery times, order accuracy, supplier performance, productivity, costs, and environmental compliance to manage their operations.
- Human resource managers measure employee satisfaction, track turnover, training costs, employee satisfaction, turnover, market innovation, training effectiveness, and skills development.
- Within the federal government, economists analyze unemployment rates, manufacturing capacity and global economic indicators to provide forecasts and trends.
- Hospitals track many different clinical outcomes for regulatory compliance reporting and for their own analysis.
- Schools analyze test performance and state boards of education use statistical performance data to allocate budgets to school districts.

Data support a variety of company purposes, such as planning, reviewing company performance, improving operations, and comparing company performance with competitors' or "best practices" benchmarks. Data that organizations use should focus on critical success factors that lead to competitive advantage. An example from the

Boeing Company shows the value of having good business data and analysis capabilities.<sup>1</sup> In the early 1990s, Boeing's assembly lines were morasses of inefficiency. A manual numbering system dating back to World War II bomber days was used to keep track of an airplane's four million parts and 170 miles of wiring; changing a part on a 737's landing gear meant renumbering 464 pages of drawings. Factory floors were covered with huge tubs of spare parts worth millions of dollars. In an attempt to grab market share from rival Airbus, the company discounted planes deeply and was buried by an onslaught of orders. The attempt to double production rates, coupled with implementation of a new production control system, resulted in Boeing being forced to shut down its 737 and 747 lines for 27 days in October 1997, leading to a \$178 million loss and a shakeup of top management. Much of the blame was focused on Boeing's financial practices and lack of real-time financial data. With a new Chief Financial Officer and finance team, the company created a "control panel" of vital measures, such as materials costs, inventory turns, overtime, and defects, using a color-coded spreadsheet. For the first time, Boeing was able to generate a series of charts showing which of its programs were creating value and which were destroying it. The results were eye-opening and helped formulate a growth plan. As one manager noted, "The data will set you free."

Data also provide key inputs to decision models. A **decision model** is a logical or mathematical representation of a problem or business situation that can be developed from theory or observation. Decision models establish relationships between actions that decision makers might take and results that they might expect, thereby allowing the decision makers to predict what might happen based on the model. For instance, the manager of a grocery store might want to know how best to use price promotions, coupons, and advertising to increase sales. In the past, grocers have studied the relationship of sales volume to programs such as these by conducting controlled experiments to identify the relationship between actions and sales volumes.<sup>2</sup> That is, they implement different combinations of price promotions, coupons, and advertising (the decision variables), and then observe the sales that result. Using the data from these experiments, we can develop a predictive model of sales as a function of these decision variables. Such a model might look like the following:

$$\text{Sales} = a + b \times \text{Price} + c \times \text{Coupons} + d \times \text{Advertising} + e \times \text{Price} \times \text{Advertising}$$

where  $a, b, c, d$ , and  $e$  are constants that are estimated from the data. By setting levels for price, coupons, and advertising, the model estimates a level of sales. The manager can use the model to help identify effective pricing, promotion, and advertising strategies.

Because of the ease with which data can be generated and transmitted today, managers, supervisors, and front-line workers can easily be overwhelmed. Data need to be summarized in a quantitative or visual fashion. One of the most important tools for doing this is **statistics**, which David Hand, former president of the Royal Statistical Society in the UK, defines as *both the science of uncertainty and the technology of extracting information from data*.<sup>3</sup> Statistics involve collecting, organizing, analyzing, interpreting, and presenting data. A **statistic** is a summary measure of data. You are undoubtedly familiar with the concept of statistics in daily life as reported in newspapers and the media; baseball batting averages, airline on-time arrival performance, and economic statistics such as Consumer Price Index are just a few examples. We can easily google statistical information about investments and financial markets, college loans and home mortgage rates, survey results about national political issues, team and individual

<sup>1</sup> Jerry Useem, "Boeing versus Boeing," *Fortune*, October 2, 2000, 148–160.

<sup>2</sup> "Flanking in a Price War," *Interfaces*, Vol. 19, No. 2, 1989, 1–12.

<sup>3</sup> David Hand, "Statistics: An Overview," in Miodrag Lovric, Ed., *International Encyclopedia of Statistical Science*, Springer Major Reference; <http://www.springer.com/statistics/book/978-3-642-04897-5>, p. 1504.

sports performance, and well, just about anything. To paraphrase Apple, “There’s a stat for that!” Modern spreadsheet technology, such as Microsoft Excel, has made it quite easy to apply statistical tools to organize, analyze, and present data to make them more understandable.

Most organizations have traditionally focused on financial and market information, such as profit, sales volume, and market share. Today, however, many organizations use a wide variety of measures that provide a comprehensive view of business performance. For example, the Malcolm Baldrige Award Criteria for Performance Excellence, which many organizations use as a high-performance management framework, suggest that high-performing organizations need to measure results in five basic categories:

1. *Product and process outcomes*, such as reliability, performance, defect levels, service errors, response times, productivity, production flexibility, setup times, time to market, waste stream reductions, innovation, emergency preparedness, strategic plan accomplishment, and supply chain effectiveness.
2. *Customer-focused outcomes*, such as customer satisfaction and dissatisfaction, customer retention, complaints and complaint resolution, customer perceived value, and gains and losses of customers.
3. *Workforce-focused outcomes*, such as workforce engagement and satisfaction, employee retention, absenteeism, turnover, safety, training effectiveness, and leadership development.
4. *Leadership and governance outcomes*, such as communication effectiveness, governance and accountability, environmental and regulatory compliance, ethical behavior, and organizational citizenship.
5. *Financial and market outcomes*. Financial outcomes might include revenue, profit and loss, net assets, cash-to-cash cycle time, earnings per share, and financial operations efficiency (collections, billings, receivables). Market outcomes might include market share, business growth, and new products and service introductions.

Understanding key relationships among these types of measures can help organizations make better decisions. For example, Sears, Roebuck and Company provided a consulting group with 13 financial measures, hundreds of thousands of employee satisfaction data points, and millions of data points on customer satisfaction. Using advanced statistical tools, the analysts discovered that employee attitudes about the job and the company are key factors that predict their behavior with customers, which, in turn, predicts the likelihood of customer retention and recommendations, which, in turn, predict financial performance. As a result, Sears was able to predict that if a store increases its employee satisfaction score by five units, customer satisfaction scores will go up by two units and revenue growth will beat the stores’ national average by 0.5%.<sup>4</sup> Such an analysis can help managers make decisions, for instance, on improved human resource policies.

## SOURCES AND TYPES OF DATA

Data may come from a variety of sources: internal record-keeping, special studies, and external databases. Internal data are routinely collected by accounting, marketing, and operations functions of a business. These might include production output, material costs, sales, accounts receivable, and customer demographics. Other data must be generated through special efforts. For example, customer satisfaction data are often acquired by mail,

---

<sup>4</sup> “Bringing Sears into the New World,” *Fortune*, October 13, 1997, 183–184.

Internet, or telephone surveys; personal interviews; or focus groups. External databases are often used for comparative purposes, marketing projects, and economic analyses. These might include population trends, interest rates, industry performance, consumer spending, and international trade data. Such data can be found in annual reports, Standard & Poor's Compustat data sets, industry trade associations, or government databases.

One example of a comprehensive government database is FedStats ([www.fedstats.gov](http://www.fedstats.gov)), which has been available to the public since 1997. FedStats provides access to the full range of official statistical information produced by the Federal Government without having to know in advance which Federal agency produces which particular statistic. With convenient searching and linking capabilities to more than 100 agencies—which provide data and trend information on such topics as economic and population trends, crime, education, health care, aviation safety, energy use, farm production, and more—FedStats provides one location for access to the full breadth of Federal statistical information.

The use of data for analysis and decision making certainly is not limited to business. Science, engineering, medicine, and sports, to name just a few, are examples of professions that rely heavily on data. Table 1.1 provides a list of data files that are available in the *Statistics Data Files* folder on the Companion Website accompanying this book. All are saved in Microsoft Excel workbooks. These data files will be used throughout this book to illustrate various issues associated with statistics and data analysis and also for many of the questions and problems at the end of the chapters. They show but a sample of the wide variety of applications for which statistics and data analysis techniques may be used.

## Metrics and Data Classification

A **metric** is a unit of measurement that provides a way to objectively quantify performance. For example, senior managers might assess overall business performance using such metrics as net profit, return on investment, market share, and customer satisfaction. A supervisor in a manufacturing plant might monitor the quality of a production process for a polished faucet by visually inspecting the products and counting the number of surface defects. A useful metric would be the percentage of faucets that have surface defects. For a web-based retailer, some useful metrics are the percentage of orders filled accurately and the time taken to fill a customer's order. **Measurement** is the act of obtaining data associated with a metric. **Measures** are numerical values associated with a metric.

Metrics can be either discrete or continuous. A **discrete metric** is one that is derived from counting something. For example, a part dimension is either within tolerance or out of tolerance; an order is complete or incomplete; or an invoice can have one, two, three, or any number of errors. Some discrete metrics associated with these examples would be the proportion of parts whose dimensions are within tolerance, the number of incomplete orders for each day, and the number of errors per invoice. **Continuous metrics** are based on a continuous scale of measurement. Any metrics involving dollars, length, time, volume, or weight, for example, are continuous.

A key performance dimension might be measured using either a continuous or a discrete metric. For example, an airline flight is considered on time if it arrives no later than 15 minutes from the scheduled arrival time. We could evaluate on-time performance by counting the number of flights that are late, or by measuring the number of minutes that flights are late. Discrete data are usually easier to capture and record, but provide less information than continuous data. However, one generally must collect a larger amount of discrete data to draw appropriate statistical conclusions as compared to continuous data.

**TABLE 1.1** Data Files Available on Companion Website**Business and Economics**

Accounting Professionals	House Sales
Atlanta Airline Data	Housing Starts
Automobile Quality	Insurance Survey
Baldridge	Internet Usage
Banking Data	Microprocessor Data
Beverage Sales	Mortgage Rates
Call Center Data	New Account Processing
Cell Phone Survey	New Car Sales
Cereal Data	Nuclear Power
China Trade Data	Prime Rate
Closing Stock Prices	Quality Control Case Data
Coal Consumption	Quality Measurements
Coal Production	Refrigerators
Concert Sales	Residential Electricity Data
Consumer Price Index	Restaurant Sales
Consumer Transportation Survey	Retail Electricity Prices
Credit Approval Decisions	Retirement Portfolio
Customer Support Survey	Room Inspection
Customer Survey	S&P 500
DJIA December Close	Salary Data
EEO Employment Report	Sales Data
Employees Salaries	Sampling Error Experiment
Energy Production & Consumption	Science and Engineering Jobs
Federal Funds Rate	State Unemployment Rates
Gas & Electric	Statistical Quality Control Problems
Gasoline Prices	Surgery Infections
Gasoline Sales	Syringe Samples
Google Stock Prices	Treasury Yield Rates
Hatco	Unions and Labor Law Data
Hi-Definition Televisions	University Grant Proposals
Home Market Value	

**Behavioral and Social Sciences**

Arizona Population	Freshman College Data
Blood Pressure	Graduate School Survey
Burglaries	Infant Mortality
California Census Data	MBA Student Survey
Census Education Data	Ohio Education Performance
Church Contributions	Ohio Prison Population
Colleges and Universities	Self-Esteem
Death Cause Statistics	Smoking & Cancer
Demographics	Student Grades
Facebook Survey	Vacation Survey

**Science and Engineering**

Pile Foundation	Surface Finish
Seattle Weather	Washington, DC, Weather

**Sports**

Baseball Attendance	NASCAR Track Data
Golfing Statistics	National Football League
Major League Baseball	Olympic Track and Field Data

A	B	C	D	E
1	2010 J.D. Power and Associates Initial Quality Statistics			
2				
3	Problems per 100 Vehicles			
4	Acura	86		
5	Audi	111		
6	BMW	113		
7	Buick	114		
8	Cadillac	111		
9	Chevrolet	111		

**FIGURE 1.1** Example of Cross-Sectional, Univariate Data  
(Portion of Automobile Quality)

When we deal with data, it is important to understand the type of data in order to select the appropriate statistical tool or procedure. One classification of data is the following:

**1.** Types of data

- *Cross-sectional*—data that are collected over a single period of time
- *Time series*—data collected over time

**2.** Number of variables

- *Univariate*—data consisting of a single variable
- *Multivariate*—data consisting of two or more (often related) variables

Figures 1.1–1.4 show examples of data sets from Table 1.1 representing each combination from this classification.

Another classification of data is by the type of measurement scale. Failure to understand the differences in measurement scales can easily result in erroneous or misleading analysis. Data may be classified into four groups:

**1. Categorical (nominal) data**, which are sorted into categories according to specified characteristics. For example, a firm's customers might be classified by their geographical region (North America, South America, Europe, and Pacific); employees might be classified as managers, supervisors, and associates. The categories bear no quantitative relationship to one another, but we usually assign an arbitrary number to each category to ease the process of managing the data and computing statistics. Categorical data are usually counted or expressed as proportions or percentages.

A	B	C	D	E	F
1	Banking Data				
2					
3	Median Age	Median Years Education	Median Income	Median Home Value	Median Household Wealth
4	35.9	14.8	\$91,033	\$183,104	\$220,741
5	37.7	13.8	\$86,748	\$163,843	\$223,152
6	36.8	13.8	\$72,245	\$142,732	\$176,926
7	35.3	13.2	\$70,639	\$145,024	\$166,260
8	35.3	13.2	\$64,879	\$135,951	\$148,868
9	34.8	13.7	\$75,591	\$155,334	\$188,310
10	39.3	14.4	\$80,615	\$181,265	\$201,743
11					\$38,766

**FIGURE 1.2** Example of Cross-Sectional, Multivariate Data  
(Portion of Banking Data)

A	B	C	D	E	F
1	Gasoline Prices - Average U.S. Prices for Regular Gasoline				
2					
3	Date	Price			
4	Jan 03, 2000	\$ 1.26			
5	Jan 10, 2000	\$ 1.25			
6	Jan 17, 2000	\$ 1.27			
7	Jan 24, 2000	\$ 1.31			
8	Jan 31, 2000	\$ 1.31			
9	Feb 07, 2000	\$ 1.32			
10	Feb 14, 2000	\$ 1.35			
11	Feb 21, 2000	\$ 1.40			

**FIGURE 1.3** Example of Time-Series, Univariate Data  
(Portion of *Gasoline Prices*)

**2. Ordinal data**, which are ordered or ranked according to some relationship to one another. A common example in business is data from survey scales; for example, rating a service as poor, average, good, very good, or excellent. Such data are categorical but also have a natural order, and consequently, are ordinal. Other examples include ranking regions according to sales levels each month and NCAA basketball rankings. Ordinal data are more meaningful than categorical data because data can be compared to one another (“excellent” is better than “very good”). However, like categorical data, statistics such as averages are meaningless even if numerical codes are associated with each category (such as your class rank), because ordinal data have no fixed units of measurement. In addition, meaningful numerical statements about differences between categories cannot be made. For example, the difference in strength between basketball teams ranked 1 and 2 is not necessarily the same as the difference between those ranked 2 and 3.

**3. Interval data**, which are ordered, have a specified measure of the distance between observations but have no natural zero. Common examples are time and temperature. Time is relative to global location, and calendars have arbitrary starting dates. Both the Fahrenheit and Celsius scales represent a specified measure of distance—degrees—but have no natural zero. Thus we cannot take meaningful ratios; for example, we cannot say that  $50^{\circ}$  is twice as hot as  $25^{\circ}$ . Another example is SAT or GMAT scores. The scores can be used to rank students, but only differences between scores provide information on how much better one student performed over another; ratios make little sense. In contrast to ordinal data, interval data allow meaningful comparison of ranges, averages, and other statistics.

In business, data from survey scales, while technically ordinal, are often treated as interval data when numerical scales are associated with the categories (for instance,

A	B	C	D	E	F	G	H	I	J	K	L	
1	Daily Treasury Yield Curve Rates											
2												
3	Date	1 mo	3 mo	6 mo	1 yr	2 yr	3 yr	5 yr	7 yr	10 yr	20 yr	30 yr
4	1/2/2008	3.09	3.26	3.32	3.17	2.88	2.89	3.28	3.54	3.91	4.39	4.35
5	1/3/2008	3.19	3.24	3.29	3.13	2.83	2.85	3.26	3.54	3.91	4.41	4.37
6	1/4/2008	3.22	3.2	3.22	3.06	2.74	2.75	3.18	3.47	3.88	4.4	4.36
7	1/7/2008	3.27	3.27	3.29	3.11	2.76	2.76	3.16	3.46	3.86	4.37	4.34
8	1/8/2008	3.31	3.25	3.27	3.09	2.76	2.76	3.16	3.47	3.86	4.39	4.35
9	1/9/2008	3.34	3.22	3.22	3.04	2.69	2.69	3.1	3.4	3.82	4.35	4.32
10	1/10/2008	3.37	3.24	3.21	3.04	2.71	2.74	3.16	3.49	3.91	4.47	4.44

**FIGURE 1.4** Example of Time-Series, Multivariate Data (Portion of *Treasury Yield Rates*)

1 = poor, 2 = average, 3 = good, 4 = very good, 5 = excellent). Strictly speaking, this is not correct, as the “distance” between categories may not be perceived as the same (respondents might perceive a larger distance between poor and average than between good and very good, for example). Nevertheless, many users of survey data treat them as interval when analyzing the data, particularly when only a numerical scale is used without descriptive labels.

**4. Ratio data**, which have a natural zero. For example, dollar has an absolute zero. Ratios of dollar figures are meaningful. Thus, knowing that the Seattle region sold \$12 million in March while the Tampa region sold \$6 million means that Seattle sold twice as much as Tampa. Most business and economic data fall into this category, and statistical methods are the most widely applicable to them.

This classification is hierarchical in that each level includes all of the information content of the one preceding it. For example, ratio information can be converted to any of the other types of data. Interval information can be converted to ordinal or categorical data but cannot be converted to ratio data without the knowledge of the absolute zero point. Thus, a ratio scale is the strongest form of measurement.

The managerial implications of this classification are in understanding the choice and validity of the statistical measures used. For example, consider the following statements:

- Sales occurred in March (categorical).
- Sales were higher in March than in February (ordinal).
- Sales increased by \$50,000 in March over February (interval).
- Sales were 20% higher in March than in February (ratio).

A higher level of measurement is more useful to a manager because more definitive information describes the data. Obtaining ratio data can be more expensive than categorical data, especially when surveying customers, but it may be needed for proper analysis. Thus, before data are collected, consideration must be given to the type of data needed.

## STATISTICAL THINKING

The importance of applying statistical concepts to make good business decisions and improve performance cannot be overemphasized. **Statistical thinking** is a philosophy of learning and action for improvement that is based on the following principles:

- All work occurs in a system of interconnected processes.
- Variation exists in all processes.
- Better performance results from understanding and reducing variation.<sup>5</sup>

Work gets done in any organization through **processes**—systematic ways of doing things that achieve desired results. Understanding processes provides the context for determining the effects of variation and the proper type of action to be taken. Any process contains many sources of variation. In manufacturing, for example, different batches of material vary in strength, thickness, or moisture content. Cutting tools have inherent variation in their strength and composition. During manufacturing, tools experience wear, vibrations cause changes in machine settings, and electrical fluctuations cause variations in power. Workers may not position parts on fixtures consistently, and physical and emotional stress may affect workers’ consistency. In addition, measurement gauges and human inspection capabilities are not uniform, resulting in variation in measurements even when there is little variation in the true value. Similar phenomena occur in

<sup>5</sup>Galen Britz, Don Emerling, Lynne Hare, Roger Hoerl, and Janice Shade, “How to Teach Others to Apply Statistical Thinking,” *Quality Progress*, June 1997, 67–79.

service processes because of variation in employee and customer behavior, application of technology, and so on.

While variation exists everywhere, many managers do not often recognize it or consider it in their decisions. For example, if sales in some region fell from the previous year, the regional manager might quickly blame her sales staff for not working hard, even though the drop in sales may simply be the result of uncontrollable variation. How often do managers make decisions based on one or two data points without looking at the pattern of variation, see trends when they do not exist, or try to manipulate financial results they cannot truly control? Unfortunately, the answer is “quite often.” Usually, it is simply a matter of ignorance of how to deal with data and information. A more educated approach would be to formulate a theory, test this theory in some way, either by collecting and analyzing data or developing a model of the situation. Using statistical thinking can provide better insight into the facts and nature of relationships among the many factors that may have contributed to the event and enable managers to make better decisions.

In recent years, many organizations have implemented Six Sigma initiatives. **Six Sigma** can be best described as a business process improvement approach that seeks to find and eliminate causes of defects and errors, reduce cycle times and cost of operations, improve productivity, better meet customer expectations, and achieve higher asset use and returns on investment in manufacturing and service processes. The term *six sigma* is actually based on a statistical measure that equates to 3.4 or fewer errors or defects per million opportunities. Six Sigma is based on a simple problem-solving methodology—**DMAIC**, which stands for Define, Measure, Analyze, Improve, and Control—that incorporates a wide variety of statistical and other types of process improvement tools. Six Sigma has heightened the awareness and application of statistics among business professionals at all levels in organizations, and the material in this book will provide the foundation for more advanced topics commonly found in Six Sigma training courses.

## **Populations and Samples**

One of the most basic applications of statistics is drawing conclusions about populations from sample data. A **population** consists of all items of interest for a particular decision or investigation, for example, *all* married drivers over the age of 25 in the United States, *all* first-year MBA students at a college, or *all* stockholders of Google. It is important to understand that a population can be anything we define it to be, such as all customers who have purchased from Amazon over the past year or individuals who do not own a cell phone. A company like Amazon keeps extensive records on its customers, making it easy to retrieve data about the entire population of customers with prior purchases. However, it would probably be impossible to identify all individuals who do not own cell phones. A population may also be an existing collection of items (for instance, all teams in the National Football League) or the potential, but unknown, output of a process (such as automobile engines produced on an assembly line).

A **sample** is a subset of a population. For example, a list of individuals who purchased a CD from Amazon in the past year would be a sample from the population of all customers who purchased from the company. Whether this sample is representative of the population of customers—which depends on how the sample data are intended to be used—may be debatable; nevertheless, it is a sample. Sampling is desirable when complete information about a population is difficult or impossible to obtain. For example, it may be too expensive to send all previous customers a survey. In other situations, such as measuring the amount of stress needed to destroy an automotive tire, samples are necessary even though the entire population may be sitting in a warehouse. Most of

the data files in Table 1.1 represent samples, although some, like the major league baseball data, represent populations.

We use samples because it is often not possible or cost-effective to gather population data. We are all familiar with survey samples of voters prior to and during elections. A small subset of potential voters, if properly chosen on a statistical basis, can provide accurate estimates of the behavior of the voting population. Thus, television network anchors can announce the winners of elections based on a small percentage of voters before all votes can be counted. Samples are routinely used for business and public opinion polls—magazines such as *Business Week* and *Fortune* often report the results of surveys of executive opinions on the economy and other issues. Many businesses rely heavily on sampling. Producers of consumer products conduct small-scale market research surveys to evaluate consumer response to new products before full-scale production, and auditors use sampling as an important part of audit procedures. In 2000, the U.S. Census began using statistical sampling for estimating population characteristics, which resulted in considerable controversy and debate.

**Statistics** are summary measures of population characteristics computed from samples. In business, statistical methods are used to present data in a concise and understandable fashion, to estimate population characteristics, to draw conclusions about populations from sample data, and to develop useful decision models for prediction and forecasting. For example, in the 2010 J.D. Power and Associates' Initial Quality Study, Porsche led the industry with a reported 83 problems per 100 vehicles. The number 83 is a statistic based on a sample that summarizes the total number of problems reported per 100 vehicles and suggests that the entire population of Porsche owners averaged less than one problem ( $83/100$  or 0.83) in their first 90 days of ownership. However, a particular automobile owner may have experienced zero, one, two, or perhaps more problems.

The process of collection, organization, and description of data is commonly called **descriptive statistics**. **Statistical inference** refers to the process of drawing conclusions about unknown characteristics of a population based on sample data. Finally, **predictive statistics**—developing predictions of future values based on historical data—is the third major component of statistical methodology. In subsequent chapters, we will cover each of these types of statistical methodology.

## USING MICROSOFT EXCEL

Spreadsheet software for personal computers has become an indispensable tool for business analysis, particularly for the manipulation of numerical data and the development and analysis of decision models. In this text, we will use Microsoft Excel 2010 for Windows to perform spreadsheet calculations and analyses. Some key differences exist between Excel 2010 and Excel 2007. We will often contrast these differences, but if you use an older version, you should be able to apply Excel easily to problems and exercises. In addition, we note that Mac versions of Excel do not have the full functionality that Windows versions have.

Although Excel has some flaws and limitations from a statistical perspective, its widespread availability makes it the software of choice for many business professionals. We do wish to point out, however, that better and more powerful statistical software packages are available, and serious users of statistics should consult a professional statistician for advice on selecting the proper software.

We will briefly review some of the fundamental skills needed to use Excel for this book. This is not meant to be a complete tutorial; many good Excel tutorials can be found online, and we also encourage you to use the Excel help capability (by clicking the question mark button at the top right of the screen).

## Basic Excel Skills

To be able to apply the procedures and techniques we will study in this book, it is necessary for you to know many of the basic capabilities of Excel. We will assume that you are familiar with the most elementary spreadsheet concepts and procedures:

- Opening, saving, and printing files
- Moving around a spreadsheet
- Selecting ranges
- Inserting/deleting rows and columns
- Entering and editing text, numerical data, and formulas
- Formatting data (number, currency, decimal places, etc.)
- Working with text strings
- Performing basic arithmetic calculations
- Formatting data and text
- Modifying the appearance of the spreadsheet
- Sorting data

Excel has extensive online help, and many good manuals and training guides are available both in print and online, and we urge you to take advantage of these. However, to facilitate your understanding and ability, we will review some of the more important topics in Excel with which you may or may not be familiar. Other tools and procedures in Excel that are useful in statistics, data analysis, or decision modeling will be introduced as we need them.

### SKILL-BUILDER EXERCISE 1.1

Sort the data in the Excel file *Automobile Quality* from lowest to highest number of problems per 100 vehicles using the sort capability in Excel.

Menus and commands in Excel 2010 reside in the “ribbon” shown in Figure 1.5. Menus and commands are arranged in logical *groups* under different *tabs* (*File*, *Home*, *Insert*, and so on); small triangles pointing downward indicate *menus* of additional choices. We will often refer to certain commands or options and where they may be found in the ribbon.

### Copying Formulas and Cell References

Excel provides several ways of copying formulas to different cells. This is extremely useful in building decision models, because many models require replication of formulas for different periods of time, similar products, and so on. One way is to select the cell with the formula to be copied, click the *Copy* button from the *Clipboard* group under the *Home* tab (or simply press Ctrl-C on your keyboard), click on the cell you wish to

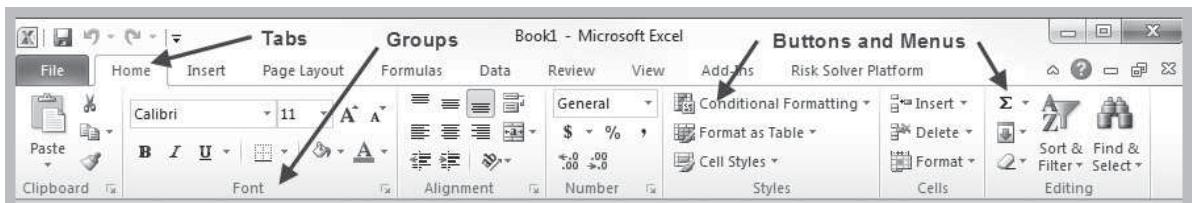


FIGURE 1.5 Excel 2010 Ribbon

	A	B	C	D	E
1	Total science and engineering jobs in thousands: 2000 and projected 2010				
2					
3	Occupation	2000	2010	Difference	
4	Scientists	3,241	5,301	2,060	
5	Life scientists	184	218	34	
6	Mathematical/computer scientists	2,408	4,308	1,900	
7	Computer specialists	2,318	4,213	1,895	
8	Mathematical scientists	89	95	6	
9	Physical scientists	239	283	44	
10	Social scientists	410	492	82	
11	Engineers	1,465	1,603	138	
12	All occupations	145,571	167,754		

**FIGURE 1.6** Copying Formulas by Dragging

copy to, and then click the *Paste* button (or press Ctrl-V). You may also enter a formula directly in a range of cells without copying and pasting by selecting the range, typing in the formula, and pressing Ctrl-Enter.

To copy a formula from a single cell or range of cells down a column or across a row, first select the cell or range, then click and hold the mouse on the small square in the lower right-hand corner of the cell (the “fill handle”), and drag the formula to the “target” cells you wish to copy to. To illustrate this technique, suppose we wish to compute the differences in projected employment for each occupation in the Excel file *Science and Engineering Jobs*. In Figure 1.6, we have added a column for the difference and entered the formula =C10-B10 in the first row. Highlight cell D4 and then simply drag the handle down the column. Figure 1.7 shows the results.

### SKILL-BUILDER EXERCISE 1.2

Modify the Excel file *Science and Engineering Jobs* to compute the percent increase in the number of jobs for each occupational category.

In any of these procedures, the *structure* of the formula is the same as in the original cell, but the cell references have been changed to reflect the *relative addresses* of the formula in the new cells. That is, the new cell references have the same relative relationship to the new formula cell(s) as they did in the original formula cell. Thus, if a formula is copied (or moved) one cell to the right, the relative cell addresses will have their column label increased by one; if we copy or move the formula two cells down, the row

	A	B	C	D	E
1	Total science and engineering jobs in thousands: 2000 and projected 2010				
2					
3	Occupation	2000	2010	Difference	
4	Scientists	3,241	5,301	2,060	
5	Life scientists	184	218	34	
6	Mathematical/computer scientists	2,408	4,308	1,900	
7	Computer specialists	2,318	4,213	1,895	
8	Mathematical scientists	89	95	6	
9	Physical scientists	239	283	44	
10	Social scientists	410	492	82	
11	Engineers	1,465	1,603	138	
12	All occupations	145,571	167,754		

**FIGURE 1.7** Results of Dragging Formulas

	A	B	C	D
3	Occupation	2000	2010	Difference
4	Scientists	3241	5301	=C4-B4
5	Life scientists	184	218	=C5-B5
6	Mathematical/computer scientists	2408	4308	=C6-B6
7	Computer specialists	2318	4213	=C7-B7
8	Mathematical scientists	89	95	=C8-B8
9	Physical scientists	239	283	=C9-B9
10	Social scientists	410	492	=C10-B10
11	Engineers	1465	1603	=C11-B11
12	All occupations	145571	167754	

**FIGURE 1.8** Formulas for *Science and Engineering Jobs Worksheet*

number is increased by 2. Figure 1.8 shows the formulas for the *Science and Engineering Jobs* spreadsheet example. For example, note that the formulas in each row are the same, except for the column reference.

Sometimes, however, you do not want to change the relative addressing because you would like all the copied formulas to point to a certain cell. We do this by using a \$ before the column and/or row address of the cell. This is called an *absolute address*. For example, suppose we wish to compute the percent of the total for each occupation for 2010. In cell E4, enter the formula =C4/\$C\$12. Then, if we copy this formula down column E for other months, the numerator will change to reference each occupation, but the denominator will still point to cell C12 (see Figure 1.9). You should be very careful to use relative and absolute addressing appropriately in your models.

## Functions

Functions are used to perform special calculations in cells. Some of the more common functions that we will use in statistical applications include the following:

MIN(*range*)—finds the smallest value in a range of cells

MAX(*range*)—finds the largest value in a range of cells

SUM(*range*)—finds the sum of values in a range of cells

AVERAGE(*range*)—finds the average of the values in a range of cells

COUNT(*range*)—finds the number of cells in a range that contain numbers

COUNTIF(*range, criteria*)—finds the number of cells within a range that meet specified criteria

Other more advanced functions often used in decision models are listed below:

AND(*condition 1, condition 2...*)—a logical function that returns TRUE if all conditions are true, and FALSE if not

	A	B	C	D	E
3	Occupation	2000	2010	Difference	Percent Increase
4	Scientists	3241	5301	=C4-B4	=C4/\$C\$12
5	Life scientists	184	218	=C5-B5	=C5/\$C\$12
6	Mathematical/computer scientists	2408	4308	=C6-B6	=C6/\$C\$12
7	Computer specialists	2318	4213	=C7-B7	=C7/\$C\$12
8	Mathematical scientists	89	95	=C8-B8	=C8/\$C\$12
9	Physical scientists	239	283	=C9-B9	=C9/\$C\$12
10	Social scientists	410	492	=C10-B10	=C10/\$C\$12
11	Engineers	1465	1603	=C11-B11	=C11/\$C\$12
12	All occupations	145571	167754		

**FIGURE 1.9** Example of Absolute Address Referencing

*OR(condition 1, condition 2...)*—a logical function that returns TRUE if any condition is true, and FALSE if not

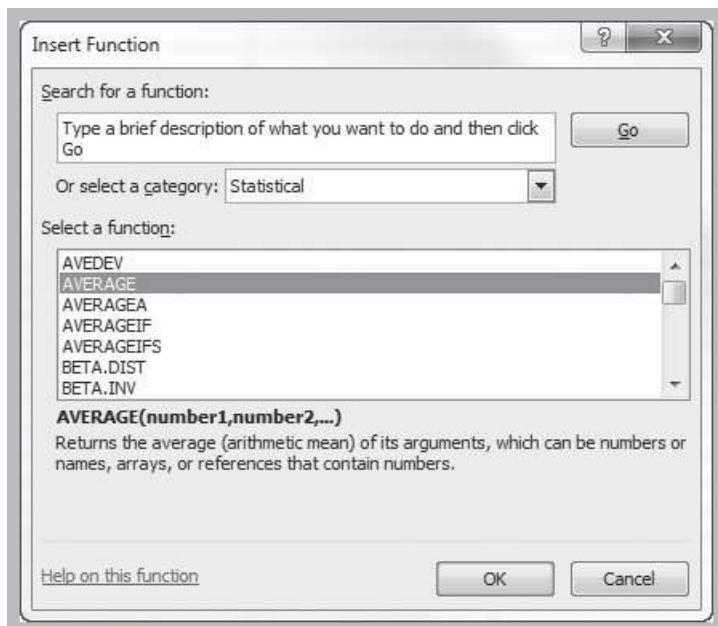
*IF(condition, value if true, value if false)*—a logical function that returns one value if the condition is true and another if the condition is false

*VLOOKUP(value, table range, column number)*—looks up a value in a table

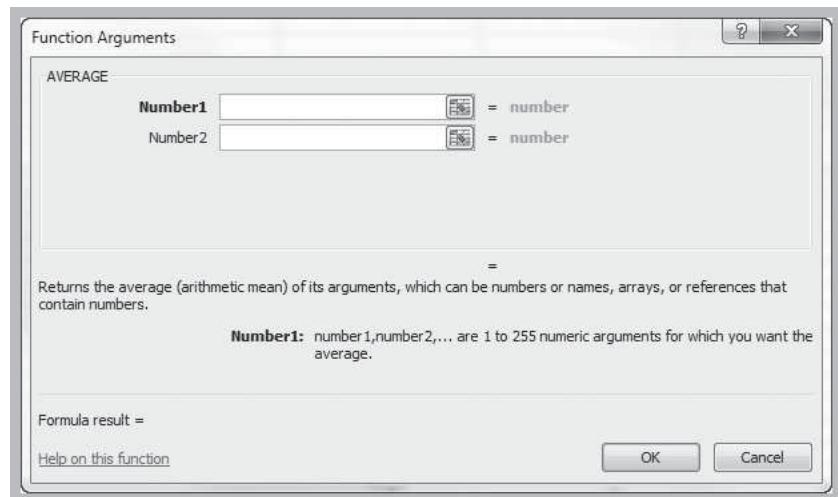
Excel has a wide variety of other functions for statistical, financial, and other applications, many of which we will use throughout the text. The easiest way to locate a particular function is to select a cell and click on the *Insert function* button [ $f_x$ ], which can be found under the ribbon next to the formula bar and also in the *Function Library* group in the *Formulas* tab. This is particularly useful even if you know what function to use but you are not sure of what arguments to enter. Figure 1.10 shows the dialog box from which you may select the function you wish to use, in this case, the AVERAGE function. Once this is selected, the dialog box in Figure 1.11 appears. When you click in an input cell, a description of the argument is shown. Thus, if you were not sure what to enter for the argument *number 1*, the explanation in Figure 1.11 will help you. For further information, you could click on the *Help on this function* link button in the lower left-hand corner.

The IF function, *IF(condition, value if true, value if false)*, allows you to choose one of two values to enter into a cell. If the specified *condition* is true, value A will be put in the cell. If the condition is false, value B will be entered. For example, if cell C2 contains the function =IF(A8=2,7,12), it states that if the value in cell A8 is 2, the number 7 will be assigned to cell C2; if the value in cell A8 is not 2, the number 12 will be assigned to cell C2. “Conditions” may include the following:

- = equal to
- > greater than
- < less than



**FIGURE 1.10** Insert Function Dialog



**FIGURE 1.11** Function Arguments Dialog for Average

- >= greater than or equal to
- <= less than or equal to
- <> not equal to

You may “nest” up to seven IF functions by replacing *value-if-true* or *value-if-false* in an IF function with another IF function, for example:

=IF(A8=2,(IF(B3=5,"YES","")),15)

This says that if cell A8 equals 2, then check the contents of cell B3. If cell B3 is 5, then the value of the function is the text string YES; if not, it is a blank space (a text string that is blank). However, if cell A8 is not 2, then the value of the function is 15 no matter what cell B3 is. You may use AND and OR functions as the *condition* within an IF function, for example: =IF(AND(B1=3,C1=5),12,22). Here, if cell B1 = 3 and cell C1 = 5, then the value of the function is 12, otherwise it is 22.

### SKILL-BUILDER EXERCISE 1.3

In the Excel file *Residential Electricity Data*, use Excel functions to find the maximum, minimum, and total for the Number of Consumers and Average Monthly Consumption for all census divisions.

### Other Useful Excel Tips

- **Split Screen.** You may split the worksheet horizontally and/or vertically to view different parts of the worksheet at the same time. The vertical splitter bar is just to the right of the bottom scroll bar, and the horizontal splitter bar is just above the right-hand scroll bar. Position your cursor over one of these until it changes shape, click, and drag the splitter bar to the left or down.
- **Paste Special.** When you normally copy (one or more) cells and paste them in a worksheet, Excel places an exact copy of the formulas or data in the cells (except for relative addressing). Often you simply want the *result* of formulas, so the data will remain constant even if other parameters used in the formulas change. To do this, use the *Paste Special* option found within the *Paste* menu in the *Clipboard*

group under the *Home* tab instead of the *Paste* command. Choosing *Paste Values* will paste the result of the formulas from which the data were calculated.

- **Column and Row Widths.** Many times a cell contains a number that is too large to display properly because the column width is too small. You may change the column width to fit the largest value or text string anywhere in the column by positioning the cursor to the right of the column label so that it changes to a cross with horizontal arrows, and then double-click. You may also move the arrow to the left or right to manually change the column width. You may change the row heights in a similar fashion by moving the cursor below the row number label. This can be especially useful if you have a very long formula to display. To break a formula within a cell, position the cursor at the break point in the formula bar and press Alt-Enter.
- **Displaying Formulas in Worksheets.** Choose *Show Formulas* in the *Formula Auditing* group under the *Formulas* tab. You will probably need to change the column width to display the formulas properly.
- **Displaying Grid Lines and Row and Column Headers for Printing.** Check the *Print* boxes for gridlines and headings in the *Sheet Options* group under the *Page Layout* tab. Note that the *Print* command can be found by clicking on the *Office* button.
- **Filling a Range with a Series of Numbers.** Suppose you want to build a worksheet for entering 100 data values. It would be tedious to have to enter the numbers from 1 to 100 one at a time. Simply fill in the first few values in the series and highlight them. Now click and drag the small square (fill handle) in the lower right-hand corner down (Excel will show a small pop-up window that tells you the last value in the range) until you have filled in the column to 100; then release the mouse.

## Excel Add-Ins

Microsoft Excel will provide most of the computational support required for the material in this book. Excel (Windows only) provides an add-in called the *Analysis Toolpak*, which contains a variety of tools for statistical computation, and *Solver*, which is used for optimization. These add-ins are not included in a standard Excel installation. To install them in Excel 2010, click the *File* tab and then *Options* in the left column. Choose *Add-Ins* from the left column. At the bottom of the dialog, make sure *Excel Add-ins* is selected in the *Manage:* box and click *Go*. In the *Add-Ins* dialog, if *Analysis Toolpak*, *Analysis Toolpak VBA*, and *Solver Add-in* are not checked, simply check the boxes and click *OK*. You will not have to repeat this procedure every time you run Excel in the future.

Four other add-ins available with this book provide additional capabilities and features not found in Excel and will be used in various chapters in this book. Prentice-Hall's *PHStat2* (which we will simply refer to as *PHStat*) add-in provides useful statistical support that extends the capabilities of Excel.<sup>6</sup> Refer to the installation procedures on the Companion Website. *PHStat* will be used in Chapters 1–8 and in Chapter 11. The student version of *Crystal Ball* provides a comprehensive set of tools for performing risk analysis simulations. *Crystal Ball* will be used in Chapter 10. *TreePlan* provides Excel support for decision trees and will be used in Chapter 11. Finally, Frontline Systems' *Risk Solver Platform*<sup>7</sup> provides a replacement (called *Premium Solver*) for the default *Solver* in Excel and will be used in Chapters 13 and 14. The Companion Website also includes an Excel workbook, *SimQuick-v2.xls*, which will be used for process simulation in Chapter 12.

<sup>6</sup>The latest version of *PHStat*, *PHStat2*, is included on the Companion Website. Enhanced versions and updates may be published on the *PHStat* Web site at [www.prenhall.com/phstat](http://www.prenhall.com/phstat). To date, *PHStat* is not available for Mac.

<sup>7</sup>*Risk Solver Platform* is a full-featured package that contains many other tools similar to other add-ins we use in this book. However, we will use only the *Premium Solver* component.

	A	B	C	D	E	F	G	H
1	Process Capability							
2								
3	5.21	5.87	4.85	4.95	5.07	4.96	4.96	5.11
4	5.02	5.33	4.82	4.86	4.82	4.96	5.06	5.11
5	4.90	5.11	5.02	5.13	5.03	4.94	4.86	5.08
6	5.00	5.07	4.90	4.95	4.85	5.19	4.96	5.03
7	5.16	4.93	4.73	5.22	4.89	4.91	4.99	4.94
8	5.03	4.99	5.04	4.81	4.82	5.01	4.94	4.88
9	4.96	5.04	5.07	4.91	5.18	4.93	5.06	4.91
10	5.04	5.14	4.81	4.95	5.02	5.05	4.95	4.86
11	4.98	5.09	5.04	4.94	5.05	4.96	5.02	4.89
12	5.07	5.06	5.03	4.81	4.88	4.92	5.01	4.91
13	5.02	4.85	5.01	5.11	5.08	4.95	5.04	4.87
14	5.08	4.93	5.14	4.81	4.98	5.08	5.01	4.93
15	4.85	5.04	5.12	4.97	5.02	4.97	5.02	5.14
16	4.90	5.09	4.89	5.07	4.99	5.04	5.03	4.87
17	4.97	5.07	4.91	5.03	5.02	4.94	5.18	4.98
18	5.09	4.99	4.97	4.81	5.03	4.98	5.08	4.88
19	4.89	5.01	4.98	4.95	5.02	5.03	5.14	4.88
20	4.87	4.88	5.01	4.89	5.07	5.05	4.92	5.01
21	5.01	4.93	5.01	5.08	4.95	4.91	4.97	4.93
22	4.97	5.10	5.09	4.93	4.95	5.09	4.92	4.93
23	4.76	4.94	4.93	4.99	4.94	5.21	5.14	4.99
24	4.94	4.88	5.04	4.94	5.12	4.87	4.92	4.91
25	4.92	4.89	5.11	5.13	5.08	5.02	5.03	4.96
26	4.91	4.89	5.07	5.02	4.91	4.81	4.98	4.78
27	4.96	5.02	5.13	5.13	4.92	4.98	4.89	4.88

**FIGURE 1.12** Excel Worksheet Process Capability

Throughout this book, we will provide many notes that describe how to use specific features of Microsoft Excel, *PHStat*, or other add-ins. These are summarized in chapter appendixes and are noted in the text by a margin icon when they will be useful to supplement examples and discussions of applications. It is important to read these notes and apply the procedures described in them in order to gain a working knowledge of the software features to which they refer.

We will illustrate the use of one of the *PHStat* procedures. In many cases, data on Excel worksheets may not be in the proper form to use a statistical tool. Figure 1.12, for instance, shows the worksheet *Process Capability* from the Excel file *Quality Measurements*, which we use for a case problem later in this book. Some tools in the *Analysis Toolpak* require that the data be organized in a single column in the worksheet. As a user, you have two choices. You can manually move the data within the worksheet, or you can use a utility from the *Data Preparation* menu in *PHStat* called *Stack Data* (see the note *Using the Stack Data and Unstack Data Tools* in the Appendix to this chapter).

The tool creates a new worksheet called “Stacked” in your Excel workbook, a portion of which is shown in Figure 1.13. If the original data columns have group labels (headers), then the column labeled “Group” will show them; otherwise, as in this example, the columns are simply labeled as Group1, Group2, and so on. In this example, Group1 refers to the data in the first column. If you apply the *Unstack Data* tool to the data in Figure 1.13, you will put the data in its original form.

#### SKILL-BUILDER EXERCISE 1.4

Use the *PHStat Stack* tool to stack the sample observations for the first shift in the Excel file *Syringe Samples*. Then, modify the Excel file *Automobile Quality* to label each car brand as either Foreign or Domestic, use the *PHStat Unstack* tool to group them.



**Spreadsheet Note**

	A	B	C
1	Group	Value	
2	Group1	5.21	
3	Group1	5.02	
4	Group1	4.90	
5	Group1	5.00	
6	Group1	5.16	
7	Group1	5.03	
8	Group1	4.96	
9	Group1	5.04	
10	Group1	4.98	
11	Group1	5.07	

**FIGURE 1.13** Portion of Stacked Worksheet

## DISPLAYING DATA WITH EXCEL CHARTS

The Excel file *EEO Employment Report* provides data on the employment in the state of Alabama for 2006. Figure 1.14 shows a portion of this data set. Raw data such as these are often difficult to understand and interpret. Graphs and charts provide a convenient way to visualize data and provide information and insight for making better decisions.

Microsoft Excel provides an easy way to create charts within your spreadsheet (see the section on *Creating Charts in Excel* in Appendix 1.1). These include vertical and horizontal bar charts, line charts, pie charts, area charts, scatter plots, three-dimensional charts, and many other special types of charts. We generally will not guide you through every application but will provide some guidance for new procedures as appropriate.



Spreadsheet Note

### Column and Bar Charts

Excel distinguishes between vertical and horizontal bar charts, calling the former *column charts* and the latter *bar charts*. A *clustered column chart* compares values across categories using vertical rectangles; a *stacked column chart* displays the contribution of each value to the total by stacking the rectangles; and a *100% stacked column chart* compares the percentage that each value contributes to a total. An example of a clustered column chart is

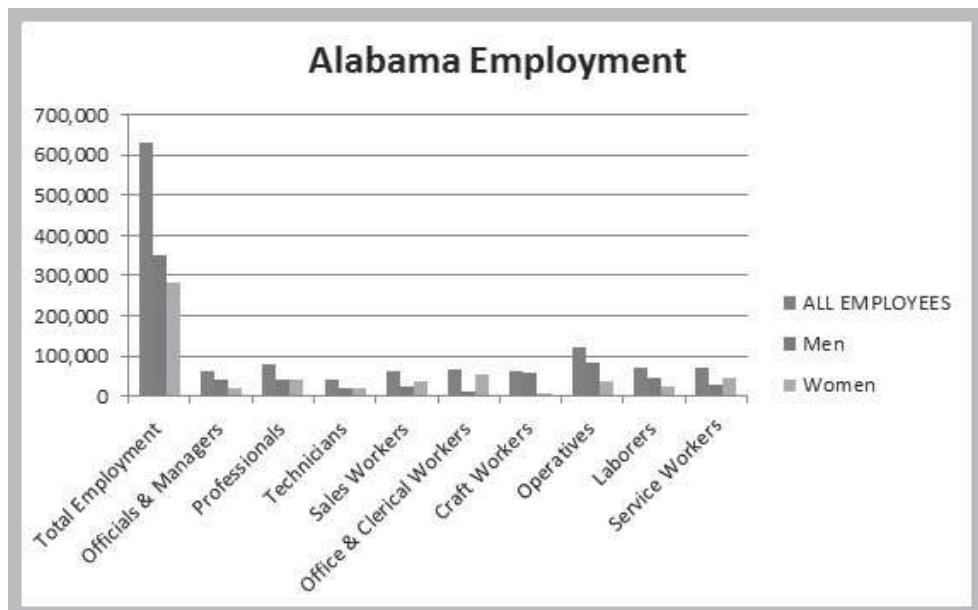
	A	B	C	D	E	F	G	
1	Equal Employment Opportunity Commission Report - Number Employed in State of Alabama, 2006							
2	Racial/Ethnic Group and Gender	Total Employment	Officials & Managers	Professionals	Technicians	Sales Workers	Office & Clerical Workers	
3	ALL EMPLOYEES	632,329	60,258	80,733	39,868	62,019	67,014	
4	Men	349,353	41,777	39,792	19,848	23,727	11,293	
5	Women	282,976	18,481	40,941	20,020	38,292	55,721	
6	WHITE	407,545	51,252	67,622	28,830	41,091	44,565	
7	Men	237,516	36,536	34,842	16,004	17,756	7,656	
8	Women	170,029	14,716	32,780	12,826	23,335	36,909	
9	MINORITY	224,784	9,006	13,111	11,038	20,928	22,449	
10	Men	111,837	5,241	4,950	3,844	5,971	3,637	
11	Women	112,947	3,765	8,161	7,194	14,957	18,812	

**FIGURE 1.14** Portion of EEO Commission Employment Report

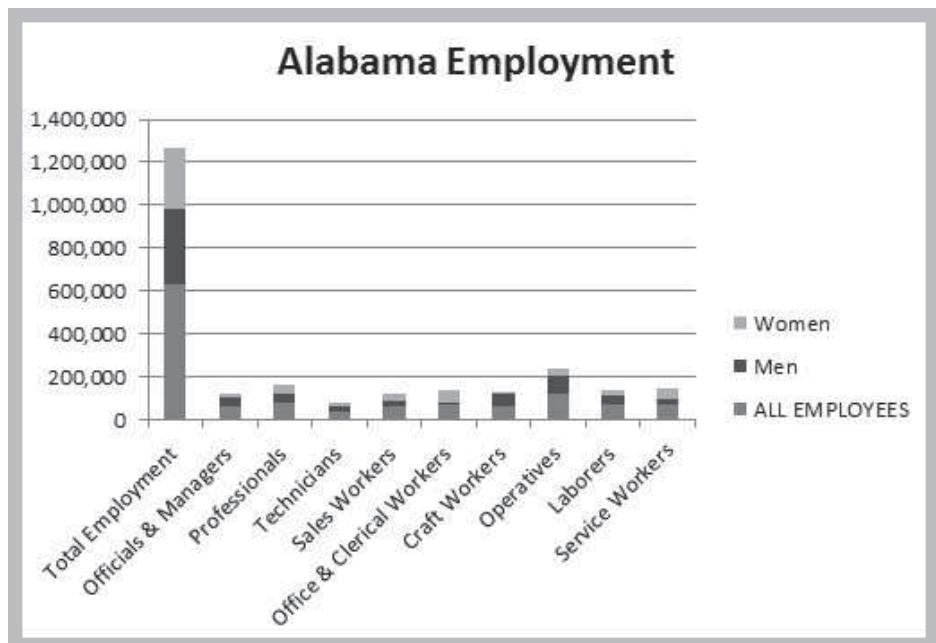
shown in Figure 1.15 for the Alabama employment data shown previously; Figure 1.16 shows a stacked column chart for the same data. Bar charts present information in a similar fashion, only horizontally instead of vertically.

### SKILL-BUILDER EXERCISE 1.5

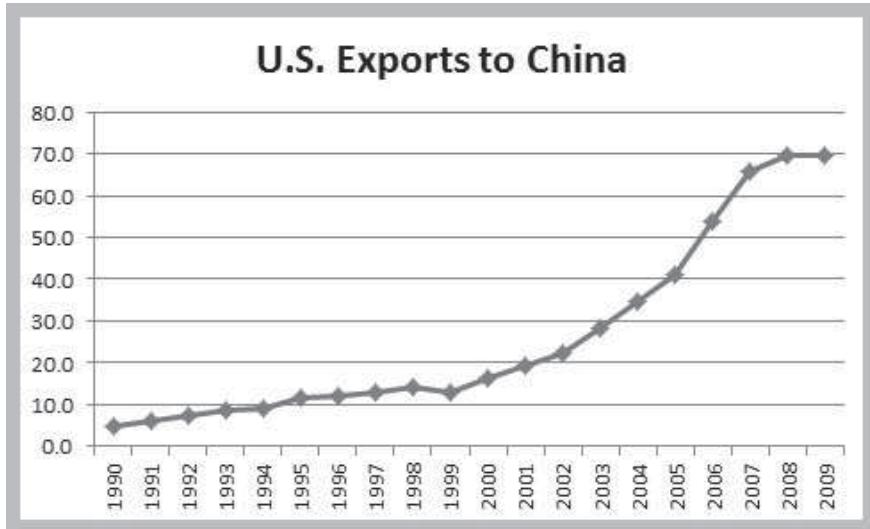
Create the column chart shown in Figure 1.15 for the *EEO Employment Report* data.



**FIGURE 1.15** Column Chart for Alabama Employment Data



**FIGURE 1.16** Stacked Column Chart



**FIGURE 1.17** Line Chart for U.S. to China Exports

## Line Charts

Line charts provide a useful means for displaying data over time. For instance, a line chart showing the amount of U.S. exports to China in billions of dollars from the Excel file *China Trade Data* is shown in Figure 1.17. The chart clearly shows a significant rise in exports starting in the year 2000, which began to level off around 2008. You may plot multiple data series in line charts; however, they can be difficult to interpret if the magnitude of the data values differs greatly. In this case, it would be advisable to create separate charts for each data series.

### SKILL-BUILDER EXERCISE 1.6

Create line charts for the closing prices in the Excel file *S&P 500*.

## Pie Charts

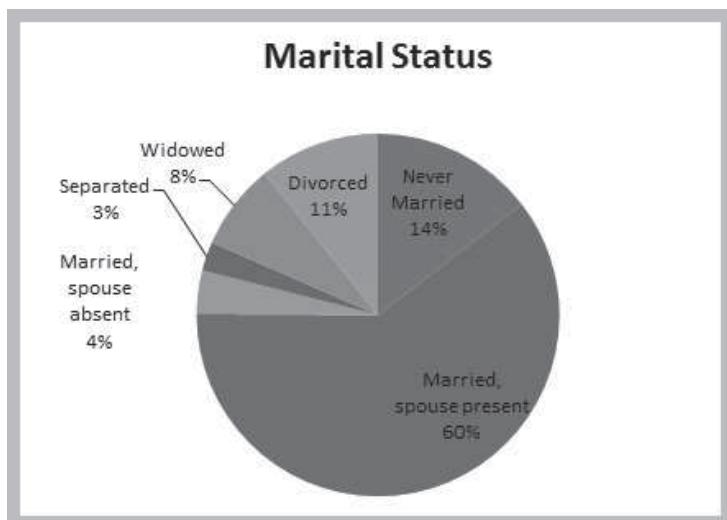
For many types of data, we are interested in understanding the relative proportion of each data source to the total. For example, consider the marital status of individuals in the U.S. population in the Excel file *Census Education Data*, a portion of which is shown in Figure 1.18. To show the relative proportion in each category, we can use a **pie chart**, as shown in Figure 1.19. This chart uses a layout option that shows the labels associated with the data, but not the actual values or proportions. A different layout that shows both can also be chosen.

### SKILL-BUILDER EXERCISE 1.7

Create a pie chart showing the breakdown of occupations in the *Science and Engineering Jobs* Excel file.

	A	B
18	<b>Marital Status</b>	
19	Never Married	25,752,000
20	Married, spouse present	107,008,000
21	Married, spouse absent	6,844,000
22	Separated	4,605,000
23	Widowed	13,577,000
24	Divorced	19,030,000

**FIGURE 1.18** Portion of *Census Education Data*



**FIGURE 1.19** Pie Chart for Marital Status

## Area Charts

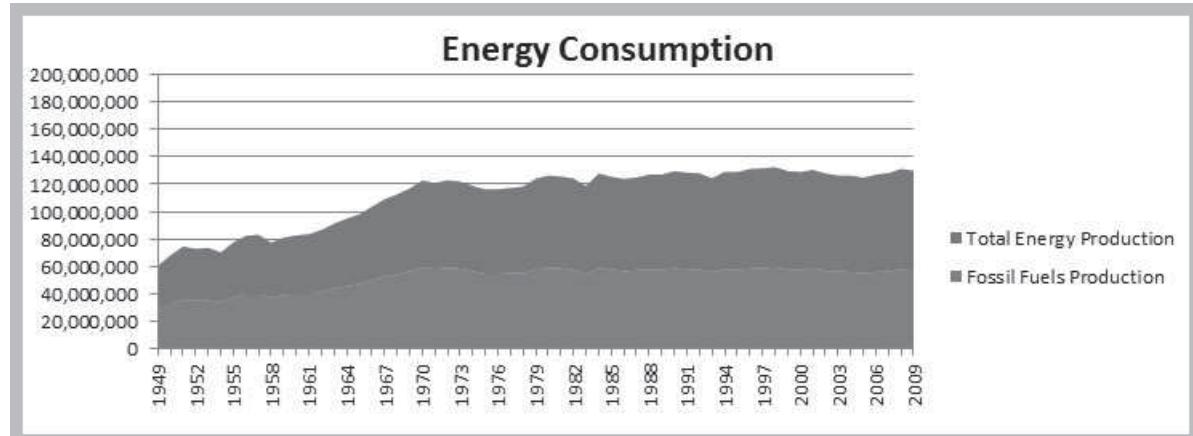
An **area chart** combines the features of a pie chart with those of line charts. For example, Figure 1.20 displays total energy consumption (billion Btu) and consumption of fossil fuels from the Excel file *Energy Production & Consumption*. This chart shows that while total energy consumption has grown since 1949, the relative proportion of fossil fuel consumption has remained generally consistent at about half of the total, indicating that alternative energy sources have not replaced a significant portion of fossil fuel consumption. Area charts present more information than pie or line charts alone but may clutter the observer's mind with too many details if too many data series are used; thus, they should be used with care.

## Scatter Diagrams

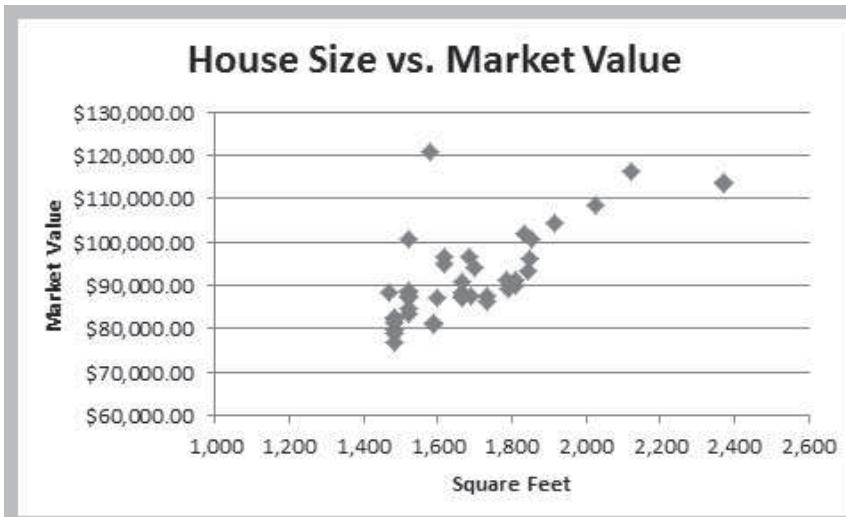
**Scatter diagrams** show the relationship between two variables. Figure 1.21 shows a scatter diagram of house size (in square feet) versus the home market value from the Excel file *Home Market Value*. The data show that higher market values are associated with larger homes. In Chapter 2 we shall see how to describe such a relationship numerically.

### SKILL-BUILDER EXERCISE 1.8

Create a scatter diagram showing the relationship between Hours online/week and Log-ins/day in the *Facebook Survey* data.



**FIGURE 1.20** Area Chart for Energy Consumption



**FIGURE 1.21** Scatter Diagram of House Size Versus Market Value

## Miscellaneous Excel Charts

Excel provides several additional charts for special applications (see Figure 1.22). A **stock chart** allows you to plot stock prices, such as the daily high, low, and close. It may also be used for scientific data such as temperature changes. A **surface chart** shows three-dimensional data. A **doughnut chart** is similar to a pie chart but can contain more than one data series. A **bubble chart** is a type of scatter chart in which the size of the data marker corresponds to the value of a third variable; consequently, it is a way to plot three variables in two dimensions. Finally, a **radar chart** allows you to plot multiple dimensions of several data series.

## Ethics and Data Presentation

In summary, tables of numbers often hide more than they inform. Graphical displays clearly make it easier to gain insights about the data. Thus, graphs and charts are a means of converting raw data into useful managerial information. However, it can be easy to distort data by manipulating the scale on the chart. For example, Figure 1.23 shows the U.S. exports to China in Figure 1.17 displayed on a different scale. The pattern looks much flatter and suggests that the rate of exports is not increasing as fast as it

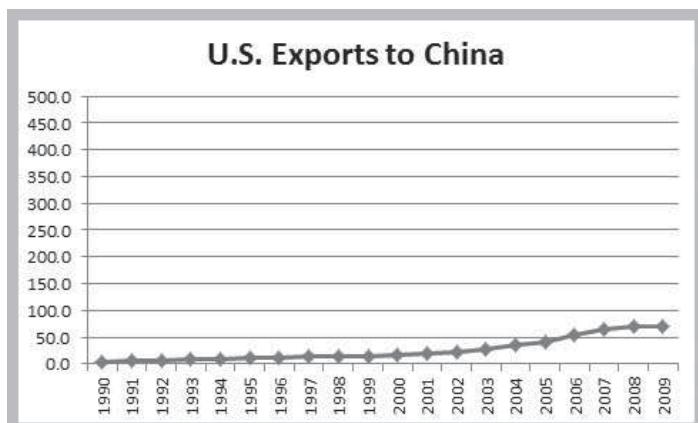


**FIGURE 1.22** Other Excel Charts

really is. It is not unusual to see distorted graphs in newspapers and magazines that are intended to support the author's conclusions. Creators of statistical displays have an ethical obligation to report data honestly and without attempts to distort the truth.

### SKILL-BUILDER EXERCISE 1.9

Create a bubble chart for the first five colleges in the Excel file *Colleges and Universities* for which the x-axis is the Top 10% HS, y-axis is Acceptance Rate, and bubbles represent the Expenditures per Student.



**FIGURE 1.23** An Alternate View of U.S. Exports to China

## Basic Concepts Review Questions

1. Explain the importance of statistics in business.
2. Explain the difference between data and information.
3. What do the terms “descriptive statistics” and “statistical inference” refer to?
4. What is a population? Why is sample information generally used to make conclusions about the population?
5. What is a metric, and how does it differ from a measure?
6. Explain the difference between a discrete and a continuous metric.
7. Explain the differences between categorical, ordinal, interval, and ratio data.
8. Explain the difference between cross-sectional and time-series data.
9. What is the Six Sigma initiative?
10. What is the difference between a population and a sample?
11. List the different types of charts available in Excel, and explain characteristics of data sets that make each chart most appropriate to use.
12. What kind of graphical displays are appropriate for representing the frequencies of several mutually exclusive categories? What are the options for representing such data over the levels of another variable, such as year, gender or race?

## Problems and Applications

1. For the Excel file *Surface Finish*, identify each of the variables as categorical, ordinal, interval, and ratio.
2. Consider the data given in the Excel file *MBA student survey*, and look at the variable Undergraduate Concentration.
  - a. Construct a column chart that visually represents the frequencies of each of the undergraduate concentrations.
  - b. Consider a pie chart showing the proportion of individuals in each concentration.
3. Construct a column chart for the data in the Excel file *Restaurant Sales* to allow a comparison of the delivery sales with the lunch sales and dinner sales. What other charts may be used to display this information?
4. Data from the 2000 U.S. Census show the following distribution of ages for residents of Ohio:

Total Households	4,445,773
Family households (families)	2,993,023
With own children under 18 years	1,409,912
Married-couple family	2,285,798
With own children under 18 years	996,042
Female householder, no husband present	536,878
With own children under 18 years	323,095
Nonfamily households	1,452,750
Householder living alone	1,215,614
Householder 65 years and over	446,396

- a. Construct a column chart to visually represent these data.
- b. Construct a stacked bar chart to display the sub categories where relevant. (Note that you will have to compute additional subcategories, for instance, under Family households, the number of families

- without children under 18, so that the total of the subcategories equals the major category total. The sum of all categories does not equal the total.)
- c. Construct a pie chart showing the proportion of households in each category.
5. The Excel file *Energy Production & Consumption* provides various energy data since 1949.
- a. Construct an area chart showing the fossil fuel consumption as a proportion of total energy consumption.
  - b. Construct line charts for total energy production and total energy imports.
  - c. Construct a scatter diagram for total energy consumption and total energy production.
6. Consider the information contained in the Excel file *Gas and Electric*.
- a. Construct appropriate charts to visually display this information.
  - b. What conclusions can you draw from these charts?
7. Consider the information contained in the Excel file *Mortgage Rates*.
- a. Construct appropriate charts to visually display this information.
  - b. What conclusions can you draw from these charts?
8. Construct whatever charts you deem appropriate to convey the information contained in the Excel file *Census Education Data*. What conclusions can you draw from these?
9. Construct whatever charts you deem appropriate to convey the information contained in the Excel file *Science and Engineering Jobs*. What conclusions can you draw from these?
10. Construct an appropriate chart to visually display the information contained in the Excel file *Coal Consumption*.
11. Modify the Excel file *Hi-Definition Television* to identify televisions that belong to the big screen/projection and LCD/Plasma categories. Use Excel functions to find the minimum and maximum values for the overall scores for each type of television. Count the number of televisions in the big screen/projection category and the LCD/Plasma category.

## Case

### A Data Collection and Analysis Project

Develop a simple questionnaire to gather data that include a set of both categorical variables and ratio variables. In developing the questionnaire, think about some meaningful questions that you would like to address using the data. The questionnaire should pertain to any subject of interest to you, for example, customer satisfaction with products or school-related issues, investments, hobbies, leisure activities, and so on—be creative! (Several Web sites provide examples of questionnaires that may help you. You might want to check out [www.samplequestionnaire.com](http://www.samplequestionnaire.com) or [www.examplequestionnaire.com](http://www.examplequestionnaire.com) for some ideas.) Aim for a total of 6–10 variables. Obtain a sample of at least 20 responses from fellow

students or coworkers. Record the data on an Excel worksheet and construct appropriate charts that visually convey the information you gathered, and draw any conclusions from your data. Then, as you learn new material in Chapters 2–7, apply the statistical tools as appropriate to analyze your data and write a comprehensive report that describes how you drew statistical insights and conclusions, including any relevant Excel output to support your conclusions. (Hint: a good way to embed portions of an Excel worksheet into a Word document is to copy it and then use the *Paste Special* feature in Word to paste it as a picture. This allows you to size the picture by dragging a corner.)

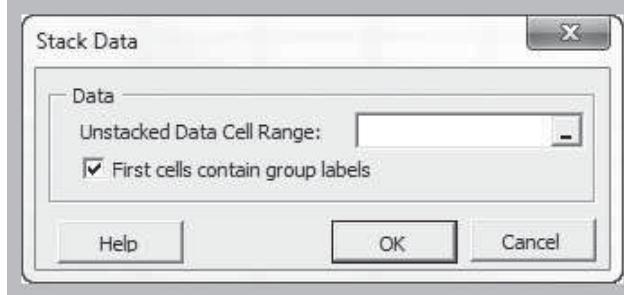
## APPENDIX 1.1

### Excel and PHStat Notes

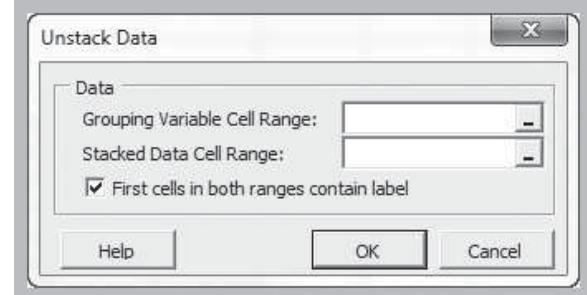
#### A. Using the PHStat Stack Data and Unstack Data Tools

From the *PHStat* menu, select *Data Preparation* then either *Stack Data* (to create a single column from multiple columns) or *Unstack Data* (to split a single column into multiple according to a grouping label). Figures 1A.1 and 1A.2 show

the dialog boxes that appear. To stack data in columns (with optional column labels), enter the range of the data in the *Unstacked Data Cell Range*. If the first row of the range contains a label, check the box *First cells contain group labels*. These labels will appear in the first column of the stacked data to help you identify the data if appropriate.



**FIGURE 1A.1** Stack Data Dialog



**FIGURE 1A.2** Unstack Data Dialog

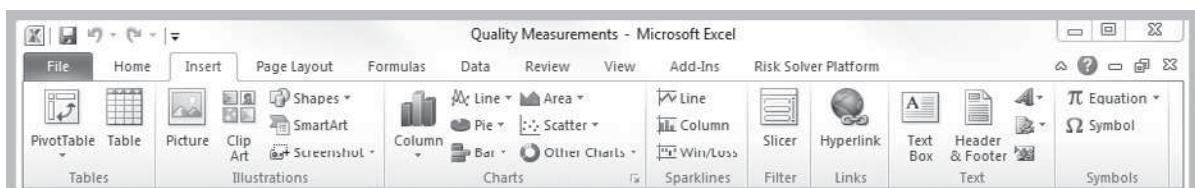
To unstack data in a single column and group them according to a set of labels in another column, enter the range of the column that contains the labels for the grouping variable in the *Grouping variable cell range* box and the range of the data in the *Stacked data cell range* box. If the top row contains descriptive labels, check the *First cells contain labels* box. This tool is useful when you wish to sort data into different groups.

## B. Creating Charts in Excel 2010

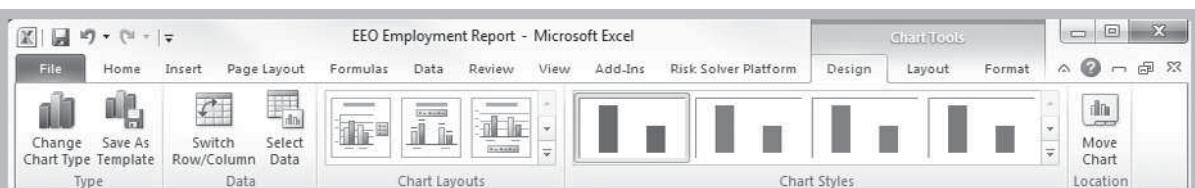
Excel provides a very comprehensive charting capability with many features. With a little experimentation, you can create very professional charts for business presentations. It is best to first highlight the range of the data you wish to chart. The Excel Help files provide guidance on formatting your data for a particular type of chart. Click the *Insert* tab in the Excel ribbon (Figure 1B.1). From the *Charts* group,

click the chart type, and then click a chart subtype that you want to use. Once a basic chart is created, you may use the options within the *Chart Tools* tabs to customize your chart (Figure 1B.2). In the *Design* tab, you can change the type of chart, data included in the chart, chart layout, and styles. From the *Layout* tab, you can modify the layout of titles and labels, axes and gridlines, and other features. The *Format* tab provides various formatting options. Many of these options can also be invoked by right-clicking on elements of the chart.

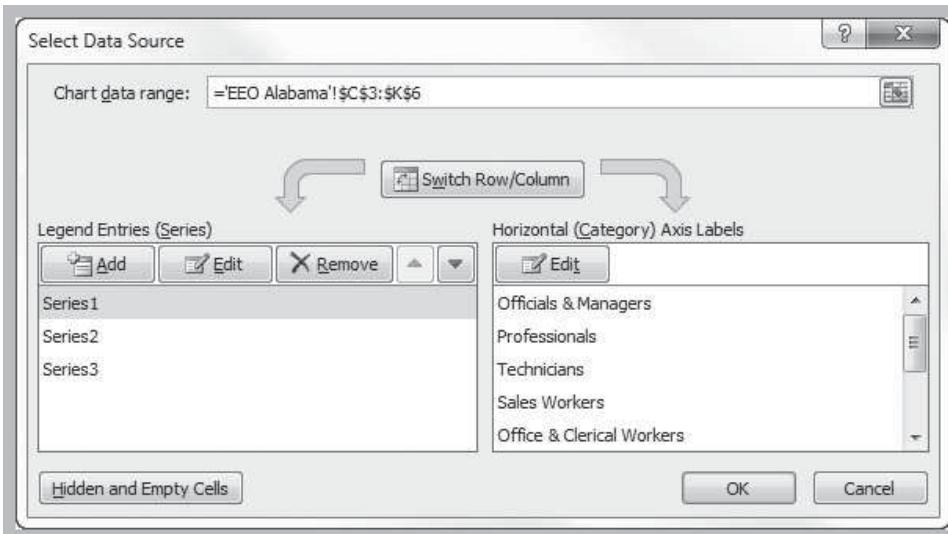
We will illustrate a simple bar chart for the various employment categories for all employees in the Excel file *EEO Employment Report*. First, highlight the range C3:K6, which includes the headings and data for each category. Click on the *Column Chart* button and then on the first chart type in the list (a clustered column chart). To add a title, click on the first icon in the *Chart Layouts* group. Click on "Chart Title" in the chart and change it to "EEO



**FIGURE 1B.1** Excel Insert Tab



**FIGURE 1B.2** Excel Chart Tools



**FIGURE 1B.3** Select Data Source Dialog

Employment Report—Alabama.” The names of the data series can be changed by clicking on the *Select Data* button in the *Data* group of the *Design* tab. In the *Select Data Source* dialog (see Figure 1B.3), click on “Series1” and then the *Edit* button. Enter the name of the data series, in this case “All

Employees.” Change the names of the other data series to “Men” and “Women” in a similar fashion. You can also change the order in which the data series are displayed on the chart using the up and down buttons. The final chart is shown in Figure 1.15.

## *Chapter 2*

# Descriptive Statistics and Data Analysis

- INTRODUCTION 56
- DESCRIPTIVE STATISTICS 56
- FREQUENCY DISTRIBUTIONS, HISTOGRAMS, AND DATA PROFILES 57
  - Categorical Data 58
  - Numerical Data 58
  - Data Profiles 62
- DESCRIPTIVE STATISTICS FOR NUMERICAL DATA 63
  - Measures of Location 63
  - Measures of Dispersion 64
  - Measures of Shape 67
  - Excel Descriptive Statistics Tool 68
  - Measures of Association 69
- DESCRIPTIVE STATISTICS FOR CATEGORICAL DATA 71
- VISUAL DISPLAY OF STATISTICAL MEASURES 73
  - Box Plots 73
  - Dot-Scale Diagrams 73
  - Outliers 74
- DATA ANALYSIS USING PIVOTTABLES 74
- BASIC CONCEPTS REVIEW QUESTIONS 78
- PROBLEMS AND APPLICATIONS 78
- CASE: THE MALCOLM BALDRIGE AWARD 81
- APPENDIX 2.1: DESCRIPTIVE STATISTICS: THEORY AND COMPUTATION 83
  - A. Mean, Variance, and Standard Deviation 83
  - B. Statistical Measures for Grouped Data 84
  - C. Skewness and Kurtosis 84
  - D. Correlation 85
- APPENDIX 2.2: EXCEL AND *PHSTAT* NOTES 85
  - A. Creating a Frequency Distribution and Histogram 85
  - B. Using the Descriptive Statistics Tool 85
  - C. Using the Correlation Tool 86
  - D. Creating Box Plots 87
  - E. Creating PivotTables 87
  - F. One- and Two-Way Tables and Charts 87

## INTRODUCTION

In Chapter 1, we discussed the role of data in modern organizations and how data can be visualized using charts. In this chapter, we discuss how to effectively summarize data quantitatively and perform some basic analyses for useful managerial information and insight. Our focus is on learning how to understand and incorporate these tools to make better decisions, as well as becoming proficient with the capabilities of Microsoft Excel.

## DESCRIPTIVE STATISTICS

The Excel file *Facebook Survey* provides data from a sample of students about their Facebook habits. The data include categorical variables for gender and number of views/day, and numerical variables for the number of hours spent online/week and their estimated number of friends (see Figure 2.1).

What can we learn from these data? We might want to get a “big picture” view of what the data tell us. For example, we might want to determine what is the typical number of friends, if any differences exist by gender, or how the number of views/day might be related to the number of friends. Statistical measures provide an effective and efficient way of obtaining meaningful information from data. **Descriptive statistics** refers to a collection of quantitative measures and ways of describing data. This includes *measures of central tendency* (mean, median, mode, proportion), *measures of dispersion* (range, variance, standard deviation), and *frequency distributions and histograms*.

Statistical support within Microsoft Excel can be accomplished in three ways:

1. Using statistical functions that are entered in worksheet cells directly or embedded in formulas.
2. Using the Excel *Analysis Toolpak* add-in to perform more complex statistical computations.<sup>1</sup>
3. Using the *Prentice-Hall* statistics add-in, *PHStat*, to perform analyses not designed into Excel.

Table 2.1 summarizes many of the descriptive statistics functions and tools that we will use in this chapter. One important point to note about the use of the tools in the *Analysis Toolpak* versus Excel functions is that while functions dynamically change as

	A	B	C	D	E
1	Facebook Survey				
2					
3	Student	Gender	Views/day	Hours online/week	Friends
4	1	female	6-10	4	150
5	2	female	11-15	10	400
6	3	male	1-5	7	120
7	4	male	21-25	15	500
8	5	female	11-15	9	260
9	6	female	1-5	5	70
10	7	female	1-5	7	90
11	8	male	6-10	5	250
12	9	female	11-15	12	110
13	10	female	1-5	2	30

**FIGURE 2.1** Portion of Excel File *Facebook Survey*

<sup>1</sup> Note to Mac users: Excel for the Mac does not support the *Analysis Toolpak*. Some of these procedures are available in the free edition of StatPlus:mac LE ([www.analystsoft.com](http://www.analystsoft.com)). A more complete version, StatPlus:mac Pro, can also be purchased. Some significant differences, however, exist in the tools between the versions.

**TABLE 2.1** Excel Statistical Functions and Tools

	<b>Description</b>
<b>Excel 2010 Functions</b>	
AVERAGE( <i>data range</i> )	Computes the average value (arithmetic mean) of a set of data
MEDIAN( <i>data range</i> )	Computes the median (middle value) of a set of data
MODE.SNGL( <i>data range</i> )	Computes the single most frequently occurring value in a set of data
MODE.MULT( <i>data range</i> )	Computes the most frequently occurring values of a set of data
VAR.S( <i>data range</i> )	Computes the variance of a set of data, assumed to be a sample
VAR.P( <i>data range</i> )	Computes the variance of a set of data, assumed to be an entire population
STDEV.S( <i>data range</i> )	Computes the standard deviation of a set of data, assumed to be a sample
STDEV.P( <i>data range</i> )	Computes the standard deviation of a set of data, assumed to be an entire population
SKEW( <i>data range</i> )	Computes the skewness, a measure of the degree to which a distribution is not symmetric around its mean
PERCENTILE.INC( <i>array, k</i> )	Computes the <i>k</i> th percentile of data in a range, where <i>k</i> is in the range 0–1, inclusive
KURT( <i>data range</i> )	Computes the kurtosis, a measure of the peakedness or flatness of a distribution
QUARTILE.INC( <i>array, quart</i> )	Computes the quartile of a distribution, based on percentile values from 0 to 1, inclusive
COVARIANCE.P( <i>array1, array2</i> )	Computes the covariance, assuming population data
COVARIANCE.S( <i>array1, array2</i> )	Computes the covariance, assuming sample data
CORREL( <i>array1, array2</i> )	Computes the correlation coefficient between two data sets
<b>Analysis Toolpak Tools</b>	
Descriptive Statistics	Provides a summary of a variety of basic statistical measures
Histogram	Creates a frequency distribution and graphical histogram for a set of data
Rank and Percentile	Computes the ordinal and percentage rank of each value in a data set
Correlation	Computes the correlation coefficient between two data sets
<b>PHStat Add-In</b>	
Box-and-Whisker Plot	Creates a box-and-whisker plot of a data set
Stem-and-Leaf Display	Creates a stem-and-leaf display of a data set
Dot-Scale Diagram	Creates a dot-scale diagram of a data set
Frequency Distribution	Creates a table of frequency counts and percentage frequency values
Histogram & Polygons	Creates a frequency table and histogram and optional frequency polygons

the data in the spreadsheet are changed, the results of the *Analysis Toolpak* tools do not. For example, if you compute the average value of a range of numbers directly using the function AVERAGE(*range*), then changing the data in the range will automatically update the result. However, you would have to rerun the *Descriptive Statistics* tool after changing the data.

## FREQUENCY DISTRIBUTIONS, HISTOGRAMS, AND DATA PROFILES

A **frequency distribution** is a table that shows the number of observations in each of several nonoverlapping groups. We may construct frequency distributions for both categorical and numerical data.

**TABLE 2.2** Frequency Distribution of Views/Day

Views/Day	Frequency
1–5	9
6–10	13
11–15	5
16–20	3
21–25	3
Total	33

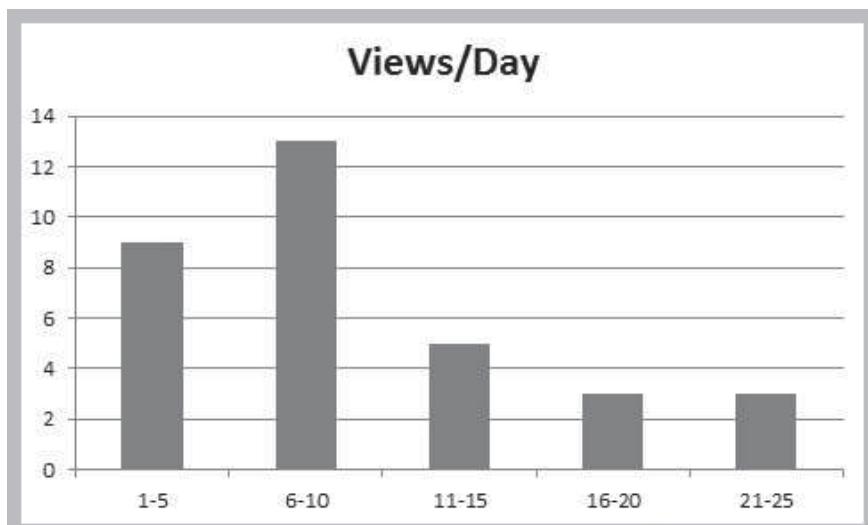
### Categorical Data

Categorical variables naturally define the groups in a frequency distribution; we need only count the number of observations that appear in each category. For the *Facebook Survey* data, for example, we may count the number of students who access Facebook 1–5 times, 6–10 times, and so on, using the Excel COUNTIF function. For instance, to count the number of students who view Facebook 1–5 times/day, use the function =COUNTIF(\$C\$4:\$C\$36,"1-5"). A frequency distribution for this categorical variable is shown in Table 2.2. We may construct a column chart to visualize these frequencies as shown in Figure 2.2.

We may express the frequencies as a fraction or proportion, of the total; this is called a **relative frequency distribution** and is shown in Table 2.3. Thus, the relative frequency of 1–5 views/day is calculated as  $9/33 = 0.273$ . Note that the sum of the relative frequencies must equal 1.0. A pie chart of the frequencies will visually show these proportions.

### Numerical Data

For numerical data that consist of a small number of discrete values, we may construct a frequency distribution similar to the way we did for categorical data, that is, simply count the frequencies of each discrete value. For example, in the Facebook data, all

**FIGURE 2.2** Column Chart of Views/Day

**TABLE 2.3** Relative Frequency Distribution

Views/Day	Frequency	Relative Frequency
1–5	9	0.273
6–10	13	0.394
11–15	5	0.152
16–20	3	0.091
21–25	3	0.091
Total	33	1.000

the numbers of hours online/week are whole numbers between 2 and 15. A frequency distribution for these data is shown in Table 2.4. A graphical depiction of a frequency distribution for numerical data in the form of a column chart is called a **histogram**. A histogram for hours online/week is shown in Figure 2.3.

Frequency distributions and histograms can be created using the *Analysis Toolpak* in Excel (see Appendix 2.2A, “Creating a Frequency Distribution and Histogram”). *PHStat* also provides tools for creating frequency distributions and histograms in the *Descriptive Statistics* menu option.

For numerical data that have many different discrete values with little repetition or are continuous, a frequency distribution requires that we define groups (called “bins” in Excel) by specifying the number of groups, the width of each group, and the upper and lower limits of each group. It is important to remember that the groups may not overlap so that each value is counted in exactly one group.

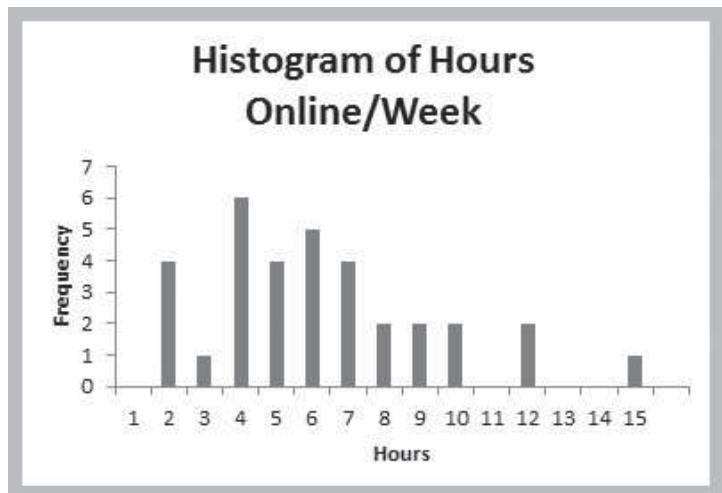
Figure 2.4 shows a portion of the *Facebook Survey* worksheet to which we added a bin range for Friends. Using these values, we are placing each value of Friends into groups from 0 to 50, more than 50 and up to 100, more than 100 and up to 150, and so on. Figure 2.5 shows the result of applying the Excel *Histogram* tool. The left column of the frequency distribution shows the upper limit of each cell in which the data fall. We see



Spreadsheet Note

**TABLE 2.4** Frequency Distribution for Hours Online/Week

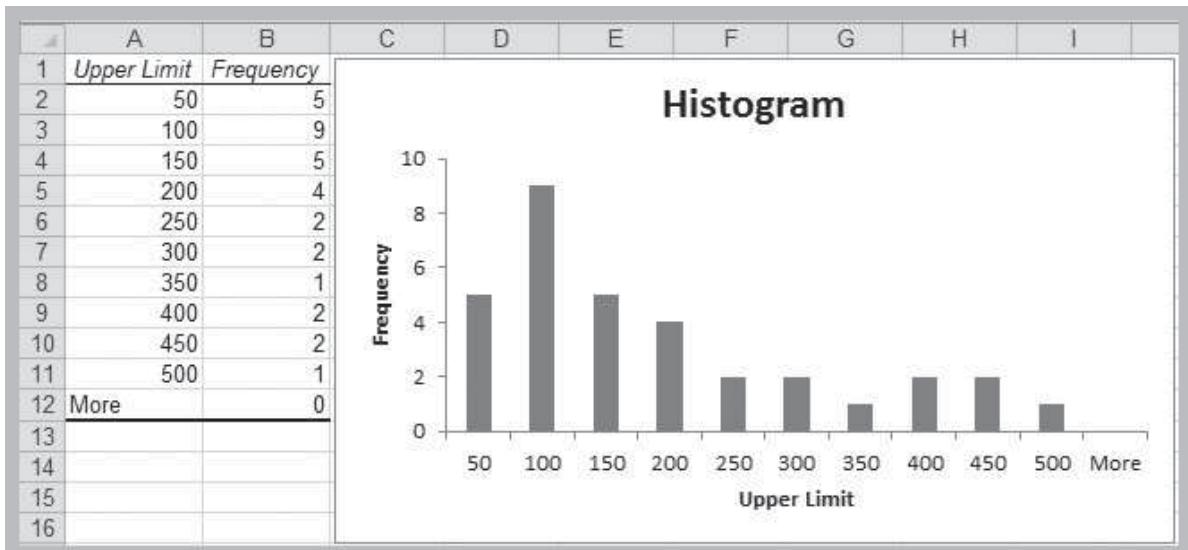
Hours Online/Week	Frequency
1	0
2	4
3	1
4	6
5	4
6	5
7	4
8	2
9	2
10	2
11	0
12	2
13	0
14	0
15	1
Total	33



**FIGURE 2.3** Histogram of Hours Online/Week

Facebook Survey					
					Bin Range
	Student	Gender	Views/day	Hours online/week	Friends
4	1	female	6-10	4	150
5	2	female	11-15	10	400
6	3	male	1-5	7	120
7	4	male	21-25	15	500
8	5	female	11-15	9	260
9	6	female	1-5	5	70
10	7	female	1-5	7	90
11	8	male	6-10	5	250
12	9	female	11-15	12	110
13	10	female	1-5	2	500

**FIGURE 2.4** Facebook Survey Worksheet with Bin Range



**FIGURE 2.5** Frequency Distribution and Histogram for Facebook Friends

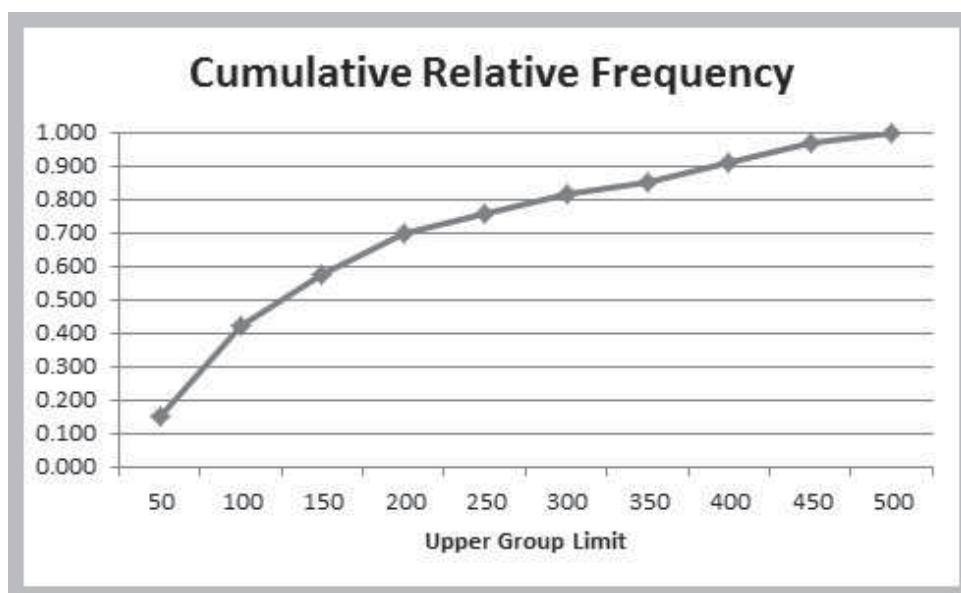
**TABLE 2.5** Relative and Cumulative Relative Frequencies for Facebook Friends

Upper Limit	Frequency	Relative Frequency	Cumulative Relative Frequency
50	5	0.152	0.152
100	9	0.273	0.424
150	5	0.152	0.576
200	4	0.121	0.697
250	2	0.061	0.758
300	2	0.061	0.818
350	1	0.030	0.848
400	2	0.061	0.909
450	2	0.061	0.970
500	1	0.030	1.000

that we have five observations below or equal to 50, nine observations greater than 50 and less than or equal to 100, and so on. The histogram shows that fewer students have large numbers of Facebook Friends and that most have 200 or less.

For numerical data, we may also compute the relative frequency called an ogive of observations in each group. By summing all the relative frequencies at or below each upper limit, we obtain the **cumulative relative frequency** (see Table 2.5). The cumulative relative frequency represents the proportion of the total sample that falls at or below the upper limit value. For example, we see that 0.697 or about 70% of students have 200 or fewer friends on Facebook. Note that since relative frequencies must be between 0 and 1 and must add up to 1, the cumulative frequency for the last group must equal 1.

Figure 2.6 shows a chart for the cumulative relative frequency, called an ogive. From this chart, you can easily estimate the proportion of observations that falls below a certain value. For example, you can see that about 70% of the data falls at or below 200 and about 90% of the data falls at or below 400, and so on.

**FIGURE 2.6** Chart of Cumulative Relative Frequency

## SKILL-BUILDER EXERCISE 2.1

---

Construct frequency distributions and histograms for the numerical data in the Excel file *Cell Phone Survey*. Also, compute the relative frequencies and cumulative relative frequencies. Plot the cumulative relative frequencies on a line chart similar to Figure 2.6.

A limitation of the Excel *Histogram* tool is that the frequency distribution and histogram are not linked to the data; thus, if you change any of the data, you must repeat the entire procedure to construct a new frequency distribution and histogram. An alternative is to use Excel's FREQUENCY function and the *Chart Wizard*. First, define the bins as you would by using the *Histogram* tool. Select the range of cells adjacent to the bin range and add one additional empty cell below it (this provides an overflow cell). Then enter the formula =FREQUENCY(*data range , bin range*) and press *Ctrl-Shift-Enter* simultaneously. This is necessary because FREQUENCY is an array function in Excel. This will create the frequency distribution. You may then construct a histogram using a column chart, customizing it as appropriate. Now, if the data are changed, the frequency distribution and histogram will be updated automatically. The following Skill-Builder exercises asks you to try this.

## SKILL-BUILDER EXERCISE 2.2

---

Use the FREQUENCY function to construct a frequency distribution for the data in the *Cell Phone Survey* Excel file.

## Data Profiles

Data are often expressed as *percentiles* and *quartiles*. You are no doubt familiar with percentiles from standardized tests used for college or graduate school entrance examinations (SAT, ACT, GMAT, GRE, etc.). Percentiles specify the percentage of other test takers who scored at or below the score of a particular individual. Generally speaking, the *kth percentile* is a value at or below which at least *k* percent of the observations lie. However, the way by which percentiles are calculated is not standardized. The most common way to compute the *kth percentile* is to order the *N* data values from smallest to largest and calculate the rank of the *kth percentile* using the formula  $Nk/100 + 0.5$ , round to the nearest integer, and then take the value corresponding to this rank as the *kth percentile*. For example, in the Facebook Friends data, the rank of the 90th percentile would be computed as  $33(90)/100 + 0.5 = 30.2$  or rounded, 30. Thus, the 30th ordered value is 400, and would be the 90th percentile. However, statistical software such as Excel uses different methods that often involve interpolating between ranks instead of rounding, thus producing different results. The Excel 2010 function PERCENTILE.INC(E4:E36,0.9), which is the same as the PERCENTILE function in older versions, calculates the 90th percentile as 396.<sup>2</sup>

Excel has a *Data Analysis* tool called *Rank and Percentile* for sorting data from high to low and computing percentiles associated with each value. You need only specify the range of the data to use the tool. For the Facebook Friends data, results are shown in Figure 2.7. You can see that the 90th percentile is close to 400.

---

<sup>2</sup> Older Excel functions will still work in Excel 2010.

	A	B	C	D
1	Point	Friends	Rank	Percent
2	4	500	1	100.00%
3	18	450	2	96.80%
4	26	430	3	93.70%
5	2	400	4	90.60%
6	25	380	5	87.50%
7	17	340	6	84.30%
8	22	280	7	81.20%
9	5	260	8	78.10%
10	8	250	9	75.00%
11	14	240	10	71.80%
12	13	200	11	68.70%
13	21	180	12	65.60%
14	28	170	13	59.30%
15	33	170	13	59.30%
16	1	150	15	53.10%
17	15	150	15	53.10%
18	3	120	17	46.80%
19	20	120	17	46.80%
20	9	110	19	43.70%
21	24	100	20	40.60%
22	7	90	21	31.20%
23	16	90	21	31.20%
24	29	90	21	31.20%
25	11	80	24	25.00%
26	27	80	24	25.00%
27	6	70	26	18.70%
28	32	70	26	18.70%
29	23	60	28	15.60%
30	19	50	29	6.20%
31	30	50	29	6.20%
32	31	50	29	6.20%
33	10	30	32	0.00%
34	12	30	32	0.00%

**FIGURE 2.7 Rank and Percentile Tool**  
Results

*Quartiles* represent the 25th percentile (called the first quartile,  $Q_1$ ), 50th percentile (second quartile,  $Q_2$ ), 75th percentile (third quartile,  $Q_3$ ), and 100th percentile (fourth quartile,  $Q_4$ ). One-fourth of the data is below the first quartile, and three-fourths of the data are below the third quartile. We may compute quartiles using the Excel 2010 function QUARTILE.INC (same as the older function QUARTILE). For example, in the Facebook Friends data, QUARTILE.INC(E4:E36,1) = 80 =  $Q_1$  and QUARTILE.INC(E4:E36,3) = 120 =  $Q_3$ . You can also identify these values in Figure 2.7.

We can extend these ideas to other divisions of the data. For example, *deciles* divide the data into 10 sets: the 10th percentile, 20th percentile, and so on. All of these types of measures are called **data profiles** (or **fractiles**).

## DESCRIPTIVE STATISTICS FOR NUMERICAL DATA

While frequency distributions and histograms provide a basic summary, we often need more information about data. Numerical measures of location, dispersion, shape, and association provide more specific quantitative information about data.

### Measures of Location

Measures of location provide estimates of a single value that in some fashion represents “centering” of the entire set of data. The most common is the *average*. We all use averages routinely in our lives, for example, to measure student accomplishment in a

class, to measure the performance of sports teams, and to measure performance in business. The average is formally called the **arithmetic mean** (or simply the *mean*), which is the sum of the observations divided by the number of observations. The mean of a population is denoted by the Greek letter  $\mu$ , and the mean of a sample is denoted by  $\bar{x}$ . Thus, for the *Facebook Survey* data, the sum of the data in column D for hours online/week is 206. By dividing this by the number of students (33), we compute the mean as  $\bar{x} = 206/33 = 6.24$ . We may also calculate the mean in Excel using the function `AVERAGE(data range)`. Using `=AVERAGE(E4:E36)` for the *Facebook Survey* data, we find that the mean number of friends in a student's network is 176.97. The mean is unique for every set of data and is meaningful for both interval and ratio data. However, it can be affected by **outliers**—observations that are radically different from the rest.

Another measure of location is the **median**, the middle value when the data are arranged from smallest to largest. For an odd number of observations, the median is the middle of the sorted numbers. For an even number of observations, the median is the mean of the two middle numbers. Note that the median is the same as the 50th percentile and the second quartile. We could have used the *Sort* option in Excel to rank order the values in the columns of the Facebook data and then determine the median. Since we have 33 observations, the median would be the 17th observation. Half the data are below the median, and half the data are above it. In Figure 2.7, in which the data are ranked, you can identify the 17th observation as 120. Thus, the median number of Facebook Friends is 120. The Excel function `MEDIAN(data range)` would also provide this result. The median is meaningful for ratio, interval, and ordinal data. As opposed to the mean, the median is *not* affected by outliers. In this case, the median is very close in value to the mean.

A third measure of location is the **mode**. The mode is the observation that occurs most frequently. The mode can easily be identified from a frequency distribution or histogram. For instance, in Table 2.4 and Figure 2.3, we see that the largest frequency of hours online/week is 6, which corresponds to a value of 4 hours/week. You may also use the Excel 2010 function `MODE.SNGL(data range)`, which is the same as the older `MODE` function. For frequency distributions and histograms of grouped data, the mode is the group with the largest frequency. Thus, in Figure 2.5, we see that more students have between 50 and 100 friends than any other group. The mode is most useful for data sets that consist of a relatively small number of unique values. When a data set has few repeating values, the mode is not very useful. Some data sets have multiple modes; to identify these you can use the Excel 2010 function `MODE.MULT(data range)`, which returns an array of modal values.

Another measure of location that some people use is the **midrange**. This is simply the average of the largest and smallest values in the data set. For the Facebook data, the maximum value for hours online/week is 15 and the minimum value is 2. Thus, the midrange is  $(15 + 2)/2 = 8.5$ . Caution must be exercised when using this statistic because extreme values easily distort the result. Note that the midrange uses only two pieces of data, while the mean uses *all* the data; thus, it is usually a much rougher estimate than the mean and is often used for only small sample sizes.

## Measures of Dispersion

**Dispersion** refers to the degree of variation in the data, that is, the numerical spread (or compactness) of the data. For instance, comparing the histograms of the Facebook data clearly shows more variation in Friends than in the number of hours online/week. Several statistical measures characterize dispersion: the *range*, *variance*, and *standard deviation*. The **range** is the simplest and is computed as the difference between the maximum value and the minimum value in the data set. Although Excel does not provide a function for the range, it can be computed easily by the formula `=MAX(data range) - MIN(data range)`.

Like the midrange, the range is affected by outliers and thus, is often only used for very small data sets. To avoid outliers, the **interquartile range (IQR)** is often used, which is simply  $Q_3 - Q_1$ . This is the range of the middle 50% of the data.

A more commonly used measure of dispersion is the **variance**, whose computation depends on *all* the data. The larger the variance, the more the data are “spread out” from the mean, and the more variability one can expect in the observations. The formula used for calculating the variance is different for populations and samples (we will discuss the theory behind this later in the chapter). The Excel function VAR.S(*data range*) may be used to compute the sample variance, which is denoted as  $s^2$ , while the Excel function VAR.P(*data range*) is used to compute the variance of a population, which is denoted as  $\sigma^2$ . A related measure, which is perhaps the most popular and useful measure of dispersion, is the **standard deviation**, which is defined as the square root of the variance. The Excel function STDEV.P(*data range*) calculates the standard deviation for a population ( $\sigma$ ); the function STDEV.S(*data range*) calculates it for a sample ( $s$ ). (Prior to Excel 2010, the functions VAR and STDEV applied to samples, and VARP and STDEVP applied to populations.)

Using these Excel functions for the *Facebook Survey* data, we find that the sample variance for hours online/week is 9.81 and the sample standard deviation is 3.13, and that the sample variance for the number of friends is 17996.78 and the sample standard deviation is 134.15.

**USING THE STANDARD DEVIATION** The standard deviation is generally easier to interpret than the variance because its units of measure are the same as the units of the data. Thus, it can be more easily related to the mean or other statistics measured in the same units. The standard deviation is a useful measure of risk, particularly in financial analysis. For example, the Excel file *Closing Stock Prices* (see Figure 2.8) lists daily closing prices for four stocks and the Dow Jones Industrial Average over a one-month period. The average closing price for Intel Corporation (INTC) and General Electric (GE) are quite similar, \$18.81 and \$16.19, respectively. However, the standard deviation of INTC’s price over this time frame was \$0.50, while GE’s was \$0.35. GE had less variability and, therefore, less risk. A larger standard deviation implies that while a greater potential exists of a higher return, there is also greater risk of realizing a lower return. Many investment publications and Web sites provide standard deviations of stocks and mutual funds to help investors assess risk in this fashion. We will learn more about risk in Part II of this book.

One of the more important results in statistics is **Chebyshev’s theorem**, which states that for *any set of data*, the proportion of values that lie within  $k$  standard deviations ( $k > 1$ ) of the mean is at least  $1 - 1/k^2$ . Thus, for  $k = 2$  at least three-fourths of the data lie within two standard deviations of the mean; for  $k = 3$  at least  $8/9$ , or 89% of the data lie within three standard deviations of the mean.

For many data sets encountered in practice, the percentages are generally much higher than Chebyshev’s theorem specifies. These are reflected in what are called the **empirical rules**:

1. Approximately 68% of the observations will fall within one standard deviation of the mean, or between  $\bar{x} - s$  and  $\bar{x} + s$ .
2. Approximately 95% of the observations will fall within two standard deviations of the mean, or within  $\bar{x} \pm 2s$ .
3. Approximately 99.7% of the observations will fall within three standard deviations of the mean, or within  $\bar{x} \pm 3s$ .

Depending on the data, the actual percentages may be higher or lower. For the *Facebook Survey* friends data, one standard deviation around the mean yields the interval [42.82, 311.12]. If we count the number of observations within this interval, we find that 25 of 33,

	A	B	C	D	E	F
1	Closing Stock Prices					
2						
3	Date	IBM	INTC	CSCO	GE	DJ Industrials
4	9/3/2010	127.58	18.43	21.04	15.392	10447.93
5	9/7/2010	125.95	18.12	20.58	15.44	10340.69
6	9/8/2010	126.08	17.9	20.64	15.7	10387.01
7	9/9/2010	126.36	18	20.61	15.91	10415.24
8	9/10/2010	127.99	17.97	20.62	15.98	10462.77
9	9/13/2010	129.61	18.557	21.26	16.25	10544.13
10	9/14/2010	128.85	18.74	21.45	16.16	10526.49
11	9/15/2010	129.43	18.72	21.59	16.34	10572.73
12	9/16/2010	129.67	18.97	21.93	16.23	10594.83
13	9/17/2010	130.19	18.81	21.863	16.29	10607.85
14	9/20/2010	131.79	18.93	21.75	16.55	10753.62
15	9/21/2010	131.98	19.14	21.64	16.52	10761.03
16	9/22/2010	132.57	19.01	21.67	16.5	10739.31
17	9/23/2010	131.67	18.98	21.53	16.14	10662.42
18	9/24/2010	134.11	19.423	22.09	16.66	10860.26
19	9/27/2010	134.65	19.235	22.11	16.43	10812.04
20	9/28/2010	134.89	19.505	21.863	16.44	10858.14
21	9/29/2010	135.48	19.24	21.87	16.36	10835.28
22	9/30/2010	134.14	19.2	21.9	16.25	10788.05
23	10/1/2010	135.64	19.32	21.91	16.36	10829.68
24	Mean	130.9315	18.81	21.4958	16.1951	10639.975
25	Standard Deviation	3.223518	0.499559	0.522015	0.3509	171.9448152

**FIGURE 2.8** Excel File *Closing Stock Prices*

or 75.8% fall within one standard deviation of the mean. A three standard deviation interval is  $[-205.4, 599.42]$ , and you can easily see that all data fall within it.

Two or three standard deviations around the mean are commonly used to describe the variability of most practical sets of data. For example, suppose that a retailer knows that on average, an order is delivered via standard ground transportation in 8 days with a standard deviation of 1 day. Using the second empirical rule, the retailer can therefore tell a customer with confidence that their package should arrive between 6 and 10 days. As another example, capability of a manufacturing process, which is characterized by the expected variation of output, is generally quantified as the mean  $\pm$  three standard deviations. This range is used in many quality control and Six Sigma applications.

### SKILL-BUILDER EXERCISE 2.3

---

Calculate the percentage of observations that fall within one, two, and three standard deviations of the mean for each of the stocks in the Excel file *Closing Stock Prices*. How do these compare with the empirical rules?

The **coefficient of variation (CV)** provides a relative measure of the dispersion in data relative to the mean and is defined as:

$$\text{CV} = \text{Standard Deviation}/\text{Mean}$$

(2.1)

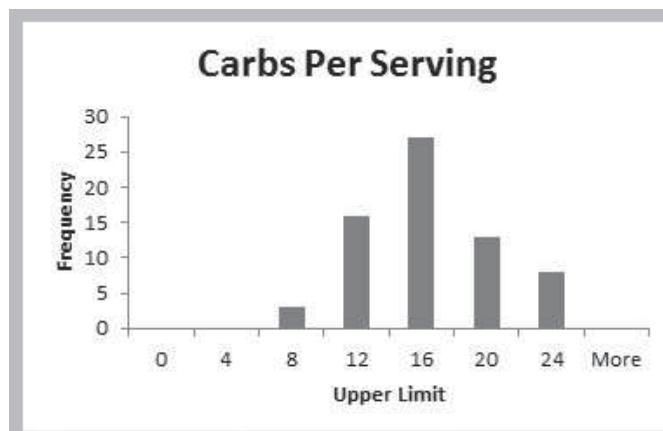
Sometimes the CV is multiplied by 100 to express it as a percentage. This statistic is useful for comparing the variability of two or more data sets when their scales differ. One practical application of the CV is in comparing stock prices. For example, by examining only the standard deviations in the *Closing Stock Prices* worksheet, we might conclude that IBM is more risky than the other stocks. However, the mean stock price of IBM is much larger than the other stocks. Thus, comparing standard deviations directly provides little information. The CV provides a more comparable measure. Using the data in Figure 2.8 for the Stock Price data, we calculate the coefficients of variation for IBM as 0.025, for INTC as 0.027, for Cisco as 0.024, for GE as 0.022, and for the Dow Jones industrial average (DJIA) as 0.016. We see that the coefficients of variation of the stocks are not very different; in fact, INTC is just slightly more risky than IBM relative to its average price. However, an index fund based on the Dow Industrials would be less risky than any of the individual stocks.

## Measures of Shape

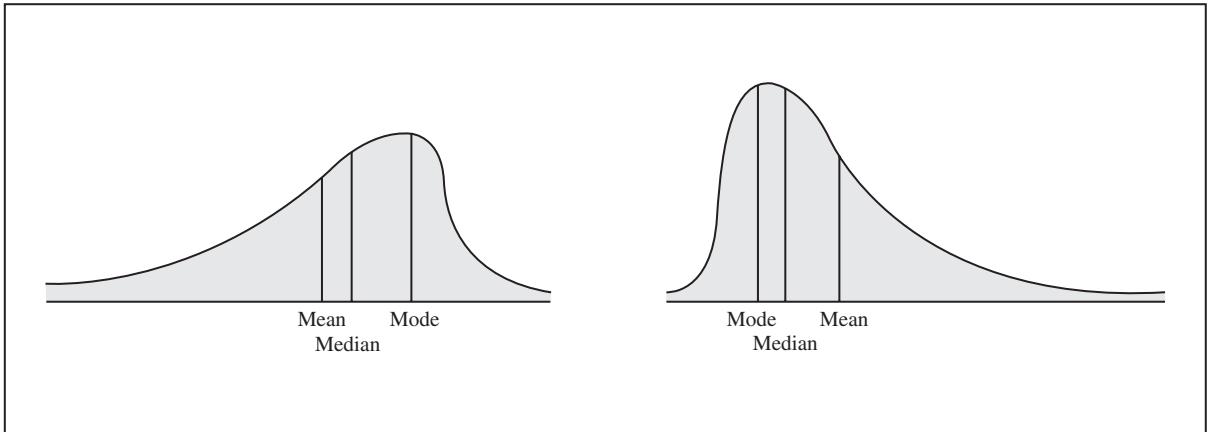
Histograms of sample data can take on a variety of different shapes. Figure 2.9 shows a histogram for carbohydrates per serving from the Excel file *Cereal Data*. Compare this to the histogram of the Facebook Friends data in Figure 2.5. The carbohydrate histogram is relatively symmetric, having its modal value in the middle and falling away from the center in roughly the same fashion on either side. The Facebook histogram is asymmetrical or *skewed*, that is, more of the mass is concentrated on one side and the distribution of values “tails off” to the other. Those that tail off to the right, like the Facebook example, are called *positively skewed*; those that tail off to the left are said to be *negatively skewed*.

The **coefficient of skewness (CS)**, which can be found using the Excel function `SKEW(data range)` measures the degree of asymmetry of observations around the mean. If CS is positive, the distribution of values is positively skewed; if negative, it is negatively skewed. The closer the CS is to 0, the lesser the degree of skewness. A CS greater than 1 or less than -1 suggests a high degree of skewness. A value between 0.5 and 1 or between -0.5 and -1 represents moderate skewness. Coefficients between 0.5 and -0.5 indicate relative symmetry. The CS for carbohydrate data is 0.15, while for the *Facebook Survey* friends data is 1.03, indicating positive skewness.

Histograms that have only one “peak,” such as Carbs per Serving, are called **unimodal**. (If a histogram has exactly two peaks, we call it **bimodal**. This often signifies a mixture of samples from different populations.) For unimodal histograms that are relatively symmetric, the mode is a fairly good estimate of the mean. For example,



**FIGURE 2.9** Histogram of Carbohydrates per Serving for Cereals



**FIGURE 2.10** Characteristics of Skewed Distributions

in Figure 2.9, we see that the mode occurs in the cell (12, 16); thus, the midpoint, 14, would be a good estimate of the mean (the true mean is 14.77). On the other hand, for the Facebook data in Figure 2.5, the mode occurs in the cell (50, 100). The midpoint estimate, 75, is not very close to the true mean of 176. Skewness pulls the mean away from the mode.

Comparing measures of location can sometimes reveal information about the shape of the distribution of observations. For example, if it were perfectly symmetrical and unimodal, the mean, median, and mode would all be the same. If it is negatively skewed, we would generally find that  $\text{mean} < \text{median} < \text{mode}$ , while a positive skewness would suggest that  $\text{mode} < \text{median} < \text{mean}$  (see Figure 2.10).

**Kurtosis** refers to the peakedness (i.e., high, narrow) or flatness (i.e., short, flat-topped) of a histogram. The **coefficient of kurtosis** (CK) measures the degree of kurtosis of a population and can be computed using the Excel function KURT(range). Distributions with values of CK less than 3 are more flat with a wide degree of dispersion; those with values of CK greater than 3 are more peaked with less dispersion. The higher the kurtosis, the more area the histogram has in the tails rather than in the middle. This measure has some applicability in risk analysis, but is generally not used very often in business applications.

### Excel Descriptive Statistics Tool

Excel provides a useful tool for basic data analysis, *Descriptive Statistics* (see Appendix 2.2B, “Using the Descriptive Statistics Tool”), which provides a summary of statistical measures that describe location, dispersion, and shape for sample data (not a population). Applying this tool to the *Facebook Survey* data, we obtain the results shown in Figure 2.11. The tool provides all the measures we have discussed except for one—the standard error—which we will discuss in Chapter 3, along with the minimum, maximum, sum and count.



Spreadsheet Note

#### SKILL-BUILDER EXERCISE 2.4

Use the *Descriptive Statistics* tool to summarize the numerical data in the Excel file *New Account Processing*.

A	B	C	D
1 Hours online/week		Friends	
2			
3 Mean	6.242424242	Mean	176.969697
4 Standard Error	0.545349316	Standard Error	23.35287946
5 Median		6 Median	120
6 Mode		4 Mode	90
7 Standard Deviation	3.132793313	Standard Deviation	134.152079
8 Sample Variance	9.814393939	Sample Variance	17996.7803
9 Kurtosis	0.682212964	Kurtosis	-0.018620284
10 Skewness	0.864609885	Skewness	1.031675419
11 Range		13 Range	470
12 Minimum		2 Minimum	30
13 Maximum		15 Maximum	500
14 Sum	206	Sum	5840
15 Count	33	Count	33

**FIGURE 2.11** Facebook Survey Data Descriptive Statistics Summary

## Measures of Association

Two variables have a strong statistical relationship with one another if they appear to move together. We see many examples on a daily basis; for instance, attendance at baseball games is often closely related to the win percentage of the team, and ice cream sales likely have a strong relationship with daily temperature. Figure 1.21 showed a scatter diagram that suggested that the house size was related to market value. When two variables appear to be related, you might suspect a cause-and-effect relationship. Sometimes, however, statistical relationships exist even though a change in one variable is not *caused* by a change in the other. For example, the *New York Times* reported a strong statistical relationship between the golf handicaps of corporate CEOs and their companies' stock market performance over three years. Chief executive officers (CEOs) who were better-than-average golfers were likely to deliver above-average returns to shareholders.<sup>3</sup> Clearly, the ability to golf would not cause better business performance. Therefore, you must be cautious in drawing inferences about causal relationships based solely on statistical relationships. (On the other hand, you might want to spend more time out on the practice range!)

Understanding the relationships between variables is extremely important in making good business decisions, particularly when cause-and-effect relationships can be justified. When a company understands how internal factors such as product quality, employee training, and pricing factors affect such external measures as profitability and customer satisfaction, it can make better decisions. Thus, it is helpful to have statistical tools for measuring these relationships.

The Excel file *Colleges and Universities*, a portion of which is shown in Figure 2.12, contains data from 49 top liberal arts and research universities across the United States. Several questions might be raised about statistical relationships among these variables. For instance, does a higher percentage of students in the top 10% of their high school class suggest a higher graduation rate? Is acceptance rate related to the amount spent per student? Do schools with lower acceptance rates tend to accept students with higher SAT scores? Questions such as these can be addressed by computing the correlation between the variables.

**Correlation** is a measure of a linear relationship between two variables, *X* and *Y*, and is measured by the **correlation coefficient**. The correlation coefficient is a number

<sup>3</sup> Adam Bryant, "CEOs' Golf Games Linked to Companies' Performance," *Cincinnati Enquirer*, June 7, 1998, E1.

	A	B	C	D	E	F	G
1	Colleges and Universities						
2							
3	School	Type	Median SAT	Acceptance Rate	Expenditures/Student	Top 10% HS	Graduation %
4	Amherst	Lib Arts	1315	22%	\$ 26,636	85	93
5	Barnard	Lib Arts	1220	53%	\$ 17,653	69	80
6	Bates	Lib Arts	1240	36%	\$ 17,554	58	88
7	Berkeley	University	1176	37%	\$ 23,665	95	68
8	Bowdoin	Lib Arts	1300	24%	\$ 25,703	78	90
9	Brown	University	1281	24%	\$ 24,201	80	90
10	Bryn Mawr	Lib Arts	1255	56%	\$ 18,847	70	84

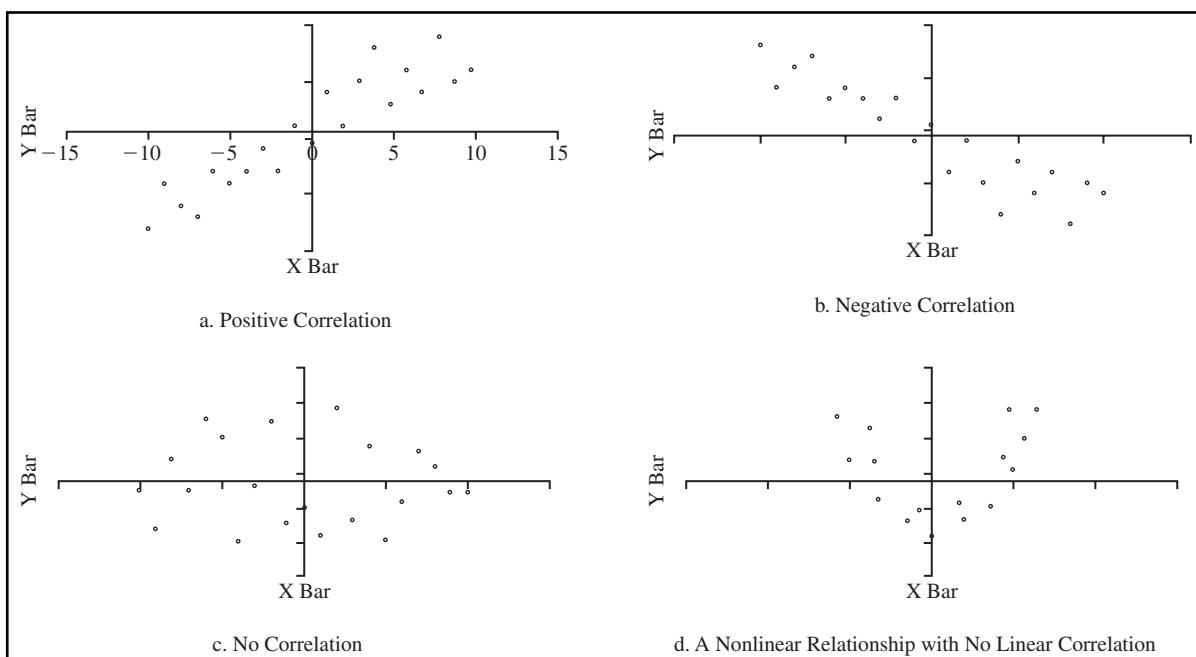
**FIGURE 2.12** Portion of Excel File Colleges and Universities

between  $-1$  and  $+1$ . A correlation of  $0$  indicates that the two variables have no linear relationship to each other. Thus, if one changes, we cannot reasonably predict what the other variable might do. A positive correlation coefficient indicates a linear relationship for which one variable increases as the other also increases. A negative correlation coefficient indicates a linear relationship for one variable that increases while the other decreases. In economics, for instance, a price-elastic product has a negative correlation between price and sales; as price increases, sales decrease, and vice versa. These relationships are illustrated in Figure 2.13. Note that although Figure 2.13(d) has a clear relationship between the variables, the relationship is not linear and the correlation is  $0$ .

Excel's CORREL function computes the correlation coefficient of two data arrays, and the *Data Analysis Correlation* tool computes correlation coefficients for more than two arrays (see Appendix 2.2C, "Using the Correlation Tool"). The correlation matrix among all the variables in the *Colleges and Universities* worksheet is shown in Figure 2.14. None of the correlations are very high; however, we see a moderate positive correlation



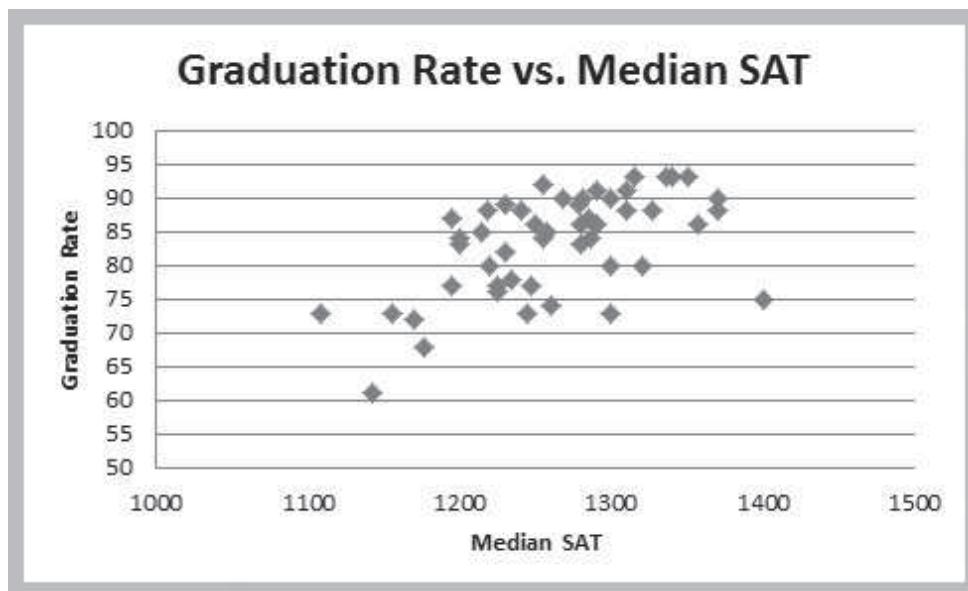
**Spreadsheet Note**



**FIGURE 2.13** Examples of Correlation

	A	B	C	D	E	F
1		Median SAT	Acceptance Rate	Expenditures/Student	Top 10% HS	Graduation %
2	Median SAT	1				
3	Acceptance Rate	-0.601901959	1			
4	Expenditures/Student	0.572741729	-0.284254415	1		
5	Top 10% HS	0.503467995	-0.609720972	0.505782049	1	
6	Graduation %	0.564146827	-0.55037751	0.042503514	0.138612667	1

**FIGURE 2.14** Correlation Results for Colleges and Universities Data



**FIGURE 2.15** Scatter Chart of Graduation Rate Versus Median SAT

between the graduation rate and SAT score (see Figure 2.15 for a chart of these data), indicating that schools with higher median SATs have higher graduation rates. We also see a moderate negative correlation between acceptance rate and graduation rate, indicating that schools with lower acceptance rates have higher graduation rates. We also see that acceptance rate is negatively correlated with the median SAT and Top 10% HS, suggesting that schools with lower acceptance rates have higher student profiles. The correlations with Expenditures/Student also suggest that schools with higher student profiles spend more money per student.

### SKILL-BUILDER EXERCISE 2.5

Compute the correlations between all pairs of numerical variables in the Excel file *Major League Baseball*. Drawing upon your knowledge of the game, explain why the results make sense.

## DESCRIPTIVE STATISTICS FOR CATEGORICAL DATA

Statistics such as means and variances are not appropriate for categorical data. Instead, we are generally interested in the fraction of data that have a certain characteristic. The formal statistical measure is called the **sample proportion**, usually denoted as  $p$ . Proportions

are key descriptive statistics for categorical data, such as defects or errors in quality control applications or consumer preferences in market research. For example, in the *Facebook Survey* Excel file, column B lists the gender of each respondent. The proportion of females is  $p = 20/33 = 0.606$ . The Excel function  $=\text{COUNTIF}(\text{data range}, \text{criteria})$  is useful in determining how many observations meet specified characteristics. For instance, to find the number of females, we used the function  $=\text{COUNTIF}(B4:B36, \text{"female"})$ . The criterion field can also be numerical, such as " $> 15$ " or " $= 0$ " and so on.

### SKILL-BUILDER EXERCISE 2.6

Use the COUNTIF function to find the proportions of respondents who use different types of cell phones in the *Cell Phone Survey* Excel file.

One of the most basic statistical tools used to summarize categorical data and examine the relationship between two categorical variables is cross-tabulation. A **cross-tabulation** is a tabular method that displays the number of observations in a data set for different subcategories of two categorical variables. A cross-tabulation table is often called a **contingency table**. The subcategories of the variables must be mutually exclusive and exhaustive, meaning that each observation can be classified into only one subcategory and, taken together over all subcategories, they must constitute the complete data set.

To illustrate, suppose we wish to identify the number of students who are in the different categories of views/day by Gender in the *Facebook Survey* data. A contingency table for these data is shown in Table 2.6. If we convert the data into the proportions of each views/day group for females and males, we obtain the result shown in Table 2.7. For example, this states that 30% of females are in the 1–5 views/day group, and that 20% of males are in the 1–5 group. Looking at the differences by gender, we see some differences, particularly in the 21–25 group, where the proportion of males is three times higher than the proportion of females. Cross-tabulations are commonly used in marketing research to provide insight into characteristics of different market segments using categorical variables such as gender, educational level, marital status, and so on.

**TABLE 2.6 A Contingency Table for Gender and Views/Day**

Gender	Views/Day					Total
	1–5	6–10	11–15	16–20	21–25	
Female	6	7	4	2	1	20
Male	3	6	1	1	2	13
<b>Total</b>	<b>9</b>	<b>13</b>	<b>5</b>	<b>3</b>	<b>3</b>	<b>33</b>

**TABLE 2.7 Proportions of Students in Views/Day Groups by Gender**

Gender	Views/Day					Total
	1–5	6–10	11–15	16–20	21–25	
Female	0.3	0.35	0.2	0.1	0.05	1
Male	0.2	0.46	0.08	0.08	0.15	1

## VISUAL DISPLAY OF STATISTICAL MEASURES

Statisticians use other types of graphs to visually display statistical measures. Two useful tools are *box plots* and *dot-scale diagrams*, both of which are available in the *PHStat* Excel add-in.

### Box Plots

**Box plots** (sometimes called **box-and-whisker plots**) graphically display five key statistics of a data set—the minimum, first quartile, median, third quartile, and maximum—and are very useful in identifying the shape of a distribution and outliers in the data. Box plots can be created in Excel using *PHStat* (see Appendix 2.2D, “Creating Box Plots”).

A box plot along with the five-number summary is shown in Figure 2.16 for friends in the *Facebook Survey* data. The “whiskers” extend on either side of the box to represent the minimum and maximum values in a data set, and the box encloses the first and third quartiles (the IQR), with a line inside the box representing the median. Very long whiskers suggest possible outliers in the data. Since the box is somewhat off center to the left and the median line is also slightly off center within the box, the distribution appears to be positively skewed (this was verified by the CS calculated earlier).



Spreadsheet Note

### Dot-Scale Diagrams

A **dot-scale diagram** is another visual display that shows a histogram of data values as dots corresponding to individual data points, along with the mean, median, first and third quartiles, and  $\pm 1, 2$ , and 3 standard deviation ranges from the mean. The mean essentially acts as a fulcrum as if the data were balanced along an axis. Figure 2.17 shows a dot-scale diagram for Facebook Friends data generated by *PHStat* from the *Descriptive Statistics* menu item. Dot-scale diagrams provide a better visual picture and understanding of the data than either box-and-whisker plots or stem-and-leaf displays.

Visual displays such as box-and-whisker plots and dot-scale diagrams give more complete pictures of data sets. They are highly useful tools in exploring the characteristics of data before computing other statistical measures.

### SKILL-BUILDER EXERCISE 2.7

Construct a box plot and dot-scale diagrams for the numerical data in the Excel file *Vacation Survey*.

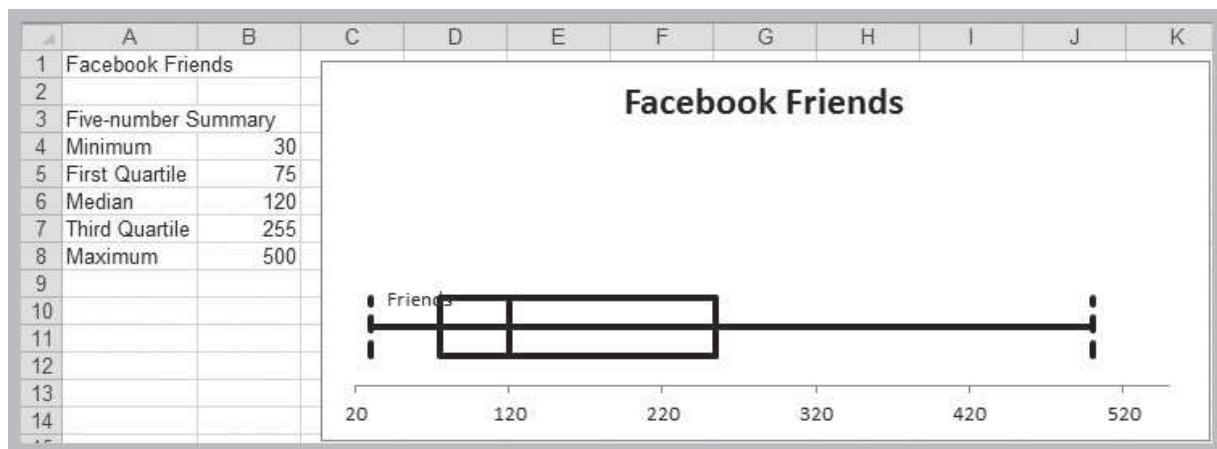
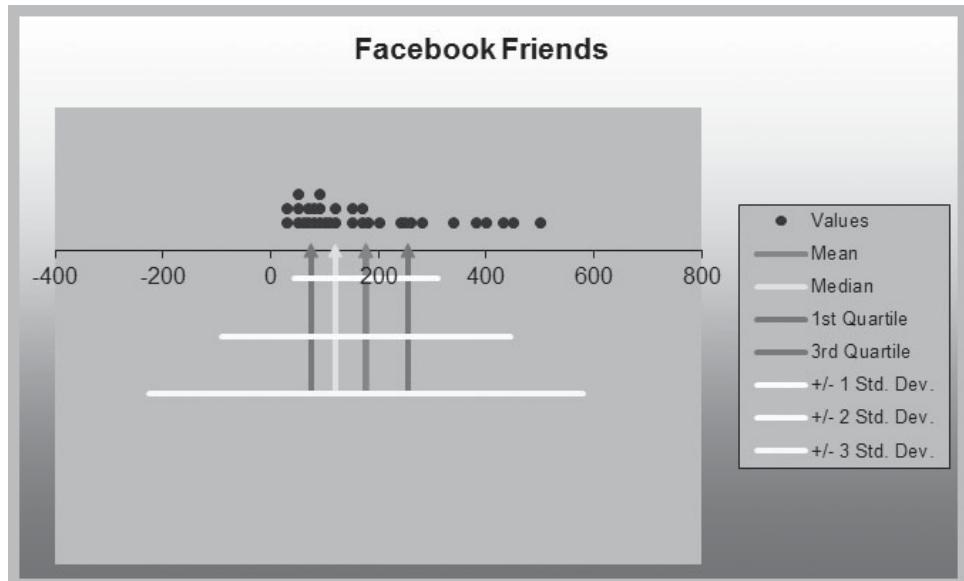


FIGURE 2.16 Box Plot and Five-Number Summary



**FIGURE 2.17** Dot-Scale Diagram

## Outliers

Earlier we had noted that the mean and range are sensitive to outliers in the data. Outliers can make a significant difference in the results we obtain from statistical analyses. An important statistical question is how to identify them. The first thing to do from a practical perspective is to check the data for possible errors, such as a misplaced decimal point or an incorrect transcription to a computer file. Box plots and dot-scale diagrams can help identify possible outliers visually. We might use the empirical rule to identify an outlier as one that is more than three standard deviations from the mean. We can also identify outliers based on the IQR. “Mild” outliers are often defined as being between  $1.5 \times \text{IQR}$  and  $3 \times \text{IQR}$  to the left of  $Q_1$  or to the right of  $Q_3$ , and “extreme” outliers as more than  $3 \times \text{IQR}$  away from these quartiles. Basically, there is no standard definition of what constitutes an outlier other than an unusual observation as compared with the rest.

While individual variables might not exhibit outliers, combinations of them might. We see this in Figure 1.21 for the *Home Market Value* data. The last observation has a high market value (\$120,700) but a relatively small house size (1581 square feet). The point on the scatter diagram does not seem to coincide with the rest of the data.

The question is what to do with outliers. They should not be blindly eliminated unless there is a legitimate reason for doing so—for instance, if the last home in the *Home Market Value* example has an outdoor pool that makes it significantly different from the rest of the neighborhood. Statisticians often suggest that analyses should be run with and without the outliers so that the results can be examined critically.

## DATA ANALYSIS USING PIVOTTABLES

Excel provides a powerful tool for distilling a complex data set into meaningful information: PivotTables. PivotTables allows you to create custom summaries and charts (see Chapter 2) of key information in the data. To apply PivotTables, you need a data set with column labels in the first row. The data set in the Excel file *Accounting Professionals*, shown in Figure 2.18, which provides the results of a survey of 27 employees in a tax division of a Fortune 100 company, satisfies this condition. Select any cell in the data set,

	A	B	C	D	E	F	G
1	Accounting Department Survey Data						
2	Employee	Gender	Years of Service	Years Undergraduate Study	Graduate Degree?	CPA?	Age Group
3	1	F	17	4	N	Y	41-45
4	2	F	6	2	N	N	26-30
5	3	M	8	4	Y	Y	31-35
6	4	F	8	4	Y	N	31-35
7	5	M	16	4	Y	Y	36-40
8	6	F	21	1	N	Y	51-55
9	7	M	27	4	N	N	51-55
10	8	F	7	4	Y	Y	26-30
11	9	M	8	4	N	N	31-35
12	10	M	23	2	N	Y	41-45

**FIGURE 2.18** Portion of Excel File Accounting Professionals

and choose *PivotTable* and *PivotChart Report* from the *Data* tab, and follow the steps of the wizard (see the note “Creating PivotTables” in Appendix 2.2). Excel then creates a blank PivotTable as shown in Figure 2.19.

You should first decide what types of tables you wish to create. For example, in the *Accounting Department Survey* data, suppose you wish to count the average number of years of service for males and females with and without a graduate degree. If you drag the variable *Gender* from the *PivotTable Field List* in Figure 2.19 to the *Row Labels* area, the variable *Graduate Degree?* into the *Column Labels* area, and the variable *Years of Service* into the *Values* area, then you will have created the PivotTable shown in Figure 2.20. However, the sum of years of service (default) is probably not what you would want.



Spreadsheet Note

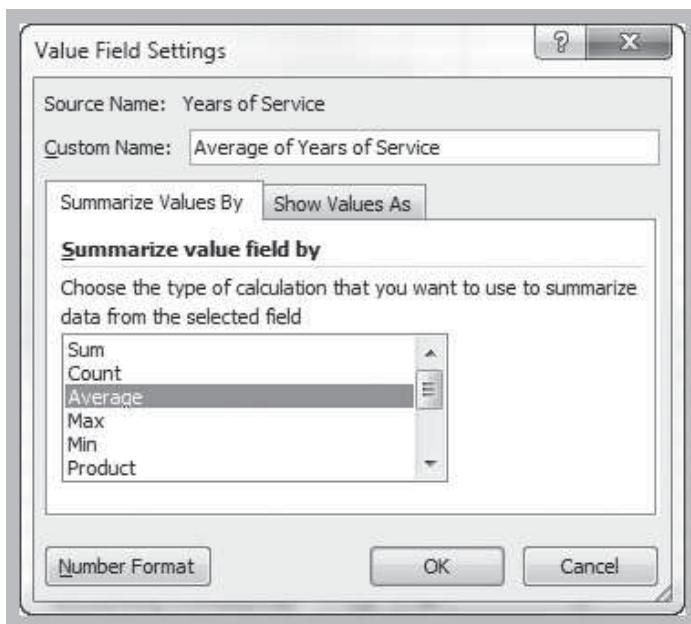
**FIGURE 2.19** Blank PivotTable

	A	B	C	D
1				
2				
3	Sum of Years of Service	Column Labels		
4	Row Labels	N	Y	Grand Total
5	F	95	46	141
6	M	168	88	256
7	Grand Total	263	134	397

**FIGURE 2.20** PivotTable for the Sum of Years of Service

In the *Options* tab under *PivotTable Tools* in the menu bar, click on the *Active Field* group and choose *Value Field Settings*. You will be able to change the summarization method in the PivotTable in the dialog shown in Figure 2.21. Selecting *Average* results in the PivotTable shown in Figure 2.22, we see that the average number of years of service is not much different for holders of graduate degrees, but that females have much fewer years of service than males.

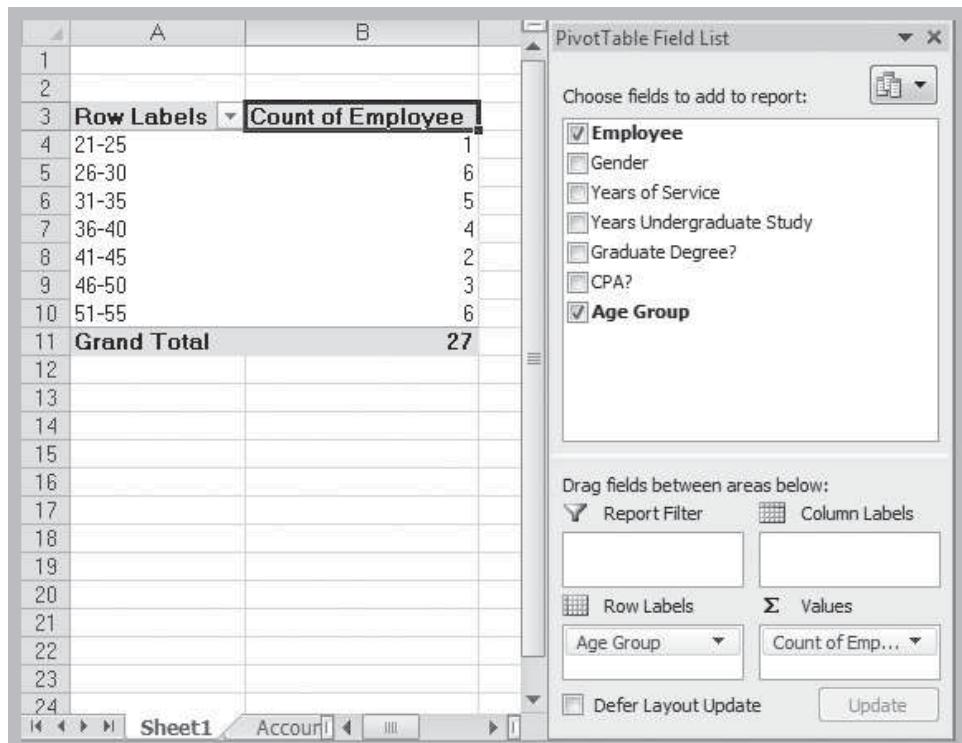
The beauty of PivotTables is that if you wish to change the analysis, you can simply uncheck the boxes in the *PivotTable Field List* or drag the variable names to different



**FIGURE 2.21** Value Field Settings Dialog

	A	B	C	D
1				
2				
3	Average of Years of Service	Column Labels		
4	Row Labels	N	Y	Grand Total
5	F	10.55555556	9.2	10.07142857
6	M	21	17.6	19.69230769
7	Grand Total	15.47058824	13.4	14.7037037

**FIGURE 2.22** PivotTable for Average Years of Service



**FIGURE 2.23** Count of Number of Employees by Age Group

field areas. You may easily add multiple variables in the fields to create different views of the data. Figure 2.23 shows a count of the number of employees by age group. The best way to learn about PivotTables is simply to experiment with them!

### SKILL-BUILDER EXERCISE 2.8

Create PivotTables to find the average number of years of undergraduate study for each age group with and without a CPA and the number of employees in each age group with and without a CPA in the Excel file *Accounting Professionals*.

PivotTables also provide an easy method of constructing cross-tabulations for categorical data. Figure 2.24 shows a PivotTable that was created by extending the one shown in Figure 2.23. Simply drag the field *Graduate Degree?* into the *Column Label* box in the *PivotTable Field List*. This PivotTable is a cross-tabulation of the number of employees in each age group who do or do not possess a graduate degree.

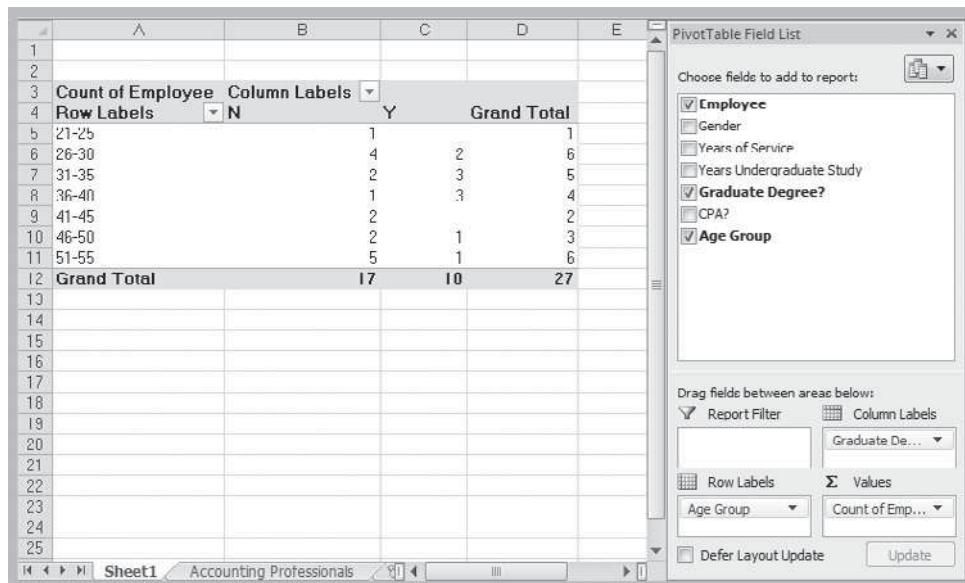
*PHStat* includes two procedures that facilitate the process of creating certain types of PivotTables by simplifying the input process and automatically creating charts (see Appendix 2.2D, “One- and Two-Way Tables and Charts”). This tool also provides a quick way of creating cross-tabulations.



Spreadsheet Note

### SKILL-BUILDER EXERCISE 2.9

Develop a cross-tabulation and chart of the categorical variables *Student* and *Gender* in the Facebook Survey Excel file using the *PHStat Two-Way Tables & Charts* tool.



**FIGURE 2.24** Cross-Tabulation as a PivotTable

## Basic Concepts Review Questions

1. Explain the principal types of descriptive statistics measures that are used for describing data.
2. What are frequency distributions and histograms? What information do they provide?
3. Provide some examples of data profiles.
4. Explain how to compute the relative frequency and cumulative relative frequency.
5. Discuss the different measures of location and compare their properties. In particular, discuss the situations where the median may be preferred over the mean, and vice versa.
6. What does skewness measure? Interpret the value of the coefficient of skewness as obtained in Excel.
7. Explain the importance of the standard deviation in interpreting and drawing conclusions about risk.
8. What does Chebyshev's theorem state and how can it be used in practice?
9. Explain the coefficient of variation and how it can be used.
10. What does kurtosis measure? Interpret the value of the coefficient of kurtosis as obtained in Excel.
11. What does the correlation coefficient measure? How do you interpret a correlation coefficient of zero?
12. What is a proportion? Provide some practical examples where proportions are used in business.
13. What is an outlier? How can one be identified?
14. Explain the information contained in box plots and dot-scale diagrams.
15. What is a PivotTable? Describe some of the key features that PivotTables have.
16. Explain how to compute the mean and variance of a sample and a population. How would you explain the formulas in simple English?
17. How can one estimate the mean and variance of data that are summarized in a grouped frequency distribution? Why are these only estimates?
18. Explain why the sign of the covariance measure indicates the direction of the relationship between the variables of interest.

## Problems and Applications

1. A community health status survey obtained the following demographic information from the respondents:

Age	Frequency
18–29	297
30–45	661
46–64	634
65+	369

Compute the relative frequency and cumulative relative frequency of the age groups. Also, estimate the average age of the sample of respondents. What assumptions do you have to make to do this?

2. The Excel file *MBA Student Survey* provides demographic data and responses to questions on the number of nights out per week and study hours per week for a group of MBA students. Construct frequency distributions and compute the relative frequencies for the

- categorical variables of gender, international student status, and undergraduate concentration. What conclusions can you draw?
3. Construct a frequency distribution and histogram for driving accuracy (%) in the Excel file *Golfing Statistics* using the Excel *Histogram* tool and appropriate bin ranges. Find the relative frequencies and cumulative relative frequencies for each bin, and estimate the average accuracy using the frequency distribution.
  4. Construct frequency distributions and histograms using the Excel *Histogram* tool for the total nuclear power production for the US, Canada and France in the Excel file *Nuclear Power*. Define appropriate bin ranges for each variable.
  5. Find the 10th and 90th percentiles of the number of hits by the different teams in the Excel file *Major League Baseball*.
  6. Find the first, second, and third quartiles for the variables Current Salary and Education in the Excel file *Salary Data*. What is the interquartile range for each of these?
  7. Find the 10th and 90th percentiles and the first and third quartiles for the taxi-in time in the Excel file *Atlanta Airline Data*.
  8. Compute the mean, median, variance, and standard deviation using the appropriate Excel functions for all the variables in the Excel file *National Football League*. Note that the data represent a population. Apply the *Descriptive Statistics* tool to these data, what differences do you observe? Why did this occur?
  9. The Excel file *Student Grades* provides information about the performance of students in a midterm examination and in a final exam.
    - a. Considering these data as a sample from the population of all students in this institution, compute the mean, variance, and standard deviation for each of these variables using the formulas (2.A2), (2.A5), and (2.A7).
    - b. Compute the coefficient of variation for both midterm and final exam scores. Which one has the greater relative dispersion?
  10. The Excel file *Burglaries* contains data on the number of burglaries before and after a Citizen Police program. Apply the *Descriptive Statistics* tool to these data. Does Chebyshev's theorem hold for the number of monthly burglaries before and after the citizen-police program?
  11. The Excel file *Baseball Attendance* shows the attendance in thousands at San Francisco Giants baseball games for the 10 years before the Oakland A's moved to the Bay Area in 1968, as well as the combined attendance for both teams for the next 11 years. What is the mean and standard deviation of the number of baseball fans attending before and after the A's move to the San Francisco area? What conclusions might you draw?
  12. For the Excel file *University Grant Proposals*, compute descriptive statistics for all proposals and also for the proposals that were funded and those that were rejected. Are any differences apparent?
  13. Compute descriptive statistics for liberal arts colleges and research universities in the Excel file *Colleges and Universities*. Compare the two types of colleges. What can you conclude?
  14. Compute descriptive statistics for all colleges and branch campuses for each year in the Excel file *Freshman College Data*. Are any differences apparent from year to year?
  15. The data in the Excel file *University Grant Proposals* presents the amount of funding requested by each grant proposal. Using formulas (2.A8) and (2.A10), estimate the mean and standard deviation of the requested funding in the grant proposals, assuming these data represent the entire population of all grant proposals.
  16. In a chess tournament involving 64 players, the ages of the contestants were summarized in a frequency table which gave the following information:
    - Ages between 15 and 24: 6 contestants
    - Ages between 25 and 34: 25 contestants
    - Ages between 35 and 44: 18 contestants
    - Ages between 45 and 54: 10 contestants
    - Ages between 55 and 64: 5 contestantsEstimate the sample mean and sample standard deviation of the age of the contestants using formulas (2.A9) and (2.A11).
  17. Data from the 2000 U.S. Census in the Excel file *California Census Data* show the distribution of ages for residents of California. Estimate the mean age and standard deviation of age for California residents using formulas (2A.9) and (2A.11), assuming these data represent a sample of current residents.
  18. A deep-foundation engineering contractor has bid on a foundation system for a new world headquarters building for a Fortune 500 company. A part of the project consists of installing 311 auger cast piles. The contractor was given bid information for cost-estimating purposes, which consisted of the estimated depth of each pile; however, actual drill footage of each pile could not be determined exactly until construction was performed. The Excel file *Pile Foundation* contains the estimates and actual pile lengths after the project was completed. Compute the correlation coefficient between the estimated and actual pile lengths. What does this tell you?
  19. Call centers have high turnover rates because of the stressful environment. The national average is approximately 50%. The director of human resources for a large bank has compiled data from about 70 former employees at one of the bank's call centers (see the Excel file *Call Center Data*). For this sample, how strongly is length of service correlated with starting age?
  20. A national homebuilder builds single-family homes and condominium-style townhouses. The Excel file *House Sales* provides information on the selling price, lot cost, type of home, and region of the country (M = Midwest, S = South) for closings during one month.
    - a. Construct a scatter diagram showing the relationship between sales price and lot cost. Does there

- appear to be a linear relationship? Compute the correlation coefficient.
- b. Construct scatter diagrams showing the relationship between sales price and lot cost *for each region*. Do linear relationships appear to exist? Compute the correlation coefficients.
- c. Construct scatter diagrams showing the relationship between sales price and lot cost for each type of house. Do linear relationships appear to exist? Compute the correlation coefficients.
21. The Excel file *Infant Mortality* provides data on infant mortality rate (deaths per 1,000 births), female literacy (percentage who read), and population density (people per square kilometer) for 85 countries. Compute the correlation matrix for these three variables. What conclusions can you draw?
22. The Excel file *Refrigerators* provides data on various brands and models. Compute the correlation matrix for the variables. What conclusions can you draw?
23. The worksheet *Mower Test* in the Excel file *Quality Measurements* shows the results of testing 30 samples of 100 lawn mowers prior to shipping. Find the proportion of units that failed the test for each sample. What proportion failed overall?
24. The Excel file *EEO Employment Report* shows the number of people employed in different professions for various racial and ethnic groups. Find the proportion of men and women in each ethnic group for the total employment and in each profession.
25. A mental health agency measured the self-esteem score for randomly selected individuals with disabilities who were involved in some work activity within the past year. The Excel file *Self Esteem* provides the data, including the individuals' marital status, length of work, type of support received (direct support includes job-related services such as job coaching and counseling), education, and age. Construct a cross-tabulation of the number of individuals within each classification of marital status and support level.
26. Construct cross-tabulations of Gender versus Carrier and Type versus Usage in the Excel file *Cell Phone Survey*. What might you conclude from this analysis?
27. The Excel file *Unions and Labor Law Data* reports the percentage of public and private sector employees in unions in 1982 for each state, along with indicators of whether the states had a bargaining law that covered public employees or right-to-work laws.
- a. Compute the proportion of employees in unions in each of the four categories: public sector with bargaining laws, public sector without bargaining laws, private sector with bargaining laws, and private sector without bargaining laws.
- b. Compute the proportion of employees in unions in each of the four categories: public sector with right-to-work laws, public sector without right-to-work laws, private sector with right-to-work laws, and private sector without right-to-work laws.
- laws, private sector with right-to-work laws, and private sector without right-to-work laws.
- c. Construct a cross-tabulation of the number of states within each classification of having or not having bargaining laws and right-to-work laws.
28. Construct box plots and dot-scale diagrams for each of the variables in the data set *Ohio Education Performance*. What conclusions can you draw from them? Are any possible outliers evident?
29. A producer of computer-aided design software for the aerospace industry receives numerous calls for technical support. Tracking software is used to monitor response and resolution times. In addition, the company surveys customers who request support using the following scale:
- 0—Did not exceed expectations
  - 1—Marginally met expectations
  - 2—Met expectations
  - 3—Exceeded expectations
  - 4—Greatly exceeded expectations
- The questions are as follows:
- Q1: Did the support representative explain the process for resolving your problem?
  - Q2: Did the support representative keep you informed about the status of progress in resolving your problem?
  - Q3: Was the support representative courteous and professional?
  - Q4: Was your problem resolved?
  - Q5: Was your problem resolved in an acceptable amount of time?
  - Q6: Overall, how did you find the service provided by our technical support department?
- A final question asks the customer to rate the overall quality of the product using this scale:
- 0—Very poor
  - 1—Poor
  - 2—Good
  - 3—Very good
  - 4—Excellent
- A sample of survey responses and associated resolution and response data are provided in the Excel file *Customer Support Survey*. Use descriptive statistics, box plots, and dot-scale diagrams as you deem appropriate to convey the information in these sample data and write a report to the manager explaining your findings and conclusions.
30. Call centers have high turnover rates because of the stressful environment. The national average is approximately 50%. The director of human resources for a large bank has compiled data from about 70 former employees at one of the bank's call centers (see the Excel file *Call Center Data*). Use PivotTables to find these items:
- a. The average length of service for males and females in the sample.

- b. The average length of service for individuals with and without a college degree.
- c. The average length of service for males and females with and without prior call center experience.

What conclusions might you reach from this information?

31. The Excel file *University Grant Proposals* provides data on the dollar amount of proposals, gender of the researcher, and whether the proposal was funded or not. Construct a PivotTable to find the average amount of proposals by gender and outcome.
32. A national homebuilder builds single-family homes and condominium-style townhouses. The Excel file *House Sales* provides information on the selling price, lot cost, type of home, and region of the country (M = Midwest, S = South) for closings during one month. Use PivotTables to find the average selling price and lot cost for each type of home in each region of the market. What conclusions might you reach from this information?
33. The Excel file *MBA Student Survey* provides data on a sample of students' social and study habits. Use PivotTables to find the average age, number of nights out per week, and study hours per week by gender, whether the student is international or not, and undergraduate concentration.
34. A mental health agency measured the self-esteem score for randomly selected individuals with disabilities

who were involved in some work activity within the past year. The Excel file *Self-Esteem* provides the data including the individuals' marital status, length of work, type of support received (direct support includes job-related services such as job coaching and counseling), education, and age. Use PivotTables to find the average length of work and self-esteem score for individuals in each classification of marital status and support level. What conclusions might you reach from this information?

35. The Excel file *Cell Phone Survey* reports opinions of a sample of consumers regarding the signal strength, value for the dollar, and customer service for their cell phone carriers. Use PivotTables to find the following:
- a. The average signal strength by type of carrier.
  - b. Average value for the dollar by type of carrier and usage level.
  - c. Variance of perception of customer service by carrier and gender.

What conclusions might you reach from this information?

36. The Excel file *Freshman College Data* shows data for four years at a large urban university. Use PivotTables to examine differences in student high school performance and first-year retention among different colleges at this university. What conclusions do you reach?

## Case

### The Malcolm Baldrige Award

The Malcolm Baldrige Award recognizes U.S. companies that excel in high-performance management practice and have achieved outstanding business results. The award is a public-private partnership, funded primarily through a private foundation and administered through the National Institute of Standards and Technology (NIST) in cooperation with the American Society for Quality (ASQ) and is presented annually by the President of the United States. It was created to increase the awareness of American business for quality and good business practices and has become a worldwide standard for business excellence. See the Program Web site at [www.nist.gov/baldrige](http://www.nist.gov/baldrige) for more information.

The award examination is based on a rigorous set of criteria, called the *Criteria for Performance Excellence*, which consists of seven major categories: Leadership; Strategic Planning; Customer Focus; Measurement, Analysis, and Knowledge Management; Workforce Focus; Process Management; and Results. Each category consists

of several *items* that focus on major requirements on which businesses should focus. For example, the two items in the Leadership category are Senior Leadership and Governance and Social Responsibilities. Each item, in turn, consists of a small number of *areas to address*, which seek specific information on approaches used to ensure and improve competitive performance, the deployment of these approaches, or results obtained from such deployment. The current year's criteria may be downloaded from the Web site.

Applicants submit a 50-page document that describes their management practices and business results that respond to the criteria. The evaluation of applicants for the award is conducted by a volunteer board of examiners selected by NIST. In the first stage, each application is reviewed by a team of examiners. They evaluate the applicant's response to each criteria item, listing major strengths and opportunities for improvement relative to the criteria. Based on these comments, a score

(continued)

from 0 to 100 in increments of 10 is given to each item. Scores for each examination item are computed by multiplying the examiner's score by the maximum point value that can be earned for that item, which varies by item. These point values weight the importance of each item in the criteria. Then the examiners share information on a secure Web site and discuss issues via telephone conferencing to arrive at consensus comments and scores. The consensus stage is an extremely important step of the process. It is designed to smooth out variations in examiners' scores, which inevitably arise because of different perceptions of the applicants' responses relative to the criteria, and provide useful feedback to the applicants. In many cases, the insights of one or two judges may sway opinions, so consensus scores are not simple averages. A national panel of judges then reviews the scores and selects the highest-scoring applicants for site visits.

At this point, a team of examiners visits the company for the greater part of a week to verify information contained in the written application and resolve issues that are unclear or about which the team needs to learn more. The results are written up and sent to the judges who use the site visit reports and discussions with the team leaders to recommend award recipients to the Secretary of Commerce.

Statistics and data analysis tools can be used to provide a summary of the examiners' scoring profiles and to help the judges review the scores. Figure 2.25 illustrates a hypothetical example (Excel file *Baldrige*).<sup>4</sup> Your task is to apply the concepts and tools discussed in this chapter to analyze the data and provide the judges with appropriate statistical measures and visual information to facilitate their decision process regarding a site visit recommendation.

A	B	C	D	E	F	G	H	I	J	K	
1	Baldrige Examination Scores										
2											
Individual Assessment Percentage Scores											
4	Item	Maximum Points	Examiner 1	Examiner 2	Examiner 3	Examiner 4	Examiner 5	Examiner 6	Examiner 7	Examiner 8	Consensus Score
5	1.1	70	80	80	50	60	60	70	70	50	75
6	1.2	50	30	50	30	40	40	60	60	50	50
7	2.1	40	50	70	50	50	40	60	70	40	65
8	2.2	45	30	40	50	50	60	40	30	50	55
9	3.1	40	30	60	40	60	50	30	50	30	45
10	3.2	45	30	50	60	60	60	50	30	60	50
11	4.1	45	40	70	50	60	40	30	20	50	50
12	4.2	45	30	20	40	40	30	30	10	30	40
13	5.1	45	70	50	60	40	40	60	60	50	60
14	5.2	40	50	20	40	40	70	40	40	20	40
15	6.1	35	50	60	50	50	50	40	30	40	45
16	6.2	50	40	40	60	50	40	30	60	50	50
17	7.1	100	60	70	70	70	80	70	70	70	75
18	7.2	70	50	60	70	50	70	50	70	70	70
19	7.3	70	50	40	50	50	70	30	30	50	50
20	7.4	70	40	50	50	50	50	40	20	60	45
21	7.5	70	70	70	60	70	50	60	80	50	75
22	7.6	70	60	80	70	60	70	40	60	70	70
23	Weighted score		499.5	565.5	546.5	543	564	478.5	503	523	585

**FIGURE 2.25** Baldrige Examination Scores

<sup>4</sup>The criteria undergo periodic revision, so the items and maximum points will not necessarily coincide with the current year's criteria.

## APPENDIX 2.1

### Descriptive Statistics: Theory and Computation

In this appendix, we summarize some basic theory and mathematical basis for descriptive statistics calculations. While you will most often use the capabilities of Excel to perform calculations, understanding the basic theory and formulas is also important, particularly when Excel procedures are not available.

#### A. Mean, Variance, and Standard Deviation

It is common practice in statistics to use Greek letters to represent population measures and Roman letters to represent sample statistics. We will use  $N$  to represent the number of items in a population, and  $n$  to represent the number of observations in a sample. If a population consists of  $N$  observations  $x_1, \dots, x_N$ , population mean,  $\mu$  is calculated as:

$$\mu = \frac{\sum_{i=1}^N x_i}{N} \quad (2A.1)$$

The mean of a sample of  $n$  observations,  $x_1, \dots, x_n$ , denoted by “ $\bar{x}$ -bar” is calculated as:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (2A.2)$$

Note that the calculations for the mean are the same; only the notation differs between a sample and a population. One property of the mean is that the sum of the deviations of each observation from the mean is 0:

$$\sum_i (x_i - \bar{x}) = 0 \quad (2A.3)$$

The formula for the variance of a population is:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} \quad (2A.4)$$

where  $x_i$  is the value of the  $i$ th item,  $N$  is the number of items in the population, and  $\mu$  is the population mean. Essentially, the variance is the average of the squared deviations of the observations from the mean.

A major difference exists between the variance of a population and that of a sample. The variance of a sample is calculated using the formula:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (2A.5)$$

where  $n$  is the number of items in the sample, and  $\bar{x}$  is the sample mean. It may seem peculiar to use a different denominator to “average” the squared deviations from the mean for populations and samples, but statisticians have shown that the formula for the sample variance provides a more accurate representation of the true population variance. We will discuss this more formally in Chapter 4. For now, simply understand that the proper calculations of the population and sample variance use different denominators based on the number of observations in the data.

The standard deviation is the square root of the variance. For a population, the standard deviation is computed as:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} \quad (2A.6)$$

and for samples, it is:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} \quad (2A.7)$$

#### SKILL-BUILDER EXERCISE 2.10

Recognizing that the data in the Excel file *Major League Baseball* represents population data, develop a spreadsheet to calculate the mean, variance, and standard deviation using formulas (2A.1), (2A.4), and (2A.6).

## B. Statistical Measures for Grouped Data

When sample data are summarized in a frequency distribution, the mean of a population may be computed using the formula:

$$\mu = \frac{\sum_{i=1}^N f_i x_i}{N} \quad (2A.8)$$

For samples, the formula is similar:

$$\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{n} \quad (2A.9)$$

where  $f_i$  is the frequency of observation  $x_i$ .

To illustrate this, consider the Hours online/week in the *Facebook Survey* data. The calculations are shown below:

Hours Online/Week	Frequency	Hours × Frequency
1	0	0
2	4	8
3	1	3
4	6	24
5	4	20
6	5	30
7	4	28
8	2	16
9	2	18
10	2	20
11	0	0
12	2	24
13	0	0
14	0	0
15	1	15
Sum		206

$$\text{Mean} = 206/33 = 6.24$$

### SKILL-BUILDER EXERCISE 2.11

Construct a frequency distribution for the number of vacations per year in the Excel file *Vacation Survey*. Use the frequency distribution to calculate the mean and standard deviation, and verify your calculations using appropriate Excel functions.

## C. Skewness and Kurtosis

The Coefficient of Skewness, CK, is computed as:

$$CS = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^3}{\sigma^3} \quad (2A.12)$$

If the data are grouped into  $k$  cells in a frequency distribution, we can use modified versions of these formulas to estimate the mean by replacing  $x_i$  with a representative value (such as the midpoint) for all the observations in each cell. Thus, using the Facebook Friends data, we would have:

Upper Limit	Midpoint	Frequency	Midpoint × Frequency
50	25	5	125
100	75	9	675
150	125	5	625
200	175	4	700
250	225	2	450
300	275	2	550
350	325	1	325
400	375	2	750
450	425	2	850
500	475	1	475
Sum		5,525	

$$\text{Estimation of the mean} = 5,525/33 = 167.42$$

Note that this is not identical, but close, to the true mean of 176.97. This is because we have not used all the original data, but only representative values for each cell. Although most statistics are simple concepts, they must be applied correctly, and we need to understand how to interpret them properly.

We may use similar formulas to compute the population variance for grouped data:

$$\sigma^2 = \frac{\sum_{i=1}^N f_i (x_i - \mu)^2}{N} \quad (2A.10)$$

and sample variance:

$$s^2 = \frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{n - 1} \quad (2A.11)$$

For sample data, replace the population mean and standard deviation with the corresponding sample statistics.

The Coefficient of Kurtosis, CK, is computed as:

$$CK = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^4}{\sigma^4} \quad (2A.13)$$

Again, for sample data, use the sample statistics instead of the population measures.

## D. Correlation

The correlation coefficient for a population is computed as:

$$\rho_{xy} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \quad (2A.14)$$

The numerator is called the **covariance** and is the average of the products of deviations of each observation from its respective mean:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{N} \quad (2A.15)$$

To understand this, examine the formula for the covariance. This is the average of the product of the deviations of each pair of observations from their respective means. Suppose that large (small) values of  $X$  are generally associated with large (small) values of  $Y$ . Then in most cases, both  $x_i$  and  $y_i$  are either above or below their respective means. If so, the product of the deviations from the means will be a

positive number, and when added together and averaged, it will give a positive value for the covariance. On the other hand, if small (large) values of  $X$  are associated with large (small) values of  $Y$ , then one of the deviations from the mean will generally be negative while the other is positive. When multiplied together, a negative value results, and the value of the covariance will be negative. The Excel function COVAR computes the covariance of a population.

In a similar fashion, the **sample correlation coefficient** is computed as:

$$r_{xy} = \frac{\text{cov}(X, Y)}{s_x s_y} \quad (2A.16)$$

However, the sample covariance is computed as:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \quad (2A.17)$$

Like the sample variance, note the use of  $n - 1$  in the denominator. The Excel 2010 function COVARIANCE.P computes the covariance for a population; the Excel function COVARIANCE.S computes the covariance for a sample.

## APPENDIX 2.2

### Excel and PHStat Notes

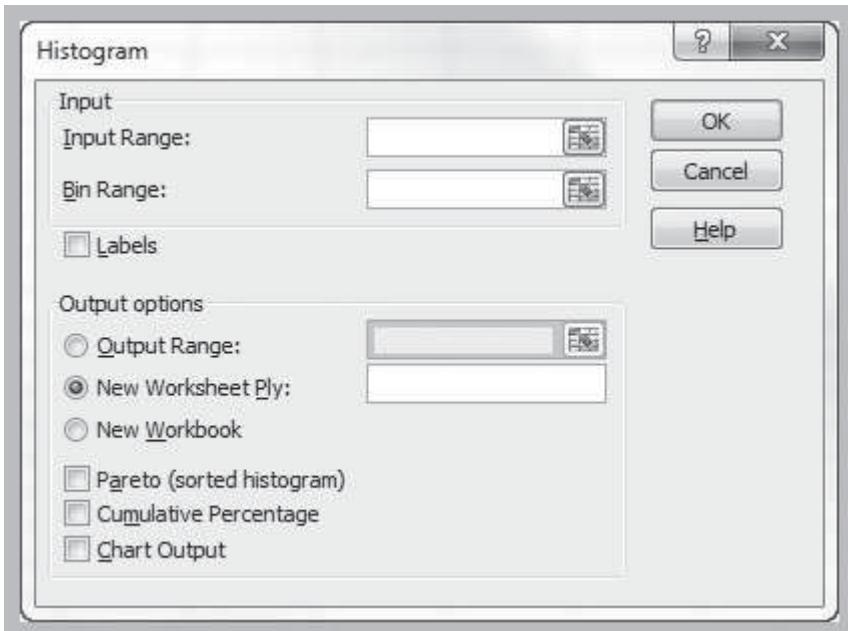
#### A. Creating a Frequency Distribution and Histogram

Click the *Data Analysis* tools button in the *Analysis* group under the *Data* tab in the Excel menu bar, and select *Histogram* from the list. In the dialog box (see Figure 2A.1), specify the *Input Range* corresponding to the data. If you include the column header, then also check the *Labels* box, so Excel knows that the range contains a label. If you do not specify a *Bin Range*, Excel will automatically determine bin values for the frequency distribution and histogram, which often results in a rather poor histogram. We recommend that you define your own bin values by specifying the upper cell limits of each interval in a column in your worksheet after examining the range of the data. Generally, you should choose between 5 and 15 cells, and bin ranges should be of equal width. The more data you have, the more cells you should generally use. Note that with fewer cells, the cell widths will be wider. Wider cell widths provide a “coarse” histogram. Sometimes you need to experiment to find the best number of cells that provide a useful visualization of the data. Choose the width by calculating

(Max. value – Min. value)/Number of cells, and round the result to a reasonable value. (If you check the *Data Labels* box, be sure you include a column label such as “Bin” or “Upper Cell Limit” in the bin range column or else Excel will treat your first value as a label.) If you have a small number of unique values, use discrete values for the bin range. Check the *Chart Output* box to display a histogram in addition to the frequency distribution.

#### B. Using the Descriptive Statistics Tool

Click on *Data Analysis* in the *Analysis* group under the *Data* tab in the Excel menu bar. Select *Descriptive Statistics* from the list of tools. The *Descriptive Statistics* dialog shown in Figure 2A.2 will appear. You need only enter the range of the data, which must be in a *single row or column*. If the data are in multiple columns, the tool treats each row or column as a separate data set, depending on which you specify. This means that if you have a single data set arranged in a matrix format, you would have to stack the data (e.g., using the *PHStat Stack Data* tool described in Chapter 1) in a single

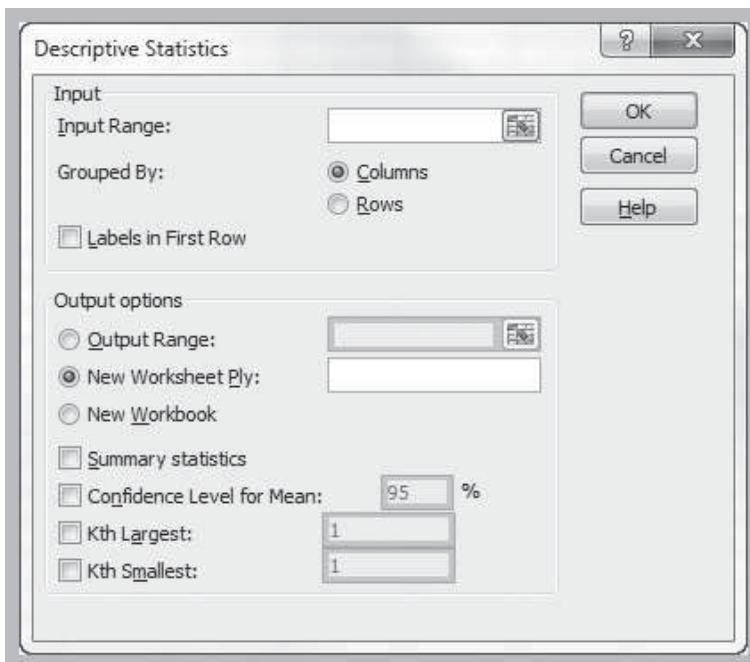


**FIGURE 2A.1** Histogram Tool Dialog

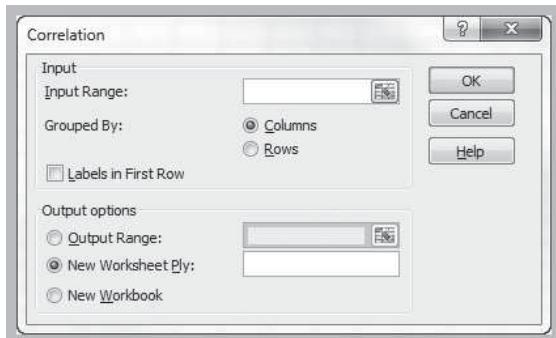
column before applying the *Descriptive Statistics* tool. Check the box *Labels in First Row* if labels are included in the input range. You may choose to save the results in the current worksheet or in a new one. For basic summary statistics, check the box *Summary statistics*; you need not check any others.

### C. Using the Correlation Tool

Select *Correlation* from the *Data Analysis* tool list. The dialog box is shown in Figure 2A.3. You need only input the range of the data (which must be in contiguous columns; if not, you must move them in your worksheet), specify whether



**FIGURE 2A.2** Descriptive Statistics Tool Dialog

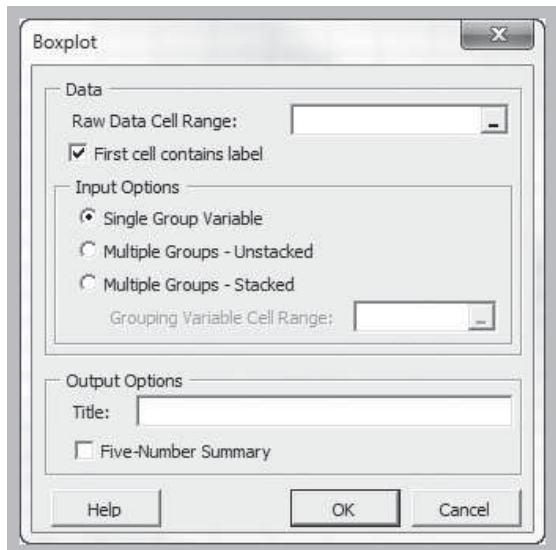


**FIGURE 2A.3** Correlation Tool Dialog

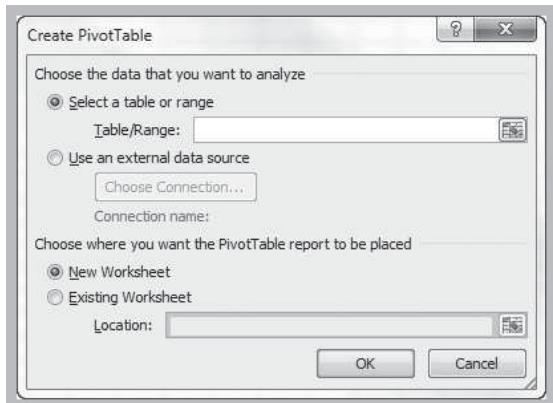
the data are grouped by rows or columns (most applications will be grouped by columns), and indicate whether the first row contains data labels. The output of this tool is a matrix giving the correlation between each pair of variables. This tool provides the same output as the CORREL function for each pair of variables.

#### D. Creating Box Plots

From the *PHStat* menu, select *Descriptive Statistics* and then *Box Plot*. The dialog box is shown in Figure 2A.4. In the *Raw Data Cell Range* box, enter the range of the data; if the first cell contains a label, check the box below. For a single data set, check the *Single Group Variable* radio button. For multiple groups of data, check the appropriate button (see the *PHStat* note on stacked and unstacked data in Chapter 1). In the *Output Options* section, you may enter a title for the chart. Checking the *Five-Number Summary* box will provide a worksheet with the minimum, first quartile, median, third quartile, and maximum values of the data set(s).



**FIGURE 2A.4** PHStat Box Plot Dialog



**FIGURE 2A.5** Create PivotTable Dialog

#### E. Creating PivotTables

Choose *PivotTable* from the *Tables* group under the *Insert* tab. The *Create PivotTable* dialog (Figure 2A.5) asks you for the range of the data. If you click on any cell within the data matrix before inserting a PivotTable, Excel will generally default to the complete range of your list. You may either put the PivotTable into a new worksheet or in a blank range of the existing worksheet.

You may create other PivotTables without repeating all the steps in the Wizard. Simply copy and paste the first table.

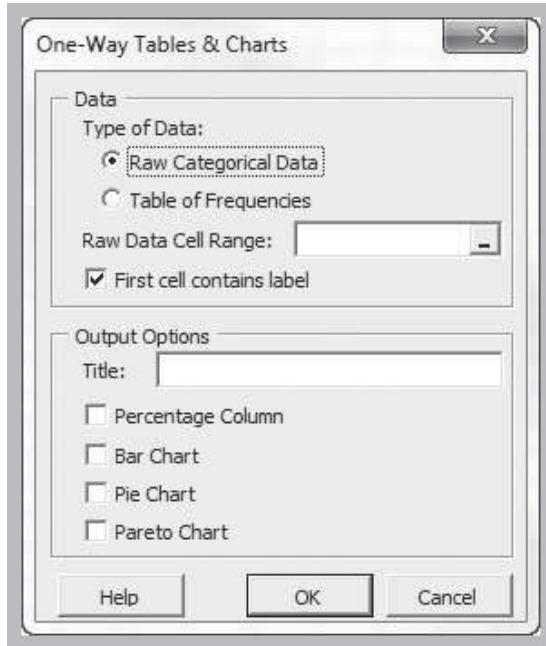
#### F. One- and Two-Way Tables and Charts

To generate a one-way PivotTable for a set of categorical data, select *Descriptive Statistics* from the *PHStat* menu and then *One-Way Tables & Charts*. The dialog box in Figure 2A.6 prompts you for the type and location of the data and optional charts that you would like to create. The type of data may be either:

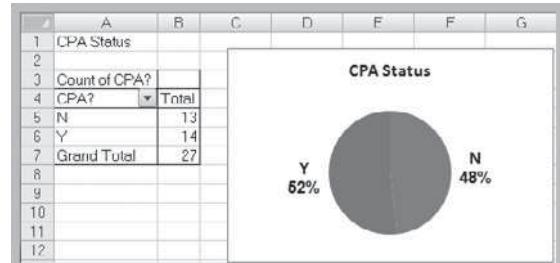
- *Raw Categorical Data*. If selected, the single-column cell range entered in the dialog box contains raw, unsummarized data. (Selected by default.)
- *Table of Frequencies*. If selected, the two-column cell range entered in the dialog box contains a frequency table containing categories and frequency counts for each category.

Enter the single-column (raw data) or two-column (frequency table) cell range containing the categorical data to be summarized in the *Raw Data Cell Range* field, and check *First cell contains label* if the contents of the first cell (or first row) are treated as a descriptive label and not as a data value. *PHStat* creates a new worksheet for the PivotTable and the chart. Figure 2A.7 shows the table and pie chart generated for the CPA status in the Excel file *Accounting Professionals* (placed on one worksheet).

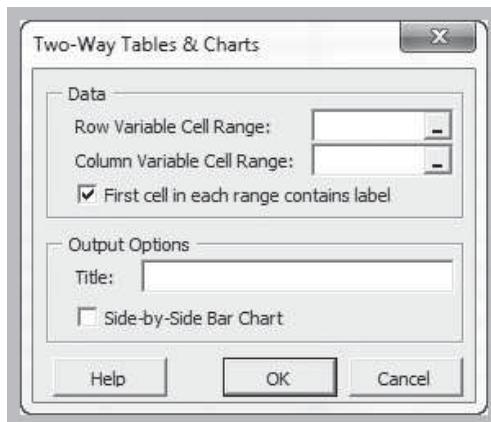
A two-way table may also be constructed by selecting *Two-Way Tables & Charts* from the menu. Click on *PHStat*



**FIGURE 2A.6** *PHStat One-Way Tables & Charts* Dialog



**FIGURE 2A.7** One-Way Table and Chart for CPA Status



**FIGURE 2A.8** *PHStat Two-Way Tables & Charts* Dialog

from the *Menu Commands* group under the *Add-Ins* tab. In the *Descriptive Statistics* menu, select *Two-Way Tables & Charts*. In the dialog (Figure 2A.8), enter the ranges of the categorical variables for the contingency table. If you want a bar chart to display the results, check the box. However, we note that the PivotTables created by this tool include

only the variables selected and does not allow you to change the variables, thus, limiting the flexibility of creating different views of PivotTables as can be done using the Excel PivotTable tool.

## *Chapter 3*

# Probability Concepts and Distributions

- INTRODUCTION 90
- BASIC CONCEPTS OF PROBABILITY 90
  - Basic Probability Rules and Formulas 91
  - Conditional Probability 92
- RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS 94
- DISCRETE PROBABILITY DISTRIBUTIONS 97
  - Expected Value and Variance of a Discrete Random Variable 98
  - Bernoulli Distribution 99
  - Binomial Distribution 99
  - Poisson Distribution 100
- CONTINUOUS PROBABILITY DISTRIBUTIONS 102
  - Uniform Distribution 104
  - Normal Distribution 105
  - Triangular Distribution 108
  - Exponential Distribution 109
  - Probability Distributions in *PHStat* 110
  - Other Useful Distributions 110
- JOINT AND MARGINAL PROBABILITY DISTRIBUTIONS 113
- BASIC CONCEPTS REVIEW QUESTIONS 114
- PROBLEMS AND APPLICATIONS 114
- CASE: PROBABILITY ANALYSIS FOR QUALITY MEASUREMENTS 118
- APPENDIX 3.1: PROBABILITY DISTRIBUTIONS: THEORY AND COMPUTATION 118
  - A. Expected Value and Variance of a Random Variable 118
  - B. Binomial Distribution 118
  - C. Poisson Distribution 119
  - D. Uniform Distribution 119
  - E. Normal Distribution 119
  - F. Exponential Distribution 120
- APPENDIX 3.2: EXCEL AND *PHStat* NOTES 120
  - A. Normal Probability Tools 120
  - B. Generating Probabilities in *PHStat* 121

## INTRODUCTION

Most business decisions involve some elements of uncertainty and randomness. For example, in models of manufacturing operations, times of job arrivals, job types, processing times, times between machine breakdowns, and repair times all involve uncertainty. Similarly, a model to predict the future return of an investment portfolio requires a variety of assumptions about uncertain economic conditions and market behavior. Specifying the nature of such assumptions is a key modeling task that relies on fundamental knowledge of probability concepts and probability distributions—the subject of this chapter.

The notion of probability is used everywhere, both in business and in our daily lives; from market research and stock market predictions, to the World Series of Poker and weather forecasts. Probability quantifies the uncertainty that we encounter all around us, and is an important element of analytical modeling and decision making. Probability concepts and probability distributions are also important in applying statistics to analyze sample data from business processes. Thus, we will also examine characteristics of sampling distributions in this chapter and discuss errors associated with sampling.

## BASIC CONCEPTS OF PROBABILITY

Managers often wish to know such things as the likelihood that a new product will be profitable, or the chances that a project will be completed on time. **Probability** is the likelihood that an outcome—such as whether a new product will be profitable or not, or whether a project will be completed within 15 weeks—occurs. Probabilities are expressed as values between 0 and 1, although many people convert them to percentages. The statement “there is a 10% chance that oil prices will rise next quarter” is another way of stating that “the probability of a rise in oil prices is 0.1.” The closer the probability is to 1, the more likely it is that the outcome will occur.

To formally discuss probability, we need some new terminology. An **experiment** is a process that results in some outcome. An experiment might be as simple as rolling two dice, observing and recording weather conditions, conducting a market research study, or watching the stock market. The **outcome** of an experiment is a result that we observe; it might be the sum of two dice, a description of the weather, the proportion of consumers who favor a new product, or the change in the Dow Jones Industrial Average (DJIA) at the end of a week. The collection of all possible outcomes of an experiment is called the **sample space**. For instance, if we roll two fair dice, the possible outcomes are the numbers 2 through 12; if we observe the weather, the outcome might be clear, partly cloudy, or cloudy; the outcomes for customer reaction to a new product in a market research study would be favorable or unfavorable; and the weekly change in the DJIA can theoretically be any positive or negative real number. Note that a sample space may consist of a small number of discrete outcomes or an infinite number of outcomes.

Probability may be defined from one of three perspectives. First, if the process that generates the outcomes is known, probabilities can be deduced from theoretical arguments; this is the *classical definition* of probability. For example, if we examine all possible outcomes associated with rolling two dice, we can easily determine that out of 36 possible outcomes, one outcome will be the number 2, two outcomes will be the number 3 (you can roll a 1 on the first die and 2 on the second, and vice versa), six outcomes will be the number 7, and so on. Thus, the probability of rolling any number is the ratio of the number of ways of rolling that number to the total number of possible outcomes. For instance, the probability of rolling a 2 is  $1/36$ , the probability of rolling a 3 is  $2/36 = 1/18$ , and the probability of rolling a 7 is  $6/36 = 1/6$ . Similarly, if two consumers are asked their opinion about a new product, there could be four possible outcomes:

1. (Favorable, Favorable)
2. (Favorable, Not Favorable)
3. (Not Favorable, Favorable)
4. (Not Favorable, Not Favorable)

	A	B	C	D	E	F
1	Seattle Weather					
2						
3		Average Temperature	Average Rainfall	Clear	Partly Cloudy	Cloudy
4						
5	January	41.3	5.4	3	5	23
6	February	44.3	4	3	6	19
7	March	46.6	3.8	4	8	19
8	April	50.4	2.5	5	9	16
9	May	56.1	1.8	7	10	14
10	June	61.4	1.6	7	8	15
11	July	65.3	0.9	12	10	9
12	August	65.7	1.2	10	10	11
13	September	60.8	1.9	9	8	13
14	October	53.5	3.3	5	8	18
15	November	46.3	5.7	3	6	21
16	December	41.6	6	3	5	23

**FIGURE 3.1** Excel File Seattle Weather

If these are assumed to be equally likely, the probability that at least one consumer would respond unfavorably is  $3/4$ .

The second approach to probability, called the *relative frequency definition*, is based on empirical data. For example, a sample of weather in the Excel file *Seattle Weather* shows that on average in January in Seattle, 3 days were clear, 5 were partly cloudy, and 23 were cloudy (see Figure 3.1). Thus, the probability of a clear day in Seattle in January would be computed as  $3/31 = 0.097$ . As more data become available (or if global weather changes), the distribution of outcomes, and hence, the probability may change.

Finally, the *subjective definition* of probability is based on judgment, as financial analysts might do in predicting a 75% chance that the DJIA will increase 10% over the next year, or as sports experts might predict a one-in-five chance (0.20 probability) of a certain team making it to the Super Bowl at the start of the football season.

Which definition to use depends on the specific application and the information we have available. We will see various examples that draw upon each of these perspectives.

## Basic Probability Rules and Formulas

Suppose we label the  $n$  elements of a sample space as  $O_1, O_2, \dots, O_n$ , where  $O_i$  represents the  $i$ th outcome in the sample space. Let  $P(O_i)$  be the probability associated with the outcome  $O_i$ . Two basic facts govern probability:

- The probability associated with any outcome must be between 0 and 1, or

$$0 \leq P(O_i) \leq 1 \text{ for each outcome } O_i \quad (3.1)$$

- The sum of the probabilities over all possible outcomes must be 1.0, or

$$P(O_1) + P(O_2) + \dots + P(O_n) = 1$$

An **event** is a collection of one or more outcomes from a sample space. An example of an event would be rolling a 7 or an 11 with two dice, having a clear or partly cloudy day in Seattle, or obtaining a positive weekly change in the DJIA. This leads to the following rule:

**Rule 1.** *The probability of any event is the sum of the probabilities of the outcomes that compose that event.*

For example, consider the event of rolling a 7 or an 11 on two dice. The probability of rolling a 7 is  $6/36$  and the probability of rolling an 11 is  $2/36$ ; thus, the probability of rolling a 7 or 11 is  $6/36 + 2/36 = 8/36$ . Similarly, the probability of a clear or partly cloudy day in January in Seattle is  $3/31 + 5/31 = 8/31$ .

If  $A$  is any event, the **complement** of  $A$ , denoted as  $A^c$ , consists of all outcomes in the sample space not in  $A$ .

**Rule 2.** The probability of the complement of any event  $A$  is  $P(A^c) = 1 - P(A)$ .

For example, if  $A = \{7, 11\}$  in the dice example, then  $A^c = \{2, 3, 4, 5, 6, 8, 9, 10, 12\}$ . Thus, the probability  $P(A^c) = 1 - 8/36 = 28/36$ .

The union of two events contains all outcomes that belong to either of the two events. To illustrate this with rolling two dice, let  $A$  be the event  $\{7, 11\}$  and  $B$  be the event  $\{2, 3, 12\}$ . The union of  $A$  and  $B$  is the event  $\{2, 3, 7, 11, 12\}$ . If  $A$  and  $B$  are two events, the probability that some outcome in either  $A$  or  $B$  (i.e., the union of  $A$  and  $B$ ) occurs is denoted as  $P(A \text{ or } B)$ . Finding this probability depends on whether the events are mutually exclusive or not.

Two events are **mutually exclusive** if they have no outcomes in common. The events  $A$  and  $B$  in the dice example are mutually exclusive. When events are mutually exclusive, the following rule applies:

**Rule 3.** If events  $A$  and  $B$  are mutually exclusive, then  $P(A \text{ or } B) = P(A) + P(B)$ .

For the dice example, the probability of event  $A$  is  $P(A) = 8/36$ , and the probability of event  $B$  is  $P(B) = 4/36$ . Therefore, the probability that either event  $A$  or  $B$  occurs, that is, the roll of the dice is either 2, 3, 7, 11, or 12, is  $8/36 + 4/36 = 12/36$ .

If two events are *not* mutually exclusive, then adding their probabilities would result in double counting some outcomes, so an adjustment is necessary. This leads to the following rule:

**Rule 4.** If two events  $A$  and  $B$  are not mutually exclusive, then  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ .

Here,  $(A \text{ and } B)$  represents the intersection of events  $A$  and  $B$ ; that is, all outcomes belonging to both  $A$  and  $B$ . For instance, in the dice example, if  $A = \{2, 3, 12\}$  and  $B = \{\text{even number}\}$ , then  $A$  and  $B$  are not mutually exclusive because both have the numbers 2 and 12 in common. Therefore,  $P(A \text{ or } B) = P(\{2, 3, 12\}) + P(\{\text{even number}\}) - P(A \text{ and } B) = 4/36 + 18/36 - 2/36 = 20/36$ .

## Conditional Probability

**Conditional probability** is the probability of occurrence of one event  $A$ , given that another event  $B$  is known to be true or have already occurred. Suppose that a sample of 100 individuals who were asked to evaluate their preference for three new proposed energy drinks in a blind taste test yielded the following cross-tabulation:

Cross-Tabulation	Brand 1	Brand 2	Brand 3	Total
Male	25	17	21	63
Female	9	6	22	37
Total	34	23	43	100

Consider the following events:

$M$  = respondent is male

$F$  = respondent is female

$B1$  = respondent prefers brand 1

$B2$  = respondent prefers brand 2

$B3$  = respondent prefers brand 3

Note that  $P(M) = 63/100$ ,  $P(F) = 37/100$ ,  $P(B1) = 34/100$ ,  $P(B2) = 23/100$ , and  $P(B3) = 43/100$ . Also note that the numbers within the table represent the number of outcomes in the

intersection of the gender and brand events. For example, the probability that an individual is male and prefers brand 1 is  $P(M \text{ and } B1) = 25/100$ .

Now suppose that we know that a respondent is male. What is the probability that he prefers brand 1? Note that there are only 63 males in the group, and of these, 25 prefer brand 1. Therefore, the probability that a male respondent prefers brand 1 is  $25/63$ . This is called a conditional probability since it depends on the knowledge of one of the events.

In general, the **conditional probability** of an event  $A$  given that event  $B$  is known to have occurred is:

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)} \quad (3.2)$$

We read the notation  $P(A | B)$  as “the probability of  $A$  given  $B$ .” Thus, if we substitute  $B1$  for  $A$  and  $M$  for  $B$  in this formula, we have

$$P(B1 | M) = P(B1 \text{ and } M) / P(M) = (25/100) / (63/100) = 25/63 = 0.397.$$

Similarly, the probability of preferring brand 1 if the respondent is female is

$$P(B1 | F) = P(B1 \text{ and } F) / P(F) = (9/100) / (37/100) = 9/37 = 0.243.$$

The table below summarizes the conditional probabilities of brand preference given gender:

$P(\text{Brand}   \text{Gender})$	Brand 1	Brand 2	Brand 3
Male	0.397	0.270	0.333
Female	0.243	0.162	0.595

Such information can be important in marketing efforts. Knowing that there is a difference in preference by gender can help focus advertising. For example, we see that about 40% of males prefer brand 1 while only about 24% of females do, and a higher proportion of females prefer brand 3. This suggests that it would make more sense to focus on advertising brand 1 more in male-oriented media and brand 3 in female-oriented media.

The conditional probability formula may be used in other ways. For example, multiplying both sides of equation (3.2) by  $P(B)$ , we obtain  $P(A \text{ and } B) = P(A | B) P(B)$ . Note that we may switch the roles of  $A$  and  $B$  and write  $P(B \text{ and } A) = P(B | A) P(A)$ . But  $P(B \text{ and } A)$  is the same as  $P(A \text{ and } B)$ ; thus we can express  $P(A \text{ and } B)$  in two ways:

$$P(A \text{ and } B) = P(A | B) P(B) = P(B | A) P(A) \quad (3.3)$$

This is often called the **multiplication law of probability**.

For example, suppose that in a game of Texas Hold 'Em, a player receives an ace on his first card. The probability that he will end up with “pocket aces” (two aces in the hand) is  $P(\text{ace on first card and ace on second card}) = P(\text{ace on second card} | \text{ace on first card}) \times P(\text{ace on first card})$ . Since the probability of an ace on the first card is  $4/52$ , and the probability of an ace on the second card if he has already drawn an ace is  $3/51$ , we have:

$$\begin{aligned} & P(\text{ace on first card and ace on second card}) \\ &= P(\text{ace on second card} | \text{ace on first card}) \times P(\text{ace on first card}) \\ &= (3/51) \times (4/52) = 0.004525 \end{aligned}$$

In the marketing example above, we see that the probability of preferring a brand depends on gender. We may say that brand preference and gender are not independent. We may formalize this concept by defining the notion of independent events:

Two events  $A$  and  $B$  are **independent** if

$$P(A | B) = P(A) \quad (3.4)$$

Applying this to the example, we see that while  $P(B1 | M) = 0.397$ ,  $P(B1) = 0.34$ ; thus, these two events are not independent.

Finally, we see that if two events are independent, then we can simplify the multiplication law of probability in equation (3.3) by substituting  $P(A)$  for  $P(A | B)$ :

$$P(A \text{ and } B) = P(B)P(A) = P(A)P(B) \quad (3.5)$$

For example, if  $A$  is the event that a 6 is first rolled on a pair of dice, and  $B$  is the event of rolling a 2, 3, or 12 on the next roll, then  $P(A \text{ and } B) = P(A)P(B) = (5/36)(4/36) = 20/1296$ , because the roll of a pair of dice does not depend on the previous role.

### SKILL-BUILDER EXERCISE 3.1

Develop a spreadsheet for computing the joint probabilities from the cross-tabulation data for the energy drink example. Use the spreadsheet to compute the marginal probabilities, and all conditional probabilities  $P(\text{Brand} | \text{Gender})$ .

## RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

Some experiments naturally have numerical outcomes, as rolls of dice or the weekly change in the DJIA. For other experiments, such as observing the weather or obtaining consumer response to a new product, the sample space is categorical. To have a consistent mathematical basis for dealing with probability, we would like the outcomes of all experiments to be numerical. A **random variable** is a numerical description of the outcome of an experiment. Formally, a random variable is a function that assigns a real number to each element of a sample space. If we have categorical outcomes, we can associate an arbitrary numerical value to them, such as 0 = clear, 1 = partly cloudy, and 2 = cloudy, but there is no physical or natural meaning to this scheme. Similarly, a favorable product reaction in a market research study might be assigned a value of 1, and an unfavorable reaction a value of 0. Random variables are usually denoted by capital Roman letters, such as  $X$  or  $Y$ .

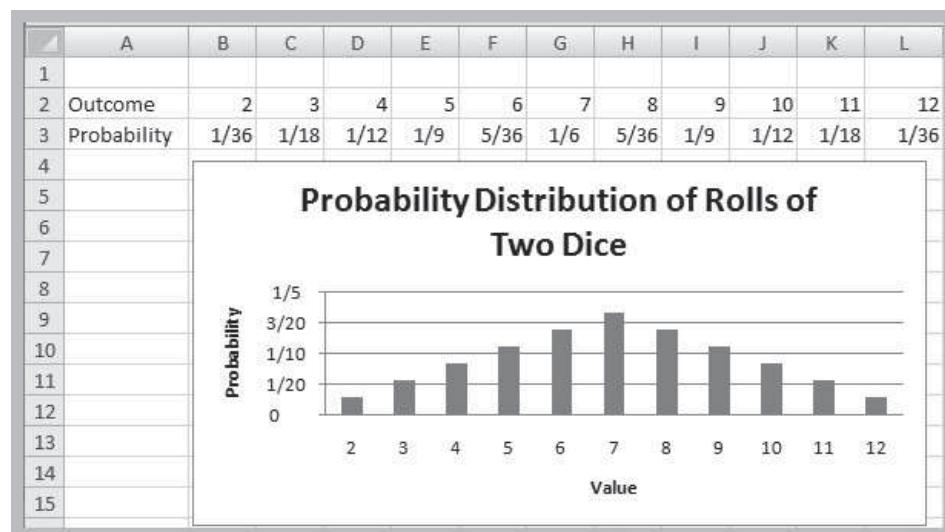
Random variables may be discrete or continuous. A **discrete random variable** is one for which the number of possible outcomes can be counted. For example, the outcomes of rolling dice, the type of weather for the next day, and customer reactions to a product are discrete random variables. The number of outcomes may be finite or theoretically infinite, such as the number of hits on a Web site link during some period of time—we cannot place a guaranteed upper limit on this number—nevertheless, the outcomes can be counted. A **continuous random variable** has outcomes over one or more continuous intervals of real numbers, such as the weekly change in the DJIA, which may assume any positive or negative value. Other examples of continuous random variables include the daily temperature, the time to complete a task, the time between failures of a machine, and the return on an investment.

A **probability distribution** is a characterization of the possible values that a random variable may assume along with the probability of assuming these values. A probability distribution can be either discrete or continuous, depending on the nature

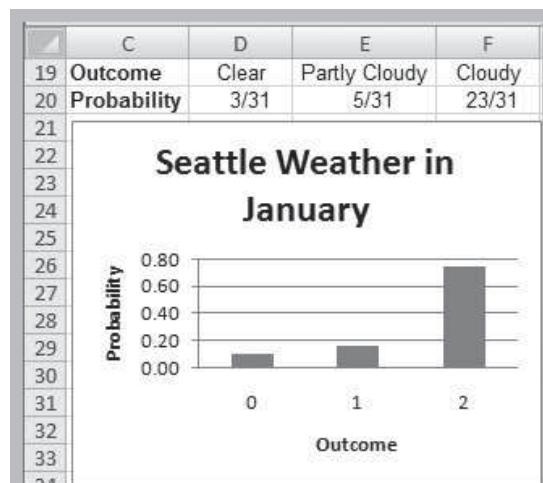
of the random variable it models. Discrete distributions are easier to understand and work with, and we will deal with them first.

We may develop a probability distribution using any one of the three perspectives of probability. First, if we can quantify the probabilities associated with the values of a random variable from theoretical arguments, then we can easily define the probability distribution. For example, the probabilities of the outcomes for rolling two dice, calculated by counting the number of ways to roll each number divided by the total number of possible outcomes, along with an Excel column chart depicting the probability distribution are shown in Figure 3.2.

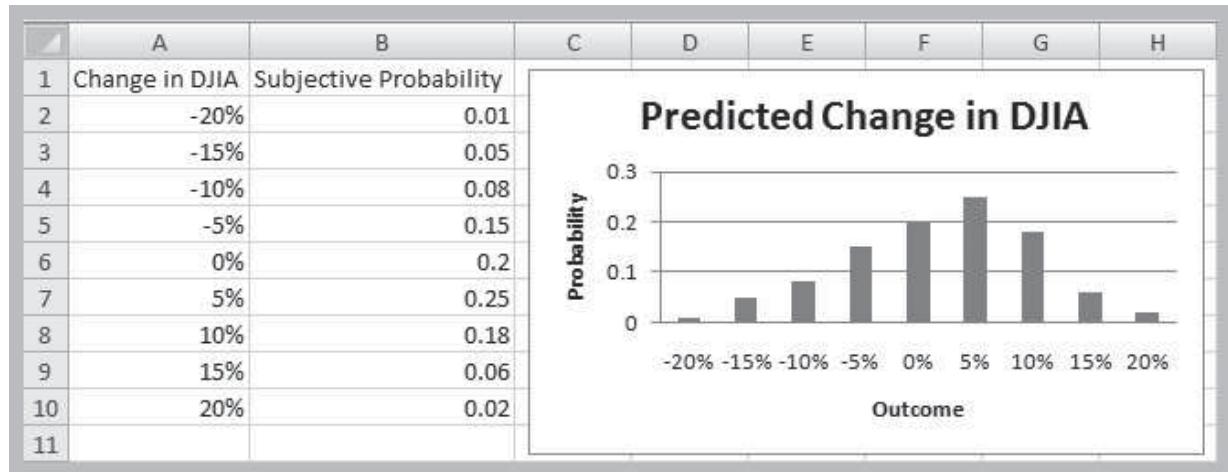
Second, we can calculate the relative frequencies from a sample of empirical data to develop a probability distribution. Figure 3.3 shows the distribution of weather in Seattle in January based on the data in the Excel file *Seattle Weather*. Because this is based on sample data, we usually call this an **empirical probability distribution**. An empirical probability distribution is an approximation of the probability distribution of the associated random



**FIGURE 3.2** Probability Distribution of Rolls of Two Dice



**FIGURE 3.3** Empirical Probability Distribution of Seattle Weather



**FIGURE 3.4** Subjective Probability Distribution of DJIA Change

variable, whereas the probability distribution of a random variable, such as one derived from counting arguments, is a theoretical model of the random variable.

Finally, we could simply specify a probability distribution using subjective values and expert judgment. This is often done in creating decision models for phenomena for which we have no historical data. We will see many examples of this in Part II of this book. Figure 3.4 shows a hypothetical example of the distribution of one expert's assessment of the how the DJIA might change in the next year.

Researchers have identified many common types of probability distributions that are useful in a variety of applications. A working knowledge of common families

**TABLE 3.1** Basic Probability Distribution Support in Excel

		Description
<b>EXCEL 2010 FUNCTION</b>		
BINOM.DIST( <i>number_s</i> , <i>trials</i> , <i>probability_s</i> , <i>cumulative</i> )		Returns the individual term binomial distribution
POISSON.DIST( <i>x</i> , <i>mean</i> , <i>cumulative</i> )		Returns the Poisson distribution
NORM.DIST( <i>x</i> , <i>mean</i> , <i>standard_cumulative</i> )		Returns the normal cumulative distribution for <i>deviation</i> , the specified mean and standard deviation
NORM.S.DIST( <i>z</i> )		Returns the standard normal cumulative distribution ( <i>mean</i> = 0, standard deviation = 1)
NORM.INV( <i>probability</i> , <i>mean</i> , <i>standard_dev</i> )		Returns the inverse of the normal cumulative distribution for the specified mean and standard deviation
NORM.S.INV( <i>probability</i> )		Returns the inverse of the standard normal cumulative distribution
STANDARDIZE( <i>x</i> , <i>mean</i> , <i>standard_deviation</i> )		Returns a normalized value for a distribution characterized by a mean and standard deviation
EXPON.DIST( <i>x</i> , <i>lambda</i> , <i>cumulative</i> )		Returns the exponential distribution
<b>PHSTAT ADD-IN</b>		
Binomial Probabilities		Computes binomial probabilities and histogram
Poisson Probabilities		Computes Poisson probabilities and histogram
Normal Probabilities		Computes normal probabilities
Exponential Probabilities		Computes exponential probabilities
Simple and Joint Probabilities		Computes simple and joint probabilities for a $2 \times 2$ cross-tabulation

of probability distributions is important for several reasons. First, it can help you to understand the underlying process that generates sample data. We will investigate the relationship between distributions and samples later in this chapter. Second, many phenomena in business and nature follow some theoretical distribution and, therefore, are useful in building decision models. In Chapter 9, we will discuss how to fit sample data to the best theoretical distribution. Finally, working with distributions is essential in computing probabilities of occurrence of outcomes to assess risk and make decisions.

We will focus our discussions on applying Excel and *PHStat* in working with probability distributions. Appendix 3.1 provides some of the underlying theory and analytical calculations associated with these procedures. Excel and *PHStat* have a variety of functions and tools for working with many of the distributions that we will introduce in this chapter. The most useful ones are summarized in Table 3.1.

## DISCRETE PROBABILITY DISTRIBUTIONS

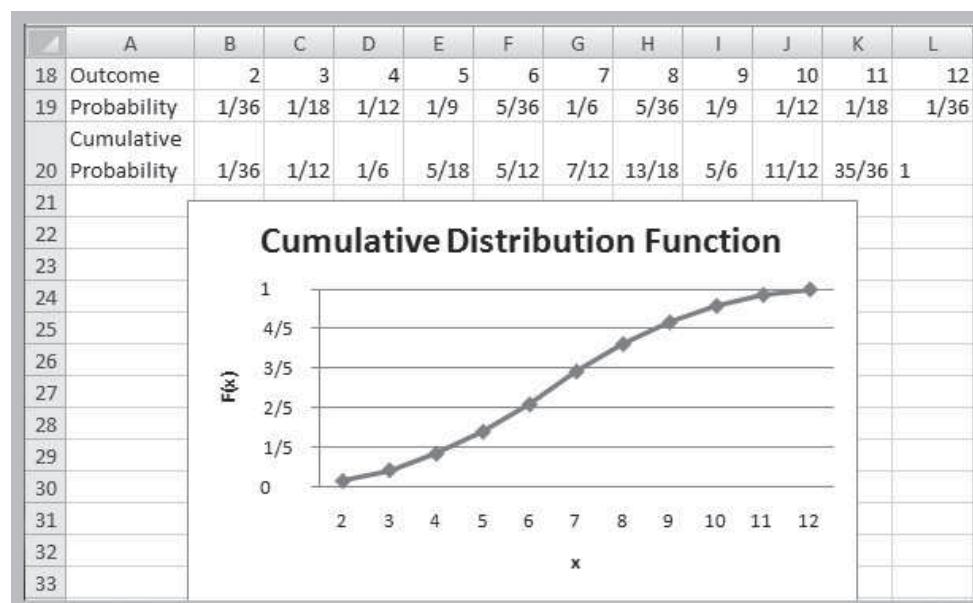
For a discrete random variable  $X$ , the probability distribution of the discrete outcomes is called a **probability mass function**, and is denoted by a mathematical function  $f(x)$ . The symbol  $x_i$  represents the  $i$ th value of the random variable  $X$  and  $f(x_i)$  is the probability. For instance, in Figure 3.2 for the dice example,  $x_1 = 2$  and  $f(x_1) = 1/36$ ;  $x_2 = 3$  and  $f(x_2) = 1/18$ , and so on. A probability mass function has the properties that (1) the probability of each outcome must be between 0 and 1 and (2) the sum of all probabilities must add to 1; that is:

$$0 \leq f(x_i) \leq 1 \quad \text{for all } i \quad (3.6)$$

$$\sum_i f(x_i) = 1 \quad (3.7)$$

You may easily verify that this holds in each of the examples in Figures 3.2–3.4.

A **cumulative distribution function**,  $F(x)$ , specifies the probability that the random variable  $X$  will assume a value *less than or equal to* a specified value,  $x$ . This is also denoted as  $P(X \leq x)$  and read as “the probability that the random variable  $X$  is less than or equal to  $x$ .” For example, the cumulative distribution function for rolling two dice is shown in Figure 3.5 along with an Excel line chart. To use this, suppose we want



**FIGURE 3.5** Cumulative Distribution Function for Rolling Dice

to know the probability of rolling a 6 or less. We simply look up the cumulative probability for 6, which is 5/12. Alternatively, we could locate the point for  $x = 6$  in the chart and estimate the probability from the graph. Also note that if the probability of rolling a 6 or less is 5/12, then the probability of the complementary event (rolling a 7 or more) is  $1 - 5/12 = 7/12$ . We can also use the cumulative distribution function to find probabilities over intervals. For example, to find the probability of rolling a number between 4 and 8,  $P(4 \leq X \leq 8)$ , we can find  $P(X \leq 8)$  and subtract  $P(X \leq 3)$ , that is:

$$P(4 \leq X \leq 8) = P(X \leq 8) - P(X \leq 3) = 13/18 - 1/12 = 23/36$$

*A word of caution.* Be careful with the endpoints because 4 is included in the interval we wish to compute, and we need to subtract  $P(X \leq 3)$ , not  $P(X \leq 4)$ !

### Expected Value and Variance of a Discrete Random Variable

The **expected value** of a random variable corresponds to the notion of the mean, or average, for a sample. For a discrete random variable  $X$ , the expected value, denoted as  $E[X]$ , is the weighted average of all possible outcomes, where the weights are the probabilities:

$$E[X] = \sum x_i f(x_i) \quad (3.8)$$

Applying this formula to Figure 3.2, we see that the expected value of rolling two dice is

$$\begin{aligned} E[X] &= 2(1/36) + 3(1/18) + 4(1/12) + 5(1/9) + 6(5/36) + 7(1/6) + 8(5/36) \\ &\quad + 9(1/9) + 10(1/12) + 11(1/18) + 12(1/36) = 7 \end{aligned}$$

As another example, suppose that you play a lottery in which you buy a ticket for \$50 and are told you have a 1 in 1,000 chance of winning \$25,000. The random variable  $X$  is your net winnings, and its probability distribution is:

$x$	$f(x)$
-\$50	0.999
\$24,950	0.001

The expected value,  $E[X]$ , is  $-$50(0.999) + \$24,950(0.001) = -$25$ . This means that if you played this game repeatedly over the long run, you would lose an average of \$25 each time you play. Of course, for any *one* game you would either lose \$50 or win \$24,950.

We may also compute the variance,  $V[X]$ , of a discrete random variable  $X$  as a weighted average of the squared deviations from the expected value:

$$V[X] = \sum (x - E[X])^2 f(x) \quad (3.9)$$

For the lottery example, the variance is calculated as:

$$V[X] = (-50 - [-25])^2(0.999) + (24,950 - [-25])^2(.001) = 624,375$$

Similar to our discussion in Chapter 2, the variance measures the uncertainty of the random variable; the higher the variance, the higher the uncertainty of the outcome. Although variances are easier to work with mathematically, we usually measure the variability of a random variable by its standard deviation, which is simply the square root of the variance. Thus, the standard deviation for the lottery example is  $\sqrt{624,375} = 790.17$ .

## SKILL-BUILDER EXERCISE 3.2

Develop a spreadsheet for computing the expected value and variance for the probability distribution of the rolls of two dice (see Figure 3.2).

### Bernoulli Distribution

The *Bernoulli distribution* characterizes a random variable having two possible outcomes, each with a constant probability of occurrence. Typically, these outcomes represent “success” ( $x = 1$ ) having probability  $p$ , and “failure” ( $x = 0$ ) having probability  $1 - p$ . A “success” can be any outcome you define. For example, in attempting to boot a new computer just off the assembly line, we might define a “success” as “does not boot up” in defining a Bernoulli random variable to characterize the probability distribution of failing to boot. Thus, “success” need not be a positive result in the traditional sense.

The probability mass function of the Bernoulli distribution is:

$$f(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases} \quad (3.10)$$

where  $p$  represents the probability of success. The expected value is  $p$ , and the variance is  $p(1 - p)$ .

A Bernoulli distribution might be used to model whether an individual responds positively ( $x = 1$ ) or negatively ( $x = 0$ ) to a telemarketing promotion. For example, if you estimate that 20% of customers contacted will make a purchase, the probability distribution that describes whether or not a particular individual makes a purchase is Bernoulli with  $p = 0.2$ . Think of the following experiment. Suppose that you have a box with 100 marbles: 20 red and 80 white. For each customer, select one marble at random (and then replace it). The outcome will have a Bernoulli distribution. If a red marble is chosen, then that customer makes a purchase; if it is white, the customer does not make a purchase.

### Binomial Distribution

The *binomial distribution* models  $n$  independent replications of a Bernoulli experiment, each with a probability  $p$  of success. The random variable  $X$  represents the number of successes in these  $n$  experiments. In the telemarketing example, suppose that we call  $n = 10$  customers, each of which has a probability  $p = 0.2$  of making a purchase. Then the probability distribution of the number of positive responses obtained from 10 customers is binomial. Using the binomial distribution, we can calculate the probability that exactly  $x$  customers out of the 10 will make a purchase.

The formula for the probability mass function for the binomial distribution is rather complex (see Appendix 3.1), and binomial probabilities are tedious to compute by hand; however, they can be computed in Excel 2010 easily using the function:

`BINOM.DIST(number_s, trials, probability_s, cumulative)`

In this function, *number\_s* plays the role of  $x$ , and *probability\_s* is the same as  $p$ . If *cumulative* is set to TRUE, then this function will provide cumulative probabilities; otherwise the default is FALSE, and it provides values of the probability mass function,  $f(x)$ .

Figure 3.6 shows the results of using this function to compute the distribution for this example. For instance, the probability that exactly four individuals will make a purchase is  $\text{BINOM.DIST}(4,10,0.2,\text{FALSE}) = 0.088080 = f(4)$ , and the probability that four

Binomial Probabilities					
n	10		=BINOMDIST(A7,\$B\$3,\$B\$4, FALSE)		
p	0.2		=BINOMDIST(A7,\$B\$3,\$B\$4,TRUE)		
x	f(x)	F(x)			
0	0.107374	0.107374			
1	0.268435	0.375810			
2	0.301990	0.677800			
3	0.201327	0.879126			
4	0.088080	0.967207			
5	0.026424	0.993631			
6	0.005505	0.999136			
7	0.000786	0.999922			
8	0.000074	0.999996			
9	0.000004	1.000000			
10	0.000000	1.000000			

**FIGURE 3.6** Computing Binomial Probabilities in Excel

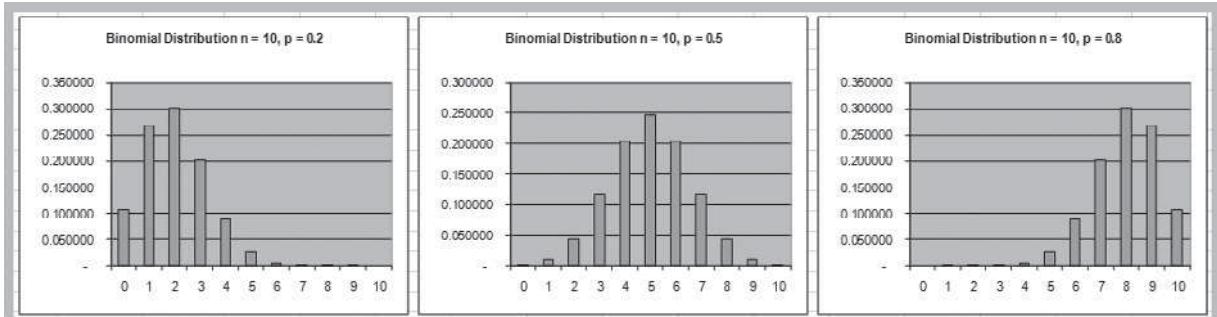
or less individuals will make a purchase is  $\text{BINOM.DIST}(4,10,0.2,\text{TRUE}) = 0.967207 = F(4)$ . Correspondingly, the probability that more than 4 out of 10 individuals will make a purchase is  $1 - F(4) = 1 - 0.967207 = 0.032793$ . A binomial distribution might also be used to model the results of sampling inspection in a production operation or the effects of drug research on a sample of patients.

The expected value of the binomial distribution is  $np$ , and the variance is  $np(1 - p)$ . The binomial distribution can assume different shapes and amounts of skewness, depending on the parameters. Figure 3.7 shows two examples. When  $p = 0.5$ , the distribution is symmetric. For larger values of  $p$ , the binomial distribution is negatively skewed; for smaller values, it is positively skewed.

### Poisson Distribution

The Poisson distribution is a discrete distribution used to model the number of occurrences in some unit of measure, for example, the number of events occurring in an interval of time, number of items demanded per customer from an inventory, or the number of errors per line of software code. For example, suppose that the average number of customers arriving at an ATM during lunch hour is 12 customers per hour. The probability that exactly  $x$  customers will arrive during the hour is given by a Poisson distribution with a mean of 12.

The Poisson distribution assumes no limit on the number of occurrences (meaning that the random variable  $X$  may assume any nonnegative integer value), that occurrences are independent, and that the average number of occurrences per unit is a constant,  $\lambda$  (Greek lowercase lambda). The expected value of the Poisson distribution is  $\lambda$ , and the variance also is equal to  $\lambda$ .

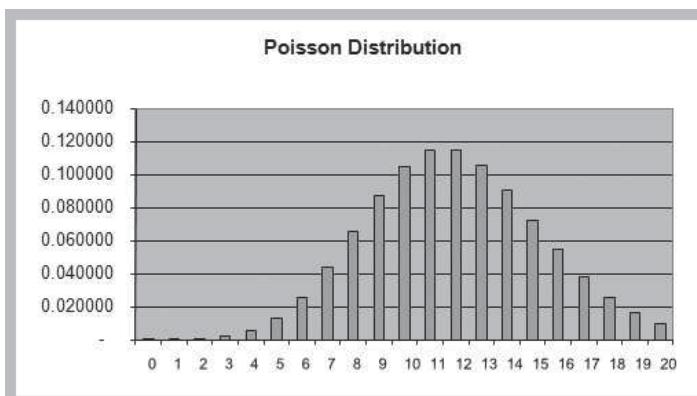


**FIGURE 3.7** Examples of the Binomial Distribution

A	B	C	D	E	F
1	Poisson Probabilities				
2			=POISSON(A7,\$B\$3,FALSE)		
3	Mean	12		=POISSON(A7,\$B\$3,TRUE)	
6	x	f(x)	F(x)		
7	0	0.000006	0.000006		
8	1	0.000074	0.000080		
9	2	0.000442	0.000522		
10	3	0.001770	0.002292		
11	4	0.005309	0.007600		
12	5	0.012741	0.020341		
13	6	0.025481	0.045822		
14	7	0.043682	0.089504		
15	8	0.065523	0.155028		
16	9	0.087364	0.242392		
17	10	0.104837	0.347229		
18	11	0.114368	0.461597		
19	12	0.114368	0.575965		
20	13	0.105570	0.681536		
21	14	0.090489	0.772025		
22	15	0.072391	0.844416		
23	16	0.054293	0.898709		
24	17	0.038325	0.937034		
25	18	0.025550	0.962584		
26	19	0.016137	0.978720		
27	20	0.009682	0.988402		

**FIGURE 3.8** Computing Poisson Probabilities in Excel

Like the binomial, Poisson probabilities are cumbersome to compute by hand (see Appendix 3.1). Probabilities can easily be computed in Excel using the function  $\text{POISSON.DIST}(x, \text{mean}, \text{cumulative})$ . Figure 3.8 shows the results of using this function to compute the distribution for the ATM example with  $\lambda = 12$ . Thus, the probability of exactly one arrival during the lunch hour is calculated by the Excel function  $=\text{POISSON.DIST}(1,12,\text{FALSE}) = 0.000074 = f(1)$ ; the probability of 4 arrivals or less is calculated by  $=\text{POISSON.DIST}(4,12,\text{TRUE}) = 0.007600 = F(4)$ , and so on. Because the possible values of a Poisson random variable are infinite, we have not shown the complete distribution in Figure 3.8. As  $x$  gets large, the probabilities become quite small. Figure 3.9 shows this Poisson distribution. Like the binomial, the specific shape depends on the value of the parameter  $\lambda$ ; the distribution is more skewed for smaller values.



**FIGURE 3.9** Poisson Distribution for  $\lambda = 12$

### SKILL-BUILDER EXERCISE 3.3

Construct charts for the probability mass function for a Poisson distribution with means of 1, 5, and 20. Compare these to Figure 3.8. What conclusion can you reach?

## CONTINUOUS PROBABILITY DISTRIBUTIONS

As we noted earlier, a continuous random variable is defined over one or more intervals of real numbers, and therefore, has an infinite number of possible outcomes. Suppose that the expert who predicted the probabilities associated with next year's change in the DJIA in Figure 3.4 kept refining the estimates over larger and larger ranges of values. Figure 3.10 shows what such a probability distribution might look like using 2.5% increments rather than 5%. Notice that the distribution is similar in shape to the one in Figure 3.4 but simply has more outcomes. If this refinement process continues, then the distribution will approach the shape of a smooth curve as shown in the figure. Such a curve that characterizes outcomes of a continuous random variable is called a **probability density function**, and is described by a mathematical function  $f(x)$ . A probability density function has the properties given below:

- (1)  $f(x) \geq 0$  for all values of  $x$
- (2) The total area under the function above the  $x$ -axis is 1.0

A simple example of a probability density function is shown in Figure 3.11. Suppose that the random variable  $X$  represents the amount of overflow (in ounces) from a cereal box-filling machine. The random variable  $X$  is defined over the interval from 0 to 1, and the form of the density function suggests that the probability of overfilling a box gets

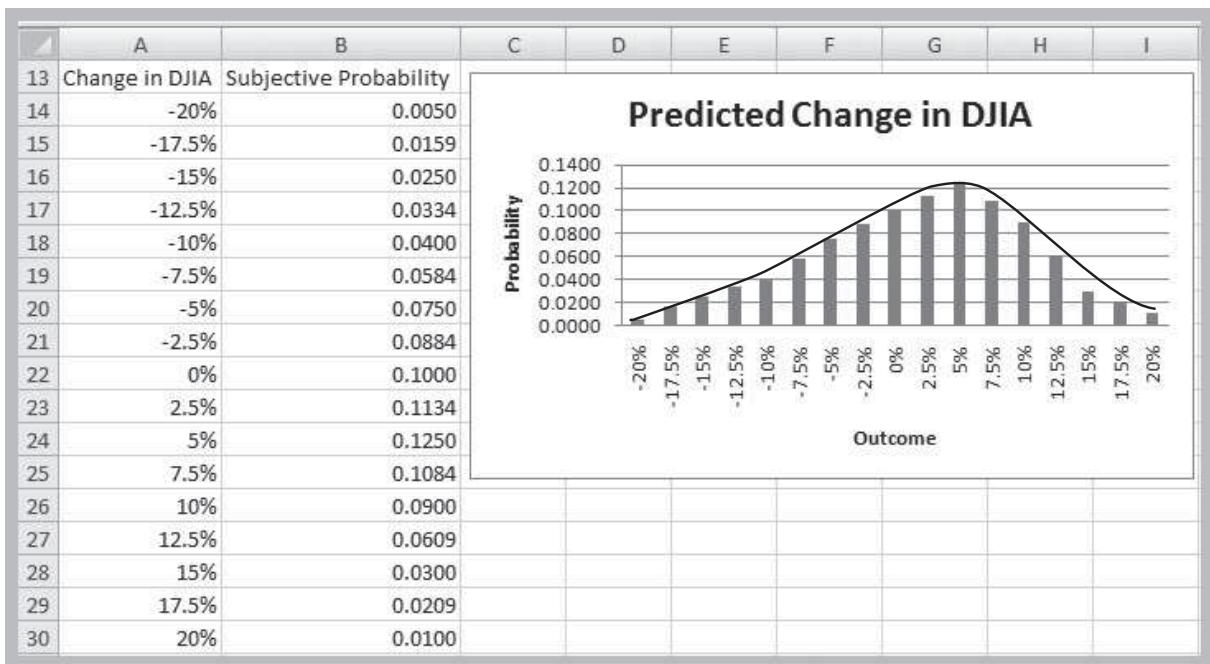
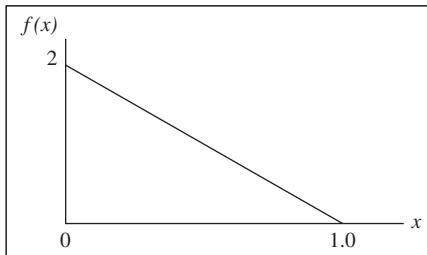


FIGURE 3.10 Refined Probability Distribution of DJIA Change



**FIGURE 3.11** A Continuous Probability Distribution

smaller for larger values of  $X$ . Using algebra, you should be able to determine that the mathematical form of this density function is:

$$f(x) = -2x + 2 \quad \text{for } 0 \leq x \leq 1 \quad (3.11)$$

Note that  $f(x) \geq 0$  for all values of  $x$  between 0 and 1, and that the area under the function above the  $x$ -axis is 1.0 (since the area of a triangle is one-half the altitude times the base, or  $0.5 \times 2 \times 1 = 1$ ).

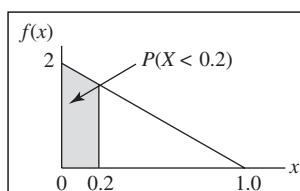
For continuous random variables, it does not make mathematical sense to attempt to define a probability for a specific value of  $x$  because there are an infinite number of values; thus,  $P(X = x) = 0$ . For continuous random variables, probabilities are only defined over intervals, such as  $P(a \leq X \leq b)$  or  $P(X > c)$ . The probability that a random variable will assume a value between  $a$  and  $b$  is given by the area under the density function between  $a$  and  $b$ .

Thus, in the cereal filling example, we may compute the probability that the overfill will be less than 0.2 ounces, greater than 0.8 ounces, or between 0.2 and 0.8 ounces. For example, Figure 3.12 shows the probability that the overfill will be less than 0.2 ounces. From equation (3.11), we see that the height of the density function at  $x = 0.2$  is  $f(0.2) = -2(0.2) + 2 = 1.6$ . Using geometry, the area of the triangular region from  $x = 0.2$  to  $x = 1$  is  $0.5(1.6)(0.8) = 0.64$ . Therefore, the probability that the overfill will be less than 0.2 ounces is  $1 - 0.64 = 0.36$ . We leave it as an exercise to find the other probabilities using similar arguments.

The cumulative distribution function for a continuous random variable is denoted the same way as for discrete random variables,  $F(x)$ , and represents the probability that the random variable  $X$  is less than or equal to  $x$ ,  $P(X \leq x)$ . Intuitively,  $F(x)$  represents the area under the density function to the left of  $x$ .  $F(x)$  can often be derived mathematically from  $f(x)$  using techniques of calculus. For the cereal overfilling example, we can show that  $F(x) = -x^2 + 2x$ . Thus,  $P(X < 0.2) = F(0.2) = -(0.2)^2 + 2(0.2) = 0.36$  as we found using geometry.

Knowing  $F(x)$  makes it easy to compute probabilities over intervals for continuous distributions. The probability that  $X$  is between  $a$  and  $b$  is equal to the difference of the cumulative distribution function evaluated at these two points; that is:

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) \quad (3.12)$$



**FIGURE 3.12** Probability that  $X$  is Less than 0.2

For continuous distributions we need not be concerned about the endpoints as we were with discrete distributions because  $P(a \leq X \leq b)$  is the same as  $P(a < X < b)$ .

The formal definitions of expected value and variance for a continuous random variable are similar; however, to understand them, we must rely on notions of calculus, so we will not discuss them in this book. It is important to understand that the expected value, variance, and standard deviation of random variables are not sample statistics like the mean, sample variance, and sample standard deviation we introduced in Chapter 2. Rather, they are measures associated with the set of *all* possible outcomes of the random variable.

Continuous probability distributions depend on one or more parameters. Many continuous distributions can assume different shapes and sizes, depending on the value of the parameters, similar to what we will see for the binomial and Poisson. There are three basic types of parameters. A **shape parameter** controls the basic shape of the distribution. For certain distributions, changing the shape parameter will cause major changes in the form of the distribution. For others, the changes will be less severe. A **scale parameter** controls the unit of measurement within the range of the distribution. Changing the scale parameter either contracts or expands the distribution along the horizontal axis. Finally, a **location parameter** specifies the location of the distribution relative to zero on the horizontal axis. The location parameter may be the midpoint or the lower endpoint of the range of the distribution. Not all distributions will have all three parameters; some may have more than one shape parameter. Understanding the effects of these parameters is important in selecting distributions as inputs to decision models. The distributions we have chosen to describe are incorporated into the *Crystal Ball* software that we will use in Part II of this book when discussing decision modeling and risk analysis.

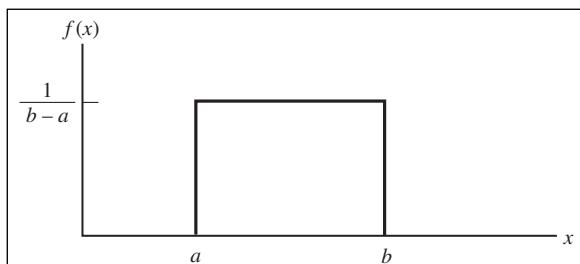
## Uniform Distribution

The uniform distribution characterizes a continuous random variable for which all outcomes between some minimum value  $a$  and maximum value  $b$  are equally likely. The density function for the uniform distribution is shown in Figure 3.13. You can easily verify that the area under the density function is one by using simple geometry. The expected value and variance of a uniform random variable  $X$  are computed as follows:

$$EV[X] = (a + b)/2 \quad (3.13)$$

$$V[X] = (b - a)^2/12 \quad (3.14)$$

Note that  $a$  can be considered to be a location parameter since it controls the location of the distribution along the horizontal axis. If  $a$  is fixed, the value of  $b$  plays the role of a scale parameter. Increasing  $b$  elongates the distribution; decreasing  $b$  compresses it. There is no shape parameter since any uniform distribution is flat. A variation of the



**FIGURE 3.13** Uniform Probability Density Function

uniform distribution is one for which the random variable is restricted to be integer values between  $a$  and  $b$  (also integers); this is called a **discrete uniform distribution**.

The uniform distribution is often used when little knowledge about a random variable is available; the parameters  $a$  and  $b$  are chosen judgmentally to reflect a modeler's best guess about the range of the random variable. Although Excel does not provide a function to compute uniform probabilities, the formula is simple enough to incorporate into a spreadsheet. See Appendix 3.1 for an example.

## Normal Distribution

The *normal distribution* is a continuous distribution that is described by the familiar bell-shaped curve and is perhaps the most important distribution used in statistics. The normal distribution is observed in many natural phenomena. Errors of various types, such as deviations from specifications of machined items, often are normally distributed. Thus, the normal distribution finds extensive applications in quality control. Processing times in some service systems also follow a normal distribution. Another useful application is that the distribution of the averages of random variables having any distribution tends to be normal as the number of random variables increases.

The normal distribution is characterized by two parameters: the mean,  $\mu$  (the location parameter), and the variance,  $\sigma^2$  (the scale parameter). Thus, as  $\mu$  changes, the location of the distribution on the  $x$ -axis also changes, and as  $\sigma^2$  is decreased or increased, the distribution becomes narrower or wider, respectively. Figure 3.14 provides a sketch of a special case of the normal distribution called the **standard normal distribution**—the normal distribution with  $\mu = 0$  and  $\sigma^2 = 1$ . This distribution is important in performing many probability calculations. A standard normal random variable is usually denoted by  $Z$ , and its density function by  $f(z)$ . The scale along the  $z$ -axis represents the number of standard deviations from the mean of 0.

The normal distribution is symmetric and has the property that the median equals the mean. Thus, half the area falls above the mean and half below it. Although the range of  $x$  is unbounded, meaning that the tails of the distribution extend to negative and positive infinity, most of the density is close to the mean; in fact, over 99% of the area is within (i.e., plus or minus) three standard deviations of the mean.

Two Excel 2010 functions are used to compute normal probabilities: NORM.DIST( $x, mean, standard\_deviation, cumulative$ ) and NORM.S.DIST( $z$ ). NORM.DIST( $x, mean, standard\_deviation, TRUE$ ) calculates the cumulative probability  $F(x) = P(X \leq x)$  for a specified mean and standard deviation. (If *cumulative* is set to *FALSE*, the function simply calculates the value of the density function  $f(x)$ , which has little practical application.) NORM.S.DIST( $z$ ) generates the cumulative probability for a standard normal distribution.

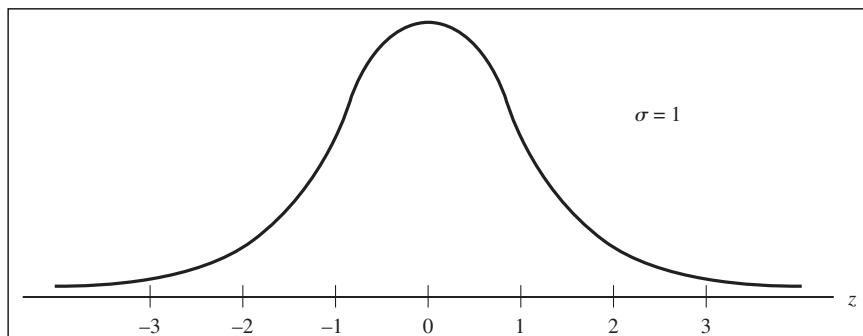


FIGURE 3.14 Standard Normal Distribution

A	B	C	D	E	F
1 Normal Probabilities					
2					
3 Mean	750				
4 Standard Deviation	100				
5					=NORMDIST(A7,\$B\$3,\$B\$4,TRUE)
6 x		F(x)			
7 500	0.0062				
8 550	0.0228				
9 600	0.0668				
10 650	0.1587				
11 700	0.3085				
12 750	0.5000				
13 800	0.6915				
14 850	0.8413				
15 900	0.9332				
16 950	0.9772				
17 1000	0.9938				

**FIGURE 3.15** Normal Probability Calculations

To illustrate the application of the normal distribution, suppose that a company has determined that the distribution of customer demand ( $X$ ) is normal with a mean of 750 units/month and a standard deviation of 100 units/month. Figure 3.15 shows some cumulative probabilities calculated with the NORM.DIST function. The company would like to know the following:

1. What is the probability that demand will be at most 900 units?
2. What is the probability that demand will exceed 700 units?
3. What is the probability that demand will be between 700 and 900 units?
4. What level of demand would be exceeded at most 10% of the time?

To answer the questions, first draw a picture. Figure 3.16(a) shows the probability that demand will be at most 900 units, or  $P(X < 900)$ . This is simply the cumulative probability for  $x = 900$ , which can be calculated using the Excel function =NORM.DIST(900,750,100,TRUE) = 0.9332.

Figure 3.16(b) shows the probability that demand will exceed 700 units,  $P(X > 700)$ . Using the principles we have previously discussed, this can be found by subtracting  $P(X < 700)$  from 1, or:

$$P(X > 700) = 1 - P(X < 700) = 1 - F(700) = 1 - 0.3085 = 0.6915$$

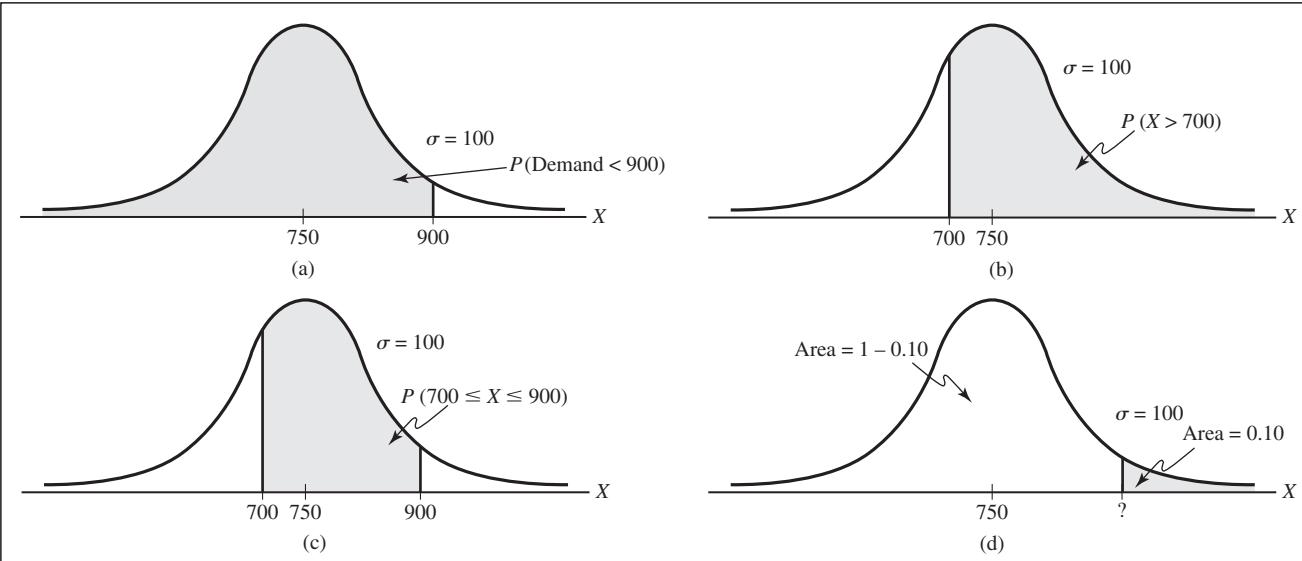
This can be computed in Excel using the formula =1 - NORM.DIST(700,750,100,TRUE).

The probability that demand will be between 700 and 900,  $P(700 < X < 900)$ , is illustrated in Figure 3.16(c). This is calculated by:

$$\begin{aligned} P(700 < X < 900) &= P(X < 900) - P(X < 700) \\ &= F(900) - F(700) = 0.9332 - 0.3085 = 0.6247 \end{aligned}$$

In Excel, we would use the formula =NORM.DIST(900,750,100,TRUE) - NORM.DIST(700,750,100,TRUE).

The third question is a bit tricky. We wish to find the level of demand that will be exceeded only 10% of the time; that is, find the value of  $x$  so that  $P(X > x) = 0.10$ .



**FIGURE 3.16** Computing Normal Probabilities

This is illustrated in Figure 3.16(d). An upper-tail probability of 0.10 is equivalent to a cumulative probability of 0.90. From Figure 3.15, we can see that the correct value must be somewhere between 850 and 900 because  $F(850) = 0.8413$  and  $F(900) = 0.9332$ . The Excel function `NORM.INV(probability, mean, standard_dev)` can be used to find the exact value. In this function, *probability* is the cumulative probability value corresponding to the value of  $x$  we seek. Thus, the proper formula is `=NORM.INV(0.90,750,100) = 878.155`; a demand of approximately 878 will satisfy the criterion.

All of these questions can also be answered using the *PHStat Normal* probability tool. See Appendix 3.2A on *Normal Probability Tools*.

The Excel function `NORM.S.DIST(z)` finds probabilities for the standard normal distribution. To illustrate the use of this function, let us find the areas under the standard normal distribution within one, two, and three standard deviations of the mean. Figure 3.17 shows calculations using `NORM.S.DIST`. Therefore, the probability within one standard deviation of the mean is found using the Excel formula `=NORM.S.DIST(1) - NORM.S.DIST(-1) = P(-1 < Z < 1) = 0.8413 - 0.1587 = 0.6826`. In a similar fashion,



Spreadsheet Note

A	B	C	D	E
1	Standard Normal Probabilities			
2				
3	$z$	$F(z)$		$=NORMSDIST(A4)$
4	-3	0.0013		
5	-2	0.0228		
6	-1	0.1587		
7	0	0.5000		
8	1	0.8413		
9	2	0.9772		
10	3	0.9987		

**FIGURE 3.17** Standard Normal Probabilities Using the `NORMSDIST` Function

you should verify that the area within two standard deviations of the mean is 0.9544 and the area within three standard deviations of the mean is 0.9973. Notice that these values are about what we described as the empirical rules in Chapter 2. These are important characteristics of the normal distribution.

As a final note, we can use the normal distribution with  $\mu = np$  and  $\sigma^2 = np(1 - p)$  to approximate the binomial distribution. This approximation holds well when  $np \geq 5$  and  $n(1 - p) \geq 5$ .

### SKILL-BUILDER EXERCISE 3.4

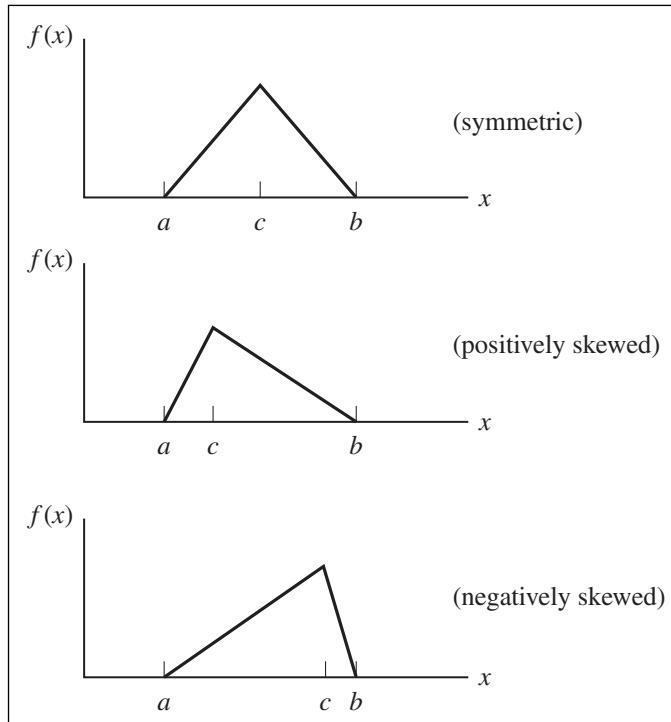
Develop a spreadsheet to compute the probability over any interval for the normal distribution, using any input values of the mean, standard deviation, and endpoints of the interval.

### Triangular Distribution

The triangular distribution is a continuous distribution defined by three parameters: the minimum,  $a$ ; maximum,  $b$ ; and most likely,  $c$ . Outcomes near the most likely value have a higher chance of occurring than those at the extremes. By varying the position of the most likely value relative to the extremes, the triangular distribution can be symmetric or skewed in either direction, as shown in Figure 3.18. From Figure 3.18, you can see that  $a$  is the location parameter,  $b$  is the scale parameter, and  $c$  is the shape parameter. The expected value and the variance of a triangular random variable  $X$  are given by the following:

$$EV[X] = (a + b + c)/3 \quad (3.15)$$

$$Var[X] = (a^2 + b^2 + c^2 - ab - ac - bc)/18 \quad (3.16)$$



**FIGURE 3.18** Examples of Triangular Distributions

The triangular distribution is often used as a rough approximation of other distributions, such as the normal, or when no data are available and a distribution must be assumed judgmentally. Because it depends on three simple parameters and can assume a variety of shapes—for instance, it can be skewed in either direction by changing the value of  $c$ —it is very flexible in modeling a wide variety of assumptions. One drawback, however, is that it is bounded, thereby eliminating the possibility of extreme values that might possibly occur. Excel does not have a function to compute triangular probabilities.

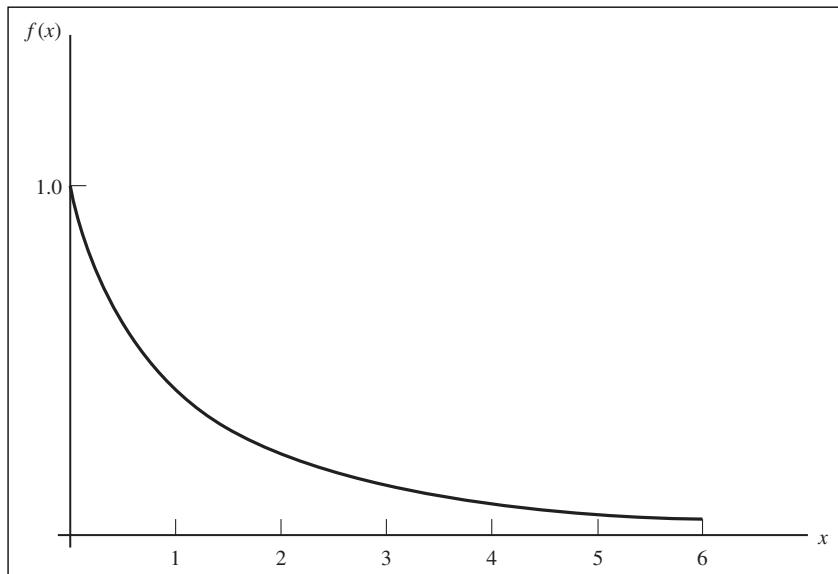
## Exponential Distribution

The exponential distribution is a continuous distribution that models the time between randomly occurring events. Thus, it is often used in such applications as modeling the time between customer arrivals to a service system, or the time to failure of machines, lightbulbs, and other mechanical or electrical components. A key property of the exponential distribution is that it is *memoryless*; that is, the current time has no effect on future outcomes. For example, the length of time until a machine failure has the same distribution no matter how long the machine has been running.

Similar to the Poisson distribution, the exponential distribution has one parameter  $\lambda$ . In fact, the exponential distribution is closely related to the Poisson; if the number of events occurring during an interval of time has a Poisson distribution, then the time between events is exponentially distributed. For instance, if the number of arrivals at a bank is Poisson distributed, say with mean 12/hour, then the time between arrivals is exponential, with mean 1/12 hour, or 5 minutes.

The expected value of the exponential distribution =  $1/\lambda$  and the variance =  $(1/\lambda)^2$ . The exponential distribution has no shape or location parameters;  $\lambda$  is the scale parameter. Figure 3.19 provides a sketch of the exponential distribution. The exponential distribution has the properties that it is bounded below by 0, it has its greatest density at 0, and the density declines as  $x$  increases.

The Excel function EXPON.DIST( $x$ , *lambda*, cumulative) can be used to compute exponential probabilities. To illustrate the exponential distribution, suppose that the mean time to failure of a critical component of an engine is  $1/\lambda = 8,000$  hours. Figure 3.20



**FIGURE 3.19** Example of an Exponential Distribution ( $\lambda = 1$ )

	A	B	C	D	E	F
1	Exponential Probabilities					
2						
3	Mean	8000				
4						=EXPONDIST(A6,1/\$B\$3,TRUE)
5	x	F(x)				
6	1000	0.117503				
7	2000	0.221199				
8	3000	0.312711				
9	4000	0.393469				
10	5000	0.464739				
11	6000	0.527633				
12	7000	0.583138				
13	8000	0.632121				
14	9000	0.675348				
15	10000	0.713495				
16	11000	0.74716				
17	12000	0.77687				
18	13000	0.803088				
19	14000	0.826226				
20	15000	0.846645				

**FIGURE 3.20** Exponential Probabilities in Excel

shows a portion of the cumulative distribution function. Note that we used the mean in cell B3 as an input in the worksheet; thus, we entered  $\lambda$  as  $1/\$B$3$  in the EXPON.DIST function. The probability that the component will fail before  $x$  hours is given by the cumulative distribution function  $F(x)$ . For example, the probability of failing before 5,000 hours is  $F(5000) = 0.465$ .

### Probability Distributions in PHStat

*PHStat* has several routines for generating probabilities of distributions we have discussed. This allows you to compute probabilities without requiring you to develop a detailed worksheet in Excel. The following distributions are available:

- Normal
- Binomial
- Exponential
- Poisson
- Hypergeometric (which we do not cover in this text)



Spreadsheet Note

Appendix 3.2B describes the binomial option; other probabilities can be generated in a similar fashion.

### Other Useful Distributions

Many other probability distributions, especially those distributions that assume a wide variety of shapes, find application in decision modeling for characterizing a wide variety of phenomena. Such distributions provide a great amount of flexibility in representing both empirical data or when judgment is needed to define an appropriate distribution. We provide a brief description of these distributions; further details may be found in more advanced texts on probability and statistics.

- **Lognormal Distribution.** If the natural logarithm of a random variable  $X$  is normal, then  $X$  has a lognormal distribution. Because the lognormal distribution is positively skewed and bounded below by zero, it finds applications in modeling phenomena that have low probabilities of large values and cannot have negative values, such as the time to complete a task. Other common examples include stock prices and real estate prices. The lognormal distribution is also often used for “spiked” service times, that is, when the probability of 0 is very low but the most likely value is just greater than 0.

- **Gamma Distribution.** The gamma distribution is a family of distributions defined by a shape parameter  $\alpha$ , a scale parameter  $\beta$ , and a location parameter  $L$ .  $L$  is the lower limit of the random variable  $X$ ; that is, the gamma distribution is defined for  $X > L$ . Gamma distributions are often used to model the time to complete a task, such as customer service or machine repair. It is used to measure the time between the occurrence of events when the event process is not completely random. It also finds application in inventory control and insurance risk theory.

A special case of the gamma distribution when  $\alpha = 1$  and  $L = 0$  is called the *Erlang distribution*. The Erlang distribution can also be viewed as the sum of  $k$  independent and identically distributed exponential random variables. The mean is  $k/\lambda$ , and the variance is  $k/\lambda^2$ . When  $k = 1$ , the Erlang is identical to the exponential distribution. For  $k = 2$ , the distribution is highly skewed to the right. For larger values of  $k$ , this skewness decreases, until for  $k = 20$ , the Erlang distribution looks similar to a normal distribution. One common application of the Erlang distribution is for modeling the time to complete a task when it can be broken down into independent tasks, each of which has an exponential distribution.

- **Weibull Distribution.** The Weibull distribution is another probability distribution capable of taking on a number of different shapes defined by a scale parameter  $\alpha$ , a shape parameter  $\beta$ , and a location parameter  $L$ . Both  $\alpha$  and  $\beta$  must be greater than 0. When the location parameter  $L = 0$  and  $\beta = 1$ , the Weibull distribution is the same as the exponential distribution with  $\lambda = 1/\alpha$ . By choosing the location parameter  $L$  different from 0, you can model an exponential distribution that has a lower bound different from 0. When  $\beta = 3.25$ , the Weibull approximates the normal distribution. Weibull distributions are often used to model results from life and fatigue tests, equipment failure times, and times to complete a task.

- **Beta Distribution.** One of the most flexible distributions for modeling variation over a fixed interval from 0 to a positive value  $s$  is the beta. The beta distribution is a function of two shape parameters,  $\alpha$  and  $\beta$ , both of which must be positive. The parameter  $s$  is the scale parameter. Note that  $s$  defines the upper limit of the distribution range. If  $\alpha$  and  $\beta$  are equal, the distribution is symmetric. If either parameter is 1.0 and the other is greater than 1.0, the distribution is in the shape of a J. If  $\alpha$  is less than  $\beta$ , the distribution is positively skewed; otherwise, it is negatively skewed. These properties can help you to select appropriate values for the shape parameters.

- **Geometric Distribution.** This distribution describes the number of trials until the first success where the probability of a success is the same from trial to trial. An example would be the number of parts manufactured until a defect occurs, assuming that the probability of a defect is constant for each part.

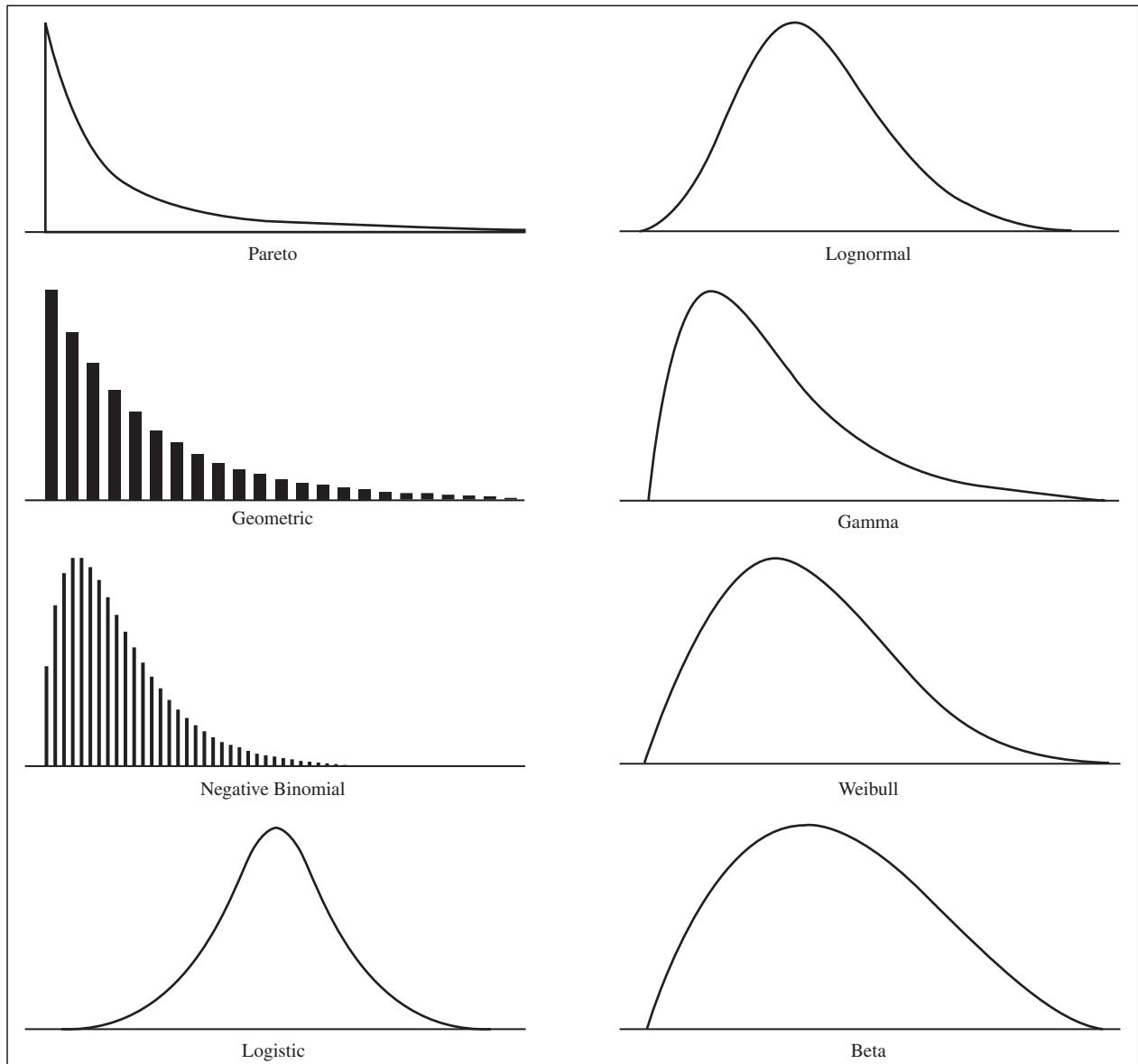
- **Negative Binomial Distribution.** Like the geometric distribution, the negative binomial distribution models the distribution of the number of trials until the  $r$ th success, for example, the number of sales calls needed to sell 10 orders.

- **Hypergeometric Distribution.** This is similar to the binomial, except that it applies to sampling without replacement. The hypergeometric distribution is often used in quality control inspection applications.

- **Logistic Distribution.** This is commonly used to describe growth of a population over time.

- **Pareto Distribution.** This describes phenomena in which a small proportion of items accounts for a large proportion of some characteristic. For example, a small number of cities constitutes a large proportion of the population. Other examples include the size of companies, personal incomes, and stock price fluctuations.
- **Extreme Value Distribution.** This describes the largest value of a response over a period of time, such as rainfall, earthquakes, and breaking strengths of materials.

Figure 3.21 is a graphical summary of some of the distributions we have discussed. Understanding the shapes of these distributions makes it easier to select an appropriate distribution for a decision-modeling application. Several other probability distributions, notably the  $t$ ,  $F$ , and chi-square ( $\chi^2$ ) distributions, are used in statistical inference applications but have limited application in modeling. We will introduce these as needed in the next several chapters.



**FIGURE 3.21** Shapes of Some Useful Probability Distributions

## JOINT AND MARGINAL PROBABILITY DISTRIBUTIONS

Many applications involve the simultaneous consideration of two or more random variables. For example, we might observe the movement of both the Dow Jones Industrial Average (DJIA) and NASDAQ stock indexes on a daily basis and whether the index was up, down, or unchanged. We might define a random variable  $X$  to be the outcome associated with the DJIA, where  $X = 1$  represents up,  $X = -1$  represents down, and  $X = 0$  represents unchanged; and define a random variable  $Y$  in a similar fashion for the NASDAQ. Suppose that we observed the outcomes associated with the stock market for a long period of time and recorded the percentage of days each of the following outcomes occurred:

1. DJIA up; NASDAQ up: 26%
2. DJIA down; NASDAQ up: 5%
3. DJIA unchanged; NASDAQ up: 3%
4. DJIA up; NASDAQ down: 10%
5. DJIA down; NASDAQ down: 42%
6. DJIA unchanged; NASDAQ down: 5%
7. DJIA up; NASDAQ unchanged: 4%
8. DJIA down; NASDAQ unchanged: 3%
9. DJIA unchanged; NASDAQ unchanged: 2%

A probability distribution that specifies the probabilities of outcomes of two different random variables,  $X$  and  $Y$ , that occur at the same time, or jointly, is called a **joint probability distribution**. A joint probability distribution for the changes in the DJIA and NASDAQ indexes would be:

		DJIA Up	DJIA Unchanged	DJIA Down	Marginal Probability
Nasdaq Up	Joint Probabilities	$X = 1$	$X = 0$	$X = -1$	
	$Y = 1$	0.26	0.03	0.05	0.29
Nasdaq unchanged	$Y = 0$	0.04	0.02	0.03	0.06
	$Y = -1$	0.10	0.05	0.42	0.15
Marginal Probability		0.40	0.10	0.50	1

We denote a joint probability distribution (for discrete random variables) as  $P(X = x, Y = y) = f(x, y)$ . For example, the probability that both indexes go up on one day is  $P(X = 1, Y = 1) = f(1,1) = 0.26$ . **Marginal probabilities** represent the probability associated with the outcomes of each random variable regardless of the value of the other. If we add the joint probabilities across the rows and down the columns, we obtain marginal probability distributions as shown in the table. The marginal probabilities in the right-hand column represent the probability distribution of  $Y$  independent of the value of  $X$ , and those in the bottom row represent the probability distribution of  $X$  independent of the value of  $Y$ . Thus, the probability that the NASDAQ goes up no matter what the DJIA does is 0.29.

Similar to the notion of independent events that we described earlier in this chapter, we may say that two random variables  $X$  and  $Y$  are **statistically independent** if  $f(x,y) = f(x)f(y)$  for all values of  $x$  and  $y$ . This means that the value of one random variable does not depend on the value of the other. This is easily determined by checking to see if the product of the marginal probabilities for every row and column equals the joint probability in the respective row and column. For example, if we multiply the marginal probability that the NASDAQ goes up (0.29) by the probability that the DJIA

goes up (0.40), we obtain  $(0.29)(0.40) = 0.116$ , which is not equal to the joint probability 0.26. Therefore, we may readily conclude that these two random variables are not statistically independent.

---

## Basic Concepts Review Questions

1. Explain the concepts of an experiment, outcome, sample space, and event. Provide some examples different from those in the chapter.
2. Define probability and explain its three perspectives. Provide an example of each.
3. What is the complement of an event and how do you find its probability?
4. Explain the concept of mutually exclusive events. How do you compute the probability  $P(A \text{ or } B)$  when  $A$  and  $B$  are, and are not, mutually exclusive?
5. Explain the notion of conditional probability and how it is computed.
6. What is meant by the independence of two events,  $A$  and  $B$ ? If  $A$  and  $B$  are mutually exclusive, can they be independent?
7. What is a random variable? Explain the difference between discrete and continuous random variables.
8. What is the multiplicative rule of probability?
9. Why can we not define probability for a point?
10. What are the basic rules of probability?
11. Briefly summarize the important characteristics and applications of the Bernoulli, binomial, and Poisson distributions.
12. What properties must a continuous probability distribution have? How are probabilities defined for continuous distributions?
13. Explain the role of shape, scale, and location parameters in continuous probability distributions.
14. Explain practical situations in which the uniform, normal, triangular, and exponential distributions might be used.
15. What is the standard normal distribution? How and under what conditions can the normal distribution be used to find binomial probabilities?
16. Describe the joint probability distribution of two random variables. How are marginal probabilities calculated? When are the random variables independent?

---

## Problems and Applications

1. Consider a regular deck of cards where the cards ace, jack, queen, and king are numbered 1, 11, 12, and 13 respectively. Consider the experiment of drawing two cards randomly from this deck without replacement.
  - a. Describe the outcomes of this experiment. List the elements of the sample space.
  - b. What is the probability of obtaining a total of 6 for the two cards?
  - c. Let  $A$  be the event "total card value is 4 or less". Find  $P(A)$  and  $P(A^c)$ .
2. A fair coin is tossed three times.
  - a. List all possible outcomes in the sample space. Find the probability associated with each outcome.
  - b. Let  $A$  be the event "exactly 3 heads." Find  $P(A)$ .
  - c. Let  $B$  be the event "at most 2 heads." Find  $P(B)$ .
  - d. Let  $C$  be the event "at least 2 heads." Find  $P(C)$ .
  - e. Are the events  $A$  and  $B$  mutually exclusive? Find  $P(A \text{ or } B)$ .
  - f. Are the events  $A$  and  $C$  mutually exclusive? Find  $P(A \text{ or } C)$ .
3. A company places 3 digit serial numbers on each part that is made. Any number between 0 and 9 may be used in the digits. How many different serial numbers are possible if
  - a. Digits can be repeated?
  - b. Digits cannot be repeated?
4. Roulette is played at a table similar to the one in Figure 3.22. A wheel with the numbers 1 through 36 (evenly distributed with the colors red and black) and two green numbers 0 and 00 rotates in a shallow bowl with a curved wall. A small ball is spun on the inside of the wall and drops into a pocket corresponding to one of the numbers. Players may make 11 different types of bets by placing chips on different areas of the table. These include bets on a single number, two adjacent numbers, a row of three numbers, a block of four numbers, two adjacent rows of six numbers, and the five number combinations of 0, 00, 1, 2, and 3; bets on the numbers 1–18 or 19–36; the first, second, or third group of 12 numbers; a column of 12 numbers; even or odd; and red or black. Payoffs differ by bet. For instance, a single-number bet pays 35 to 1 if it wins; a three-number bet pays 11 to 1; a column bet pays 2 to 1; and a color bet pays even money. Define the following events:  $C1 = \text{column 1 number}$ ,  $C2 = \text{column 2 number}$ ,  $C3 = \text{column 3 number}$ ,  $O = \text{odd number}$ ,  $E = \text{even number}$ ,  $G = \text{green number}$ ,

**FIGURE 3.22** Layout of a Typical Roulette Table

$F_{12}$  = first 12 numbers,  $S_{12}$  = second 12 numbers, and  $T_{12}$  = third 12 numbers.

- Find the probability of each of these events.
  - Find  $P(G \text{ or } O)$ ,  $P(O \text{ or } F_{12})$ ,  $P(C_1 \text{ or } C_3)$ ,  $P(E \text{ and } F_{12})$ ,  $P(E \text{ or } F_{12})$ ,  $P(S_{12} \text{ and } T_{12})$ , and  $P(O \text{ or } C_2)$ .
5. A regular six faced die is rolled 100 times, leading to the following frequencies.

Number	Frequency
1	20
2	16
3	18
4	10
5	14
6	22

- What is the probability of getting the number 2?
  - What is the probability of getting a number greater than 4?
6. A survey of 150 residents of a certain city found that 120 residents subscribed to the morning newspaper, 75 subscribed to the evening newspaper, and 50 subscribed to both. A resident is chosen at random from the group of these 150 residents.
- What is the probability that the selected resident subscribes to the morning newspaper?
  - What is the probability that the resident subscribes neither to the morning nor to the evening newspaper?
  - What is the probability that the resident subscribes to either of the newspapers?
7. Row 27 of the Excel file *Census Education Data* gives the number of unemployed persons having a specific educational level.
- Find the probability that an employed person has attained each of the educational levels listed in the data.
  - Suppose that  $A$  is the event "has at least an Associate's Degree" and  $B$  is the event "is at least a high school
- graduate". Find the probabilities of these events. Are they mutually exclusive? Why or why not? Find the probability  $P(A \text{ or } B)$ .
- Use the Civilian Labor Force data in the Excel file *Census Education Data* to find the following:
    - $P(\text{Unemployed and Advanced Degree})$
    - $P(\text{Unemployed} \mid \text{Advanced Degree})$
    - $P(\text{Not a High School Grad} \mid \text{Unemployed})$
    - Are the events "Unemployed" and "at least a High School Graduate" independent?
  - A research laboratory proposes a medical test for a certain disease. It is known that 5% of the individuals of the population have this disease. Suppose the probability that the test indicates that an individual is healthy when he/she actually has the disease is 0.1, while the probability that the test indicates a diseased status when the individual is healthy is 0.02. An individual is randomly chosen from the population and administered the test. The result turns out to be positive. What is the probability that the individual is diseased?
  - A survey based on the general adult population of an American city shows that among the surveyed individuals, 60% are cigarette smokers and 3% of the smokers have lung cancer. On the other hand, only 1% of the non-smokers have lung cancer. Construct the joint probability distribution table of the variables Smoking Status and Lung Cancer Status.
  - Construct the probability distribution for the value of a 2-card hand dealt from a standard deck of 52 cards (all face cards have a value of 10 and an ace has a value of 11).
    - What is the probability of being dealt 21?
    - What is the probability of being dealt 16?
    - Construct a chart for the cumulative distribution function. What is the probability of being dealt a 16 or less? Between 12 and 16? Between 17 and 20?
    - Find the expected value and standard deviation of a two-card hand.
  - Based on the data in the Excel file *Facebook Survey*, develop a probability mass function and cumulative distribution function (both tabular and as charts) for the random

variable Hours Online/Week. What is the probability that an individual in this survey is online on Facebook for two hours per week? At most four hours? Ten or more?

13. Using the data in the Excel file *Room Inspection*, construct the probability mass function and the cumulative distribution function (both tabular and as charts) for the random variable Number of Nonconforming Rooms. What is the probability that the number of nonconforming rooms is greater than five? Greater than three but less than or equal to five? Less than or equal to 3?

14. A bakery has the following probability distribution for the daily demand for its cakes:

Demand, $x$ (in hundreds)	Probability, $f(x)$
0	0.04
1	0.16
2	0.18
3	0.28
4	0.34
5 or more	0.00

Find the expectation, variance and the standard deviation of the daily demand for cakes.

15. A major application of data mining in marketing is determining the attrition of customers. Suppose that the probability of a long-distance carrier's customer leaving for another carrier from one month to the next is 0.15. What distribution models the retention of an individual customer? What is the expected value and standard deviation?

16. What type of distribution models the random variable Prior Call Center Experience in the file *Call Center Data*? Define the parameter(s) for this distribution based on the data.

17. A multiple choice test has 30 problems, and 4 choices for each problem. A student taking the test does not know the correct answer for any of the problems and guesses the answer for each. What is the probability that the student will answer at least 10 questions correctly?

18. Pairs of chips, one each from brands *A* and *B*, are tested for their longevity. In 100 such experiments, brand *A* came out on top 65 times. What is the probability that this performance or better can happen just by chance?

19. A popular resort hotel has 300 rooms and is usually fully booked. About 4% of the time a reservation is cancelled before the 6:00 p.m. deadline with no penalty. What is the probability that at least 280 rooms will be occupied? Use the binomial distribution to find the exact value and the normal approximation to the binomial and compare your answers.

20. In a shooting competition, a participant hits the target at any shot with a probability of 0.5. What is the minimum number of shots needed to make the probability of at least 8 hits equal to or greater than 0.9?

21. A barber must serve 8 customers everyday to cover the infrastructural costs. His number of customers per day has a mean of 10. What is the probability that he will serve less than 8 customers a day?

22. The following table provides the frequencies of the number of alpha particles emitted by a certain radioactive system in 100 intervals of length 1/100 seconds.

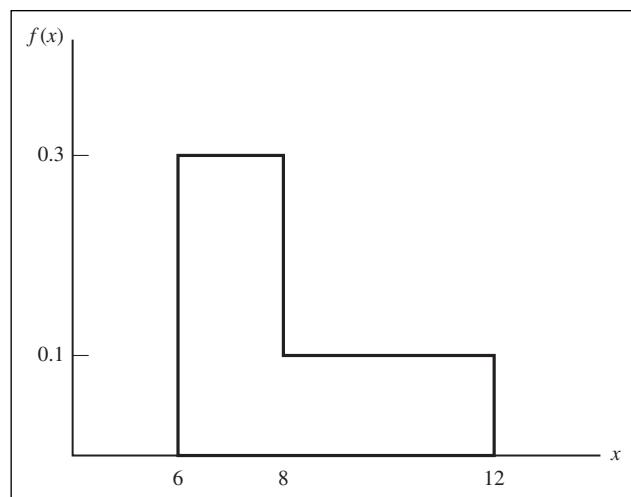
Number	Frequency
0	0
1	1
2	4
3	13
4	19
5	16
6	15
7	9
8	12
9	8
10	3
Total	100

- a. Find the probability of emission of 0 to 10 alpha particles using these data.

- b. Assuming a Poisson distribution and using the mean number of alpha particles per time interval of 1/100 seconds from the empirical data, compute the probabilities of observing 0 through 10 alpha particles in a given interval. Compare these to your answer to part (a). Does a Poisson distribution model this phenomenon satisfactorily?

23. Verify that the function corresponding to the figure below is a valid probability density function. Then find the following probabilities:

- a.  $P(x < 8)$
- b.  $P(x > 7)$
- c.  $P(6 < x < 10)$
- d.  $P(8 < x < 11)$



24. The weight of a box of pasta in a production process is uniformly distributed between 280 and 320 grams.
- Find the expected value and the variance of the weight of the pasta boxes.
  - What is the probability that a randomly chosen box of pasta will weigh less than 290 grams?
  - What is the probability that a randomly chosen box of pasta will weigh more than 305 grams?
25. The minimum volume of a regular can of soda is 10 ounces. It is also known that 60% of the volume of similar cans is less than 13 ounces. If the volume of the cans is uniformly distributed, what will be the parameters of the distribution?
26. The birth weight  $X$  (in grams) of babies in the United States is normally distributed with a mean of 3315 and standard deviation of 575. Find the following:
- $P(X < 3890)$
  - $P(2740 < X)$
  - $P(2165 < X < 4465)$
  - The birth weight that the upper 5% of babies achieve.
27. The weight of high school boys in a certain city is normally distributed, with mean  $\mu = 50$  kg and standard deviation  $\sigma = 4$  kg. Find the probability that an individual student's weight is
- Less than 45 kg.
  - Between 48 and 51 kg.
  - Greater than 63 kg.
28. The specifications of a particular component in a manufacturing process requires that the length of the component be 25 mm. Due to the variation in the manufacturing process the components actually have a mean of 25 mm, a standard deviation of 2 mm and are normally distributed.
- If the length of the component is more than 30 mm, it will be discarded by the quality control check. What proportion of components discarded by the quality control check?
  - What is the probability that the length of component will be smaller than 22 mm?
29. The weight of a particular chip in the manufacturing process of a certain notepad computer must be between 40 and 50 grams. It is known that the standard deviation of the weight of the chip, which is normally distributed, is 3 grams.
- If the actual mean weight is 45 grams, what proportion of chips will meet the specifications?
  - If the mean weight is recalibrated to 48 grams what proportion of the chips will meet the specifications?
  - If the mean weight is 45 grams, how small should the standard deviation be so that 99% of the chips meet the specifications?
30. It is known that people who buy lottery tickets regularly spend approximately \$30 a month on them, with a standard deviation of \$5. Find the probability that such a person will spend at least \$20 per month on lottery tickets. If 100 individuals who buy lottery tickets regularly are sampled, find the mean and standard deviation of those who spend at least \$20 per month on lottery tickets.
31. Compute the expected values and the variance for each of the parameter sets of the triangular distribution.
- $a = 0, b = 4, c = 2$
  - $a = 0, b = 4, c = 3$
  - $a = 0, b = 4, c = 1$
32. The amount of time (in seconds) spent by bees in a certain flower patch is exponential with a mean of 90. What is the probability that a particular bee will take more than 100 seconds to collect honey?
33. The length of time between arrivals of cars in a highway tollbooth is exponentially distributed with a mean of 45 seconds.
- What is the probability that an arrival time will be greater than a minute?
  - What proportion of arrivals will be in less than 30 seconds?
34. Using the data on excel file *Census Education data*, construct the probability distribution for race and educational status.
- What is the probability that a person is white and has a bachelor's degree?
  - What is the probability that someone known to have an associate's degree is black?
  - What is the probability that an individual of the Other race has at least an associate's degree?
  - If it is known that an individual is female, what is the probability that she drives an SUV?
  - Determine whether the random variables "gender" and the event "vehicle driven" are statistically independent. What would this mean for advertisers?
35. Using the data on excel file *Census Education data*, construct the probability distribution for race and educational status.
- What is the probability that a person is white and has a bachelor's degree?
  - What is the probability that someone known to have an associate's degree is black?
  - What is the probability that an individual of the Other race has at least an associate's degree?
36. Compute the joint probability distribution and marginal probabilities for age and educational status in the Excel file *Census Education Data*, and verify that these random variables are not statistically independent. Then, using the marginal probabilities, determine the joint probabilities that would make them statistically independent. Convert the joint probabilities to numbers of people and compare them to the data. What can you conclude?

## Case

### Probability Analysis for Quality Measurements

A manufacturer of home and industrial lawn and garden equipment collects a variety of data from special studies, many of which are related to quality control. The company routinely collects data about functional test performance of its mowers after assembly; results from the past 30 days are given in the worksheet *Mower Test* in the Excel file *Quality Measurements*. In addition, many in-process measurements are taken to ensure that manufacturing processes remain in control and can produce according to design specifications. The worksheet *Process Capability* provides the results of 200 samples of blade weights taken from the manufacturing process that produces mower blades. You have been asked you to evaluate these data. Specifically,

1. What fraction of mowers fails for each of the 30 samples in the worksheet *Mower Test*? What distribution might be appropriate to model the failure of an individual mower? Using these data, estimate the sampling distribution of the mean, the overall fraction of failures, and the standard error of the mean. Is a normal distribution an appropriate assumption for the sampling distribution of the mean?

2. What fraction of mowers fails the functional performance test using all the data in the worksheet *Mower Test*? Using this result, what is the probability of having  $x$  failures in the next 100 mowers tested, for  $x$  from 0 to 20?
3. Do the data in the worksheet *Process Capability* appear to be normally distributed? (Construct a frequency distribution and histogram and use these to draw a conclusion.) If not, based on the histogram, what distribution might better represent the data?
4. Estimate the mean and standard deviation for the data in the worksheet *Process Capability*. Using these values, and assuming that the process capability data are normal, find the probability that blade weights from this process will exceed 5.20. What is the probability that weights will be less than 4.80? What is the actual percentage of weights that exceed 5.20 or are less than 4.80 from the data in the worksheet? How do the normal probability calculations compare? What do you conclude?

Summarize all your findings to these questions in a well-written report.

## APPENDIX 3.1

### Probability Distributions: Theory and Computation

In this appendix, we expand upon the basic concepts developed earlier in this chapter and provide details regarding the formulas used to perform many of the calculations involving probability distributions.

#### A. Expected Value and Variance of a Random Variable

As noted earlier in the chapter, the expected value of a discrete random variable,  $E[X]$ , is the sum of the product of the values of the random variable and their probabilities.  $E[X]$  is computed using the formula:

$$E[X] = \sum_{i=1}^{\infty} x_i f(x_i) \quad (3A.1)$$

Note the similarity to computing the population mean using Formula (2A.9) in Chapter 2:

$$\mu = \frac{\sum_{i=1}^N f_i x_i}{N}$$

If we write this as the sum of  $x_i$  times ( $f_i/N$ ), then we can think of  $f_i/N$  as the probability of  $x_i$ . Then this expression for the mean has the same basic form as the expected value formula. Similarly, the variance of a random variable is computed as:

$$\text{Var}[X] = \sum_{i=1}^{\infty} (x_i - E[X])^2 f(x_i) \quad (3A.2)$$

Note the similarity between this and Formula (2A.10) in Chapter 2.

#### B. Binomial Distribution

The probability mass function for the binomial distribution is:

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & \text{for } x = 0, 1, 2, \dots, n \\ 0, & \text{otherwise} \end{cases} \quad (3A.3)$$

The notation  $\binom{n}{x}$  represents the number of ways of choosing  $x$  distinct items from a group of  $n$  items and is computed as:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad (3A.4)$$

where  $n!$  ( $n$  factorial) =  $n(n-1)(n-2)\dots(2)(1)$ , and  $0!$  is defined to be 1.

For example, if the probability that any individual will react positively to a new drug is 0.8, then the probability distribution that  $x$  individuals will react positively out of a sample of 10 is:

$$f(x) = \begin{cases} \binom{10}{x} (0.8)^x (0.2)^{10-x}, & \text{for } x = 0, 1, 2, \dots, 10 \\ 0, & \text{otherwise} \end{cases}$$

If  $x = 4$ , for example, we have:

$$\begin{aligned} f(4) &= \binom{10}{4} (0.8)^4 (0.2)^{10-4} = \frac{10!}{4!6!} (0.4096) (0.000064) \\ &= 0.005505 \end{aligned}$$

## C. Poisson Distribution

The probability mass function for the Poisson distribution is:

$$f(x) = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!}, & \text{for } x = 0, 1, 2, \dots \\ 0, & \text{otherwise} \end{cases} \quad (3A.5)$$

where the mean number of occurrences in the defined unit of measure is  $\lambda$ .

Suppose that on average, 12 customers arrive at an ATM during lunch hour. The probability that exactly  $x$  customers will arrive during the hour would be calculated using the formula:

$$f(x) = \begin{cases} \frac{e^{-12} 12^x}{x!}, & \text{for } x = 0, 1, 2, \dots \\ 0, & \text{otherwise} \end{cases} \quad (3A.6)$$

For example, the probability that exactly 5 customers will arrive is  $e^{-12}(12^5)/5!$  or 0.12741.

## D. Uniform Distribution

For a uniform distribution with a minimum value  $a$  and a maximum value  $b$ , the density function is:

$$f(x) = \frac{1}{b-a} \quad \text{if } a \leq x \leq b \quad (3A.7)$$

and the cumulative distribution function is:

$$F(x) = \begin{cases} 0, & \text{if } x < a \\ \frac{x-a}{b-a}, & \text{if } a \leq x \leq b \\ 1, & \text{if } b < x \end{cases} \quad (3A.8)$$

For example, suppose that sales revenue for a product varies uniformly each week between  $a = \$1,000$  and  $b = \$2,000$ . Then the probability that sales revenue will be less than  $x = \$1,300$  is:

$$F(1,300) = \frac{1,300 - 1,000}{2,000 - 1,000} = 0.3$$

Similarly, the probability that revenue will be between \$1,500 and \$1,700 is

$$F(1,700) - F(1,500) = 0.7 - 0.5 = 0.2.$$

## E. Normal Distribution

The probability density function for the normal distribution is quite complex:

$$f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \quad (3A.9)$$

In addition, the cumulative distribution function cannot be expressed mathematically, only numerically. A numerical tabulation of the standard normal distribution is provided in the appendix at the end of this book. Table A.1 presents a table of the cumulative probabilities for values of  $z$  from  $-3.9$  to  $3.99$ . To illustrate the use of this table, let us find the area under the normal distribution within one standard deviation of the mean. Note that  $P(-1 < Z < 1) = F(1) - F(-1)$ . Using Table A.1, we find  $F(1) = 0.8413$  and  $F(-1) = 0.1587$ . Therefore,  $P(-1 < Z < 1) = 0.8413 - 0.1587 = 0.6826$ .

To simplify the calculation of normal probabilities for any random variable  $X$  with an arbitrary mean  $\mu$  and standard deviation  $\sigma$ , we may transform any value of the random variable  $X$  to a random variable  $Z$  having a standard normal distribution by applying the following formula:

$$z = \frac{x - \mu}{\sigma} \quad (3A.10)$$

These are called **standardized normal values**, or sometimes simply  **$z$ -values**. Standardized  $z$ -values are expressed in units of standard deviations of  $X$ . Thus, a  $z$ -value of 1.5 means that the associated  $x$ -value is 1.5 standard deviations above the mean  $\mu$ ; similarly, a  $z$ -value of  $-2.0$  corresponds to an  $x$ -value that is two standard deviations below the mean  $\mu$ .

Standardized  $z$ -values are particularly useful in solving problems involving arbitrary normal distributions and allow us to find probabilities using the standard normal table in Table A.1. For example, suppose that we determine

that the distribution of customer demand ( $X$ ) is normal with a mean of 750 units/month and a standard deviation of 100 units/month. To find this probability using Table A.1, we transform this into a standard normal distribution by finding the  $z$ -value that corresponds to  $x = 900$ :

$$z = \frac{900 - 750}{100} = 1.5$$

This means that  $x = 900$  is 1.5 standard deviations above the mean of 750. From Table A.1,  $F(1.5)$  is 0.9332. Therefore, the probability that  $Z$  exceeds 1.5 (equivalently, the probability that  $X$  exceeds 900) is  $1 - 0.9332 = 0.0668$ . To summarize,

$$\begin{aligned} P(X > 900) &= P(Z > 1.5) = 1 - P(Z < 1.5) \\ &= 1 - 0.9332 \\ &= 0.0668 \end{aligned}$$

Another common calculation involving the normal distribution is to find the value of  $x$  corresponding to a specified probability, or area. For this example, suppose that we wish to find the level of demand that will be exceeded only 10% of the time; that is, find the value of  $x$  so that  $P(X > x) = 0.10$ . An upper-tail probability of 0.10 is equivalent to a cumulative probability of 0.90. Using Table A.1, we search for 0.90 as close

as possible in the *body* of the table and obtain  $z = 1.28$ , corresponding to 0.8997. This means that a value of  $z$  that is 1.28 standard deviations above the mean has an upper-tail area of approximately 0.10. Using the standard normal transformation,

$$z = \frac{x - 750}{100} = 1.28$$

and solving for  $x$ , we find  $x = 878$ .

## F. Exponential Distribution

The exponential distribution has the density function:

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0 \quad (3A.11)$$

and its cumulative distribution function is:

$$F(x) = 1 - e^{-\lambda x}, \quad x \geq 0 \quad (3A.12)$$

Suppose that the mean time to failure of a critical component of an engine is  $1/\lambda = 8,000$  hours. The probability that the component will fail before  $x$  hours is given by the cumulative distribution function. Thus, the probability of failing before 5,000 hours is:

$$F(5000) = 1 - e^{-(1/8000)(5000)} = 1 - e^{-5/8} = 0.465$$

## APPENDIX 3.2

### Excel and PHStat Notes

#### A. Normal Probability Tools

*PHStat* has a useful tool for computing probabilities for normal distributions. From the *Probability & Prob. Distributions* menu, select *Normal*. The *Normal Probability Distribution* dialog, shown in Figure 3A.1, allows you to compute probabilities for any interval and also find the value of  $X$  for a given

cumulative percentage similar to what the *NORM.INV* function did. The dialog box is filled out to answer the same questions as we did in the normal probability example. The results are shown in Figure 3A.2. The tool also calculates  $z$ -values for the standard normal distribution.

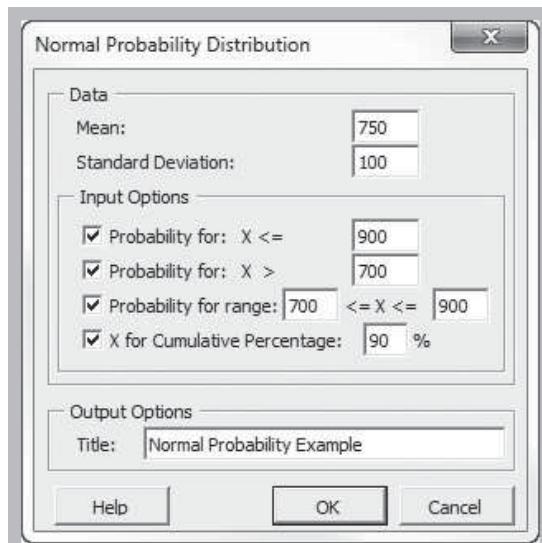


FIGURE 3A.1 Normal Probability Distribution Dialog

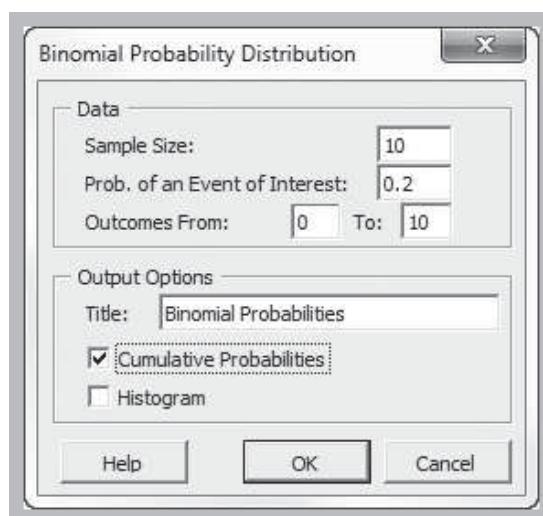
A	B	C	D	E
1	Normal Probability Example			
2				
3	Common Data			
4	Mean	750		
5	Standard Deviation	100		
6				
7	Probability for $X \leq$		Probability for a Range	
8	X Value	900	From X Value	700
9	Z Value	1.5	To X Value	900
10	$P(X \leq 900)$	0.9331928	Z Value for 700	-0.5
11			Z Value for 900	1.5
12	Probability for $X >$		$P(X \leq 700)$	0.3085
13	X Value	700	$P(X \leq 900)$	0.9332
14	Z Value	-0.5	$P(700 < X \leq 900)$	0.6247
15	$P(X > 700)$	0.6915		
16			Find X and Z Given Cum. Pctage.	
17	Probability for $X < 900$ or $X > 700$		Cumulative Percentage	90.00%
18	$P(X < 900 \text{ or } X > 700)$	1.6247	Z Value	1.281552
			X Value	878.1552

**FIGURE 3A.2** PHStat Normal Probability Example Results

## B. Generating Probabilities in PHStat

From the PHStat menu, select *Probability & Prob. Distributions* and then *Binomial*. The dialog box in Figure 3A.3 prompts you for the distribution's parameters and range of outputs desired. By checking the boxes for *Cumulative Probabilities*,

the routine provides additional information as shown in Figure 3A.4, specifically columns D through G. Checking the *Histogram* box provides a graphical chart of the distribution over the range of outcomes specified.



**FIGURE 3A.3** Binomial Probability Distribution Dialog in PHStat

	A	B	C	D	E	F	G
1	<b>Binomial Probabilities</b>						
2							
3	<b>Data</b>						
4	<b>Sample size</b>	10					
5	<b>Probability of an event of interest</b>	0.2					
6							
7	<b>Statistics</b>						
8	<b>Mean</b>	2					
9	<b>Variance</b>	1.6					
10	<b>Standard deviation</b>	1.264911					
11							
12	<b>Binomial Probabilities Table</b>						
13		X	P(X)	P( $\leq$ X)	P( $<$ X)	P( $>$ X)	P( $\geq$ X)
14		0	0.107374	0.107374	0	0.892626	1
15		1	0.268435	0.37581	0.107374	0.62419	0.892626
16		2	0.30199	0.6778	0.37581	0.3222	0.62419
17		3	0.201327	0.879126	0.6778	0.120874	0.3222
18		4	0.08808	0.967207	0.879126	0.032793	0.120874
19		5	0.026424	0.993631	0.967207	0.006369	0.032793
20		6	0.005505	0.999136	0.993631	0.000864	0.006369
21		7	0.000786	0.999922	0.999136	7.79E-05	0.000864
22		8	7.37E-05	0.999996	0.999922	4.2E-06	7.79E-05
23		9	4.1E-06	1	0.999996	1.02E-07	4.2E-06
24		10	1.02E-07	1	1	0	1.02E-07

**FIGURE 3A.4** Output from *PHStat* Binomial Probability Distribution Option

## *Chapter 4*

# Sampling and Estimation

- INTRODUCTION 124
- STATISTICAL SAMPLING 124
  - Sample Design 124
  - Sampling Methods 125
  - Errors in Sampling 127
- RANDOM SAMPLING FROM PROBABILITY DISTRIBUTIONS 127
  - Sampling from Discrete Probability Distributions 128
  - Sampling from Common Probability Distributions 129
  - A Statistical Sampling Experiment in Finance 130
- SAMPLING DISTRIBUTIONS AND SAMPLING ERROR 131
  - Applying the Sampling Distribution of the Mean 134
- SAMPLING AND ESTIMATION 134
  - Point Estimates 135
  - Unbiased Estimators 136
  - Interval Estimates 137
- CONFIDENCE INTERVALS: CONCEPTS AND APPLICATIONS 137
  - Confidence Interval for the Mean with Known Population Standard Deviation 138
  - Confidence Interval for the Mean with Unknown Population Standard Deviation 140
  - Confidence Interval for a Proportion 142
  - Confidence Intervals for the Variance and Standard Deviation 143
  - Confidence Interval for a Population Total 145
- USING CONFIDENCE INTERVALS FOR DECISION MAKING 146
- CONFIDENCE INTERVALS AND SAMPLE SIZE 146
- PREDICTION INTERVALS 148
- ADDITIONAL TYPES OF CONFIDENCE INTERVALS 149
  - Differences Between Means, Independent Samples 149
  - Differences Between Means, Paired Samples 149
  - Differences Between Proportions 150
- BASIC CONCEPTS REVIEW QUESTIONS 150
- PROBLEMS AND APPLICATIONS 150
- CASE: ANALYZING A CUSTOMER SURVEY 153

■ APPENDIX 4.1: THEORETICAL FOUNDATIONS OF CONFIDENCE INTERVALS 153

- A. Theory Underlying Confidence Intervals 153
  - B. Sampling Distribution of the Proportion 154
  - C. Sample Size Determination 155
  - D. Additional Confidence Intervals 155
- APPENDIX 4.2: EXCEL AND PHSTAT NOTES 158
- A. Excel-Based Random Sampling Tools 158
  - B. Using the VLOOKUP Function 159
  - C. Sampling from Probability Distributions 159
  - D. Confidence Intervals for the Mean 160
  - E. Confidence Intervals for Proportions 160
  - F. Confidence Intervals for the Population Variance 161
  - G. Determining Sample Size 161

## INTRODUCTION

In Chapters 1–3, we discussed the use of data for managing and decision making, introduced methods for data visualization and descriptive statistics, and gained an understanding of some important probability distributions used in statistics and decision modeling. These topics were focused on how to view data to gain better understanding and insight. However, managers need to go further than simply understanding what data tell; they need to be able to *draw conclusions* about population characteristics from the data to make effective decisions. Sampling methods and estimation, as well as hypothesis testing and statistical inference, which we focus on in the next chapter, provide the means for doing this.

## STATISTICAL SAMPLING

Sampling approaches play an important role in providing information for making business decisions. Market researchers, for example, need to sample from a large base of customers or potential customers; auditors must sample among large numbers of transactions; and quality control analysts need to sample production output to verify quality levels. Most populations, even if they are finite, are generally too large to deal with effectively or practically. For instance, it would be impractical as well as too expensive to survey the entire population of TV viewers in the United States. Sampling is also clearly necessary when data must be obtained from destructive testing or from a continuous production process. Thus, the purpose of sampling is to obtain sufficient information to draw a valid inference about a population.

### Sample Design

The first step in sampling is to design an effective sampling plan that will yield representative samples of the populations under study. A **sampling plan** is a description of the approach that will be used to obtain samples from a population prior to any data collection activity. A sampling plan states the objectives of the sampling activity, the target population, the population **frame** (the list from which the sample is selected), the method of sampling, the operational procedures for collecting the data, and the statistical tools that will be used to analyze the data.

The objective of a sampling study might be to estimate key parameters of a population, such as a mean, proportion, or standard deviation. For example, *USA Today* reported on May 19, 2000, that the U.S. Census Bureau began a statistical sampling procedure to

estimate the number and characteristics of people who might have been missed in the traditional head count. Another application of sampling is to determine if significant differences exist between two populations. For instance, the Excel file *Burglaries* provides data about monthly burglaries in an urban neighborhood before and after a citizen-police program was instituted. You might wish to determine whether the program was successful in reducing the number of burglaries.

The ideal frame is a complete list of all members of the target population. However, for practical reasons, a frame may not be the same as the target population. For example, a company's target population might be all golfers in America, which might be impossible to identify, whereas a practical frame might be a list of golfers who have registered handicaps with the U.S. Golf Association. Understanding how well the frame represents the target population helps us to understand how representative of the target population the actual sample is and, hence, the validity of any statistical conclusions drawn from the sample. In a classic example, *Literary Digest* polled individuals from telephone lists and membership rolls of country clubs for the 1936 presidential election and predicted that Alf Landon would defeat Franklin D. Roosevelt. The problem was that the frame—individuals who owned telephones and belonged to country clubs—was heavily biased toward Republicans and did not represent the population at large.

## Sampling Methods

Sampling methods can be *subjective* or *probabilistic*. Subjective methods include **judgment sampling**, in which expert judgment is used to select the sample (survey the "best" customers), and **convenience sampling**, in which samples are selected based on the ease with which the data can be collected (survey all customers I happen to visit this month). Probabilistic sampling involves selecting the items in the sample using some random procedure. Probabilistic sampling is necessary to draw valid statistical conclusions.

The most common probabilistic sampling approach is simple random sampling. **Simple random sampling** involves selecting items from a population so that every subset of a given size has an equal chance of being selected. If the population data are stored in a database, simple random samples can generally be obtained easily. For example, consider the Excel file *Cereal Data*, which contains nutritional and marketing information for 67 cereals, a portion of which is shown in Figure 4.1. Suppose that we wish to sample 10 of these cereals. *PHStat* and Excel provide tools to generate a random set of values from a given population size (see Appendix 4.2A, *Excel-Based Random Sampling Tools*). An example of a random sample of 10 cereals from this list, obtained by the *PHStat Random Sample Generator* tool, is shown in Figure 4.2.



Spreadsheet Note

	A	B	C	D	E	F	G	H
1	Cereal Data							
2								
3	Product	Cereal Name	Manufacturer	Calories	Sodium	Fiber	Carbs	Sugars
4	1	100% Bran	Nabisco	70	130	10	5	6
5	2	All-Bran	Kellogg	70	260	9	7	5
6	3	All-Bran w/Extra Fiber	Kellogg	50	140	14	8	0
7	4	Almond Delight	Ralston Purina	110	200	1	14	8
8	5	Apple Cinn Cheerios	General Mills	110	180	1.50	10.50	10
9	6	Apple Jacks	Kellogg	110	125	1	11	14
10	7	Basic 4	General Mills	130	210	2	18	8
11	8	Bran Chex	Ralston Purina	90	200	4	15	6
12	9	Bran Flakes	Post	90	210	5	13	5
13	10	Cap'n'Crunch	Quaker	120	220	0	12	12

FIGURE 4.1 Portion of the Excel File *Cereal Data*

A
1 Random Cereal Samples
2 Lucky Charms
3 Apple Jacks
4 Apple Cinn Cheerios
5 Crispy Wheat & Raisins
6 Honey Graham Ohs
7 Puffed Wheat
8 Honey Nut Cheerios
9 Trix
10 Post Nat. Raisin Bran
11 Almond Delight

**FIGURE 4.2** Random Sample  
of Ten Cereals

Other methods of sampling include the following:

- **Systematic (Periodic) Sampling.** This is a sampling plan that selects items periodically from the population. For example, to sample 250 names from a list of 400,000, the first name could be selected at random from the first 1,600, and then every 1,600th name could be selected. This approach can be used for telephone sampling when supported by an automatic dialer that is programmed to dial numbers in a systematic manner. However, systematic sampling is not the same as simple random sampling because for any sample, every possible sample of a given size in the population does not have an equal chance of being selected. In some situations, this approach can induce significant bias if the population has some underlying pattern. For instance, sampling orders received every seven days may not yield a representative sample if customers tend to send orders on certain days every week.
- **Stratified Sampling.** This type of sampling applies to populations that are divided into natural subsets (strata) and allocates the appropriate proportion of samples to each stratum. For example, a large city may be divided into political districts called wards. Each ward has a different number of citizens. A stratified sample would choose a sample of individuals in each ward proportionate to its size. This approach ensures that each stratum is weighted by its size relative to the population and can provide better results than simple random sampling if the items in each stratum are not homogeneous. However, issues of cost or significance of certain strata might make a disproportionate sample more useful. For example, the ethnic or racial mix of each ward might be significantly different, making it difficult for a stratified sample to obtain the desired information.
- **Cluster Sampling.** This is based on dividing a population into subgroups (clusters), sampling a set of clusters, and (usually) conducting a complete census within the clusters sampled. For instance, a company might segment its customers into small geographical regions. A cluster sample would consist of a random sample of the geographical regions, and all customers within these regions would be surveyed (which might be easier because regional lists might be easier to produce and mail).
- **Sampling from a Continuous Process.** Selecting a sample from a continuous manufacturing process can be accomplished in two main ways. First, select a time at random; then select the next  $n$  items produced after that time. Second, select  $n$  times at random; then select the next item produced after these times. The first approach generally ensures that the observations will come from a homogeneous population; however, the second approach might include items from different populations if the characteristics of the process should change over time, so caution should be used.

## Errors in Sampling

The purpose of sampling is to obtain statistics that estimate population parameters. Sample design can lead to two sources of errors. The first type of error, **nonsampling error**, occurs when the sample does not represent the target population adequately. This is generally a result of poor sample design, such as using a convenience sample when a simple random sample would have been more appropriate or choosing the wrong sampling frame. The other type of error, **sampling (statistical) error**, occurs because samples are only a subset of the total population. Sampling error is inherent in any sampling process, and although it can be minimized, it cannot be totally avoided.

Sampling error depends on the size of the sample relative to the population. Thus, determining the number of samples to take is essentially a statistical issue that is based on the accuracy of the estimates needed to draw a useful conclusion. We discuss this later in this chapter. However, from a practical standpoint, one must also consider the cost of sampling and sometimes make a trade-off between cost and the information that is obtained.

## RANDOM SAMPLING FROM PROBABILITY DISTRIBUTIONS

Many applications in both statistics and decision modeling require random samples from specific probability distributions. For example, some statistical problems cannot be solved (or may be very difficult to solve) using mathematics. An alternative is to find an answer using sampling experiments. Many decision models contain outcome variables that are complicated functions of several input random variables. For example, in a financial model, we might be interested in the distribution of the cumulative discounted cash flow over several years when sales, sales growth rate, operating expenses, and inflation factors are all uncertain and are described by probability distributions. Understanding the probability distribution of cash flow can only be accomplished by sampling procedures (which we will address in Part II of this book). Finally, sampling experiments can be used to develop and better understand important theories in statistics.

The basis for generating random samples from probability distributions is the concept of a random number. A **random number** is one that is uniformly distributed between 0 and 1. Technically speaking, computers cannot generate truly random numbers since they must use a predictable algorithm. However, the algorithms are designed to generate a sequence of numbers that appear to be random. In Excel, we may generate a random number within any cell using the function RAND(). This function has no arguments; therefore, nothing should be placed within the parentheses (but the parentheses are required). Table 4.1 shows a table of 100 random numbers generated in Excel. You

**TABLE 4.1** One Hundred Random Numbers

0.007120	0.215576	0.386009	0.201736	0.457990	0.127602	0.387275	0.639298	0.757161	0.285388
0.714281	0.165519	0.768911	0.687736	0.466579	0.481117	0.260391	0.508433	0.528617	0.755016
0.226987	0.454259	0.487024	0.269659	0.531411	0.197874	0.527788	0.613126	0.716988	0.747900
0.339398	0.434496	0.398474	0.622505	0.829964	0.288727	0.801157	0.373983	0.095900	0.041084
0.692488	0.137445	0.054401	0.483937	0.954835	0.643596	0.970131	0.864186	0.384474	0.134890
0.962794	0.808060	0.169243	0.347993	0.848285	0.216635	0.779147	0.216837	0.768370	0.371613
0.824428	0.919011	0.820195	0.345563	0.989111	0.269649	0.433170	0.369070	0.845632	0.158662
0.428903	0.470202	0.064646	0.100007	0.379286	0.183176	0.180715	0.008793	0.569902	0.218078
0.951334	0.258192	0.916104	0.271980	0.330697	0.989264	0.770787	0.107717	0.102653	0.366096
0.635494	0.395185	0.320618	0.003049	0.153551	0.231191	0.737850	0.633932	0.056315	0.281744

should be aware that unless the automatic recalculation feature is suppressed, whenever any cell in the spreadsheet is modified, the values in any cell containing the RAND() function will change. Automatic recalculation can be changed to manual by choosing *Calculation Options* in the *Calculation* group under the *Formulas* tab. Under manual recalculation mode, the worksheet is recalculated only when the F9 key is pressed.

## **Sampling from Discrete Probability Distributions**

Sampling from discrete probability distributions using random numbers is quite easy. We will illustrate this process using the probability distribution for rolling two dice that we developed in Chapter 3. The probability mass function and cumulative distribution are shown below.

$x$	$f(x)$	$F(x)$
2	0.028	0.028
3	0.056	0.083
4	0.083	0.167
5	0.111	0.278
6	0.139	0.417
7	0.167	0.583
8	0.139	0.722
9	0.111	0.833
10	0.083	0.917
11	0.056	0.972
12	0.028	1.000

Notice that the values of  $F(x)$  divide the interval from 0 to 1 into smaller intervals that correspond to the probabilities of the outcomes. For example, the interval between 0 and 0.028, inclusive, has a probability of 0.028 and corresponds to the outcome  $x = 2$ ; the interval between 0.028 and up to and including 0.083 has a probability of 0.056 and corresponds to the outcome  $x = 3$ ; and so on. This is summarized as follows:

Interval		Outcome	
0	to	0.028	2
0.028	to	0.083	3
0.083	to	0.167	4
0.167	to	0.278	5
0.278	to	0.417	6
0.417	to	0.583	7
0.583	to	0.722	8
0.722	to	0.833	9
0.833	to	0.917	10
0.917	to	0.972	11
0.972	to	1.000	12

Any random number, then, must fall within one of these intervals. Thus, to generate an outcome from this distribution, all we need to do is to select a random number and determine the interval into which it falls. Suppose we use the first column in Table 4.1. The first random number is 0.007120. This falls in the first interval; thus, the first sample outcome is  $x = 2$ . The second random number is 0.714281. This number

falls in the seventh interval, generating a sample outcome  $x = 8$ . Essentially, we have developed a technique to roll dice on a computer! If this is done repeatedly, the frequency of occurrence of each outcome should be proportional to the size of the random number range (i.e., the probability associated with the outcome) because random numbers are uniformly distributed. We can easily use this approach to generate outcomes from any discrete distribution; the VLOOKUP function in Excel can be used to implement this on a spreadsheet (see Appendix 4.2B, *Using the VLOOKUP Function*).



Spreadsheet Note

### SKILL-BUILDER EXERCISE 4.1

Use the *PHStat Random Sample Generator* to generate a list of 10 random integers from among 67 in the Excel file *Cereal Data*, and then use the VLOOKUP function to extract the names of the corresponding cereals.

## Sampling from Common Probability Distributions

This approach of generating random numbers and transforming them into outcomes from a probability distribution may be used to sample from most any distribution. A value randomly generated from a specified probability distribution is called a **random variate**. For example, it is quite easy to transform a random number into a random variate from a uniform distribution between  $a$  and  $b$ . Consider the formula:

$$U = a + (b - a)*\text{RAND}() \quad (4.1)$$

Note that when  $\text{RAND} = 0$ ,  $U = a$ , and when  $\text{RAND}$  approaches 1,  $U$  approaches  $b$ . For any other value of  $\text{RAND}$  between 0 and 1,  $(b - a)*\text{RAND}()$  represents the same proportion of the interval  $(a, b)$  as  $\text{RAND}$  does of the interval  $(0, 1)$ . Thus, all real numbers between  $a$  and  $b$  can occur. Since  $\text{RAND}$  is uniformly distributed, so also is  $U$ .

While this is quite easy, it is certainly not obvious how to generate random variates from other distributions such as a normal or exponential distribution. We will not describe the technical details of how this is done, but rather just describe the capabilities available in Excel.

Excel allows you to generate random variates from discrete distributions and certain others using the *Random Number Generation* option in the *Analysis Toolpak* (see Appendix 4.2C, *Sampling from Probability Distributions*). However, one disadvantage with using the *Random Number Generation* tool is that you must repeat the process to generate a new set of sample values; pressing the recalculation (F9) key will not change the values. This can make it difficult to use this tool to analyze decision models.

Excel 2010 also has several functions that may be used to generate random variates. For the normal distribution, use the following:

- $\text{NORM.INV}(probability, mean, standard\_deviation)$ —normal distribution with a specified mean and standard deviation
- $\text{NORM.S.INV}(probability)$ —standard normal distribution

And for some advanced distributions, you might see:

- $\text{LOGNORM.INV}(probability, mean, standard\_deviation)$ —lognormal distribution, where  $\ln(X)$  has the specified mean and standard deviation
- $\text{BETA.INV}(probability, alpha, beta, A, B)$ —beta distribution
- $\text{GAMMA.INV}(probability, alpha, beta)$ —gamma distribution

To use these functions, simply enter  $\text{RAND}()$  in place of *probability* in the function. For example,  $\text{NORM.INV}(\text{RAND}(), 5, 2)$  will generate random variates from a normal distribution with mean 5 and standard deviation 2. Each time the worksheet is recalculated, a new random number and, hence, a new random variate are generated. These



Spreadsheet Note



## Spreadsheet Note

functions may be embedded in cell formulas and will generate new values whenever the worksheet is recalculated.

*PHStat* also includes the ability to generate samples from a uniform (0, 1) distribution, standard normal distribution, and an arbitrary discrete distribution. These are also described in Appendix 4.2C, *Sampling from Probability Distributions*. As with the Excel *Random Number Generation* tool, this *PHStat* tool generates the samples “off line”; that is, they cannot be embedded directly into other cell formulas.

### A Statistical Sampling Experiment in Finance

In finance, one way of evaluating capital budgeting projects is to compute a profitability index (*PI*), which is defined as the ratio of the present value of future cash flows (*PV*) to the initial investment (*I*):

$$PI = PV/I \quad (4.2)$$

Because the cash flow and initial investment that may be required for a particular project are often uncertain, the profitability index is also uncertain. If we can characterize *PV* and *I* by some probability distributions, then we would like to know the probability distribution for *PI*. For example, suppose that *PV* is estimated to be normally distributed with a mean of \$12 million and a standard deviation of \$2.5 million, and the initial investment is also estimated to be normal with a mean of \$3 million and standard deviation of \$0.8 million. Intuitively, one might believe that the *PI* is also normally distributed with a mean of \$12 million/\$3 million = \$4 million; however, as we shall see, this is not the case. We can use a sampling experiment to identify the probability distribution of *PI* for these assumptions.

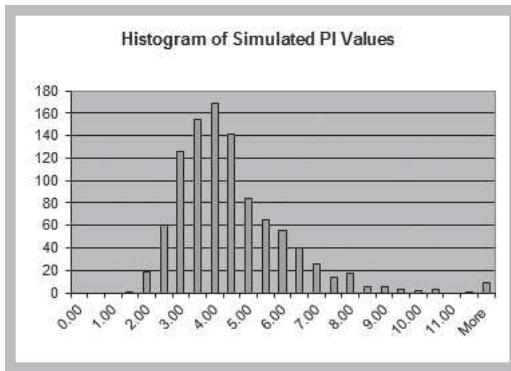
Figure 4.3 shows a simple model. For each experiment, the values of *PV* and *I* are sampled from their assumed normal distributions using the NORMINV function. *PI* is calculated in column D, and the average value for 1,000 experiments is shown in cell E8. We clearly see that this is not equal to 4 as previously suspected. The histogram in Figure 4.4 also demonstrates that the distribution of *PI* is not normal but is skewed to the right. This experiment confirms that the ratio of two normal distributions is not normally distributed.

#### SKILL-BUILDER EXERCISE 4.2

Replicate the statistical sampling experiment to find the mean and distribution of the profitability index as described in this chapter (Figure 4.3). How closely do your results coincide with the results discussed and Figure 4.4?

	A	B	C	D	E
1	Profitability Index Analysis				
2					
3		Mean	Standard Deviation		
4	PV	12	2.5		
5	I	3	0.8		
6					
7	Experiment	PV	I	PI	Mean
8	1	8.396743042	3.573822001	2.349513501	4.762285
9	2	11.7446542	3.66554571	3.204067043	
10	3	11.76586862	3.554538257	3.310097619	
11	4	11.44456518	3.33708406	3.429510606	
12	5	9.373641185	3.692222659	2.538752955	
13	6	10.47906344	2.598868941	4.0321631	
14	7	14.31716958	3.203954788	4.46859289	
15	8	8.901052248	0.729081227	12.20858791	
16	9	13.99414343	3.180751244	4.399634662	
17	10	12.5758327	3.513579887	3.579207847	

**FIGURE 4.3** Sampling Experiment for Profitability Index



**FIGURE 4.4** Histogram of Profitability Index

## SAMPLING DISTRIBUTIONS AND SAMPLING ERROR

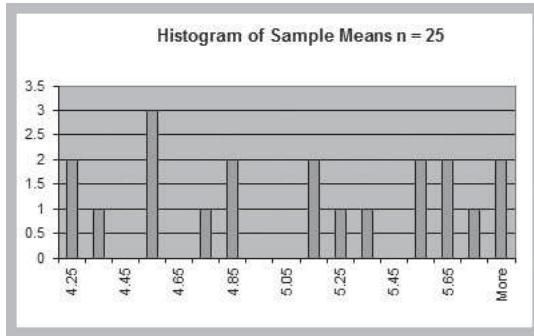
Whenever we collect data, we are essentially taking a sample from some generally unknown probability distribution. Usually, the goal is to estimate a population parameter, such as the mean. An important statistical question is: How good is the estimate obtained from the sample? We could gain some insight into this question by performing a sampling experiment.

Let us assume that a random variable is uniformly distributed between 0 and 10. First, we will compute the characteristics of this random variable. Using the formulas introduced in Chapter 3, we find that the expected value is  $(0 + 10)/2 = 5$ , the variance is  $(10 - 0)^2/12 = 8.33$ , and the standard deviation is 2.89. Suppose we generate a sample of size  $n = 25$  for this random variable using the Excel formula = 10\*RAND() and compute the sample mean. Note that each of the 25 values is uniformly distributed between 0 and 10. When averaged together, we would expect the sample mean to be close to 5, but probably not exactly equal to 5 because of the randomness in the sampled values.

Now suppose that we repeat this experiment, say 20 times, and obtain a set of 20 sample means. Figure 4.5 shows a portion of a spreadsheet for this experiment. (The Excel file is available on the Companion Website as *Sampling Error Experiment* if you wish to experiment further.) Figure 4.6 shows a histogram of 20 sample means generated in this fashion. For a sample size of 25, the sample means seem to be rather uniformly spread out.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
1	Sampling Error Experiment																					
2	Instructions:	The worksheet is designed for 20 samples with sample sizes of up to 500. To change the sample size, simply change the range in the formulas in row 6 for computing the sample mean to include the appropriate number of observations.																				
3																						
4																						
5	Experiment	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
6	Sample Mean	5.011	4.774	4.983	4.808	5.130	5.030	5.118	4.958	5.155	5.128	5.031	5.014	4.821	4.838	4.853	5.070	5.072	5.020	4.953	5.109	
7																						
8	Sample	1	1.528	8.061	7.232	2.881	6.074	8.949	9.507	2.190	7.463	2.487	2.878	2.504	2.927	4.982	6.996	0.999	8.756	6.911	7.891	2.572
9		2	5.299	7.018	0.975	1.298	6.561	0.731	1.612	6.376	5.861	5.613	6.556	3.836	1.231	6.969	7.960	7.285	5.870	4.508	1.678	7.679
10		3	6.174	2.333	4.188	4.832	2.992	7.205	8.786	1.440	4.328	7.471	0.459	2.217	7.020	3.243	6.356	7.152	2.631	2.845	9.526	8.879
11		4	3.609	4.155	3.552	0.188	0.619	1.766	5.173	4.625	0.339	4.067	5.250	9.360	7.693	0.239	6.761	0.449	5.816	7.534	8.348	0.074
12		5	9.000	0.636	8.796	0.916	9.754	1.690	4.458	5.460	2.032	9.930	9.107	0.522	1.899	1.672	8.205	6.523	7.115	5.267	0.558	8.224

**FIGURE 4.5** Portion of Spreadsheet for Sampling Error Experiment

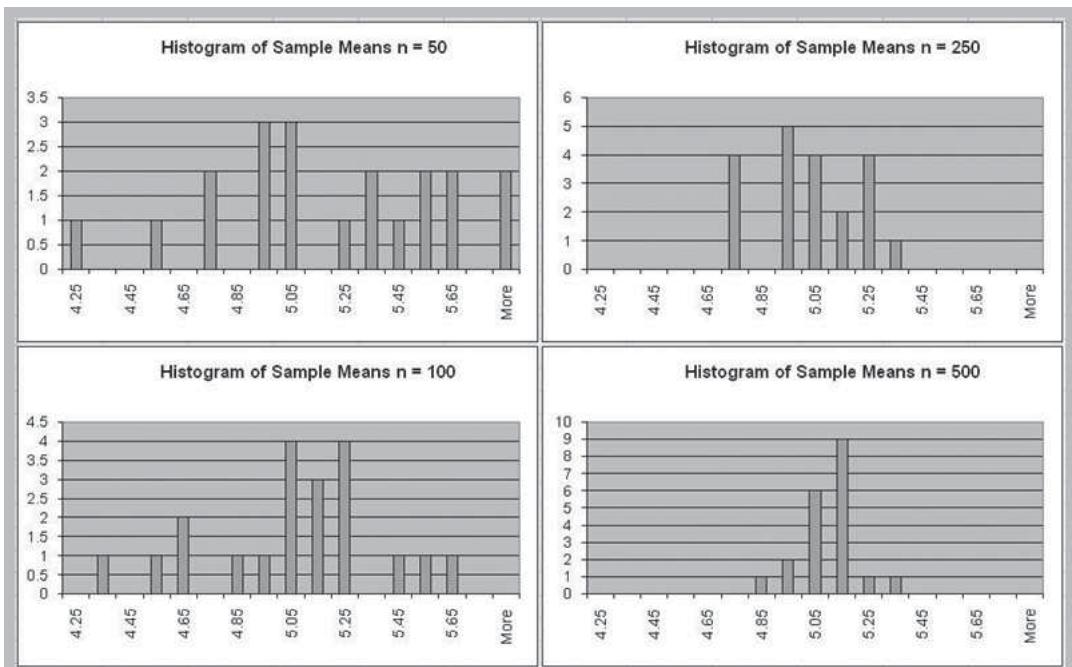


**FIGURE 4.6** Histogram of 20 Sample Means for  $n = 25$

**TABLE 4.2** Results from Sampling Error Experiment

Sample Size	Average of 20 Sample Means	Standard Deviation of 20 Sample Means
25	5.166	0.639449
50	4.933	0.392353
100	4.948	0.212670
250	5.028	0.167521
500	4.997	0.115949

Now let us repeat this experiment for larger sample sizes. Table 4.2 shows some results. Notice that as the sample size gets larger, the average of the 20 sample means seems to be getting closer to the expected value of 5, and also, the standard deviation of the 20 sample means becomes smaller, meaning that the means of these 20 samples are clustered closer together around the true expected value. Figure 4.7 shows histograms of



**FIGURE 4.7** Histograms of Sample Means for Increasing Sample Sizes

the sample means for each of these cases. These illustrate the conclusions we just made, and also, perhaps more surprisingly, the distribution of the sample means appears to assume the shape of a normal distribution for larger sample sizes!

In our experiment, we used only 20 sample means. If we had used a much larger number, the distributions would have been more well defined. The means of *all possible* samples of a fixed size  $n$  from some population will form a distribution that we call the **sampling distribution of the mean**. Statisticians have shown that the sampling distribution of the mean has two key properties. First, the standard deviation of the sampling distribution of the mean is called the **standard error of the mean** and is computed as:

$$\text{Standard Error of the Mean} = \sigma / \sqrt{n} \quad (4.3)$$

where  $\sigma$  is the standard deviation of the population from which the individual observations are drawn, and  $n$  is the sample size. From this formula, we see that as  $n$  increases, the standard error decreases, just as our experiment demonstrated. This suggests that the estimates of the mean that we obtain from larger sample sizes provide greater accuracy in estimating the true population mean. In other words, larger sample sizes have *less sampling error*.

For our experiment, we know that the variance of the population is 8.33 (because the values were uniformly distributed). Therefore, the standard deviation of the population is  $\sigma = 2.89$ . We may compute the standard error of the mean for each of the sample sizes in our experiment using Formula (4.3). This yields:

Sample Size, $n$	Standard Error of the Mean
25	0.577
50	0.408
100	0.289
250	0.183
500	0.129

The standard deviations shown in Table 4.2 are *estimates* of the standard error of the mean based on the limited number of 20 samples. If we compare these estimates with the theoretical values above, we see that they are close but not exactly the same. This is simply because the true standard error is based on all possible sample means in the sampling distribution, whereas we used only 20. If you repeat the experiment with a larger number of samples, the observed values of the standard error would be closer to these theoretical values.

In practice, we will never know the true population standard deviation, and generally only take a limited sample of  $n$  observations. However, we may estimate the standard error of the mean using the sample data by simply dividing the sample standard deviation by the square root of  $n$ .

The second property that statisticians have shown is called the **central limit theorem**, one of the most important practical results in statistics. The central limit theorem states that if the sample size is large enough, the sampling distribution of the mean is approximately normally distributed, *regardless* of the distribution of the population, and that the mean of the sampling distribution will be the same as that of the population. This is exactly what we observed in our experiment. The distribution of the population was uniform, yet the sampling distribution of the mean converges to a normal distribution as the sample size increases. The central limit theorem also states that if the population is normally distributed, then the sampling distribution of the mean will also be normal for *any* sample size. The central limit theorem allows us to use the theory we learned about calculating probabilities for normal distributions to draw conclusions about sample means.

## SKILL-BUILDER EXERCISE 4.3

---

Generate 20 groups of 10 uniformly distributed numbers, and calculate the mean of each group. Compute the mean and variance of all 200 values, as well as the mean and variance of the 20 means. Compute the standard error of the mean using the 20 sample means and compare this to  $s/\sqrt{n}$  for the entire sample. Explain your results.

### Applying the Sampling Distribution of the Mean

Suppose that the size of individual customer orders (in dollars),  $X$ , from a major discount book publisher Web site is normally distributed with a mean of \$36 and standard deviation of \$8. The probability that the next individual who places an order at the Web site will spend more than \$40 can be found by calculating:

$$1 - \text{NORM.DIST}(40, 36, 8, \text{TRUE}) = 1 - 0.6915 = 0.3085$$

Now suppose that a sample of 16 customers is chosen. What is the probability that the *mean purchase* for these 16 customers will exceed \$40? To find this, we must realize that we must use the sampling distribution of the mean to carry out the appropriate calculations. The sampling distribution of the mean will have a mean of \$36, but a standard error of  $\$8/\sqrt{16} = \$2$ . Then the probability that the mean purchase exceeds \$40 for a sample size of  $n = 16$  is:

$$1 - \text{NORM.DIST}(40, 36, 2, \text{TRUE}) = 1 - 0.9772 = 0.0228$$

While about 30% of individuals will make purchases exceeding \$40, the chance that 16 customers will collectively average more than \$40 is much smaller. It would be very unlikely for 16 customers to all make high-volume purchases, as some individual purchases would as likely be less than \$36 as more, making the variability of the mean purchase amount for the sample of 16 much smaller than for individuals.

The key to applying sampling distribution correctly is to understand whether the probability that you wish to compute relates to an individual observation or to the mean of a sample. If it relates to the mean of a sample, then you must use the sampling distribution of the mean, whose standard deviation is the standard error,  $\sigma/\sqrt{n}$ . Understanding the standard error of the mean and characteristics of the sampling distribution is also important for designing sampling plans and performing various statistical tests. We address these issues in this and subsequent chapters.

## SAMPLING AND ESTIMATION

Sample data provide the basis for many useful analyses to support decision making. **Estimation** involves assessing the value of an unknown population parameter—such as a population mean, population proportion, or population variance—using sample data. When we sample, the estimators we use—such as a sample mean, sample proportion, or sample variance—are random variables that are characterized by some sampling distribution. By knowing the sampling distribution of the estimator, we can use probability theory to quantify the uncertainty associated with the estimator.

We use two types of estimates in statistics. **Point estimates** are single numbers used to estimate the value of a population parameter. However, because of sampling error, it is unlikely that a point estimate will equal the true value of the population parameter, and the point estimate alone does not provide any information of the magnitude of the sampling error. **Confidence interval estimates** provide a range of values between which the value of the population parameter is believed to be, and also provide an assessment of sampling error by specifying a probability that the interval correctly estimates the

**TABLE 4.3 Sampling and Estimation Support in Excel**

Excel 2010 Function	Description
CONFIDENCE.NORM( <i>alpha</i> , <i>standard_dev</i> , <i>size</i> )	Returns the confidence interval for a population mean using a normal distribution
CONFIDENCE.T( <i>alpha</i> , <i>standard_dev</i> , <i>size</i> )	Returns the confidence interval for a population mean using a <i>t</i> -distribution
T.INV( <i>probability</i> , <i>deg_freedom</i> )	Returns the left-tailed inverse of the <i>t</i> -distribution
CHISQ.DIST( <i>x</i> , <i>deg_freedom</i> )	Returns the probability above <i>x</i> for a given value of degrees of freedom.
CHISQ.INV( <i>probability</i> , <i>deg_freedom</i> )	Returns the value of <i>x</i> that has a right-tail area equal to <i>probability</i> for a specified degree of freedom.
Analysis Toolpak Tools	Description
SAMPLING	Creates a simple random sample with replacement or a systematic sample from a population
PHStat Add-In	Description
Random Sample Generator	Generates a random sample without replacement
Confidence Intervals	Computes confidence intervals for means with $\sigma$ known or unknown, proportions, and population total
Sample Size	Determines sample sizes for means and proportions

true (unknown) population parameter. Microsoft Excel and the *PHStat* add-in provide several options for supporting these analyses, as summarized in Table 4.3.

## Point Estimates

The most common point estimates are the descriptive statistical measures we described in Chapter 2 and summarized in Table 4.4 along with their corresponding population parameters. They are used to estimate the population parameters, also listed in Table 4.4.

Figure 4.8 shows the 10 samples from the *Cereal Data* file that were randomly selected in Figure 4.2. We calculated the sample mean and sample standard deviation for calories, sodium, fiber, carbohydrates, and sugars in the sample as well as the population mean and population standard deviation for the entire data set. The sample statistics are point estimates. Notice that there are some considerable differences as compared to the population parameters because of sampling error. A point estimate alone does not provide any indication of the magnitude of the potential error in the estimate. A major metropolitan newspaper reported that college professors were the highest-paid workers in the region, with an average of \$150,004, based on a Bureau of Labor Statistics survey. Actual averages for two local universities were less than \$70,000. What happened? As reported in a follow-up story, the sample size was very small and included a large

**TABLE 4.4 Common Point Estimates**

Point Estimate	Population Parameter
Sample mean, $\bar{x}$	Population mean, $\mu$
Sample variance, $s^2$	Population variance, $\sigma^2$
Sample standard deviation, $s$	Population standard deviation, $\sigma$
Sample proportion, $\hat{p}$	Population proportion, $\pi$

	A	B	C	D	E	F
1	Cereal Name	Calories	Sodium	Fiber	Carbs	Sugars
2	Lucky Charms	110	180	0	12	12
3	Apple Jacks	110	125	1	11	14
4	Apple Cinn Cheerios	110	180	1.50	10.50	10
5	Crispy Wheat & Raisins	100	140	2	11	10
6	Honey Graham Ohs	120	220	1	12	11
7	Puffed Wheat	50	0	1	10	0
8	Honey Nut Cheerios	110	250	1.50	11.50	10
9	Trix	110	140	0	13	12
10	Post Nat. Raisin Bran	120	200	6	11	14
11	Almond Delight	110	200	1	14	8
12	Sample Mean	105	163.5	1.5	11.6	10.1
13	Sample Standard Deviation	20.1384	69.284	1.7	1.197	4.0125
14						
15	Population Mean	105.522	167.313	2.187	14.77	6.9552
16	Population Standard Deviation	18.631	79.9782	2.487	3.831	4.3758

**FIGURE 4.8** Point Estimates for Cereal Samples

number of highly paid medical school faculty; as a result, the sampling error was huge. Interval estimates, which we will discuss soon, provide better information than point estimates alone.

### Unbiased Estimators

It seems quite intuitive that the sample mean should provide a good point estimate for the population mean. However, it may not be clear why the formula for the sample variance that we introduced in Chapter 2 has a denominator of  $n - 1$ , particularly because it is different from the formula for the population variance [see Formulas (2A.5) and (2A.6) in Appendix 2.1]. Recall that the population variance is computed by:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{N}$$

whereas the sample variance is computed by the formula:

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

Why is this so? Statisticians develop many types of estimators, and from a theoretical as well as a practical perspective, it is important that they “truly estimate” the population parameters they are supposed to estimate. Suppose that we perform an experiment in which we repeatedly sampled from a population and computed a point estimate for a population parameter. Each individual point estimate will vary from the population parameter; however, we would hope that the long-term average (expected value) of all possible point estimates would equal the population parameter. If the expected value of an estimator equals the population parameter it is intended to estimate, the estimator is said to be **unbiased**. If this is not true, the estimator is called **biased** and will not provide correct results.

Fortunately, all the estimators in Table 4.4 are unbiased and, therefore, are meaningful for making decisions involving the population parameter. In particular, statisticians have shown that the denominator  $n - 1$  used in computing  $s^2$  is necessary to provide an unbiased estimator of  $\sigma^2$ . If we simply divide by the number of observations, the estimator would tend to underestimate the true variance.

## SKILL-BUILDER EXERCISE 4.4

In the Excel file *Cereal Data*, we have computed the population standard deviation of sugars as 4.376. Using a method to generate random samples, generate 20 random samples of 30 cereals and compute the standard deviation of each sample using both the population formula and sample formula. Compare your results with the true population standard deviation and comment about the fact that the sample formula is an unbiased estimator.

### Interval Estimates

An **interval estimate** provides a range for a population characteristic based on a sample. Intervals are quite useful in statistics because they provide more information than a point estimate. Intervals specify a range of plausible values for the characteristic of interest and a way of assessing “how plausible” they are. In general, a **100(1 –  $\alpha$ )% probability interval** is any interval  $[A, B]$  such that the probability of falling between A and B is  $1 - \alpha$ . Probability intervals are often centered on the mean or median. For instance, in a normal distribution, the mean  $\pm$  1 standard deviation describes an approximate 68% probability interval around the mean. As another example, the 5th and 95th percentiles in a data set constitute a 90% probability interval.

We see interval estimates in the news all the time when trying to estimate the mean or proportion of a population. Interval estimates are often constructed by taking a point estimate and adding and subtracting a margin of error that is based on the sample size. For example, a Gallup poll might report that 56% of voters support a certain candidate with a margin of error of  $\pm 3\%$ . We would conclude that the true percentage of voters that support the candidate is most likely between 53% and 59%. Therefore, we would have a lot of confidence in predicting that the candidate would win a forthcoming election. If, however, the poll showed a 52% level of support with a margin of error of  $\pm 4\%$ , we might not be as confident in predicting a win because the true percentage of supportive voters is likely to be somewhere between 48% and 56%. The question you might be asking at this point is how to calculate the margin of error. In national surveys and political polls, such margins of error are usually stated, but they are never clearly explained! To understand them, we need to introduce the concept of confidence intervals.

### CONFIDENCE INTERVALS: CONCEPTS AND APPLICATIONS

A **confidence interval** is an interval estimate that also specifies the likelihood that the interval contains the true population parameter. This probability is called the **level of confidence**, denoted by  $1 - \alpha$ , where  $\alpha$  is a number between 0 and 1. The level of confidence is usually expressed as a percentage; common values are 90%, 95%, or 99%. (Note that if the level of confidence is 90%, then  $\alpha = 0.1$ .) The margin of error depends on the level of confidence and the sample size. For example, suppose that the margin of error for some sample size and a level of confidence of 95% is calculated to be 2.0. One sample might yield a point estimate of 10. Then a 95% confidence interval would be [8, 12]. However, this interval may or may not include the true population mean. If we take a different sample, we will most likely have a different point estimate, say 10.4, which, given the same margin of error, would yield the interval estimate [8.4, 12.4]. Again, this may or may not include the true population mean. If we chose 100 different samples, leading to 100 different interval estimates, we would expect that 95% of them—the level of confidence—would contain the true population mean. We would say we are “95% confident” that the interval we obtain from sample data contains the true population mean. The higher the confidence level, the more assurance we have that the interval contains the true population parameter. As the confidence level increases,

**TABLE 4.5** Common Confidence Intervals

Type of Confidence Interval	PHStat Tool
Mean, standard deviation known	Estimate for the mean, sigma known
Mean, standard deviation unknown	Estimate for the mean, sigma unknown
Proportion	Estimate for the proportion
Variance	Estimate for the population variance
Population total	Estimate for the population total

the confidence interval becomes larger to provide higher levels of assurance. You can view  $\alpha$  as the risk of incorrectly concluding that the confidence interval contains the true mean.

When national surveys or political polls report an interval estimate, they are actually confidence intervals. However, the level of confidence is generally not stated because the average person would probably not understand the concept or terminology. While not stated, you can probably assume that the level of confidence is 95%, as this is the most common value used in practice.

Many different types of confidence intervals may be developed. The formulas used depend on the population parameter we are trying to estimate and possibly other characteristics or assumptions about the population. Table 4.5 provides a summary of the most common types of confidence intervals and *PHStat* tools available for computing them (no tools are available in Excel for computing confidence intervals). All tools can be found in the *Confidence Intervals* menu within *PHStat*. We will discuss other types of confidence intervals later in this chapter and in Appendix 4.1.

### Confidence Interval for the Mean with Known Population Standard Deviation

The simplest type of confidence interval is for the mean of a population where the standard deviation is assumed to be known. You should realize, however, that in nearly all practical sampling applications, the population standard deviation will *not* be known. However, in some applications, such as measurements of parts from an automated machine, a process might have a very stable variance that has been established over a long history, and it can reasonably be assumed that the standard deviation is known.

To illustrate this type of confidence interval, we will use the cereal samples in Figure 4.8, as we have already calculated the population standard deviation of the cereal characteristics. For example, the point estimate for calories for the 10 samples of cereals was calculated as 105.0. The population standard deviation is known to be 18.631. In most practical applications, samples are drawn from very large populations. If the population is relatively small compared to the sample size, a modification must be made to the confidence interval. Specifically, when the sample size,  $n$ , is larger than 5% of the population size,  $N$ , a correction factor is needed in computing the margin of error. In this example, the population is only 67, so the 10 samples represent about 15% of the population. Therefore, in using the *PHStat* tool (see Appendix 4.2D, *Confidence Intervals for the Mean*), we checked the box for the finite population correction (FPC).

Figure 4.9 shows the *PHStat* results. The 95% confidence interval (including the FPC) is [94.27, 115.73]. This means that we believe that the population mean of the number of calories per serving of breakfast cereals is somewhere between 94.27 and 115.73, with only a small likelihood (0.05) that the population mean is outside of this interval.



Spreadsheet Note

A	B
1	Calories
2	
3	<b>Data</b>
4	Population Standard Deviation      18.631
5	Sample Mean      105
6	Sample Size      10
7	Confidence Level      95%
8	
9	Intermediate Calculations
10	Standard Error of the Mean      5.891639509
11	Z Value      -1.95996398
12	Interval Half Width      11.54740125
13	
14	<b>Confidence Interval</b>
15	Interval Lower Limit      93.45259875
16	Interval Upper Limit      116.5474012
17	
18	
19	<b>Finite Populations</b>
20	Population Size      67
21	FPC Factor      0.929320377
22	Interval Half Width      10.73123528
23	Interval Lower Limit      94.26876472
24	Interval Upper Limit      115.7312353

**FIGURE 4.9** *PHStat* Results for a 95% Confidence Interval for Calories

To fully understand these results, it is necessary to examine the formula used to compute the confidence interval. A  $100(1 - \alpha)\%$  confidence interval for the population mean  $\mu$  is given by:

$$\bar{x} \pm z_{\alpha/2}(\sigma / \sqrt{n}) \quad (4.4)$$

Note that this formula is simply the sample mean (point estimate) plus or minus a margin of error. The margin of error is a number  $z_{\alpha/2}$  times the standard error of the sampling distribution of the mean,  $\sigma / \sqrt{n}$ . The value  $z_{\alpha/2}$  represents the value of a standard normal random variable that has a cumulative probability of  $\alpha/2$  (the reasoning is explained in Appendix 4.1). It may be found from the standard normal table (see Table A.1 in the appendix at the end of the book) or may be computed in Excel using the function NORMSINV( $\alpha/2$ ). These values are shown in the *Intermediate Calculations* section of the *PHStat* results. The standard error is 5.8916 and  $z_{\alpha/2}$  (z-value) for  $\alpha = 0.05$  is approximately  $-1.96$ . Note that the z-value is negative because  $\alpha/2$  represents a small area in the *left* tail of the standard normal distribution; however, to calculate the margin of error, the positive value is used. Therefore, the margin of error, or interval half-width, is  $(5.8916)(1.96) = 11.547$ , resulting in the confidence interval  $105 \pm 11.547$  or  $[93.453, 116.547]$ , without consideration of the FPC. For finite populations, the FPC factor is:

$$\sqrt{\frac{N - n}{N - 1}}$$

In this example, the FPC factor is 0.929 and is multiplied by the standard error in order to find the adjusted interval half-width and the confidence interval. That is, the adjusted standard error of the mean is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}} \quad (4.5)$$

Note that if  $n$  is small relative to  $N$ , then the term  $\sqrt{(N - n)/(N - 1)}$  is approximately 1 and the difference is minor. Also, note that the confidence interval in Figure 4.9 is slightly smaller when the FPC factor is included.

Note that as the level of confidence,  $1 - \alpha$ , decreases,  $z_{\alpha/2}$  decreases, and the confidence interval becomes smaller. For example, a 90% confidence interval (with  $z_{0.05} = 1.64$ ) will be smaller than a 95% confidence interval (with  $z_{0.025} = 1.96$ ). Similarly, a 99% confidence interval (with  $z_{0.005} = 2.76$ ) will be larger than a 95% confidence interval. Essentially, you must trade off a higher level of accuracy with the risk that the confidence interval does not contain the true mean. Smaller risk will result in a larger confidence interval. However, you can also see that as the sample size increases, the standard error decreases, making the confidence interval smaller and providing a more accurate interval estimate for the same level of risk. So if you wish to reduce the risk, you should consider increasing the sample size.

### SKILL-BUILDER EXERCISE 4.5

Generate 50 random samples of size 10 from the Excel file *Cereal Data*, and compute a 90% confidence interval for the mean of Carbs for each sample, using the known population standard deviation of 3.831. Determine how many confidence intervals actually contain the true population mean 14.77.

### Confidence Interval for the Mean with Unknown Population Standard Deviation

In most practical applications, the standard deviation of the population is unknown, and we need to calculate the confidence interval differently. For example, in the Excel file *Credit Approval Decisions*, a bank has sample data used in making credit approval decisions (see Figure 4.10). Suppose that we want to estimate the mean number revolving balance for the population of applicants that own a home (which we assume is large so that the FPC factor is not required). After sorting the data by Homeowner, we could use the *PHStat* tool *Estimate for the Mean, Sigma Unknown* to find a confidence interval using these data. Figure 4.11 shows the results. The tool calculates the sample statistics and the confidence interval using the intermediate calculations. The confidence interval is the sample mean plus or minus the interval half-width:  $\$12,630.37 \pm \$2,133.55$  or  $[\$10,496.82, \$14,763.928]$ .

You will notice that the intermediate calculations are somewhat different from the case in which the population standard deviation was known. Instead of using  $z_{\alpha/2}$  based on the normal distribution, the tool uses a “*t*-value” with which to multiply the standard error to compute the interval half-width. The *t*-value comes from a new probability distribution called the ***t-distribution***. The *t*-distribution is actually a family of

	A	B	C	D	E	F
1	Credit Approval Decisions					
2						
3	Homeowner	Credit Score	Years of Credit History	Revolving Balance	Revolving Utilization	Decision
4	Y	725	20	\$ 11,320	25%	Approve
5	Y	573	9	\$ 7,200	70%	Reject
6	Y	677	11	\$ 20,000	55%	Approve
7	N	625	15	\$ 12,800	65%	Reject
8	N	527	12	\$ 5,700	75%	Reject
9	Y	795	22	\$ 9,000	12%	Approve
10	N	733	7	\$ 35,200	20%	Approve

FIGURE 4.10 Portion of Excel File *Credit Approval Decisions*

A	B
1 Revolving Balance - Homeowners	
2	
3 Data	
4 Sample Standard Deviation	5393.384467
5 Sample Mean	12630.37037
6 Sample Size	27
7 Confidence Level	95%
8	
9 Intermediate Calculations	
10 Standard Error of the Mean	1037.957325
11 Degrees of Freedom	26
12 t Value	2.055529439
13 Interval Half Width	2133.551837
14	
15 Confidence Interval	
16 Interval Lower Limit	10496.82
17 Interval Upper Limit	14763.92

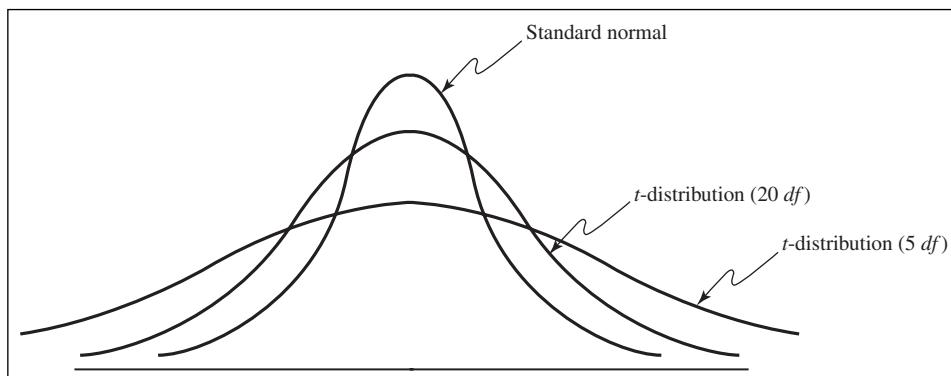
**FIGURE 4.11** Confidence Interval for Revolving Balance of Homeowners

probability distributions with a shape similar to the standard normal distribution. Different *t*-distributions are distinguished by an additional parameter, **degrees of freedom (df)**. The *t*-distribution has a larger variance than the standard normal, thus making confidence intervals wider than those obtained from the standard normal distribution, in essence correcting for the uncertainty about the true standard deviation. As the number of degrees of freedom increases, the *t*-distribution converges to the standard normal distribution (Figure 4.12). When sample sizes get to be as large as 120, the distributions are virtually identical; even for sample sizes as low as 30–35, it becomes difficult to distinguish between the two. Thus, for large sample sizes, many people use *z*-values to establish confidence intervals even when the standard deviation is unknown. We must point out, however, that for any sample size, the *true* sampling distribution of the mean is the *t*-distribution, so when in doubt, use the *t*.

The concept of “degrees of freedom” can be puzzling. It can best be explained by examining the formula for the sample variance:

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

Note that to compute  $s^2$  we need to first compute the sample mean,  $\bar{x}$ . If we know the value of the mean, then we need only know  $n - 1$  distinct observations; the  $n$ th is



**FIGURE 4.12** Comparison of the *t*-Distribution to the Standard Normal Distribution

completely determined. (For instance, if the mean of 3 values is 4, and you know that two of the values are 2 and 4, you can easily determine that the third number must be 6.) The number of sample values that are free to vary defines the number of degrees of freedom; in general,  $df$  equals the number of sample values minus the number of estimated parameters. Because the sample variance uses one estimated parameter, the mean, the  $t$ -distribution used in confidence interval calculations has  $n - 1$  degrees of freedom. Because the  $t$ -distribution explicitly accounts for the effect of the sample size in estimating the population variance, it is the proper one to use for any sample size. However, for large samples, the difference between  $t$ - and  $z$ -values is very small, as we noted earlier.

The formula for a  $100(1 - \alpha)\%$  confidence interval for the mean  $\mu$  when the population standard deviation is unknown is:

$$\bar{x} \pm t_{\alpha/2,n-1} \left( s / \sqrt{n} \right) \quad (4.6)$$

where  $t_{\alpha/2,n-1}$  is the value from the  $t$ -distribution with  $n - 1$  degrees of freedom, giving an upper-tail probability of  $\alpha/2$ . We may find values of  $t_{\alpha/2,n-1}$  in Table A.2 in the appendix at the end of the book or by using the Excel function T.INV( $1 - \alpha/2, n - 1$ ). (Note that the Excel function uses  $1 - \alpha/2$  as an argument, not  $\alpha/2$  in order to obtain the upper tail value.) Thus, in our example,  $t_{0.025,26} = \text{T.INV}(0.975, 26) = 2.055$ , and the standard error  $s / \sqrt{n} = 1.037.96$ , resulting in the confidence interval  $\$12,630.37 \pm 2.055(\$1,037.95)$ .

### Confidence Interval for a Proportion

For categorical variables having only two possible outcomes, such as good or bad, male or female, and so on, we are usually interested in the *proportion* of observations in a sample that has a certain characteristic. An unbiased estimator of a population proportion  $\pi$  is the statistic  $\hat{p} = x/n$  (the **sample proportion**), where  $x$  is the number in the sample having the desired characteristic and  $n$  is the sample size. For example, the last column in the Excel file *Insurance Survey* (see Figure 4.13) describes whether a sample of

	A	B	C	D	E	F	G
1	Insurance Survey						
2							
3	Age	Gender	Education	Marital Status	Years Employed	Satisfaction*	Premium/Deductible**
4	36	F	Some college	Divorced	4	4	N
5	55	F	Some college	Divorced	2	1	N
6	61	M	Graduate degree	Widowed	26	3	N
7	65	F	Some college	Married	9	4	N
8	53	F	Graduate degree	Married	6	4	N
9	50	F	Graduate degree	Married	10	5	N
10	28	F	College graduate	Married	4	5	N
11	62	F	College graduate	Divorced	9	3	N
12	48	M	Graduate degree	Married	6	5	N
13	31	M	Graduate degree	Married	1	5	N
14	57	F	College graduate	Married	4	5	N
15	44	M	College graduate	Married	2	3	N
16	38	M	Some college	Married	3	2	N
17	27	M	Some college	Married	2	3	N
18	56	M	Graduate degree	Married	4	4	Y
19	43	F	College graduate	Married	5	3	Y
20	45	M	College graduate	Married	15	3	Y
21	42	F	College graduate	Married	12	3	Y
22	29	M	Graduate degree	Single	10	5	N
23	28	F	Some college	Married	3	4	Y
24	36	M	Some college	Divorced	15	4	Y
25	49	F	Graduate degree	Married	2	5	N
26	46	F	College graduate	Divorced	20	4	N
27	52	F	College graduate	Married	18	2	N
28	*Measured from 1-5 with 5 being highly satisfied.						
29	**Would you be willing to pay a lower premium for a higher deductible?						

FIGURE 4.13 Excel File: Insurance Survey

A	B
1 Insurance Survey	
2	
3 Data	
4 Sample Size	24
5 Number of Successes	6
6 Confidence Level	95%
7	
8 Intermediate Calculations	
9 Sample Proportion	0.25
10 Z Value	-1.95996398
11 Standard Error of the Proportion	0.088388348
12 Interval Half Width	0.173237978
13	
14 Confidence Interval	
15 Interval Lower Limit	0.076762022
16 Interval Upper Limit	0.423237978

**FIGURE 4.14** Confidence Interval for the Proportion

employees would be willing to pay a lower premium for a higher deductible for their health insurance. Suppose we are interested in the proportion of individuals who answered yes. We may easily confirm that 6 out of the 24 employees, or 25%, answered yes. Thus, a point estimate for the proportion answering yes is  $\hat{p} = 0.25$ . Using the *PHStat* tool *Estimate for the Proportion* (see Appendix 4.2E, *Confidence Interval for Proportions*), we find that a 95% confidence interval for the proportion of employees answering yes is [0.077, 0.423]. This is shown in Figure 4.14. Notice that this is a fairly large confidence interval, suggesting that we have quite a bit of uncertainty as to the true value of the population proportion. This is because of the relatively small sample size.

These calculations are based on the following: a  $100(1 - \alpha)$  confidence interval for the proportion is:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad (4.7)$$

Notice that as with the mean, the confidence interval is the point estimate plus or minus some margin of error. In this case,  $\sqrt{\hat{p}(1 - \hat{p})}/n$  is the standard error for the sampling distribution of the proportion.

## Confidence Intervals for the Variance and Standard Deviation

Understanding variability is critical to effective decisions. Thus, in many situations, one is interested in obtaining point and interval estimates for the variance or standard deviation. For example, the Excel file *Home Market Value* provides some data on a sample of houses in a residential neighborhood (see Figure 4.15). The standard deviation of market value provides information on the spread of home prices in this neighborhood. Using the sample standard deviation, \$10,553, and a 2-standard deviation spread, we might predict that about 95% of home prices (assuming a normal distribution) would deviate about  $2(\$10,553) = \$21,106$  from the average value. However, because of sampling error, this value could be quite different.

*PHStat* has a tool for computing confidence intervals for the variance and standard deviation (see Appendix 4.2F, *Confidence Intervals for the Population Variance*). The calculations assume that the population from which the sample was drawn has an approximate normal distribution. If this assumption is not met, the confidence interval may not be accurate for the confidence level chosen. Using the *PHStat* tool for the *Home Market Value* data, we obtain the results in Figure 4.16. A 95% confidence interval for the standard deviation of market values is [\$8,683, \$13,457]. In other words, the standard deviation might be as low as about \$8,600 or even \$13,000. Thus, a 2-standard deviation



Spreadsheet Note



Spreadsheet Note

	A	B	C
1	Home Market Value		
2			
3	House Age	Square Feet	Market Value
4	33	1,812	\$90,000.00
5	32	1,914	\$104,400.00
6	32	1,842	\$93,300.00
7	33	1,812	\$91,000.00
8	32	1,836	\$101,900.00
9	33	2,028	\$108,500.00
10	32	1,732	\$87,600.00
11	33	1,850	\$96,000.00

**FIGURE 4.15** Portion of Excel File Home Market Value

spread of home prices from the average might be as small as  $2(\$8,683) = \$17,366$  or as large as  $2(\$13,457) = \$26,914$ .

Calculation of the confidence intervals for variance and standard deviation is quite different from the other confidence intervals we have studied. Although we use the sample standard deviation  $s$  as a point estimate for  $\sigma$ , the sampling distribution of  $s$  is not normal, but is related to a special distribution called the **chi-square ( $\chi^2$ ) distribution**. The chi-square distribution is characterized by degrees of freedom, similar to the  $t$ -distribution. Table A.3 in the appendix in the back of this book provides critical values of the chi-square distribution for selected values of  $\alpha$ . The Excel function CHISQ.DIST( $x, \text{deg\_freedom}$ ) returns the probability above  $x$  for a given value of degrees of freedom. Also, the Excel function CHISQ.INV( $\text{probability}, \text{deg\_freedom}$ ) returns the value of  $x$  that has a right-tail area equal to  $\text{probability}$  for a specified degree of freedom.

However, unlike the normal or  $t$ -distributions, the chi-square distribution is not symmetric, which means that the confidence interval is not simply a point estimate plus or minus some number of standard errors. The point estimate is always

	A	B	C	D	E
1	Home Market Value				
2					
3	Data				
4	Sample Size	42			
5	Sample Standard Deviation	10553.1			
6	Confidence Level	95%			
7					
8	Intermediate Calculations				
9	Degrees of Freedom	41			
10	Sum of Squares	4.57E+09			
11	Single Tail Area	0.025			
12	Lower Chi-Square Value	25.21452			
13	Upper Chi-Square Value	60.56057			
14					
15	Results				
16	Interval Lower Limit for Variance	7.5E+07			
17	Interval Upper Limit for Variance	1.8E+08			
18					
19	Interval Lower Limit for Standard Deviation	8683.14			
20	Interval Upper Limit for Standard Deviation	13456.9			
21					
22	Assumption:				
23	Population from which sample was drawn has an approximate normal distribution.				

**FIGURE 4.16** PHStat Results for Confidence Interval for Population Variance

closer to the left endpoint of the interval. A  $100(1 - \alpha)\%$  confidence interval for the variance is:

$$\left[ \frac{(n-1)s^2}{\chi_{n-1, \alpha/2}^2}, \frac{(n-1)s^2}{\chi_{n-1, 1-\alpha/2}^2} \right] \quad (4.8)$$

We may compute  $\chi_{41, 0.025}^2$  using the Excel function CHISQ.INV(0.025, 41) = 60.56057 and  $\chi_{41, 0.975}^2$  by CHISQ.INV(0.975, 41) = 25.21452, as shown in the Intermediate Calculations section of Figure 4.16. The confidence interval limits for the standard deviation is simply the square root of the confidence interval limits of the variance.

### Confidence Interval for a Population Total

In some applications, we might be more interested in estimating the *total* of a population rather than the mean. For instance, an auditor might wish to estimate the total amount of receivables by sampling a small number of accounts. If a population of  $N$  items has a mean  $\mu$ , then the population total is  $N\mu$ . We may estimate a population total from a random sample of size  $n$  from a population of size  $N$  by the point estimate  $N\bar{x}$ . For example, suppose that an auditor in a medical office wishes to estimate the total amount of unpaid reimbursement claims for 180 accounts over 60 days old. A sample of 20 from this population yielded a mean amount of unpaid claims of \$185 and the sample standard deviation is \$22. Using the *PHStat* confidence interval tool *Estimate for the Population Proportion*, we obtain the results shown in Figure 4.17. A 95% confidence interval is [\$31,547.78, \$35,052.22].

These calculations are based on the following: a  $100(1 - \alpha)\%$  confidence interval for the population total is:

$$N\bar{x} \pm t_{\alpha/2, n-1} N \frac{s}{\sqrt{n}} \sqrt{\frac{N-1}{N-1}} \quad (4.9)$$

If you examine this closely, it is almost identical to the formula used for the confidence interval for a mean with an unknown population standard deviation and a finite

A	B
1	Unpaid Reimbursement Claims
2	
3	Data
4	Population Size                    180
5	Sample Mean                      185
6	Sample Size                        20
7	Sample Standard Deviation      22
8	Confidence Level                 95%
9	
10	Intermediate Calculations
11	Population Total                33300.00
12	FPC Factor                    0.945438918
13	Standard Error of the Total    837.1700134
14	Degrees of Freedom            19
15	t Value                        2.09302405
16	Interval Half Width            1752.22
17	
18	Confidence Interval
19	Interval Lower Limit          31547.78
20	Interval Upper Limit          35052.22

**FIGURE 4.17** Confidence Interval for Population Total

population correction factor, except that both the point estimate and the interval half-width are multiplied by the population size  $N$  to scale the result to a total, rather than an average.

## USING CONFIDENCE INTERVALS FOR DECISION MAKING

Confidence intervals can be used in many ways to support business decisions. For example, in packaging some commodity product such as laundry soap, the manufacturer must ensure that the packages contain the stated amount to meet government regulations. However, variation may occur in the filling equipment. Suppose that the required weight for one product is 64 ounces. A sample of 30 boxes is measured and the sample mean is calculated to be 63.82 with a standard deviation of 1.05. Does this indicate that the equipment is underfilling the boxes? Not necessarily. A 95% confidence interval for the mean is [63.43, 64.21]. Although the sample mean is less than 64, the sample does not provide sufficient evidence to draw that conclusion that the population mean is less than 64 because 64 is contained within the confidence interval. In fact, it is just as plausible that the population mean is 64.1 or 64.2. However, suppose that the sample standard deviation was only 0.46. The confidence interval for the mean would be [63.65, 63.99]. In this case, we would conclude that it is highly unlikely that the population mean is 64 ounces because the confidence interval falls completely below 64; the manufacturer should check and adjust the equipment to meet the standard.

As another example, suppose that an exit poll of 1,300 voters found that 692 voted for a particular candidate in a two-person race. This represents a proportion of 53.23% of the sample. Could we conclude that the candidate will likely win the election? A 95% confidence interval for the proportion is [0.505, 0.559]. This suggests that the population proportion of voters who favor this candidate will be larger than 50%, so it is safe to predict the winner. On the other hand, suppose that only 670 of the 1,300 voters voted for the candidate, indicating a sample proportion of 0.515. The confidence interval for the population proportion is [0.488, 0.543]. Even though the sample proportion is larger than 50%, the sampling error is large, and the confidence interval suggests that it is reasonably likely that the true population proportion will be less than 50%, so it would not be wise to predict the winner based on this information.

We also point out that confidence intervals are most appropriate for cross-sectional data. For time-series data, confidence intervals often make little sense because the mean and/or variance of such data typically change over time. However, for the case in which time-series data are *stationary*—that is, they exhibit a constant mean and constant variance—then confidence intervals can make sense. A simple way of determining whether time-series data are stationary is to plot them on a line chart. If the data do not show any trends or patterns and the variation remains relatively constant over time, then it is reasonable to assume the data are stationary. However, you should be cautious when attempting to develop confidence intervals for time-series data because high correlation between successive observations (called autocorrelation) can result in misleading confidence interval results.

## CONFIDENCE INTERVALS AND SAMPLE SIZE

An important question in sampling is the size of the sample to take. Note that in all the formulas for confidence intervals, the sample size plays a critical role in determining the width of the confidence interval. As the sample size increases, the width of the confidence interval decreases, providing a more accurate estimate of the true population parameter. In many applications, we would like to control the margin of error in a confidence interval. For example, in reporting voter preferences, we might wish to ensure that the margin of error is  $\pm 2\%$ . Fortunately, it is relatively easy to determine the



#### Spreadsheet Note

appropriate sample size needed to estimate the population parameter within a specified level of precision. *PHStat* provides tools for computing sample sizes for estimating means and proportions (see Appendix 4.2G, *Determining Sample Size*).

For example, when discussing the use of confidence intervals for decision making, we presented an example of filling boxes of soap, which resulted in a confidence interval for the mean of [63.43, 64.21]. The width of the confidence interval is  $\pm 0.39$ , which represents the sampling error. Suppose the manufacturer would like the sampling error to be at most 0.15, resulting in a confidence interval width of 0.30. Using the *PHStat* tool for sample size determination for the mean with an estimated population standard deviation of 1.05 obtained from the original sample of 30, we find that 189 samples would be needed (see Figure 4.18). To verify this, Figure 4.19 shows that if a sample of 189 resulted in the same sample mean and standard deviation, the sampling error for the confidence interval [63.67, 63.97] is indeed  $\pm 0.15$ .

Of course, we generally do not know the population standard deviation prior to finding the sample size. A commonsense approach would be to take an initial sample to estimate the population standard deviation using the sample standard deviation  $s$  and determine the required sample size, collecting additional data if needed. If the half-width of the resulting confidence interval is within the required margin of error, then

A	B
1	Sample Size Determination
2	
3 Data	
4 Population Standard Deviation	1.05
5 Sampling Error	0.15
6 Confidence Level	95%
7	
8 Intermediate Calculations	
9 Z Value	-1.95996398
10 Calculated Sample Size	188.2314822
11	
12 Result	
13 Sample Size Needed	189

**FIGURE 4.18** *PHStat* Results for Sample Size Determination for the Mean

A	B
1	Confidence Interval Estimate for the Mean
2	
3 Data	
4 Sample Standard Deviation	1.05
5 Sample Mean	63.82
6 Sample Size	189
7 Confidence Level	95%
8	
9 Intermediate Calculations	
10 Standard Error of the Mean	0.076376262
11 Degrees of Freedom	188
12 t Value	1.972662649
13 Interval Half Width	0.150664599
14	
15 Confidence Interval	
16 Interval Lower Limit	63.67
17 Interval Upper Limit	63.97

**FIGURE 4.19** Results for Confidence Interval for the Mean Using Sample Size = 189

A	B
1	Sample Size for the Proportion
2	
3 Data	
4 Estimate of True Proportion	0.5
5 Sampling Error	0.02
6 Confidence Level	95%
7	
8 Intermediate Calculations	
9 Z Value	-1.95996398
10 Calculated Sample Size	2400.911763
11	
12 Result	
13 Sample Size Needed	2401

**FIGURE 4.20** *PHStat Results for Sample Size Determination for the Proportion*

we clearly have achieved our goal. If not, we can use the new sample standard deviation  $s$  to determine a new sample size and collect additional data as needed. Note that if  $s$  changes significantly, we still might not have achieved the desired precision and might have to repeat the process. Usually, however, this will be unnecessary.

For the voting example in the previous section, suppose that we wish to determine the number of voters to poll to ensure a sampling error of at most  $\pm 2\%$ . The *PHStat* tool requires an estimate of the true proportion. In practice, this value will not be known. You could use the sample proportion from a preliminary sample as an estimate to plan the sample size, but this might require several iterations and additional samples to find the sample size that yields the required precision. When no information is available, the most conservative approach is to use 0.5 for the estimate of the true proportion. This will result in the sample size that will guarantee the required precision no matter what the true proportion is. However, if we do have a good estimate for the population proportion, then we should use it because it will result in smaller required sample sizes that will usually be less expensive to obtain. For example, using 0.5 as an estimate of the true proportion, to estimate the number of voters to poll to obtain a 95% confidence interval on the proportion of voters that choose a particular candidate with a precision of  $\pm 0.02$  or less, we would need a sample of 2,401 voters (see Figure 4.20).

## PREDICTION INTERVALS

Another type of interval used in estimation is a prediction interval. A **prediction interval** is one that provides a range for predicting the value of a new observation from the same population. This is different from a confidence interval, which provides an interval estimate of a population parameter, such as the mean or proportion. A confidence interval is associated with the *sampling distribution* of a statistic, but a prediction interval is associated with the distribution of the random variable itself.

When the population standard deviation is unknown, a  $100(1 - \alpha)\%$  prediction interval for a new observation is

$$\bar{x} \pm t_{\alpha/2, n-1}(s\sqrt{1 + 1/n}) \quad (4.10)$$

Note that this interval is wider than the confidence interval in equation (4.6) by virtue of the additional value of 1 under the square root. This is because, in addition to estimating the population mean, we must also account for the variability of the new observation

around the mean. Using the example of estimating the revolving balance in the Excel file *Credit Approval Decisions* in Figure 4.11, we may state a 95% prediction interval for the revolving balance of a new homeowner as

$$\$12,630.37 \pm 2.055(\$5,393.38) \sqrt{1+1/27} \text{ or } [\$1,343.59, \$23,917.15]$$

## ADDITIONAL TYPES OF CONFIDENCE INTERVALS

Most confidence intervals have the same basic form: a point estimate of the population parameter of interest plus or minus some number of standard errors. Thus, establishing confidence intervals requires choosing the proper point estimate for a population parameter as well as an understanding of the sampling distribution of the parameter being estimated and, in particular, the standard error. In this section we summarize several additional types of confidence intervals. However, spreadsheet-based tools for computing them are not available in *PHStat* and you must resort to calculating them using the formulas, some of which are rather complex. Appendix 4.1 discusses the formulas and theory behind them. Optional skill-building exercises will ask you to create Excel templates for calculating these confidence intervals.

### Differences Between Means, Independent Samples

In many situations, we are interested in differences between two populations. For example, in the *Accounting Professionals* Excel file, we might be interested in the difference in mean years of service between females and males. Similarly, in the *Burglaries* Excel file, we might be interested in the difference between the mean number of burglaries per month before and after the citizen-police program was instituted. In both these examples, samples are drawn independently from the two populations.

To illustrate the application of this type of confidence interval, the means and standard deviations for both male and female employees in the *Accounting Professionals* Excel file were computed. The mean years of service for females is 10.07, and the mean years of service for males is 19.69. Thus, a point estimate for the difference in years of service is  $10.07 - 19.69 = -9.62$ , indicating that, on average, males have been working at the company over 9 years longer than females. Using the confidence interval formula described in Appendix 4.1, we find a 95% confidence interval for the mean difference in years of service between females and males for the population of accounting workers in the company is  $[-15.118, -4.122]$ . Because the entire confidence interval lies below zero, this suggests that the male workers have more experience than females.

### Differences Between Means, Paired Samples

A second situation involves *paired samples*. For example, a deep-foundation engineering contractor has bid on a foundation system for a new world headquarters building for a Fortune 500 company. A part of the project consists of installing 311 auger cast piles. The contractor was given bid information for cost estimating purposes, which consisted of the estimated depth of each pile; however, actual drill footage of each pile could not be determined exactly until construction was performed. The Excel file *Pile Foundation* contains the estimates and actual pile lengths after the project was completed. We might be interested in the difference between the means of the actual and estimated pile lengths. From the sample data, the mean difference is found to be 6.38, indicating that, on average, the actual lengths were underestimated. Using the formula developed in Appendix 4.1, a 95% confidence interval for the mean difference is  $[5.234, 7.526]$ . This states that the true population difference does not appear to be zero, indicating a bias in estimating the pile depth.

## Differences Between Proportions

A final type of confidence interval that has useful applications is the difference between proportions. For example in the *Accounting Professionals* data, we see that the proportion of females who have Certified Public Accountant (CPA) credentials is  $8/14 = 0.57$ , while the proportion of males having a CPA is  $6/13 = 0.46$ . While this sample data suggests that a higher proportion of females have a CPA, a 95% confidence interval for the difference in proportions between females and males, using the formula provided in Appendix 4.2, is  $[-0.2650, 0.4850]$ . This suggests that we cannot conclusively state that the proportion of females having a CPA is higher than males, because a difference of zero falls within the confidence interval.

## Basic Concepts Review Questions

1. Explain the importance of sampling from a managerial perspective.
2. What is a sampling plan and what elements should be included in one?
3. How does a frame differ from a target population?
4. Describe the difference between subjective and probabilistic sampling methods. What are the advantages and disadvantages of each?
5. Explain how the following sampling approaches work:
  - a. Simple random sampling
  - b. Systematic sampling
  - c. Stratified sampling
  - d. Cluster sampling
  - e. Sampling from a continuous process
6. What is the difference between nonsampling error and sampling error? Why might each type of error occur?
7. Define a random number. How is it different from a random variate?
8. Explain the sampling distribution of the mean. What properties does it have?
9. What is the standard error of the mean? How does it relate to the standard deviation of the population from which a sample is taken?
10. Why is it always reasonable to use the normal distribution to solve probability questions regarding a sample mean

based on a sample from any arbitrary distribution when the sample size is large? Which famous theorem justifies the use of the normal distribution in the above situation?

11. When is an estimator called an unbiased estimator?
12. What do we mean by an unbiased estimator? Why is this important?
13. What is a confidence interval? How do you properly interpret the level of confidence,  $1 - \alpha$ ?
14. How does the *t*-distribution differ from the standard normal distribution?
15. When is it important to apply the finite population correction factor to the standard error when developing confidence intervals?
16. Summarize the different types of confidence intervals that one may construct, and provide a practical application for each.
17. Discuss how confidence intervals can help in making decisions. Provide some examples different from those in the chapter.
18. Under what circumstances can confidence intervals be applied to time-series data?
19. Explain how a confidence interval changes with changes in the level of confidence and sample size.
20. What is a prediction interval and how does it differ from a confidence interval?

## Problems and Applications

1. Data have to be collected to measure the effectiveness of an advertising campaign in a particular morning newspaper in an American city. For this situation, describe how one can collect samples using different sampling plans such as simple random sampling, stratified sampling, and cluster sampling.
2. Compute the mean and standard deviation of the runs scored by the 30 teams in major league baseball in the year 2010 (data given in the Excel file *Major League Baseball*). Now draw a random sample of six teams using an appropriate random number generator in Excel. Compute the mean and standard deviation of the runs scored by the selected teams. How does it compare with the population quantities?

3. A bakery has the following probability distribution for the daily demand for its cakes.

Demand, $x$	Probability, $f(x)$
0	0.04
1	0.16
2	0.18
3	0.28
4	0.34
5 or more	0.00

Using the first column of random numbers in Table 4.1, generate 20 samples from this distribution and construct a histogram of the results.

4. Suppose that we conduct an experiment in which samples of size  $n$  are generated from a normal distribution having a known standard deviation  $\sigma$ . If we compute the range of each sample, we can estimate the distribution of the statistic  $R/\sigma$ . The expected value of this statistic is a factor that statisticians have labeled as  $d_2$ . If we know this value and a sample range, then we can estimate  $\sigma$  by  $R/d_2$ . The values of  $d_2$  are shown below for sample sizes from 2 through 5.

$n$	$d_2$
2	1.128
3	1.693
4	2.059
5	2.326

Develop a sampling experiment on a spreadsheet to estimate these values of  $d_2$  by generating 1,000 samples of  $n$  random variates from a normal distribution with a mean of 0 and standard deviation of 3 (using the Excel function NORMINV). For each of the 1,000 samples, compute the range, and the value of  $R/\sigma$ . Use the average value of  $R/\sigma$  to estimate  $d_2$  for sample sizes  $n = 2$  through 5. Compare your results to published factors shown above. How well did your experiment perform?

5. A mechanical system needs 10 identical machine parts, and functions as long as all 10 parts are functioning. The lifetime of each of these machine parts has a normal distribution with a mean of 400 hours and a standard deviation of 12 hours. Devise and implement a sampling experiment for estimating the distribution of the minimum lifetime of these 10 machine parts and its expected value.
6. The speed of cars on a certain motorway has a normal distribution with mean 90 km/hr and a standard deviation of 12 km/hr. A radar unit is used to measure speeds of cars on this motorway. On a particular day, 9 cars are picked up at random on this motorway and their speeds are measured.
- a. What is the distribution of the mean speed of the cars in this sample?
  - b. What is the probability that the sample mean will be greater than 96 km/hr?
  - c. What is the probability that the sample mean will be less than 88 km/hr?
7. A candy maker produces mints that have a label weight of 20 grams. Because of variations in the manufacturing process, the candies have a mean weight of 20 grams and a standard deviation of 0.25 grams, normally distributed.
- a. If a customer samples 25 mints, what is the probability that the mean weight is less than 19.88 grams?

- b. What will be the value which the mean weight of the 25 candies will exceed only 10% of the time?
8. Using the data in the Excel file *Atlanta Airline Data*, find and interpret 95% confidence intervals for the following:
- a. Mean difference between scheduled time of arrival and actual arrival time.
  - b. Proportion of flights that arrive early.
9. Using the data in the Excel file *MBA Student Survey*, find 95% confidence intervals for the mean number of nights out per week and mean number of study hours per week by gender. Based on the confidence intervals, would you conclude that there is a difference in social and study habits between males and females?
10. Using the data in the worksheet *Consumer Transportation Survey*, develop 95% and 99% confidence intervals for the following:
- a. The mean hours per week that individuals spend in their vehicles
  - b. The average number of miles driven per week
  - c. The proportion of individuals who are satisfied with their vehicle
  - d. The proportion of individuals who have at least one child
- Explain the differences as the level of confidence increases.
11. The Excel file *Restaurant Sales* provides sample information on lunch, dinner, and delivery sales for a local Italian restaurant. Develop 95% confidence intervals for the mean of each of these variables, as well as for weekday and weekend sales. What conclusions can you reach?
12. The Excel file *Golfing Statistics* provides data on the driving accuracy of 25 contestants. Develop a 90% confidence interval for driving accuracy.
13. Average daily hotel room rates in two different American cities are given as follows:

	Average Room Rate	Standard Deviation	Sample Size
City I	\$112	\$9	10
City II	\$122	\$12	12

- If we cannot assume the variances to be equal, construct a 99% confidence interval for the two differences of the two means.
14. If, based on a sample of size 300, a political candidate found that 165 people would vote for him in a two-person contest, can she be reasonably certain of victory? Find a 99% confidence interval and explain.
15. If based on a sample of size 400, a political candidate found that 240 people would vote for him in a two-person race, what is the 90% confidence interval for his proportion of votes? Can he be reasonably certain of victory? Find a 90% confidence interval and explain.
16. The Excel file *Blood Pressure* shows diastolic blood pressure readings before and after a new medication. Find 95% confidence intervals for the variance for each

of these groups. Based on these confidence intervals, would you conclude that the medication has kept the reading more stable?

17. Using data in the Excel file *Colleges and Universities*, find 95% confidence intervals for the standard deviation of the median SAT for each of the two groups: liberal arts colleges and research universities. Based on these confidence intervals, does there appear to be a difference in the variation of the median SAT scores between the two groups?
18. In a sample of 64 different restaurants belonging to a fast food chain, the average monthly electricity bill is \$2000 with the standard deviation of \$300. Find the 99% confidence interval for the mean electricity bill of the population of different restaurants of this fast food chain.
19. The Excel file *New Account Processing* provides data for a sample of employees in a company. Assume that the company has 125 people in total assigned to new account processing. Find a 95% confidence interval for the total sales of the population of account representatives.
20. It is known that the standard deviation of the lifetime of light bulbs manufactured by a company is 30 hours. How large should the sample size be so that we are 90% confident that the estimated mean lifetime of these light bulbs is no more than  $\pm 3$  hours away from the true lifetime?
21. A researcher wants to estimate the proportion of secretaries of Fortune 500 companies who have personal computers at their workstations. How large should the sample be if the researcher is to be 95% confident that the difference between the sample estimate and the actual proportion is no more than  $\pm 2\%$ ?
22. The Excel file *Baseball Attendance* shows the attendance in thousands at San Francisco Giants' baseball games for the 10 years before the Oakland A's moved to the Bay Area in 1968, as well as the combined attendance for both teams for the next 11 years.
  - a. Do the data appear to be stationary?
  - b. Develop 95% confidence intervals for the mean attendance of each of the two groups. Based on these confidence intervals, would you conclude that attendance has changed after the move?
23. The state of Ohio Department of Education has a mandated ninth-grade proficiency test that covers writing, reading, mathematics, citizenship (social studies), and science. The Excel file *Ohio Education Performance* provides data on success rates (defined as the percentage of students passing) in school districts in the greater Cincinnati metropolitan area along with state averages. Find 50% and 90% probability intervals centered on the median for each of the variables in the data.
24. Using the data in the worksheet *Consumer Transportation Survey*, develop 95% and 99% prediction intervals for the following:
  - a. the hours per week that an individual will spend in his or her vehicle

- b. the number of miles driven per week

Compare these to the confidence intervals developed in Problem 10.

25. The Excel file *Restaurant Sales* provides sample information on lunch, dinner, and delivery sales for a local Italian restaurant. Develop 95% prediction intervals for the dollar sales of each of these variables for next Saturday.
26. For the Excel file *Burglaries*, find 95% confidence and prediction intervals for the number of burglaries before and after the Citizen Police meetings. How do they compare?

*Note: The following problems require material about Additional Confidence Intervals found in Appendix 4.1.*

27. An experiment was conducted to compare people's reaction times to a red light versus a green light. When signaled with either the red or the green light, the subject was asked to hit a switch to turn off the light. When the switch was hit, a clock was turned off and the reaction time in seconds was recorded. The following result gives the reaction times for eight subjects.

Subject	Red	Green
1	0.44	0.45
2	0.23	0.32
3	0.29	0.56
4	0.51	0.47
5	0.24	0.33
6	0.39	0.41
7	0.42	0.45
8	0.47	0.47

Find a 90% confidence interval for the mean difference of the reactions in the populations.

28. Suppose that scores on a standardized test in mathematics taken by students from large and small high schools are  $N(\mu_X, \sigma^2)$  and  $N(\mu_Y, \sigma^2)$  respectively, where  $\sigma^2$  is unknown. A random sample of  $n = 16$  students from large high schools yielded an average and standard deviation of 81.31 and 7.8 respectively. A random sample of  $n = 16$  students from small high schools yielded an average of 78.61 and sample standard deviation of 7.0. Find a 95% confidence interval for the difference of the population means.
29. The Excel file *Mortgage Rates* contains time-series data on rates of three different mortgage instruments. Assuming that the data are stationary, construct a 95% confidence interval for the mean difference between the 30-year and 15-year fixed rate mortgage rates. Based on this confidence interval, would you conclude that there is a difference in the mean rates?
30. The Excel file *Student Grades* contains data on midterm and final exam grades in one section of a large statistics course. Construct a 95% confidence interval for the mean difference in grades between the midterm and final exams.

## Case

### Analyzing a Customer Survey

A supplier of industrial equipment has conducted a survey of customers for one of its products in its principal marketing regions: North America, South America, Europe, and in its emerging market in China. The data, which tabulate the responses on a scale from 1 to 5 on dimensions of quality, ease of use, price, and service, are in the Excel file *Customer Survey*. Use point and interval

estimates, as well as other data analysis tools such as charts, PivotTables, and descriptive statistics, to analyze these data and write a report to the marketing vice president. Specifically, you should address differences among regions and proportions of “top box” survey responses (which is defined as scale levels 4 and 5) for each of the product dimensions.

## APPENDIX 4.1

### Theoretical Foundations of Confidence Intervals

In this appendix, we will present the theory behind confidence intervals and also introduce several additional types of confidence intervals that do not have spreadsheet support tools.

#### A. Theory Underlying Confidence Intervals

Recall that the scale of the standard normal distribution is measured in units of standard deviations. We will define  $z_\alpha$  to represent the value from the standard normal distribution that provides an upper tail probability of  $\alpha$ . That is, the area to the right of  $z_\alpha$  is equal to  $\alpha$ . Some common values that we will use often include  $z_{0.025} = 1.96$  and  $z_{0.05} = 1.645$ . You should check the standard normal table in Table A.1 in the appendix at the end of the book to verify where these numbers come from.

We stated that the sample mean,  $\bar{x}$ , is a point estimate for the population mean  $\mu$ . We can use the central limit theorem to quantify the sampling error in  $\bar{x}$ . Recall that the central limit theorem states that no matter what the underlying population, the distribution of sample means is approximately normal with mean  $\mu$  and standard deviation (standard error)  $\sigma_{\bar{x}} = \sigma / \sqrt{n}$ . Therefore, based on our knowledge about the normal distribution, we can expect approximately 95% of sample means to fall within  $\pm 2$  standard errors of  $\mu$ , and more than 99% of them to fall within  $\pm 3$  standard errors of  $\mu$ . More specifically,  $100(1 - \alpha)\%$  of sample means will fall within  $\pm z_{\alpha/2}\sigma_{\bar{x}}$  of the population mean  $\mu$  as illustrated in Figure 4A1.1.

However, we do not know  $\mu$  but estimate it by  $\bar{x}$ . Suppose that we construct an interval around  $\bar{x}$  by adding

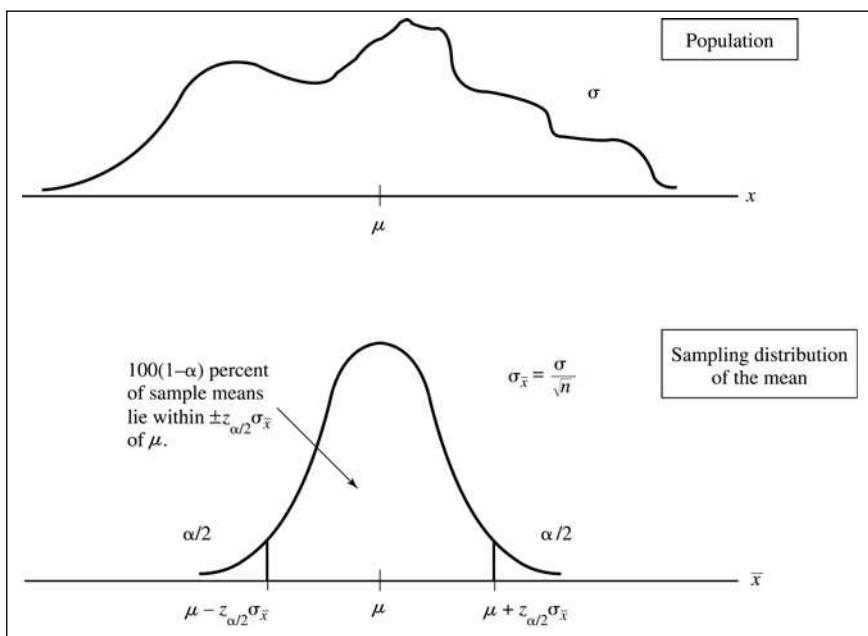
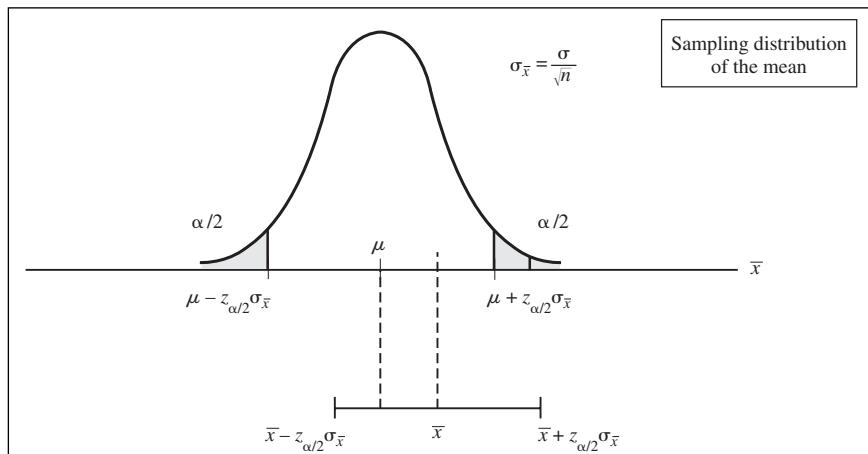
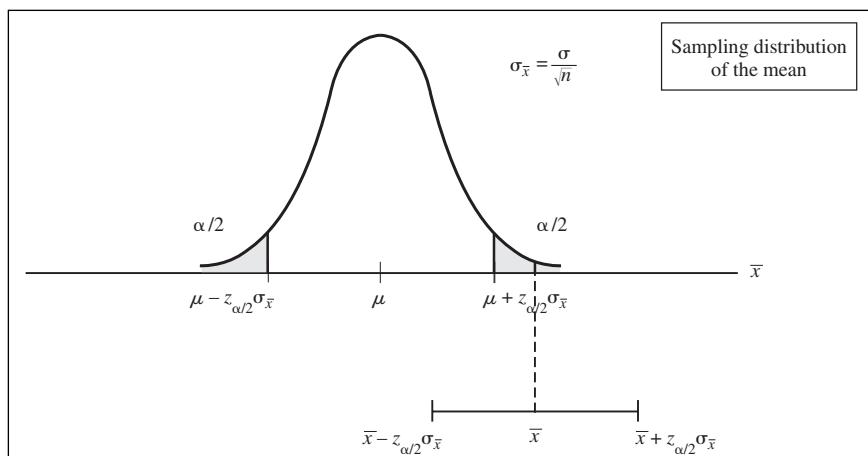


FIGURE 4A1.1 Sampling Distribution of the Mean



**FIGURE 4A1.2** An Interval Estimate that Contains the True Population Mean



**FIGURE 4A1.3** An Interval Estimate that Does Not Contain the True Population Mean

and subtracting  $z_{\alpha/2}\sigma_{\bar{x}}$ . If  $\bar{x}$  lies within  $\pm z_{\alpha/2}\sigma_{\bar{x}}$  of the true mean  $\mu$  as shown in Figure 4A1.2, then you can see that this interval will contain the population mean  $\mu$ . On the other hand, if  $\bar{x}$  lies farther away than  $z_{\alpha/2}\sigma_{\bar{x}}$  from the true mean  $\mu$  (in one of the shaded regions in Figure 4A1.3), then we see that the interval estimate does *not* contain the true population mean. Because  $100(1 - \alpha)\%$  of sample means will fall within  $z_{\alpha/2}\sigma_{\bar{x}}$  of the population mean  $\mu$ , we can see that precisely  $100(1 - \alpha)\%$  of the intervals we construct in this fashion will contain  $\mu$ . Therefore, a  $100(1 - \alpha)\%$  confidence interval for the population mean  $\mu$  is given by:

$$\bar{x} \pm z_{\alpha/2}(\sigma / \sqrt{n}) \quad (4A.1)$$

This confidence interval will contain the true population mean  $100(1 - \alpha)\%$  of the time.

distribution to construct the confidence intervals. To establish a confidence interval for a proportion, we need to know the sampling distribution of the proportion and its standard error. The **sampling distribution of the proportion** is analogous to the sampling distribution of the mean, and is the probability distribution of all possible values of  $\hat{p}$ . If we are sampling with replacement from a finite population, the sampling distribution of  $\hat{p}$  follows the binomial distribution with mean  $n\pi$  and variance  $n\pi(1 - \pi)$ . It follows that the sampling distribution of  $\hat{p} = x/n$  has mean  $n\pi/n = \pi$  and variance  $n\pi(1 - \pi)/n^2 = \pi(1 - \pi)/n$ . Thus, the standard error of the proportion is  $\sqrt{\pi(1-\pi)/n}$ . When  $n\pi$  and  $n(1 - \pi)$  are at least 5, the sampling distribution of  $\hat{p}$  approaches the normal distribution as a consequence of the central limit theorem. Therefore, under these conditions, we may use  $z$ -values to determine the range of sampling error for a specified confidence level. Because we generally use  $\hat{p}$  as an estimate for  $\pi$ , the confidence interval becomes:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad (4A.2)$$

## B. Sampling Distribution of the Proportion

For confidence intervals for the mean, we used the standard error based on the central limit theorem and the normal

## C. Sample Size Determination

The formulas for determining sample sizes to achieve a given margin of error are based on the confidence interval half-widths. For example, consider the confidence interval for the mean with a known population standard deviation:

$$\bar{x} \pm z_{\alpha/2} \left( \sigma / \sqrt{n} \right) \quad (4A.3)$$

Suppose we want the width of the confidence interval on either side of the mean to be at most  $E$ . In other words,

$$E \geq z_{\alpha/2} \left( \sigma / \sqrt{n} \right)$$

Solving for  $n$ , we find:

$$n \geq (z_{\alpha/2})^2 (\sigma^2) / E^2 \quad (4A.4)$$

In a similar fashion, we can compute the sample size required to achieve a desired confidence interval half-width for a proportion by solving the following equation for  $n$ :

$$E \geq z_{\alpha/2} \sqrt{\frac{\pi(1 - \pi)}{n}}$$

This yields:

$$n \geq \frac{(z_{\alpha/2})^2 \pi(1 - \pi)}{E^2} \quad (4A.5)$$

In practice, the value of  $\pi$  will not be known. You could use the sample proportion from a preliminary sample as an estimate of  $\pi$  to plan the sample size, but this might require several iterations and additional samples to find the sample size that yields the required precision. When no information is available, the most conservative estimate is to set  $\pi = 0.5$ . This maximizes the quantity  $\pi(1 - \pi)$  in the formula, resulting in the sample size that will guarantee the required precision no matter what the true proportion is.

## D. Additional Confidence Intervals

In this section, we develop the formulas for additional types of confidence intervals for which no *PHStat* tool is available.

### Difference in Means

An important type of confidence interval is one for estimating the difference in means of two populations. The method of constructing confidence intervals differs depending on whether the samples are independent or paired, and whether

the variances of the populations can be assumed to be equal or not.

We will assume that we have random samples from two populations with the following:

	Population 1	Population 2
Mean	$\mu_1$	$\mu_2$
Standard deviation	$\sigma_1$	$\sigma_2$
Point estimate	$\bar{x}_1$	$\bar{x}_2$
Sample size	$n_1$	$n_2$

A point estimate for the difference in means,  $\mu_1 - \mu_2$ , is given by  $\bar{x}_1 - \bar{x}_2$ . We consider two different cases: when the variances of the two populations are unequal, and when they can be assumed to be equal.

### Independent Samples with Unequal Variances

A confidence interval for independent samples with unequal variances is:

$$\bar{x}_1 - \bar{x}_2 \pm (t_{\alpha/2, df^*}) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (4A.6)$$

where the degrees of freedom for the *t*-distribution,  $df^*$ , are computed as:

$$df^* = \frac{\left[ \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2}{\left[ \frac{(s_1^2/n_1)^2}{n_1 - 1} \right] + \left[ \frac{(s_2^2/n_2)^2}{n_2 - 1} \right]} \quad (4A.7)$$

and fractional values are rounded down. This calculation may be eliminated by using a conservative estimate of the number of degrees of freedom as the minimum of  $n_1$  and  $n_2$ , which results in a larger confidence interval.

To illustrate, we use the *Accounting Professionals* data. Sorting the data by gender, the sample means and standard deviations were calculated. We found  $s_1 = 4.39$  and  $n_1 = 14$  (females), and  $s_2 = 8.39$  and  $n_2 = 13$  (males). Calculating  $df^*$ , we obtain  $df^* = 17.81$ , so use 17 as the degrees of freedom. A 95% confidence interval for the difference in years of service is:

$$10.07 - 19.69 \pm 2.1098 \sqrt{\frac{19.2721}{14} + \frac{70.3921}{13}} \\ = -9.62 \pm 5.498 \\ \text{or } [-15.118, -4.122]$$

## SKILL-BUILDER EXERCISE 4.6

Develop an Excel template that will calculate a confidence interval for the difference in means for independent samples with unequal variances.

## Independent Samples with Equal Variances

When we can assume that the variance of the two populations are equal, we can estimate a common ("pooled") standard deviation that is a weighted combination of the individual sample standard deviations,  $s_p$ :

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (4A.8)$$

Then the sampling distribution of  $\bar{x}_1 - \bar{x}_2$  has a  $t$ -distribution with  $n_1 + n_2 - 2$  degrees of freedom and standard error:

$$s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (4A.9)$$

Therefore, a  $100(1 - \alpha)\%$  confidence interval is:

$$\bar{x}_1 - \bar{x}_2 \pm \left(t_{\alpha/2, n_1 + n_2 - 2}\right) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (4A.10)$$

If we assume equal population variances for the *Accounting Professionals* data, the pooled standard deviation is:

$$s_p = \sqrt{\frac{(14 - 1)(4.39)^2 + (13 - 1)(8.39)^2}{14 + 13 - 2}} = 6.62$$

Then, a 95% confidence interval for the difference in mean years of service between females and males is:

$$10.07 - 19.69 \pm (2.0595)6.62 \sqrt{\frac{1}{14} + \frac{1}{13}} \\ = -9.62 \pm 5.25 \text{ or } [-14.87, -4.37]$$

Note that there is little difference in the confidence interval from the unequal variance case for this example. In general, assume equal population variances unless you have evidence that the variances are significantly different.

## SKILL-BUILDER EXERCISE 4.7

Develop an Excel template that will calculate a confidence interval for the difference in means with equal variances.

## Paired Samples

For paired samples, we first compute the difference between each pair of observations,  $D_i$ , for  $i = 1, \dots, n$ . If we average these differences, we obtain  $\bar{D}$ , a point estimate for the mean difference between the populations. The standard deviation of the differences is similar to calculating an ordinary standard deviation:

$$s_D = \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n - 1}} \quad (4A.11)$$

A  $100(1 - \alpha)\%$  confidence interval is:

$$\bar{D} \pm \left(t_{n-1, \alpha/2}\right) s_D / \sqrt{n} \quad (4A.12)$$

For the *Pile Foundation* data described in the main text, we compute the difference for each pile by subtracting the estimated value from the actual value, as shown in Figure 4A1.4, then calculated  $\bar{D} = 6.38$  and  $s_D = 10.31$ . Note that because the sample size exceeds the

Pile Foundation Data				
	A	B	C	D
1				
2				
3	Pile Number	Estimated Pile Length (ft.)	Actual Pile Length (ft.)	Actual - Estimated
4				
5	1	10.58	18.58	8.00
6	2	10.58	18.58	8.00
7	3	10.58	18.58	8.00
8	4	10.58	18.58	8.00
9	5	10.58	28.58	18.00
10	6	10.58	26.58	16.00
11	7	10.58	17.58	7.00
12	8	10.58	27.58	17.00
13	9	10.58	27.58	17.00
14	10	10.58	37.58	27.00

FIGURE 4A1.4 Difference Calculations for Portion of *Pile Foundation Data*

largest degrees of freedom listed in the table, we must use the critical value of  $t$  with an infinite number of degrees of freedom in Table A.2 in the appendix at the end of the book. For  $\alpha/2 = 0.025$ , this value is 1.96,

which is the same as the  $z$ -value. Thus, a 95% confidence interval is:

$$6.38 \pm 1.96(10.31/\sqrt{311}) = 6.38 \pm 1.146 \text{ or } (5.234, 7.526)$$

## SKILL-BUILDER EXERCISE 4.8

Develop an Excel template that will calculate a confidence interval for the difference in means for paired samples.

### Differences between Proportions

Let  $\hat{p}_1$  and  $\hat{p}_2$  be sample proportions from two populations using sample sizes  $n_1$  and  $n_2$ , respectively. For reasonably large sample sizes, that is, when  $n_i\hat{p}_i$  and  $n_i(1 - \hat{p}_i)$  are greater than 5 for  $i = 1, 2$ , the distribution of the statistic  $\hat{p}_1 - \hat{p}_2$  is approximately normal. A confidence interval for differences between proportions of two populations is computed as follows:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \quad (4A.13)$$

For example, in the *Accounting Professionals* data, the proportion of females having a CPA is  $8/14 = 0.57$ , while the proportion of males having a CPA is  $6/13 = 0.46$ . A 95% confidence interval for the difference in proportions between females and males is:

$$0.57 - 0.46 \pm 1.96 \sqrt{\frac{0.57(1 - 0.57)}{14} + \frac{0.46(1 - 0.46)}{13}} \\ = 0.11 \pm 0.3750 \text{ or } [-0.2650, 0.4850]$$

## SKILL-BUILDER EXERCISE 4.9

Develop an Excel template that will calculate a confidence interval for the difference in proportions.

Table 4A.1 provides a complete summary of all confidence interval formulas we have discussed.

**TABLE 4A.1 Summary of Confidence Interval Formulas**

Type of Confidence Interval	Formula
Mean, standard deviation known	$\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$
Mean, standard deviation unknown	$\bar{x} \pm t_{\alpha/2, n-1}(s/\sqrt{n})$
Proportion	$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$
Population total	$N\bar{x} \pm t_{\alpha/2, n-1}N \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$
Difference between means, independent samples, equal variances	$\bar{x}_1 - \bar{x}_2 \pm (t_{\alpha/2, n_1+n_2-2})s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
	$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$
Difference between means, independent samples, unequal variances	$\bar{x}_1 - \bar{x}_2 \pm (t_{\alpha/2, df^*}) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

(Continued)

**TABLE 4A.1 (Continued)**

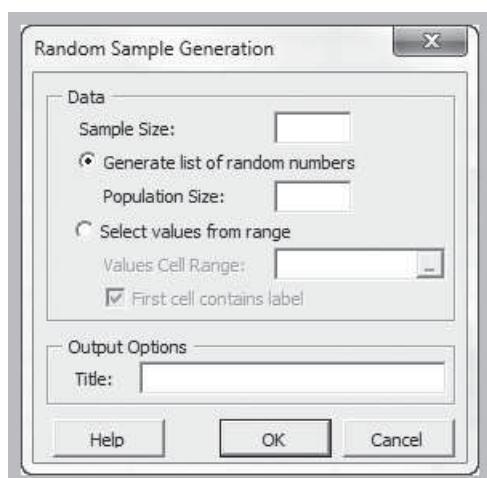
Type of Confidence Interval	Formula
	$df^* = \frac{\left[ \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2}{\left[ \frac{(s_1^2/n_1)^2}{n_1 - 1} \right] + \left[ \frac{(s_2^2/n_2)^2}{n_2 - 1} \right]}$
Difference between means, paired samples	$\bar{D} \pm (t_{n-1, \alpha/2})s_D / \sqrt{n}$
Differences between proportions	$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} + \sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
Variance	$\left[ \frac{(n-1)s^2}{\chi^2_{n-1, \alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1, 1-\alpha/2}} \right]$

## APPENDIX 4.2

### Excel and PHStat Notes

#### A. Excel-Based Random Sampling Tools

PHStat provides a tool that can be used to generate a random list of integers between one and a specified population size or to randomly select values from a range of data on a worksheet without replacement. From the PHStat menu, select *Sampling*, then *Random Sample Generation*. Figure 4A2.1 shows the dialog box that appears. Enter the sample size

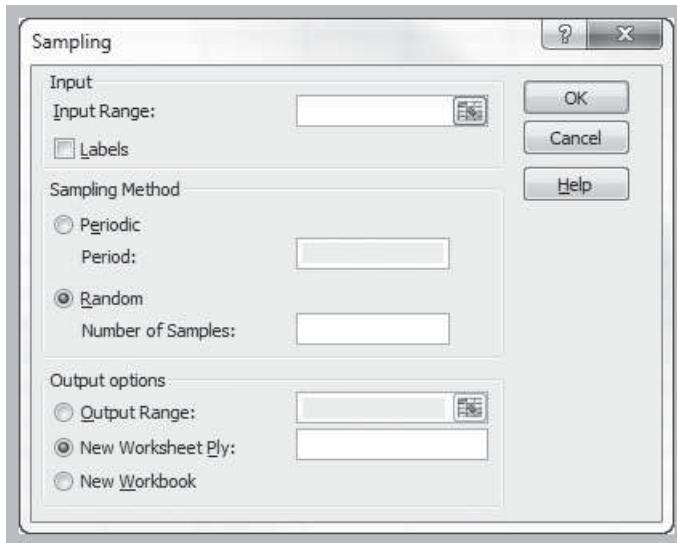


**FIGURE 4A2.1** PHStat Random Sample Generation Dialog

desired in the *Sample Size* box. Click the first radio button if you want a list of random integers, and enter the population size in the box below this option. Click the second radio button to select a sample from data on a worksheet. The range of the data must be entered in the *Values Cell Range* box (checking *First cell contains label* if appropriate). This range must be a single column containing the values from which to draw the random sample.

Excel provides a similar tool for generating random samples. Click on *Data Analysis* in the *Analysis* group of the *Data* tab and select *Sampling*. This brings up the dialog box shown in Figure 4A2.2. In the *Input Range* box, you specify the data range from which the sample will be taken. (This tool requires that the data sampled be numeric; the PHStat *Random Sample Generation* tool does not have this restriction.) The *Labels* box can be checked if the first row is a data set label. There are two options for sampling:

1. Sampling can be *periodic*, and you will be prompted for the *Period*, which is the interval between sample observations from the beginning of the data set. For instance, if a period of 5 is used, observations 5, 10, 15, etc. will be selected as samples.
2. Sampling can also be *random*, and you will be prompted for the *Number of Samples*. Excel will then randomly select this number of samples (with replacement) from the specified data set.



**FIGURE 4A2.2** Excel Sampling Tool Dialog

However, this tool generates random samples *with replacement*, so you must be careful to check for duplicate observations in the sample. The *PHStat* tool samples without replacement, so it will not generate duplicates.

## B. Using the VLOOKUP Function

The VLOOKUP function allows you to look up a value in a table. The VLOOKUP function is useful in generating random variates from a discrete probability distribution. It is similar in concept to the IF function, but it allows you to pick an answer from an entire table of values. VLOOKUP stands for a vertical lookup table; (a similar function, HLOOKUP—for horizontal lookup table, is also available). The function  $VLOOKUP(A,X:Y,B)$  uses three arguments:

1. The value to look up ( $A$ )
2. The table range ( $X:Y$ )
3. The number of the column whose value we want ( $B$ )

To illustrate this, suppose we want to place in cell G14 a number 1, 2, or 3, depending on the contents of cell B9. If the value in cell B9 is 0.55 or less, then G14 should be 1; if it is greater than 0.55 but 0.85 or less, then G14 should be 2; and if it is greater than 0.85, then cell G14 should be 3. We must first put the data in a table in cells B4 through C6:

B	C
4	0
5	0.55
6	0.85

Now consider the formula in cell G14:

$$= VLOOKUP(B9,B4:C6,2)$$

The VLOOKUP function takes the value in cell B9 and searches for a corresponding value in the first column of the range B4:C6. The search ends when the first value greater than the value in cell B9 is found. The function then returns to the previous row in column B, picks the number found in the cell in the second column of that row in the table range, and enters it in cell G14. Suppose the number placed in cell B9 is 0.624. The function would search column B until it finds the first number larger than 0.624. This is 0.85 in cell B6. Then it returns to row 5 and picks the number in column C as the value of the function. Thus, the value placed in cell G14 is 2.

## C. Sampling from Probability Distributions

Excel and *PHStat* have procedures for generating random variates from some standard types of probability distributions. From the *Data* tab in Excel, select *Data Analysis* in the *Analysis* group and then *Random Number Generation*. The *Random Number Generation* dialog box, shown in Figure 4A2.3, will appear. From the *Random Number Generation* dialog box, you may select from seven distributions: uniform, normal, Bernoulli, binomial, Poisson, and patterned, as well as discrete. (The patterned distribution is characterized by a lower and upper bound, a step, a repetition rate for values, and a repetition rate for the sequence.) You are asked to specify the upper-left cell reference of the output table that will store the outcomes, the number of variables (columns of values you want generated), number of random numbers (the number of data points you want generated for each variable), and the type of distribution. The default distribution is the discrete distribution, which we illustrate. To use the discrete distribution, the spreadsheet must contain a table with two columns: the left column containing the outcomes and the right column



**FIGURE 4A2.3** Excel Random Number Generation Dialog

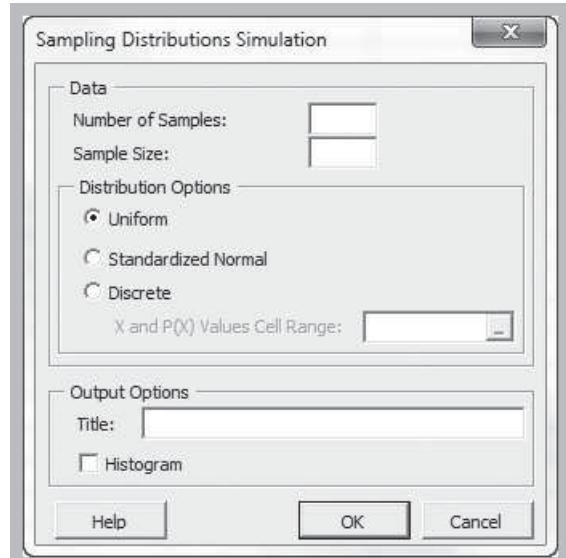
	A	B	C
1	x	f(x)	Simulated Dice Rolls
2	2	0.027778	9
3	3	0.055556	10
4	4	0.083333	9
5	5	0.111111	8
6	6	0.138889	6
7	7	0.166667	5
8	8	0.138889	3
9	9	0.111111	7
10	10	0.083333	7
11	11	0.055556	10
12	12	0.027778	

**FIGURE 4A2.4** Simulating Dice Rolls with the Random Number Generation Tool

containing the probabilities associated with the outcomes (which must sum to 1.0). Figure 4A2.4 shows an example of simulating dice rolls.

The dialog box in Figure 4A2.3 also allows you the option of specifying a random number seed. A **random number seed** is a value from which a stream of random numbers is generated. By specifying the same seed, you can produce the same random numbers at a later time. This is desirable when we wish to reproduce an identical sequence of "random" events in a simulation in order to test the effects of different policies or decision variables under the same circumstances.

From the *PHStat* menu, select *Sampling*, then *Sampling Distributions Simulation*. The dialog box shown in Figure 4A2.5 appears. You must enter the number of samples



**FIGURE 4A2.5** PHStat Sampling Distributions Simulation Dialog

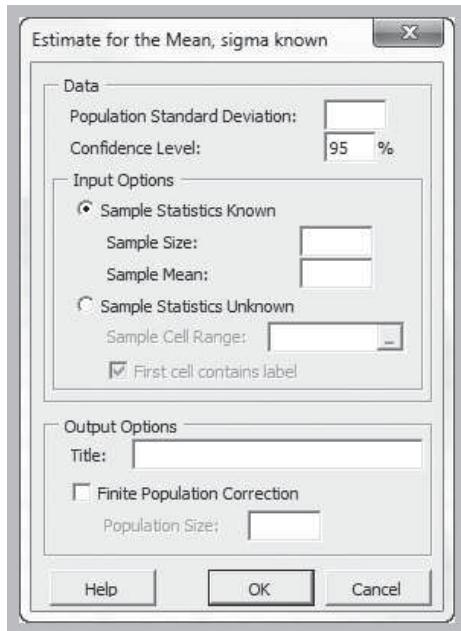
to be generated, the sample size, and the type of distribution (uniform, standardized normal, or discrete). If you select the discrete distribution, you also need to enter the range in a worksheet that contains the probability mass function. You may also opt for a histogram as part of the output. The procedure creates a new worksheet with the sample output in columns, along with the mean of each sample, overall mean, and standard error (to be discussed shortly).

## D. Confidence Intervals for the Mean

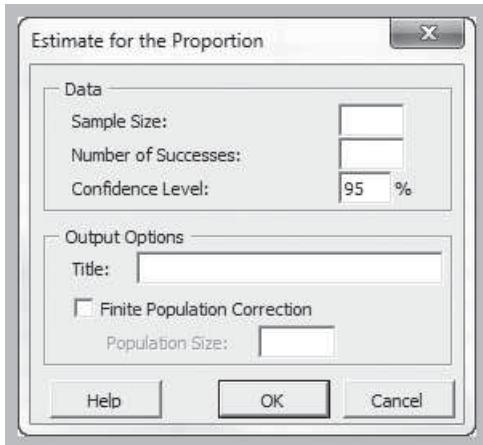
*PHStat* has two tools for finding confidence intervals for the mean: one assumes that the population standard deviation is known; the second assumes that it is unknown. Both input dialogs are similar. If the population standard deviation is assumed known, choose *Confidence Intervals* from the *PHStat* menu and select *Estimate for the Mean, Sigma Known*. The dialog box is shown in Figure 4A2.6. Enter the known population standard deviation. If the sample statistics (sample mean) has been calculated, you may enter it along with the sample size; otherwise, you may check the radio button for *Sample Statistics Unknown* and enter the range of the data, and the tool will perform the calculations. If the FPC is needed, check the box and enter the population size.

## E. Confidence Intervals for Proportions

From the *PHStat Confidence Intervals* menu, select *Estimate for the Proportion*. The dialog is shown in Figure 4A2.7. Enter the sample size and number of successes, that is, the number of observations having the characteristic of interest.



**FIGURE 4A2.6** PHStat Dialog for Confidence Interval Estimate for the Mean, Sigma Known

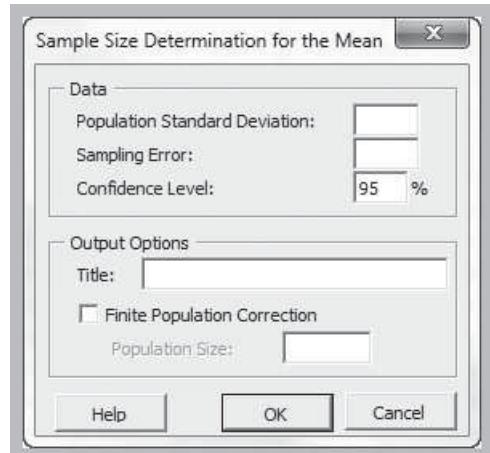


**FIGURE 4A2.7** PHStat Dialog for Confidence Interval Estimate for the Proportion

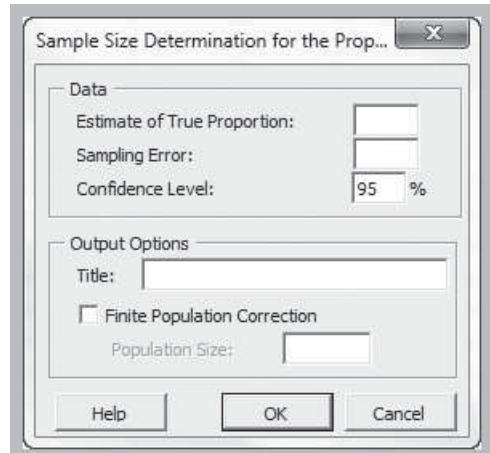
level. The tool provides confidence intervals for both the variance and standard deviation.

## G. Determining Sample Size

From the *PHStat* menu, select *Sample Size* and either *Determination for the Mean* or *Determination for the Proportion*. The dialog boxes are shown in Figures 4A2.8 and 4A2.9. You need to enter either the population standard deviation (or at least an estimate if it is known) or estimate of the true proportion, sampling error desired, and confidence level. The sampling error is the desired half-width of the confidence interval. The output options also allow you to incorporate a FPC factor if appropriate.



**FIGURE 4A2.8** PHStat Dialog for Sample Size Determination for the Mean



**FIGURE 4A2.9** PHStat Dialog for Sample Size Determination for the Proportion

## F. Confidence Intervals for the Population Variance

From the *PHStat* menu, select *Confidence Intervals* then *Estimate for the Population Variance*. The dialog box is simple, requiring only the sample size, sample standard deviation (which must be calculated beforehand), and confidence

## *Chapter 5*

# Hypothesis Testing and Statistical Inference

- INTRODUCTION 163
- BASIC CONCEPTS OF HYPOTHESIS TESTING 163
  - Hypothesis Formulation 164
  - Significance Level 165
  - Decision Rules 166
  - Spreadsheet Support for Hypothesis Testing 169
- ONE-SAMPLE HYPOTHESIS TESTS 169
  - One-Sample Tests for Means 169
  - Using  $p$ -Values 171
  - One-Sample Tests for Proportions 172
  - One Sample Test for the Variance 174
- TYPE II ERRORS AND THE POWER OF A TEST 175
- TWO-SAMPLE HYPOTHESIS TESTS 177
  - Two-Sample Tests for Means 177
  - Two-Sample Test for Means with Paired Samples 179
  - Two-Sample Tests for Proportions 179
  - Hypothesis Tests and Confidence Intervals 180
  - Test for Equality of Variances 181
- ANOVA: TESTING DIFFERENCES OF SEVERAL MEANS 182
  - Assumptions of ANOVA 184
  - Tukey–Kramer Multiple Comparison Procedure 184
- CHI-SQUARE TEST FOR INDEPENDENCE 186
- BASIC CONCEPTS REVIEW QUESTIONS 188
- PROBLEMS AND APPLICATIONS 188
- CASE: *HATCO, INC.* 191
- APPENDIX 5.1: HYPOTHESIS-TESTING THEORY AND COMPUTATION 191
  - A. Two-Sample Tests for Differences in Means 191
  - B. Two-Sample Test for Differences in Proportions 192
  - C. Test for Equality of Variances 192
  - D. Theory of Analysis of Variance 192

## ■ APPENDIX 5.2: EXCEL AND PHSTAT NOTES 193

- A. One-Sample Test for the Mean, Sigma Unknown 193
- B. One-Sample Test for Proportions 193
- C. Using Two-Sample *t*-Test Tools 193
- D. Testing for Equality of Variances 194
- E. Single-Factor Analysis of Variance 195
- F. Chi-Square Test for Independence 195

## INTRODUCTION

**Statistical inference** is the process of drawing conclusions about populations from sample data. For instance, a producer of computer-aided design software for the aerospace industry receives numerous calls for technical support. In the past, the average response time has exceeded 25 minutes. The company has upgraded its information systems and believes that this will help reduce response time. As a result, it believes that the average response time can be reduced to less than 25 minutes. A sample of 44 customers after the upgrade (given in the Excel file *Customer Support Survey*) revealed an average response time of 21.91 minutes. Can we conclude that the population mean from which this sample was drawn is truly less than 25? Similarly, in the Excel file *Burglaries*, local government officials might wish to determine whether implementing a federally funded citizen-police program will have a significant effect in reducing the rate of burglaries. A sample collected months before the program showed that the average number of monthly burglaries was 64.32, while the average of a sample after the program began was 60.65. Can we conclude that the program has significantly reduced the rate of burglaries?

In both cases, the sample means suggest an improvement. However, it's quite possible that for the customer support example, the population mean truly is 25 or more, and that we were just lucky to draw a sample whose mean was smaller. Similarly, although the average number of monthly burglaries appears to have fallen, we cannot tell whether this is significant or simply due to sampling error. Because of potential sampling error, it would be dangerous to conclude that the company was meeting its goal just by looking at the sample mean without better statistical evidence. In the two examples, one does show a significant statistical difference and the other does not. Can you guess which? **Hypothesis testing** is a technique that allows you to draw valid statistical conclusions about the value of population parameters or differences among them. In this chapter, we will learn how to use hypothesis testing and other important tools for statistical inference.

## BASIC CONCEPTS OF HYPOTHESIS TESTING

**Hypothesis testing** involves drawing inferences about two contrasting propositions (hypotheses) relating to the value of a population parameter, such as a mean, proportion, standard deviation, or variance. One of these propositions (called the **null hypothesis**) describes an existing theory (response time before the information systems upgrade exceeds 25 minutes) or a belief (the rate of burglaries has not improved). The second proposition (called the **alternative hypothesis**) is based on new information provided by sample data (response times after the system upgrade or burglary rates after the implementation of the citizen-police program). Based on the new sample data, we either (1) reject the null hypothesis and conclude that the sample data provide sufficient statistical evidence to support the alternative hypothesis or (2) fail to reject the null hypothesis and conclude that the sample data does not support the alternative hypothesis. If we fail to reject the null hypothesis, then we can only assume that the existing theory or belief is true, although we haven't proven it.

A good analogy of hypothesis testing is the U.S. legal system. In our system of justice, a defendant is innocent until proven guilty. The null hypothesis—our belief in the absence of any contradictory evidence—is "not guilty," while the alternative hypothesis

is “guilty.” If the evidence (sample data) strongly indicates that the defendant is guilty, then we reject the assumption of innocence. If the evidence is not sufficient to indicate guilt, then we cannot reject the “not guilty” hypothesis; however, we haven’t *proven* that the defendant is innocent.

Conducting a hypothesis test involves several steps:

1. Formulating the hypotheses to test
2. Selecting a *level of significance*, which defines the risk of drawing an incorrect conclusion about the assumed hypothesis that is actually true
3. Determining a decision rule on which to base a conclusion
4. Collecting data and calculating a test statistic
5. Applying the decision rule to the test statistic and drawing a conclusion

## Hypothesis Formulation

Hypothesis testing begins by defining two alternative, mutually exclusive propositions about one or more population parameters. The null hypothesis, denoted by  $H_0$ , represents an existing theory or belief that is accepted as correct in the absence of contradictory data. The alternative hypothesis, denoted by  $H_1$ , must be true if we reject the null hypothesis. In the *Customer Support Survey* example, the null and alternative hypotheses would be:

$$H_0: \text{mean response time} \geq 25 \text{ minutes}$$

$$H_1: \text{mean response time} < 25 \text{ minutes}$$

If we can find evidence that the mean response time is less than 25 minutes, then we would reject the null hypothesis and conclude that  $H_1$  is true. If we cannot find such evidence, we would simply have to assume that the mean response time is still 25 minutes or more because of the lack of contradictory data; however, we will not have proven this hypothesis in a statistical sense. Because of sampling error, it might be true that the true mean is indeed less than 25 minutes, but the sample we chose does not confirm this!

This example hypothesis test involves a single population parameter—the mean response time—and is called a *one-sample hypothesis test* because we will base our conclusion on one sample drawn from the population. We could also formulate hypotheses about the parameters of two populations, called *two-sample tests*, which involve drawing two samples, one from each of the two populations. For instance, in the burglaries example, we might define the null hypothesis to be:

$$\begin{aligned} H_0: & \text{Mean number of burglaries after program} \\ & - \text{mean number of burglaries before program} \geq 0 \\ H_1: & \text{Mean number of burglaries after program} \\ & - \text{mean number of burglaries before program} < 0 \end{aligned}$$

If we find statistical evidence to reject  $H_0$ , then we can conclude that  $H_1$  is true and that the mean number of burglaries after the program was implemented has been reduced.

Table 5.1 summarizes the types of one-sample and two-sample hypothesis tests we may conduct and the proper way to formulate the hypotheses. One-sample tests always compare a population parameter to some constant. For one-sample tests, note that the statements of the null hypotheses are expressed as either  $\geq$ ,  $\leq$ , or  $=$ . It is *not correct* to formulate a null hypothesis using  $>$ ,  $<$ , or  $\neq$ . For two-sample tests, the proper statement of the null hypothesis is always a difference between the population parameters, again expressing the null hypotheses as either  $\geq$ ,  $\leq$ , or  $=$ . In most applications, we compare the difference in means to zero; however, any constant may be used on the right-hand side of the hypothesis.

**TABLE 5.1 Types of Hypothesis Tests****One-Sample Tests**

- $H_0$ : population parameter  $\geq$  constant vs.  $H_1$ : population parameter  $<$  constant  
 $H_0$ : population parameter  $\leq$  constant vs.  $H_1$ : population parameter  $>$  constant  
 $H_0$ : population parameter = constant vs.  $H_1$ : population parameter  $\neq$  constant

**Two-Sample Tests**

- $H_0$ : population parameter (1) – population parameter (2)  $\geq 0$  vs.  
 $H_1$ : population parameter (1) – population parameter (2)  $< 0$   
 $H_0$ : population parameter (1) – population parameter (2)  $\leq 0$  vs.  
 $H_1$ : population parameter (1) – population parameter (2)  $> 0$   
 $H_0$ : population parameter (1) – population parameter (2) = 0 vs.  
 $H_1$ : population parameter (1) – population parameter (2)  $\neq 0$

How do we determine the proper form of the null and alternative hypotheses? Hypothesis testing always *assumes* that  $H_0$  is true and uses sample data to determine whether  $H_1$  is more likely to be true. Statistically, we cannot “prove” that  $H_0$  is true; we can only *fail to reject* it. Thus, if we cannot reject the null hypothesis, we have only shown that there is insufficient evidence to conclude that the alternative hypothesis is true. However, rejecting the null hypothesis provides strong evidence (in a statistical sense) that the null hypothesis is not true and that the alternative hypothesis is true. The legal analogy of “burden of proof” that we discussed earlier provides a way to understand this.

Burden of proof revolves around the alternative hypothesis. For example, if we wish to show statistically that the citizen-police program has had an effect in reducing the rate of burglaries, it would be *incorrect* to state the hypotheses as:

$$\begin{aligned} H_0: & \text{Mean number of burglaries after program} \\ & - \text{mean number of burglaries before program} \leq 0 \\ H_1: & \text{Mean number of burglaries after program} \\ & - \text{mean number of burglaries before program} > 0 \end{aligned}$$

If we fail to reject  $H_0$ , we have no conclusive evidence to support it; we have only failed to find evidence to support  $H_1$ . Therefore, we cannot conclude that the mean after the program started is necessarily any smaller than before. However, using the original forms of our hypotheses, if we have evidence that shows that the mean number of burglaries after the program began has decreased, then we can conclude that the program was beneficial. A useful way of thinking about this is whatever you would like to prove to be true should define the *alternative* hypothesis. Thus, in the *Customer Support Survey* example, the claim that the firm is meeting its goal of a mean response time of less than 25 minutes defines  $H_1$ .

**Significance Level**

Hypothesis testing can result in four different outcomes:

1. The null hypothesis is actually true, and the test correctly fails to reject it.
2. The null hypothesis is actually false, and the hypothesis test correctly reaches this conclusion.

3. The null hypothesis is actually true, but the hypothesis test incorrectly rejects it (called **Type I error**).
4. The null hypothesis is actually false, but the hypothesis test incorrectly fails to reject it (called **Type II error**).

The probability of making a Type I error,  $P(\text{Rejecting } H_0 | H_0 \text{ is true})$ , is generally denoted by  $\alpha$  and is called the **level of significance** of the test. This probability is essentially the risk that you can afford to take in making the incorrect conclusion that the alternative hypothesis is true when in fact the null hypothesis is true. The **confidence coefficient** is  $1 - \alpha$ , which is the probability of *correctly failing to reject* the null hypothesis, or  $P(\text{Not rejecting } H_0 | H_0 \text{ is true})$ . For a confidence coefficient of 0.95, we mean that we expect 95 out of 100 samples to support the null hypothesis rather than the alternate hypothesis. Commonly used levels for  $\alpha$  are 0.10, 0.05, and 0.01, resulting in confidence levels of 0.90, 0.95, and 0.99, respectively.

The probability of a Type II error,  $P(\text{Not rejecting } H_0 | H_0 \text{ is false})$ , is denoted by  $\beta$ . Unlike  $\alpha$ , this cannot be specified in advance but depends on the true value of the (unknown) population parameter. To see this, consider the hypotheses in the customer survey example:

$$H_0: \text{mean response time} \geq 25 \text{ minutes}$$

$$H_1: \text{mean response time} < 25 \text{ minutes}$$

If the true mean response from which the sample is drawn is, say, 15 minutes, we would expect to have a much smaller probability of incorrectly concluding that the null hypothesis is true than when the true mean response is 24 minutes, for example. In the first case, the sample mean would very likely be much less than 25, leading us to reject  $H_0$ . If the true mean is 24, however, even though the true mean response time is less than 25 minutes, we would have a much higher probability of failing to reject  $H_0$  because a higher likelihood exists that the sample mean would be greater than 25 due to sampling error. Thus, the farther away the true mean response time is from the hypothesized value, the smaller is  $\beta$ . Generally, as  $\alpha$  decreases,  $\beta$  increases, so the decision maker must consider the trade-offs of these risks.

The value  $1 - \beta$  is called the **power of the test** and represents the probability of *correctly rejecting* the null hypothesis when it is indeed false, or  $P(\text{Rejecting } H_0 | H_0 \text{ is false})$ . We would like the power of the test to be high to allow us to make a valid conclusion. The power of the test is sensitive to the sample size; small sample sizes generally result in a low value of  $1 - \beta$ . The power of the test can be increased by taking larger samples, which enable us to detect small differences between the sample statistics and population parameters with more accuracy. However, a larger sample size incurs higher costs, giving more meaning to the adage, “There is no such thing as a free lunch.” Table 5.2 summarizes this discussion.

## Decision Rules

The decision to reject or fail to reject a null hypothesis is based on computing a test statistic from sample data that is a function of the population parameter of interest and comparing it to a critical value from the hypothesized sampling distribution of the test

**TABLE 5.2 Error Types in Hypothesis Testing**

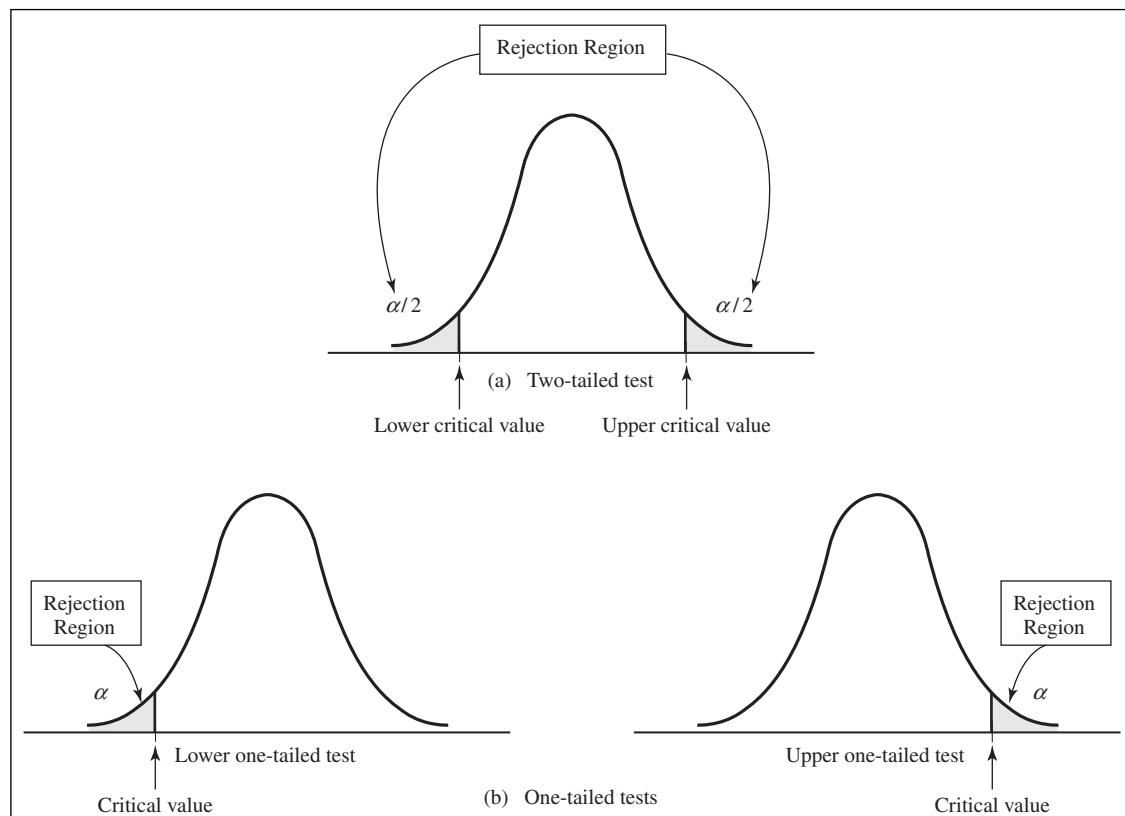
	Test Rejects $H_0$	Test Fails to Reject $H_0$
Alternative hypothesis ( $H_1$ ) is true	Correct	Type II error ( $\beta$ )
Null hypothesis ( $H_0$ ) is true	Type I error ( $\alpha$ )	Correct

statistic. The sampling distribution is usually the normal distribution, *t*-distribution, or some other well-known distribution. The sampling distribution is divided into two parts, a *rejection region* and a *nonrejection region*. If the null hypothesis is false, it is more likely that the test statistic will fall into the rejection region. If it does, we reject the null hypothesis; otherwise, we fail to reject it. The rejection region is chosen so that the probability of the test statistic falling into it if  $H_0$  is true is the probability of a Type I error,  $\alpha$ .

The rejection region generally occurs in the tails of the sampling distribution of the test statistic and depends on the structure of the hypothesis test (Table 5.1). If the null hypothesis is structured as “=” and the alternative hypothesis as “≠,” then we would reject  $H_0$  if the test statistic is either significantly high or low. In this case, the rejection region will occur in both the upper and lower tails of the distribution [see Figure 5.1(a)]. This is called a **two-tailed test of hypothesis**. Because the probability that the test statistic falls into the rejection region, given that  $H_0$  is true, the combined area of both tails must be  $\alpha$ . Usually, each tail has an area of  $\alpha/2$ .

The other types of hypothesis, which specifies a direction of relationship (where  $H_0$  is either “ $\geq$ ” or “ $\leq$ ,” are called **one-tailed tests of hypothesis**. In this case, the rejection region occurs only in one tail of the distribution [see Figure 5.1(b)]. Determining the correct tail of the distribution to use as the rejection region for a one-tailed test is easy. If  $H_1$  is stated as “ $<$ ,” the rejection region is in the lower tail; if  $H_1$  is stated as “ $>$ ,” the rejection region is in the upper tail (just think of the inequality as an arrow pointing to the proper tail direction!).

The rejection region is defined by a *critical value* (see Figure 5.1), which is the value that divides the rejection region from the rest of the distribution. Two-tailed tests have both upper and lower critical values, while one-tailed tests have either a lower or upper



**FIGURE 5.1** Illustration of Rejection Regions in Hypothesis Testing

**TABLE 5.3** Common Types of Hypothesis Tests

Type of Test	Excel/PHStat Procedure
One-sample test for the mean, $\sigma$ known	PHStat: One-Sample Test—Z-test for the Mean, Sigma Known
One-sample test for the mean, $\sigma$ unknown	PHStat: One-Sample Test—t-test for the Mean, Sigma Unknown
One-sample test for the proportion	PHStat: One-Sample Test—Z-test for the Proportion
One-sample test for the variance	PHStat: One-Sample Test—Chi-Square Test for the Variance
Two-sample test for means, $\sigma^2$ known	Excel Z-test: Two-Sample for Means PHStat: Two-Sample Tests—Z-test for Differences in Two Means
Two-sample test for means, $\sigma^2$ unknown, unequal	Excel t-test: Two-Sample Assuming Unequal Variances PHStat: Two-Sample Tests—Separate-Variance t-test
Two-sample test for means, $\sigma^2$ unknown, assumed equal	Excel t-test: Two-Sample Assuming Equal Variances PHStat: Two Sample Tests—Pooled Variance t-test
Paired two-sample test for means	Excel t-test: Paired Two-Sample for Means
Two-sample test for proportions	PHStat: Two-Sample Tests—Z-test for Differences in Two Proportions
Two-sample test for equality of variances	Excel F-test Two-Sample for Variances PHStat: Two-Sample Tests—F-Test for Differences in Two Variances

**TABLE 5.4** Additional Excel Support for Hypothesis Testing and Statistical Inference

Excel 2010 Function	Description
CHISQ.TEST(actual_range, expected_range)	Returns the test for independence, the value of the chi-square distribution, and the appropriate degrees of freedom
T.TEST(array1, array2, tails, type)	Returns the probability associated with a t-test
T.DIST(x, deg_freedom, cumulative)	Returns the left-tailed t-distribution
T.DIST.2T(x, deg_freedom)	Returns the two-tailed t-distribution
T.DIST.RT(x, deg_freedom)	Returns the right-tailed t-distribution
Z.TEST(array, x, sigma)	Returns the two-tailed p-value of a Z-test
F.TEST(array1, array2)	Returns the result of an F-test, the two-tailed probability that the variances in Array1 and Array2 are not significantly different

Analysis Toolpak Tools	Description
ANOVA: Single Factor	Tests hypothesis that means of two or more samples measured on one factor are equal

PHStat Add-In	Description
One Sample Tests	Hypothesis tests for the mean, proportion, and variance
Two Sample Tests	Hypothesis tests for differences in means, proportions, and variances
Multiple-Sample Tests/One-Way ANOVA	Tests hypothesis that means of two or more samples measured on one factor are equal
Multiple-Sample Tests/Chi-Square Test	Performs chi-square test of independence

critical value. For standard normal and *t*-distributions, which have a mean of 0, lower-tail critical values are negative; upper-tail critical values are positive.

Critical values make it easy to determine whether or not the test statistic falls in the rejection region of the proper sampling distribution. For example, for an upper one-tailed test, if the test statistic is greater than the critical value, the decision would be to reject the null hypothesis. Similarly, for a lower one-tailed test, if the test statistic is less than the critical value, we would reject the null hypothesis. For a two-tailed test, if the test statistic is either greater than the upper critical value or less than the lower critical value, the decision would be to reject the null hypothesis.

## Spreadsheet Support for Hypothesis Testing

Both Excel and *PHStat* have numerous tools for conducting hypothesis tests. In some cases, only one procedure is available; in others, both Excel and *PHStat* provide tools. Table 5.3 summarizes some of the more common types of hypothesis tests and Excel and *PHStat* tools for conducting these tests. Each test can be applied to different forms of the hypotheses (that is, when the null hypothesis is stated as “ $\geq$ ,” “ $\leq$ ,” or “ $=$ ”). The challenge is to identify the proper test statistic and decision rule and to understand the information provided in the output. Table 5.4 provides a summary of additional spreadsheet support for hypothesis testing. We will illustrate these tests through examples in the next several sections. Many other types of hypothesis tests exist, some of which are described in the appendix to this chapter; however, spreadsheet-based procedures are not available.

## ONE-SAMPLE HYPOTHESIS TESTS

In this section, we will discuss several hypothesis tests for means, proportions, and variances involving a single sample.

### One-Sample Tests for Means

We will first consider one-sample tests for means. The appropriate sampling distribution and test statistic depends on whether the population standard deviation is known or unknown. If the population standard deviation is known, then the sampling distribution of the mean is normal; if not, we use a *t*-distribution, which was introduced in Chapter 4 when discussing confidence intervals for the mean with an unknown population standard deviation. In most practical applications, the population standard deviation will not be known but is estimated from the sample, so we will only illustrate this case.

For the *Customer Support Survey* data, we will test the hypotheses:

$$H_0: \text{mean response time} \geq 25 \text{ minutes}$$

$$H_1: \text{mean response time} < 25 \text{ minutes}$$

with a level of significance of 0.05. This is a lower-tailed, one-sample test for the mean with an unknown standard deviation.

We will use the *PHStat* procedure for a one-sample test for the mean with an unknown population standard deviation (see Appendix 5.2A, “One-Sample Test for the Mean, Sigma Unknown”). From the 44 observations in the Excel file, we computed the sample mean to be 21.91 and the sample standard deviation as 19.49. Figure 5.2 shows the output provided by *PHStat*. The *Data* portion of the output simply summarizes the hypothesis we are testing, level of significance specified, and sample statistics. The *t-Test Statistic* is calculated using the formula:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad (5.1)$$



Spreadsheet Note

A	B	C	D	E
1 Response Time				
2				
3 Data				
4 Null Hypothesis $\mu =$	25			
5 Level of Significance	0.05			
6 Sample Size	44			
7 Sample Mean	21.91			
8 Sample Standard Deviation	19.49			
9				
10 Intermediate Calculations				
11 Standard Error of the Mean	2.938228053			
12 Degrees of Freedom	43			
13 t Test Statistic	-1.051654243			
14				
15 Lower-Tail Test			Calculations Area	
16 Lower Critical Value	-1.681070703	For one-tailed tests:		
17 p-Value	0.149416269	TDIST value	0.149416	
18 Do not reject the null hypothesis		1-TDIST value	0.850584	

**FIGURE 5.2** Results for One-Sample Test for the Mean, Sigma Unknown

where  $\mu_0$  is the hypothesized value and  $s/\sqrt{n}$  is the standard error of the sampling distribution of the mean. Applied to this example, we have:

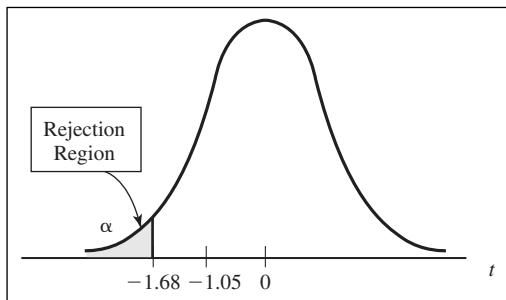
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{21.91 - 25}{19.49/\sqrt{44}} = \frac{-3.09}{20938} = -1.05$$

Observe that the numerator is the distance between the sample mean (21.91) and the hypothesized value (25). By dividing by the standard error, the value of  $t$  represents the number of standard errors the sample mean is from the hypothesized value. In this case, the sample mean is 1.05 standard errors below the hypothesized value of 25.

This notion provides the fundamental basis for the hypothesis test—if the sample mean is “too far” away from the hypothesized value, then the null hypothesis should be rejected. The decision is based on the level of significance,  $\alpha$ . For a one-tailed test, the critical value is the number of standard errors away from the hypothesized value for which the probability of exceeding the critical value is  $\alpha$ . If  $\alpha = 0.05$ , for example, then we are saying that there is only a 5% chance that a sample mean will be that far away from the hypothesized value purely because of sampling error, and that the small likelihood of this occurring suggests that the true population mean is different from what was hypothesized.

The *t-Test Statistic* has a  $t$ -distribution with  $n - 1$  degrees of freedom. If the level of significance is 0.05, then the critical value for a lower-tail test is the value of the  $t$ -distribution with  $n - 1$  degrees of freedom that provides a lower tail area of 0.05; that is,  $t_{\alpha, n-1}$ . We may find  $t$ -values in Table A.2 in the appendix at the end of the book or by using the Excel function T.INV( $\alpha, n - 1$ ). Hence, to find the critical value, we find  $t_{0.05, 43} = \text{T.INV}(0.05, 43) = -1.68$ . (The T.INV function returns the left-tail critical value; for the right-tail critical value for an upper-tail test, use T.INV(1 -  $\alpha, n - 1$ )).

By comparing the *t-Test Statistic* with the *Lower Critical Value*, we see that the test statistic does not fall below the critical value (that is,  $-1.05 > -1.68$ ) and is not in the rejection region. Therefore, we cannot reject  $H_0$  and cannot conclude that the mean response time is less than 25 minutes. Figure 5.3 illustrates the conclusion we reached. Even though the sample mean is less than 25, we cannot conclude that the mean response time is less than 25 minutes because of the large amount of sampling error.



**FIGURE 5.3** *t*-Test for Mean Response

## Using *p*-Values

In the *PHStat* output in Figure 5.2, we see something called a *p-value*. An alternative approach to comparing a test statistic to a critical value in hypothesis testing is to find the probability of obtaining a test statistic value equal to or more extreme than that obtained from the sample data when the null hypothesis is true. This probability is commonly called a **p-value**, or **observed significance level**. For example, the *t-test Statistic* for the hypothesis test in the response time example is  $-1.05$ . If the true mean is really  $25$ , then what is the probability of obtaining a test statistic of  $-1.05$  or less (the area to the left of  $-1.05$  in Figure 5.3)? Equivalently, what is the probability that a sample mean from a population with a mean of  $25$  will be at least  $1.05$  standard errors below  $25$ ? We can calculate this using the Excel function  $\text{T.DIST}(-1.05, 43, \text{TRUE}) = 0.149416$ . In other words, there is about a  $15\%$  chance that the test statistic would be  $-1.05$  or smaller if the null hypothesis were true. This is a fairly high probability, so it would be difficult to conclude that the true mean is less than  $25$ , and we could attribute the fact that the test statistic is less than the hypothesized value to sampling error alone and not reject the null hypothesis. In general, compare the *p*-value to the chosen level of significance; whenever  $p < \alpha$ , reject the null hypothesis. *p*-values make it easy to draw conclusions about hypothesis tests.

Next, we illustrate a two-tailed test using the data for the respondents in the Excel file *Vacation Survey*. Suppose that the sponsor of the survey wanted to target individuals who were approximately  $35$  years old. Thus, we wish to test whether the average age of respondents is equal to  $35$ . The hypothesis to test is:

$$H_0: \text{mean age} = 35$$

$$H_1: \text{mean age} \neq 35$$

The sample mean is computed to be  $39.12$ , and the sample standard deviation is  $7.53$ . Figure 5.4 shows the results using the *PHStat t-Test for the Mean, Sigma Unknown*.

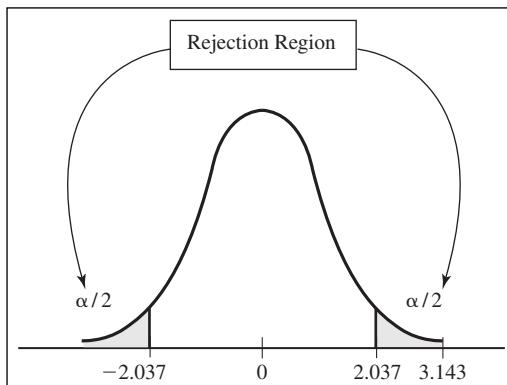
The test statistic is the same as in the previous example:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = (39.12 - 35)/(7.53/\sqrt{33}) = 3.143$$

In this case, the sample mean is  $3.143$  standard errors above the hypothesized mean of  $35$ . However, because this is a two-tailed test, the rejection region and decision rule are different. For a level of significance  $\alpha$ , we reject  $H_0$  if the *t-Test Statistic* falls either below the negative critical value,  $-t_{\alpha/2, n-1}$ , or above the positive critical value,  $t_{\alpha/2, n-1}$ . Using the Excel function  $\text{T.INV}(0.025, 32)$  to calculate  $t_{0.025, 32}$ , we obtain  $-2.0369$ . Thus, the critical values are  $\pm 2.0369$ . Because the *t-Test Statistic* does not fall between these values, we must reject the null hypothesis that the average age is  $35$  (see Figure 5.5).

A	B
1 Age	
2	
3 Data	
4 Null Hypothesis $\mu =$	35
5 Level of Significance	0.05
6 Sample Size	33
7 Sample Mean	39.12121212
8 Sample Standard Deviation	7.532253878
9	
10 Intermediate Calculations	
11 Standard Error of the Mean	1.311197099
12 Degrees of Freedom	32
13 t Test Statistic	3.143091244
14	
15 Two-Tail Test	
16 Lower Critical Value	-2.036933343
17 Upper Critical Value	2.036933343
18 p-Value	0.003592445
19	Reject the null hypothesis

**FIGURE 5.4** PHStat Results for Two-Tailed  $t$ -Test



**FIGURE 5.5** Illustration of Two-Tailed  $t$ -Test

The  $p$ -value for this test is 0.00359, which can also be computed by the Excel function T.DIST.2T(3.143,32). Note that the probability to the right of 3.143 is actually only T.DIST.RT(3.143,32) = 0.001795, which is half of the  $p$ -value. So we cannot compare 0.001795 to  $\alpha = 0.05$  to draw a conclusion, since the tail area to the right is only  $\alpha/2 = 0.025$ . Because this is a two-tailed test, we must double this probability to find the correct  $p$ -value to compare against  $\alpha$ . Since the  $p$ -value is smaller than the chosen significance level of 0.05, we reject  $H_0$ .

### One-Sample Tests for Proportions

Many important business measures, such as market share or the fraction of deliveries received on time, are expressed as proportions. For example, in generating the *Customer Support Survey* data, one question asks the customer to rate the overall quality of the company's software product using a scale of

- 0—Very poor
- 1—Poor
- 2—Good
- 3—Very good
- 4—Excellent

These data are in column G in the Excel file *Customer Support Survey*. The firm tracks customer satisfaction of quality by measuring the proportion of responses in the top two categories. Over the past, this proportion has averaged about 75%. For these data, 30 of the 44 responses, or 68.2%, are in the top two categories. Is there sufficient evidence to conclude that this satisfaction measure has dropped below 75%? Answering this question involves testing the hypotheses about the population proportion  $\pi$ :

$$H_0: \pi \geq 0.75$$

$$H_1: \pi < 0.75$$

*PHStat* has a tool for performing this test (see Appendix 5.2B, “One-Sample Test for Proportions”), but Excel does not. Applying this tool to the *Customer Support Survey* data, we obtain the results shown in Figure 5.6. The Z-Test Statistic for a one-sample test for proportions is:

$$Z = \frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}} \quad (5.2)$$

where  $\pi_0$  is the hypothesized value and  $\hat{p}$  is the sample proportion. The denominator represents the standard error for the sampling distribution of the proportion and is shown in the intermediate calculations. For this example,

$$Z = \frac{0.682 - 0.75}{\sqrt{0.75(1 - 0.75)/44}} = -1.04$$

Similar to the test statistic for means, the Z-Test Statistic shows the number of standard errors that the sample proportion is from the hypothesized value. In this case, the sample proportion of 0.68 is 1.04 standard errors below the hypothesized value of 0.75. Because this is a lower one-tailed test, we reject  $H_0$  if the Z-Test Statistic is less than the lower critical value. The sampling distribution of Z is a standard normal; therefore, for a level of significance of 0.05, the critical value of Z is found by the Excel function NORM.S.INV(0.05) = -1.645. Because the Z-Test Statistic is not less than the Lower Critical Value, we cannot reject the null hypothesis that the proportion is at least 0.75. We would attribute the low proportion of responses in the top two boxes to sampling error



Spreadsheet Note

A	B
1	Overall Quality Satisfaction
2	
Data	
4	Null Hypothesis $\pi =$ 0.75
5	Level of Significance 0.05
6	Number of Items of Interest 30
7	Sample Size 44
8	
Intermediate Calculations	
10	Sample Proportion 0.681818182
11	Standard Error 0.065279121
12	Z Test Statistic -1.044465936
13	
Lower-Tail Test	
15	Lower Critical Value -1.644853627
16	p-Value 0.148134936
17	Do not reject the null hypothesis

FIGURE 5.6 *PHStat* Results for Z-Test for Proportions

and the relatively small sample size. Note that the  $p$ -value is greater than the significance level, leading to the same conclusion of not rejecting the null hypothesis.

## One Sample Test for the Variance

In many applications, we are interested in testing a hypothesis about the value of a population variance. For example, in manufacturing, understanding variability is important to ensure that a process consistently produces output that meets desired specifications. If the variance of the process changes significantly, then it is quite possible that undesirable defects will result. For example, in the *Process Capability* worksheet in the Excel file *Quality Measurements*, we have 200 samples of blade weights taken from the manufacturing process that produces mower blades. When the manufacturing process is working well, the variance in the blade weights should be no greater than 0.010. The variance of this sample is calculated to be 0.0136095. Can the manufacturer conclude that the variance has significantly increased?

The hypothesis we are testing is:

$$H_0: \sigma^2 \leq 0.10$$

$$H_1: \sigma^2 > 0.10$$

*PHStat* has a tool for performing this test, *Chi-Square Test for the Variance* (see Appendix 5.2C). Recall that in Chapter 4, we observed that a confidence interval for the variance was based on a chi-square distribution; thus, this hypothesis test also uses a chi-square test statistic:

$$\chi^2 = (n - 1)s^2/\sigma_0^2 \quad (5.3)$$

Here,  $s^2$  is the sample variance, and  $\sigma_0^2$  is the hypothesized variance. This statistic has a chi-square distribution with  $n - 1$  degrees of freedom. For the data, the value of the test statistic is:

$$\chi^2 = 199(0.0136095)/0.01 = 270.83$$

If  $\chi^2$  exceeds the critical value,  $\chi_{\alpha}^2$ , which can be found in Table A.3 for small sample sizes or using the Excel function *CHISQ.INV( $\alpha$ , degrees of freedom)*, then we reject the null hypothesis. Figure 5.7 shows the results of using the *PHStat* tool. We see that the variance of the manufacturing process appears to have increased.

A	B
Chi-Square Test of Variance	
4 Null Hypothesis	$\sigma^2 =$
5 Level of Significance	0.05
6 Sample Size	200
7 Sample Standard Deviation	0.1166597
Intermediate Calculations	
10 Degrees of Freedom	199
11 Half Area	0.025
12 Chi-Square Statistic	270.8287635
Upper-Tail Test	
15 Upper Critical Value	232.9118218
16 $p$ -Value	0.00053211
Reject the null hypothesis	

FIGURE 5.7 *PHStat* Results for Chi-Square Test of Variance

Spreadsheet Note



Choosing the correct critical value depends on the type of hypothesis test. For a lower-tailed test, reject  $H_0$  if  $\chi^2 < \chi^2_{(1-\alpha)}$ , and for a two-tailed test, reject  $H_0$  if  $\chi^2 < \chi^2_{(1-\alpha/2)}$ , or if  $\chi^2 < \chi^2_{(\alpha/2)}$ , using the appropriate degrees of freedom. As always, reject  $H_0$  if the  $p$ -value is less than  $\alpha$ .

## TYPE II ERRORS AND THE POWER OF A TEST

The probability of a Type I error,  $\alpha$ , can be specified by the experimenter. However, the probability of a Type II error,  $\beta$  (the probability of failing to reject  $H_0$  when it indeed is false), and the power of the test ( $1 - \beta$ ) are the result of the hypothesis test itself. Understanding the power of a test is important to interpret and properly apply the results of hypothesis testing. The power of the test depends on the true value of the population mean, the level of confidence used, and the sample size. This is illustrated in Figure 5.8. Suppose that we are testing the null hypothesis  $H_0: \mu \geq \mu_0$  against the alternative  $H_1: \mu < \mu_0$ , and suppose that the true mean is actually  $\mu_1$ . We specify the probability of a Type I error,  $\alpha$ , in the lower tail of the hypothesized distribution; this defines the rejection region. However, the sampling distributions overlap, and it is possible that the test statistic will fall into the acceptance region, even though the sample comes from a distribution with mean  $\mu_1$  leading us to not reject  $H_0$ . The area that overlaps the hypothesized distribution in the acceptance region is  $\beta$ , the probability of a Type II error.

From the figure, it would seem intuitive that  $\beta$  would be smaller the farther away the true mean is from  $\mu_0$ . In other words, if the true mean is close to  $\mu_0$ , it would be difficult to distinguish much of a difference, and the probability of concluding that  $H_0$  is true would be high. However, if the true mean is far away, the chances that sampling error alone would lead us to conclude that  $H_0$  was true would probably be quite small. This is illustrated in Figure 5.9.

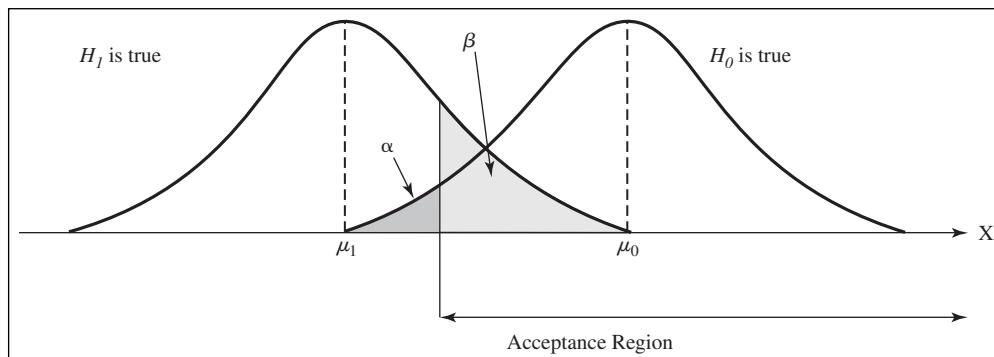
In a similar fashion,  $\beta$  also depends on the sample size. We know that the standard error of the mean decreases as the sample size increases. This makes the sampling distribution narrower and reduces the overlap shown in Figure 5.10, thus reducing the value of  $\beta$  (see Figure 5.10).

Clearly, we would like the power of the test to be high so that we can more easily discriminate between the two hypotheses and avoid Type II errors. However, because the true mean is unknown, we cannot determine  $\beta$  exactly. All we can do is to calculate it for different values of the true mean and assess the potential for committing a Type II error. This is relatively easy to do.

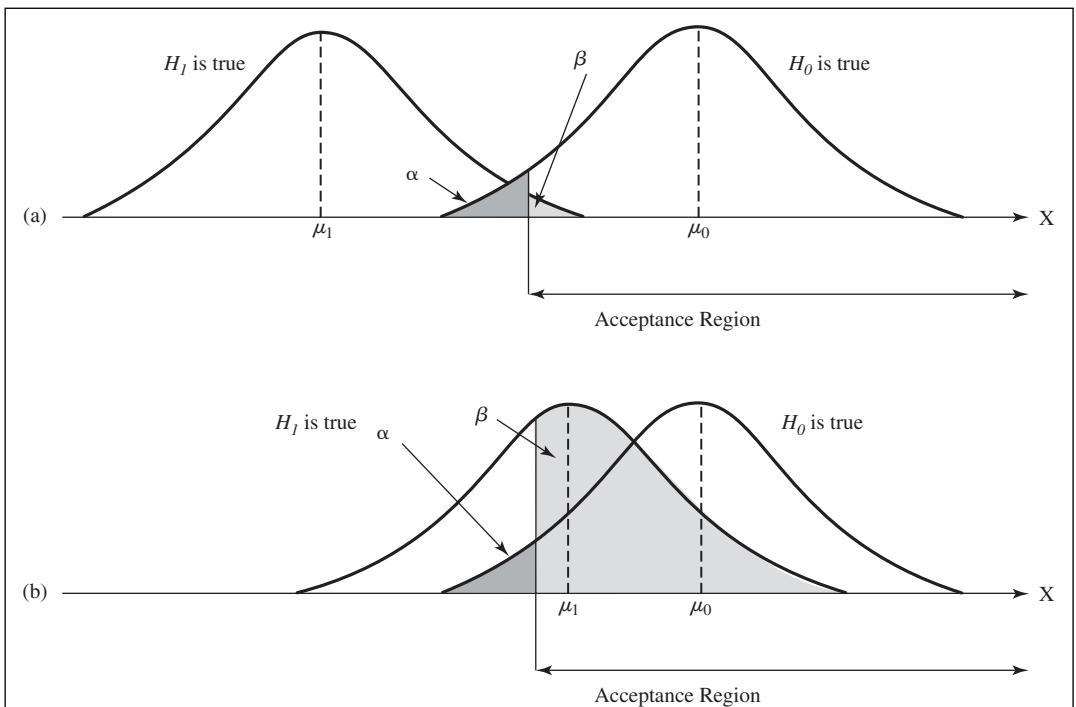
We will illustrate this using the *Customer Support Survey* data in the context of Figure 5.9. Recall that the hypothesis test is:

$$H_0: \text{mean response time} \geq 25 \text{ minutes}$$

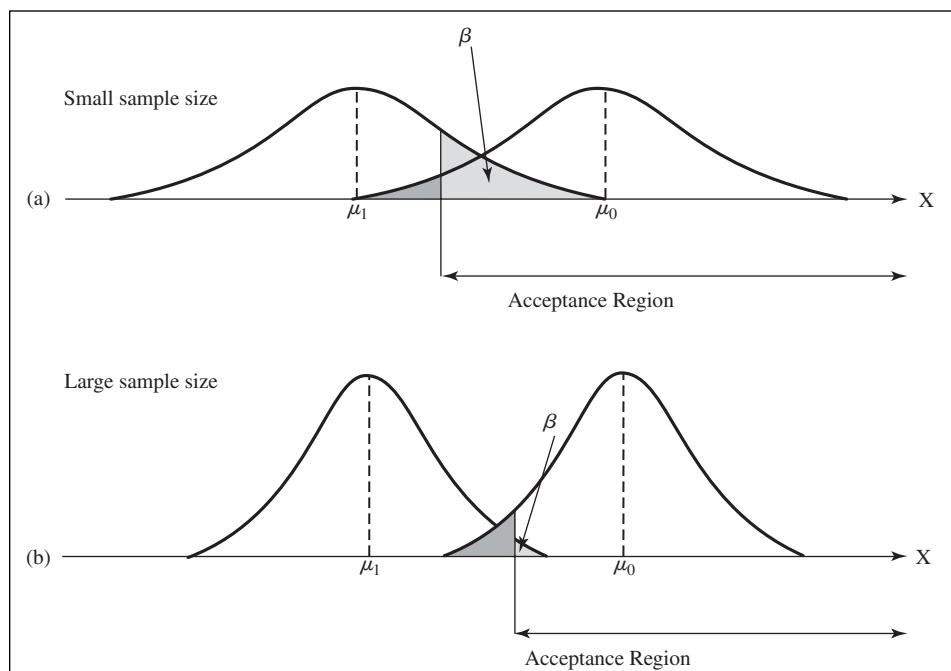
$$H_1: \text{mean response time} < 25 \text{ minutes}$$



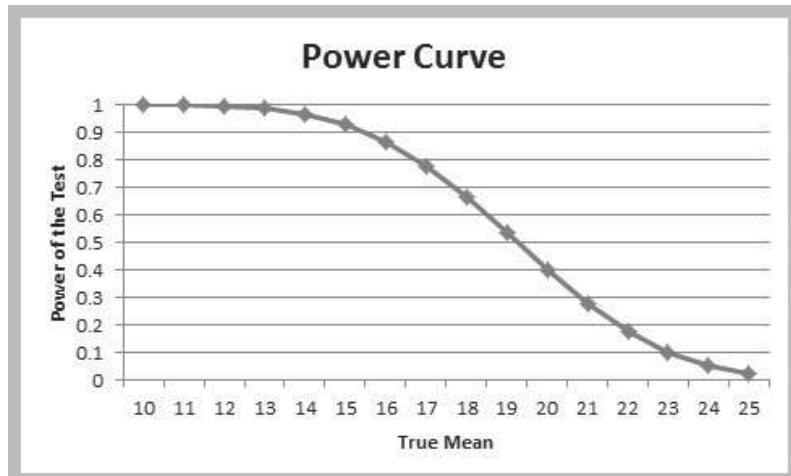
**FIGURE 5.8** Finding the Probability of a Type II Error



**FIGURE 5.9** How  $\beta$  Depends on  $H_1$



**FIGURE 5.10** How  $\beta$  Depends on Sample Size



**FIGURE 5.11** Power Curve for Customer Support Survey Hypothesis

and that the standard error of the mean was calculated to be 2.938. Because the sample size is large, we will use  $z$ -values instead of  $t$ . If  $\alpha = 0.05$ , then the left-hand limit of the acceptance region is 1.96 standard errors to the left of 25, or  $25 - 1.96(2.938) = 19.24$ . Now suppose that the true mean is  $\mu_1 = 23$ . The area to the right of 19.24 for a distribution with a mean of 23 is  $\beta$ . The  $z$ -value is calculated as  $z = (19.24 - 23)/2.938 = -1.28$ . Because  $\beta = 1 - \text{NORM.S.DIST}(-1.28)$ , we find that is  $1 - 0.1003 = 0.8997$ . Now suppose that the true mean is only 15. Then the  $z$ -value corresponding to 19.24 for this distribution is  $z = (19.24 - 15)/2.938 = 1.443$ . The area to the right of this value is  $\beta = 1 - \text{NORM.S.DIST}(1.443) = 0.0745$ . If we perform these calculations for a range of values for  $\mu_1$  (using the NORM.S.DIST function as we have illustrated), we could draw a **power curve** that shows  $(1 - \beta)$  as a function of  $\mu_1$ , shown in Figure 5.11. The relative steepness of the curve will help assess the risk of a Type II error.

### SKILL-BUILDER EXERCISE 5.1

For the *Customer Support Survey* data, calculate the probability of a Type II error if the hypothesized value is 30 and the true mean is either 20 or 25. Compare your answers to the power curve in Figure 5.11.

## TWO-SAMPLE HYPOTHESIS TESTS

Hypothesis testing finds wide applicability in comparing two populations for differences in means, proportions, or variances. The hypothesis-testing procedures are similar to those previously discussed in the sense of formulating one- or two-tailed tests and identifying the critical test statistic and rejection region. The formulas for the test statistics are more complicated than for one-sample tests, and are discussed in Appendix 5.1 at the end of this chapter.

### Two-Sample Tests for Means

In the Excel file *Burglaries* that we described earlier, we might wish to test the hypotheses:

$$H_0: \text{mean burglaries/month after program } (\mu_1) \\ - \text{mean burglaries/month before program } (\mu_2) \geq 0$$

$$H_1: \text{mean burglaries/month after program } (\mu_1) - \text{mean burglaries/month before program } (\mu_2) < 0$$

Rejecting the null hypothesis suggests that the citizen-police program was effective in reducing the number of burglaries. If we cannot reject the null hypothesis, then even though the mean number of monthly burglaries is smaller since the program began, the difference would most likely be due to sampling error and is not statistically significant.

Selection of the proper test statistic for a two-sample test for means depends on whether the population standard deviations are known, and if not, whether they are assumed to be equal. *PHStat* and Excel have procedures for conducting these tests (see Appendix 5.2D, “Using Two-Sample *t*-Test Tools”). Please read this carefully because care must be taken in interpreting the Excel output.



Spreadsheet Note

For the burglary data, the population variances are unknown, but the sample variances are close enough that we will assume equality for selecting the proper tool. Figure 5.12 shows the output from both the *PHStat* and Excel tests for the difference in means with unknown but equal variances. From the *PHStat* output, we see that the *t-Test Statistic* is  $-0.768$ . When compared to the lower critical value of  $-1.67252$ , we cannot reject the null hypothesis and conclude that there is insufficient evidence to claim that the mean number of burglaries has decreased after the citizen-police program was initiated. The *p-value* of  $0.222808$  confirms this, as it is much larger than the level of significance.

The Excel results are more difficult to interpret properly. As stated in the note in Appendix 5.2, for a lower-tailed test, we must change the sign on the critical value (*t Critical one-tail*) to make it negative ( $-1.67252$ ), and then compare this to the *t Stat* value of  $-0.768$  to draw the correct conclusion. Note that for a two-tailed test, the critical value is specified for the upper tail only, and you must also realize that the lower-tail critical value is the negative; however, the *p-value* provided is correct.

A	B	C	D	E	F	G
1 Pooled-Variance <i>t</i> Test for the Difference Between Two Means 2 (assumes equal population variances)				t-Test: Two-Sample Assuming Equal Variances		
3 Data						
4 Hypothesized Difference	0				Variable 1	Variable 2
5 Level of Significance	0.05			Mean	60.64705882	64.31707317
6 Population 1 Sample				Variance	253.8676471	282.4719512
7 Sample Size	17			Observations	17	41
8 Sample Mean	60.647			Pooled Variance	274.2992929	
9 Sample Standard Deviation	15.933			Hypothesized Mean Difference	0	
10 Population 2 Sample				df	56	
11 Sample Size	41			t Stat	-0.768170828	
12 Sample Mean	64.317			P(T<=t) one-tail	0.222806507	
13 Sample Standard Deviation	16.807			t Critical one-tail	1.672522303	
14				P(T<=t) two-tail	0.445613014	
15 Intermediate Calculations				t Critical two-tail	2.003240719	
16 Population 1 Sample Degrees of Freedom	16					
17 Population 2 Sample Degrees of Freedom	40					
18 Total Degrees of Freedom	56					
19 Pooled Variance	274.2996					
20 Difference in Sample Means	-3.67					
21 <i>t</i> Test Statistic	-0.76817					
22						
23 Lower-Tail Test						
24 Lower Critical Value	-1.67252					
25 <i>p</i> -Value	0.222808					
26 Do not reject the null hypothesis						

FIGURE 5.12 Excel and *PHStat* Results for Differences in Means

## Two-Sample Test for Means with Paired Samples

The Excel file *Pile Foundation* contains the estimates used in a bid and actual auger cast pile lengths for a foundation engineering project. The contractor's past experience suggested that the bid information was generally accurate, so the average difference between the actual pile lengths and estimated lengths should be close to 0. After this project was completed, the contractor found that the average difference was 6.38. Did he receive poor bid information, or was this difference simply the result of sampling error?

In this situation, the data from the two samples are naturally paired. When this is the case, a paired *t*-test is more accurate than assuming that the data come from independent populations, as was the case for the two-sample tests for means that we discussed earlier. The hypothesis to test is:

$$H_0: \text{average difference} = 0$$

$$H_1: \text{average difference} \neq 0$$

Excel has a *Data Analysis* tool, *t-Test: Paired Two Sample for Means*. In the dialog, you need only enter the variable ranges and hypothesized mean difference. Figure 5.13 shows the results.

We compute the difference for each pile,  $D_i$ , by subtracting the estimated value from the actual value and finding the average difference,  $\bar{D}$ , and standard deviation of the differences,  $s_D$ . The test statistic is:

$$t = \frac{\bar{D} - \mu_D}{s_D / \sqrt{n}} \quad (5.4)$$

with  $n - 1$  degrees of freedom. Here,  $\mu_D$  is the hypothesized difference between the means, or 0. For the data, we found the average difference to be 6.3 and  $s_D = 10.31$ . Therefore,  $t = 6.38 / (10.31 / \sqrt{311}) = 10.91$ . This is a two-tailed test, so in Figure 5.13 we interpret the results using only the two-tail information. As before, Excel only provides the upper-tail critical value. Because  $t$  is much larger than the upper critical value of 1.96 (assuming a 0.05 significance level), we must reject the null hypothesis and conclude that the mean of the estimates is significantly different from the mean of the actual pile lengths. Note that the *p*-value is essentially 0, verifying this conclusion.

## Two-Sample Tests for Proportions

Similar to means, we may conduct hypothesis tests for differences in proportions. For example, suppose that a company wishes to test whether the proportion of

	A	B	C
1	t-Test: Paired Two Sample for Means		
2			
3		Variable 1	Variable 2
4	Mean	34.55623794	28.17755627
5	Variance	267.0113061	255.8100385
6	Observations	311	311
7	Pearson Correlation	0.79692836	
8	Hypothesized Mean Difference	0	
9	df	310	
10	t Stat	10.91225025	
11	P(T<=t) one-tail	5.59435E-24	
12	t Critical one-tail	1.649783823	
13	P(T<=t) two-tail	1.11887E-23	
14	t Critical two-tail	1.967645863	

**FIGURE 5.13** Results for Paired Two-Sample *t*-Test for Means

females who have a graduate degree is the same as the proportion of males. We test the hypotheses:

$$H_0: \pi_1 - \pi_2 = 0$$

$$H_1: \pi_1 - \pi_2 \neq 0$$

where  $\pi_1$  is the population proportion of females who have a graduate degree and  $\pi_2$  is the population proportion of males who have a graduate degree. PHStat has a tool for conducting this test, *Z-Test for Differences in Two Proportions*, found in the *Two-Sample Tests* menu. The dialog prompts you to enter the hypothesized difference, and the number of successes and sample size for each sample. Assume that 46 out of 141 females have a graduate degree, and 88 out of 256 males do. Figure 5.14 shows the results, including the optional calculation of a confidence interval.

The formula used for calculating the *Z-Test Statistic* is explained in Appendix 5.1. Note that the test statistic value falls between the lower and upper critical values, indicating that we should not reject the null hypothesis that the proportion of females who have a graduate degree is the same as the proportion of males. Again, the *p*-value exceeds the level of significance, leading to the same conclusion. The confidence interval shows that the estimated difference in the population proportions can be either positive or negative.

### Hypothesis Tests and Confidence Intervals

You may have been thinking that a relationship exists between hypothesis tests and confidence intervals because they rely on essentially the same information, and indeed, this is true. For instance, in the previous example for differences in proportions, we observed that the confidence interval did not conclusively show that the difference in proportions between females and males is either positive or negative, and we did not

A	B	C	D	E
1	Graduate Degree Differences by Gender			
2				
3	Data			
4	Hypothesized Difference	0		
5	Level of Significance	0.05		
6	Group 1			
7	Number of Successes	46		
8	Sample Size	141		
9	Group 2			
10	Number of Successes	88		
11	Sample Size	256		
12				
13	Intermediate Calculations			
14	Group 1 Proportion	0.326241135		
15	Group 2 Proportion	0.34375		
16	Difference in Two Proportions	-0.017508865		
17	Average Proportion	0.337531486		
18	Z Test Statistic	-0.353063254		
19				
20	Two-Tail Test			
21	Lower Critical Value	-1.959963985		
22	Upper Critical Value	1.959963985		
23	p-Value	0.72404102		
24	Do not reject the null hypothesis			

Confidence Interval Estimate of the Difference Between Two Proportions	
Data	
Confidence Level	95%
Intermediate Calculations	
Z Value	-1.959963985
Std. Error of the Diff. between two Proportions	0.049397531
Interval Half Width	0.096817381
Confidence Interval	
Interval Lower Limit	-0.114326246
Interval Upper Limit	0.079308516

FIGURE 5.14 PHStat Results for Differences in Proportions

reject the null hypothesis. In contrast for the paired test for means with the pile foundation data, we rejected the null hypothesis that the population means are the same, and found in the previous chapter that a 95% confidence interval for the mean difference was (5.234, 7.526). This confidence interval suggests a positive difference between the actual and estimated pile lengths. Thus, in testing the difference between two population parameters, if the confidence interval for the difference contains 0, then we would not reject the corresponding null hypothesis. However, if both confidence limits are either positive or negative, then we would reject the null hypothesis.

For one-tail hypothesis tests, we need to examine on which side of the hypothesized value the confidence interval falls. For example, suppose that we tested the hypotheses:

$$H_0: \text{mean response time} \geq 30 \text{ minutes}$$

$$H_1: \text{mean response time} < 30 \text{ minutes}$$

for the *Customer Support Survey* data using a level of significance of 0.05. You may verify that we would reject this null hypothesis and conclude that the mean is less than 30 minutes. If we construct a 95% confidence interval for the mean response time, we obtain (15.98, 27.84). Note that this confidence interval lies entirely below 30; therefore, we can reach the same conclusion.

## Test for Equality of Variances

As we have seen, Excel supports two different *t*-tests for differences in means, one assuming equal variances and the other assuming unequal variances. We can test for equality of variances between two samples using a new type of test, the *F*-test. To use this test, we must assume that both samples are drawn from normal populations. To illustrate the *F*-test, we will use the data in the Excel file *Burglaries* and test whether the variance in the number of monthly burglaries is the same before and after the citizen-police program began.

$$H_0: \sigma_{\text{before}}^2 = \sigma_{\text{after}}^2$$

$$H_1: \sigma_{\text{before}}^2 \neq \sigma_{\text{after}}^2$$

The *F*-test can be applied using either the *PHStat* tool found in the menu by choosing *Two-Sample Tests...F-test for Differences in Two Variances*, or the Excel *Data Analysis* tool *F-test for Equality of Variances* (see Appendix 5.2E, “Testing for Equality of Variances”).

Figure 5.15 shows the output from both tests. Both tests compute an *F*-test statistic. The *F*-test statistic is the ratio of the variances of the two samples:

$$F = \frac{s_1^2}{s_2^2} \tag{5.5}$$

For this example, the test statistic is  $F = s_1^2/s_2^2 = 282.47/253.87 = 1.113$ . This statistic is compared to critical values from the *F*-distribution for a given confidence level. Like the *t*-distribution, the *F*-distribution is characterized by degrees of freedom. However, the *F*-statistic has *two* values of degrees of freedom—one for the sample variance in the numerator and the other for the sample variance in the denominator. In both cases, the number of degrees of freedom is equal to the respective sample size minus 1. The *PHStat* output provides both lower and upper critical values from the *F*-distribution. We see that the *F*-Test Statistic falls between these critical values; therefore, we do not reject the null hypothesis. The *p*-value of 0.847 confirms this, as it is larger than the level of significance.

Proper interpretation of the Excel results depends on how we take the ratio of sample variances. As suggested in the note, we recommend ensuring that variable 1 has



Spreadsheet Note

A	B	C	D	E	F
1 PHStat Results for Equality of Variances					Excel Results for Equality of Variances
2					
3 Data					F-Test Two-Sample for Variances
4 Level of Significance	0.05				
5 Population 1 Sample					
6 Sample Size	41				Variable 1
7 Sample Standard Deviation	16.81				Variable 2
8 Population 2 Sample					Mean
9 Sample Size	17				64.31707317
10 Sample Standard Deviation	15.93				Variance
11					282.4719512
12 Intermediate Calculations					Observations
13 F Test Statistic	1.113535				41
14 Population 1 Sample Degrees of Freedom	40				df
15 Population 2 Sample Degrees of Freedom	16				F
16					1.112674082
17 Two-Tail Test					P(F<=f) one-tail
18 Lower Critical Value	0.464213				0.424390536
19 Upper Critical Value	2.508529				F Critical one-tail
20 p-Value	0.847369				2.508529216
21 Do not reject the null hypothesis					

**FIGURE 5.15** *PHStat* and Excel Results for Equality of Variances Test

the larger variance. Note that when we do this,  $F > 1$ . If the variances differ significantly, we would expect  $F$  to be much larger than 1; the closer  $F$  is to 1, the more likely it is that the variances are the same. Therefore, we need only compare  $F$  to the upper-tail critical value. However, because Excel provides results for a one-tail test only, we must use  $\alpha/2$  as the input value in the Excel dialog.

In comparing the results, note that the computed  $F$ -test statistic is the same (to within some differences because of rounding the calculations of the sample standard deviations in *PHStat*) and that the upper critical values are both the same. Therefore, the same conclusions are reached by comparing the test statistic to the upper-tail critical value. However, note that the  $p$ -values differ, because in *PHStat* the level of significance is entered as  $\alpha$ , whereas in Excel, the level of significance was entered as  $\alpha/2$  for a one-tailed test. Thus, the  $p$ -value for Excel is half that of *PHStat*. These are both correct, as long as you realize that you compare the  $p$ -values to the proper levels of significance; that is, in *PHStat*,  $0.847369 > 0.05$ , while in Excel,  $0.42439 > 0.025$ . The same conclusions are reached in both cases.

### SKILL-BUILDER EXERCISE 5.2

Use the burglary data to run the test for equality of variances by switching the variables so that the value of  $F$  is less than 1. Using both the *PHStat* and Excel tools, explain how to interpret the Excel results properly in this situation.

## ANOVA: TESTING DIFFERENCES OF SEVERAL MEANS

To this point, we have discussed hypothesis tests that compare a population parameter to a constant value or that compare the means of two different populations. Often, we would like to compare the means of several different groups to determine if all are equal, or if any are significantly different from the rest. For example, in the Excel file *Insurance Survey*, we might be interested in whether any significant differences exist in satisfaction among individuals with different levels of education. In statistical terminology, educational level

is called a **factor**, and we have three categorical levels of this factor. Thus, it would appear that we will have to perform three different pairwise tests to establish whether any significant differences exist among them. As the number of factor levels increases, you can easily see that the number of pairwise tests grows large very quickly. Fortunately, other statistical tools exist that eliminate the need for such a tedious approach. **Analysis of variance (ANOVA)** provides a tool for doing this. The null hypothesis is that the population means of all groups are equal; the alternative hypothesis is that at least one mean differs from the rest:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_m$$

*H<sub>1</sub>: at least one mean is different from the others*

Let us apply ANOVA to test the hypothesis that the mean satisfaction for all educational levels in the Excel file *Insurance Survey* are equal against the alternative that at least one mean is different. The table below shows the summary statistics:

College Graduate	Graduate Degree	Some College
5	3	4
3	4	1
5	5	4
3	5	2
3	5	3
3	4	4
3	5	4
4	5	
2		
Average	3.444	4.500
Count	9	8
		3.143
		7

Excel provides a *Data Analysis* tool, *ANOVA: Single Factor* (see Appendix 5.2F, “Single-Factor Analysis of Variance”) to conduct analysis of variance. The results for the example are given in Figure 5.16. The output report begins with a summary report of basic statistics for each group. The ANOVA section reports the details of the hypothesis test. ANOVA derives its name from the fact that we are analyzing variances in the data. The theory is described more fully in Appendix 5.1, but essentially ANOVA computes a measure of the variance between the means of each group, and a measure of the variance within the groups. These measures are in the MS column (MS stands for “mean square”). If the null hypothesis is true, the between-group variance should be small, as each of the group means should be close to one another. If the means among groups are significantly different, then MS Between Groups will be significantly larger than MS Within Groups. The ratio of the variances, MS Between Groups/MS Within Groups, is a test statistic from an *F*-distribution (similar to the test for equality of variances). Thus, when MS Between Groups is large relative to MS Within Groups, *F* will be large. If the *F*-statistic is large enough based on the level of significance chosen and exceeds a critical value, we would reject the null hypothesis. In this example,  $F = 3.939484/1.003779 = 3.92$  and the critical value (*F crit*) from the *F*-distribution is 3.4668. Here  $F > F crit$ ; therefore, we must reject the null hypothesis and conclude that there are significant differences in the means of the groups; that is, the mean satisfaction is not the same among the three educational levels. Alternatively, we see that the *p*-value is smaller than the chosen level of significance, 0.05, leading to the same conclusion.



Spreadsheet Note

	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	Groups	Count	Sum	Average	Variance		
5	College Graduate	9	31	3.444444444	1.027777778		
6	Graduate Degree	8	36	4.5	0.571428571		
7	Some College	7	22	3.142857143	1.476190476		
8							
9							
10	ANOVA						
11	Source of Variation	SS	df	MS	F	P-value	F crit
12	Between Groups	7.878968254	2	3.939484127	3.924651732	0.035635398	3.466800112
13	Within Groups	21.07936508	21	1.003779289			
14							
15	Total	28.95833333	23				

**FIGURE 5.16** ANOVA Results for *Insurance Survey Data*

### Assumptions of ANOVA

ANOVA requires assumptions that the  $m$  groups or factor levels being studied represent populations whose outcome measures

1. Are randomly and independently obtained
2. Are normally distributed
3. Have equal variances

If these assumptions are violated, then the level of significance and the power of the test can be affected. Usually, the first assumption is easily validated when random samples are chosen for the data. ANOVA is fairly robust to departures from normality, so in most cases this isn't a serious issue. The third assumption is required in order to pool the variances within groups. If sample sizes are equal, this assumption does not have serious effects on the statistical conclusions; however, with unequal sample sizes, it can. When you suspect this, you can use the *Levene test* to investigate the hypothesis:

$$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_m^2$$

$$H_1: \text{Not all } \sigma_j^2 \text{ are equal}$$

The Levene test can be found in the *PHStat Multiple Sample Test* menu.

When the assumptions underlying ANOVA are violated, you may use a **nonparametric test** that does not require these assumptions; for example, the Kruskal-Wallis rank test for determining whether multiple populations have equal medians. This test is available in *PHStat* also, and we refer you to more comprehensive texts on statistics for further information and examples.

### Tukey-Kramer Multiple Comparison Procedure

Figure 5.17 shows an ANOVA for the customer ratings of four product dimensions in the Excel file *Customer Survey*. The  $F$ -test statistic is much larger than the critical value (and the  $p$ -value is essentially 0), leading us to reject the null hypothesis. Although ANOVA can identify a difference among the means of multiple populations, it cannot determine which of the means are significantly different from the rest. To do this, we may use the **Tukey-Kramer multiple comparison procedure**. This method compares

	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	Groups	Count	Sum	Average	Variance		
5	Quality	200	879	4.395	0.581884		
6	Ease of Use	200	833	4.165	0.610829		
7	Price	200	734	3.67	1.136784		
8	Service	200	828	4.14	0.794372		
9							
10							
11	ANOVA						
12	Source of Variation	SS	df	MS	F	P-value	F crit
13	Between Groups	55.505	3	18.50167	23.6907	1.08E-14	2.616089
14	Within Groups	621.65	796	0.780967			
15							
16	Total	677.155	799				

**FIGURE 5.17** ANOVA results for Customer Survey Data

the absolute value of the difference in means for all pairs of groups and compares these to a critical range. Values exceeding the critical range identify those populations whose means differ. *PHStat* provides this procedure if you check the box in the *One-Way ANOVA* tool. Figure 5.18 shows the results of the procedure for the customer survey data. We see that in comparing one group to another, all means are significantly different except for Group 2 and Group 4 (Ease of Use and Service). Note that although the means of Quality, Ease of Use, and Service all appear to be close to one another, the large sample size provides the ability to differentiate between these seemingly similar customer ratings.

Analysis of variance may be extended to more than one factor. For example, suppose that a company wants to investigate whether changes in temperature and pressure settings affect output quality or yield in some production process. Temperature and pressure represent two factors, and multiple samples might be selected for each of three different combinations of temperature levels and pressure settings. This would be an example of a two-factor ANOVA and would allow the investigator to test hypotheses about whether differences exist among the levels of each factor individually and also whether any interactions exist between the factors; that is, whether the effect of one factor depends on the level of the other. Further discussion is beyond the scope of this book, but additional information may be found in more comprehensive statistics texts and books devoted exclusively to the subject.

	A	B	C	D	E	F	G	H	I	J	K
1	Customer Survey										
2											
3		Sample	Sample			Absolute	Std. Error	Critical			
4	Group	Mean	Size		Comparison	Difference	of Difference	Range	Results		
5	1	4.395	200		Group 1 to Group 2	0.23	0.06248869	0.2268	Means are different		
6	2	4.165	200		Group 1 to Group 3	0.725	0.06248869	0.2268	Means are different		
7	3	3.67	200		Group 1 to Group 4	0.255	0.06248869	0.2268	Means are different		
8	4	4.14	200		Group 2 to Group 3	0.495	0.06248869	0.2268	Means are different		
9					Group 2 to Group 4	0.025	0.06248869	0.2268	Means are not different		
10	Other Data				Group 3 to Group 4	0.47	0.06248869	0.2268	Means are different		
11	Level of significance	0.05									
12	Numerator d.f.	4									
13	Denominator d.f.	796									
14	MSW	0.780967									
15	Q Statistic	3.63									

**FIGURE 5.18** Results from Tukey–Kramer Procedure

## CHI-SQUARE TEST FOR INDEPENDENCE

A common problem in business is to determine whether two categorical variables are independent. We discussed the notion of statistical independence in Chapter 3 and why it is important in marketing applications. For example, a consumer study might collect data on preferences for three different energy drinks of both male and female high school students. The objective of the study might be to determine if energy drink preferences are independent of gender. Independence would mean that the proportion of individuals who prefer one drink over another would be essentially the same no matter if the individual is male or female. On the other hand, if males have different preferences than females, the variables would be dependent. Knowing this can help marketing personnel better target advertising campaigns to different demographic groups.

In Chapter 3, we described how to determine if two random variables are statistically independent by examining the joint and marginal probabilities. However, with empirical data, sampling error can make it difficult to properly assess independence of categorical variables. We would never expect the joint probabilities to be exactly the same as the product of the marginal probabilities because of sampling error even if the two variables are statistically independent. However, we can draw a conclusion using a hypothesis test called the *chi-square test for independence*. The chi-square test for independence tests the following hypotheses:

$$H_0: \text{the two categorical variables are independent}$$

$$H_1: \text{the two categorical variables are dependent}$$

For example, the data below show the sample data used in Chapter 3 for brand preferences of energy drinks.

Cross-Tabulation	Brand 1	Brand 2	Brand 3	Total
MALE	25	17	21	63
FEMALE	9	6	22	37
TOTAL	34	23	43	100

The chi-square test for independence tests whether the proportion of males who prefer a particular brand is no different from the proportion of females. For instance, of the 63 male students, 25 (40%) prefer Brand 1. If gender and brand preference are indeed independent, we would expect that about the same proportion of the sample of female students would also prefer Brand 1. In actuality, only 9 of 37 (24%) prefer Brand 1. However, we do not know whether this is simply due to sampling error or represents a significant difference.

 **PHStat** provides a tool for conducting this test (see the note in Appendix 5.2F, "Chi-square Test for Independence"). Figure 5.19 shows the completed worksheet for the test. The test calculates the expected frequencies that should be observed if the null hypothesis is true. For example, if the variables are independent, we would expect to find about 21 or 22 males and about 12 or 13 females who prefer Brand 1. To compute the expected frequency for a particular cell in the table, simply multiply the row total by the column total and divide by the grand total. Thus, the expected frequency for Male and Brand 1 is  $(63)(34)/100 = 21.42$ . One caution: All expected frequencies should be at least 1.0. If not, then you should combine rows or columns to

	A	B	C	D	E	
1	Brand Preferences					
2						
3	Observed Frequencies					
4		Column variable				
5	Row variable	Brand 1	Brand 2	Brand 3	Total	
6	Male	25	17	21	63	
7	Female	9	6	22	37	
8	Total	34	23	43	100	
9						
10	Expected Frequencies					
11		Column variable				
12	Row variable	Brand 1	Brand 2	Brand 3	Total	
13	Male	21.42	14.49	27.09	63	
14	Female	12.58	8.51	15.91	37	
15	Total	34	23	43	100	
16						
17	Data					
18	Level of Significance	0.05				
19	Number of Rows	2				
20	Number of Columns	3				
21	Degrees of Freedom	2				
22						
23	Results					
24	Critical Value	5.99146				
25	Chi-Square Test Statistic	6.49243				
26	p -Value	0.03892				
27	Reject the null hypothesis					
28						
29	<i>Expected frequency assumption</i> is met.					
30						

**FIGURE 5.19** Chi-Square Test for Independence Results

meet this requirement. Drawing from the discussion in Chapter 3, you can verify that the product of the marginal probabilities equals the joint probabilities in the Expected Frequencies table.

The procedure uses the observed and expected frequencies to compute a test statistic, called a **chi-square statistic**, which is the sum of the squares of the differences between observed frequency,  $f_o$ , and expected frequency,  $f_e$ , divided by the expected frequency in each cell:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \quad (5.6)$$

The closer the observed frequencies are to the expected frequencies, the smaller will be the value of the chi-square statistic. We compare this statistic for a specified level of significance  $\alpha$  to the critical value from a chi-square distribution with  $(r - 1)(c - 1)$  degrees of freedom, where  $r$  and  $c$  are the number of rows and columns in the contingency table, respectively. If the test statistic exceeds the critical value for a specified level of significance, we reject  $H_0$ . In this case, the chi-square test statistic is 6.49243, and the critical value is 5.99146. Because the test statistic exceeds the critical value, we reject the null hypothesis that the two categorical variables are independent. The small  $p$ -value

of 0.03892 also confirms this conclusion. The Excel function CHISQ.TEST(*actual\_range*, *expected\_range*) computes the *p*-value for the chi-square test, but you must calculate the expected frequencies.

### SKILL-BUILDER EXERCISE 5.3

Use formula (5.6) to verify the calculation of the chi-square test statistic for the example in Figure 5.19.

## Basic Concepts Review Questions

- How is the hypothesis testing problem different from that of estimation? Explain the general construction and steps in a hypothesis test.
- Explain the difference between the null and alternative hypothesis. Which one can be proven in a statistical sense?
- Explain why it is always necessary to specify a fixed null value in the one-sample hypothesis test, while such specifications are generally not necessary for the two-sample hypothesis test.
- Define the significance level of a hypothesis test, and the two kinds of errors associated with testing.
- Explain how the rejection region is determined for a hypothesis test. How does it differ between one- and two-tailed tests?
- How can you determine when to use a lower one-tailed test of hypothesis versus an upper one-tailed test?
- What is the difference between the significance level and the observed significance level?
- Explain the difference between the rejection region and *p*-value approaches for reaching a decision in a hypothesis test.

- How are Type II errors influenced by the true (unknown) population mean and sample size used in a hypothesis test?
- Explain the peculiar nuances associated with the Excel tools for two-sample *t*-tests. What issues must you consider to use the tests and interpret the results properly?
- What is the difference between paired and independent samples?
- In many statistical applications, we may be interested in finding confidence intervals for, or testing hypotheses about, the same set of parameters. Explain the link between the confidence level and the significance level in these cases.
- Explain how to interpret the Excel results for a hypothesis test for equality of variances.
- What is analysis of variance? What hypothesis does it test? Provide some practical examples.
- What are the key assumptions of ANOVA? What should be done if these are seriously violated?
- Explain the purpose of the chi-square test for independence. Provide some practical examples where this test might be used in business.

## Problems and Applications

For all hypothesis tests, assume that the level of significance is 0.05 unless otherwise stated.

- A company claims that the average nicotine content of its cigarettes is no more than 13 milligrams. An experimenter has selected a sample of 100 cigarettes from this company. Set up an appropriate hypothesis test to verify the company's claim.
- The state of Ohio Department of Education has a mandated ninth-grade proficiency test that covers writing, reading, mathematics, citizenship (social studies), and science. The Excel file *Ohio Education Performance* provides data on success rates (defined as the percentage of students passing) in school districts in the greater Cincinnati metropolitan area along with state averages. Test the null hypothesis that the average score in the Cincinnati area is equal to the state average in each test and for the composite score.
- Formulate and test a hypothesis to determine if statistical evidence suggests that the graduation rate for either top liberal arts colleges or research universities in the sample *Colleges and Universities* exceeds 90%. Do the data support a conclusion that the graduation rate exceeds 85%? Would your conclusions change if the level of significance were 0.01 instead of 0.05?
- The Excel file *Sales Data* provides data on a sample of customers. The company believes that its average profit is no less than \$5000. However, some analysts feel that its average profit may actually be lower. Test a hypothesis to verify this. If, for the same sample and the same value of the sample mean, the sample standard deviation equals \$1000, would the conclusion change?

5. The Excel file *Call Center Data* provides data on the background of call center employees. The company would like at least 40% of its employees to have a college degree. Perform a test with this data to determine whether there is evidence to claim that the company's requirements are not met.
6. It is claimed by some sources that, on the basis of the numbers presented in the Excel file *Atlanta Airline Data*, 65% or more of the total number of flights arrive early. Perform a test of hypothesis for this data to verify whether this claim can be supported.
7. In the Excel file *New Account Processing*, test the null hypothesis that the variance in the Close Ratio for the sample of employees is at most 1.5%.
8. In the Excel file *Gasoline Prices*, determine if there is evidence to conclude that the variance of weekly prices during 2009 is less than 10%.
9. Using the data in the Excel file *Consumer Transportation Survey*, test the following null hypotheses:
- Individuals spend at least 10 hours per week in their vehicles.
  - Individuals drive an average of 450 miles per week.
  - The average age of SUV drivers is no greater than 35.
  - At least 75% of individuals are satisfied with their vehicles.
10. An Internal Revenue Service auditor ran a test on a sample of income tax returns filed in two successive years to check whether the average refund for taxpayers is larger this year compared to the last. The mean and standard deviation for last year's and this year's samples are as given below:

	Last year	This year
Mean	340	390
Standard Deviation	28	31
Sample Size	500	500

Test whether the average refund increased this year compared to the last.

11. Do males and females have the same number of friends? Use the data in the Excel file *Facebook Survey* and perform an appropriate test to determine whether the mean number of friends for the two genders is the same.
12. Determine if there is evidence to conclude that the mean number of vacations taken by married individuals is less than the number taken by single/divorced individuals using the data in the Excel file *Vacation Survey*.
13. The Excel file *Accounting Professionals* provides the results of a survey of 27 employees in a tax division of a *Fortune 100* company.
- Test the null hypothesis that the average number of years of service is the same for males and females.
  - Test the null hypothesis that the average years of undergraduate study is the same for males and females.
14. In the Excel file *Cell Phone Survey*, test the hypothesis that the mean responses for Value for the Dollar and Customer Service do not differ by gender.
15. For the data in the Excel file *Graduate School Survey*, perform a test for proportions at level 10% to determine whether the proportion of individuals who plan to attend graduate school is the same for married and unmarried groups.
16. A study of nonfatal occupational injuries in the United States found that about 31% of all injuries in the service sector involved the back. The National Institute for Occupational Safety and Health (NIOSH) recommended conducting a comprehensive ergonomics assessment of jobs and workstations. In response to this information, Mark Glassmeyer developed a unique ergonomic handcart to help field service engineers be more productive and also to reduce back injuries from lifting parts and equipment during service calls. Using a sample of 382 field service engineers who were provided with these carts, Mark collected the following data:

	Year 1 (without cart)	Year 2 (with cart)
Average call time	8.05 hours	7.84 hours
Standard deviation call time	1.39 hours	1.34 hours
Proportion of back injuries	0.018	0.010

- Determine if there is statistical evidence that average call time has decreased as a result of using the cart. What other factors might account for any changes?
  - Determine if there is statistical evidence that the proportion of back injuries has decreased as a result of using the cart.
17. The director of human resources for a large bank has compiled data on about 70 former employees at one of the bank's call centers (see the Excel file *Call Center Data*). For each of the following, assume equal variances of the two populations.
- Test the null hypothesis that the average length of service for males is the same as for females.
  - Test the null hypothesis that the average length of service for individuals without prior call center experience is the same as those with experience.
  - Test the null hypothesis that the average length of service for individuals with a college degree is the same as for individuals without a college degree.
  - Now conduct tests of hypotheses for equality of variances. Were your assumptions of equal variances valid? If not, repeat the test(s) for means using the unequal variance test.
18. A producer of computer-aided design software for the aerospace industry receives numerous calls for technical support. Tracking software is used to monitor

response and resolution times. In addition, the company surveys customers who request support using the following scale: 0—Did not exceed expectations; 1—Marginally met expectations; 2—Met expectations; 3—Exceeded expectations; 4—Greatly exceeded expectations. The questions are as follows:

- Q1: Did the support representative explain the process for resolving your problem?
- Q2: Did the support representative keep you informed about the status of progress in resolving your problem?
- Q3: Was the support representative courteous and professional?
- Q4: Was your problem resolved?
- Q5: Was your problem resolved in an acceptable amount of time?
- Q6: Overall, how did you find the service provided by our technical support department?

A final question asks the customer to rate the overall quality of the product using a scale of 0—Very poor; 1—Poor; 2—Good; 3—Very good; 4—Excellent. A sample of survey responses and associated resolution and response data are provided in the Excel file *Customer Support Survey*.

- a. The company has set a service standard of one day for the mean resolution time. Does evidence exist that the response time is less than one day? How do the outliers in the data affect your result? What should you do about them?
  - b. Test the hypothesis that the average service index is equal to the average engineer index.
19. Jim Aspenwall, a NASCAR enthusiast, has compiled some key statistics for NASCAR Winston Cup racing tracks across the United States. These tracks range in shape, length, and amount of banking on the turns and straightaways (see the Excel file *NASCAR Track Data*). Test the hypothesis that there is no difference in qualifying record speed between oval and other shapes of tracks.
20. Some analysts feel that the Ohio Education Program is doing well in reading and writing, but not so in science and math. Perform a paired sample test to determine whether the mean math score is significantly smaller than the mean writing score.
21. The Excel file *Unions and Labor Law Data* reports the percentage of public- and private-sector employees in unions in 1982 for each state, along with indicators whether the states had a bargaining law that covered public employees or right-to-work laws.
- a. Test the hypothesis that the percentage of employees in unions for both the public sector and private sector is the same for states having bargaining laws as for those who do not.
  - b. Test the hypothesis that the percentage of employees in unions for both the public sector and private
- sector is the same for states having right-to-work laws as for those who do not.
22. For the data in the Excel file *Golfing Statistics*, test whether the variance of the driving distance of the group having driving accuracy less than 60% is the same as that of the group having accuracy greater than 60%.
23. Consider the data given in the Excel file *Unions and Labor Laws*. Perform a test of hypothesis for testing the equality of variances of the percentages of employees in unions in the public sector:
  - a. for states having bargaining laws and states not having bargaining laws.
  - b. for states having right-to-work laws and states not having right-to-work laws.
24. For the data in the Excel file *Student Grades*, perform an appropriate test to determine whether the average scores for the midterm and the final are the same.
25. For the data in the Excel file *Cell Phone Survey*, apply ANOVA to determine if the mean signal strengths are the same for the three different types of cell phones.
26. An engineer measured the surface finish of 35 parts produced on a lathe, noting the revolutions per minute of the spindle and the type of tool used (see the Excel file *Surface Finish*). Use ANOVA to test the hypothesis that the mean surface finish is the same for each tool. If the null hypothesis is rejected, apply the Tukey–Kramer multiple comparison procedure to identify significant differences.
27. Using the data in the Excel file *Freshman College Data*, use ANOVA to determine whether the mean retention rate is the same for all colleges over the 4-year period. Second, use ANOVA to determine if the mean ACT and SAT scores are the same each year over all colleges. If the null hypothesis is rejected, apply the Tukey–Kramer multiple comparison procedure to identify significant differences.
28. A mental health agency measured the self-esteem score for randomly selected individuals with disabilities who were involved in some work activity within the past year. The Excel file *Self Esteem* provides the data, including the individuals' marital status, length of work, type of support received (direct support includes job-related services such as job coaching and counseling), education, and age.
- a. Apply ANOVA to determine if self-esteem is the same for all marital status levels. If the null hypothesis is rejected, apply the Tukey–Kramer multiple comparison procedure to identify significant differences.
  - b. Use the chi-square test to determine if marital status is independent of support level.
29. For the data in the file *Accounting Professionals*, perform a chi-square test to determine whether gender and CPA status are independent.

30. For the data in the Excel file *Graduate School Survey* perform a chi-square test for independence to determine whether plans to attend graduate school are independent of marital status.

31. For the data in the Excel file *New Account Processing* test for the independence of gender and prior industry background.

## Case

### HATCO, Inc.

The Excel file *HATCO*<sup>1</sup> consists of data related to predicting the level of business obtained from a survey of purchasing managers of customers of an industrial supplier, HATCO. The variables are as follows:

- *Delivery Speed*—amount of time it takes to deliver the product once an order is confirmed
- *Price Level*—perceived level of price charged by product suppliers
- *Price Flexibility*—perceived willingness of HATCO representatives to negotiate price on all types of purchases
- *Manufacturing Image*—overall image of the manufacturer or supplier
- *Overall Service*—overall level of service necessary for maintaining a satisfactory relationship between supplier and purchaser
- *Sales Force Image*—overall image of the manufacturer's sales force
- *Product Quality*—perceived level of quality of a particular product

- *Size of Firm*—size relative to others in this market (0 = small; 1 = large)
- *Usage Level*—percentage of total product purchased from HATCO

Responses to the first seven variables were obtained using a graphic rating scale, where a 10-centimeter line was drawn between endpoints labeled “poor” and “excellent.” Respondents indicated their perceptions using a mark on the line, which was measured from the left endpoint. The result was a scale from 0 to 10 rounded to one decimal place. Management considers a score of at least 7 to be its benchmark; anything less will trigger some type of intervention.

You have been asked to analyze these data. Your analysis should include a descriptive summary of the data using appropriate charts and visual displays, recommendations on whether any interventions are necessary based on the benchmark targets, whether any differences exist between firm size and customer ratings, and between firm size and usage level. Use appropriate hypothesis tests to support your conclusions and summarize your findings in a formal report to the Vice President of Purchasing.

<sup>1</sup> Adapted from Hair, Anderson, Tatham, and Black in *Multivariate Analysis*, 5th ed., Prentice-Hall, 1998.

## APPENDIX 5.1

### Hypothesis-Testing Theory and Computation

In this appendix, we provide some additional details for two-sample hypothesis tests and the theory behind analysis of variance.

#### A. Two-Sample Tests for Differences in Means

The test statistics for the various two sample tests for means are summarized in the following text:

- **Population Variance Known.** When the population variance is known, the sampling distribution of the

difference in means is normal. Thus, the test statistic is a Z-value, and the critical value is found in the normal distribution table.

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \quad (5A.1)$$

- **Population Variance Unknown, Assumed Equal.** When the population variance is unknown, the sampling distribution of the difference in means has a t-distribution with  $n_1 + n_2 - 2$  degrees of freedom.

If we assume that the two population variances are equal, the test statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left( \frac{n_1 + n_2}{n_1 n_2} \right)}} \quad (5A.2)$$

- **Population Variance Unknown, Unequal.** When the population variance is unknown but unequal, the sampling distribution of the difference in means also has a *t*-distribution, but the degrees of freedom are more complicated to compute. The test statistic is:

$$t = (\bar{x}_1 - \bar{x}_2) / \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (5A.3)$$

and the degrees of freedom for the *t*-distribution,  $df^*$ , are computed as:

$$df^* = \frac{\left[ \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2}{\left[ \frac{(s_1^2/n_1)^2}{n_1 - 1} \right] + \left[ \frac{(s_2^2/n_2)^2}{n_2 - 1} \right]} \quad (5A.4)$$

## B. Two-Sample Test for Differences in Proportions

For a two-sample test for the difference in proportions, the test statistic is:

$$Z = \frac{p_1 - p_2}{\sqrt{\bar{p}(1 - \bar{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (5A.5)$$

where  $\bar{p}$  = number of successes in both samples/ $(n_1 + n_2)$ . This statistic has an approximate standard normal distribution; therefore, the critical value is chosen from a standard normal distribution.

## C. Test for Equality of Variances

In this chapter, we explained that the ratio of variances of two samples is an *F*-statistic with  $df_1 = n_1 - 1$  (numerator) and  $df_2 = n_2 - 1$  (denominator) degrees of freedom. If *F* exceeds the critical value  $F_{\alpha/2, df_1, df_2}$  of the *F*-distribution, then we reject  $H_0$ . As noted, this is really a two-tailed test, but as long as *F* is the ratio of the larger to the smaller variance, we can base our conclusion on the upper-tail value only. *PHStat* provides both the lower and upper critical values. Table A.4 in the appendix at the end of the book provides only upper-tail critical values, and the distribution is *not* symmetric. To find the lower-tail critical value, reverse the degrees of freedom, find the upper-tail value, and take the reciprocal. Thus, for the example in the chapter lower-tail critical value is approximately  $1/F_{0.025, 16, 40} = 1/2.18 = 0.46$  (we used the

numerator  $df = 15$  to give an approximation for the critical value because 16 is not included in Table A.4). This is close to the value shown in the *PHStat* results.

Suppose we took the ratio of the smaller variance to the larger one; that is,  $F = 253.87/282.47 = 0.899$ . In this case, the closer that *F* is to 0, the greater the likelihood that the population variances differ, so we need only compare *F* to the lower-tail critical value and reject if *F* is less than this value. Thus, in this case, with a numerator  $df = 16$  and the denominator  $df = 40$ , the lower-tail critical value is 0.464. Since  $F = 0.899 > 0.464$ , we cannot reject the null hypothesis and reach the same conclusion as before.

## D. Theory of Analysis of Variance

We define  $n_j$  as the number of observations in sample *j*. ANOVA examines the variation among and within the *m* groups or factor levels. Specifically, the total variation in the data is expressed as the variation between groups plus the variation within groups:

$$SST = SSB + SSW \quad (5A.6)$$

where

$SST$  = total variation in the data

$SSB$  = variation between groups

$SSW$  = variation within groups

We compute these terms using the following formulas.

$$SST = \sum_{j=1}^m \sum_{i=1}^{n_j} \left( X_{ij} - \bar{\bar{X}} \right)^2 \quad (5A.7)$$

$$SSB = \sum_{j=1}^m n_j \left( \bar{X}_j - \bar{\bar{X}} \right)^2 \quad (5A.8)$$

$$SSW = \sum_{j=1}^m \sum_{i=1}^{n_j} \left( X_{ij} - \bar{X}_j \right)^2 \quad (5A.9)$$

where

$\bar{\bar{X}}$  = total number of observations

$\bar{X}$  = overall or grand mean

$X_{ij}$  = *i*th observation in group *j*

$\bar{X}_j$  = sample mean of group *j*

From these formulas, you can see that each term is a “sum of squares” of elements of the data; hence, the notation “SST,” which can be thought of as the “Sum of Squares Total,”  $SSB$  is the “Sum of Squares Between” groups, and  $SSW$  is the “Sum of Squares Within” groups. Observe that if the means of each group are indeed equal ( $H_0$  is true), then the sample means of each group will be essentially the same as the overall mean, and  $SSB$  would be very small, and most of the total variation in the data is due to sampling variation within groups. The sum of squares is computed in the ANOVA section of the Excel output (Figure 5.18). By dividing the sums of squares by the degrees of freedom, we compute the mean squares (MS).

## SKILL-BUILDER EXERCISE 5.4

Verify the calculations for the sums of squares and mean square calculations for the ANOVA example in this chapter using formulas (5A.7), (5A.8), and (5A.9).

## APPENDIX 5.2

### Excel and PHStat Notes

#### A. One-Sample Test for the Mean, Sigma Unknown

From the *PHStat One-Sample Tests* menu, select *One-Sample Tests* then *t-test for the Mean, Sigma Unknown*. The dialog box, shown in Figure 5A.1, first asks you to input the value of the null hypothesis and significance level. For the Customer Support Survey example in the chapter, we would input 25 in the *Null Hypothesis* box and 0.05 for the *Level of Significance*. You may either specify the sample statistics or let the tool compute them from the data by specifying the range of the data. Under *Test Options* you may choose among a two-tailed test, upper one-tailed test, or lower one-tailed test. The *Title* box in the *Output Options* names the Excel worksheet.

#### B. One-Sample Test for Proportions

One-sample tests involving a proportion can be found by selecting the menu item *One-Sample Tests* and choosing *Z-Test*.

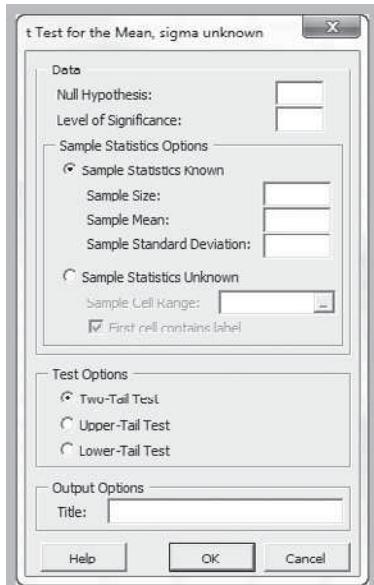


FIGURE 5A.1 PHStat Dialog for t-Test for the Mean, Sigma Unknown

for the Proportion. Figure 5A.2 shows the dialog box. The tool requires you to enter the value for the null hypothesis, significance level, number of items of interest, sample size, and the type of test. The sample proportion is the *Number of Items of Interest/Sample Size*. Note that you cannot enter the proportion as a number itself because the test depends on the sample size which must be specified. Thus, if you only know the sample proportion and the sample size, you must convert it into the Number of Items of Interest by multiplying the sample proportion by the sample size.

#### C. Using Two-Sample t-Test Tools

To test differences between the means of two populations, three types of hypothesis tests are available.

1. *Population variance is known.* Both Excel and *PHStat* have tools for a two-sample test for means with known

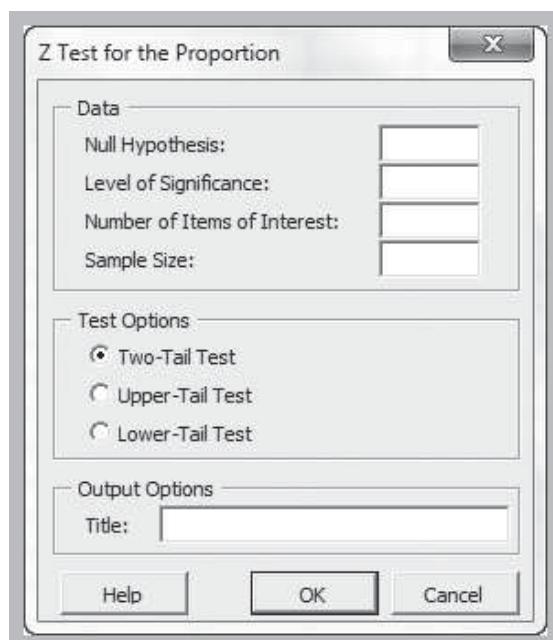


FIGURE 5A.2 PHStat Dialog for One-Sample Test for Proportions

population variances. The dialogs are shown in Figures 5A.3 and 5A.4. In Excel, choose *Z-Test: Two Sample for Means* from the *Data Analysis* menu. The dialog prompts you for the range of each variable, hypothesized mean difference, known variances for each variable, whether the ranges have labels, and the level of significance ( $\alpha$ ). If you leave the box *Hypothesized Mean Difference* blank or enter 0, the test is for equality of means. However, the tool allows you to specify a value  $d$  to test the hypothesis  $H_0: \mu_1 - \mu_2 = d$ . In *PHStat*, from the *Two-Sample Test* menu, choose *Z-Test for Differences in Two Means*. However, *PHStat* requires you to enter a value for the sample mean instead of calculating it from the data range. You must also enter the hypothesized difference and level of significance. For *PHStat*, you may specify the type of test (two-tail, upper-tail, or lower-tail).

- Population variance is unknown and unequal.* Excel has a tool for a two-sample test for means with unequal variances. From the *Data Analysis* menu, choose *t-test: Two-Sample Assuming Unequal Variances*. Data input is straightforward: The dialog box is similar to Figure 5A.3, and prompts you to enter the ranges for each sample, whether the ranges have labels, the hypothesized mean difference, and the level of significance ( $\alpha$ ), which defaults to 0.05. As in the Z-test, the tool also allows you to specify a value  $d$  to test the hypothesis  $H_0: \mu_1 - \mu_2 = d$ . In *PHStat*, choose *Separate-Variance t Test* from the *Two Sample Tests* menu. The dialog is identical to Figure 5A.4.
- Population variance unknown but assumed equal.* In Excel, choose *t-test: Two-Sample Assuming Equal Variance*; in *PHStat*, choose *Pooled Variance t-test*. These tools conduct the test for equal variances only. The dialogs are similar to those in Figures 5A.3 and 5A.4. *PHStat* also provides an option to compute a confidence interval.

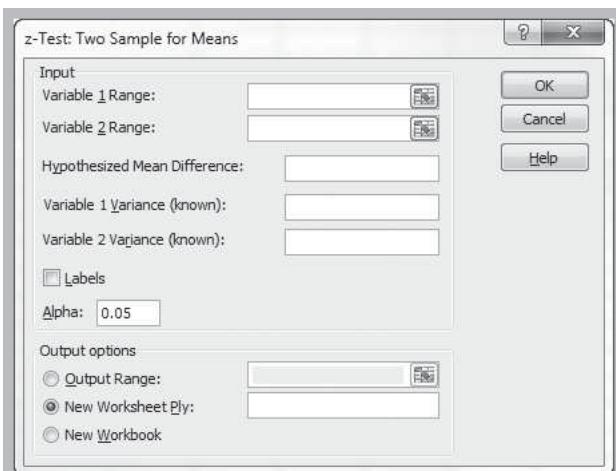


FIGURE 5A.3 Excel Dialog for Z-Test Two Sample for Means

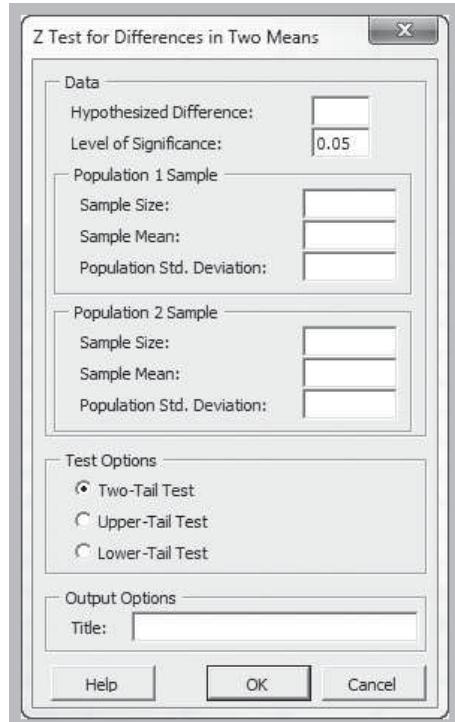


FIGURE 5A.4 PHStat Dialog for Z-Test for Differences in Two Means

Although the *PHStat* output is straightforward, you must be *very careful* in interpreting the information from Excel. If  $t_{\text{Stat}}$  is negative,  $P(T \leq t)$  provides the correct  $p$ -value for a lower-tail test; however, for an upper-tail test, you must subtract this number from 1.0 to get the correct  $p$ -value. If  $t_{\text{Stat}}$  is nonnegative, then  $P(T \leq t)$  provides the correct  $p$ -value for an upper-tail test; consequently, for a lower-tail test, you must subtract this number from 1.0 to get the correct  $p$ -value. Also, for a lower-tail test, you must change the sign on  $t_{\text{Critical one-tail}}$ .

## D. Testing for Equality of Variances

The *PHStat* tool *F-test for Differences in Two Variances* requires that you enter the level of significance and the sample size and sample standard deviation for each sample, which you must compute from the data. As with all *PHStat* tools, you can specify whether the test is two tail, upper tail, or lower tail.

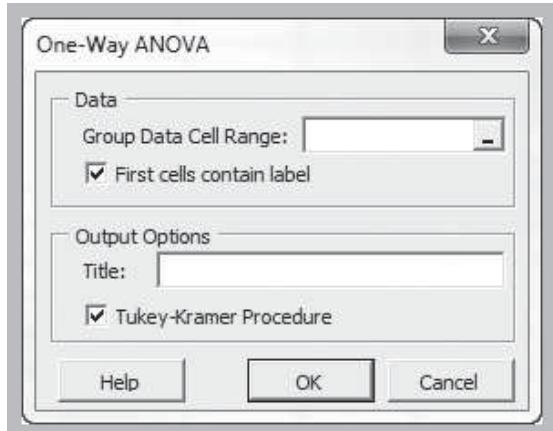
To use the Excel tool, select *F-test for Equality of Variances* from the *Data Analysis* menu. Specify the *Variable 1 Range* and the *Variable 2 Range* for both data sets and a value for the significance level. Note, however, that this tool provides results for only a one-tailed test. Thus, for a two-tailed test of equality of variances, you must use  $\alpha/2$  for the significance level in the Excel dialog box. If the variance of variable 1 is greater than the variance of variable 2, the output will specify the upper tail; otherwise, you obtain the lower-tail information.

For ease of interpretation, we suggest that you ensure that variable 1 has the larger variance. See the example for how to properly interpret the Excel results.

## E. Single-Factor Analysis of Variance

To use ANOVA to test for difference in sample means in Excel, select *ANOVA: Single Factor* from the *Data Analysis* tools. This displays the dialog box shown in Figure 5A.5. You need only specify the input range of the data (in contiguous columns) and whether it is stored in rows or columns (i.e., whether each factor level or group is a row or column in the range). You must also specify the level of significance (alpha level) and the output options.

In *PHStat*, choose *One-Way ANOVA* from the *Multiple Sample Tests* menu. The dialog is shown in Figure 5A.6. To invoke the Tukey-Kramer Procedure, check the box in the dialog. *PHStat* prompts you to enter a value for “Q-statistic” manually in the worksheet before displaying the conclusions of the test. This value is based on the numerator  $df$  and denominator  $df$  shown in the output and not the degrees

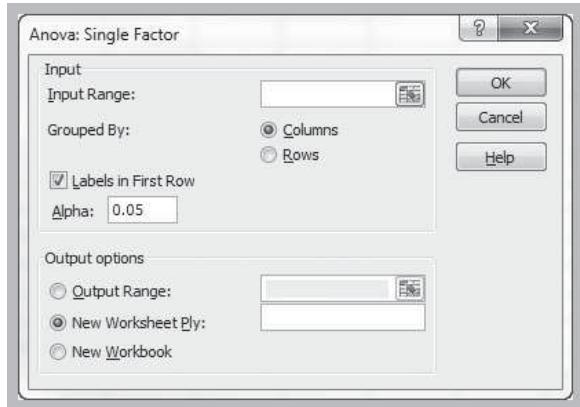


**FIGURE 5A.6** *PHStat One-Way ANOVA* Dialog

of freedom for the ANOVA test, which are different. The Q-statistic may be found in Table A.5 in the appendix at the back of this book.

## F. Chi-Square Test for Independence

From the *PHStat Multiple Sample Tests* menu, select *Chi-Square Test*. In the dialog box, enter the level of significance, number of rows, and number of columns in the contingency table. *PHStat* will create a worksheet in which you will need to enter the data for the observed frequencies. You may also customize the names of the rows and columns for your specific application. After you complete the table, the calculations are performed automatically. *PHStat* includes an optional check box for the *Marascuilo procedure*. This makes comparisons among all pairs of groups to determine if the proportions are significantly different, similar to the way the Tukey-Kramer procedure does for ANOVA. This is useful if there are more than two levels of a category so you can identify which pairs are significantly different from each other.



**FIGURE 5A.5** *ANOVA: Single-Factor* Dialog

## *Chapter 6*

# Regression Analysis

- INTRODUCTION 197
- SIMPLE LINEAR REGRESSION 198
  - Least-Squares Regression 200
  - A Practical Application of Simple Linear Regression to Investment Risk 202
- SIMPLE LINEAR REGRESSION IN EXCEL 203
  - Regression Statistics 204
  - Regression as Analysis of Variance 205
  - Testing Hypotheses for Regression Coefficients 205
  - Confidence Intervals for Regression Coefficients 206
  - Confidence and Prediction Intervals for X-Values 206
- RESIDUAL ANALYSIS AND REGRESSION ASSUMPTIONS 206
  - Standard Residuals 208
  - Checking Assumptions 208
- MULTIPLE LINEAR REGRESSION 210
  - Interpreting Results from Multiple Linear Regression 212
  - Correlation and Multicollinearity 212
- BUILDING GOOD REGRESSION MODELS 214
  - Stepwise Regression 217
  - Best-Subsets Regression 217
  - The Art of Model Building in Regression 218
- REGRESSION WITH CATEGORICAL INDEPENDENT VARIABLES 220
  - Categorical Variables with More Than Two Levels 223
- REGRESSION MODELS WITH NONLINEAR TERMS 225
- BASIC CONCEPTS REVIEW QUESTIONS 228
- PROBLEMS AND APPLICATIONS 228
- CASE: HATCO 231
- APPENDIX 6.1: REGRESSION THEORY AND COMPUTATION 231
  - A. Regression as Analysis of Variance 231
  - B. Standard Error of the Estimate 233
  - C. Adjusted  $R^2$  233
  - D. Confidence Intervals 233
  - E. Prediction Intervals 233

## ■ APPENDIX 6.2: EXCEL AND PHSTAT NOTES 233

- A. Using the Trendline Option 233
- B. Using Regression Tools 233
- C. Using the Correlation Tool 235
- D. Stepwise Regression 235
- E. Best-Subsets Regression 236

## INTRODUCTION

In Chapter 2, we discussed statistical relationships and introduced correlation as a measure of the strength of a linear relationship between two numerical variables. Decision makers are often interested in predicting the value of a dependent variable from the value of one or more independent, or explanatory, variables. For example, the market value of a house is typically related to its size. In the Excel file *Home Market Value* (see Figure 6.1), data obtained from a county auditor provides information about the age, square footage, and current market value of houses in a particular subdivision. We might wish to predict the market value as a function of the size of the home. We might use a simple linear equation:

$$\text{Market Value} = a + b \times \text{Square Feet}$$

Market Value is the dependent variable, and Square Feet is the independent variable. Using the data, we could determine values of the constants,  $a$  and  $b$ , that would best explain Market Value as a function of size.

As another example, many colleges try to predict students' performance as a function of several characteristics. In the Excel file *Colleges and Universities* (see Figure 6.2), we might wish to predict the graduation rate as a function of the other variables—median SAT, acceptance rate,

	A	B	C
1	Home Market Value		
2			
3	House Age	Square Feet	Market Value
4	33	1,812	\$90,000.00
5	32	1,914	\$104,400.00
6	32	1,842	\$93,300.00
7	33	1,812	\$91,000.00
8	32	1,836	\$101,900.00
9	33	2,028	\$108,500.00
10	32	1,732	\$87,600.00

**FIGURE 6.1** Portion of *Home Market Value*

	A	B	C	D	E	F	G
1	Colleges and Universities						
2							
3	School	Type	Median SAT	Acceptance Rate	Expenditures/Student	Top 10% HS	Graduation %
4	Amherst	Lib Arts	1315	22%	\$ 26,636	85	93
5	Barnard	Lib Arts	1220	53%	\$ 17,653	69	80
6	Bates	Lib Arts	1240	36%	\$ 17,554	58	88
7	Berkeley	University	1176	37%	\$ 23,665	95	68
8	Bowdoin	Lib Arts	1300	24%	\$ 25,703	78	90

**FIGURE 6.2** Portion of Excel File *Colleges and Universities*

expenditures/student, and the top 10% of their high school class. We might use the following equation:

$$\begin{aligned}\text{Graduation\%} &= a + b \times \text{Median SAT} + c \times \text{Acceptance Rate} \\ &\quad + d \times \text{Expenditures/Student} + e \times \text{Top 10\% HS}\end{aligned}$$

Here, the graduation rate would be the dependent variable, and the remaining variables would be the independent variables. The constants  $a$ ,  $b$ ,  $c$ ,  $d$ , and  $e$  would be determined to provide the best equation that would predict the graduation rate for these independent variables.

**Regression analysis** is a tool for building statistical models that characterize relationships among a dependent variable and one or more independent variables, all of which are numerical. A regression model that involves a single independent variable is called *simple regression*. A regression model that involves two or more independent variables is called *multiple regression*.

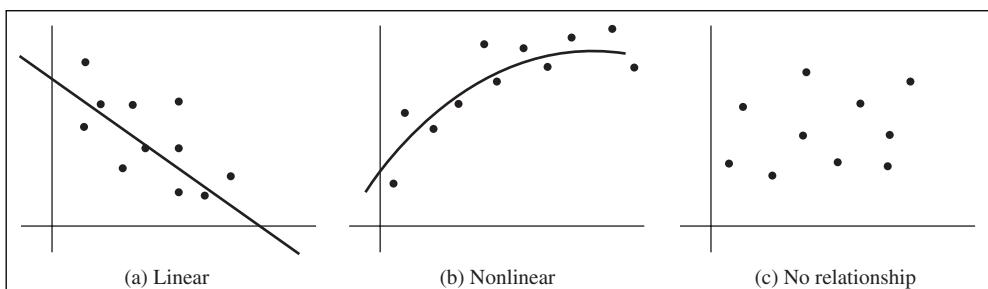
Two broad categories of regression models are used often in business settings: (1) regression models of cross-sectional data, such as those just described, and (2) regression models of time-series data, in which the independent variables are time or some function of time and the focus is on predicting the future. Time-series regression is an important tool in *forecasting*, which will be discussed in Chapter 7.

In this chapter, we describe how to develop and analyze both simple and multiple regression models. Our principal focus is to gain a basic understanding of the assumptions of regression models, statistical issues associated with interpreting regression results, and practical issues in using regression as a tool for making and evaluating decisions.

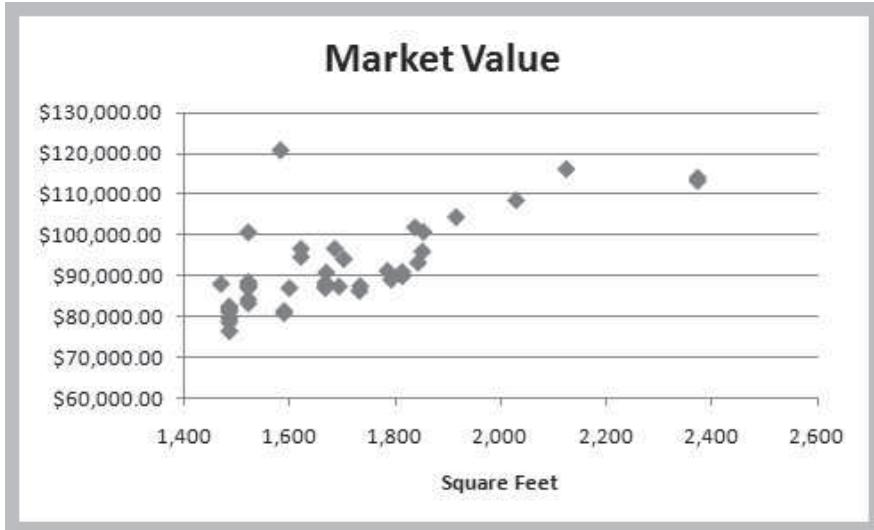
## SIMPLE LINEAR REGRESSION

The simplest type of regression model involves one independent variable,  $X$ , and one dependent variable,  $Y$ . The relationship between two variables can assume many forms, as illustrated in Figure 6.3. The relationship may be linear or nonlinear, or there may be no relationship at all. We will focus our discussion on linear regression models; therefore, the first thing to do is to verify that the relationship is linear as in Figure 6.3(a). If the relationship is clearly nonlinear as in Figure 6.3(b), then alternate approaches must be used, and if no relationship is evident as in Figure 6.3(c), then it is pointless to even consider developing a regression model.

To determine if a linear relationship exists between the variables, we recommend that you create a scatter chart first (see Chapter 2). For example, Figure 6.4 shows a scatter chart of the market value in relation to the size of the home in the Excel file *Home Market Value*. The independent variable,  $X$ , is the number of square feet, and the dependent variable,  $Y$ , is the market value. In general, we see that higher market values are associated with larger home sizes. The relationship is clearly not perfect; but our goal is to build a good model to estimate the market value as a function of the number of square feet by finding the best-fitting straight line that represents the data.



**FIGURE 6.3** Examples of Variable Relationships



**FIGURE 6.4** Scatter Chart of Market Value Versus Home Size

### SKILL-BUILDER EXERCISE 6.1

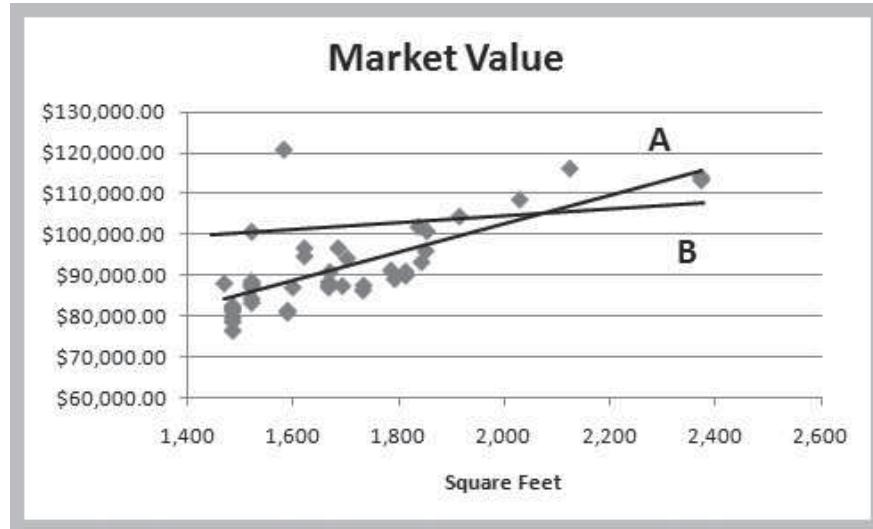
Using the data in the Excel in the file *Home Market Value*, construct the scatter chart in Figure 6.4 and add a linear trendline.

The value of regression can be explained as follows. Suppose we wanted to estimate the home market value for any home in the population from which the sample data were gathered. If all we knew were the market values, then the best estimate of the market value for any home would simply be the sample mean, which is \$92,069. Thus, no matter if the house has 1,500 square feet, or 2,200 square feet, the best estimate of market value would still be \$92,069. Because the market values vary from about \$75,000 to over \$120,000, there is quite a bit of uncertainty in using the mean as the estimate. However, from the scatter chart, we see that larger homes tend to have higher market values. Therefore, if we know that a home has 2,200 square feet, we would expect the market value estimate to be higher, probably over \$100,000, than one that has only 1,500 square feet. By knowing the home size, we can reduce some of the variation in our estimation than simply using the mean.

In regression analysis, we assume that the sample data are drawn from some unknown population for each value of  $X$ . For example, in the *Home Market Value* data, the first and fourth observations come from a population of homes having 1,812 square feet; the second observation comes from a population of homes having 1,914 square feet; and so on. Because we are assuming that a linear relationship exists, the expected value of  $Y$  is  $\beta_0 + \beta_1 X$  for each value of  $X$ . The coefficients  $\beta_0$  and  $\beta_1$  are population parameters that represent the intercept and slope, respectively, of the population from which a sample of observations is taken. The intercept is the mean value of  $Y$  when  $X = 0$ , and the slope is the change in the mean value of  $Y$  as  $X$  changes by one unit.

Thus, for a specific value of  $X$ , we have many possible values of  $Y$  that vary around the mean. To account for this, we add an error term,  $\varepsilon$  (the Greek letter epsilon), to the mean. This defines a simple linear regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (6.1)$$



**FIGURE 6.5** Two Possible Regression Lines

Because we don't know the entire population, we don't know the true values of  $\beta_0$  and  $\beta_1$ . In practice, we must estimate these as best as we can from the sample data. Using the *Home Market Value* data, note that each point in Figure 6.4 represents a paired observation of the square footage of the house and its market value. If we draw a straight line through the data, some of the points will fall above the line, some below it, and a few might fall on the line itself. Figure 6.5 shows two possible straight lines to represent the relationship between  $X$  and  $Y$ . Clearly, you would choose A as the better fitting line over B because all the points are "closer" to the line and the line appears to be in the "middle" of the data. The only difference between the lines is the value of the slope and intercept; thus, we seek to determine the values of the slope and intercept that provide the best-fitting line. We do this using a technique called least-squares regression.

### Least-Squares Regression

Define  $b_0$  and  $b_1$  to be estimates of  $\beta_0$  and  $\beta_1$ . Thus, the estimated simple linear regression equation is:

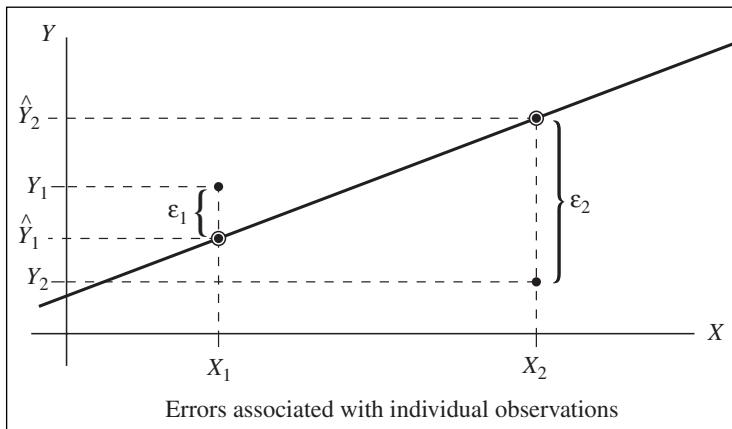
$$\hat{Y} = b_0 + b_1 X \quad (6.2)$$

Let  $(X_i, Y_i)$  represent the  $i$ th observation. When the value of the independent variable is  $X_i$ , then  $\hat{Y}_i = b_0 + b_1 X_i$  is the estimated value of  $Y$  for  $X_i$ .

One way to quantify the relationship between each point and the estimated regression equation is to measure the vertical distance between them,  $Y_i - \hat{Y}_i$  (see Figure 6.6). We can think of these differences as the observed errors (often called **residuals**),  $e_i$ , associated with estimating the value of the dependent variable using the regression line. The best-fitting line should minimize some measure of these errors. Because some errors will be negative and others positive, we might take their absolute value, or simply square them. Mathematically, it is easier to work with the squares of the errors.

Adding the squares of the errors, we obtain the following function:

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (6.3)$$



**FIGURE 6.6** Measuring the Errors in a Regression Model

If we can find the best values of the slope and intercept that minimizes the sum of squares (hence the name, “least squares”) of the observed errors,  $e_i$ , we will have found the best-fitting regression line. Note that  $X_i$  and  $Y_i$  are the known data and that  $b_0$  and  $b_1$  are unknowns in equation 6.2). Using calculus, we can show that the solution that minimizes the sum of squares of the observed errors is:

$$b_1 = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2} \quad (6.4)$$

$$b_0 = \bar{Y} - b_1 \bar{X} \quad (6.5)$$

Although the calculations for the least-squares coefficients appear to be somewhat complicated, they can easily be performed on an Excel spreadsheet. Even better, Excel has built-in capabilities to do this. Table 6.1 summarizes useful spreadsheet functions and tools that we will use in this chapter. For example, you may use the functions INTERCEPT

**TABLE 6.1** Spreadsheet Functions and Tools for Regression Analysis

	Description
<b>Excel Function</b>	
INTERCEPT( <i>known_y's, known_x's</i> )	Calculates the intercept for a least-squares regression line
SLOPE( <i>known_y's, known_x's</i> )	Calculates the slope of a linear regression line
TREND( <i>known_y's, known_x's, new_x's</i> )	Computes the value on a linear regression line for specified values of the independent variable
<b>Analysis Toolpak Tools</b>	
Regression	Performs linear regression using least squares
<b>PHStat Add-In</b>	
Simple Linear Regression	Generates a simple linear regression analysis
Multiple Regression	Generates a multiple linear regression analysis
Best Subsets	Generates a best-subsets regression analysis
Stepwise Regression	Generates a stepwise regression analysis

and SLOPE to find the least-squares coefficients  $b_0$  and  $b_1$ . Using these functions for the *Home Market Value* data,  $b_0 = 32673$  and the slope,  $b_1 = 35.036$ . Thus, the least-squares regression equation is

$$\hat{Y} = 32,673 + 35.036X$$

The slope tells us that for every additional square foot, the market value increases by \$35.036. Such a model can be used to estimate the market value for data not present in the sample. Thus, for a house with 1,750 square feet, the estimated market value is  $32,673 + 35.036(1,750) = \$93,986$ . This can also be found by using the Excel function TREND(*known\_y's*, *known\_x's*, *new\_x's*). Thus, using the *Home Market Value* data, we could estimate the value of a house with 1,750 square feet using the function =TREND(C4:C45, B4:B45, 1,750).

### SKILL-BUILDER EXERCISE 6.2

For the *Home Market Value* data, use formulas (6.4) and (6.5) on a spreadsheet to calculate the least-squares regression coefficients.



Spreadsheet Note

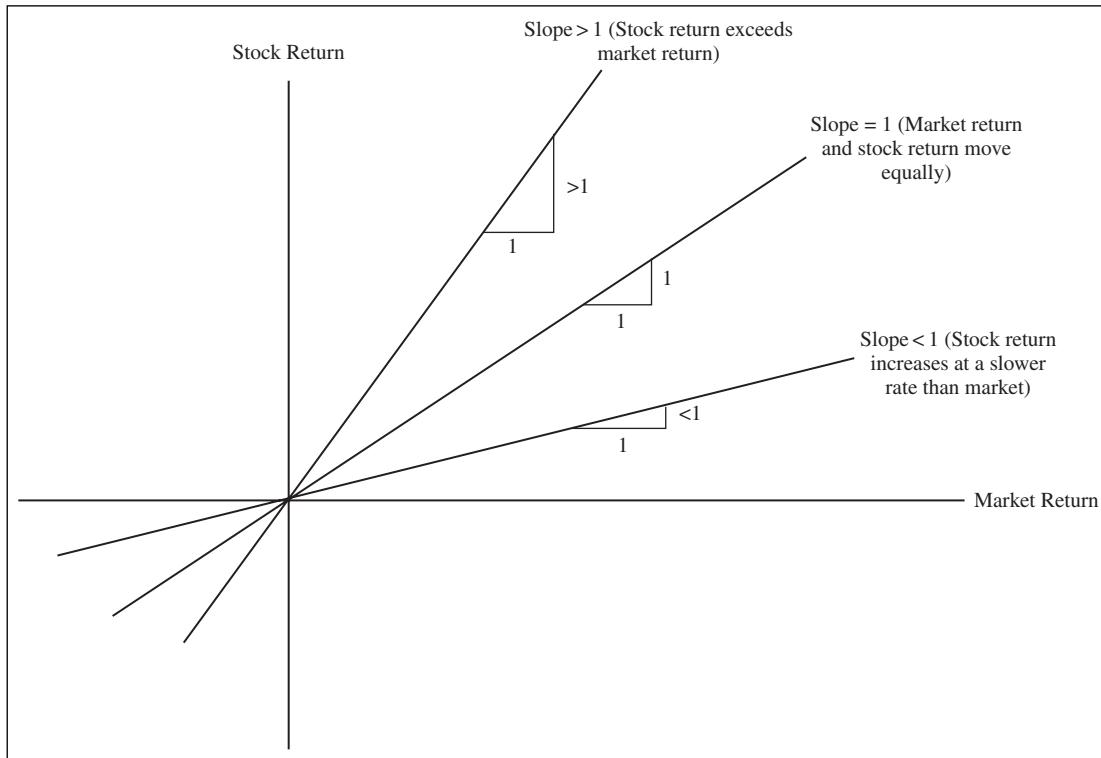
We may also determine and plot the least-squares regression line from the data on a scatter chart using the *Trendline* option (see the note in Appendix 6.2A, “Using the Trendline Option”). For the *Home Market Value* data, line A in Figure 6.5 is actually the least-squares regression line that was found in this fashion.

*One important caution:* it is dangerous to extrapolate a regression model outside the ranges covered by the observations. For instance, if you wanted to predict the market value of a house that has 3,000 square feet, the results may or may not be accurate, because the regression model estimates did not use any observations larger than 2,400 square feet. We cannot be sure that a linear extrapolation will hold and should not use the model to make such predictions.

## A Practical Application of Simple Regression to Investment Risk

Investing in the stock market is highly attractive to everyone. However, stock investments do carry an element of risk. Risk associated with an individual stock can be measured in two ways. The first is **systematic risk**, which is the variation in stock price explained by the market—as the market moves up or down, the stock tends to move in the same direction. The Standard & Poor’s (S&P’s) 500 index is most commonly used as a measure of the market. For example, we generally see that stocks of consumer products companies are highly correlated with the S&P index, while utility stocks generally show less correlation with the market. The second type of risk is called **specific risk** and is the variation that is due to other factors, such as the earnings potential of the firm, acquisition strategies, and so on. Specific risk is measured by the standard error of the estimate.

Systematic risk is characterized by a measure called *beta*. A beta value equal to 1.0 means that the specific stock will match market movements, a beta less than 1.0 indicates that the stock is less volatile than the market, and a beta greater than 1.0 indicates that the stock has greater variance than the market. Thus, stocks with large beta values are riskier than those with lower beta values. Beta values can be calculated by developing a regression model of a particular stock’s returns (the dependent variable) against the average market returns (the independent variable). The slope of the regression line is the beta risk. This can be explained through the chart in Figure 6.7. If we plot the daily market returns



**FIGURE 6.7** Illustration of Beta Risk

against the returns of the individual stock and find the regression line, we would observe that if the slope equals 1, the stock changes at the same rate as the market. However, if the stock price changes are less than the market changes, the slope of the regression line would be less than 1, while the slope would be greater than 1 when the stock price changes exceed that of the market. A negative slope would indicate a stock that moves in the opposite direction to the market (e.g., if the market goes up, the stock price goes down).

For example, suppose that we collected data on a specific stock and the S&P 500 index over an extended period of time and computed the percentage change from one day to the next for both variables. If there appears to be a linear relationship between the change in the S&P 500 index and change in the stock price, we could fit the model:

$$\text{Daily Change in Stock Price} = \beta_0 + \beta_1 \text{S\&P Change}$$

Suppose that the least-squares fit resulted in the model:

$$\text{Daily Change in Stock Price} = 0.002 + 1.205 \text{S\&P Change}$$

The slope of the regression line, 1.205, is the beta risk of the stock. This indicates that the stock is more risky than the average S&P 500 stock.

## SIMPLE LINEAR REGRESSION IN EXCEL

Regression analysis software tools available in Excel and *PHStat* provide a variety of information concerning regression analysis (see Appendix 6.2B, “Using Regression Tools”). In this section, we will describe how to interpret the results from these tools using the *Home Market Value* example.



Spreadsheet Note

A	B	C	D	E	F	G
1 Home Market Value						
2						
3 <i>Regression Statistics</i>						
4 Multiple R	0.731255223					
5 R Square	0.534734202					
6 Adjusted R Square	0.523102557					
7 Standard Error	7287.722712					
8 Observations	42					
9						
10 ANOVA						
11	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12 Regression	1	2441633669	2441633669	45.97236277	3.79802E-08	
13 Residual	40	2124436093	53110902.32			
14 Total	41	4566069762				
15						
16	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17 Intercept	32673.2199	8831.950745	3.699434116	0.000649604	14823.18178	50523.25802
18 Square Feet	35.03637258	5.16738385	6.780292234	3.79802E-08	24.59270036	45.48004481

**FIGURE 6.8** Basic Regression Analysis Output for *Home Market Value* Example

Figure 6.8 shows the basic regression analysis output provided by both the Excel (by default) and *PHStat* tools. The least-squares estimates of the slope and intercept are found in the *Coefficients* column in the bottom section of the output. We see that the intercept is 32,673, and the slope (coefficient of the independent variable, Square Feet) is 35.036.

### SKILL-BUILDER EXERCISE 6.3

Use the Excel or *PHStat* regression tool for the data in the *Home Market Value* file to confirm the results in Figure 6.8.

## Regression Statistics

The regression analysis tools provide some useful regression statistics (see rows 3–8 in Figure 6.8). *Multiple R* is another name for the sample correlation coefficient, *r* that was introduced in Chapter 2. Values of *r* range from  $-1$  to  $+1$ , where the sign is determined by the sign of the slope of the regression line. A *Multiple R* value of  $+1$  indicates perfect positive correlation, that is, as the independent variable increases, the dependent variable also does; a value of  $-1$  indicates perfect negative correlation—as *X* increases, *Y* decreases. A value of  $0$  indicates no correlation.

*R Square* ( $R^2$ ) is called the **coefficient of determination**, and gives the proportion of variation in the dependent variable that is explained by the independent variable of the regression model.  $R^2$  is a measure of the “fit” of the regression line to the data. The value of  $R^2$  will be between  $0$  and  $1$ . A value of  $1.0$  indicates a perfect fit and all data points would lie on the regression line, while a value of  $0$  indicates that no relationship exists. For the market value data,  $R^2 = 0.5347$ . This means that approximately  $53\%$  of the variation in Market Value is explained by Square Feet. The remaining variation is due to other factors that were not included in the model. Although we would like high values of  $R^2$ , it is difficult to specify a “good” value that signifies a strong relationship because this depends on the application. For example, in scientific applications such as calibrating physical measurement equipment,  $R^2$  values close to  $1$  would be expected;

in marketing research studies, an  $R^2$  of 0.6 or more is considered very good; however, in many social science applications, values in the neighborhood of 0.3 might be considered acceptable.

*Adjusted R-Square* is a statistic that modifies the value of  $R^2$  by incorporating the sample size and the number of explanatory variables in the model. Although it does not give the actual percentage of variation explained by the model as  $R^2$  does, it is useful when comparing this model with other models that include additional explanatory variables. We will discuss it in more detail in the context of multiple linear regression later in this chapter.

*Standard Error* in the Excel output is the variability of the observed  $Y$ -values from the predicted values, ( $\hat{Y}$ ). This is formally called the **standard error of the estimate**,  $S_{yx}$ . If the data are clustered close to the regression line, then the standard error will be small; the more scattered the data are, the larger the standard error. For this example, the standard error of the estimate is \$7,287.72. The computed standard deviation of the market values is \$10,553. You can see that the variation around the regression line is less than the variation around the mean because the independent variable explains some of the variation.

## Regression as Analysis of Variance

In Chapter 5, we introduced analysis of variance (ANOVA), which conducts an  $F$ -test to determine whether variation due to a particular factor, such as the differences in sample means, is significantly larger than that due to error. ANOVA is commonly applied to regression to test for *significance of regression*. This is shown in rows 10–14 in Figure 6.8. For a simple linear regression model, **significance of regression** is simply a hypothesis test of whether the regression coefficient  $\beta_1$  (slope of the independent variable) is 0:

$$\begin{aligned} H_0: \beta_1 &= 0 \\ H_1: \beta_1 &\neq 0 \end{aligned}$$

If we reject the null hypothesis, then we may conclude that the slope of the independent variable is not 0, and therefore is statistically significant in the sense that it explains some of the variation of the dependent variable around the mean. The value of *Significance F* is the  $p$ -value for the  $F$ -test; if this is less than the level of significance, we would reject the null hypothesis. In this example the  $p$ -value is essentially 0 ( $3.798 \times 10^{-8}$ ) and would lead us to conclude that home size is a significant variable in explaining the variation in market value.

## Testing Hypotheses for Regression Coefficients

Rows 17 and 18 of the Excel output, in addition to specifying the least-squares coefficients, provide additional information for testing hypotheses associated with the intercept and slope. Specifically, we may test the null hypothesis that  $\beta_0$  or  $\beta_1$  equals 0. Usually, it makes little sense to test or interpret the hypothesis that  $\beta_0 = 0$  unless the intercept has a significant physical meaning in the context of the application. Testing the null hypothesis  $H_0: \beta_1 = 0$  is the same as the significance of regression test that we described earlier.

The  $t$ -test for the slope is similar to the one-sample test for the mean that we described in Chapter 5. The test statistic is

$$t = \frac{b_1 - 0}{\text{Standard Error}} \tag{6.6}$$

and is given in the column labeled *t Stat* in the Excel output. For example, the test statistic associated with Square Feet is  $35.036 / 5.167 = 6.780$ . Although the critical value of the

*t*-distribution is not provided, the output does provide the *p*-value for the test. Note that the *p*-value associated with the test for the slope coefficient, Square Feet, is equal to the *Significance F* value. This will always be true for a regression model with one independent variable because it is the only explanatory variable. However, as we shall see, this will not be the case for multiple regression models. The small *p*-value leads us to reject the null hypothesis.

### Confidence Intervals for Regression Coefficients

Confidence intervals (*Lower 95%* and *Upper 95%* values in the output) provide information about the unknown values of the true regression coefficients, accounting for sampling error. For example, a 95% confidence interval for the slope is [24.59, 45.48]. Similarly, a 95% confidence interval for the intercept is [14,823, 50,523]. Although the regression model is  $\hat{Y} = 32,673 + 35.036X$ , the confidence intervals suggest a bit of uncertainty about predictions using the model. Thus, although we estimated that a house with 1,750 square feet has a market value of  $32,673 + 35.036(1,750) = \$93,986$ , if the true population parameters are at the extremes of the confidence intervals, the estimate might be as low as  $14,823 + 24.59(1,750) = \$57,855$  or as high as  $50,523 + 45.48(1,750) = \$130,113$ . Tighter confidence intervals provide more accuracy in our predictions.

### Confidence and Prediction Intervals for X-Values

Recall that we noted that for each *X*, there is a population of *Y*-values from which the sample data come. Therefore, when we estimate the value of *Y* for a given value of *X* (such as the market value of a home with *X* = 1,750 square feet), we only obtain a point estimate. We may develop a confidence interval for the mean value of *Y* for a specific value of *X*, and a prediction interval for an individual response *Y* for a specific value of *X*. It is important to understand the difference in these intervals. A confidence interval for the mean value of *Y* quantifies the uncertainty about the *mean value* of the dependent variable (e.g., the mean market value of the population of homes having 1,750 square feet); a prediction interval is an interval that quantifies the uncertainty in the dependent variable for a *single future observation* (e.g., the market value of an individual home with 1,750 square feet).

*PHStat* provides these interval estimates by checking the appropriate box in the regression tool and entering the value of the independent variable. Figure 6.9 shows the results for the *Home Market Value* data. A 95% confidence interval for the mean market value is [\$91,643, \$96,330], while a 95% prediction interval for the market value of an individual home is [\$79,073, \$108,901]. Notice that the prediction interval is wider than a confidence interval because the variance of the mean value is smaller than the variance of individual values (recall our discussion of the difference between the standard error of the mean for the sampling distribution and the standard deviation of individual observations in Chapter 4).

## RESIDUAL ANALYSIS AND REGRESSION ASSUMPTIONS

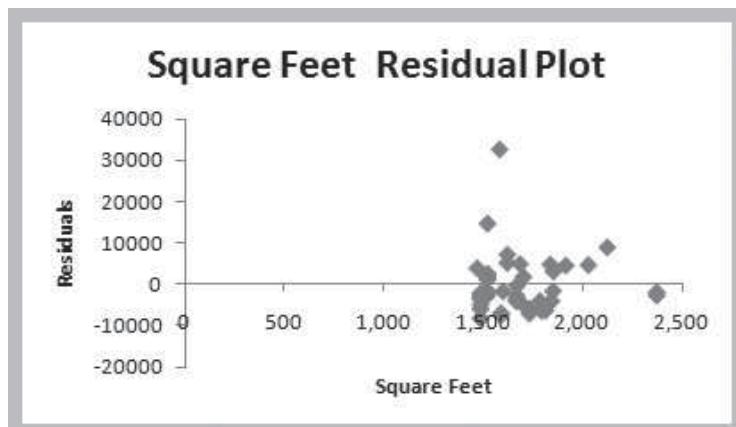
Recall that residuals are the observed errors, which are the differences between the actual values and the estimated values of the dependent variable using the regression equation. Figure 6.10 shows a portion of the residual table from the Excel tool. The residual output includes, for each observation, the predicted value using the estimated regression equation and the residual. For example, the first home has a market value of \$90,000 and the regression model predicts \$96,159.13. Thus, the residual is  $90,000 - 96,159.13 = -\$6,159.13$ . Both Excel and *PHStat* provide options for displaying a plot of the residuals. Figure 6.11 shows the residual plot from the Excel tool.

A	B	
1	<b>Confidence Interval Estimate</b>	
2		
3	<b>Data</b>	
4	X Value	
5	1750	
6	Confidence Level	
7	95%	
8	Intermediate Calculations	
9	Sample Size	42
10	Degrees of Freedom	40
11	tValue	2.021075
12	XBar, Sample Mean of X	1695.262
13	Sum of Squared Differences from XBar	1989034
14	Standard Error of the Estimate	7287.723
15	h Statistic	0.025316
16	Predicted Y (YHat)	93986.87
17	<b>For Average Y</b>	
18	Interval Half Width	2343.533
19	Confidence Interval Lower Limit	91643.3
20	Confidence Interval Upper Limit	96330.4
21		
22	<b>For Individual Response Y</b>	
23	Interval Half Width	14914.31
24	Prediction Interval Lower Limit	79072.6
25	Prediction Interval Upper Limit	108901

**FIGURE 6.9** PHStat Interval Estimates

A	B	C	D	
22	<b>RESIDUAL OUTPUT</b>			
23				
24	Observation	Predicted Market Value	Residuals	Standard Residuals
25	1	96159.12702	-6159.127018	-0.855636403
26	2	99732.83702	4667.162978	0.64837022
27	3	97210.2182	-3910.218196	-0.543214164
28	4	96159.12702	-5159.127018	-0.716714702
29	5	96999.99996	4900.00004	0.680716341

**FIGURE 6.10** Portion of Residual Output



**FIGURE 6.11** Residual Plot for Square Feet

## Standard Residuals

Standard residuals are residuals divided by their standard deviation; this standard deviation is calculated as  $S_{YX}\sqrt{1 - h_i}$ , where  $h_i$  is computed using equation (6A.5) in Appendix 6.1. (*PHStat* does not provide standard residuals). Standard residuals describe how far each residual is from its mean in units of standard deviations (similar to a  $z$ -value for a standard normal distribution). For example, the first observation is about 0.85 standard deviations below the regression line. Standard residuals are useful in checking assumptions underlying regression analysis, which we will address shortly and to detect outliers that may bias the results.

Recall from Chapter 2 that an outlier is an extreme value that is different from the rest of the data. A single outlier can make a significant difference in the regression equation, changing the slope and intercept, and hence, how it would be interpreted and used in practice. Some consider a standardized residual outside of  $\pm 2$  standard deviations as an outlier. A more conservative rule of thumb would be to consider outliers outside of a  $\pm 3$  standard error range. (Commercial software have more sophisticated techniques for identifying outliers.)

For example, if you look back at Figure 6.4, you will notice that one point (with a market value of around \$120,000) appears to be quite different from the rest of the data. In fact, the standard residual for this observation is 4.53 (you can see that it appears to be an outlier in the residual plot in Figure 6.11). You might question whether this observation belongs in the data, because the house has large value despite a relatively small size. A reason might be an outdoor pool or an unusually large plot of land. Because this value will influence the regression results and may not be representative of the other homes in the neighborhood, you might consider dropping this observation and recomputing the results.

### SKILL-BUILDER EXERCISE 6.4

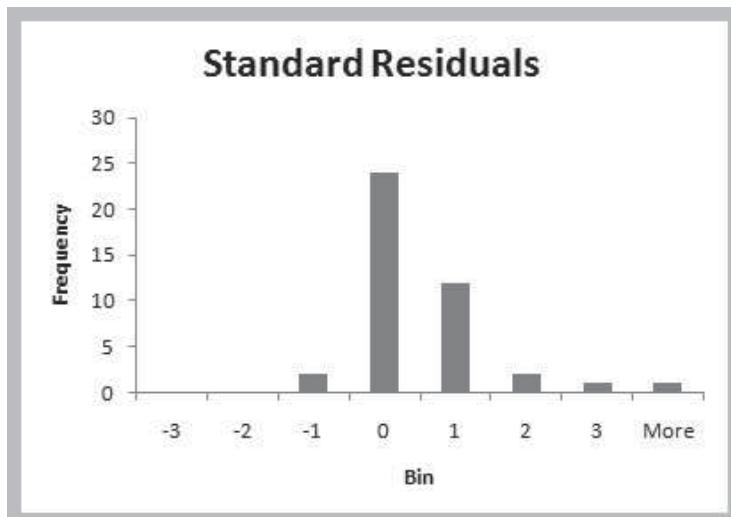
Drop the apparent outlier from the data in *Home Market Value* and rerun the regression analysis. How do the results compare?

## Checking Assumptions

The statistical hypothesis tests associated with regression analysis are predicated on some key assumptions about the data.

1. **Linearity.** This is usually checked by examining a scatter diagram of the data or examining the residual plot. If the model is appropriate, then the residuals should appear to be randomly scattered about zero, with no apparent pattern. If the residuals exhibit some well-defined pattern, such as a linear trend, a parabolic shape, and so on, then there is good evidence that some other functional form might better fit the data. The scatter diagram of the market value data appears to be linear; the residual plot in Figure 6.11 also confirms no pattern in the residuals.

2. **Normality of errors.** Regression analysis assumes that the errors for each individual value of  $X$  are normally distributed, with a mean of 0. This can be verified by examining a histogram of the standard residuals and inspecting for a bell-shaped distribution or using more formal goodness-of-fit tests. Figure 6.12 shows a histogram of the standard residuals for the market value data. The distribution appears to be somewhat positively skewed (particularly with the outlier), but does not appear to be a serious departure from normality. It is usually difficult to evaluate normality with small sample



**FIGURE 6.12** Histogram of Standard Residuals

sizes. However, regression analysis is fairly robust against departures from normality, so in most cases, this is not a serious issue.

**3. Homoscedasticity.** The third assumption is **homoscedasticity**, which means that the variation about the regression line is constant for all values of the independent variable. This can also be evaluated by examining the residual plot and looking for large differences in the variances at different values of the independent variable. In Figure 6.12, we see no serious differences in the spread of the data for different values of  $X$ , particularly if the outlier is eliminated. Caution should be exercised when looking at residual plots. In many applications, the model is derived from limited data, and multiple observations for different values of  $X$  are not available, making it difficult to draw definitive conclusions about homoscedasticity. If this assumption is seriously violated, then other techniques should be used instead of least squares for estimating the regression model.

**4. Independence of errors.** Finally, residuals should be independent for each value of the independent variable. For cross-sectional data, this assumption is usually not a problem. However, when time is the independent variable, this is an important assumption. If successive observations appear to be correlated—for example, by becoming larger over time or exhibiting a cyclical type of pattern—then this assumption is violated. Correlation among successive observations over time is called **autocorrelation** and can be identified by residual plots having clusters of residuals with the same sign. Autocorrelation can be evaluated more formally using a statistical test based on the **Durbin–Watson statistic**. The Durbin–Watson statistic is:

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2} \quad (6.7)$$

This is a ratio of the squared differences in successive residuals to the sum of the squares of all residuals.  $D$  will range from 0 to 4. When successive residuals are positively autocorrelated,  $D$  will approach 0. Critical values of the statistic have been tabulated based on the sample size and number of independent variables that allow you to conclude

that there is either evidence of autocorrelation, no evidence of autocorrelation, or the test is inconclusive. *PHStat* computes the Durbin–Watson statistic, and for most practical purposes, values below 1 suggest autocorrelation; values above 1.5 and below 2.5 suggest no autocorrelation; and values above 2.5 suggest negative autocorrelation. This can become an issue when using regression in forecasting, which we will discuss in the next chapter.

When assumptions of regression are violated, statistical inferences drawn from the hypothesis tests may not be valid. Thus, before drawing inferences about regression models and performing hypothesis tests, these assumptions should be checked. However, other than linearity, these assumptions are not needed solely for model fitting and estimation purposes.

## MULTIPLE LINEAR REGRESSION

A linear regression model with more than one independent variable is called a **multiple linear regression** model. Simple linear regression is just a special case of multiple linear regression. Multiple regression has been effectively used in many business applications. For example, Kimes and Fitzsimmons<sup>1</sup> developed a model for La Quinta Motor Inns to evaluate proposed sites for new motels. This model had 35 variables that included 6 variables about competition, 18 variables about demand, 3 demographic variables, 4 market-related variables, and 4 physical variables. The characteristics of each proposed site could be entered into a spreadsheet containing the regression model and evaluated immediately.

Consider the data in the Excel file *Colleges and Universities* (see Figure 6.2). We might believe that the graduation rate is related to the other variables. For example, it is logical to propose that schools with students who have higher SAT scores, a lower acceptance rate, a larger budget, and a higher percentage of students in the top 10% of their high school classes will tend to retain and graduate more students. We could investigate this by fitting simple linear regression models for each of these variables (see the exercise below). However, none of them alone would be a very good predictor of the graduation rate. Instead, we will try to find the best model that includes multiple independent variables.

### SKILL-BUILDER EXERCISE 6.5

For the *Colleges and Universities* data, plot scatter charts for each variable individually against the graduation rate and add linear trendlines, finding both the equation and  $R^2$  values.

A multiple linear regression model has the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \quad (6.8)$$

where

$Y$  is the dependent variable

$X_1, \dots, X_k$  are the independent (explanatory) variables

$\beta_0$  is the intercept term

$\beta_1, \dots, \beta_k$  are the regression coefficients for the independent variables

$\varepsilon$  is the error term

<sup>1</sup>S.E. Kimes and J.A. Fitzsimmons, "Selecting Profitable Hotel Sites at La Quinta Motor Inns," *Interfaces* 19, No. 6, 1990, 83–94.

Similar to simple linear regression, we estimate the regression coefficients—called **partial regression coefficients**— $b_0, b_1, b_2, \dots, b_k$ , then use the model:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k \quad (6.9)$$

to predict the value of the dependent variable. The partial regression coefficients represent the expected change in the dependent variable when the associated independent variable is increased by one unit *while the values of all other independent variables are held constant.*

For the *Colleges and Universities* data, the proposed model would be:

$$\text{Graduation\%} = b_0 + b_1 \text{SAT} + b_2 \text{Acceptance} + b_3 \text{Expenditures} + b_4 \text{Top 10\% HS}$$

Thus,  $b_2$  would represent an estimate of the change in the graduation rate for a unit increase in the acceptance rate while holding all other variables constant.

As with simple linear regression, multiple linear regression uses least squares to estimate the intercept and slope coefficients that minimize the sum of squared error terms over all observations. The principal assumptions discussed for simple linear regression also hold here. The Excel *Regression* and *PHStat Multiple Regression* tools can easily perform multiple linear regression; you need only specify the full range for the independent variable data. One caution when using the tools: *The independent variables in the spreadsheet must be in contiguous columns.* So you may have to manually move the columns of data around before applying either of these tools.

The multiple regression results for the *College and Universities* data are shown in Figure 6.13. From the *Coefficients* section, we see that the model is:

$$\begin{aligned}\text{Graduation\%} &= 17.92 + 0.072 \text{SAT} - 24.859 \text{Acceptance} \\ &\quad - 0.000136 \text{Expenditures} - 0.163 \text{Top 10\% HS}\end{aligned}$$

	A	B	C	D	E	F	G
1	Regression Analysis						
2							
<b>Regression Statistics</b>							
4	Multiple R	0.731044486					
5	R Square	0.534426041					
6	Adjusted R Square	0.492101135					
7	Standard Error	5.30833812					
8	Observations	49					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	4	1423.209266	355.8023166	12.62675098	6.33158E-07	
13	Residual	44	1239.851958	28.1784536			
14	Total	48	2663.061224				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	17.92095587	24.55722367	0.729763108	0.469402466	-31.57087575	67.4127875
18	Median SAT	0.072006285	0.017983915	4.003927007	0.000236106	0.035762085	0.108250484
19	Acceptance Rate	-24.8592318	8.315184822	-2.989618672	0.004559569	-41.61738544	-8.10107817
20	Expenditures/Student	-0.00013565	6.59314E-05	-2.057438385	0.045600176	-0.000268526	-2.77379E-06
21	Top 10% HS	-0.162764489	0.079344518	-2.051364015	0.046213846	-0.322672855	-0.002856122

**FIGURE 6.13** Multiple Regression Results for *Colleges and Universities* Data

The signs of some coefficients make sense; higher SAT scores and lower acceptance rates suggest higher graduation rates. However, we might expect that larger student expenditures and a higher percentage of top high school students would also positively influence the graduation rate. Perhaps some of the best students are more demanding and transfer schools if their needs are not being met, some entrepreneurial students might pursue other interests before graduation, or it might simply be the result of sampling error. As with simple linear regression, the model should be used only for values of the independent variables within the range of the data.

## Interpreting Results from Multiple Linear Regression

The results from the *Regression* tool are in the same format as we saw for simple linear regression. *Multiple R* and *R Square*—called the **multiple correlation coefficient** and the **coefficient of multiple determination** in the context of multiple regression—indicate the strength of association between the dependent and independent variables. The value of  $R^2$  (0.53) indicates that 53% of the variation in the dependent variable is explained by these independent variables. This suggests that other factors not included in the model, perhaps campus living conditions, social opportunities, and so on, might also influence the graduation rate.

The ANOVA section in Figure 6.13 tests for significance of the *entire model*; that is, it computes an *F*-statistic for testing the hypotheses:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1: \text{at least one } \beta_j \text{ is not } 0$$

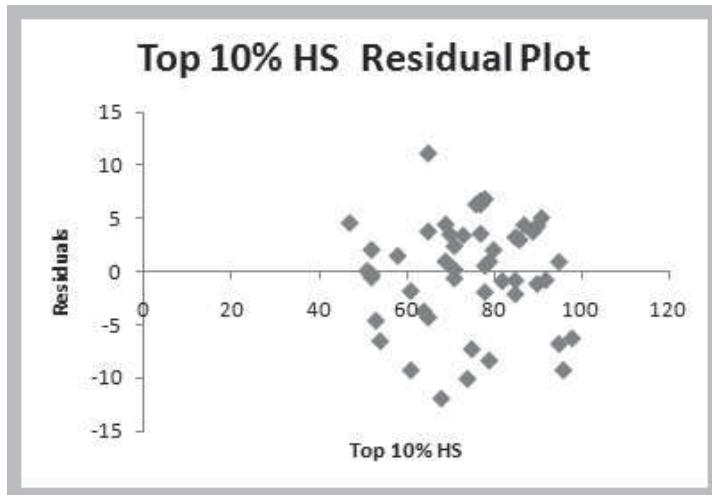
The null hypothesis states that no linear relationship exists between the dependent and *any* of the independent variables, while the alternative hypothesis states that the dependent variable has a linear relationship with *at least* one independent variable. If the null hypothesis is rejected, we cannot conclude that a relationship exists with every independent variable individually. At a 5% significance level, we reject the null hypothesis for this example because *Significance F* is essentially 0.

The last section in Figure 6.13 provides information to test hypotheses about each of the individual regression coefficients, which define the marginal contributions of the independent variables in the model. For example, to test the hypothesis that the population slope  $\beta_1$  (associated with SAT score) is 0, we examine the *p*-value and compare it to the level of significance, assumed to be 0.05. Because  $p < 0.05$ , we reject the null hypothesis that this partial regression coefficient is 0 and conclude that SAT score is significant in predicting the graduation rate. Similarly, the *p*-values for all other coefficients are less than 0.05, indicating that each of them is significant. This will not always be the case, and we will learn how to deal with large *p*-values later.

For multiple regression models, a residual plot is generated for each independent variable. This allows you to assess the linearity and homoscedasticity assumptions of regression. Figure 6.14 shows one of the residual plots from the Excel output. The assumptions appear to be met, and the other residual plots (not shown) also validate these assumptions. The normal probability plot (also not shown) does not suggest any serious departures from normality.

## Correlation and Multicollinearity

As discussed in Chapter 2, correlation, a numerical value between  $-1$  and  $+1$  measures the linear relationship between pairs of variables. The higher the absolute value of the



**FIGURE 6.14** Residual Plot for Top 10% HS Variable

correlation, the greater the strength of the relationship. The sign simply indicates whether variables tend to increase together (positive) or not (negative). Therefore, examining correlations between the dependent and independent variables can be useful in selecting variables to include in a multiple regression model because a strong correlation indicates a strong linear relationship. The Excel *Data Analysis Correlation* tool computes the correlation between all pairs of variables (see Appendix 6.2C, “Using the Correlation Tool”). However, strong correlations *among the independent variables* can be problematic.



Spreadsheet Note

Figure 6.15 shows the correlation matrix for the variables in the *Colleges and Universities* data. You can see that SAT and Acceptance Rate have moderate correlations with the dependent variable, Graduation%, but the correlation between Expenditures/ Student and Top 10% HS with Graduation% are relatively low. The strongest correlation, however, is between two independent variables: Top 10% HS and Acceptance Rate. This can potentially signify a phenomenon called **multicollinearity**, a condition occurring when two or more independent variables in the same regression model contain high levels of the same information and, consequently, are strongly correlated with one another and can predict each other better than the dependent variable. When significant multicollinearity is present, it becomes difficult to isolate the effect of one independent variable on the dependent variable, and the signs of coefficients may be the opposite of the true value, making it difficult to interpret regression coefficients. Also, *p*-values can be inflated, resulting in the conclusion not to reject the null hypothesis for significance of regression when it should be rejected.

	A	B	C	D	E	F
1		Median SAT	Acceptance Rate	Expenditures/Student	Top 10% HS	Graduation %
2	Median SAT	1				
3	Acceptance Rate	-0.601901959	1			
4	Expenditures/Student	0.572741729	-0.284254415	1		
5	Top 10% HS	0.503467995	-0.609720972	0.505782049	1	
6	Graduation %	0.564146827	-0.55037751	0.042503514	0.138612667	1

**FIGURE 6.15** Correlation Matrix for *Colleges and Universities* Data

	A	B	C	D	E
1	Regression Analysis			Regression Analysis	
2	Median SAT and all other X			Expenditures/Student and all other X	
3	<i>Regression Statistics</i>			<i>Regression Statistics</i>	
4	Multiple R	0.733444235		Multiple R	0.659705397
5	R Square	0.537940446		R Square	0.435211211
6	Adjusted R Square	0.507136476		Adjusted R Square	0.397558625
7	Standard Error	44.0015595		Standard Error	12002.17094
8	Observations	49		Observations	49
9	VIF	2.164223188		VIF	1.770573388
10					
11	Regression Analysis			Regression Analysis	
12	Top 10% HS and all other X			Acceptance Rate and all other X	
13	<i>Regression Statistics</i>			<i>Regression Statistics</i>	
14	Multiple R	0.701553357		Multiple R	0.724669621
15	R Square	0.492177112		R Square	0.525146059
16	Adjusted R Square	0.458322253		Adjusted R Square	0.49348913
17	Standard Error	9.973219924		Standard Error	0.095165693
18	Observations	49		Observations	49
19	VIF	1.969190488		VIF	2.10591071

**FIGURE 6.16** Variance Inflation Factors for Colleges and Universities Data

Correlations between independent variables exceeding an absolute value of 0.7 often indicate potential multicollinearity. However, multicollinearity is best measured using the **variance inflation factor (VIF)** for each independent variable, rather than simply examining the correlation matrix. The VIF can be computed in the *PHStat Multiple Regression* tool by checking the appropriate box in the dialog. Figure 6.16 shows the results for the *Colleges and Universities* data. If the independent variables are not correlated, then  $VIF_j = 1$ . Conservative guidelines suggest that a maximum VIF of 5 or more suggests too much multicollinearity. For the example, we see that all VIF values are below the suggested guideline. Thus, we can assume that multicollinearity is not an issue in this example. VIFs should be checked along with the assumptions of regression analysis that we discussed. If a VIF exceeds the recommended threshold, then you should consider dropping the variable from the model.

## BUILDING GOOD REGRESSION MODELS

In the *Colleges and Universities* regression example, all of the independent variables were found to be significant by evaluating the *p*-values of the regression analysis. This will not always be the case and leads to the question of how to build good regression models that include the “best” set of variables.

Figure 6.17 shows a portion of the Excel file *Banking Data*, which provides data acquired from banking and census records for different zip codes in the bank’s current market. Such information can be useful in targeting advertising for new customers or choosing locations for branch offices. The data show the median age of the population, median years of education, median income, median home value, median household wealth, and average bank balance.

Figure 6.18 shows the results for regression analysis to predict the average bank balance as a function of the other variables. While the independent variables explain over 94% of the variation in the average bank balance, you can see that at a 0.05 significance level, both Education and Home Value do not appear to be significant. A good regression model should include only significant independent variables. However, it

A	B	C	D	E	F
1	Banking Data				
2					
3	Median Age	Median Years Education	Median Income	Median Home Value	Median Household Wealth
4	35.9	14.8	\$91,033	\$183,104	\$220,741
5	37.7	13.8	\$86,748	\$163,843	\$223,152
6	36.8	13.8	\$72,245	\$142,732	\$176,926
7	35.3	13.2	\$70,639	\$145,024	\$166,260
8	35.3	13.2	\$64,879	\$135,951	\$148,868
9	34.8	13.7	\$75,591	\$155,334	\$188,310
10					\$36,708

**FIGURE 6.17** Portion of *Banking Data*

A	B	C	D	E	F	G
1	SUMMARY OUTPUT					
2						
3	<i>Regression Statistics</i>					
4	Multiple R	0.97309221				
5	R Square	0.946908448				
6	Adjusted R Square	0.944143263				
7	Standard Error	2055.64333				
8	Observations	102				
9						
10	ANOVA					
11	df	SS	MS	F	Significance F	
12	Regression	5	7235179873	1447035975	342.4394584	1.5184E-59
13	Residual	96	405664271.9	4225669.499		
14	Total	101	7640844145			
15						
16	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	-10710.64278	4260.976308	-2.513659314	0.013613178	-19168.6137
18	Age	318.6649626	60.98611242	5.225205378	1.01152E-06	197.6084892
19	Education	621.8603472	318.9595184	1.949652891	0.054135369	1254.989972
20	Income	0.146323453	0.040781001	3.588029937	0.000526666	0.065373808
21	Home Value	0.009183067	0.011038075	0.831944635	0.407504891	0.227273099
22	Wealth	0.074331533	0.011189265	6.643111131	1.84838E-09	0.052121018

**FIGURE 6.18** Regression Analysis Results for *Banking Data*

is not always clear exactly what will happen when we add or remove variables from a model; variables that are (or are not) significant in one model may (or may not) be significant in another. Therefore, you should *not* consider dropping all insignificant variables at one time, but rather take a more structured approach.

Adding an independent variable to a regression model will *always* result in  $R^2$  equal to or greater than the  $R^2$  of the original model. This is true even when the new independent variable has little true relationship with the dependent variable! Thus, trying to maximize  $R^2$  is not a useful criterion. A better way of evaluating the relative fit of different models is to use Adjusted  $R^2$ . Adjusted  $R^2$  reflects both the number of independent variables and the sample size, and may either increase or decrease when an independent variable is added or dropped, thus providing an indication of the value of adding or removing independent variables in the model. An increase in Adjusted  $R^2$  indicates that the model has improved.

This suggests a systematic approach to building good regression models:

1. Construct a model with all available independent variables. Check for significance of the independent variables by examining the  $p$ -values.
2. Identify the independent variable having the largest  $p$ -value that exceeds the chosen level of significance.
3. Remove the variable identified in step 2 from the model and evaluate Adjusted  $R^2$ . (Don't remove all variables with  $p$ -values that exceed  $\alpha$  at the same time, but only one at a time.)
4. Continue until all variables are significant.

In essence, this approach seeks to find a significant model that has the highest Adjusted  $R^2$ . If we apply this approach to the *Banking Data* example, we would remove Home Value from the model because it has the largest  $p$ -value exceeding 0.05.

Figure 6.19 shows the regression results after removing Home Value. Note that the Adjusted  $R^2$  has increased slightly, while the  $R^2$  value decreased slightly because we removed a variable from the model. Also notice that although the  $p$ -value for Education was greater than 0.05 in the first regression analysis, this variable is now significant after Home Value was removed. This now appears to be the best model.

Another criterion to determine if a variable should be removed is the  $t$ -statistic. If  $t < 1$ , then the standard error will decrease and Adjusted  $R^2$  will increase if the variable is removed. If  $t > 1$ , then the opposite will occur. In the banking regression results, we see that the  $t$ -statistic for Home Value is less than 1; therefore, we expect the Adjusted  $R^2$  to increase if we remove this variable. You can follow the same iterative approach outlined above, except using  $t$ -values instead of  $p$ -values.

These approaches using the  $p$ -values or  $t$ -statistics may involve considerable experimentation to identify the best set of variables that result in the largest Adjusted  $R^2$ . For large numbers of independent variables, the number of potential models can be overwhelming. For example, there are  $2^{10} = 1,024$  possible models that can be developed from a set of 10 independent variables. This can make it difficult to effectively screen out insignificant variables. Fortunately, automated methods—stepwise regression and best subsets—exist that facilitate this process.

A	B	C	D	E	F	G
1	SUMMARY OUTPUT					
2						
3	Regression Statistics					
4	Multiple R	0.97289551				
5	R Square	0.946525674				
6	Adjusted R Square	0.944320547				
7	Standard Error	2052.378536				
8	Observations	102				
9						
10	ANOVA					
11	df	SS	MS	F	Significance F	
12	Regression	4	7232255152	1808063788	429.2386497	9.68905E-61
13	Residual	97	408588992.5	4212257.655		
14	Total	101	7640844145			
15						
16	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	-12432.45673	3718.674319	-3.343249681	0.001177705	-19812.99569 -5051.917773
18	Age	325.0652837	60.40284468	5.381622098	5.1267E-07	205.1823604 444.9482071
19	Education	773.3800418	261.4330936	2.958233142	0.003886994	254.5077323 1292.252351
20	Income	0.159747379	0.037393587	4.272052794	4.52422E-05	0.085531461 0.233963297
21	Wealth	0.072988791	0.011054665	6.602532898	2.16051E-09	0.051048341 0.094929241

FIGURE 6.19 Regression Results without Home Value

A	B	C	D	E	F	G	H
1	Banking Data General Stepwise						
2	Table of Results for General Stepwise						
3							
4	Income entered.						
5							
6	df	SS	MS	F	Significance F		
7	Regression	1	6920338342	6920338342	960.4833602	4.43329E-53	
8	Residual	100	720505082.5	7205058.025			
9	Total	101	7640844145				
10							
11	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	
12	Intercept	4020.251548	723.8864893	5.553704355	2.31287E-07	2584.081409	5456.421688
13	Income	0.427519104	0.013794647	30.99166598	4.43329E-53	0.400150917	0.454887291
14							
15							
16	Wealth entered.						
17							
18	df	SS	MS	F	Significance F		
19	Regression	2	7088382281	3544191140	635.1115723	3.37651E-57	
20	Residual	99	552461863.7	5580422.866			
21	Total	101	7640844145				
22							
23	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	
24	Intercept	6279.90604	768.5622772	8.278696414	6.04778E-13	4774.754713	7786.056894
25	Income	0.231792518	0.037676923	6.152108514	1.62772E-08	0.157033331	0.306651705
26	Wealth	0.066901189	0.012191467	5.487542179	3.13968E-07	0.042710674	0.091091703

A	B	C	D	E	F	G	H
29	Age entered.						
30							
31	df	SS	MS	F	Significance F		
32	Regression	3	7195303082	2398454361	527.6662846	2.54304E-60	
33	Residual	98	445451062.9	4545419.01			
34	Total	101	7640844145				
35							
36	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	
37	Intercept	-3115.402529	2053.813956	-1.51686436	0.132514588	-7191.129374	960.3243157
38	Income	0.211937005	0.034249258	6.188075842	1.41707E-08	0.143970469	0.279903641
39	Wealth	0.063805426	0.010214337	5.78921007	8.5209E-08	0.041933739	0.086577112
40	Age	301.0731327	62.21535975	4.852067623	4.60905E-06	178.4087793	425.3374062
41							
42	Education entered.						
43							
44	df	SS	MS	F	Significance F		
45	Regression	4	732255152	1808083788	429.2386497	9.68905E-61	
47	Residual	97	409583992.5	4212257.655			
48	Total	101	7640844145				
49							
50	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	
51	Intercept	-12432.45673	3716.674319	-3.43249681	0.001177705	-19812.99569	-5051.917773
52	Income	0.159747379	0.037393587	4.272052794	4.52422E-05	0.085531461	0.233963297
53	Wealth	0.079887891	0.011054665	6.602528988	2.16051E-09	0.051048341	0.094929241
54	Age	325.0662037	60.40284168	5.301622098	5.1267E-07	205.1023604	444.9462071
55	Education	773.3000418	261.4330936	2.958233142	0.003086994	251.5077323	1292.252351
56							
57							

58. No other variables could be entered into the model. Stepwise ends.

**FIGURE 6.20** Stepwise Regression Results

## Stepwise Regression

Stepwise regression is a search procedure that attempts to find the best regression model without examining all possible regression models (see Appendix 6.2D, “Stepwise Regression”). In stepwise regression, variables are either added to or deleted from the current regression model at each step of the process based on the *p*-values or *t*-statistics. The process continues until no addition or removal of variables can improve the model based on the value of the adjusted  $R^2$ . Stepwise regression is usually performed in one of three ways. *Forward Selection* begins with a model having no independent variables and successively adds one at a time until no additional variable makes a significant contribution. *Backward Elimination* begins with all independent variables in the model and deletes one at a time until the best model is identified. Finally, *General Stepwise* considers deleting variables currently in the model, and if none meets the criteria for removal, then it considers adding an independent variable that is not in the model. Each option might terminate with a different model, so it is not guaranteed to always find the best.

Figure 6.20 shows the results of applying the *PHStat Stepwise Regression* tool with forward selection to the banking data. Note that using *p*-value thresholds of 0.05 results in the inclusion of Income, Wealth, Age, and Education in the final model, as we had identified earlier. Each of these variables is significant at the 0.05 level. However, note that this procedure does not provide the  $R^2$  or adjusted  $R^2$  values, so you would need to run the multiple regression analysis procedure to find these values.



Spreadsheet Note

## SKILL-BUILDER EXERCISE 6.6

Apply the *PHStat Stepwise Regression* tool to the banking data using both the general stepwise procedure and backward elimination procedure. Do you get the same model as the forward selection option?

## Best-Subsets Regression

Best-subsets regression evaluates either all possible regression models for a set of independent variables or the best subsets of models for a fixed number of independent

10	Model	Cp	k+1	R Square	Adj. R Square	Std. Error
11	X1	1132.021	2	0.319753	0.312950268	7209.482
12	X2	1153.467	2	0.307893	0.300971473	7272.06
13	X3	72.5069	2	0.905703	0.90476041	2684.224
14	X4	648.154	2	0.587349	0.583222723	5615.158
15	X5	82.72217	2	0.900054	0.899054494	2763.462
16	X1X2	744.972	3	0.534911	0.525515617	5991.298
17	X1X3	45.46647	3	0.921764	0.920183276	2457.293
18	X1X4	496.0324	3	0.672584	0.665969731	5026.931
19	X1X5	50.6053	3	0.918922	0.917283904	2501.526
20	X2X3	74.36336	3	0.905783	0.903879382	2696.611
21	X2X4	648.0156	3	0.588532	0.58021935	5635.354
22	X2X5	56.89208	3	0.915445	0.913736844	2554.599
23	X3X4	74.06647	3	0.905947	0.904046887	2694.26
24	X3X5	34.73949	3	0.927696	0.926235544	2362.292
25	X4X5	46.7294	3	0.921065	0.919470723	2468.237
26	X1X2X3	46.14721	4	0.922493	0.920120756	2458.255
27	X1X2X4	497.0943	4	0.673103	0.663095931	5048.509
28	X1X2X5	20.88464	4	0.936465	0.93451958	2225.695
29	X1X3X4	47.11486	4	0.921958	0.919569227	2466.727
30	X1X3X5	11.4155	4	0.941701	0.939916675	2131.999
31	X1X4X5	20.80774	4	0.936507	0.934563412	2224.95
32	X2X3X4	76.06601	4	0.905947	0.903068036	2707.968
33	X2X3X5	31.56207	4	0.93056	0.928433803	2326.826
34	X2X4X5	46.41904	4	0.922343	0.919965823	2460.638
35	X3X4X5	30.37914	4	0.931214	0.929108031	2315.84
36	X1X2X3X4	48.13093	5	0.922502	0.919306637	2470.75
37	X1X2X3X5	4.692132	5	0.946526	0.944320547	2052.379
38	X1X2X4X5	16.87396	5	0.939789	0.937305731	2177.83
39	X1X3X4X5	7.801146	5	0.944806	0.942530244	2085.113
40	X2X3X4X5	31.30277	5	0.931809	0.928997006	2317.652
41	X1X2X3X4X5		6	0.946908	0.944143263	2055.643

**FIGURE 6.21** Best Subsets Results for *Banking Data*



#### Spreadsheet Note

variables. *PHStat* includes a useful tool for performing best-subsets regression (see Appendix 6.2E, “Best-Subsets Regression”). Figure 6.21 shows the *PHStat* output for the banking data example ( $X_1 = \text{Age}$ ,  $X_2 = \text{Education}$ ,  $X_3 = \text{Income}$ ,  $X_4 = \text{Home Value}$ ,  $X_5 = \text{Wealth}$ ). The model having the highest Adjusted  $R^2$  can easily be identified. In Figure 6.21, this is the model with  $X_1$ ,  $X_2$ ,  $X_3$ , and  $X_5$ , the same one we identified before.

Best Subsets also evaluates models using a statistic called  $C_p$ , which is called the Bonferroni criterion.  $C_p$  estimates the bias introduced in the estimates of the responses by having an *underspecified model* (a model with important predictors missing). If  $C_p$  is much greater than  $k + 1$  (the number of independent variables plus 1), there is substantial bias. The full model always has  $C_p = k + 1$ . If all models except the full model have large  $C_p$ s, it suggests that important predictor variables are missing. Models having  $C_p$  less than  $k + 1$  or at least close to it are good models to consider. Note that the model we selected earlier is the only one that has a  $C_p$  value less than  $k + 1 = 5$ . The only other one that is close is  $X_1$ ,  $X_3$ ,  $X_4$ , and  $X_5$ .

## The Art of Model Building in Regression

The procedures we have described identify the best regression model from a purely “technical” perspective, focusing on the adjusted  $R^2$  and significance of the independent variables. Other issues, such as multicollinearity, should be considered. For instance, Figure 6.22 shows the correlation matrix for all the data in the banking example. Note that large correlations exist between Education and Home Value and also between Wealth and Income. In fact, the VIFs are Age: 1.34, Education: 2.45, Income: 14.9, Home

	A	B	C	D	E	F	G
1		Age	Education	Income	Home Value	Wealth	Balance
2	Age	1					
3	Education	0.173407147	1				
4	Income	0.4771474	0.57539402	1			
5	Home Value	0.386493114	0.753521067	0.795355158	1		
6	Wealth	0.468091791	0.469413035	0.946665447	0.698477789	1	
7	Balance	0.565466834	0.55488066	0.951684494	0.766387128	0.948711734	1

**FIGURE 6.22** Correlation Matrix for Banking Data

Value: 4.38, and Wealth: 10.71, indicating significant multicollinearity. However, after removing Home Value from the model, multicollinearity still exists, as the VIFs are Age: 1.32, Education: 1.66, Income: 12.57, and Wealth: 10.49. Thus, the model suggested by stepwise and best-subsets regression may not be adequate. The VIF values suggest that either Income or Wealth may not be appropriate variables to keep in the model.

If we remove Wealth from the model, the adjusted  $R^2$  drops to 0.9201 and all VIFs are less than 2, but we discover that Education is no longer significant. Dropping Education and leaving only Age and Income in the model results in an adjusted  $R^2$  of 0.9202. However, if we remove Income from the model instead of Wealth, the Adjusted  $R^2$  drops to only 0.9345, all VIFs are less than 2, and all remaining variables (Age, Education, and Wealth) are significant (see Figure 6.23). The  $R^2$  value for the model with these three variables is 0.936.

So what should a model builder do? The independent variables selected should make some sense in attempting to explain the dependent variable (i.e., you should have some reason to believe that changes in the independent variable will cause changes in the dependent variable even though causation cannot be proven statistically). Logic should guide your model development. In many applications, economic or physical theory might suggest that certain variables should belong in a model. Remember that additional variables do contribute to a higher  $R^2$  and, therefore, help to explain a larger proportion of the variation. Even though a variable with a large  $p$ -value is significant, it could simply be the result of sampling error and a modeler might wish to keep it.

	A	B	C	D	E	F	G
1	Regression Analysis						
2							
3	Regression Statistics						
4	Multiple R	0.967710981					
5	R Square	0.936464543					
6	Adjusted R Square	0.93451958					
7	Standard Error	2225.695322					
8	Observations	102					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	3	7155379617	2385126539	481.4819367	1.71667E-58	
13	Residual	98	485464527.3	4953719.667			
14	Total	101	7640844145				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	-17732.45142	3801.662822	-4.664393517	9.79978E-06	-25276.72737	-10188.17547
18	Age	367.8214086	64.59823831	5.693985134	1.2977E-07	239.6283103	496.0145069
19	Education	1300.308712	249.9731413	5.201793703	1.08292E-06	804.2451615	1796.372263
20	Wealth	0.116467903	0.004679827	24.88722652	3.75813E-44	0.10718094	0.125754866

**FIGURE 6.23**  
Regression Results for  
Age, Education, and  
Wealth as Independent  
Variables

Good modelers also try to have as simple a model as possible—an age-old principle known as **parsimony**—with the fewest number of explanatory variables that will provide an adequate interpretation of the dependent variable. In the physical and management sciences, some of the most powerful theories are the simplest. Thus, a model for the banking data that only includes Age, Education, and Wealth is simpler than one with four variables; because of the multicollinearity issue, there would be little gain to include Income in the model. Whether the model explains 93% or 94% of the variation in bank deposits would probably make little difference. Therefore, building good regression models relies as much on experience and judgment as it does on technical analysis.

## REGRESSION WITH CATEGORICAL INDEPENDENT VARIABLES

Some data of interest in a regression study may be ordinal or nominal. For instance, the Excel file *Employee Salaries* shown in Figure 6.24 provides salary and age data for 35 employees, along with an indicator of whether or not the employees have an MBA (Yes or No). The MBA indicator variable, however, is categorical. Since regression analysis requires numerical data, we could include categorical variables by *coding* the variables. For example, we might code “No” as 0 and “Yes” as 1. Such variables are often called **dummy variables**.

If we are interested in predicting salary as a function of the other variables, we would propose the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where

$Y$  = salary

$X_1$  = age

$X_2$  = MBA indicator (0 or 1)

After coding the MBA indicator column in the data file, we begin by running a regression on the entire data set, yielding the output shown in Figure 6.25. Note that the model explains about 95% of the variation, and the  $p$ -values of both variables are significant. The model is:

$$\text{Salary} = 893.59 + 1044.15 \times \text{Age} + 14767.23 \times \text{MBA}$$

	A	B	C	D
1	Salary Data			
2				
3	Employee	Salary	Age	MBA
4	1	\$ 28,260	25	No
5	2	\$ 43,392	28	Yes
6	3	\$ 56,322	37	Yes
7	4	\$ 26,086	23	No
8	5	\$ 36,807	32	No
9	6	\$ 57,119	57	No
10	7	\$ 48,907	45	No
11	8	\$ 34,301	32	No
12	9	\$ 31,104	25	No
13	10	\$ 60,054	57	No

**FIGURE 6.24** Portion of Excel File  
*Employee Salaries*

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.976118476					
5	R Square	0.952807278					
6	Adjusted R Square	0.949857733					
7	Standard Error	2941.914352					
8	Observations	35					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	2	5591651177	2795825589	323.0353318	6.05341E-22	
13	Residual	32	276955521.7	8654860.054			
14	Total	34	5868606699				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	893.5875971	1824.575283	0.489751015	0.627650922	-2822.950618	4610.125812
18	Age	1044.146043	42.14128238	24.77727265	1.8878E-22	958.3070603	1129.985026
19	MBA	14767.23159	1351.801764	10.92411031	2.49752E-12	12013.70151	17520.76166

**FIGURE 6.25** Initial Regression Model for Employee Salaries

Thus, a 30-year-old person with an MBA would have an estimated salary of:

$$\text{Salary} = 893.59 + 1044.15 \times 30 + 14767.23 \times 1 = \$46,985.02$$

This model suggests that having an MBA increases the salary of this group of employees by almost \$15,000. Note that by substituting either 0 or 1 for MBA, we obtain two models:

$$\text{No MBA: Salary} = 893.59 + 1044.15 \times \text{Age}$$

$$\text{MBA: Salary} = 15,660.82 + 1044.15 \times \text{Age}$$

The only difference between them is the intercept. The models suggest that the rate of salary increase for age is the same for both groups. Of course, this may not be true. Individuals with MBAs might earn relatively higher salaries as they get older. In other words, the slope of Age may *depend* on the value of MBA. Such a dependence is called an **interaction**.

We can test for interactions by defining a new variable,  $X_3 = X_1 \times X_2$  and testing whether this variable is significant, leading to an alternative model. With the interaction term, the new model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

In the worksheet, we need to create a new column (called Interaction) by multiplying MBA by Age for each observation (see Figure 6.26). The regression results are shown in Figure 6.27.

	A	B	C	D	E
1	Salary Data				
2					
3	Employee	Salary	Age	MBA	Interaction
4	1	\$ 28,260	25	0	0
5	2	\$ 43,392	28	1	28
6	3	\$ 56,322	37	1	37
7	4	\$ 26,086	23	0	0
8	5	\$ 36,807	32	0	0

**FIGURE 6.26** Portion of Employee Salaries Modified for Interaction Term

A	B	C	D	E	F	G
1	SUMMARY OUTPUT					
2						
3	Regression Statistics					
4	Multiple R	0.989321416				
5	R Square	0.978756863				
6	Adjusted R Square	0.976701076				
7	Standard Error	2005.37675				
8	Observations	35				
9						
10	ANOVA					
11		df	SS	MS	F	Significance F
12	Regression	3	5743939086	1914646362	476.098288	5.31397E-26
13	Residual	31	124667613.2	4021535.91		
14	Total	34	5868606699			
15						
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%
17	Intercept	3902.509386	1336.39766	2.920170772	0.006467654	1176.908399
18	Age	971.3090382	31.06887722	31.26308786	5.23658E-25	907.9436456
19	MBA	-2971.080074	3026.24236	-0.98177202	0.333812767	-9143.142034
20	Interaction	501.8483604	81.55221742	6.153705887	7.9295E-07	335.5215171
						668.1752038

**FIGURE 6.27** Regression Results with Interaction Term

From Figure 6.27, we see that the adjusted  $R^2$  increases; however, the  $p$ -value for the MBA indicator variable is 0.33, indicating that this variable is not significant. Therefore, we drop this variable and run a regression using only Age and the interaction term. The results are shown in Figure 6.28. Adjusted  $R^2$  increased slightly, and both Age and the interaction term are significant. The final model is:

$$\text{Salary} = 3323.11 + 984.25 \text{ Age} + 425.58 \text{ MBA} \times \text{Age}$$

The models for employees with and without an MBA are:

$$\begin{aligned}\text{No MBA: Salary} &= 3323.11 + 984.25 \times \text{Age} + 425.58(0) \times \text{Age} \\ &= 3323.11 + 984.25 \times \text{Age}\end{aligned}$$

A	B	C	D	E	F	G
1	SUMMARY OUTPUT					
2						
3	Regression Statistics					
4	Multiple R	0.98898754				
5	R Square	0.978096355				
6	Adjusted R Square	0.976727377				
7	Standard Error	2004.24453				
8	Observations	35				
9						
10	ANOVA					
11		df	SS	MS	F	Significance F
12	Regression	2	5740062823	2870031411	714.4720368	2.80713E-27
13	Residual	32	128543876.4	4016996.136		
14	Total	34	5868606699			
15						
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%
17	Intercept	3323.109564	1198.353141	2.773063675	0.009184278	882.1441051
18	Age	984.2455409	28.12039088	35.00113299	4.40388E-27	926.9661794
19	Interaction	425.5845915	24.81794165	17.14826304	1.08793E-17	375.0320988
						476.1370841

**FIGURE 6.28** Final Regression Model for Salary Data

$$\begin{aligned}\text{MBA: Salary} &= 3323.11 + 984.25 \times \text{Age} + 425.58(1) \times \text{Age} \\ &= 3323.11 + 1409.83 \times \text{Age}\end{aligned}$$

Here, we see that salary not only depends on whether an employee holds an MBA, but also on age.

### Categorical Variables with More Than Two Levels

When a categorical variable has only two levels, as in the previous example, we coded the levels as 0 and 1 and added a new variable to the model. However, when a categorical variable has  $k > 2$  levels, we need to add  $k - 1$  additional variables to the model. To illustrate this, the Excel file *Surface Finish* provides measurements of the surface finish of 35 parts produced on a lathe, along with the revolutions per minute (RPM) of the spindle and one of four types of cutting tools used (see Figure 6.29). The engineer who collected the data is interested in predicting the surface finish as a function of RPM and type of tool.

Intuition might suggest defining a dummy variable for each tool type; however, doing so will cause numerical instability in the data and cause the regression tool to crash. Instead, we will need  $k - 1 = 3$  dummy variables corresponding to three of the levels of the categorical variable. The level left out will correspond to a reference or baseline value. Therefore, because we have  $k = 4$  levels of tool type, we will define a regression model of the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

where

$Y$  = surface finish

$X_1$  = RPM

$X_2 = 1$  if tool type is B and 0 if not

$X_3 = 1$  if tool type is C and 0 if not

$X_4 = 1$  if tool type is D and 0 if not

	A	B	C	D
1	<b>Surface Finish Data</b>			
2				
3	Part	Surface Finish	RPM	Cutting Tool
4	1	45.44	225	A
5	2	42.03	200	A
6	3	50.10	250	A
7	4	48.75	245	A
8	5	47.92	235	A
9	6	47.79	237	A
10	7	52.26	265	A
11	8	50.52	259	A
12	9	45.58	221	A
13	10	44.78	218	A
14	11	33.50	224	B
15	12	31.23	212	B
16	13	37.52	248	B
17	14	37.13	260	B
18	15	34.70	243	B

**FIGURE 6.29** Portion of Surface Finish

Note that when  $X_2 = X_3 = X_4 = 0$ , then by default, the tool type is A. Substituting these values for each tool type into the model, we obtain:

$$\text{Tool Type A: } Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

$$\text{Tool Type B: } Y = \beta_0 + \beta_1 X_1 + \beta_2 + \varepsilon$$

$$\text{Tool Type C: } Y = \beta_0 + \beta_1 X_1 + \beta_3 + \varepsilon$$

$$\text{Tool Type D: } Y = \beta_0 + \beta_1 X_1 + \beta_4 + \varepsilon$$

For a fixed value of RPM ( $X_1$ ), the slopes corresponding to the dummy variables represent the difference between the surface finish using that tool type and the baseline using tool type A.

To incorporate these dummy variables into the regression model, we add three columns to the data as shown in Figure 6.30. Using these data, we obtain the regression results shown in Figure 6.31. The resulting model is:

$$\text{Surface Finish} = 24.49 + 0.098 \text{ RPM} - 13.31 \text{ Type B} - 20.49 \text{ Type C} - 26.04 \text{ Type D}$$

	A	B	C	D	E	F
1	Surface Finish Data					
2						
3	Part	Surface Finish	RPM	Type B	Type C	Type D
4	1	45.44	225	0	0	0
5	2	42.03	200	0	0	0
6	3	50.10	250	0	0	0
7	4	48.75	245	0	0	0
8	5	47.92	235	0	0	0
9	6	47.79	237	0	0	0
10	7	52.26	265	0	0	0
11	8	50.52	259	0	0	0
12	9	45.58	221	0	0	0
13	10	44.78	218	0	0	0
14	11	33.50	224	1	0	0
15	12	31.23	212	1	0	0
16	13	37.52	248	1	0	0
17	14	37.13	260	1	0	0
18	15	34.70	243	1	0	0
19	16	33.92	238	1	0	0
20	17	32.13	224	1	0	0
21	18	35.47	251	1	0	0
22	19	33.49	232	1	0	0
23	20	32.29	216	1	0	0
24	21	27.44	225	0	1	0
25	22	24.03	200	0	1	0
26	23	27.33	250	0	1	0
27	24	27.20	245	0	1	0
28	25	27.10	235	0	1	0
29	26	27.30	237	0	1	0
30	27	28.30	265	0	1	0
31	28	28.40	259	0	1	0
32	29	26.80	221	0	1	0
33	30	26.40	218	0	1	0
34	31	21.40	224	0	0	1
35	32	20.50	212	0	0	1
36	33	21.90	248	0	0	1
37	34	22.13	260	0	0	1
38	35	22.40	243	0	0	1

**FIGURE 6.30** Data Matrix for Surface Finish with Dummy Variables

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.99447053					
5	R Square	0.988924942					
6	Adjusted R Square	0.987448267					
7	Standard Error	1.089163115					
8	Observations	35					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	4	3177.784271	794.4460678	669.6973322	7.32449E-29	
13	Residual	30	35.58828875	1.186276292			
14	Total	34	3213.37256				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	24.49437244	2.473298088	9.903526211	5.73134E-11	19.4432239	29.54552099
18	RPM	0.097760627	0.010399996	9.400064035	1.89415E-10	0.076521002	0.119000251
19	Type B	-13.31056756	0.487142953	-27.32374035	9.37003E-23	-14.30544619	-12.31568893
20	Type C	-20.487	0.487088553	-42.06011387	3.12134E-28	-21.48176753	-19.49223247
21	Type D	-26.03674519	0.596886375	-43.62094073	1.06415E-28	-27.25574979	-24.81774059

**FIGURE 6.31** Surface Finish Regression Model Results

Almost 99% of the variation in surface finish is explained by the model, and all variables are significant. The models for each individual tool are:

$$\begin{aligned}\text{Tool A: Surface Finish} &= 24.49 + 0.098 \text{ RPM} - 13.31(0) - 20.49(0) - 26.04(0) \\ &= 24.49 + 0.098 \text{ RPM}\end{aligned}$$

$$\begin{aligned}\text{Tool B: Surface Finish} &= 24.49 + 0.098 \text{ RPM} - 13.31 - 20.49(0) - 26.04(0) \\ &= 11.18 + 0.098 \text{ RPM}\end{aligned}$$

$$\begin{aligned}\text{Tool C: Surface Finish} &= 24.49 + 0.098 \text{ RPM} - 13.31(0) - 20.49(1) - 26.04(0) \\ &= 4.00 + 0.098 \text{ RPM}\end{aligned}$$

$$\begin{aligned}\text{Tool D: Surface Finish} &= 24.49 + 0.098 \text{ RPM} - 13.31 - 20.49(0) - 26.04(1) \\ &= -1.55 + 0.098 \text{ RPM}\end{aligned}$$

Note that the only differences among these models are the intercepts; the slopes associated with RPM are the same. This suggests that we might wish to test for interactions between the type of cutting tool and RPM; we leave this to you as an exercise.

### SKILL-BUILDER EXERCISE 6.7

For the *Surface Finish* example, run models to examine interactions between the type of cutting tool and RPM. (*Hint:* Add interaction terms between each dummy variable and RPM.) Are any interactions significant?

## REGRESSION MODELS WITH NONLINEAR TERMS

Linear regression models are not appropriate for every situation. A scatter chart of the data might show a nonlinear relationship, or the residuals for a linear fit might result in a nonlinear pattern. In such cases, we might propose a nonlinear model to explain the relationship. For instance, a second-order polynomial model would be:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Sometimes this is called a **curvilinear regression model**. In this model,  $\beta_1$  represents the linear effect of  $X$  on  $Y$ , and  $\beta_2$  represents the curvilinear effect. However, although this model appears to be quite different from ordinary linear regression models, it is still *linear in the parameters* (the betas, which are the unknowns that we are trying to estimate). In other words, all terms are a product of a beta coefficient and some function of the data, which are simply numerical values. In such cases, we can still apply least squares to estimate the regression coefficients.

To illustrate this, the Excel file *Beverage Sales* provides data on the sales of cold beverages at a small restaurant with a large outdoor patio during the summer months (see Figure 6.32). The owner has observed that sales tend to increase on hotter days. Figure 6.33 shows linear regression results for these data. The U shape of the residual plot suggests that a linear relationship is not appropriate. To apply a curvilinear regression model, add a column to the data matrix by squaring the temperatures. Now, both temperature and temperature squared are the independent variables. Figure 6.34 shows the results for the curvilinear regression model. The model is:

$$\text{Sales} = 142850 - 3643.17 \times \text{Temperature} + 23.2 \times \text{Temperature}^2$$

Note that the adjusted  $R^2$  has increased significantly from the linear model and that the residual plots now show more random patterns.

#### SKILL-BUILDER EXERCISE 6.8

For *Beverage Sales* data, run the regression tool and then use the *Trendline* option to fit a second-order polynomial to the residual plot for temperature. Also, construct a scatter chart for the original data and add a second-order polynomial trendline.

Curvilinear regression models are also often used in forecasting when the independent variable is time. This and other applications of regression in forecasting will be discussed in the next chapter.

	A	B
1	Beverage Sales	
2		
3	Temperature	Sales
4	85	\$ 1,810
5	90	\$ 4,825
6	79	\$ 438
7	82	\$ 775
8	84	\$ 1,213
9	96	\$ 8,692

**FIGURE 6.32** Portion of Excel File *Beverage Sales*

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.922351218					
5	R Square	0.850731769					
6	Adjusted R Square	0.842875547					
7	Standard Error	1041.057399					
8	Observations	21					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	117362193.6	117362194	108.287635	2.7611E-09	
13	Residual	19	20592209.67	1083800.51			
14	Total	20	137954403.2				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	-32511.24671	3408.723477	-9.5376603	1.122E-08	-39645.78693	-25376.706
18	Temperature	408.6026284	39.26555335	10.4061345	2.7611E-09	326.4188809	490.786376

FIGURE 6.33 Linear Regression Results for Beverage Sales

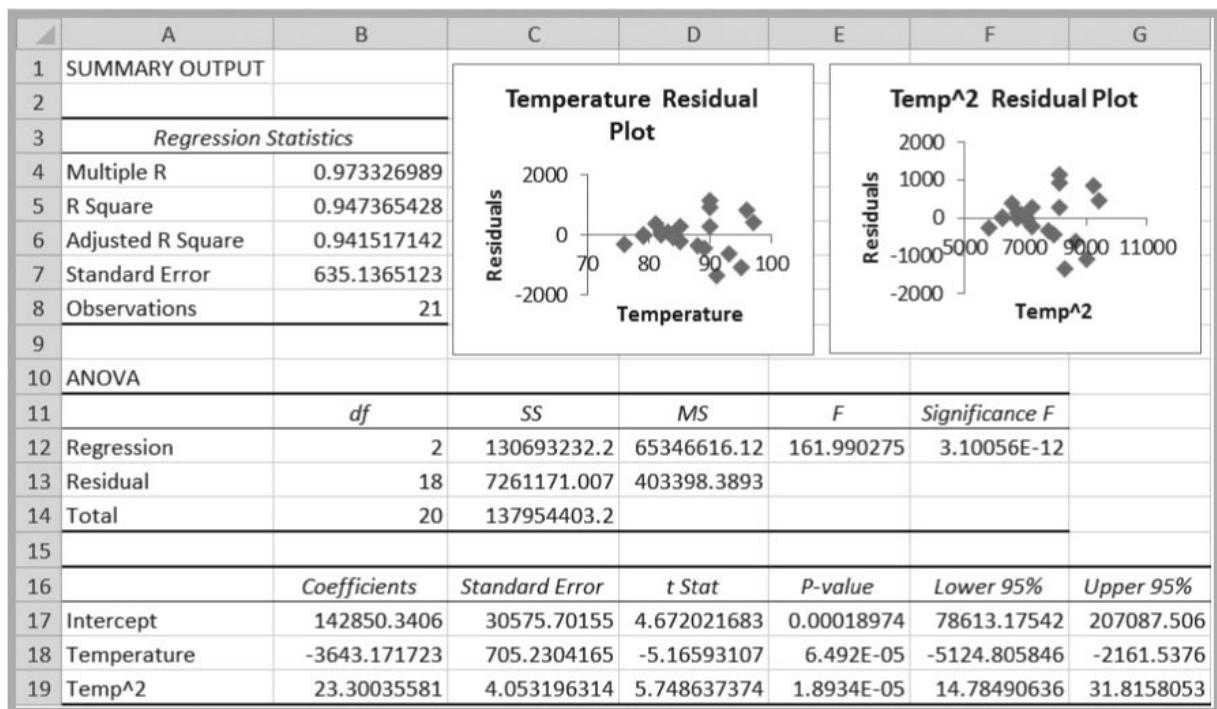


FIGURE 6.34 Curvilinear Regression Results for Beverage Sales

## Basic Concepts Review Questions

1. Interpret the intercept and slope parameters in the linear regression model.
2. What is simple linear regression? How does one find the best fitting regression line?
3. Does the regression line change when one switches the role of the independent and dependent variables? Why or why not?
4. How can regression analysis be applied to understand risk associated with stocks and mutual funds?
5. How is the efficacy of a regression model measured?
6. What is the standard error of the estimate? What information does it provide?
7. How should you interpret the value of the dependent variable generated by the fitted regression model for a specific value of the explanatory variable?
8. Explain why the prediction interval is necessarily wider than the corresponding confidence interval.
9. How are standard residuals calculated, and how can they help identify outliers?
10. Explain the assumptions of linear regression. How can you determine if each of these assumptions holds?
11. What are partial regression coefficients, and what information do they provide?
12. What are the advantages and disadvantages of using the  $R^2$  and adjusted  $R^2$  as measures of predictive fit?
13. What is the difference in the hypotheses tested in single and multiple regression using the *Significance F*-statistic?
14. What is multicollinearity, and how is it measured? What problems can it cause in regression results?
15. Describe the systematic process for building a good regression model.
16. Describe the differences and advantages/disadvantages of using stepwise and best-subsets approaches in building regression models.
17. How should you balance technical criteria and practical judgment in building regression models?
18. Explain how to include categorical variables in regression models.
19. What is interaction, and why is it important to test for it?
20. What is a curvilinear regression model? How do you know when to use one?

## Problems and Applications

1. Using the data in the Excel file *Banking Data*, check if a linear relationship exists between the variables Average Bank Balance and Median Years of Education. Construct a scatter chart and fit a trendline.
2. For the data in *China Trade Data*, set up a scatter plot and add a trendline to determine if a linear relationship exists between US imports ( $Y$ ) and US exports ( $X$ ). Also, find the estimated imports for a year when the exports were equal to 40 (billion \$).
3. The managing director of a consulting group has the following monthly data on total overhead costs and professional labor hours to bill to clients:<sup>2</sup>

Total	Billable
\$340,000	3,000
\$400,000	4,000
\$435,000	5,000
\$477,000	6,000
\$529,000	7,000
\$587,000	8,000

Develop a regression model to identify the fixed overhead costs to the consulting group.

- a. What is the constant component of the consultant group's overhead?
- b. If a special job requiring 1,000 billable hours that would contribute a margin of \$38,000 before overhead was available, would the job be attractive?
4. Use the 2010 data in the Excel files *S&P 500* and *Google Stock Prices* to find the beta risk of Google stock using simple linear regression. How do you interpret your result?
5. Choose a stock of interest and historical daily data on its closing values for 2009 through October 1, 2010. Use the Excel file *S&P 500* to find the beta risk of the stock using simple linear regression. How do you interpret your result?
6. Using the data in the Excel file *Banking Data*, run a regression analysis using Average Bank Balance as the dependent variable and Median Years of Education as the independent variable. Interpret the key regression results.
7. Using the data in the Excel file *China Trade Data*, run a regression analysis using US exports as the dependent variable and US imports as the independent variable. Interpret the key regression results.
8. The Excel file *National Football League* provides various data on professional football for the 2007 season.

<sup>2</sup> Horngren, Foster, and Datar, *Cost Accounting: A Managerial Emphasis*, 9th ed., Prentice-Hall, 1997, p. 371.

- a. Construct a scatter diagram for Points/Game and Yards/Game in the Excel file. Does there appear to be a linear relationship?
- b. Develop a regression model for predicting Points/Game as a function of Yards/Game. Explain the statistical significance of the model.
- c. Draw conclusions about the validity of the regression analysis assumptions from the residual plot and standard residuals.
- d. Find a 95% confidence interval for the mean number of Points/Game for teams with 300 Yards/Game.
- e. Find a 95% prediction interval for a team having 300 Yards/Game.
9. Consider the data in the Excel file *Olympic Track and Field Results*. The Olympic records in Discus Throw, High Jump and Long Jump all show a clear increasing trend. Can one be predicted from the others(s)? Run a regression analysis using Discus Throw as the dependent variable and High Jump as the independent variable. Interpret the key regression results.
10. A deep-foundation engineering contractor has bid on a foundation system for a new world headquarters building for a Fortune 500 company. A part of the project consists of installing 311 auger cast piles. The contractor was given bid information for cost-estimating purposes, which consisted of the estimated depth of each pile; however, actual drill footage of each pile could not be determined exactly until construction was performed. The Excel file *Pile Foundation* contains the estimates and actual pile lengths after the project was completed. Develop a linear regression model to estimate the actual pile length as a function of the estimated pile lengths. What do you conclude?
11. Consider the data in the Excel file *Restaurant Sales*. Can one reasonably predict the Delivery Sales on a given day using the Lunch Sales? Run a regression analysis using Delivery Sales as the dependent variable and Lunch Sales as the independent variable. Interpret the key regression results.
12. Consider the data in the Excel file *Restaurant Sales*. Develop a multiple linear regression model for the dependent variable Delivery Sales on a given day, using Lunch Sales and Dinner Sales as the independent variables. Compare the results with the simple linear regression which uses Lunch Sales as the only independent variable, and interpret the key regression results.
13. For the data in the Excel file *Olympic Track and Field Results*, fit a multiple linear regression model using Discus Throw as the dependent variable and High Jump and Long Jump as the independent variables. How much does the fit improve over the simple linear regression which uses High Jump as the only independent variable? Interpret the key regression results.
14. The Excel file *Cereal Data* provides a variety of nutritional information about 67 cereals and their shelf location in a supermarket. Use regression analysis to find the best model that explains the relationship between calories and the other variables. Investigate the model assumptions and clearly explain your conclusions.
15. The Excel file *Salary Data* provides information on current salary, beginning salary, previous experience (in months) when hired, and total years of education for a sample of 100 employees in a firm.
- a. Develop a multiple regression model for predicting current salary as a function of the other variables.
- b. Find the best model for predicting current salary using the *t*-value criterion.
16. The Excel file *Credit Approval Decisions* provides information on credit history for a sample of banking customers. Use regression analysis to identify the best model for predicting the credit score as a function of the other numerical variables. For the model you select, conduct further analysis to check for significance of the independent variables and for multicollinearity.
17. Using the data in the Excel file *Freshman College Data*, identify the best regression model for predicting the first year retention rate. For the model you select, conduct further analysis to check for significance of the independent variables and for multicollinearity.
18. The Excel file *Major League Baseball* provides data on the 2010 season.
- a. Construct and examine the correlation matrix. Is multicollinearity a potential problem? Find the variance inflation factors to check your intuition.
- b. Suggest an appropriate set of independent variables that predict the number of wins by examining the correlation matrix and variance inflation factors.
- c. Find the best multiple regression model for predicting the number of wins. How good is your model? Does it use the same variables you thought were appropriate in part (b)?
19. The Excel file *Golfing Statistics* provides data for a portion of the 2010 professional season for the top 25 golfers.
- a. Find the best multiple regression model for predicting earnings/event as a function of the remaining variables.
- b. Find the best multiple regression model for predicting average score as a function of the other variables except earnings and events.
20. Apply stepwise regression using each selection rule and *p*-value criterion to find a good model for predicting the number of points scored per game by football teams using the data in the Excel file *National Football League*. Compare your results.
21. Apply stepwise regression using the each selection rule and *t*-value criterion to find the best model for predicting the number of wins in the *Major League Baseball* data. Compare your results. How do the stepwise models compare with your answer to problem 18(c)?
22. Apply stepwise regression using each selection rule and *p*-value criterion to find the best models for predicting earnings/event and average score in the *Golfing*

*Statistics* data. How do the stepwise models compare with your answer to problem 19?

23. Apply best-subsets regression to find the best model for explaining the relationship between calories and the other variables in the Excel file *Cereal Data*. Use the regression tool to run your selected model, and explain all statistical results.

24. Apply best-subsets regression to find the best model for explaining the relationship between current salary and the other variables in the Excel file *Salary Data*. Use the regression tool to run your selected model, and explain all statistical results.

25. The State of Ohio Department of Education has a mandated ninth-grade proficiency test that covers writing, reading, mathematics, citizenship (social studies), and science. The Excel file *Ohio Education Performance* provides data on success rates (defined as the percentage of students passing) in school districts in the greater Cincinnati metropolitan area along with state averages.

a. Develop a multiple regression model to predict math success as a function of success in all other subjects using the systematic approach described in this chapter. Is multicollinearity a problem?

b. Suggest the best regression model to predict math success as a function of success in the other subjects by examining the correlation matrix; then run the regression tool for this set of variables.

c. Develop the best regression model to predict math success as a function of success in the other subjects using best-subsets regression.

d. Develop the best regression model to predict math success as a function of success in the other subjects using stepwise regression.

e. Compare the results for parts (a) through (d). Are the models the same? Why or why not?

26. A mental health agency measured the self-esteem score for randomly selected individuals with disabilities who were involved in some work activity within the past year. The Excel file *Self-Esteem* provides the data, including the individuals' marital status, length of work, type of support received (direct support includes job-related services such as job coaching and counseling), education, and age.

a. Use multiple linear regression for predicting self-esteem as a function of length of work, support level, education, and age.

b. Investigate possible interaction effects between support level and the other numerical variables by building an appropriate model. Determine the best model.

c. Find the best regression model for predicting self-esteem as a function of length of work, marital status, education, and age. (Note that the categorical variable marital status has more than two levels.)

27. A national homebuilder builds single-family homes and condominium-style townhouses. The Excel file *House Sales* provides information on the selling price, lot cost, type of home, and region of the country (M = Midwest, S = South) for closings during one month.

a. Develop a multiple regression model for sales price as a function of lot cost and type of home.

b. Determine if any interactions exist between lot cost and type of home, and find the best model.

c. Develop a regression model for selling price as a function of lot cost and region. (Note that the categorical variable region has more than two levels.)

28. Cost functions are often nonlinear with volume because production facilities are often able to produce larger quantities at lower rates than smaller quantities.<sup>3</sup> Using the following data, apply simple linear regression, and examine the residual plot. What do you conclude? Construct a scatter chart and use the Excel *Trendline* feature to identify the best type of trendline that maximizes  $R^2$ .

Units Produced	Costs
500	\$12,500
1,000	\$25,000
1,500	\$32,500
2,000	\$40,000
2,500	\$45,000
3,000	\$50,000

29. The Helicopter Division of Aerospatiale is studying assembly costs at its Marseilles plant.<sup>4</sup> Past data indicates the following labor hours per helicopter:

Helicopter Number	Labor Hours
1	2,000
2	1,400
3	1,238
4	1,142
5	1,075
6	1,029
7	985
8	957

Using these data, apply simple linear regression, and examine the residual plot. What do you conclude? Construct a scatter chart and use the Excel *Trendline* feature to identify the best type of trendline that maximizes  $R^2$ .

<sup>3</sup> Horngren, Foster, and Datar, *Cost Accounting: A Managerial Emphasis*, 9th ed., Prentice-Hall, 1997, p. 349.

<sup>4</sup> Horngren, Foster, and Datar, *Cost Accounting: A Managerial Emphasis*, 9th ed., Prentice Hall, 1997, p. 349.

# Case

## Hatco

The Excel file *HATCO*<sup>1</sup> consists of data related to predicting the level of business (Usage Level) obtained from a survey of purchasing managers of customers of an industrial supplier, HATCO. The following are the independent variables.

- **Delivery Speed**—amount of time it takes to deliver the product once an order is confirmed.
- **Price Level**—perceived level of price charged by product suppliers.
- **Price Flexibility**—perceived willingness of HATCO representatives to negotiate price on all types of purchases.
- **Manufacturing Image**—overall image of the manufacturer or supplier.
- **Overall Service**—overall level of service necessary for maintaining a satisfactory relationship between supplier and purchaser.
- **Sales Force Image**—overall image of the manufacturer's sales force.

- **Product Quality**—perceived level of quality of a particular product.
- **Size of Firm**—size relative to others in this market (0 = small; 1 = large).

Responses to the first seven variables were obtained using a graphic rating scale, where a 10-cm line was drawn between endpoints labeled "poor" and "excellent." Respondents indicated their perceptions using a mark on the line, which was measured from the left endpoint. The result was a scale from 0 to 10 rounded to one decimal place.

Using the tools in this chapter, conduct a complete analysis to predict Usage Level. Be sure to investigate the impact of the categorical variable Size of Firm (coded as 0 for small firms and 1 for large firms) and possible interactions. Also stratify the data by firm size to account for any differences between small and large firms. Write up your results in a formal report to HATCO management.

<sup>1</sup> Adapted from Hair, Anderson, Tatham, and Black, *Multivariate Analysis*, 5th ed., Prentice-Hall, 1998.

## APPENDIX 6.1

### Regression Theory and Computation

In this appendix, we present the basic underlying theory behind regression analysis, and also show how some of the key statistics are computed.

#### A. Regression as Analysis of Variance

The objective of regression analysis is to explain the variation of the dependent variable around its mean value as the independent variable changes. Figure 6A.1 helps to understand this. In Figure 6A.1(a), the scatter plot does not show any linear relationship. If we try to fit a regression line,  $\beta_1$  would be 0 and the model would reduce to  $Y = \beta_0 + \varepsilon$ . The intercept,  $\beta_0$ , would simply be the sample mean of the dependent variable observation,  $\bar{Y}$ . Thus, an estimate of the population mean of  $Y$  for any value of  $X$  would be  $\bar{Y}$ . We could measure the error by subtracting the estimate of the mean from each observation,  $(Y - \bar{Y})$ . In Figure 6A.1(b), we have the opposite extreme in which all points lie on the regression line, that is, all of the variation in the data around the mean is explained by the independent variable. Note that if we estimate  $Y$  using  $\bar{Y}$  alone, we would have a large amount of error. However, by knowing the value of  $X$ , we can predict the value of  $Y$  accurately using the regression line  $\hat{Y} = b_0 + b_1 X$ . In this case, the value of each observation

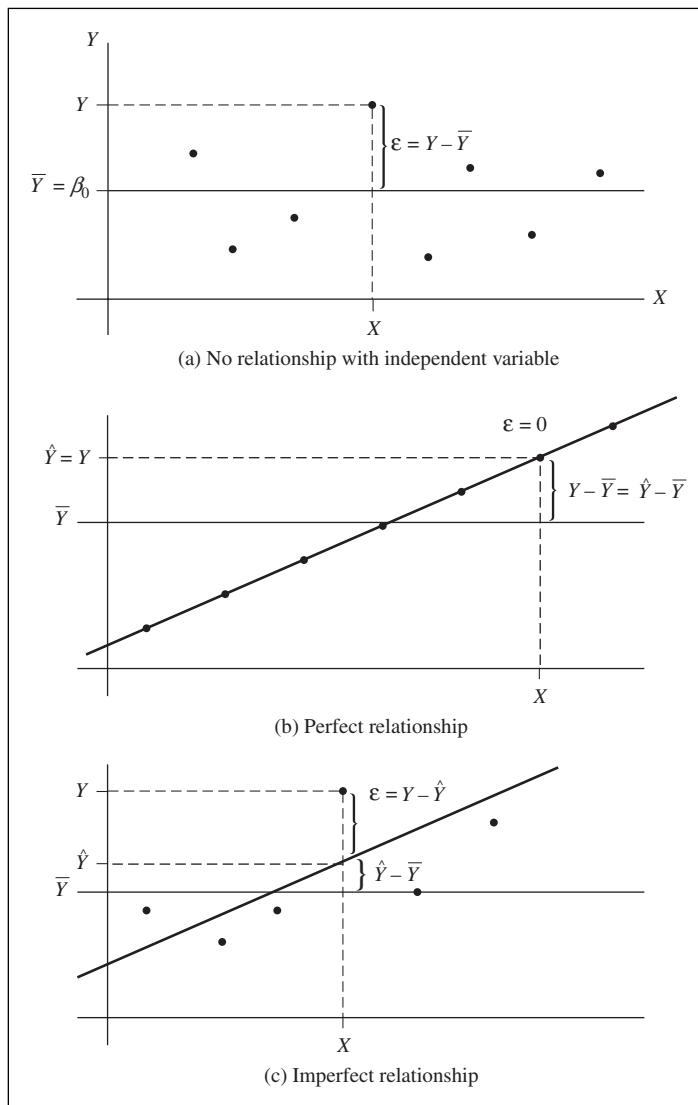
is equal to its predicted value and the error—the difference between  $\hat{Y}$  and  $Y$ —is 0. Because  $(Y - \bar{Y}) = (\hat{Y} - \bar{Y})$ , all the variation from the mean is explained by the regression line. Finally, Figure 6A.1(c) shows the typical case in which some of the variation around the mean,  $(Y - \bar{Y})$ , is explained by the independent variable,  $(\hat{Y} - \bar{Y})$ , while some is also due to error,  $(Y - \hat{Y})$ , because of the fact that the points do not lie on the regression line.

We have defined three measures of variation:

1. Variation between the observations and the mean:  $(Y - \bar{Y})$
2. Variation between the predicted values using the regression line and the mean, which is explained by the regression line:  $(\hat{Y} - \bar{Y})$
3. Variation between the individual observations and the predicted values, which is the remaining unexplained variation:  $(Y - \hat{Y})$

Because some terms will be positive and others will be negative, we need to square them to obtain a useful measure of the total variation in all the data; otherwise, they will sum to 0.

The sum of the squares of the deviations of individual observations from the mean, summed over all



**FIGURE 6A.1** Illustrations of Variation in Regression

observations,  $\sum(Y - \bar{Y})^2$ , is called the *total sum of squares* or SST. Mathematically, this can be shown to equal  $\sum(\hat{Y} - \bar{Y})^2 + \sum(Y - \hat{Y})^2$ , which is simply the sum of the squares of the variation *explained by regression*, called SSR, and the sum of squares of the errors, or *unexplained variation*, SSE. In other words,

$$SST = SSR + SSE \quad (6A.1)$$

This sum of squares is shown in the ANOVA section of the regression analysis output.

Looking back at Figure 6A.1, notice that in (a), SST=SSE because the independent variable did not explain any variation in the dependent variable. In (b), SST=SSR, all the variation is explained by the regression line. In (c), both SSR and SSE are positive, indicating that some variation is explained by regression while some error still exists. As

points become clustered closer around a straight line, SSR increases while SSE decreases.

In a simple linear regression model with one explanatory variable, the total sum of squares,  $SST = \sum(Y - \bar{Y})^2$ , has  $n - 1$  degrees of freedom because we estimate the mean. SSR has 1 degree of freedom because there is one independent variable. This leaves  $n - 2$  degrees of freedom for SSE. In Chapter 5, we noted that sums of squares divided by their appropriate degrees of freedom provide estimates of variances, which we called mean squares. Thus,  $MSR = SSR/1$  represents the variance between observations explained by regression, while  $MSE = SSE/(n - 2)$  represents the remaining variance due to error. These are shown in the MS column of the ANOVA section of the regression output. By dividing MSR by MSE, we obtain an  $F$ -statistic. If this number is higher than the critical value from the  $F$ -distribution for a chosen level of significance, then we would reject the

null hypothesis. Logically, if the null hypothesis is true, then  $SST = SSE$ , and  $SSR$  (and  $MSR$ ) would be ideally 0. Therefore, the smaller the  $F$ -ratio, the greater is the likelihood that  $H_0$  is true. Likewise, the larger the  $F$ -ratio, the greater the likelihood is that  $\beta_1 \neq 0$  and that the independent variable helps to explain the variation in the data about the mean.

## B. Standard Error of the Estimate

The standard error of the estimate,  $S_{YX}$ , is the standard deviation of the errors about the regression line. This is computed as:

$$S_{YX} = \sqrt{\frac{SSE}{n - 2}} \quad (6A.2)$$

## C. Adjusted $R^2$

Adjusted  $R^2$  is computed as:

$$R_{adj}^2 = 1 - \left[ (1 - R^2) \frac{n - 1}{n - 2} \right] \quad (6A.3)$$

## D. Confidence Intervals

We can develop a  $100(1 - \alpha)\%$  confidence interval for the *mean value* of  $Y$  using the following formula:

$$\hat{Y} \pm t_{\alpha/2, n-2} S_{YX} \sqrt{h_i} \quad (6A.4)$$

where  $\hat{Y} = b_0 + b_1 X_i$  is the predicted mean of  $Y$  for a given value,  $X_i$ , of the independent variable, and:

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2} \quad (6A.5)$$

This is a bit different from the types of confidence intervals that we computed in Chapter 4. It turns out that the true standard error and the width of the confidence interval actually depend on the value of  $X_i$  as expressed by the  $h_i$ -statistic. The further  $X_i$  deviates from the mean of  $X$ , the larger the value of  $h_i$ , and hence, the larger the confidence interval. Therefore, we actually have a collection of confidence intervals for each value of  $X$ . Such a collection of confidence intervals is called a **confidence band** around the regression line.

## E. Prediction Intervals

A  $100(1 - \alpha)\%$  prediction interval for  $\hat{Y}$  for a given value,  $X_i$ , of the independent variable is:

$$\hat{Y} \pm t_{\alpha/2, n-2} S_{YX} \sqrt{1 + h_i} \quad (6A.6)$$

In many cases, the value of  $h_i$  is quite small so that the term  $\sqrt{1 + h_i}$  is close to 1. Therefore, we can develop an *approximate*  $100(1 - \alpha)\%$  prediction interval simply by

$$\hat{Y} \pm t_{\alpha/2, n-2} S_{YX}$$

## APPENDIX 6.2

### Excel and PHStat Notes

#### A. Using the Trendline Option

First, click the chart to which you wish to add a trendline; this will display the *Chart Tools* menu. The *Trendline* option is selected from the *Analysis* group under the *Layout* tab in the *Chart Tools* menu. Click the *Trendline* button and then *More Trendline Options....* This brings up the *Format Trendline* dialog shown in Figure 6A.2. Make sure that the radio button for *Linear* is selected. We will discuss the other nonlinear models in Chapter 9. Check the boxes for *Display Equation on chart* and *Display R-squared value on chart*. Excel will display the results on the chart you have selected; you may move the equation and  $R$ -squared value for better readability by dragging them to a different location. A simpler way of doing this is to right-click on the data series in the chart and choose *Add trendline* from the pop-up menu (try it!).

#### B. Using Regression Tools

Both Excel and *PHStat* provide tools for regression analysis. The Excel *Regression* tool can be used for single or multiple linear regressions. From the *Data Analysis* menu in the *Analysis* group under the *Data* tab, select the *Regression* tool. The dialog box shown in Figure 6A.3 is displayed. In the box for the *Input Y Range*, specify the range of the dependent variable values. In the box for the *Input X Range*, specify the range for the independent variable values. Check *Labels* if your data range contains a descriptive label (we highly recommend using this). You have the option of forcing the intercept to 0 by checking *Constant is Zero*; however, you will usually not check this box because adding an intercept term allows a better fit to the data. You also can set a *Confidence Level* (the default of 95% is commonly used) to provide confidence intervals for the intercept and slope parameters.

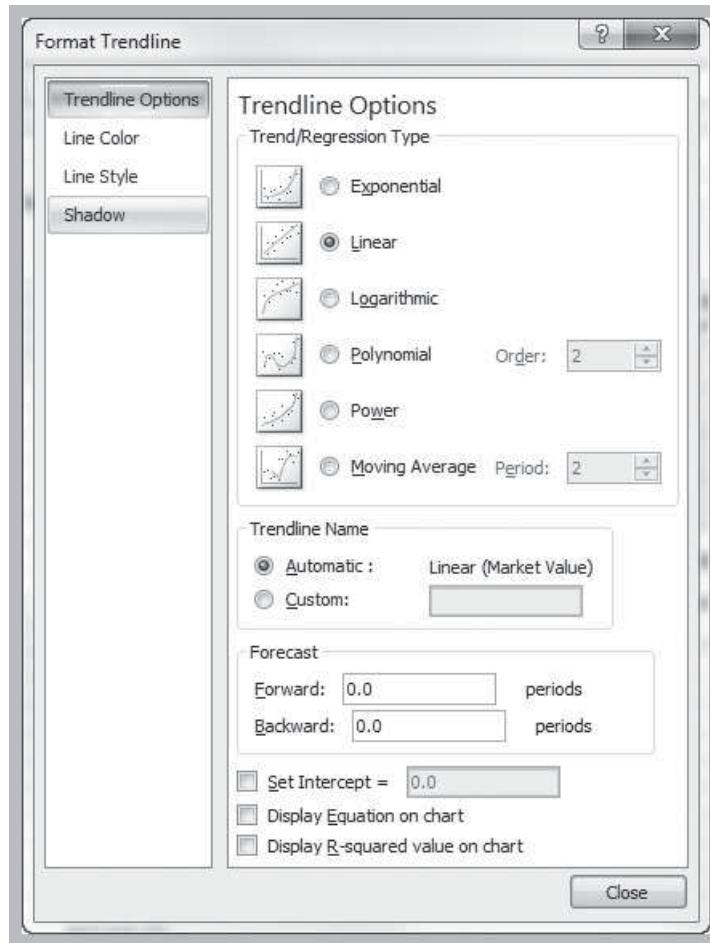


FIGURE 6A.2 Format Trendline Dialog

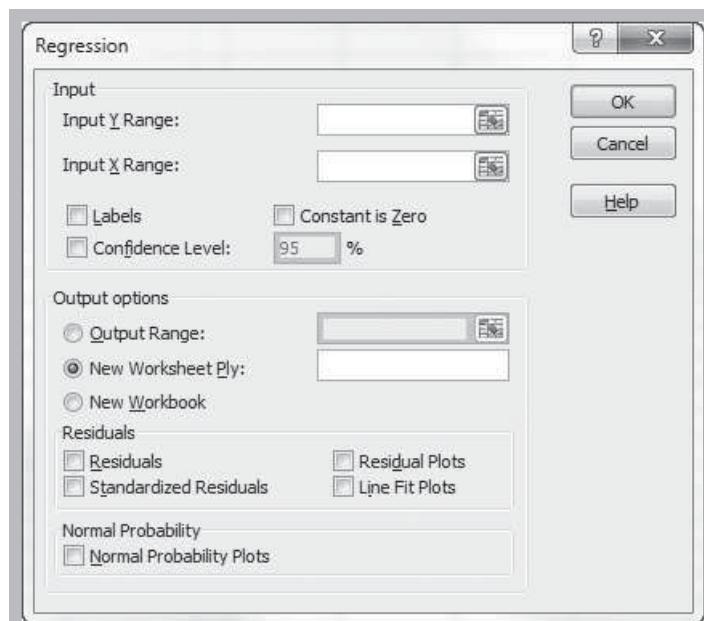


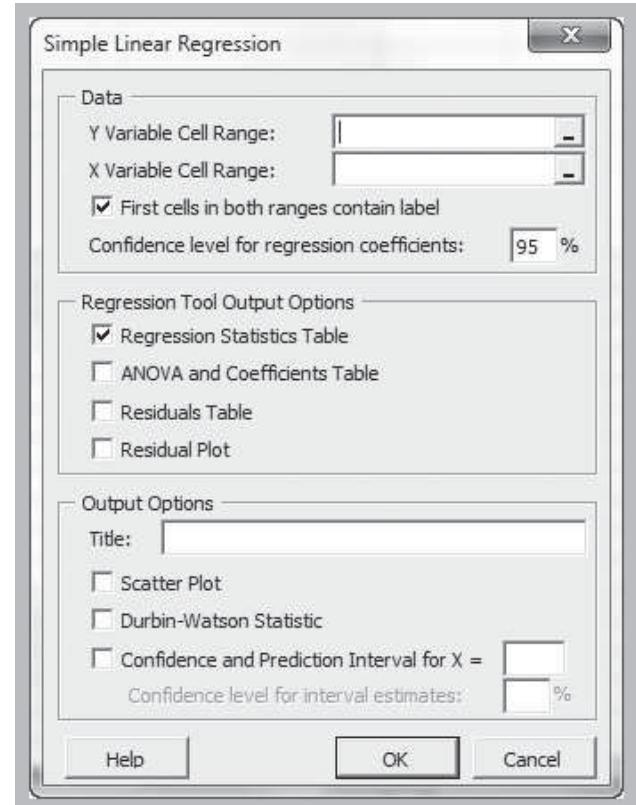
FIGURE 6A.3 Excel Regression Tool Dialog

In the *Residuals* section, you have the option of including a residuals output table by checking *Residuals*, *Standardized Residuals*, *Residual Plots*, and *Line Fit Plots*. The *Residual Plots* generates a chart for each independent variable versus the residual, and the *Line Fit Plot* generates a scatter chart with the values predicted by the regression model included (a scatter chart with an added trendline is visually superior). Finally, you may also have to choose Excel to construct a normal probability plot for the dependent variable which transforms the cumulative probability scale (vertical axis) so that the graph of the cumulative normal distribution will be a straight line. The closer the points are to a straight line, the better the fit to a normal distribution.

*PHStat* provides two separate tools, one for *Simple Linear Regression* and one for *Multiple Regression*, both found in the *Regression* menu of *PHStat*. In the *Simple Linear Regression* tool (see Figure 6A.4), the data input is identical to the Excel tool. However, the output options are somewhat different. Excel automatically provides the *Regression Statistics Table* and *ANOVA and Coefficients Table*, which are options in the *PHStat* tool. Note that while Excel provides *Standardized Residuals* and *Line Fit Plots*, *PHStat* does not. Other output options in *PHStat* include a scatter diagram, Durbin–Watson statistic (also discussed later), and confidence and prediction intervals. Results are generated on separate worksheets in the active workbook. The dialog box for the *Multiple Regression* tool is nearly identical, but has some additional check boxes for information related exclusively to multiple regression, and we will explain them later in this chapter.

### C. Using the Correlation Tool

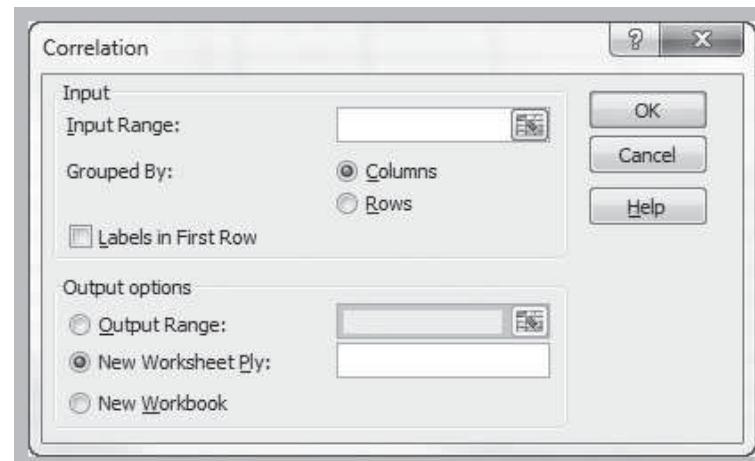
From the *Data Analysis* menu, select the *Correlation* tool. The dialog box shown in Figure 6A.5 is displayed. In the box for the *Input Range*, specify the range of the data for which you want correlations. As with other tools, check *Labels in First Row* if your data range contains a descriptive label.



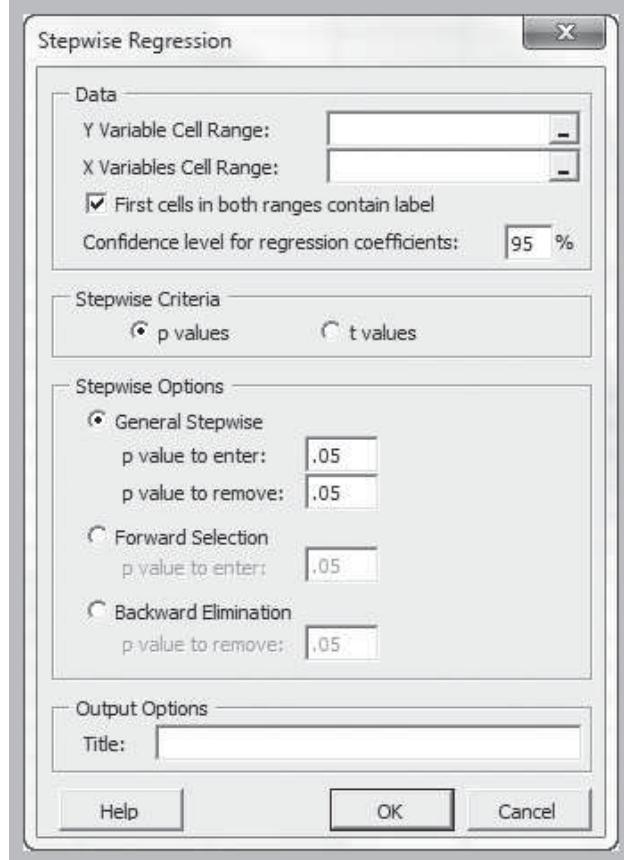
**FIGURE 6A.4** *PHStat Simple Linear Regression Tool Dialog*

### D. Stepwise Regression

From the *PHStat* menu, select *Regression*, then *Stepwise Regression*. The dialog box that appears (Figure 6A.6) prompts you to enter the range for the dependent variable and the independent variables, the confidence level for the regression, the criteria to use for selecting variables, and the type



**FIGURE 6A.5** *Correlation Tool Dialog*



**FIGURE 6A.6** *PHStat Stepwise Regression Dialog*

of stepwise procedure to run. The procedure produces two worksheets: one that contains the multiple regression model that includes all independent variables and another with a table of stepwise results.

### E. Best-Subsets Regression

From the *PHStat* menu, select *Regression*, then *Best Subsets*. The dialog box that appears prompts you to enter the range for

the dependent variable and the independent variables, as well as the confidence level for the regression. The tool creates a *Best Subsets* worksheet that contains a summary of the models analyzed. With a large number of variables, the analysis can take a significant amount of time and memory and may cause a fatal error in Excel, depending on the processor capability and amount of memory available, so this tool should be used cautiously. The tool also provides worksheets with ANOVA output for each of the combinations for further analysis.

## *Chapter 7*

# Forecasting

- INTRODUCTION 238
- QUALITATIVE AND JUDGMENTAL METHODS 238
  - Historical Analogy 239
  - The Delphi Method 239
  - Indicators and Indexes for Forecasting 239
- STATISTICAL FORECASTING MODELS 240
- FORECASTING MODELS FOR STATIONARY TIME SERIES 242
  - Moving Average Models 242
  - Error Metrics and Forecast Accuracy 244
  - Exponential Smoothing Models 246
- FORECASTING MODELS FOR TIME SERIES WITH A LINEAR TREND 248
  - Regression-Based Forecasting 248
- ADVANCED FORECASTING MODELS 249
  - Autoregressive Forecasting Models 250
  - Forecasting Models with Seasonality 252
  - Incorporating Seasonality in Regression Models 253
  - Forecasting Models with Trend and Seasonality 255
  - Regression Forecasting with Causal Variables 255
- CHOOSING AND OPTIMIZING FORECASTING MODELS USING CB PREDICTOR 257
- THE PRACTICE OF FORECASTING 262
- BASIC CONCEPTS REVIEW QUESTIONS 263
- PROBLEMS AND APPLICATIONS 264
- CASE: ENERGY FORECASTING 265
- APPENDIX 7.1: ADVANCED FORECASTING MODELS—THEORY AND COMPUTATION 265
  - A. Double Moving Average 265
  - B. Double Exponential Smoothing 265
  - C. Additive Seasonality 266
  - D. Multiplicative Seasonality 266
  - E. Holt–Winters Additive Model 266
  - F. Holt–Winters Multiplicative Model 267

■ APPENDIX 7.2: EXCEL AND CB PREDICTOR NOTES	267
□ A. Forecasting with Moving Averages	267
□ B. Forecasting with Exponential Smoothing	267
□ C. Using CB Predictor	268

## INTRODUCTION

One of the major problems that managers face is forecasting future events in order to make good decisions. For example, forecasts of interest rates, energy prices, and other economic indicators are needed for financial planning; sales forecasts are needed to plan production and workforce capacity; and forecasts of trends in demographics, consumer behavior, and technological innovation are needed for long-term strategic planning. The government also invests significant resources on predicting short-run U.S. business performance using the Index of Leading Indicators. This index focuses on the performance of individual businesses, which often is highly correlated with the performance of the overall economy, and is used to forecast economic trends for the nation as a whole.

Managers may choose from a wide range of forecasting techniques. Selecting the appropriate method depends on the characteristics of the forecasting problem, such as the time horizon of the variable being forecast, as well as available information on which the forecast will be based. Three major categories of forecasting approaches are *qualitative and judgmental techniques*, *statistical time-series models*, and *explanatory/causal methods*.

Qualitative and judgmental techniques rely on experience and intuition; they are necessary when historical data are not available or when the decision maker needs to forecast far into the future. For example, a forecast of when the next generation of a microprocessor will be available and what capabilities it might have will depend greatly on the opinions and expertise of individuals who understand the technology.

Statistical time-series models find greater applicability for short-range forecasting problems. A **time series** is a stream of historical data, such as weekly sales. Time-series models assume that whatever forces have influenced sales in the recent past will continue into the near future; thus, forecasts are developed by extrapolating these data into the future.

Explanatory/causal models, often called econometric models, seek to identify factors that explain statistically the patterns observed in the variable being forecast, usually with regression analysis. While time-series models use only time as the independent variable, explanatory/causal models generally include other factors. For example, forecasting the price of oil might incorporate independent variables such as the demand for oil (measured in barrels), the proportion of oil stock generated by OPEC countries, and tax rates. Although we can never prove that changes in these variables actually cause changes in the price of oil, we often have evidence that a strong influence exists.

Surveys of forecasting practices have shown that both judgmental and quantitative methods are used for forecasting sales of product lines or product families, as well as for broad company and industry forecasts. Simple time-series models are used for short- and medium-range forecasts, whereas regression analysis is the most popular method for long-range forecasting. However, many companies rely on judgmental methods far more than quantitative methods, and almost half judgmentally adjust quantitative forecasts. In this chapter we introduce some common methods and approaches to forecasting using judgmental techniques, statistical time-series models, regression, and exploratory/causal models.

## QUALITATIVE AND JUDGMENTAL METHODS

Qualitative, or judgmental, forecasting methods are valuable in situations for which no historical data are available or for those that specifically require human expertise and knowledge. One example might be identifying future opportunities and threats as part of a SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis within a strategic planning exercise. Another use of judgmental methods is to incorporate

nonquantitative information, such as the impact of government regulations or competitor behavior, in a quantitative forecast. Judgmental techniques range from such simple methods as a manager's opinion or a group-based jury of executive opinion to more structured approaches such as historical analogy and the Delphi method.

## **Historical Analogy**

One judgmental approach is **historical analogy**, in which a forecast is obtained through a comparative analysis with a previous situation. For example, if a new product is being introduced, the response of similar previous products to marketing campaigns can be used as a basis to predict how the new marketing campaign might fare. Of course, temporal changes or other unique factors might not be fully considered in such an approach. However, a great deal of insight can often be gained through an analysis of past experiences. For example, in early 1998, the price of oil was about \$22 a barrel. However, in mid-1998, the price of a barrel of oil dropped to around \$11. The reasons for this price drop included an oversupply of oil from new production in the Caspian Sea region, high production in non-OPEC regions, and lower-than-normal demand. In similar circumstances in the past, OPEC would meet and take action to raise the price of oil. Thus, from historical analogy, we might forecast a rise in the price of oil. OPEC members did in fact meet in mid-1998 and agreed to cut their production, but nobody believed that they would actually cooperate effectively, and the price continued to drop for a time. Subsequently, in 2000, the price of oil rose dramatically, falling again in late 2001. Analogies often provide good forecasts, but you need to be careful to recognize new or different circumstances. Another analogy is international conflict relative to the price of oil. Should war break out, the price would be expected to rise, analogous to what it has done in the past.

## **The Delphi Method**

A popular judgmental forecasting approach, called the **Delphi method**, uses a panel of experts, whose identities are typically kept confidential from one another, to respond to a sequence of questionnaires. After each round of responses, individual opinions, edited to ensure anonymity, are shared, allowing each to see what the other experts think. Seeing other experts' opinions helps to reinforce those in agreement and to influence those who did not agree to possibly consider other factors. In the next round, the experts revise their estimates, and the process is repeated, usually for no more than two or three rounds. The Delphi method promotes unbiased exchanges of ideas and discussion and usually results in some convergence of opinion. It is one of the better approaches to forecasting long-range trends and impacts.

## **Indicators and Indexes for Forecasting**

Indicators and indexes generally play an important role in developing judgmental forecasts. Indicators are measures that are believed to influence the behavior of a variable we wish to forecast. By monitoring changes in indicators, we expect to gain insight about the future behavior of the variable to help forecast the future. For example, one variable that is important to the nation's economy is the Gross Domestic Product (GDP), which is a measure of the value of all goods and services produced in the United States. Despite its shortcomings (for instance, unpaid work such as housekeeping and child care is not measured; production of poor-quality output inflates the measure, as does work expended on corrective action), it is a practical and useful measure of economic performance. Like most time series, the GDP rises and falls in a cyclical fashion. Predicting future trends in the GDP is often done by analyzing *leading indicators*—series that tend to rise and fall some predictable length of time prior to the peaks and valleys of the GDP. One example of a

leading indicator is the formation of business enterprises; as the rate of new businesses grows, one would expect the GDP to increase in the future. Other examples of leading indicators are the percent change in the money supply ( $M1$ ) and net change in business loans. Other indicators, called *lagging indicators*, tend to have peaks and valleys that follow those of the GDP. Some lagging indicators are the Consumer Price Index, prime rate, business investment expenditures, or inventories on hand. The GDP can be used to predict future trends in these indicators.

Indicators are often combined quantitatively into an index. The direction of movement of all the selected indicators are weighted and combined, providing an index of overall expectation. For example, financial analysts use the Dow Jones Industrial Average as an index of general stock market performance. Indexes do not provide a complete forecast, but rather a better picture of direction of change, and thus play an important role in judgmental forecasting.

The Department of Commerce began an Index of Leading Indicators to help predict future economic performance. Components of the index include the following:

- Average weekly hours, manufacturing
- Average weekly initial claims, unemployment insurance
- New orders, consumer goods and materials
- Vendor performance—slower deliveries
- New orders, nondefense capital goods
- Building permits, private housing
- Stock prices, 500 common stocks (Standard & Poor)
- Money supply
- Interest rate spread
- Index of consumer expectations (University of Michigan)

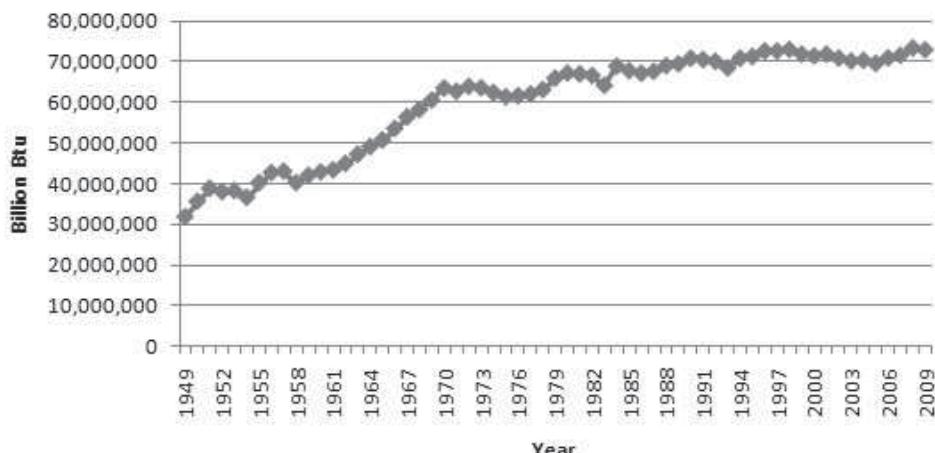
*Business Conditions Digest* included more than 100 time series in seven economic areas. This publication was discontinued in March 1990, but information related to the Index of Leading Indicators was continued in *Survey of Current Business*. In December 1995, the U.S. Department of Commerce sold this data source to The Conference Board, which now markets the information under the title *Business Cycle Indicators*; information can be obtained at its Web site ([www.conference-board.org](http://www.conference-board.org)). The site includes excellent current information about the calculation of the index, as well as its current components.

## STATISTICAL FORECASTING MODELS

Many forecasts are based on analysis of historical time-series data and are predicated on the assumption that the future is an extrapolation of the past. We will assume that a time series consists of  $T$  periods of data,  $A_t, t = 1, 2, \dots, T$ . A naive approach is to eyeball a **trend**—a gradual shift in the value of the time series—by visually examining a plot of the data. For instance, Figure 7.1 shows a chart of total energy production from the data in the Excel file *Energy Production & Consumption*. We see that energy production was rising quite rapidly during the 1960s; however, the slope appears to have decreased after 1970. It appears that production is increasing by about 500,000 each year and that this can provide a reasonable forecast provided that the trend continues.

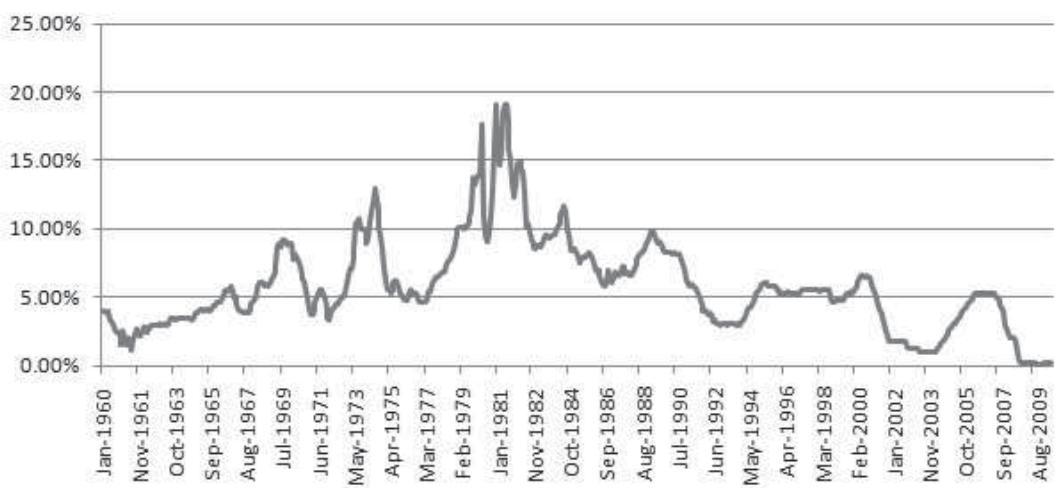
Time series may also exhibit short-term *seasonal effects* (over a year, month, week, or even a day) as well as longer-term *cyclical effects* or nonlinear trends. At a neighborhood grocery store, for instance, short-term seasonal patterns may occur over a week, with the heaviest volume of customers on weekends; seasonal patterns may also be evident during the course of a day, with higher volumes in the mornings and late afternoon. Cycles relate to much longer-term behavior, such as periods of inflation and recession or bull and bear stock market behavior. Figure 7.2 shows a chart of the data in the Excel file *Federal Funds Rate*. We see some evidence of long-term cycles in the time series.

## Total Energy Production



**FIGURE 7.1** Total Energy Production Time Series

## Federal Funds Rate



**FIGURE 7.2** Federal Funds Rate Time Series

Of course, unscientific approaches such as the “eyeball method” may be a bit unsettling to a manager making important decisions. Subtle effects and interactions of seasonal and cyclical factors may not be evident from simple visual extrapolation of data. Statistical methods, which involve more formal analyses of time series, are invaluable in developing good forecasts. A variety of statistically based forecasting methods for time series are commonly used. Among the most popular are *moving average methods*, *exponential smoothing*, and *regression analysis*. These can be implemented very easily on a spreadsheet using basic functions available in Microsoft Excel and its *Data Analysis* tools; these are summarized

**TABLE 7.1** Excel Support for Forecasting

Excel Functions	Description
TREND ( <i>known_y</i> 's, <i>known_x</i> 's, <i>new_x</i> 's, constant)	Returns values along a linear trend line
LINEST ( <i>known_y</i> 's, <i>known_x</i> 's, <i>new_x</i> 's, constant, stats)	Returns an array that describes a straight line that best fits the data
FORECAST ( <i>x</i> , <i>known_y</i> 's, <i>known_x</i> 's)	Calculates a future value along a linear trend
Analysis Toolpak	Description
Moving average	Projects forecast values based on the average value of the variable over a specific number of preceding periods
Exponential smoothing	Predicts a value based on the forecast for the prior period, adjusted for the error in that prior forecast
Regression	Used to develop a model relating time-series data to a set of variables assumed to influence the data

in Table 7.1. Moving average and exponential smoothing models work best for stationary time series. For time series that involve trends and/or seasonal factors, other techniques have been developed. These include double moving average and exponential smoothing models, seasonal additive and multiplicative models, and Holt–Winters additive and multiplicative models. We will review each of these types of models. This book also provides an Excel add-in, *CB Predictor*, which applies these methods and incorporates some intelligent technology. We will describe *CB Predictor* later in this chapter.

## FORECASTING MODELS FOR STATIONARY TIME SERIES

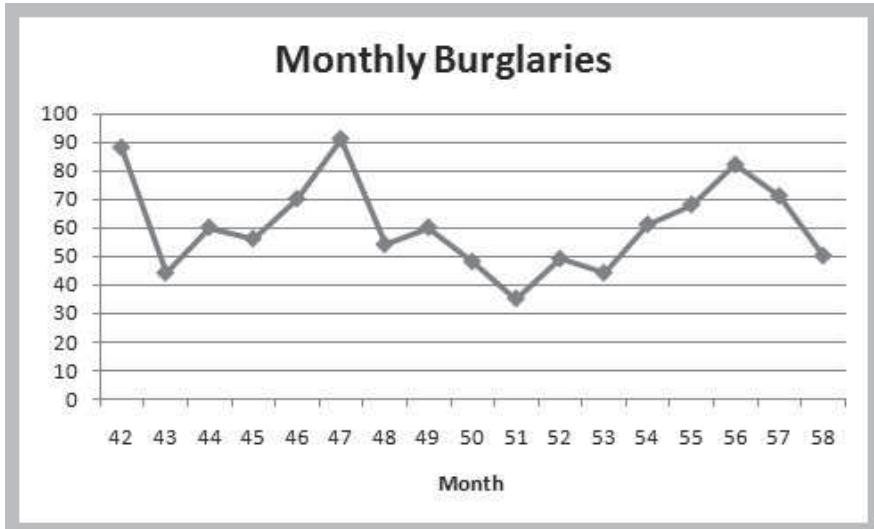
Two simple approaches that are useful over short time periods—when trend, seasonal, or cyclical effects are not significant—are moving average and exponential smoothing models.

### Moving Average Models

The **simple moving average** method is a smoothing method based on the idea of averaging random fluctuations in the time series to identify the underlying direction in which the time series is changing. Because the moving average method assumes that future observations will be similar to the recent past, it is most useful as a short-range forecasting method. Although this method is very simple, it has proven to be quite useful in stable environments, such as inventory management, in which it is necessary to develop forecasts for a large number of items.

Specifically, the simple moving average forecast for the next period is computed as the average of the most recent  $k$  observations. The value of  $k$  is somewhat arbitrary, although its choice affects the accuracy of the forecast. The larger the value of  $k$ , the more the current forecast is dependent on older data and the forecast will not react as quickly to fluctuations in the time series. The smaller the value of  $k$ , the quicker the forecast responds to changes in the time series. Also, when  $k$  is larger, extreme values have less effect on the forecasts. (In the next section, we discuss how to select  $k$  by examining errors associated with different values.)

For instance, suppose that we want to forecast monthly burglaries from the Excel file *Burglaries* since the citizen-police program began. Figure 7.3 shows a chart of these



**FIGURE 7.3** Monthly Burglaries Time Series

data. The time series appears to be relatively stable, without trend, seasonal, or cyclical effects; thus, a moving average model would be appropriate. Setting  $k = 3$ , the three-period moving average forecast for month 59 is:

$$\text{Month 59 forecast} = \frac{82 + 71 + 50}{3} = 67.67$$

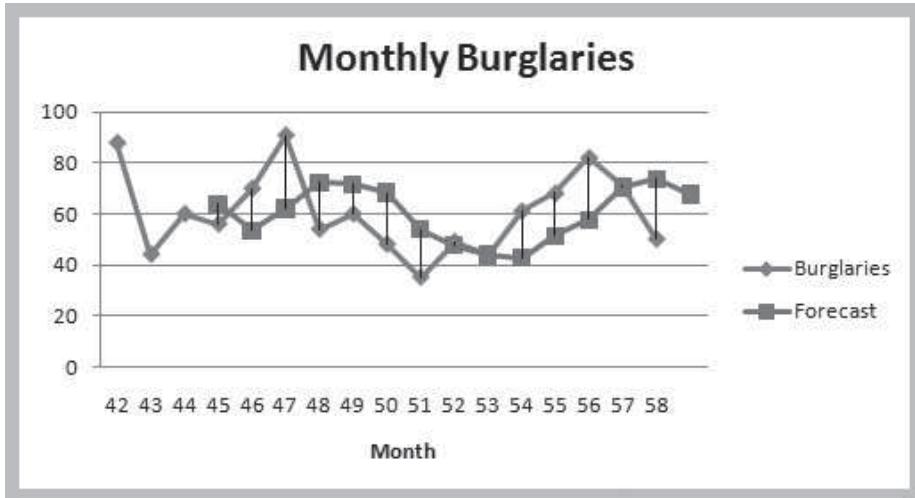
Moving average forecasts can be generated easily on a spreadsheet. Figure 7.4 shows the computations for a three-period moving average forecast of burglaries. Figure 7.5 shows a chart that contrasts the data with the forecasted values. Moving average forecasts can also be obtained from Excel's *Data Analysis* options (see Appendix 7.2A, *Forecasting with Moving Averages*).



Spreadsheet Note

C	D	E	F	
1	After Citizen-Police Program		Moving Average	
2	Month	Monthly burglaries	Forecast	
3	42	88		
4	43	44		
5	44	60		
6	45	56	64.00	Forecast for month 45 =AVERAGE(D3:D5)
7	46	70	53.33	
8	47	91	62.00	
9	48	54	72.33	
10	49	60	71.67	
11	50	48	68.33	
12	51	35	54.00	
13	52	49	47.67	
14	53	44	44.00	
15	54	61	42.67	
16	55	68	51.33	
17	56	82	57.67	
18	57	71	70.33	Forecast for month 59 =AVERAGE(D17:D19)
19	58	50	73.67	
20	59		67.67	
21				

**FIGURE 7.4** Excel Implementation of Moving Average Forecast



**FIGURE 7.5** Chart of Burglaries and Moving Average Forecast

In the simple moving average approach, the data are weighted equally. This may not be desirable because we might wish to put more weight on recent observations than on older observations, particularly if the time series is changing rapidly. Such models are called **weighted moving averages**. For example, you might assign a 60% weight to the most recent observation, 30% to the second most recent observation, and the remaining 10% of the weight to the third most recent observation. In this case, the three-period weighted moving average forecast for month 59 would be:

$$\text{Month 59 Forecast} = \frac{0.1 \times 82 + 0.3 \times 71 + 0.6 \times 50}{0.1 + 0.3 + 0.6} = \frac{59.5}{1} = 59.5$$

Different weights can easily be incorporated into Excel formulas. This leads us to the questions of how to measure forecast accuracy and also how to select the best parameters for a forecasting model.

### Error Metrics and Forecast Accuracy

The quality of a forecast depends on how accurate it is in predicting future values of a time series. The error in a forecast is the difference between the forecast and the actual value of the time series (once it is known!). In Figure 7.5, the forecast error is simply the vertical distance between the forecast and the data for the same time period. In the simple moving average model, different values for  $k$  will produce different forecasts. How do we know, for example, if a two- or three-period moving average forecast or a three-period weighted moving average model (or others) would be the best predictor for burglaries? We might first generate different forecasts using each of these models, as shown in Figure 7.6, and compute the errors associated with each model.

To analyze the accuracy of these models, we can define *error metrics*, which compare quantitatively the forecast with the actual observations. Three metrics that are commonly used are the *mean absolute deviation*, *mean square error*, and *mean absolute percentage error*. The **mean absolute deviation (MAD)** is the absolute difference between the actual value and the forecast, averaged over a range of forecasted values:

$$\text{MAD} = \frac{\sum_{t=1}^n |A_t - F_t|}{n} \quad (7.1)$$

	A	B	C	D	E	F	G	H
1	After Citizen-Police Program			3 Period				
2	Month	Monthly burglaries	k = 2	Error	k = 3	Error	Weighted	Error
3	42	88						
4	43	44						
5	44	60	66.00	-6.00				
6	45	56	52.00	4.00	64.00	-8.00	58.00	-2.00
7	46	70	58.00	12.00	53.33	16.67	56.00	14.00
8	47	91	63.00	28.00	62.00	29.00	64.80	26.20
9	48	54	80.50	-26.50	72.33	-18.33	81.20	-27.20
10	49	60	72.50	-12.50	71.67	-11.67	66.70	-6.70
11	50	48	57.00	-9.00	68.33	-20.33	61.30	-13.30
12	51	35	54.00	-19.00	54.00	-19.00	52.20	-17.20
13	52	49	41.50	7.50	47.67	1.33	41.40	7.60
14	53	44	42.00	2.00	44.00	0.00	44.70	-0.70
15	54	61	46.50	14.50	42.67	18.33	44.60	16.40
16	55	68	52.50	15.50	51.33	16.67	54.70	13.30
17	56	82	64.50	17.50	57.67	24.33	63.50	18.50
18	57	71	75.00	-4.00	70.33	0.67	75.70	-4.70
19	58	50	76.50	-26.50	73.67	-23.67	74.00	-24.00
20	59		60.50		67.67		59.50	

**FIGURE 7.6** Alternative Moving Average Forecasting Models

where  $A_t$  is the actual value of the time series at time  $t$ ,  $F_t$  is the forecast value for time  $t$ , and  $n$  is the number of forecast values (*not* the number of data points since we do not have a forecast value associated with the first  $k$  data points). MAD provides a robust measure of error and is less affected by extreme observations.

**Mean square error (MSE)** is probably the most commonly used error metric. It penalizes larger errors because squaring larger numbers has a greater impact than squaring smaller numbers. The formula for MSE is:

$$\text{MSE} = \frac{\sum_{t=1}^n (A_t - F_t)^2}{n} \quad (7.2)$$

Again,  $n$  represents the number of forecast values used in computing the average. Sometimes the square root of MSE, called the **root mean square error (RMSE)**, is used:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^n (A_t - F_t)^2}{n}} \quad (7.3)$$

Note that unlike MSE, RMSE is expressed in the same units as the data (similar to the difference between a standard deviation and a variance), allowing for more practical comparisons.

A fourth commonly used metric is **mean absolute percentage error (MAPE)**. MAPE is the average of absolute errors divided by actual observation values.

$$\text{MAPE} = \frac{\sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|}{n} \times 100 \quad (7.4)$$

**TABLE 7.2** Error Metrics for Moving Average Models of Burglary Data

	<i>k</i> = 2	<i>k</i> = 3	Three-Period Weighted
MAD	13.63	14.86	13.70
MSE	254.38	299.84	256.31
RMSE	15.95	17.32	16.01
MAPE	23.63%	26.53%	24.46%

The values of MAD and MSE depend on the measurement scale of the time-series data. For example, forecasting profit in the range of millions of dollars would result in very large MAD and MSE values, even for very accurate forecasting models. On the other hand, market share is measured in proportions; therefore, even bad forecasting models will have small values of MAD and MSE. Thus, these measures have no meaning except in comparison with other models used to forecast the same data. Generally, MAD is less affected by extreme observations and is preferable to MSE if such extreme observations are considered rare events with no special meaning. MAPE is different in that the measurement scale is eliminated by dividing the absolute error by the time-series data value. This allows a better relative comparison. Although these comments provide some guidelines, there is no universal agreement on which measure is best.

These measures can be used to compare the moving average forecasts in Figure 7.6. The results, shown in Table 7.2, verify that the two-period moving average model provides the best forecast among these alternatives.

### SKILL-BUILDER EXERCISE 7.1

Find a four-period moving average forecast for the monthly burglaries data; compute MAD, MSE, RMSE, and MAPE error metrics; and determine if this model is better than the two-period moving average shown in Table 7.2.

## Exponential Smoothing Models

A versatile, yet highly effective approach for short-range forecasting is **simple exponential smoothing**. The basic simple exponential smoothing model is:

$$\begin{aligned} F_{t+1} &= (1 - \alpha)F_t + \alpha A_t \\ &= F_t + \alpha(A_t - F_t) \end{aligned} \tag{7.5}$$

where  $F_{t+1}$  is the forecast for time period  $t + 1$ ,  $F_t$  is the forecast for period  $t$ ,  $A_t$  is the observed value in period  $t$ , and  $\alpha$  is a constant between 0 and 1, called the **smoothing constant**. To begin, the forecast for period 2 is set equal to the actual observation for period 1.

Using the two forms of the forecast equation just given, we can interpret the simple exponential smoothing model in two ways. In the first model, the forecast for the next period,  $F_{t+1}$ , is a weighted average of the forecast made for period  $t$ ,  $F_t$ , and the actual observation in period  $t$ ,  $A_t$ . The second form of the model, obtained by simply rearranging terms, states that the forecast for the next period,  $F_{t+1}$ , equals the forecast for the last period,  $F_t$ , plus a fraction  $\alpha$  of the forecast error made in period  $t$ ,  $A_t - F_t$ . Thus, to make a forecast once we have selected the smoothing constant, we need only know the previous forecast and the actual value. By repeated substitution for  $F_t$  in the equation, it is

easy to demonstrate that  $F_{t+1}$  is a decreasingly weighted average of all past time-series data. Thus, the forecast actually reflects *all* the data, provided that  $\alpha$  is strictly between 0 and 1.

For the burglary data, the forecast for month 43 is 88, which is the actual observation for month 42. Suppose we choose  $\alpha = 0.7$ ; then the forecast for month 44 would be:

$$\text{Month 44 Forecast} = (1 - 0.7)(88) + (0.7)(44) = 57.2$$

The actual observation for month 44 is 60; thus, the forecast for month 45 would be:

$$\text{Month 45 Forecast} = (1 - 0.7)(57.2) + (0.7)(60) = 59.16$$

Since the simple exponential smoothing model requires only the previous forecast and the current time-series value, it is very easy to calculate; thus, it is highly suitable for environments such as inventory systems where many forecasts must be made. The smoothing constant  $\alpha$  is usually chosen by experimentation in the same manner as choosing the number of periods to use in the moving average model. Different values of  $\alpha$  affect how quickly the model responds to changes in the time series. For instance, a value of  $\alpha = 0$  would simply repeat last period's forecast, while  $\alpha = 1$  would forecast last period's actual demand. The closer  $\alpha$  is to 1, the quicker the model responds to changes in the time series because it puts more weight on the actual current observation than on the forecast. Likewise, the closer  $\alpha$  is to 0, the more weight is put on the prior forecast, so the model would respond to changes more slowly.

An Excel spreadsheet for evaluating exponential smoothing models for the burglary data using values of  $\alpha$  between 0.1 and 0.9 is shown in Figure 7.7. A smoothing constant of  $\alpha = 0.6$  provides the lowest error for all three metrics. Excel has a *Data Analysis* tool for exponential smoothing (see Appendix 7.2B, *Forecasting with Exponential Smoothing*).



Spreadsheet Note

	C	D	E	F	G	H	I	J	K	L	M
1	After Citizen-Police Program		Smoothing Constant								
2	Month	Monthly burglaries	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3	42	88	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00
4	43	44	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00
5	44	60	83.60	79.20	74.80	70.40	66.00	61.60	57.20	52.80	48.40
6	45	56	81.24	75.36	70.36	66.24	63.00	60.64	59.16	58.56	58.84
7	46	70	78.72	71.49	66.05	62.14	59.50	57.86	56.95	56.51	56.28
8	47	91	77.84	71.19	67.24	65.29	64.75	65.14	66.08	67.30	68.63
9	48	54	79.16	75.15	74.37	75.57	77.88	80.66	83.53	86.26	88.76
10	49	60	76.64	70.92	68.26	66.94	65.94	64.66	62.86	60.45	57.48
11	50	48	74.98	68.74	65.78	64.17	62.97	61.87	60.86	60.09	59.75
12	51	35	72.28	64.59	60.45	57.70	55.48	53.55	51.86	50.42	49.17
13	52	49	68.55	58.67	52.81	48.62	45.24	42.42	40.06	38.08	36.42
14	53	44	66.60	56.74	51.67	48.77	47.12	46.37	46.32	46.82	47.74
15	54	61	64.34	54.19	49.37	46.86	45.56	44.95	44.70	44.56	44.37
16	55	68	64.00	55.55	52.86	52.52	53.28	54.58	56.11	57.71	59.34
17	56	82	64.40	58.04	57.40	58.71	60.64	62.63	64.43	65.94	67.13
18	57	71	66.16	62.83	64.78	68.03	71.32	74.25	76.73	78.79	80.51
19	58	50	66.65	64.47	66.65	69.22	71.16	72.30	72.72	72.56	71.95
20	59		64.98	61.57	61.65	61.53	60.58	58.92	56.82	54.51	52.20
21		MAD	19.33	17.16	16.15	15.36	14.93	14.71	14.72	14.88	15.36
22		MSE	496.07	390.84	359.18	346.56	340.77	338.41	339.03	343.32	352.38
23		RMSE	22.273	19.77	18.952	18.616	18.46	18.396	18.413	18.529	18.772
24		MAPE	38.28%	32.71%	30.12%	28.36%	27.54%	27.09%	27.09%	27.38%	28.23%

FIGURE 7.7 Exponential Smoothing Forecasts for Burglaries

## SKILL-BUILDER EXERCISE 7.2

Find the best value of the smoothing constant between 0.5 and 0.7 (in increments of 0.05) for exponential smoothing for the burglary data.

## FORECASTING MODELS FOR TIME SERIES WITH A LINEAR TREND

For time series with a linear trend but no significant seasonal components, **double moving average** and **double exponential smoothing** models are more appropriate than using simple moving average or exponential smoothing models. Both methods are based on the linear trend equation:

$$F_{t+k} = a_t + b_t k \quad (7.6)$$

That is, the forecast for  $k$  periods into the future from period  $t$  is a function of a base value  $a_t$ , also known as the *level*, and a *trend*, or slope,  $b_t$ . Double moving average and double exponential smoothing differ in how the data are used to arrive at appropriate values for  $a_t$  and  $b_t$ . Appendix 7.1 describes the computational procedures for these methods.

### Regression-Based Forecasting

Equation 7.6 may look familiar from simple linear regression. We introduced regression in the previous chapter as a means of developing relationships between dependent and independent variables. Simple linear regression can be applied to forecasting using time as the independent variable. For example, Figure 7.8 shows a portion of the Excel file *Coal Production*, which provides data on total tons produced from 1960 through 2007. A linear trendline shows an  $R^2$  value of 0.969 (the fitted model assumes that the years

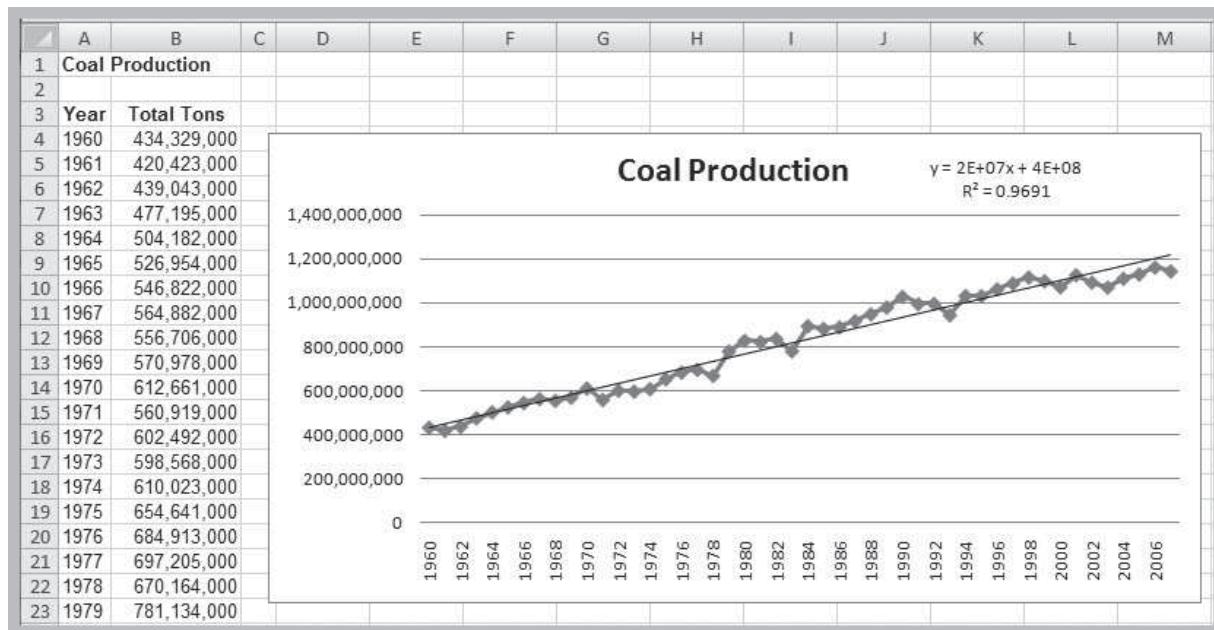


FIGURE 7.8 Regression-Based Forecast for Coal Production

are numbered 1 through 48, not as actual dates). The actual values of the coefficients in the model are shown below:

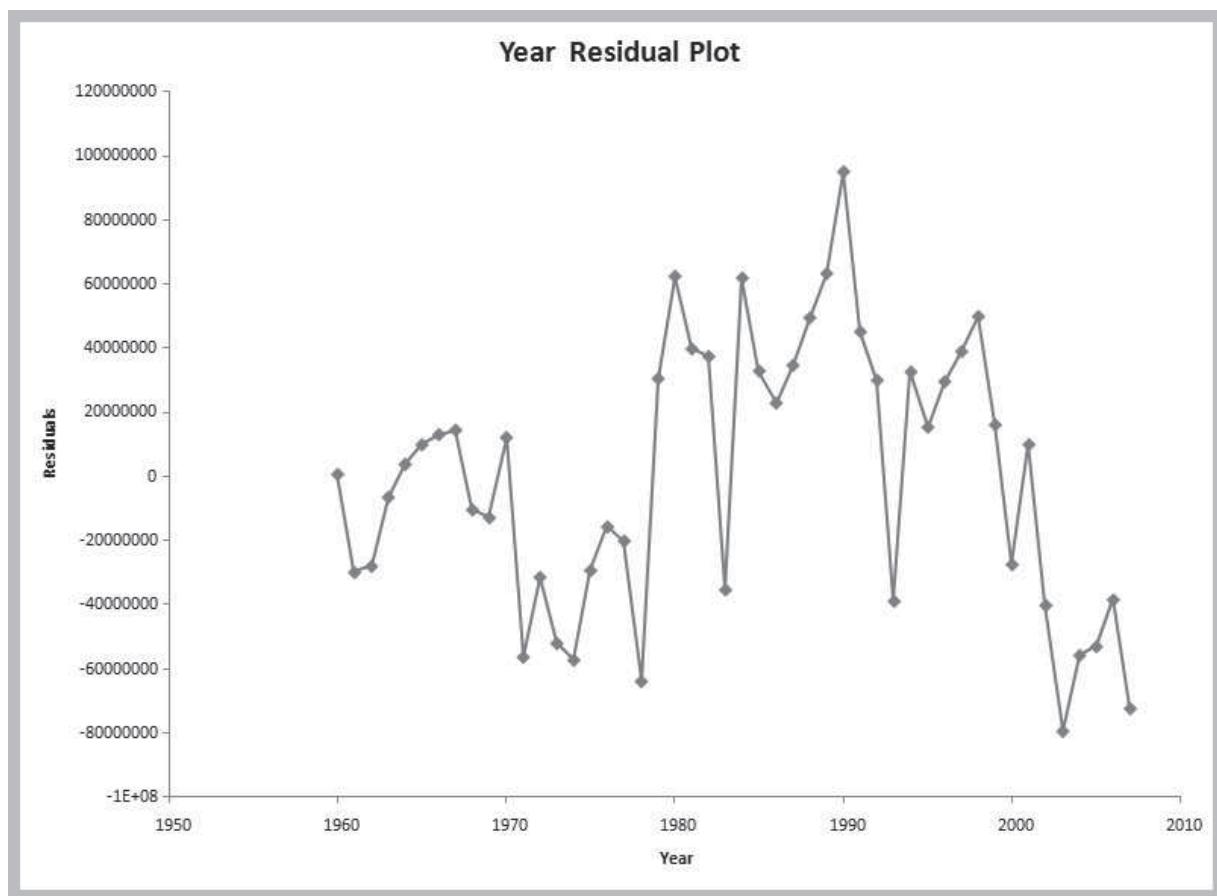
$$\text{Tons} = 416,896,322.7 + 16,685,398.57 \times \text{Year}$$

Thus, a forecast for 2008 would be:

$$\text{Tons} = 416,896,322.7 + 16,685,398.57 \times (49) = 1,234,480,853$$

## ADVANCED FORECASTING MODELS

In Chapter 6, we noted that an important assumption for using regression analysis is the lack of autocorrelation among the data. When autocorrelation is present, successive observations are correlated with one another; for example, large observations tend to follow other large observations, and small observations also tend to follow one another. This can often be seen by examining the residual plot when the data are ordered by time. Figure 7.9 shows the time-ordered residual plot from *PHStat* for the coal production regression. The residuals do not appear to be random; rather, successive observations seem to be related to one another. We introduced the Durbin–Watson statistic in Chapter 6 as a means of evaluating autocorrelation. From the *PHStat* regression tool, the Durbin–Watson statistic was computed to be 0.768, which suggests autocorrelation, indicating that other approaches, called *autoregressive models*, are more appropriate.



**FIGURE 7.9** Residual Plot for Coal Production Regression

## Autoregressive Forecasting Models

An autoregressive forecasting model incorporates correlations between consecutive values in a time series by using explanatory variables that are lagged values of the dependent variable. A first-order autocorrelation refers to the correlation among data values one period apart, a second-order autocorrelation refers to the correlation among data values two periods apart, and so on. Autoregressive models improve forecasting when autocorrelation is present in data. A **first-order autoregressive model** is:

$$Y_i = a_0 + a_1 Y_{i-1} + \delta_i \quad (7.7)$$

where  $Y_i$  is the value of the time series in period  $i$  and  $\delta_i$  is a nonautocorrelated random error term having 0 mean and constant variance. A **second-order autoregressive model** is:

$$Y_i = a_0 + a_1 Y_{i-1} + a_2 Y_{i-2} + \delta_i \quad (7.8)$$

Additional terms may be added for higher-order models.

To build an autoregressive model using multiple linear regression, we simply add additional columns to the data matrix for the dependent variable that lag the original data by some number of periods. Thus, for a third-order autoregressive model, we add columns that lag the dependent variable by one, two, and three periods. For the coal production data, a portion of this data matrix is shown in Figure 7.10. However, we must be careful to use the appropriate data when running the regression models. For example, to run a third-order model, we must not include the first three rows of the matrix, since there are no Lag 3 values for those years. Similarly, for a second-order model, do not include the data for the first two rows.

Suppose that we try a third-order model first. Using the multiple regression tool, we obtain the results shown in Figure 7.11. Note that the largest  $p$ -value (for the third-order term) exceeds  $\alpha = 0.05$ , indicating that this variable is not significant. Dropping it and re-running a second-order model yields the results shown in Figure 7.12. Thus, we can use the second-order model to forecast the coal production:

$$\text{Tons} = 43,087,157 + 0.632 \times \text{Tons}_{\text{Year}-1} + 0.341 \times \text{Tons}_{\text{Year}-2}$$

A forecast for the year 2008 would be:

$$\text{Tons} = 43,087,157 + 0.632 \times \text{Tons}_{2007} + 0.341 \times \text{Tons}_{2006}$$

	A	B	C	D	E
1	Coal Production				
2					
3	Year	Total Tons	Year - 1	Year - 2	Year - 3
4	1960	434,329,000			
5	1961	420,423,000	434,329,000		
6	1962	439,043,000	420,423,000	434,329,000	
7	1963	477,195,000	439,043,000	420,423,000	434,329,000
8	1964	504,182,000	477,195,000	439,043,000	420,423,000
9	1965	526,954,000	504,182,000	477,195,000	439,043,000
10	1966	546,822,000	526,954,000	504,182,000	477,195,000

**FIGURE 7.10** Portion of Data Matrix for Autoregressive Forecasting of Coal Production

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.988073439					
5	R Square	0.97628912					
6	Adjusted R Square	0.974554178					
7	Standard Error	35247696.59					
8	Observations	45					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	3	2.09738E+18	6.99125E+17	562.721337	2.49811E-33	
13	Residual	41	5.09384E+16	1.2424E+15			
14	Total	44	2.14831E+18				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	50332718.33	20917609.66	2.406236618	0.020709694	8088749.075	92576687.59
18	Year - 1	0.565292456	0.154196228	3.666058917	0.000700825	0.25388686	0.876698052
19	Year - 2	0.247260179	0.174243467	1.41904993	0.163442473	-0.104631637	0.599151996
20	Year - 3	0.156914847	0.151580327	1.035192694	0.306646344	-0.14920783	0.463037525

FIGURE 7.11 Results for Third-Order Autoregressive Model

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.988567258					
5	R Square	0.977265224					
6	Adjusted R Square	0.976207792					
7	Standard Error	34986337.08					
8	Observations	46					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	2	2.26249E+18	1.13125E+18	924.1877675	4.66498E-36	
13	Residual	43	5.26339E+16	1.22404E+15			
14	Total	45	2.31513E+18				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	43087157.38	19471726.51	2.212806213	0.032269724	3818678.935	82355635.83
18	Year - 1	0.63224491	0.142019472	4.451818489	5.95528E-05	0.345835353	0.918654467
19	Year - 2	0.341258327	0.141128508	2.418067984	0.019912562	0.056645569	0.625871085

FIGURE 7.12 Second-Order Autoregressive Forecasting Model

or

$$\begin{aligned}\text{Tons} &= 43,087,157 + 0.632 \times 1,145,567,167 + 0.341 \times 1,162,749,659 \\ &= 1,163,583,240\end{aligned}$$

A forecast for the next year (2009) would be:

$$\text{Tons} = 43,087,157 + 0.632 \times \text{Tons}_{2008} + 0.341 \times \text{Tons}_{2007}$$

or

$$\begin{aligned}\text{Tons} &= 43,087,157 + 0.632 \times 1,163,583,240 + 0.341 \times 1,145,567,167 \\ &= 1,169,110,169\end{aligned}$$

Finding the appropriate number of lagged variables to start with generally requires some trial and error or further analysis. More sophisticated statistical software can provide autocorrelations between the dependent variable and lagged variables; those with high autocorrelations provide a starting point for the model. See the discussion in the *CB Predictor* section at the end of this chapter.

### SKILL-BUILDER EXERCISE 7.3

Set up and fit a first-order autoregressive model for the coal production example. Compare the results to the second-order model.

As with regression in general, building regression-based forecasting models is as much of an art as science. Because the coal production data appears to have a linear trend, we might include time as an additional variable in the autoregressive model:

$$Y_i = a_0 + a_1 Y_{i-1} + a_2 Y_{i-2} + a_3 T + \delta_i$$

where  $T$  might be coded as (Year–1960); that is, the number of years since 1960, to simplify the model. A problem at the end of the chapter asks you to analyze such a model to improve the forecasts obtained from the basic autoregressive model.

## Forecasting Models with Seasonality

When time series exhibit seasonality and possibly also trend, different techniques provide better forecasts than the ones we have described. The computational theory behind these models is presented in Appendix 7.1. However, a basic understanding of these techniques is useful in order to apply *CB Predictor* software for forecasting, which we introduce later in this chapter.

Seasonal factors (with no trend) can be incorporated into a forecast by adjusting the level,  $a_t$ , in one of two ways. The **seasonal additive** model is:

$$F_{t+k} = a_t + S_{t-s+k} \tag{7.9}$$

and the **seasonal multiplicative** model is:

$$F_{t+k} = a_t S_{t-s+k} \tag{7.10}$$

In both models,  $S_{t-s+k}$  is the seasonal factor for period  $t - s + k$  and  $s$  is the number of periods in a season. A “season” can be a year, quarter, month, or even a week, depending on the application. In any case, the forecast for period  $t + k$  is adjusted up or down from a level ( $a_t$ ) by the seasonal factor. The multiplicative model is more appropriate when the seasonal factors are increasing or decreasing over time. This is evident when the amplitude of the time series changes over time.

## Incorporating Seasonality in Regression Models

Quite often time-series data exhibit seasonality, especially on an annual basis. For example, the data in the Excel file *Gas & Electric* provides two years of data for natural gas and electric usage for a residential property (see Figure 7.13). Multiple linear regression models with categorical variables can be used for time series with seasonality. To do this, we use dummy categorical variables for the seasonal components. With monthly data, as we have for natural gas usage, we have a seasonal categorical variable with  $k = 12$  levels. As discussed in Chapter 6, we construct the regression model using  $k - 1$  dummy variables. We will use January as the reference month; therefore, this variable does not appear in the model:

$$\begin{aligned}\text{Gas usage} = \beta_0 + \beta_1 \text{Time} + \beta_2 \text{February} + \beta_3 \text{March} + \beta_4 \text{April} + \beta_5 \text{May} \\ + \beta_6 \text{June} + \beta_7 \text{July} + \beta_8 \text{August} + \beta_9 \text{September} + \beta_{10} \text{October} \\ + \beta_{11} \text{November} + \beta_{12} \text{December}\end{aligned}$$

This coding scheme results in the data matrix shown in Figure 7.14. This model picks up trends from the regression coefficient for time, and seasonality from the dummy variables for each month. The forecast for the next January will be  $\beta_0 + \beta_1(25)$ . The variable coefficients (betas) for each of the other 11 months will show the adjustment relative to January. For example, forecast for next February would be  $\beta_0 + \beta_1(25) + \beta_2(1)$ , and so on.

Figure 7.15 shows the results of using the *Regression* tool in Excel after eliminating insignificant variables (Time and Feb). Because the data shows no clear linear trend,

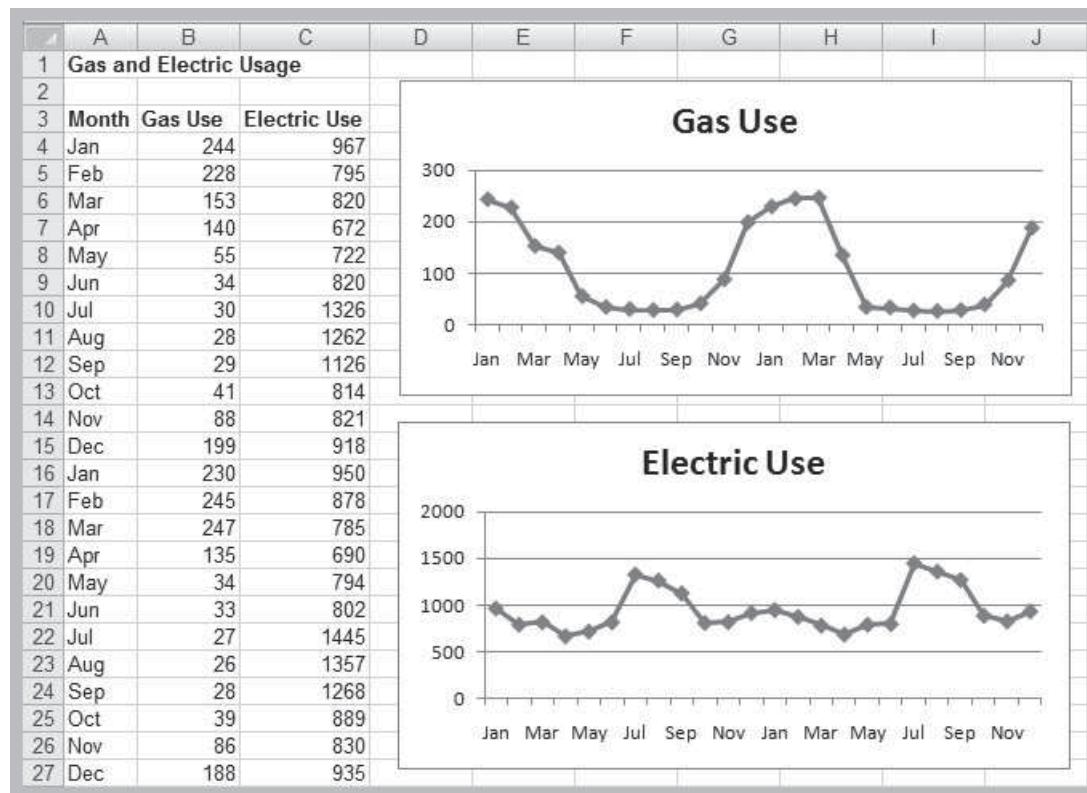


FIGURE 7.13 Gas & Electric Data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
3	Month	Gas Use	Time	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
4	Jan	244	1	0	0	0	0	0	0	0	0	0	0	0
5	Feb	228	2	1	0	0	0	0	0	0	0	0	0	0
6	Mar	153	3	0	1	0	0	0	0	0	0	0	0	0
7	Apr	140	4	0	0	1	0	0	0	0	0	0	0	0
8	May	55	5	0	0	0	1	0	0	0	0	0	0	0
9	Jun	34	6	0	0	0	0	1	0	0	0	0	0	0
10	Jul	30	7	0	0	0	0	0	1	0	0	0	0	0
11	Aug	28	8	0	0	0	0	0	0	1	0	0	0	0
12	Sep	29	9	0	0	0	0	0	0	0	1	0	0	0
13	Oct	41	10	0	0	0	0	0	0	0	0	0	1	0
14	Nov	88	11	0	0	0	0	0	0	0	0	0	0	1
15	Dec	199	12	0	0	0	0	0	0	0	0	0	0	1
16	Jan	230	13	0	0	0	0	0	0	0	0	0	0	0
17	Feb	245	14	1	0	0	0	0	0	0	0	0	0	0
18	Mar	247	15	0	1	0	0	0	0	0	0	0	0	0
19	Apr	135	16	0	0	1	0	0	0	0	0	0	0	0
20	May	34	17	0	0	0	1	0	0	0	0	0	0	0
21	Jun	33	18	0	0	0	0	1	0	0	0	0	0	0
22	Jul	27	19	0	0	0	0	0	1	0	0	0	0	0
23	Aug	26	20	0	0	0	0	0	0	1	0	0	0	0
24	Sep	28	21	0	0	0	0	0	0	0	1	0	0	0
25	Oct	39	22	0	0	0	0	0	0	0	0	0	1	0
26	Nov	86	23	0	0	0	0	0	0	0	0	0	1	0
27	Dec	188	24	0	0	0	0	0	0	0	0	0	0	1

**FIGURE 7.14** Data Matrix for Seasonal Regression Model

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.985480895					
5	R Square	0.971172595					
6	Adjusted R Square	0.948997667					
7	Standard Error	19.54432831					
8	Observations	24					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	10	167292.2083	16729.22083	43.79597661	2.33344E-08	
13	Residual	13	4965.75	381.9807692			
14	Total	23	172257.9583				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	236.75	9.772164157	24.22697738	3.33921E-12	215.6385229	257.8614771
18	Mar	-36.75	16.92588482	-2.171230656	0.04901621	-73.31615098	-0.183849024
19	Apr	-99.25	16.92588482	-5.863799799	5.55744E-05	-135.816151	-62.68384902
20	May	-192.25	16.92588482	-11.35834268	4.02824E-08	-228.816151	-155.683849
21	Jun	-203.25	16.92588482	-12.00823485	2.07264E-08	-239.816151	-166.683849
22	Jul	-208.25	16.92588482	-12.30364038	1.54767E-08	-244.816151	-171.683849
23	Aug	-209.75	16.92588482	-12.39226204	1.41949E-08	-246.316151	-173.183849
24	Sep	-208.25	16.92588482	-12.30364038	1.54767E-08	-244.816151	-171.683849
25	Oct	-196.75	16.92588482	-11.62420766	3.05791E-08	-233.316151	-160.183849
26	Nov	-149.75	16.92588482	-8.847395666	7.30451E-07	-186.316151	-113.183849
27	Dec	-43.25	16.92588482	-2.555257847	0.023953114	-79.81615098	-6.683849024

**FIGURE 7.15** Final Regression Model for Forecasting Gas Usage

the variable Time could not explain any significant variation in the data. The dummy variable for February was probably insignificant because the historical gas usage for both January and February were very close to each other. The  $R^2$  for this model is 0.971, which is very good. The final regression model is:

$$\begin{aligned}\text{Gas Usage} = & 236.75 - 36.75 \text{ March} - 99.25 \text{ April} - 192.25 \text{ May} - 203.25 \text{ June} \\ & - 208.25 \text{ July} - 209.75 \text{ August} - 208.25 \text{ September} - 196.75 \text{ October} \\ & - 149.75 \text{ November} - 43.25 \text{ December}\end{aligned}$$

### SKILL-BUILDER EXERCISE 7.4

Find the best multiple regression model for Electric Use in the *Gas & Electric* Excel file using the approach for incorporating seasonality.

### Forecasting Models with Trend and Seasonality

Many time series exhibit both trend and seasonality. Such might be the case for growing sales of a seasonal product. The methods we describe are based on the work of two researchers: C.C. Holt, who developed the basic approach; and P.R. Winters, who extended Holt's work. Hence, these approaches are commonly referred to as **Holt-Winters models**. These models combine elements of both the trend and seasonal models described above. The Holt-Winters additive model is based on the equation:

$$F_{t+1} = a_t + b_t + S_{t-s+1} \quad (7.11)$$

and the Holt-Winters multiplicative model is:

$$F_{t+1} = (a_t + b_t)S_{t-s+1} \quad (7.12)$$

The additive model applies to time series with relatively stable seasonality, while the multiplicative model applies to time series whose amplitude increases or decreases over time.

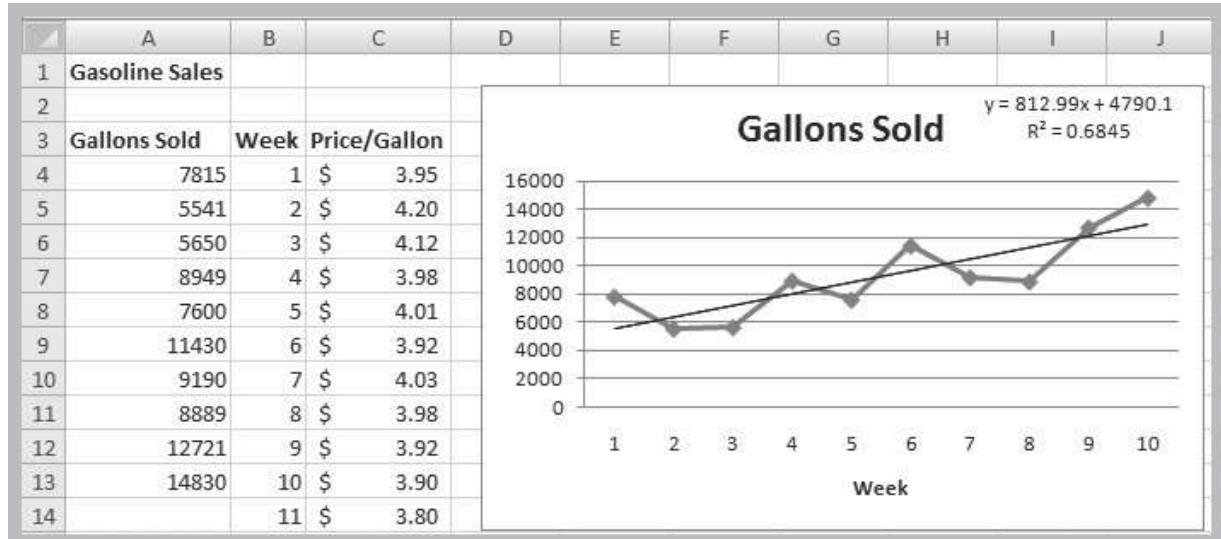
Table 7.3 summarizes the choice of non-regression-based models based on characteristics of the time series.

### Regression Forecasting with Causal Variables

In many forecasting applications, other independent variables such as economic indexes or demographic factors may influence the time series, and can be incorporated into a regression model. For example, a manufacturer of hospital equipment might include such variables as hospital capital spending and changes in the proportion of people over the age of 65 in building models to forecast future sales.

**TABLE 7.3 Forecasting Model Choice**

	No Seasonality	Seasonality
No Trend	Single moving average or single exponential smoothing	Seasonal additive or seasonal multiplicative model
Trend	Double moving average or double exponential smoothing	Holt-Winters additive or Holt-Winters multiplicative model



**FIGURE 7.16** Gasoline Sales Data and Trendline

To illustrate the use of multiple linear regression for forecasting with causal variables, suppose that we wish to forecast gasoline sales. Figure 7.16 shows the sales over 10 weeks during June through August along with the average price per gallon and a chart of the gasoline sales time series with a fitted trendline (Excel file *Gasoline Sales*). During the summer months, it is not unusual to see an increase in sales as more people go on vacations. The chart shows a linear trend, although  $R^2$  is not very high.

The trendline is:

$$\text{Sales} = 4,790.1 + 812.99 \text{ Week}$$

Using this model, we would predict sales for week 11 as:

$$\text{Sales} = 4,790.1 + 812.99(11) = 13,733 \text{ gallons}$$

However, we also see that the average price per gallon changes each week, and this may influence consumer sales. Therefore, the sales trend might not simply be a factor of steadily increasing demand, but might also be influenced by the average price per gallon. The average price per gallon can be considered as a **causal variable**. Multiple linear regression provides a technique for building forecasting models that incorporate not only time, but other potential causal variables also. Thus, to forecast gasoline sales, we propose a model using two independent variables (Week and Price/Gallon).

$$\text{Sales} = \beta_0 + \beta_1 \text{ Week} + \beta_2 \text{ Price/Gallon}$$

The results are shown in Figure 7.17 and the regression model is:

$$\text{Sales} = 72,333.08 + 508.67 \text{ Week} - 16,463.2 \text{ Price/Gallon}$$

This makes sense because as price changes, sales typically reflect the change. Notice that the  $R^2$  value is higher when both variables are included, explaining more than 86% of the variation in the data. If the company estimates that the average price for the next week will drop to \$3.80, the model would forecast the sales for week 11 as:

$$\text{Sales} = 72,333.08 + 508.67(11) - 16,463.2(3.80) = 15,368 \text{ gallons}$$

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.930528528					
5	R Square	0.865883342					
6	Adjusted R Square	0.827564297					
7	Standard Error	1235.400329					
8	Observations	10					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	2	68974748.7	34487374.35	22.59668368	0.000883465	
13	Residual	7	10683497.8	1526213.972			
14	Total	9	79658246.5				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	72333.08447	21969.92267	3.292368642	0.013259225	20382.47253	124283.6964
18	Week	508.6681395	168.1770861	3.024598364	0.019260863	110.9925233	906.3437558
19	Price/Gallon	-16463.19901	5351.082403	-3.076611005	0.017900405	-29116.49823	-3809.899789

**FIGURE 7.17** Regression Results for Gas Sales

Notice that this is higher than the pure time-series forecast because of the sensitivity to the price per gallon.

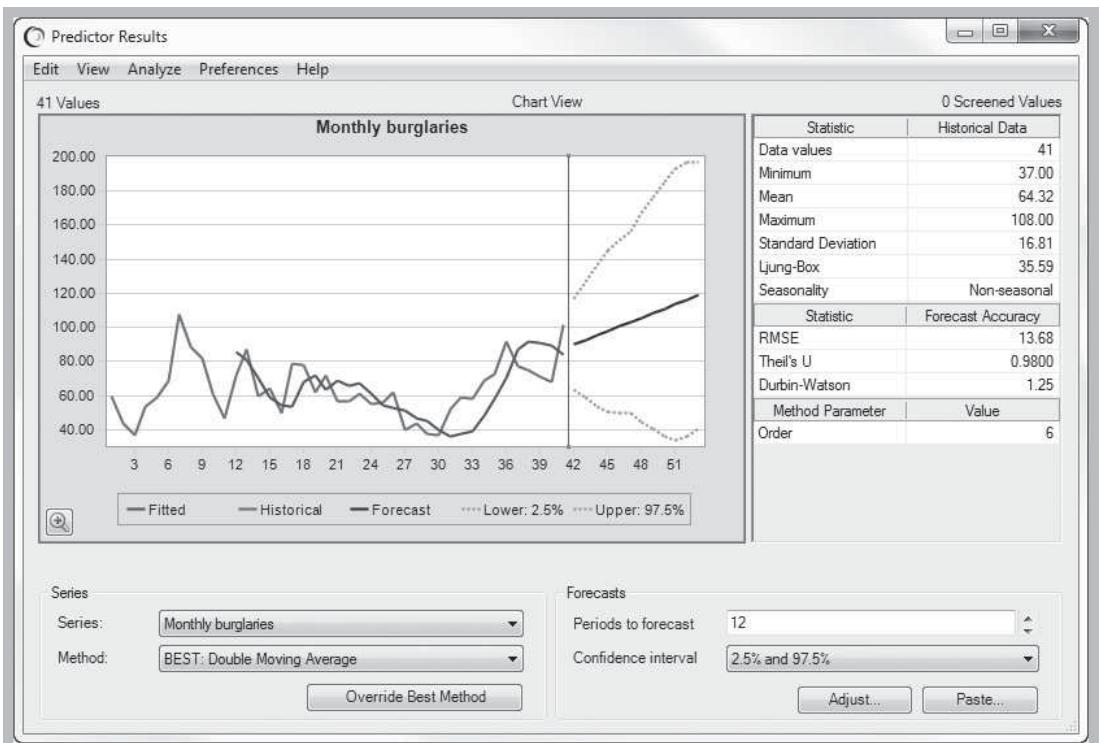
## CHOOSING AND OPTIMIZING FORECASTING MODELS USING CB PREDICTOR

*CB Predictor* is an Excel add-in for forecasting that is part of the *Crystal Ball* suite of applications. *CB Predictor* can be used as a stand-alone program for forecasting, and can also be integrated with Monte Carlo simulation, which we discuss in Chapter 10. *CB Predictor* includes all the time-series forecasting approaches we have discussed. See Appendix 7.2C, *Using CB Predictor* for basic information on using the add-in.

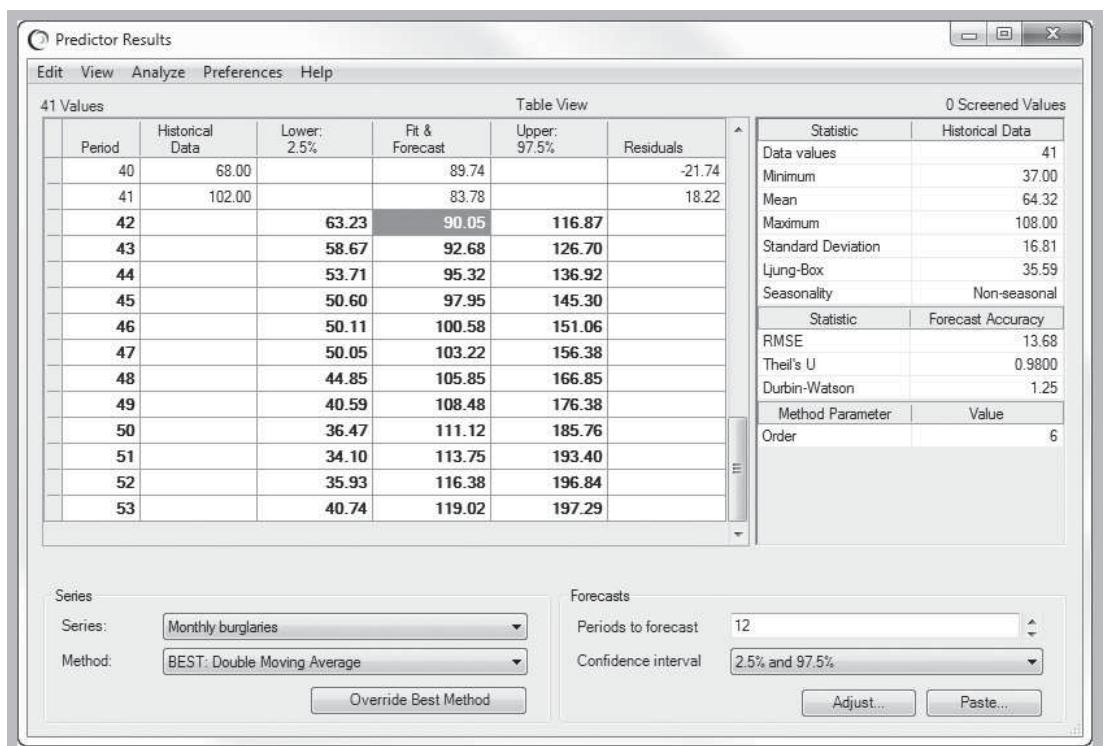
We will illustrate the use of *CB Predictor* first for the data in the worksheet *Burglaries* before the citizen-police program commenced. (Appendix 7.2C uses this example to illustrate the input data for *CB Predictor*.) Figure 7.18 shows the results. The best forecasting method that was identified is the double moving average. Statistics for the other methods considered can be found by selecting the alternative forecasting methods in the drop-down box. On the right side of the window, *CB Predictor* displays the value of RMSE, Theil's U, and Durbin-Watson statistic. Theil's U statistic is a relative error measure that compares the results with a naive forecast. A value less than 1 means that the forecasting technique is better than guessing; a value equal to 1 means that the technique is about as good as guessing; and a value greater than 1 means that the forecasting technique is worse than guessing. *CB Predictor* also identifies the best parameters for the forecasting method. In this example, number of periods for the double moving average forecast (called the "order" in the output) is 6. The forecasts for the next 12 months can be found by choosing *Table* from the *View* menu and are shown in Figure 7.19.



Spreadsheet Note



**FIGURE 7.18** CB Predictor Results for Forecasting Monthly Burglaries



**FIGURE 7.19** CB Predictor Table View for Monthly Burglary Forecasts

## SKILL-BUILDER EXERCISE 7.5

Use *CB Predictor* to find the best forecasting model for the burglary example.

As a second example, we will use the seasonal data in the Excel file *Gas & Electric* that we discussed earlier in this chapter. *CB Predictor* identifies both time series (gas use and electric use) in the *Input Data* dialog. In the *Data Attributes* dialog, the *AutoDetect* option identifies 12 month seasonality in the data. Although the data are clearly seasonal, we will select both seasonal and nonseasonal methods in the *Method Gallery* dialog. Figure 7.20 shows the results for gas use, and Figure 7.21 shows the results for electric use. For gas use, the best method identified is double exponential smoothing, with  $\alpha = 0.9990$  and  $\beta = 0.8072$ . However, Theil's U is greater than 1, indicating that the model may not be adequate. However, this may be due to the limited number of observations in the data set. For electric use, the Holt-Winters multiplicative method was found to be the best.

*CB Predictor* provides useful information to build autoregressive models. If you click on the *View Seasonality* button in the *Data Attributes* dialog (Figure 7.22), you will see another button at the lower right called *View Autocorrelations*. Clicking on this shows the autocorrelations for various lag values. Figure 7.23 illustrates this for the *Coal Production* data that we used to introduce autoregressive models. The strongest autocorrelations are shown in the table on the right; we see that these occur for the first three lags. The *p*-values show that the first two are significant. Thus, in building an autoregressive model, this suggests that you start with a second-order autoregressive model, which is

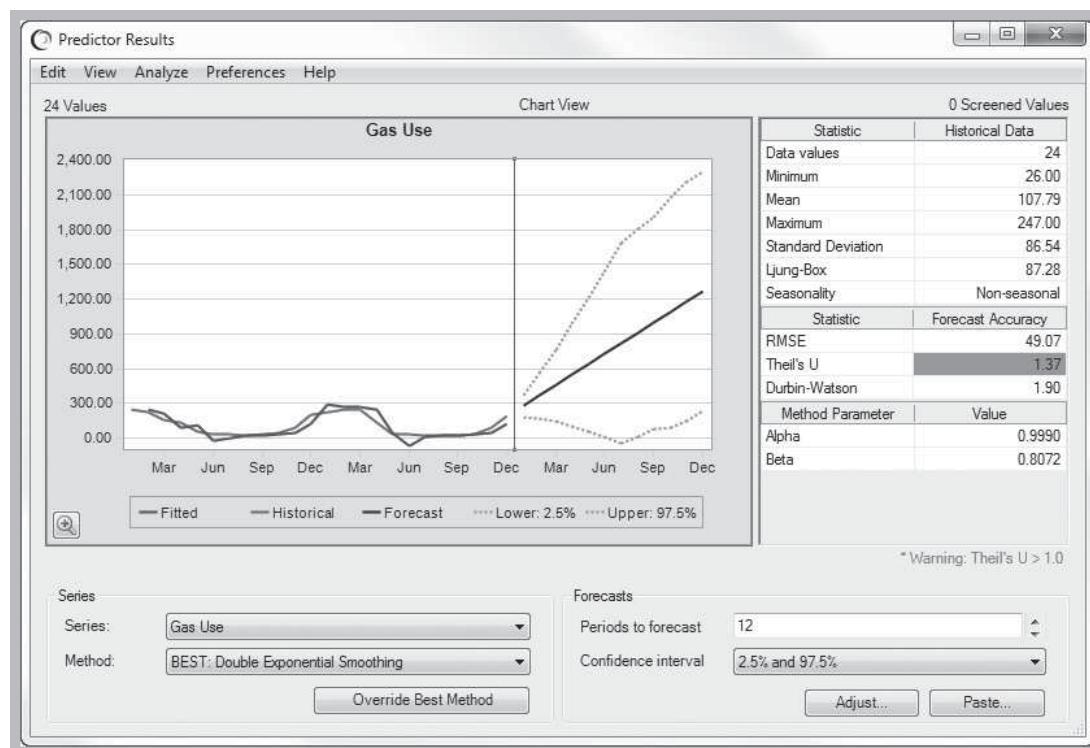
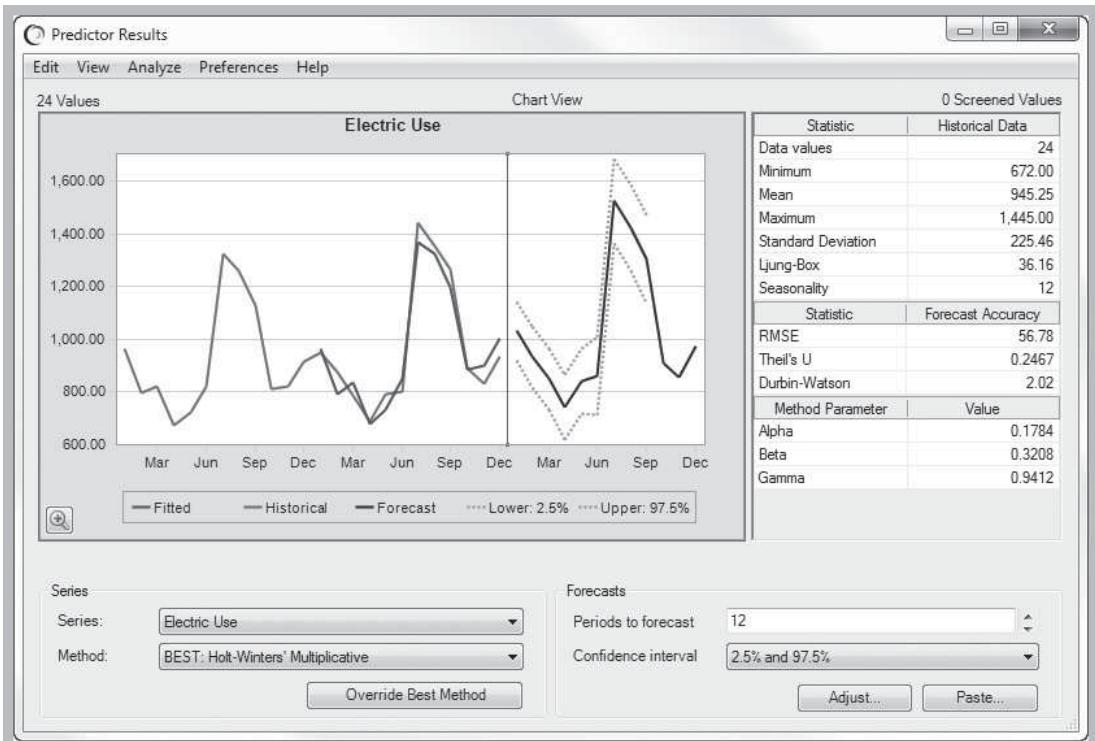


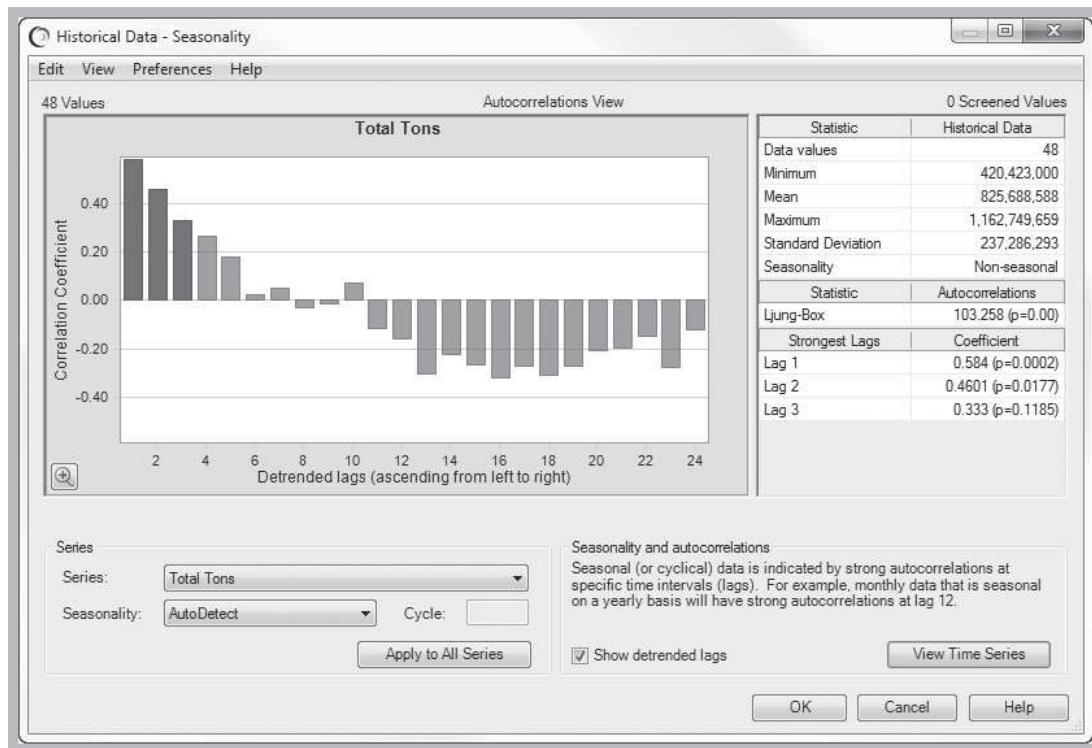
FIGURE 7.20 CB Predictor Results for Gas Usage



**FIGURE 7.21** CB Predictor Results for Electric Usage



**FIGURE 7.22** CB Predictor Seasonality Dialog



**FIGURE 7.23** Autocorrelations Calculated by *CB Predictor*

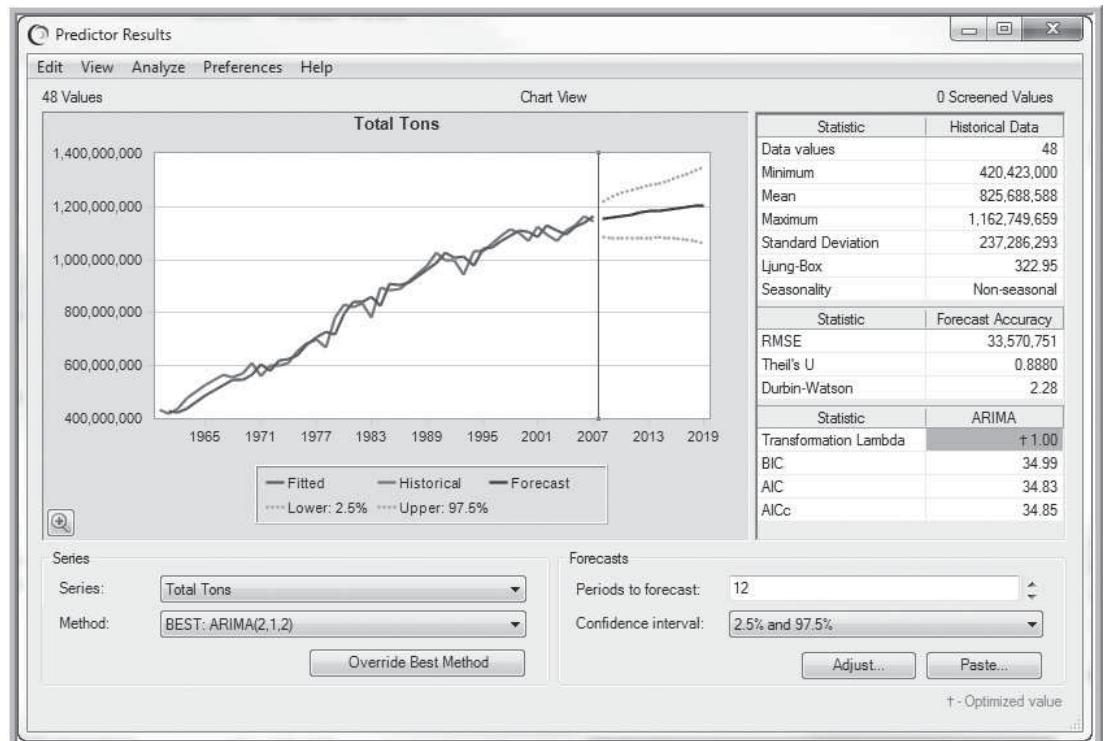
what we concluded earlier as the best for the example. An estimate of the standard error of an autocorrelation coefficient is  $1/\sqrt{n}$ , where  $n$  is the number of observations (in this case, 48). Using the empirical rules, a 95% confidence interval is approximately  $2/\sqrt{n}$ . Thus autocorrelations with absolute value more than  $2/\sqrt{n} = 0.29$  can be deemed as significant. Therefore, you might want to consider an order 3 model as a starting point.

*Crystal Ball* version 11.1.2.1, released in early 2011, adds a new option to identify the best *autoregressive integrated moving average* (ARIMA) model. The autoregressive models we discussed earlier in this chapter are special cases of the general ARIMA model. Understanding the details of the models and the procedure, particularly the optional settings, requires advanced statistical knowledge that is beyond the scope of this book. To illustrate the procedure, however, Figure 7.24 shows the results for the coal production example using the default settings. The best model identified by *CB Predictor* is an ARIMA( $p, d, q$ ) model, where  $p$  is the number of autoregressive terms,  $d$  specifies the type of growth behavior ( $d = 0$ , no trend;  $d = 1$ , linear growth;  $d = 2$ , quadratic growth, etc.), and  $q$  indicates the number of lagged forecast error terms. In this case, the procedure identified an ARIMA(2, 1, 2) model, which has the form:

$$Y_i = a_0 + a_1 Y_{i-1} + a_2 Y_{i-2} + a_3(Y_{i-1} - Y_{i-2}) - a_4 e_{i-1} - a_5 e_{i-2}$$

where  $e_{i-1}$  is the error in period  $i - 1$ . The actual fit and forecast also depends on the error metric chosen (in this case, RMSE).

This version of *Crystal Ball* also allows you to add events to forecasting with *CB Predictor*. An **event** is an identifiable occurrence that has affected historical data and could affect predicted data. These events can be one-time occurrences, such as a storm, or events that repeat on a regular basis, such as quarterly sales promotions. For example, gasoline prices have been affected by hurricanes in the Gulf of Mexico and civil unrest



**FIGURE 7.24** CB Predictor Results of ARIMA Model for Coal Production

in the Middle East. Such events help to explain unusual spikes in time series. They can be added easily in the *Data Attributes* dialog by clicking the *View Events* button.

## THE PRACTICE OF FORECASTING

In practice, managers use a variety of judgmental and quantitative forecasting techniques. Statistical methods alone cannot account for such factors as sales promotions, unusual environmental disturbances, new product introductions, large one-time orders, and so on. Many managers begin with a statistical forecast and adjust it to account for intangible factors. Others may develop independent judgmental and statistical forecasts then combine them, either objectively by averaging or in a subjective manner. It is impossible to provide universal guidance as to which approaches are best, for they depend on a variety of factors, including the presence or absence of trends and seasonality, the number of data points available, length of the forecast time horizon, and the experience and knowledge of the forecaster. Often, quantitative approaches will miss significant changes in the data, such as reversals of trends, while qualitative forecasts may catch them, particularly when using indicators as discussed earlier in this chapter.

Here we briefly highlight three practical examples of forecasting and encourage you to read the full articles cited for better insight into the practice of forecasting.

- Allied-Signal's Albuquerque Microelectronics Operation (AMO) produced radiation-hardened microchips for the U.S. Department of Energy (DOE). In 1989, a decision was made to close a plant, but operations at AMO had to be phased out over several years because of long-term contractual obligations. AMO experienced fairly erratic yields in the production of some of its complex microchips, and accurate

forecasts of yields were critical. Overestimating yields could lead to an inability to meet contractual obligations in a timely manner, requiring the plant to remain open longer. Underestimates would cause AMO to produce more chips than actually needed. AMO's yield forecasts had previously been made by simply averaging all historical data. More sophisticated forecasting techniques were implemented, resulting in improved forecasts of wafer fabrication. Using more accurate yield forecasts and optimization models, AMO was able to close the plant sooner, resulting in significant cost savings.<sup>1</sup>

- More than 70% of the total sales volume at L.L. Bean is generated through orders to its call center. Calls to the L.L. Bean call center are classified into two types: telemarketing (TM), which involves placing an order; and telephone inquiry (TI), which involves customer inquiries such as order status or order problems. Accurately forecasting TM and TI calls helps the company better plan the number of agents to have on hand at any point in time. Analytical forecasting models for both types of calls take into account historical trends, seasonal factors, and external explanatory variables such as holidays and catalog mailings. The estimated benefit from better precision from the two forecasting models is approximately \$300,000 per year.<sup>2</sup>
- DIRECTV was founded in 1991 to provide subscription satellite television. Prior to launching this product, it was vital to forecast how many homes in the United States would subscribe to satellite television, and when. A forecast was developed using the Bass diffusion model, which describes the adoption pattern of new products and technologies. The model is based on the proposition that the conditional probability of adoption by potential consumers of a new product at a given time will be a linear-increasing function of the number of previous adopters. The model was supported by forecasting analogies with cable TV. The 1992 forecast proved to be quite good in comparison with actual data over the five-year period from 1994 through 1999.<sup>3</sup>

## Basic Concepts Review Questions

1. What is a time series? Why is the analysis of a time series important? What is meant by forecasting?
2. Describe the operational format of the Delphi method.
3. What does the trend component of a time series represent?
4. Describe the seasonal effect and cyclical effect components of a time series.
5. Summarize statistical methods used in forecasting and the types of time series to which they are most appropriate.
6. Explain how a simple moving average is calculated.
7. List and define the three principal ways of measuring forecast accuracy. What are the key differences among them?
8. Explain how the exponential smoothing model can be interpreted in two different ways.
9. Which methods work well on a stationary time series?
10. What kind of forecasting models are most appropriate when significant autocorrelation is present in the data?
11. How are dummy variables used in regression forecasting models with seasonality?
12. What is a causal variable in forecasting? Provide an example from your experience of some applications where causal variables might be used in a forecast.
13. What are the advantages of using *CB Predictor* for forecasting?
14. Summarize some of the practical issues in using forecasting tools and approaches.

<sup>1</sup> D.W. Clements and R.A. Reid, "Analytical MS/OR Tools Applied to a Plant Closure," *Interfaces* 24, no. 2 (March–April, 1994): 1–12.

<sup>2</sup> B.H. Andrews and S.M. Cunningham, "L.L. Bean Improves Call-Center Forecasting," *Interfaces* 25, no. 6 (November–December, 1995): 1–13.

<sup>3</sup> Frank M. Bass, Kent Gordon, and Teresa L. Ferguson, "DIRECTV: Forecasting Diffusion of a New Technology Prior to Product Launch," *Interfaces* 31, no. 3 (May–June 2001): Part 2 of 2, S82–S93.

## Problems and Applications

1. The Excel file *Closing Stock Prices* provides data for four stocks over a one-month period.
  - a. Develop spreadsheet models for forecasting each of the stock prices using single moving average and single exponential smoothing.
  - b. Using MAD, MSE, and MAPE as guidance, find the best number of moving average periods and best smoothing constant for exponential smoothing.
2. Use the data in the Excel file *Baseball Attendance* to do the following:
  - a. Develop spreadsheet models for forecasting attendance using single moving average and single exponential smoothing.
  - b. Using MAD, MSE, and MAPE as guidance, find the best number of moving average periods and best smoothing constant for exponential smoothing.
3. For the data in the Excel file *Ohio Prison Population* do the following:
  - a. Develop spreadsheet models for forecasting both male and female populations using single moving average and single exponential smoothing.
  - b. Using MAD, MSE, and MAPE as guidance, find the best number of moving average periods and best smoothing constant for exponential smoothing.
4. For the data in the Excel file *Gasoline Prices* do the following:
  - a. Develop spreadsheet models for forecasting prices using single moving average and single exponential smoothing.
  - b. Using MAD, MSE, and MAPE as guidance, find the best number of moving average periods and best smoothing constant for exponential smoothing.
5. In the Excel file *Treasury Yield Rates*, develop spreadsheet models for forecasting the one month treasury yield. Use simple moving average models with  $k = 2$  and  $k = 3$ , and compare them on the basis of the usual error metrics.
6. Consider the data in the Excel file *Consumer Price Index*.
  - a. Use simple linear regression to forecast the data. What would be the forecasts for the next two months?
  - b. Are the data autocorrelated? Construct first- and second-order autoregressive models and compare the results to part (a).
7. In the Excel file *Surgery Infections*, develop spreadsheet models for forecasting the surgery infection rates. Use an exponential smoothing model with smoothing constant  $\alpha = 0.5$ , and weighted moving average with  $k = 3$  and weights 10%, 20% and 70%. Compare them on the basis of the usual error metrics.
8. Consider the data in the Excel file *Nuclear Power*.
  - a. Use simple linear regression to forecast the data. What would be the forecasts for the next three years?
  - b. Are the data autocorrelated? Construct first- and second-order autoregressive models and compare the results to part (a).
9. Find the best autoregressive model for the closing price of the S&P 500 using the Excel file *S&P 500*. (Hint: use *CB Predictor* to identify the strongest lags.)
10. Find the best autoregressive model for each of the variables in the Excel file *Retail Electricity Prices*. (Hint: use *CB Predictor* to identify the strongest lags.)
11. Develop a multiple regression model with categorical variables that incorporate seasonality for forecasting the temperature in Washington, D.C., using the data for years 1999 and 2000 in the Excel file *Washington DC Weather*. Use the model to generate forecasts for the next nine months and compare the forecasts to the actual observations in the data for the year 2001.
12. Develop a multiple regression model with categorical variables that incorporate seasonality for forecasting sales using the last three years of data in the Excel file *New Car Sales*.
13. Develop a multiple regression model with categorical variables that incorporate seasonality for forecasting housing starts beginning in June 2006 using the data in the Excel file *Housing Starts*.
14. Data in the Excel File *Microprocessor Data* shows the demand for one type of chip used in industrial equipment from a small manufacturer.
  - a. Construct a chart of the data. What appears to happen when a new chip is introduced?
  - b. Develop a causal regression model to forecast demand that includes both time and the introduction of a new chip as explanatory variables.
  - c. What would the forecast be for the next month if a new chip is introduced? What would it be if a new chip is not introduced?
15. Find the best moving average and exponential smoothing models for each of the stocks in the Excel file *Closing Stock Prices* using *CB Predictor*. Compare your results to your answers to Problem 1.
16. Find the best moving average and exponential smoothing models for each of the stocks in the Excel file *Baseball Attendance* using *CB Predictor*. Compare your results to your answers to Problem 2.
17. Find the best moving average and exponential smoothing models for the male and female populations in the Excel file *Ohio Prison Population* using *CB Predictor*. Compare your results to your answers to Problem 3.
18. Find the best moving average and exponential smoothing models for forecasting gasoline prices in the Excel file *Gasoline Prices* using *CB Predictor*. Compare your results to your answers to Problem 4.
19. Construct a line chart for the data in the Excel file *Arizona Population*.
  - a. Suggest the best-fitting functional form for forecasting these data.
  - b. Use *CB Predictor* to find the best forecasting model.

20. Construct a line chart for each of the variables in the data file *Death Cause Statistics*, and suggest the best forecasting technique. Then apply *CB Predictor* to find the best forecasting models for these variables.
21. The Excel file *Olympic Track and Field Data* provides the gold medal-winning distances for the high jump, discus, and long jump for the modern Olympic Games. Develop forecasting models for each of the events. What do your models predict for the next Olympics?
22. Use *CB Predictor* to find the best forecasting model for the data in the following economic time series:
- New Car Sales*
  - Housing Starts*
  - Coal Consumption*
  - DJIA December Close*
  - Federal Funds Rates*
  - Mortgage Rates*
  - Prime Rate*
  - Treasury Yield Rates*

## Case

### Energy Forecasting

The Excel file *Energy Production & Consumption* provides data on energy production, consumption, imports, and exports. You have been hired as an analyst for a government agency and have been asked to forecast these variables over the next five years. Apply forecasting tools

and appropriate visual aids, and write a formal report to the agency director that explains these data and the future forecasts. In your analysis, consider carefully how much data should be used to develop the forecast and justify your decision.

## APPENDIX 7.1

### Advanced Forecasting Models—Theory and Computation

In this appendix, we present computational formulas for advanced models for time-series forecasting. The calculations are somewhat complex, but can be implemented on spreadsheets with a bit of effort.

#### A. Double Moving Average

Double moving average involves taking averages of averages. Let  $M_t$  be the simple moving average for the last  $k$  periods (including period  $t$ ):

$$M_t = [A_{t-k+1} + A_{t-k+2} + \dots + A_t]/k \quad (7A.1)$$

The double moving average,  $D_t$ , for the last  $k$  periods (including period  $t$ ) is the average of the simple moving averages:

$$D_t = [M_{t-k+1} + M_{t-k+2} + \dots + M_t]/k \quad (7A.2)$$

Using these values, the double moving average method estimates the values of  $a_t$  and  $b_t$  in the linear trend model  $F_{t+k} = a_t + b_t k$  as:

$$\begin{aligned} a_t &= 2M_t - D_t \\ b_t &= (2/(k-1))[M_t - D_t] \end{aligned} \quad (7A.3)$$

These equations are derived essentially by minimizing the sum of squared errors using the last  $k$  periods of data. Once these parameters are determined, forecasts beyond the end of the observed data (time period  $T$ ) are calculated using the linear trend model with values of  $a_T$  and  $b_T$ . That is, for  $k$  periods beyond period  $T$ , the forecast is  $F_{T+k} = a_T + b_T k$ . For instance, the forecast for the next period would be  $F_{T+1} = a_T + b_T(1)$ .

#### B. Double Exponential Smoothing

Like double moving average, double exponential smoothing is also based on the linear trend equation,  $F_{t+k} = a_t + b_t k$ , but the estimates of  $a_t$  and  $b_t$  are obtained from the following equations:

$$\begin{aligned} a_t &= \alpha F_t + (1-\alpha)(a_{t-1} + b_{t-1}) \\ b_t &= \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \end{aligned} \quad (7A.4)$$

In essence, we are smoothing both parameters of the linear trend model. From the first equation, the estimate of the level in period  $t$  is a weighted average of the observed value at time  $t$  and the predicted value at time  $t$ ,  $a_{t-1} + b_{t-1}$  based on single exponential smoothing. For large values of  $\alpha$ , more

weight is placed on the observed value. Lower values of  $\alpha$  put more weight on the smoothed predicted value. Similarly, from the second equation, the estimate of the trend in period  $t$  is a weighted average of the differences in the estimated levels in periods  $t$  and  $t - 1$  and the estimate of the level in period  $t - 1$ . Larger values of  $\beta$  place more weight on the differences in the levels, while lower values of  $\beta$  put more emphasis on the previous estimate of the trend.

To initialize the double exponential smoothing process, we need values for  $a_1$  and  $b_1$ . One approach is to let  $a_1 = A_1$  and  $b_1 = A_2 - A_1$ ; that is, estimate the initial level with the first observation and the initial trend with the difference in the first two observations. As with single exponential smoothing, we are free to choose the values of  $\alpha$  and  $\beta$ . MAD, MSE, or MAPE may be used to find good values for these smoothing parameters. We leave it to you as an exercise to implement this model on a spreadsheet for the total energy consumption data.

## C. Additive Seasonality

The seasonal additive model is:

$$F_{t+k} = a_t + S_{t-s+k} \quad (7A.5)$$

The level and seasonal factors are estimated in the additive model using the following equations:

$$\begin{aligned} a_t &= \alpha(A_t - S_{t-s}) + (1 - \alpha)a_{t-1} \\ S_t &= \gamma(A_t - a_t) + (1 - \gamma)S_{t-s} \end{aligned} \quad (7A.6)$$

where  $\alpha$  and  $\gamma$  are smoothing constants. The first equation estimates the level for period  $t$  as a weighted average of the deseasonalized data for period  $t$ , ( $A_t - S_{t-s}$ ), and the previous period's level. The seasonal factors are updated as well using the second equation. The seasonal factor is a weighted average of the estimated seasonal component for period  $t$ , ( $A_t - a_t$ ), and the seasonal factor for the last period of that season type. Then the forecast for the next period is  $F_{t+1} = a_t + S_{t-s+1}$ . For  $k$  periods out from the final observed period  $T$ , the forecast is:

$$F_{T+k} = a_T + S_{T-s+k} \quad (7A.7)$$

To initialize the model, we need to estimate the level and seasonal factors for the first  $s$  periods (e.g., for an annual season with quarterly data this would be the first four periods; for monthly data, it would be the first 12 periods, etc.). We will use the following approach:

$$\begin{aligned} a_s &= \sum_{t=1}^s A_t / s \\ a_t &= a_s \quad \text{for } t = 1, 2, \dots, s \end{aligned} \quad (7A.8)$$

and

$$S_t = A_t - a_t \quad \text{for } t = 1, 2, \dots, s \quad (7A.9)$$

That is, we initialize the level for the first  $s$  periods to the average of the observed values over these periods and the seasonal factors to the difference between the observed data and the estimated levels. Once these have been initialized, the smoothing equations can be implemented for updating.

## D. Multiplicative Seasonality

The seasonal multiplicative model is:

$$F_{t+k} = a_t S_{t-s+k} \quad (7A.10)$$

This model has the same basic smoothing structure as the additive seasonal model but is more appropriate for seasonal time series that increase in amplitude over time. The smoothing equations are:

$$\begin{aligned} a_t &= \alpha(A_t / S_{t-s}) + (1 - \alpha)a_{t-1} \\ S_t &= \gamma(A_t / a_t) + (1 - \gamma)S_{t-s} \end{aligned} \quad (7A.11)$$

where  $\alpha$  and  $\gamma$  are again the smoothing constants. Here,  $A_t / S_{t-s}$  is the deseasonalized estimate for period  $t$ . Large values of  $\alpha$  put more emphasis on this term in estimating the level for period  $t$ . The term  $A_t / a_t$  is an estimate of the seasonal factor for period  $t$ . Large values of  $\gamma$  put more emphasis on this in the estimate of the seasonal factor.

The forecast for the period  $t + 1$  is  $F_{t+1} = a_t S_{t-s+1}$ . For  $k$  periods out from the final observed period  $T$ , the forecast is:

$$F_{t+k} = a_T S_{T-s+k} \quad (7A.12)$$

As in the additive model, we need initial values for the level and seasonal factors. We do this as follows:

$$\begin{aligned} a_s &= \sum_{t=1}^s A_t / s \\ a_t &= a_s \quad \text{for } t = 1, 2, \dots, s \end{aligned} \quad (7A.13)$$

and

$$S_t = A_t / a_t \quad \text{for } t = 1, 2, \dots, s \quad (7A.14)$$

Once these have been initialized, the smoothing equations can be implemented for updating.

## E. Holt–Winters Additive Model

The Holt–Winters additive model is based on the equation:

$$F_{t+1} = a_t + b_t + S_{t-s+1} \quad (7A.15)$$

This model is similar to the additive model incorporating seasonality that we described in the previous section, but it also includes a trend component. The smoothing equations are:

$$\begin{aligned} a_t &= \alpha(A_t - S_{t-s}) + (1 - \alpha)(a_{t-1} + b_{t-1}) \\ b_t &= \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1} \\ S_t &= \gamma(A_t - a_t) + (1 - \gamma)S_{t-s} \end{aligned} \quad (7A.16)$$

Here,  $\alpha$ ,  $\beta$ , and  $\gamma$  are the smoothing parameters for level, trend, and seasonal components, respectively. The forecast for period  $t + 1$  is:

$$F_{t+1} = a_t + b_t + S_{t-s+1} \quad (7A.17)$$

The forecast for  $k$  periods beyond the last period of observed data (period  $T$ ) is:

$$F_{T+k} = a_T + b_T k + S_{T-s+k} \quad (7A.18)$$

The initial values of level and trend are estimated in the same fashion as in the additive model for seasonality. The initial values for the trend are  $b_t = b_s$ , for  $t = 1, 2, \dots, s$ , where:

$$b_s = [(A_{s+1} - A_1)/S + (A_{s+2} - A_2)/S + \dots + (A_{s+s} - A_s)/S]/S \quad (7A.19)$$

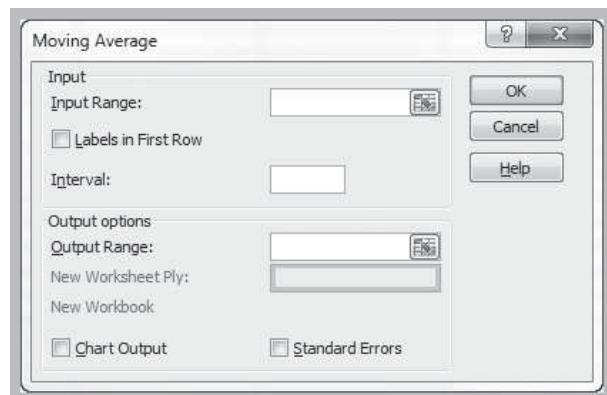
Note that each term inside the brackets is an estimate of the trend over one season. We average these over the first  $2s$  periods.

## APPENDIX 7.2

### Excel and CB Predictor Notes

#### A. Forecasting with Moving Averages

From the *Analysis* group, select *Data Analysis* then *Moving Average*. Excel displays the dialog box shown in Figure 7A.1. You need to enter the *Input Range* of the data, the *Interval* (the value of  $k$ ), and the first cell of the *Output Range*. To align the actual data with the forecasted values in the worksheet, select the first cell of the *Output Range* to be one row below the first value. You may also obtain a chart of the data and the moving averages, as well as a column of standard errors, by checking the appropriate boxes.



**FIGURE 7A.1** Excel Moving Average Tool Dialog

#### F. Holt–Winters Multiplicative Model

The Holt–Winters multiplicative model is:

$$F_{t+1} = (a_t + b_t)S_{t-s+1} \quad (7A.20)$$

This model parallels the additive model:

$$\begin{aligned} a_t &= \alpha(A_t/S_{t-s}) + (1 - \alpha)(a_{t-1} + b_{t-1}) \\ b_t &= \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1} \\ S_t &= \gamma(A_t - a_t) + (1 - \gamma)S_{t-s} \end{aligned} \quad (7A.21)$$

The forecast for period  $t + 1$  is:

$$F_{t+1} = (a_t + b_t)S_{t-s+1} \quad (7A.22)$$

The forecast for  $k$  periods beyond the last period of observed data (period  $T$ ) is:

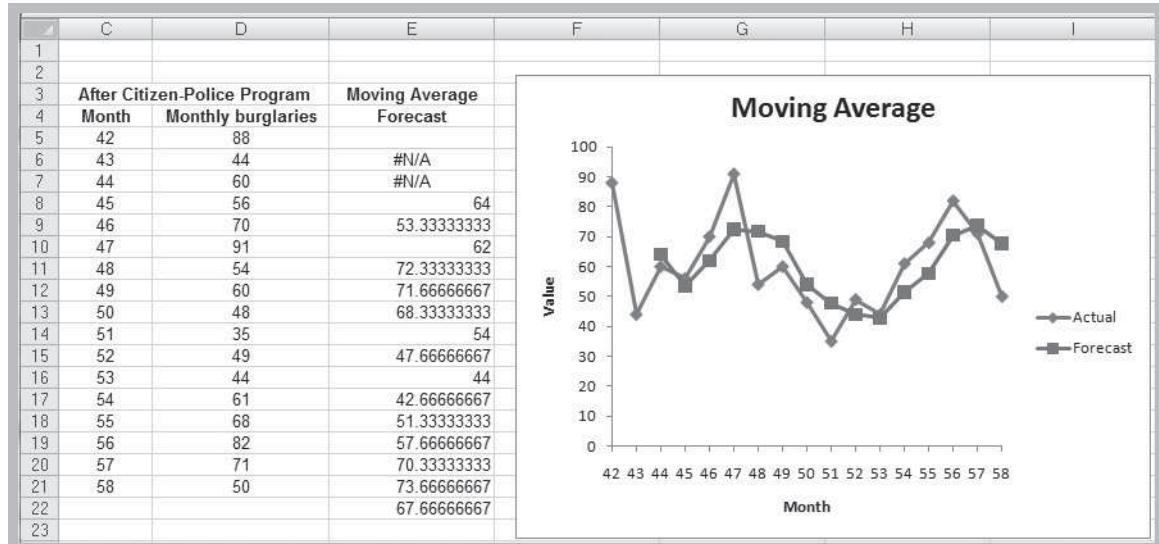
$$F_{T+k} = (a_T + b_T k)S_{T-s+k} \quad (7A.23)$$

The initial values of level and trend are estimated in the same fashion as in the multiplicative model for seasonality, and the trend component as in the Holt–Winters additive model.

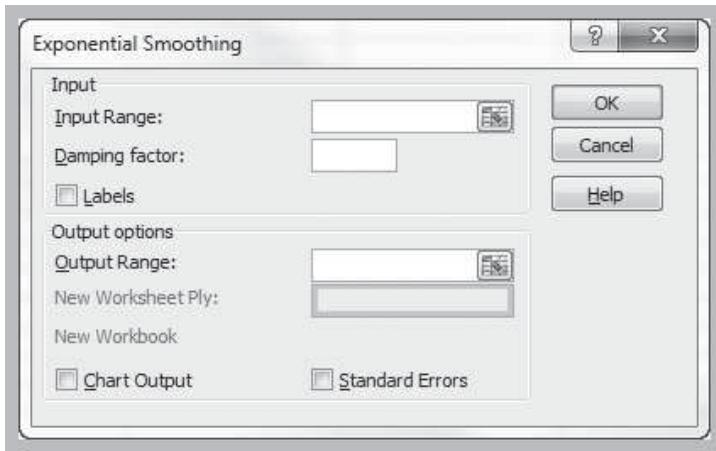
However, we do not recommend using the chart or error options because the forecasts generated by this tool are not properly aligned with the data (the forecast value aligned with a particular data point represents the forecast for the *next* month) and, thus, can be misleading. Rather, we recommend that you generate your own chart as we did in Figure 7.5. Figure 7A.2 shows the results produced by the *Moving Average* tool (with some customization of the forecast chart to show the months on the *x*-axis). Note that the forecast for month 59 is aligned with the actual value for month 58 on the chart. Compare this to Figure 7.5 and you can see the difference.

#### B. Forecasting with Exponential Smoothing

From the *Analysis* group, select *Data Analysis* then *Exponential Smoothing*. In the dialog (Figure 7A.3), as in the *Moving Average* dialog, you must enter the *Input Range* of the time-series data, the *Damping Factor* ( $1 - \alpha$ )—not the smoothing constant as we have defined it (!)—and the first cell of the *Output Range*, which should be adjacent to the first data point. You also have options for labels, to chart output, and to obtain standard errors. As opposed to the *Moving Average* tool, the chart generated by this tool does correctly align the forecasts with the actual data, as shown in Figure 7A.4. You can see that the exponential smoothing model follows the



**FIGURE 7A.2** Results of Excel Moving Average Tool (note misalignment of forecasts with actual observations in the chart)



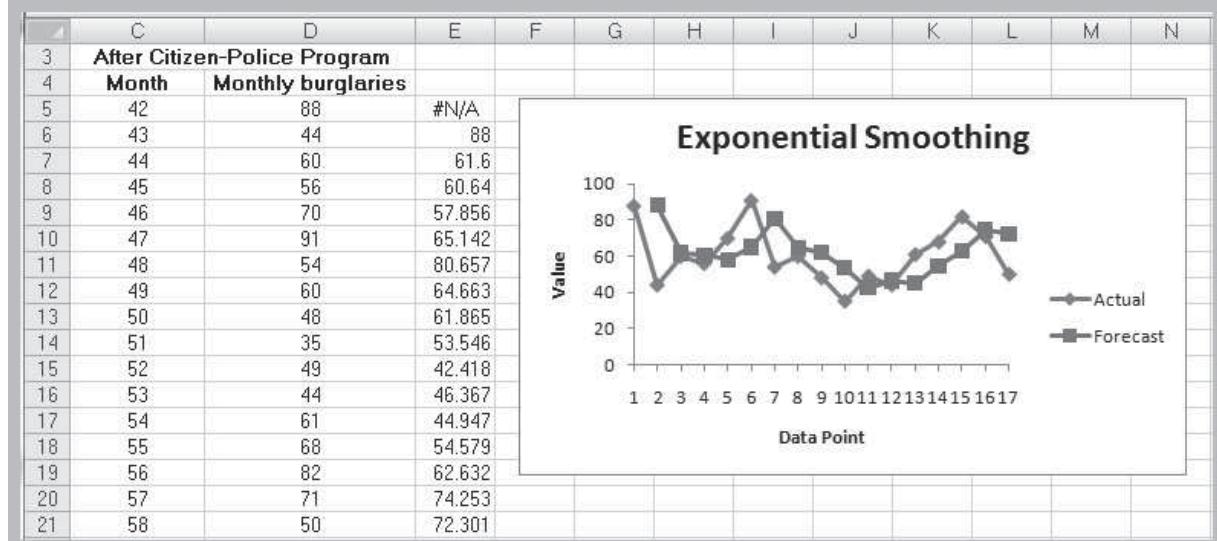
**FIGURE 7A.3** Exponential Smoothing Tool Dialog

pattern of the data quite closely, although it tends to lag with an increasing trend in the data.

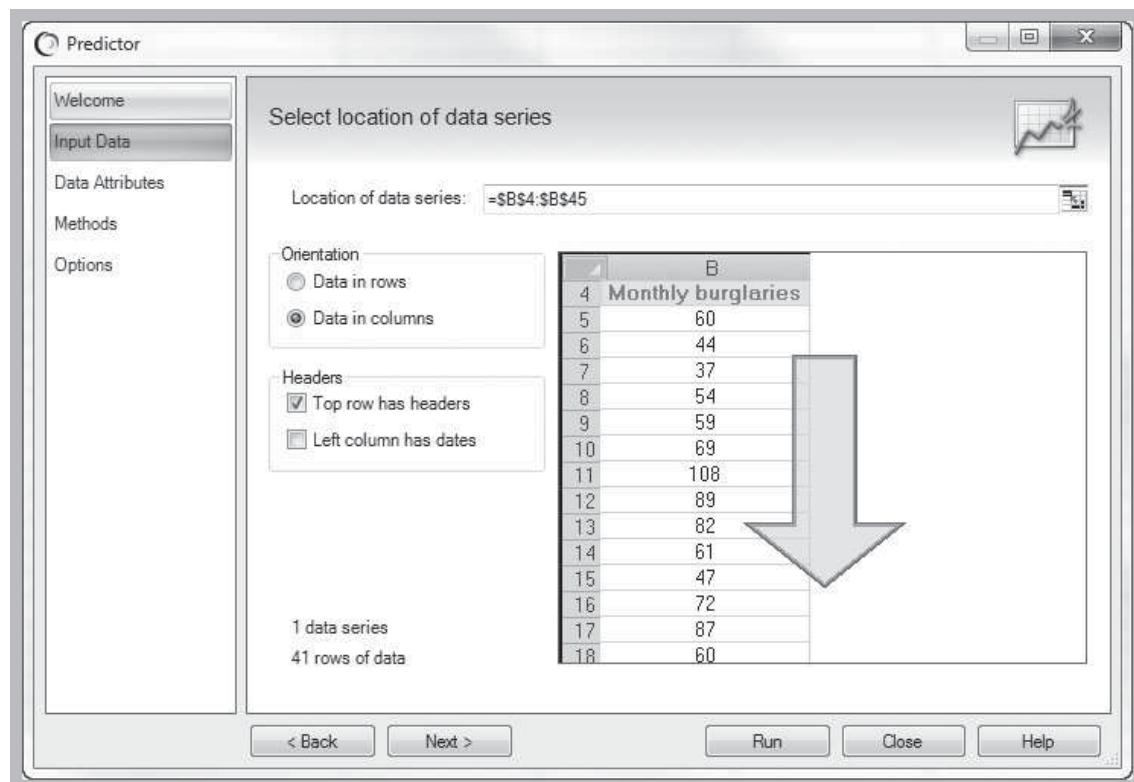
### C. Using CB Predictor

After *Crystal Ball* has been installed, *CB Predictor* may be accessed in Excel from the *Crystal Ball* tab. In the *Tools* group, select *Predictor*. *CB Predictor* guides you through four dialogs, the first of which is shown in Figure 7A.5. These can be selected by clicking the *Next* button or by clicking on the tabs. The *Input Data* dialog allows you to specify the data

range on which to base your forecast; the *Data Attributes* dialog allows you to specify the time-based nature of data (periods, months, years, etc.) and whether or not seasonality is present (see Figure 7A.6). The *AutoDetect* feature will identify seasonality if it is present. The *Method* dialog allows you to select one or more of eight nonseasonal or seasonal time-series methods—single moving average, double moving average, single exponential smoothing, double exponential smoothing, seasonal additive, seasonal multiplicative, Holt-Winters additive, or Holt-Winters multiplicative (see Figure 7A.7), or also select ARIMA or multiple regression



**FIGURE 7A.4** Exponential Smoothing Forecasts for  $\alpha = 0.6$



**FIGURE 7A.5** CB Predictor Input Data Dialog

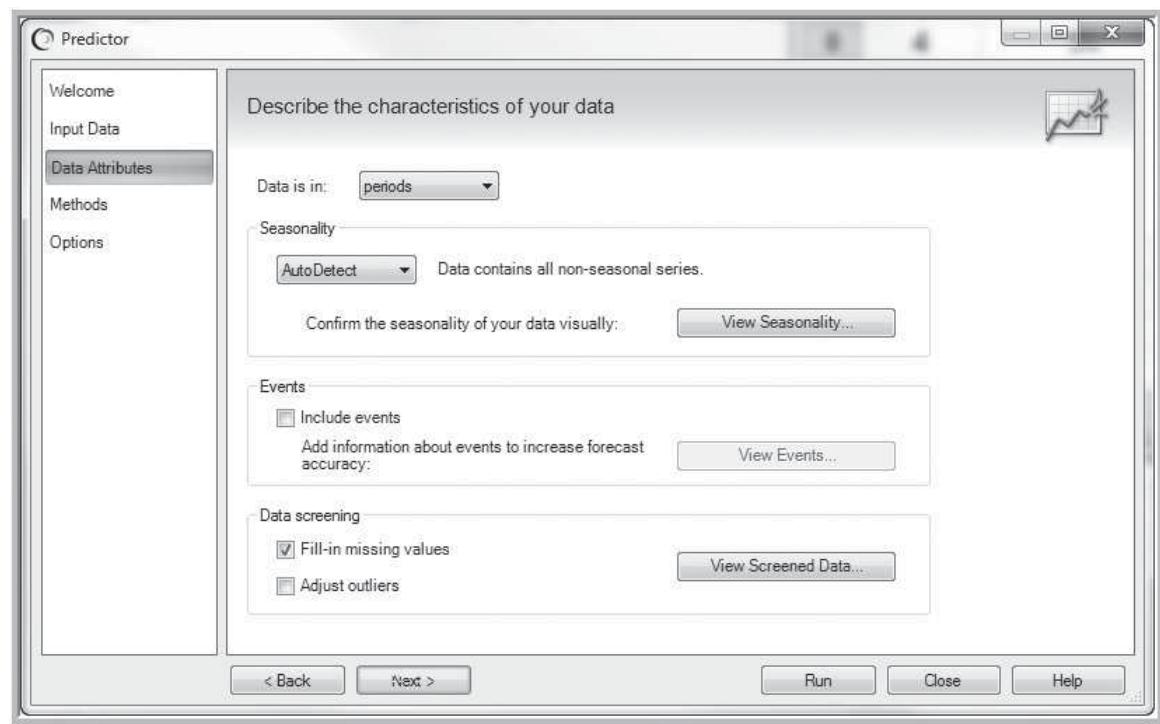


FIGURE 7A.6 CB Predictor Data Attributes Dialog

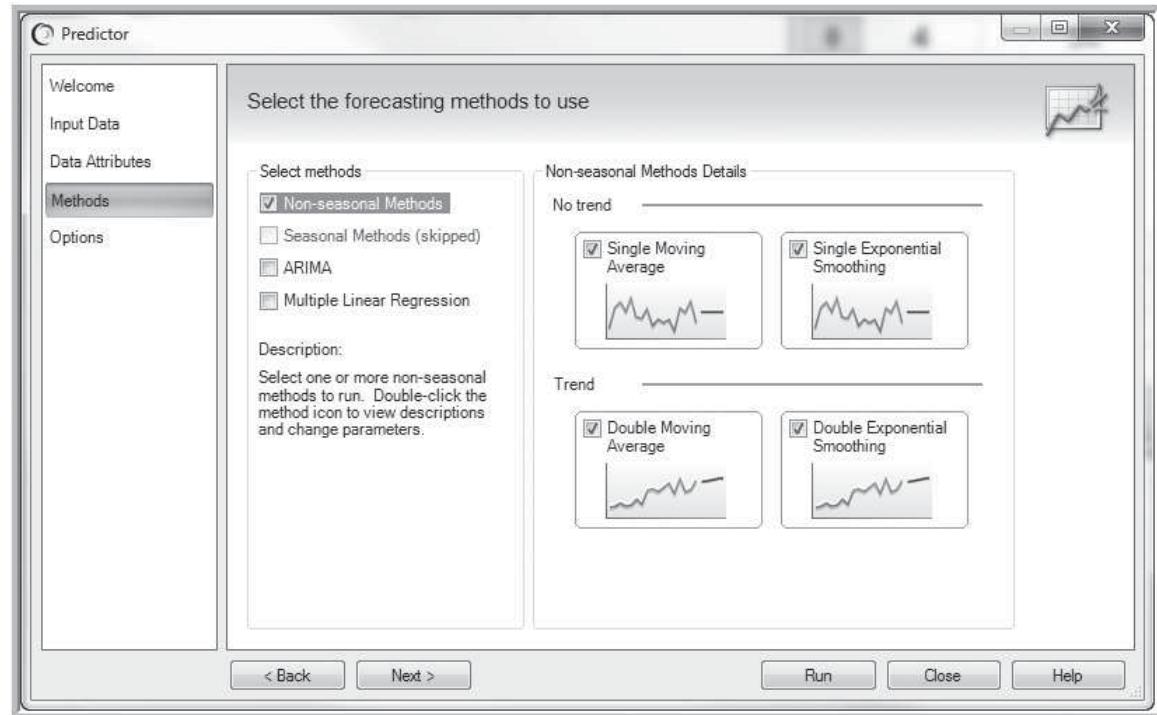
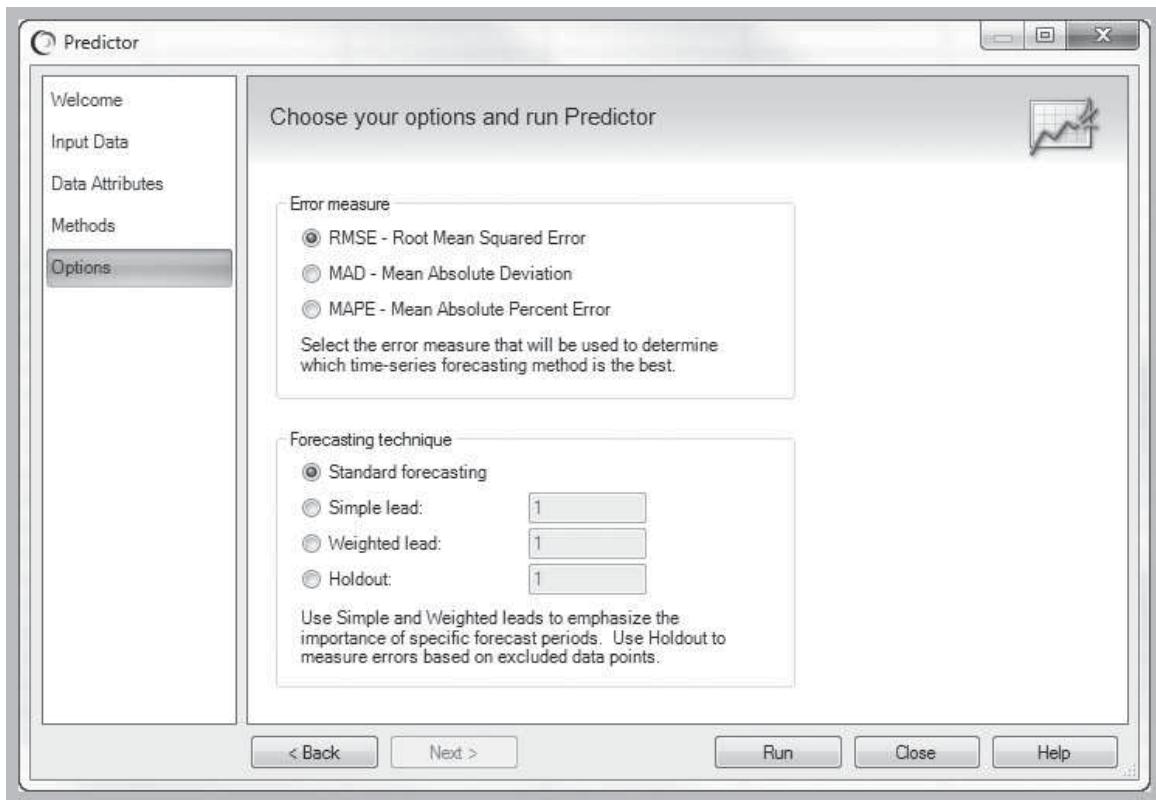


FIGURE 7A.7 CB Predictor Methods Dialog



**FIGURE 7A.8** CB Predictor Options Dialog

methods. Unless seasonality is detected or specified, only the nonseasonal methods will be displayed. The charts shown in the *Method Gallery* suggest the method that is best suited for the data similar to Table 7.3. However, *CB Predictor* can run each method you select and will recommend the one

that best forecasts your data. Not only does it select the best type of model, it also optimizes the forecasting parameters to minimize forecasting errors. The *Options* dialog allows you to select the error metric on which the models are ranked (see Figure 7A.8).

## *Chapter 8*

# Introduction to Statistical Quality Control

- INTRODUCTION 272
- THE ROLE OF STATISTICS AND DATA ANALYSIS IN QUALITY CONTROL 273
- STATISTICAL PROCESS CONTROL 274
  - Control Charts 274
  - $\bar{X}$ - and  $R$ -Charts 275
- ANALYZING CONTROL CHARTS 280
  - Sudden Shift in the Process Average 281
  - Cycles 281
  - Trends 281
  - Hugging the Center Line 281
  - Hugging the Control Limits 282
- CONTROL CHARTS FOR ATTRIBUTES 284
  - Variable Sample Size 286
- PROCESS CAPABILITY ANALYSIS 288
- BASIC CONCEPTS REVIEW QUESTIONS 290
- PROBLEMS AND APPLICATIONS 290
- CASE: QUALITY CONTROL ANALYSIS 291
- APPENDIX 8.1 PHSTAT NOTES 291
  - A. Creating  $\bar{X}$ - and  $R$ -Charts 291
  - B. Creating  $p$ -Charts 292

## **INTRODUCTION**

An important application of statistics and data analysis in both manufacturing and service operations is in the area of *quality control*. Quality control methods help employees monitor production operations to ensure that output conforms to specifications. This is important in manufactured goods since product performance depends on achieving design tolerances. It is also vital to service operations to ensure that customers receive error-free, consistent service.

Why is quality control necessary? The principal reason is that no two outputs from any production process are exactly alike. If you measure any quality characteristic—such as the diameters of machined parts, the amount of soft drink in a bottle, or the number of errors in processing orders at a distribution center—you will discover some variation. Variation is the

result of many small differences in those factors that comprise a process: people, machines, materials, methods, and measurement systems. Taken together, they are called **common causes of variation**.

Other causes of variation occur sporadically and can be identified and either eliminated or at least explained. For example, when a tool wears down, it can be replaced; when a machine falls out of adjustment, it can be reset; when a bad lot of material is discovered, it can be returned to the supplier. Such examples are called **special causes of variation**. Special causes of variation cause the distribution of process output to change over time. Using statistical tools, we can identify when they occur and take appropriate action, thus preventing unnecessary quality problems. Equally important is knowing when to leave the process alone and not react to common causes over which we have no control.

In this chapter, we introduce basic ideas of *statistical process control* and *process capability analysis*—two important tools in helping to achieve quality. The applications of statistics to quality control are far more extensive than we can present; much additional information may be found in many other sources.

## THE ROLE OF STATISTICS AND DATA ANALYSIS IN QUALITY CONTROL

We can learn a lot about the common causes of variation in a process and their effect on quality by studying process output. For example, suppose that a company such as General Electric Aircraft Engines produces a critical machined part. Key questions might be: What is the average dimension? How much variability occurs in the output of the process? What does the distribution of part dimensions look like? What proportion of output, if any, does not conform to design specifications? These are fundamental questions that can be addressed with statistics.

The role of statistics is to provide tools to analyze data collected from a process and enable employees to make informed decisions when the process needs short-term corrective action or long-term improvements. Statistical methods have been used for quality control since the 1920s, when they were pioneered at the Western Electric Company. They became a mainstay of Japanese manufacturing in the early 1950s; however, they did not become widely used in the United States until the quality management movement of the 1980s, led by pioneers such as W. Edwards Deming and Joseph M. Juran, both of whom were instrumental in the adoption of these methods in Japan. Since then, statistical quality control has been shown to be a proven means of improving customer satisfaction and reducing costs in many industries.

To illustrate the applications of statistics to quality control, we will use an example from a Midwest pharmaceutical company that manufactures individual syringes with a self-contained, single dose of an injectable drug.<sup>1</sup> In the manufacturing process, sterile liquid drug is poured into glass syringes and sealed with a rubber stopper. The remaining stage involves insertion of the cartridge into plastic syringes and the electrical “tacking” of the containment cap at a precisely determined length of the syringe. A cap that is tacked at a shorter-than-desired length (less than 4.920 inches) leads to pressure on the cartridge stopper and, hence, partial or complete activation of the syringe. Such syringes must then be scrapped. If the cap is tacked at a longer-than-desired length (4.980 inches or longer), the tacking is incomplete or inadequate, which can lead to cap loss and potentially a cartridge loss in shipment and handling. Such syringes can be reworked manually to attach the cap at a lower position. However, this process requires a 100% inspection of the tacked syringes and results in increased cost for the items. This final production step seemed to be producing more and more scrap and reworked syringes over successive weeks. At this point, statistical consultants became involved in an attempt to solve this problem and recommended statistical process control for the purpose of improving the tacking operation.

<sup>1</sup> Adapted from LeRoy A. Franklin and Samar N. Mukherjee, “An SPC Case Study on Stabilizing Syringe Lengths,” *Quality Engineering* 12, no. 1 (1999–2000): 65–71.

## STATISTICAL PROCESS CONTROL

In quality control, measures that come from counting are called *attributes*. Examples of attributes are the number of defective pieces in a shipment of components, the number of errors on an invoice, and the percent of customers rating service a 6 or 7 on a seven-point satisfaction scale. Measures that are based on a continuous measurement scale are called *variables*. Examples of variables data are the inside diameter of a drilled hole, the weight of a carton, and the time between order and delivery. This distinction is important because different statistical process control tools must be used for each type of data.

When a process operates under ideal conditions, variation in the distribution of output is due to common causes. When only common causes are present, the process is said to be **in control**. A controlled process is stable and predictable to the extent that we can predict the likelihood that future output will fall within some range once we know the probability distribution of outcomes. Special causes, however, cause the distribution to change. The change may be a shift in the mean, an increase or decrease in the variance, or a change in shape. Clearly, if we cannot predict how the distribution may change, then we cannot compute the probability that future output will fall within some range. When special causes are present, the process is said to be **out of control** and needs to be corrected to bring it back to a stable state. **Statistical process control (SPC)** provides a means of identifying special causes as well as telling us when the process is in control and should be left alone. Control charts were first used by Dr. Walter Shewhart at Bell Laboratories in the 1920s (they are sometimes called *Shewhart charts*). Shewhart was the first to make a distinction between common causes of variation and special causes of variation.

SPC consists of the following:

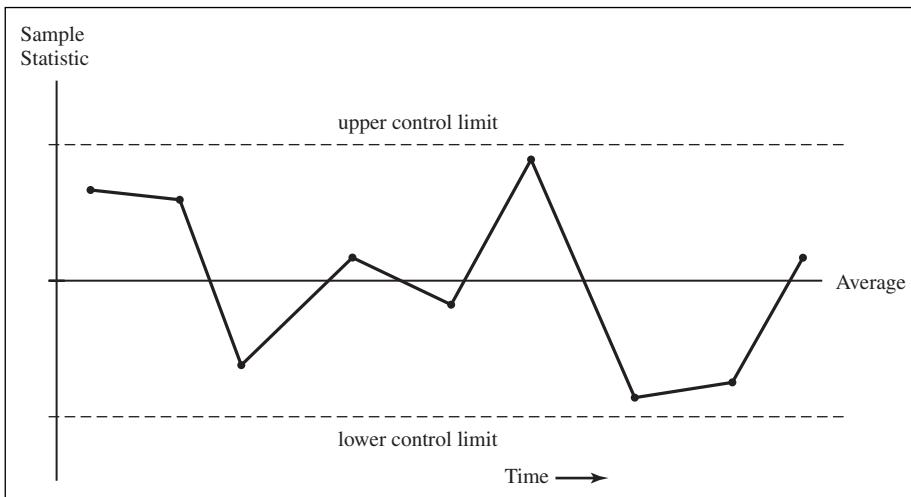
1. Selecting a sample of observations from a production or service process
2. Measuring one or more quality characteristics
3. Recording the data
4. Making a few calculations
5. Plotting key statistics on a *control chart*
6. Examining the chart to determine if any unusual patterns, called **out-of-control conditions**, can be identified
7. Determining the cause of out-of-control conditions and taking corrective action

When data are collected, it is important to clearly record the data, the time the data were collected, the measuring instruments that were used, who collected the data, and any other important information such as lot numbers, machine numbers, and the like. By having a record of such information, we can trace the source of quality problems more easily.

### Control Charts

A **run chart** is a line chart in which the independent variable is time and the dependent variable is the value of some sample statistic, such as the mean, range, or proportion. A **control chart** is a run chart that has two additional horizontal lines, called **control limits** (the upper control limit, denoted UCL, and the lower control limit, denoted LCL), as illustrated in Figure 8.1. Control limits are chosen statistically so that there is a high probability (usually greater than 0.99) that sample statistics will fall randomly within the limits *if the process is in control*.

To understand the statistical basis for control charts, let us assume that we are dealing with a variables measurement that is normally distributed with a mean  $\mu$  and standard deviation  $\sigma$ . If the process is stable, or in control, then each individual measurement will stem from this distribution. In high-volume production processes, it is generally difficult, if not impossible, to measure each individual output, so we take samples



**FIGURE 8.1** Structure of a Control Chart

at periodic intervals. For samples of a fixed size,  $n$ , we know from Chapter 4 that the sampling distribution will be normal with mean  $\mu$  and standard deviation (standard error)  $\sigma_{\bar{x}} = \sigma / \sqrt{n}$ . We would expect that about 99.7% of sample means will lie within  $\pm 3$  standard errors of the mean, or between  $\mu - 3\sigma_{\bar{x}}$  and  $\mu + 3\sigma_{\bar{x}}$ , provided the process remains in control. These values become the theoretical control limits for a control chart to monitor the centering of a process using the sample mean. Of course, we do not know the true population parameters, so we estimate them by the sample mean  $\bar{x}$  and sample standard deviation,  $s$ . Thus, the actual control limits would be:

$$\text{Lower control limit: } \bar{x} - 3s_{\bar{x}}$$

$$\text{Upper control limit: } \bar{x} + 3s_{\bar{x}}$$

In general, control limits are established as  $\pm 3$  standard errors ( $s_{\bar{x}}$ ) from the mean of the sampling distribution of the statistic we plot on the chart. Statisticians have devised various formulas for computing these limits in a practical manner that is easy for shop floor workers to understand and use. However, the theory is based on understanding the sampling distribution of the statistic we measure.

There are many different types of control charts. We will introduce three basic types of control charts in this chapter:  $\bar{x}$ - and  $R$ -charts for variables data and  $p$ -charts for attributes data. Discussions of other types of charts may be found in the references.

### **$\bar{x}$ - and $R$ -Charts**

The  $\bar{x}$ -chart monitors the centering of process output for variables data over time by plotting the mean of each sample. In manufacturing, for example, the permissible variation in a dimension is usually stated by a **nominal specification** (target value) and some **tolerance**. For example, the specifications on the syringe length are  $4.950 \pm 0.030$  inches. The nominal is 4.950 and the tolerance is  $\pm 0.030$ . Therefore, the lengths should be between 4.920 and 4.980 inches. The low value of the permissible dimension is called the **lower specification limit, LSL**; and the high value of the permissible dimension is called the **upper specification limit, USL**. Thus,  $LSL = 4.920$  and  $USL = 4.980$ . The  $\bar{x}$ -chart is used to monitor the centering of a process. The  **$R$ -chart**, or range chart, monitors the variability in the data as measured by the range of each sample. Thus, the

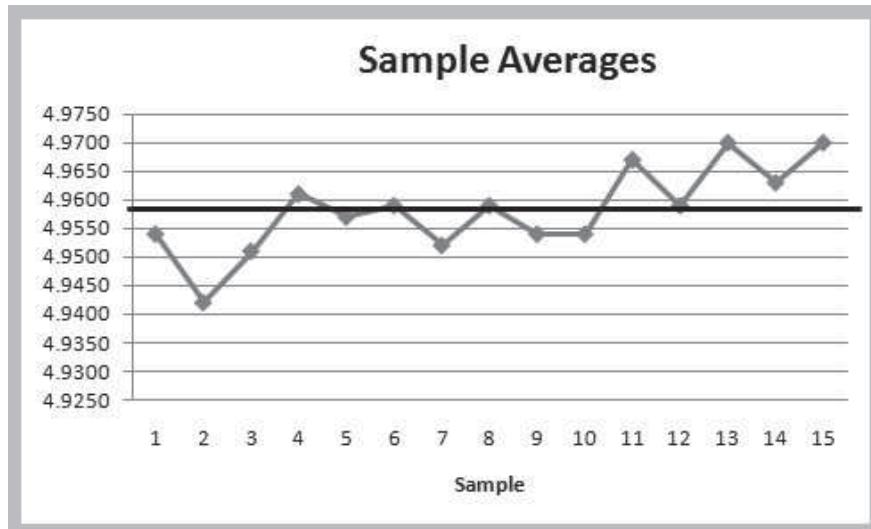
*R*-chart monitors the uniformity or consistency of the process. The smaller the value of *R*, the more uniform is the process. Any increase in the average range is undesirable; this would mean that the variation is getting larger. However, decreases in variability signify improvement. We could use the standard deviation of each sample instead of the range to provide a more accurate characterization of variability; however, for small samples (around eight or less), little differences will be apparent, and if the calculations are done manually by a worker on the shop floor, *R*-charts are much easier to apply.

The basic procedure for constructing and using any control chart is first to gather at least 25 to 30 samples of data with a fixed sample size *n* from a production process, measure the quality characteristic of interest, and record the data. We will illustrate the construction of a control chart using the data in the Excel file *Syringe Samples* (see Figure 8.2), which shows 47 samples that were taken every 15 minutes from the syringe manufacturing process over three shifts. Each sample consists of five individual observations. In column G, we calculate the mean of each sample, and in column H, the range.

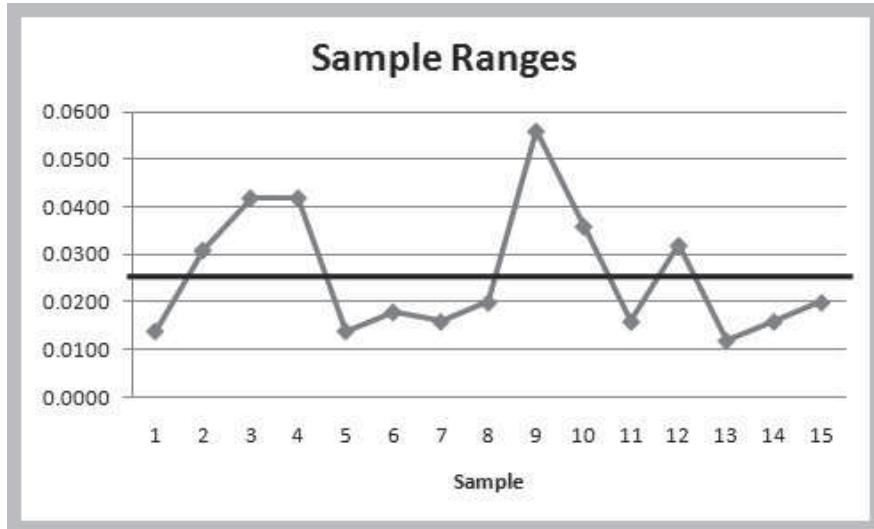
We will work with the first 15 samples (typically, it is recommended that at least 25 to 30 samples be used to construct a control chart, but we will assume that only the first 15 samples are available). After we have calculated the mean and range for each sample, we compute the average mean,  $\bar{x} = 4.9581$ , and the average range,  $\bar{R} = 0.0257$ . Figures 8.3 and 8.4 show plots of the sample means and ranges. Although the chart for sample ranges shows some variation, we cannot yet determine statistically whether this

	A	B	C	D	E	F	G	H
1	<b>Syringe Samples</b>							
2								
First Shift Data								
4	Sample	Sample Observations					Average	Range
5	1	4.9600	4.9460	4.9500	4.9560	4.9580	4.9540	0.0140
6	2	4.9580	4.9270	4.9350	4.9400	4.9500	4.9420	0.0310
7	3	4.9710	4.9290	4.9650	4.9520	4.9380	4.9510	0.0420
8	4	4.9400	4.9820	4.9700	4.9530	4.9600	4.9610	0.0420
9	5	4.9640	4.9500	4.9530	4.9620	4.9560	4.9570	0.0140

**FIGURE 8.2** Portion of Excel File *Syringe Samples*



**FIGURE 8.3** Chart of Sample Means for Syringe Data



**FIGURE 8.4** Chart of Sample Ranges for Syringe Data

variation might be due to some assignable cause or is simply due to chance. The chart for sample means appears to show an increasing trend.

The final step to complete the control charts is to compute control limits. As we explained earlier, control limits are boundaries within which the process is operating in statistical control. Control limits are based on past performance and tell us what values we can expect for  $\bar{x}$  or  $R$  as long as the process remains stable. If a point falls outside the control limits or if some unusual pattern occurs, then we should be suspicious of a special cause. The upper control limit for the  $R$ -chart is given by the formula:

$$UCL_R = D_4 \bar{R} \quad (8.1)$$

and the lower control limit for the  $R$ -chart is given by the formula:

$$LCL_R = D_3 \bar{R} \quad (8.2)$$

$D_3$  and  $D_4$  are constants that depend on the sample size and are found in Table 8.1. The theory is a bit complicated, but suffice it to say that these constants have been determined from the sampling distribution of  $R$  so that, for example,  $D_4 \bar{R} = \bar{R} + 3s_R$ , as we had described earlier.

Because the sample size is 5,  $D_4 = 2.114$ . Therefore, the upper control limit for the example is  $2.114(0.0257) = 0.0543$ . In this example,  $D_3$  for a sample size of 5 is 0; therefore, the lower control limit is 0. We then draw and label these control limits on the chart.

For the  $\bar{x}$ -chart, the control limits are given by the formulas:

$$UCL_{\bar{x}} = \bar{x} + A_2 \bar{R} \quad (8.3)$$

$$LCL_{\bar{x}} = \bar{x} - A_2 \bar{R} \quad (8.4)$$

Again, the constant  $A_2$  is determined so that  $A_2 \bar{R}$  is equivalent to three standard errors in the sampling distribution of the mean. For a sample of size 5,  $A_2 = 0.577$  from Figure 8.1. Therefore, the control limits are:

$$UCL_{\bar{x}} = 4.9581 + (0.577)(0.0257) = 4.973$$

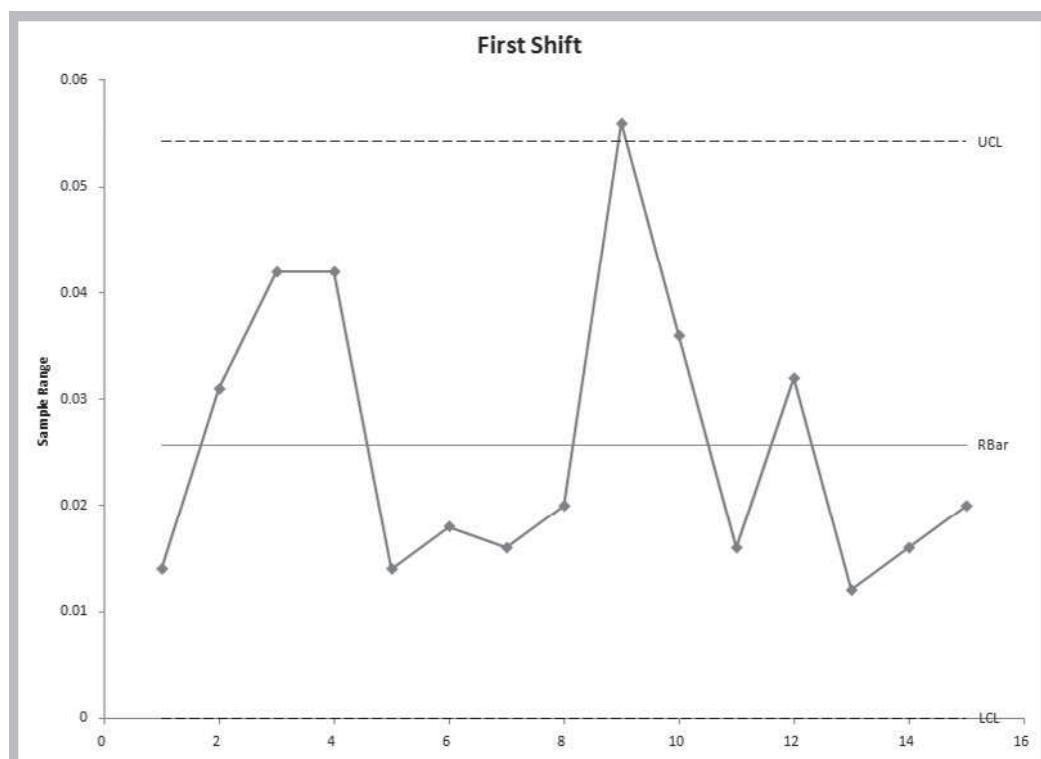
$$LCL_{\bar{x}} = 4.9581 - (0.577)(0.0257) = 4.943$$

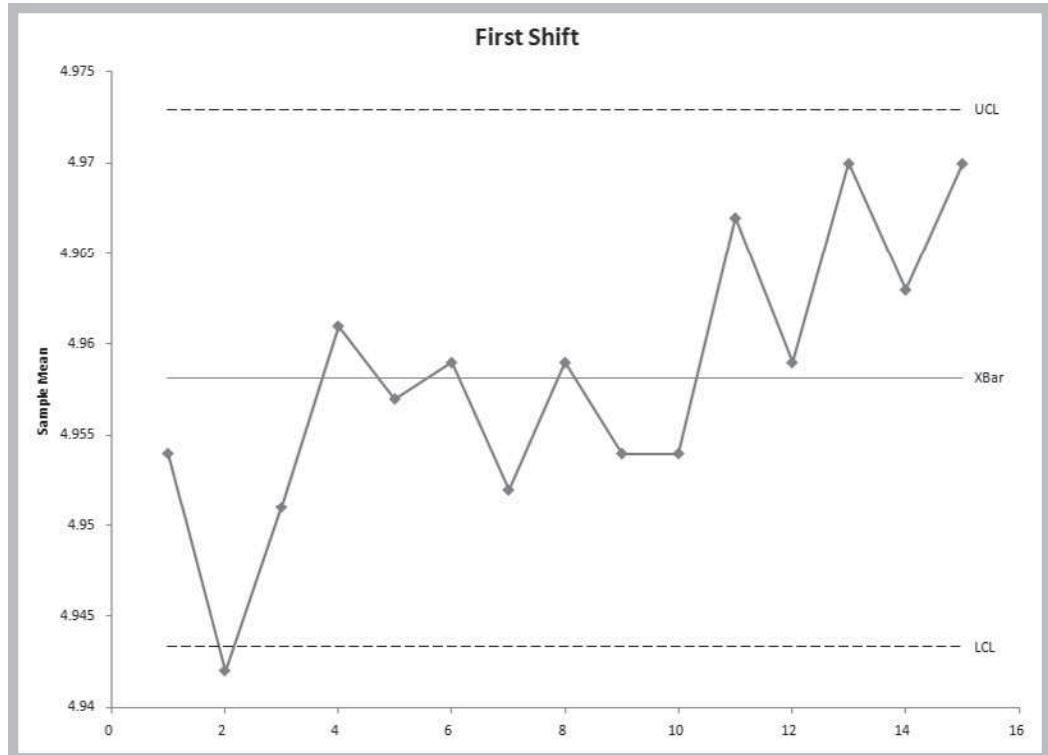
**TABLE 8.1** Control Chart Factors

Sample Size	$A_2$	$D_3$	$D_4$
2	1.880	0	3.267
3	1.023	0	2.574
4	0.729	0	2.282
5	0.577	0	2.114
6	0.483	0	2.004
7	0.419	0.076	1.924
8	0.373	0.136	1.864
9	0.337	0.184	1.816
10	0.308	0.223	1.777
11	0.285	0.256	1.744
12	0.266	0.283	1.717
13	0.249	0.307	1.693
14	0.235	0.328	1.672
15	0.223	0.347	1.653

We could draw these control limits on the charts to complete the process. *PHStat* includes a routine for constructing  $\bar{x}$ - and  $R$ -charts (see Appendix 8.1A, *Creating  $\bar{x}$ - and R-Charts*). The charts generated by this routine are shown in Figures 8.5 and 8.6. *PHStat* also creates a worksheet with the calculations using the formulas we have discussed (see Figure 8.7). The next step is to analyze the charts to determine the state of statistical control.

## Spreadsheet Note

**FIGURE 8.5** R-Chart for Syringe Data—First Shift

**First Shift****FIGURE 8.6**  $\bar{x}$ -Chart for Syringe Data—First Shift

A	B	
1	First Shift	
2		
3	Data	
4	Sample/Subgroup Size	5
5		
6	R Chart Intermediate Calculations	
7	RBar	0.025666667
8	D3 Factor	0
9	D4 Factor	2.114
10		
11	R Chart Control Limits	
12	Lower Control Limit	0
13	Center	0.025666667
14	Upper Control Limit	0.054259333
15		
16	XBar Chart Intermediate Calculations	
17	Average of Subgroup Averages	4.958133333
18	A2 Factor	0.577
19	A2 Factor * RBar	0.014809667
20		
21	XBar Chart Control Limits	
22	Lower Control Limit	4.943323667
23	Center	4.958133333
24	Upper Control Limit	4.972943

**FIGURE 8.7** PHStat Calculations Worksheet  
for R & XBar Charts

We wish to point out that using the range to monitor variability is not as accurate as using the standard deviation.  $R$ -charts were developed back in the 1930s to make it easy to implement SPC by hand because calculating a standard deviation is much more difficult. With computers today, it is easy to develop an  $s$ -chart, which plots the standard deviation of each sample instead of the range; however, *PHStat* does not provide a tool for doing this.

### SKILL-BUILDER EXERCISE 8.1

---

Build an Excel spreadsheet template for calculating control chart parameters (center line and control limits) for  $\bar{x}$ - and  $R$ -charts that allows you to enter the raw data without having to compute sample averages and ranges first. Try to design the template to automatically generate the control charts.

## ANALYZING CONTROL CHARTS

When a process is in statistical control, the points on a control chart should fluctuate at random between the control limits, and no recognizable patterns should exist. The following checklist provides a set of general rules for examining a control chart to see if a process is in control. These rules are based on the assumption that the underlying distribution of process output—and, therefore, the sampling distribution—is normal.

1. *No points are outside the control limits.* Since the control limits are set at  $\pm 3$  standard errors from the mean, the probability of a point falling outside the control limits when the process is in control is only 0.0027, under the normality assumption.
2. *The number of points above and below the center line is about the same.* If the distribution is symmetric, as a normal distribution is, we would expect this to occur. If the distribution is highly skewed, we might find a disproportionate amount on one side.
3. *The points seem to fall randomly above and below the center line.* If the distribution is stable, we would expect the same chances of getting a sample above the mean as below. However, if the distribution has shifted during the data collection process, we would expect a nonrandom distribution of sample statistics.
4. *There are no steady upward or downward trends of points moving toward either control limit.* These would indicate a gradual movement of the distribution mean.
5. *Most points, but not all, are near the center line; only a few are close to the control limits.* For a normal distribution, about 68% of observations fall within 1 standard deviation of the mean. If, for instance, we see a high proportion of points near the limits, we might suspect that the data came from two distinct distributions (visualize an inverted normal distribution).

For the syringe data, the  $R$ -chart has one point above the upper control limit. In the  $\bar{x}$ -chart, not only is one point below the lower control limit, but we see a clear upward trend. Thus, we would conclude that the process is not in control, particularly in the ability to maintain a stable average of the syringe length. It is important to keep good records of data—the time at which each sample was taken and the process conditions at that time (who was running the process, where the material came from, etc.). Most of the time, it is easy to identify a logical cause. A common reason for a point falling outside a control limit is an error in the calculation of the sample values of  $\bar{x}$  or  $R$ . Other possible causes are a sudden power surge, a broken tool, measurement error, or an incomplete or omitted operation in the process. Once in a while, however, they are a normal part of the process and occur simply by chance. When assignable causes are identified, these data should be deleted from the analysis, and new control limits should be computed.

The most common types of other out-of-control conditions are summarized next.

## **Sudden Shift in the Process Average**

When an unusual number of consecutive points fall on one side of the center line, it usually indicates that the process average has suddenly shifted. Typically, this is the result of an external influence that has affected the process; this would be a special cause. In both the  $\bar{x}$ - and  $R$ -charts, possible causes might be a new operator, a new inspector, a new machine setting, or a change in the setup or method. In the  $R$ -chart, if the shift is up, the process has become less uniform. Typical causes are carelessness of operators, poor or inadequate maintenance, or possibly a fixture in need of repair. If the shift is down in the  $R$ -chart, uniformity of the process has improved. This might be the result of improved workmanship or better machines or materials.

## **Cycles**

Cycles are short, repeated patterns in the chart, having alternative high peaks and low valleys. These are the result of causes that come and go on a regular basis. In the  $\bar{x}$ -chart, cycles may be the result of operator rotation or fatigue at the end of a shift, different gauges used by different inspectors, seasonal effects such as temperature or humidity, or differences between day and night shifts. In the  $R$ -chart, cycles can occur from maintenance schedules, rotation of fixtures or gauges, differences between shifts, or operator fatigue.

## **Trends**

A trend is the result of some cause that gradually affects the quality characteristics of the product and causes the points on a control chart to gradually move up or down from the center line. As a new group of operators gains experience on the job, for example, or as maintenance of equipment improves over time, a trend may occur. In the  $\bar{x}$ -chart, trends may be the result of improving operator skills, dirt or chip buildup in fixtures, tool wear, changes in temperature or humidity, or aging of equipment. In the  $R$ -chart, an increasing trend may be due to a gradual decline in material quality, operator fatigue, gradual loosening of a fixture or a tool, or dulling of a tool. A decreasing trend often is the result of improved operator skill, improved work methods, better purchased materials, or improved or more frequent maintenance.

## **Hugging the Center Line**

Hugging the center line occurs when nearly all the points fall close to the center line. In the control chart, it appears that the control limits are too wide. A common cause of this occurrence is the sample being taken by selecting one item systematically from each of several machines, spindles, operators, and so on. A simple example will serve to illustrate this. Suppose that one machine produces parts whose diameters average 7.508 with variation of only a few thousandths; and a second machine produces parts whose diameters average 7.502, again with only a small variation. Taken together, you can see that the range of variation would probably be between 7.500 and 7.510 and average about 7.505. Now suppose that we sample one part from each machine and compute a sample average to plot on an  $\bar{x}$ -chart. The sample averages will consistently be around 7.505 since one will always be high and the second will always be low. Even though there is a large variation in the parts taken as whole, the sample averages will not reflect this. In such a case, it would be more appropriate to construct a control chart for each machine, spindle, operator, and so on. An often overlooked cause for this pattern is miscalculation of the control limits, perhaps by using the wrong factor from the table or misplacing the decimal point in the computations.

## Hugging the Control Limits

This pattern shows up when many points are near the control limits with very few in between. It is often called a *mixture* and is actually a combination of two different patterns on the same chart. A mixture can be split into two separate patterns. A mixture pattern can result when different lots of material are used in one process or when parts are produced by different machines but fed into a common inspection group.

Quality control practitioners advocate simple rules, based on sound statistical principles, for identifying out-of-control conditions. For example, if eight consecutive points fall on one side of the center line, then you can conclude that the mean has shifted. Why? If the distribution is symmetric, then the probability that the next sample falls above or below the mean is 0.5. Because samples are independent, the probability that eight consecutive samples will fall on one side of the mean is  $(0.5)^8 = 0.0039$ —a highly unlikely occurrence. Another rule often used to detect a shift is finding 10 of 11 consecutive points on one side of the center line. The probability of this occurring can be found using the binomial distribution:

Probability of 10 out of 11 points on one side of center line

$$= \binom{11}{10} (0.5)^{10} (0.5)^1 = 0.00537$$

These examples show the value of statistics and data analysis in common production operations.

### SKILL-BUILDER EXERCISE 8.2

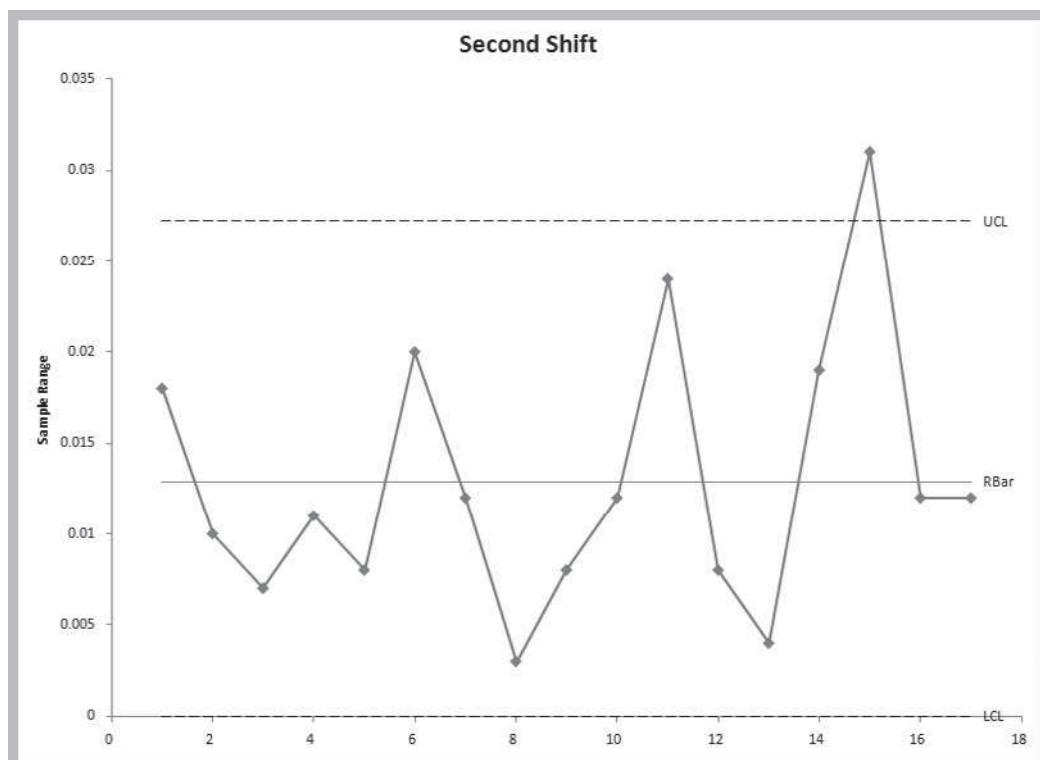
Using a spreadsheet, compute the probability of finding  $x$  consecutive points on one side of the center line if a process is in control, for  $x = 2$  to 10. Also find the probabilities of observing  $x$  out of  $n$  consecutive points on one side of the center line, where  $n$  varies from 8 to 12, and  $x$  varies from  $n - 3$  to  $n$ .

Let us return to the syringe data. After examining the first set of charts, a technician was called to adjust the machine prior to the second shift, and 17 more samples were taken. Figure 8.8 shows one of the calculation worksheets created by *PHStat* for developing the control charts. We see that the average range is now 0.0129 (versus 0.0257 in the first set of data) and the average mean is 4.9736 (versus 4.9581). Although the average dispersion appears to have been reduced, the centering of the process has gotten worse, since the target dimension is 4.950. The charts in Figures 8.9 and 8.10 also suggest that the variation has gotten out of control and that the process mean continues to drift upward.

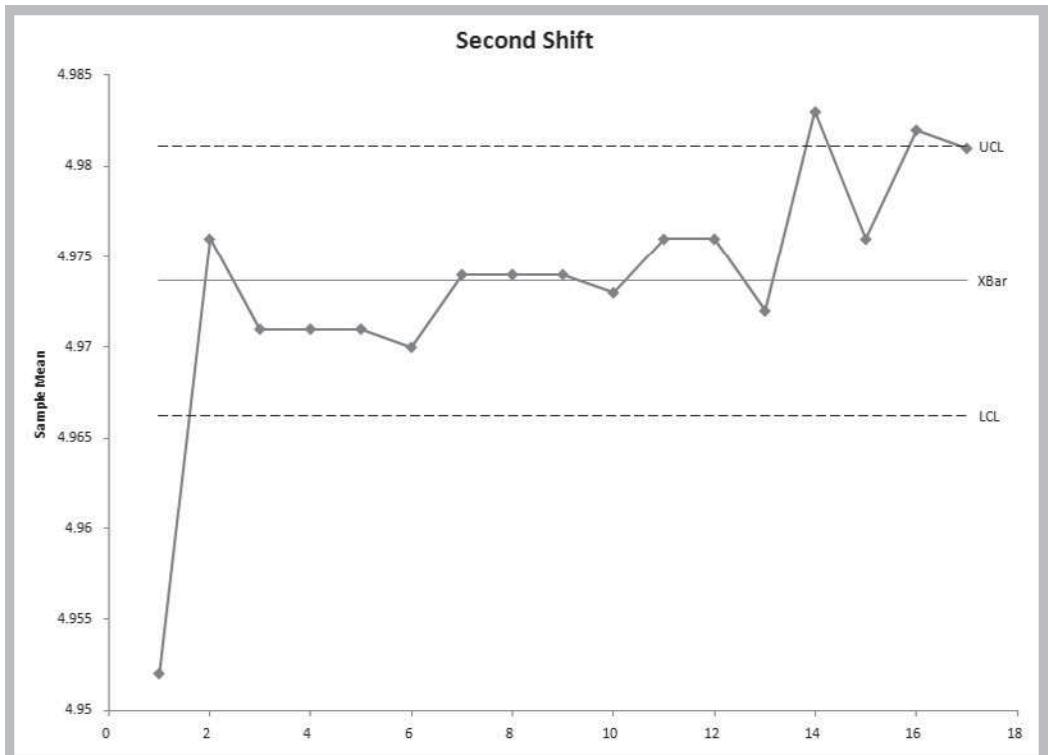
After another adjustment by a technician, the third shift collected another 15 samples. We leave it to you as an exercise to develop the control charts and verify that the  $R$ -chart appears to be in control, that the variability has stabilized, and that the  $\bar{x}$ -chart is also in control, although the average length of the syringes is 4.963, which is slightly above target. In reviewing the process, the maintenance technician discovered that as he tried to move the height adjustment stop down on its threaded shaft, it was difficult to tighten the locknut because of worn threads. As a result, the vibration from the machine loosened the locknut and adjustment cap, resulting in drifts off center of the lengths. When he set the adjustment cap a bit higher (resulting in a slightly higher average length), the threads were good enough to hold the locknut in place, reducing

A	B
1 Second Shift	
2	
3 Data	
4 Sample/Subgroup Size	5
5	
6 R Chart Intermediate Calculations	
7 RBar	0.012882353
8 D3 Factor	0
9 D4 Factor	2.114
10	
11 R Chart Control Limits	
12 Lower Control Limit	0
13 Center	0.012882353
14 Upper Control Limit	0.027233294
15	
16 XBar Chart Intermediate Calculations	
17 Average of Subgroup Averages	4.973647059
18 A2 Factor	0.577
19 A2 Factor * RBar	0.007433118
20	
21 XBar Chart Control Limits	
22 Lower Control Limit	4.966213941
23 Center	4.973647059
24 Upper Control Limit	4.981080176

**FIGURE 8.8 PHStat Calculations Worksheet  
for Second Shift Data**



**FIGURE 8.9 Second Shift R-Chart**



**FIGURE 8.10** Second Shift  $\bar{x}$ -Chart

the variation and bringing the process into control. This example shows the value of using control charts to help monitor a process and diagnose quality problems.

#### SKILL-BUILDER EXERCISE 8.3

Use *PHStat* to develop control charts for the third shift data in the syringe example. What can you conclude?

### CONTROL CHARTS FOR ATTRIBUTES

The most common control chart for attributes is the *p*-chart. A *p*-chart monitors the proportion of nonconforming items. Sometimes it is called a *fraction nonconforming* or *fraction defective* chart. For example, a *p*-chart might be used to monitor the proportion of checking account statements that are sent out with errors, FedEx packages delivered late, hotel rooms not cleaned properly, or surgical infections in a hospital.

As with variables data, a *p*-chart is constructed by first gathering 25 to 30 samples of the attribute being measured. For attributes data, it is recommended that each sample size be at least 100; otherwise, it is difficult to obtain good statistical results. The size of each sample may vary. It is usually recommended that a constant sample size be used as this makes interpreting patterns in the *p*-chart easier; however, for many applications this may not be practical or desirable.

The steps in constructing a *p*-chart are similar to those used for  $\bar{x}$ - and *R*-charts. We will first consider the case of a fixed sample size,  $n$ . Assume we have  $k$  samples, each of

size  $n$ . For each sample, we compute the fraction nonconforming,  $p$ , that is, the number of nonconforming items divided by the number in the sample. The average fraction nonconforming,  $\bar{p}$ , is computed by summing the total number of nonconforming items in all samples and dividing by the total number of items ( $= nk$  if the sample size is constant) in all samples combined. Because the number of nonconforming items in each sample follows a binomial distribution, the standard deviation is:

$$s = \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}} \quad (8.5)$$

Using the principles we described earlier in this chapter, upper and lower control limits are given by:

$$\text{UCL}_p = \bar{p} + 3s \quad (8.6)$$

$$\text{LCL}_p = \bar{p} - 3s \quad (8.7)$$

Whenever  $\text{LCL}_p$  turns out negative, we use zero as the lower control limit since the fraction nonconforming can never be negative. We may now plot the fraction nonconforming on a control chart just as we did for the averages and ranges and use the same procedures to analyze patterns in a  $p$ -chart as we did for  $\bar{x}$ - and  $R$ -charts. That is, we check that no points fall outside of the upper and lower control limits and that no peculiar patterns (runs, trends, cycles, and so on) exist in the chart.

To illustrate a  $p$ -chart, suppose that housekeeping supervisors at a hotel inspect 100 rooms selected randomly each day to determine if they were cleaned properly. Any nonconformance, such as a failure to replace used soap or shampoo or empty the wastebasket, results in the room being listed as improperly cleaned. The Excel file *Room Inspection* (shown in Figure 8.11) provides data for 25 days. The total number of nonconforming rooms is 55. Therefore, the average fraction nonconforming,  $\bar{p}$ , is  $55/2,500 = 0.022$ . This leads to the standard deviation:

$$s = \sqrt{\frac{0.022(1 - 0.022)}{100}} = 0.01467$$

The control limits are computed as:

$$\text{UCL}_p = \bar{p} + 3s = 0.022 + 3(0.01467) = 0.066$$

$$\text{LCL}_p = \bar{p} - 3s = 0.022 - 3(0.01467) = -0.022$$

Because the lower control limit is negative, we use 0.

A	B	C	D
1	Room Inspection Results		
2			
3	Sample	Rooms Inspected	Nonconforming Rooms
4	1	100	3
5	2	100	1
6	3	100	0
7	4	100	0
8	5	100	2

**FIGURE 8.11** Portion of Excel file *Room Inspection*



### Spreadsheet Note

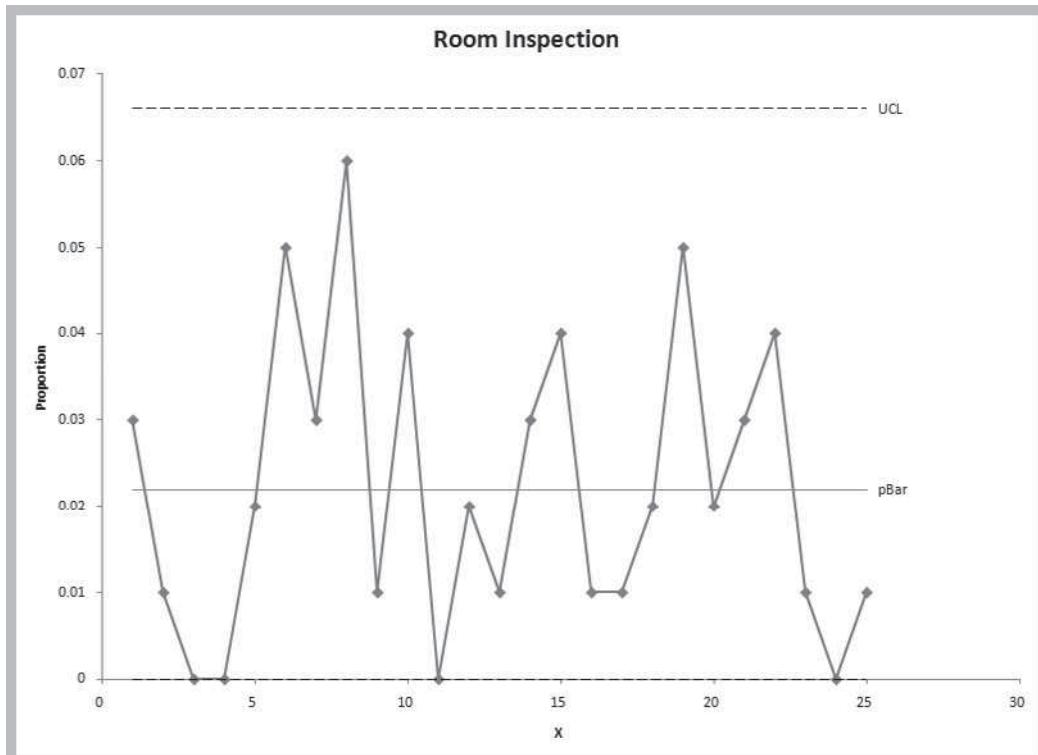
*PHStat* has a procedure for constructing  $p$ -charts (see Appendix 8.1B, *Creating p-Charts*). Figures 8.12 and 8.13 show the *Calculations* worksheet created by the tool and the resulting control chart. Although there is some variation in the proportion of rooms not cleaned properly, the chart appears to be in control, suggesting that this variation is due to common causes within the system (perhaps insufficient training of the housekeeping staff).

### Variable Sample Size

In many applications, it is desirable to use all data available rather than a sample. For example, hospitals collect monthly data on the number of infections after surgeries. To

A	B
1	Room Inspection
2	
3	Intermediate Calculations
4	Sum of Subgroup Sizes
5	2500
6	Number of Subgroups Taken
7	25
8	Average Sample/Subgroup Size
9	100
10	Average Proportion of Nonconforming Items
11	0.022
12	Three Standard Deviations
13	0.044005
10	p Chart Control Limits
11	Lower Control Limit
12	-0.022005
13	Center
	0.022
13	Upper Control Limit
	0.066005

**FIGURE 8.12**  $p$ -Chart Calculations Worksheet



**FIGURE 8.13**  $p$ -Chart for Room Inspection Data

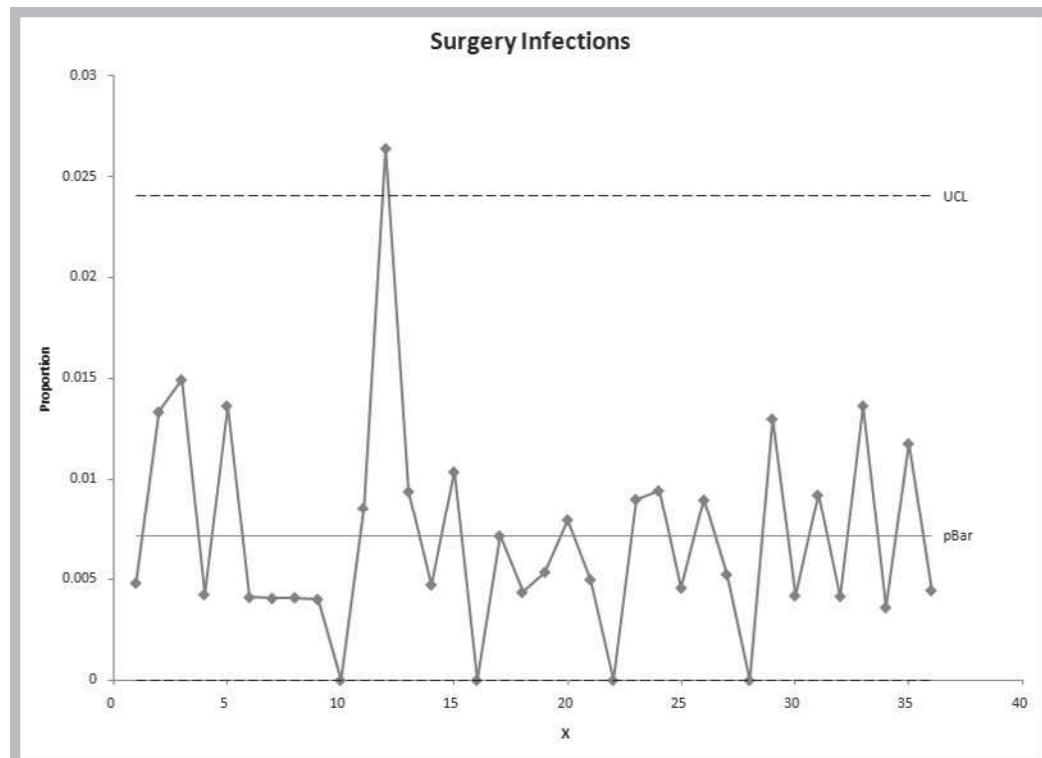
monitor the infection rate, a sample would not provide complete information. The Excel file *Surgery Infections* provides monthly data over a three-year period, a portion of which is shown in Figure 8.14. Because the sample size varies, we must modify the calculation of the standard deviation and control limits. One approach (used by *PHStat*—see the *PHStat* notes given in Chapter 6) is to compute the average sample size,  $\bar{n}$ , and use this value in the calculation of the standard deviation:

$$s = \sqrt{\frac{\bar{p}(1 - \bar{p})}{\bar{n}}} \quad (8.8)$$

Generally, this is acceptable as long as the sample sizes fall within 25% of the average. If sample sizes vary by a larger amount, then other approaches, which are beyond the scope of this book, should be used. When using this approach, note that because control limits are approximated using the average sample size, points that are actually out of control may not appear so on the chart and nonrandom patterns may be difficult to interpret; thus, some caution should be used. Figure 8.15 shows the control chart constructed

	A	B	C	D
1	<b>Surgery Infections</b>			
2				
3	Month	Surgeries	Infections	Infection Rate
4	1	208	1	0.0048
5	2	225	3	0.0133
6	3	201	3	0.0149
7	4	236	1	0.0042
8	5	220	3	0.0136

**FIGURE 8.14** Portion of Excel File *Surgery Infections*



**FIGURE 8.15** Control Chart for Surgery Infection Rate Using Average Sample Size

using *PHStat*. The chart shows that the infection rate exceeds the upper control limit in month 12, indicating that perhaps some unusual circumstances occurred at the hospital.

#### SKILL-BUILDER EXERCISE 8.4

Build an Excel spreadsheet template for calculating control chart parameters (center line and control limits) for a *p*-chart with and without variable sample sizes. Try to design the template to automatically generate the control charts.

## PROCESS CAPABILITY ANALYSIS

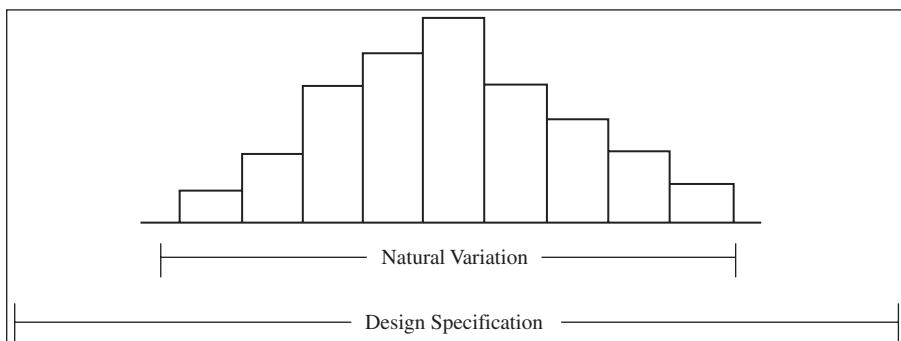
The purpose of SPC is to monitor a process over time to maintain a state of statistical control. However, just because a process is in control does not mean that it is capable of meeting specifications on the quality characteristic that is being measured. In the game of golf, for example, you might consistently shoot between 85 and 90; however, this is far from meeting the “specification”—par! **Process capability analysis** involves comparing the distribution of process output to specifications when only common causes (natural variations in materials, machines and tools, methods, operators, and the environment) determine the variation. As such, process capability is meaningless if special causes occur in the process. Therefore, before conducting a process capability analysis, control charts should be used to ensure that all special causes have been eliminated and that the process is in control.

Process capability is measured by the proportion of output that can be produced within design specifications. By collecting data, constructing frequency distributions and histograms, and computing basic descriptive statistics such as the mean and variance, we can better understand the nature of process variation and its ability to meet quality standards.

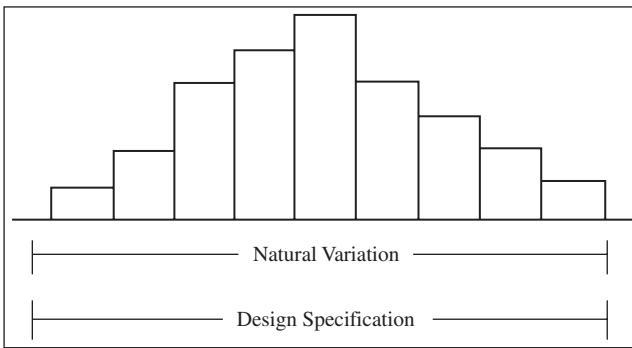
There are three important elements of process capability: the design specifications, the centering of the process, and the range of variation. Let us examine three possible situations:

**1. The natural variation in the output is smaller than the tolerance specified in the design (Figure 8.16).** The probability of exceeding the specification limits is essentially zero; you would expect that the process will almost always produce output that conforms to the specifications, as long as the process remains centered. Even slight changes in the centering or spread of the process will not affect its ability to meet specifications.

**2. The natural variation and the design specification are about the same (Figure 8.17).** A very small percentage of output might fall outside the specifications. The process should probably be closely monitored to make sure that the centering of the process does not drift and that the spread of variation does not increase.



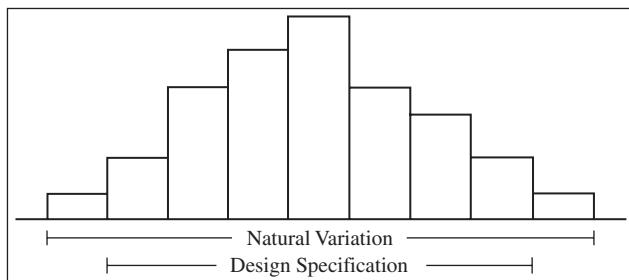
**FIGURE 8.16** Capable Process



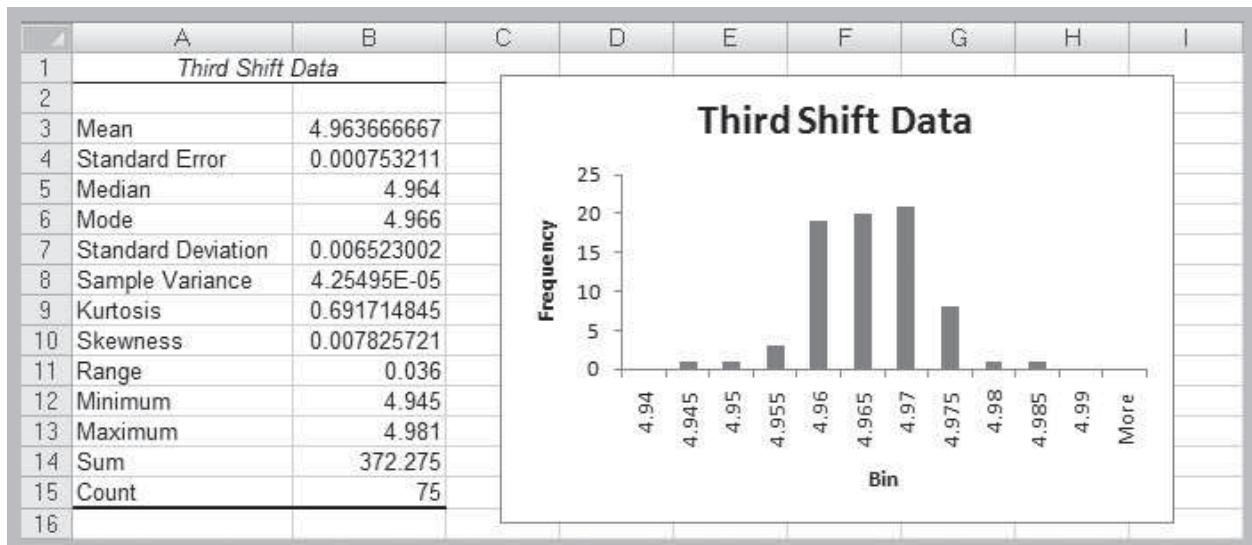
**FIGURE 8.17** Marginally Capable Process

3. The range of process variation is larger than the design specifications (Figure 8.18). The probability of falling in the tails of the distribution outside the specification limits is significant. The only way to improve product quality is to change the process.

To illustrate process capability analysis, we will use the third shift data in the Excel file *Syringe Samples*. Recall that the specifications of syringe lengths are  $4.950 \pm 0.030$  inches. Figure 8.19 shows the output of Excel's *Descriptive Statistics* and *Histogram* tools applied



**FIGURE 8.18** Incapable Process



**FIGURE 8.19** Summary Statistics and Histogram for Process Capability Analysis

to these data. The process mean is 4.963667, which is higher than the target specification for reasons described earlier. We would expect nearly all process output to fall within  $\pm 3$  standard deviations of the mean. With a standard deviation of 0.006523, this is  $4.963667 \pm 3(0.006523)$ , or between 4.944 and 4.983. This six standard deviation spread represents the capability of the process. As long as the process remains in control, we would expect the syringe lengths to fall within this range. Thus, we expect a small percentage of syringes to exceed the upper specification of 4.980. As you can see, one of the data points in this sample does exceed this specification.

The relationship between the process capability and the design specifications is often quantified by a measure called the **process capability index**, denoted by  $C_p$  (this statistic is unrelated to the  $C_p$  statistic that we discussed in the context of regression analysis in Chapter 6).  $C_p$  is simply the ratio of the specification width (that is, the difference between the upper and lower specification limits) to the process capability.

$$C_p = \frac{USL - LSL}{6s} \quad (8.9)$$

For the syringe example,  $C_p = (4.98 - 4.92)/(0.039) = 1.54$ . A  $C_p$  value greater than 1.0 means that the process is capable of meeting specifications; values lower than 1.0 mean that some nonconforming output will always be produced unless the process technology is improved. However, despite the fact that the capability is good, the process is not centered properly on the target value and some nonconforming output will occur.

### SKILL-BUILDER EXERCISE 8.5

Build a spreadsheet for computing the process capability index that allows you to enter up to 100 data points and apply it to the third shift syringe data.

## Basic Concepts Review Questions

- What kinds of variation in the production process are beyond our control and what kinds of variation require our intervention? Discuss some of the sources of each type of variation.
- What is the difference between variables and attributes?
- What is the aim of statistical process control? When do we consider a process to be *in control* and when is it *out of control*?
- Describe the steps involved in applying statistical process control.
- Explain the purpose and use of a control chart.
- What are nominal specifications and tolerances?
- Explain the differences between an  $\bar{x}$  chart and an  $R$  chart.
- Describe the common types of out-of-control conditions that one might find in a control chart.
- What are the differences between *process capability* and *process control*?

## Problems and Applications

- Find the upper and lower control limits for  $\bar{x}$  - and  $R$ -charts for the thickness of a special kind of disc when the sample grand mean (based on 40 samples of five observations each) is 1.5 millimeters and the average range is 0.03 millimeters.
- Suppose that the sample grand mean for the weight of a package of candies based on 35 samples of 12 candies each is 18 ounces with an average range of 0.8 ounces. Find the upper and lower control limits for the  $\bar{x}$  - and  $R$ -charts.
- The sample grand mean (based on 45 samples of seven observations each) for the weight of a can of beans is 250 grams with an average range of 8 grams. Find the upper and lower control limits for the  $\bar{x}$  - and  $R$ -charts.

- If 40 samples of 200 items are tested for nonconformity, and 240 of the 8000 items are defective, find the upper and lower control limits for a  $p$ -chart.
- If 30 samples of 100 items are run through a battery of tests, and 120 of the 3000 items are defective, calculate the upper and lower control limits for a  $p$ -chart.
- Suppose that an operation produces output with a standard deviation of 0.5 pounds, with an upper specification limit of 9 pounds and a lower specification limit of 4 pounds. Compute the process capability index.
- Suppose that the standard deviation of a machine manufacturing process is 0.24 inch, and while the upper and lower specification limits are 4.2 and 2.6 inches. What is the process capability index?

*Note: For the remaining problems, all data may be found in the Excel workbook Statistical Quality Control Problems. Data for each problem are found on a separate worksheet.*

- Compute control limits for  $\bar{x}$ - and  $R$ -charts for the data in the worksheet *Problem 8*. Construct and interpret the charts.
- Compute control limits for  $\bar{x}$ - and  $R$ -charts for the data in the worksheet *Problem 9*. Construct and interpret the charts.
- Hunter Nut Company produces cans of mixed nuts, advertised as containing no more than 20% peanuts. Hunter Nut Company wants to establish control limits for their process to ensure meeting this requirement.

They have taken 30 samples of 144 cans of nuts from the production process at periodic intervals, inspected each can, and identified the proportion of cans that did not meet the peanut requirement (shown in the worksheet *Problem 10*). Compute the average proportion nonconforming and the upper and lower control limits for this process. Construct the  $p$ -chart and interpret the results.

- A manufacturer of high-quality medicinal soap advertises its product as 99 and 44/100% free of medically offensive pollutants. Twenty-five samples of 100 bars of soap were gathered at the beginning of each hour of production; the numbers of bars not meeting this requirement are given in the worksheet *Problem 11*. Develop a  $p$ -chart for these data and interpret the results.
- A warehouse double-checks the accuracy of its order fulfillment process each day. The data in the worksheet *Problem 12* shows the number of orders processed each day and the number of inaccurate orders found. Construct a  $p$ -chart for these data and interpret the results.
- Refer to the data in the worksheet *Problem 8*. Suppose that the specification limits for the process are 90 (lower) and 110 (upper). Compute the process capability index and interpret. Eliminate any out-of-control points first, if any are found.
- Refer to the data in the worksheet *Problem 9*. Suppose that the specification limits for the process are 25 (lower) and 40 (upper). Compute the process capability index and interpret. Eliminate any out-of-control points first, if any are found.

## Case

### Quality Control Analysis

A manufacturer of commercial and residential lawn mowers has collected data from testing its products. These data are available in the Excel file *Quality Control Case Data*. The worksheet *Blade Weight* provides sample data from the manufacturing process of mower blades; *Mower Test* gives samples of functional performance test results;

and *Process Capability* contains additional sample data of mower blade weights. Use appropriate control charts and statistical analyses to provide a formal report to the manufacturing manager about quality issues he should understand or be concerned with.

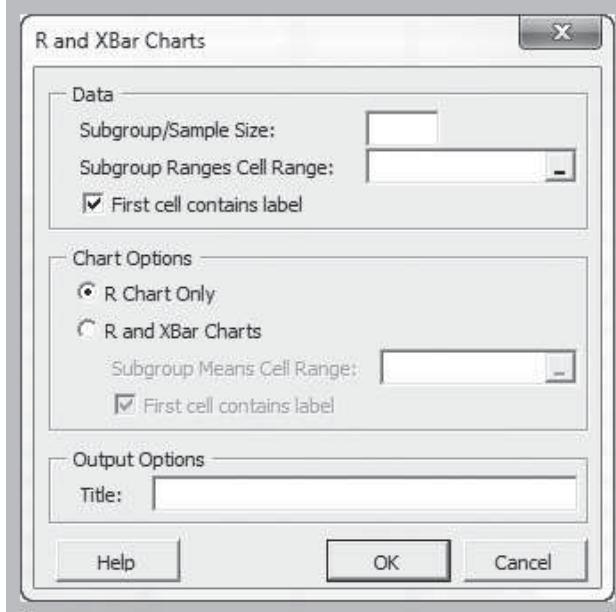
## APPENDIX 8.1

### PHStat Notes

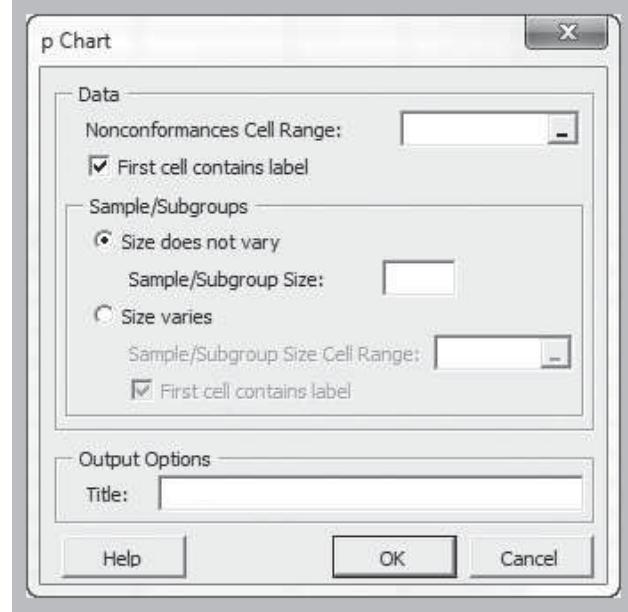
#### A. Creating $\bar{x}$ - and $R$ -Charts

From the *PHStat* menu, select *Control Charts*, followed by *R & Xbar Charts*. The dialog box that appears is shown in Figure 8A.1. Your worksheet must have already calculated the sample means and ranges. The cell ranges for these data

are entered in the appropriate boxes. You must also provide the sample size in the *Data* section and have the option of selecting only the  $R$ -chart or both the  $\bar{x}$ - and  $R$ -charts. *PHStat* will create several worksheets for calculations and the charts.



**FIGURE 8A.1** PHStat Dialog for *R* and *XBar* Charts



**FIGURE 8A.2** PHStat Dialog for *p*-Charts

## B. Creating *p*-Charts

From the *PHStat* menu, select *Control Charts* followed by *p-Chart*. The dialog box, shown in Figure 8A.2, prompts you for the cell range for the number of nonconformances and

the sample size. This procedure also allows you to have non-constant sample sizes; if so, you need to enter the cell range of the sample size data. *PHStat* creates several new worksheets for the calculations and the actual chart.

## PART II

# DECISION MODELING AND ANALYSIS

## *Chapter 9*

# Building and Using Decision Models

- INTRODUCTION 295
- DECISION MODELS 296
- MODEL ANALYSIS 299
  - What-If Analysis 299
  - Model Optimization 302
- TOOLS FOR MODEL BUILDING 304
  - Logic and Business Principles 304
  - Common Mathematical Functions 305
  - Data Fitting 306
  - Spreadsheet Engineering 308
- SPREADSHEET MODELING EXAMPLES 309
  - New Product Development 309
  - Single Period Purchase Decisions 311
  - Overbooking Decisions 312
  - Project Management 313
- MODEL ASSUMPTIONS, COMPLEXITY, AND REALISM 315
- BASIC CONCEPTS REVIEW QUESTIONS 317
- PROBLEMS AND APPLICATIONS 318
- CASE: AN INVENTORY MANAGEMENT DECISION MODEL 321
- APPENDIX 9.1: EXCEL NOTES 322
  - A. Creating Data Tables 322
  - B. Using the *Scenario Manager* 322
  - C. Using *Goal Seek* 323
  - D. Net Present Value and the NPV Function 323

### **INTRODUCTION**

Everyone makes decisions. Individuals face personal decisions such as choosing a college or graduate program, making product purchases, selecting a mortgage instrument, and investing for retirement. Managers in business organizations must determine what products to make and

how to price them, where to locate facilities, how many people to hire, where to allocate advertising budgets, whether or not to outsource a business function, and how to schedule production. Developing effective strategies to deal with these types of problems can be a difficult task. Quantitative decision models can greatly assist in these types of decisions. Part II of this book is devoted to the development and application of decision models.

Spreadsheets, in particular, provide a convenient means to manage data, construct models, and analyze them for gaining insight and supporting decisions. Although the early applications of spreadsheets were primarily in accounting and finance, spreadsheets have developed into powerful general-purpose managerial tools for decision modeling and analysis.

In this chapter, we introduce approaches for building decision models, implementing them on spreadsheets, and analyzing them to provide useful business information.

## DECISION MODELS

A **model** is an abstraction or representation of a real system, idea, or object. Models capture the most important features of a problem and present them in a form that is easy to interpret. A model can be a simple picture, a spreadsheet, or a set of mathematical relationships. A **decision model** is one that can be used to understand, analyze, or facilitate making a decision. Decision models generally have three types of inputs:

1. **Data**, which are assumed to be constant for purposes of the model. Some examples would be costs, machine capacities, and intercity distances.
2. **Uncontrollable variables**, which are quantities that can change but cannot be directly controlled by the decision maker. Some examples would be customer demand, inflation rates, and investment returns.
3. **Decision variables**, which are controllable and can be selected at the discretion of the decision maker. Some examples would be production quantities, staffing levels, and investment allocations.

Decision models characterize the relationships among the data, uncontrollable variables, and decision variables and the outputs of interest to the decision maker (see Figure 9.1). A spreadsheet is one way of expressing a decision model through the formulas entered in the cells that reflect the relationships among the model components. For any set of inputs, the spreadsheet calculates some output measures of interest. Spreadsheets are ideal vehicles for implementing decision models because of their versatility in managing data, evaluating different scenarios, and presenting results in a meaningful fashion.

Figure 9.2 shows an example of a spreadsheet for evaluating a simple outsourcing decision. Suppose that a manufacturer can produce a part for \$125/unit with a fixed cost of \$50,000. The alternative is to outsource production to a supplier at a unit cost of \$175. The decision depends on the anticipated volume of demand; for high volumes, the cost to manufacture in-house will be lower than outsourcing, because the fixed costs can be spread over a large number of units. For small volumes, it would be more economical to outsource. Knowing the cost of both alternatives and the breakeven point would facilitate the decision. The data consist of the costs associated with manufacturing the product in-house or purchasing it from an outside supplier. The key model input

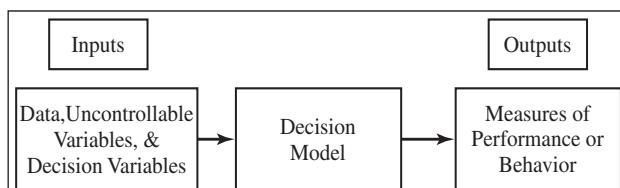


FIGURE 9.1 Nature of Decision Models

	A	B
1	Outsourcing Decision Model	
2		
3	Data	
4		
5	Manufactured in-house	
6	Fixed cost	\$ 50,000
7	Unit variable cost	\$ 125
8		
9	Purchased from supplier	
10	Unit cost	\$ 175
11		
12	Model	
13		
14	Demand volume	1500
15		
16	Total manufacturing cost	\$ 237,500
17	Total purchased cost	\$ 262,500
18	Difference	\$ (25,000)
19		
20	Decision	Manufacture

	A	B
1	Outsourcing Decision Model	
2		
3	Data	
4		
5	Manufactured in-house	
6	Fixed cost	50000
7	Unit variable cost	125
8		
9	Purchased from supplier	
10	Unit cost	175
11		
12	Model	
13		
14	Demand volume	1500
15		
16	Total manufacturing cost	=B6+B7*B14
17	Total purchased cost	=B14*B10
18	Difference	=B16-B17
19		
20	Decision	=IF(B18<=0, "Manufacture", "Outsource")

**FIGURE 9.2** Outsourcing Decision Model

is the demand volume and the outputs are the total manufacturing and purchase cost. Calculating the outputs basically consists of “stepping through” the formulas.

This model can also be expressed mathematically by defining symbols for each component of the model:

$F$  = fixed cost of in-house manufacturing

$V$  = unit variable cost of in-house manufacturing

$C$  = unit cost of outsourcing

$D$  = demand volume

Then the total manufacturing cost can be written as  $TMC = F + V \times D$ , and the total outsourcing cost as  $TOC = C \times D$ . Note the correspondence between the spreadsheet formulas and the mathematical model:

$$TMC = B6 + B7 * B14$$

$$TOC = B14 * B10$$

Thus, if you can write a spreadsheet formula, you can develop a mathematical model!

Mathematical models are easy to manipulate; for example, it is easy to find the breakeven volume by setting  $TMC = TOC$  and solving for  $D$ :

$$F + V \times D = C \times D$$

$$D = F/(C - V)$$

In contrast, it is more difficult to find the breakeven volume using trial-and-error on the spreadsheet without knowing some advanced tools and approaches. However, spreadsheets have the advantage of allowing you to easily modify the model inputs and calculate the numerical results. We will use both spreadsheets and analytical modeling approaches in our model-building applications, and it is important to be able to “speak both languages.”

Models complement decision makers' intuition and often provide insights that intuition cannot. For example, one early application of management science in marketing involved a study of sales operations. Sales representatives had to divide their time between large and small customers and between acquiring new customers and keeping old ones. The problem was to determine how the representatives should best allocate their time. Intuition suggested that they should concentrate on large customers and that it was much harder to acquire a new customer than to keep an old one. However, intuition could not tell whether they should concentrate on the 500 largest or the 5,000 largest customers or how much effort to spend on acquiring customers. Models of sales force effectiveness and customer response patterns provided the insight to make these decisions. However, it is important to understand that all models are only representations of the "real world," and as such, cannot capture every nuance that decision makers face in reality. Decision makers must often modify the policies that models suggest to account for intangible factors that they might not have been able to incorporate into the model.

Decision models take many different forms. Some models are **descriptive**; they simply describe relationships and provide information for evaluation. Regression models that we studied in Chapter 6 are examples of descriptive models; they describe relationships between the dependent and independent variables. The outsourcing model is also descriptive in that it simply allows you to evaluate the cost impact of different demand volumes. Note that demand volume is an uncontrollable input we can vary to evaluate the costs of the decision alternatives. The model itself does not include any decision variables. Descriptive models are used to explain the behavior of systems, to predict future events as inputs to planning processes, and to assist decision makers in making decisions.

Other models, called *optimization models*, are **prescriptive**; they seek to determine an optimal policy, that is, the best course of action that a decision maker should take to maximize or minimize some objective. In a highly competitive world where one percentage point can mean a difference of hundreds of thousands of dollars or more, knowing the best solution can mean the difference between success and failure. To illustrate an example of an optimization model, suppose that an airline has studied the price elasticity for the demand for a round trip between Chicago and Cancun. They discovered that when the price is \$600, daily demand is 500 passengers per day, but when the price is \$300, demand increases to 1,200 passengers per day. The airplane capacity is 300 passengers, but the airline will add additional flights if demand warrants. The fixed cost associated with each flight is \$90,000. The decision is to determine the price that maximizes profit.

To develop the optimization model, we have to first characterize demand as a function of price. Because we are provided with only two data points (\$600, 500) and (\$300, 1,200), we can assume that demand is a linear function of price and determine the equation of the straight line between them using algebra. Using the basic equation of a straight line,  $y = mx + b$ , where  $y$  is the demand,  $x$  is the price,  $m$  is the slope, and  $b$  is the  $y$ -intercept, we calculate the slope and intercept as:

$$m = (500 - 1,200)/(\$600 - \$300) = -7/3 \quad \text{and}$$

$$b = y - mx = 500 - (-7/3)(600) = 1,900$$

Thus, demand =  $1,900 - 7/3$  price. The number of flights/day would be calculated as the demand divided by the airplane capacity (300), rounded up to the next whole number. Now it is easy to calculate daily profit as:

$$\text{Profit} = \text{Demand} \times \text{Price} - \text{Fixed cost} \times \text{Flights/day}$$

Figure 9.3 shows a spreadsheet model for this scenario (Excel file *Airline Pricing Model*). The objective is to find the price that yields the largest profit. In this model, the unit price is a decision variable. The ROUNDUP function is used in cell B15 to ensure that a sufficient number of flights are scheduled to meet demand.

	A	B
1	Airline Pricing Model	
2		
3	Data	
4	Airplane capacity	300
5	Fixed cost	\$ 90,000
6	Demand function	
7	slope	-2.33
8	intercept	1900
9		
10	Model	
11		
12	Revenue	
13	Unit price	\$ 500.00
14	Demand	733
15	Number of flights/day	3
16	Total Revenue	\$366,666.67
17	Cost	
18	Fixed Cost	\$270,000.00
19		
20	Profit	\$96,666.67

	A	B
1	Airline Pricing Model	
2		
3	Data	
4	Airplane capacity	300
5	Fixed cost	90000
6	Demand function	
7	slope	=-7/3
8	intercept	1900
9		
10	Model	
11		
12	Revenue	
13	Unit price	500
14	Demand	=B8+B7*B13
15	Number of flights/day	=ROUNDUP(B14/B4,0)
16	Total Revenue	=B13*B14
17	Cost	
18	Fixed Cost	=B5*B15
19		
20	Profit	=B16-B18

**FIGURE 9.3** Airline Pricing Model Spreadsheet

## MODEL ANALYSIS

A model helps managers to gain insight into the nature of the relationships among components of a problem, aids intuition, and provides a vehicle for communication. We might be interested in:

1. Studying the impact of changes in assumptions on model outputs
2. Finding a solution such as a breakeven value
3. Determining the best solution to an optimization model
4. Evaluating risks associated with decision alternatives

Excel provides several tools for model analysis—data tables, *Scenario Manager*, and goal seek—which we introduce next. Risk analysis is the subject of the next two chapters.

## What-If Analysis

Spreadsheet models allow you to easily evaluate “what-if” questions—how specific combinations of inputs that reflect key assumptions will affect model outputs. For instance, in the outsourcing decision model described earlier, we might be interested in how different levels of fixed and variable costs affect the total manufacturing cost and the resulting decision. The process of changing key model inputs to determine their effect on the outputs is also often called **sensitivity analysis**. This is one of the most important and valuable approaches to gaining the appropriate insights to make good decisions.

Sensitivity analysis is as easy as changing values in a spreadsheet and recalculating the outputs. However, systematic approaches make this process easier and more useful. Spreadsheets facilitate sensitivity analysis. Excel 2010 provides several tools—data tables, *Scenario Manager*, and *Goal Seek*. These can be found within the *What-If Analysis* menu in the *Data* tab.

**DATA TABLES** Data tables summarize the impact of one or two inputs on a specified output. Excel allows you to construct two types of data tables (see Appendix 9.1A, *Creating Data Tables*). A **one-way data table** evaluates an output variable over a range of



values for a single input variable. **Two-way data tables** evaluate an output variable over a range of values for two different input variables.

We will illustrate the use of data tables to evaluate the impact of cost assumptions in the outsourcing decision model. Suppose we wish to create a one-way data table to evaluate the difference in manufacturing and outsourcing cost and the best decision for varying levels of fixed costs, holding all other model inputs constant. The range of these input values is shown in D4:D11 in Figure 9.4. In cell E3 enter the cell reference for the difference cell (=B18), and in cell F3 enter the cell reference for the decision (=B20). Next, select the data table range, the smallest rectangular block that includes the formula and all the values in the input range (D3:F11). In the Column Input cell of the dialog box, enter B6, the cell that contains the input value of fixed cost. Excel evaluates the difference and decision for each value in the data table as shown in Figure 9.4.

A two-way data table showing the decision for various values of fixed and variable costs is shown in Figure 9.5. Two-way data tables can only evaluate one output variable. In this case, we evaluate the decision in cell B20 of the model and reference this

A	B	C	D	E	F	G
1	Outsourcing Decision Model					
2						
3	Data					
4						
5	Manufactured in-house					
6	Fixed cost	\$ 50,000				
7	Unit variable cost	\$ 125				
8						
9	Purchased from supplier					
10	Unit cost	\$ 175				
11						
12	Model					
13						
14	Demand volume	1500				
15						
16	Total manufacturing cost	\$ 237,500				
17	Total purchased cost	\$ 262,500				
18	Difference	\$ (25,000)				
19						
20	Decision	Manufacture				

**FIGURE 9.4** One-Way Data Table

A	B	C	D	E	F	G	H	I	J	K
1	Outsourcing Decision Model									
2										
3	Data									
4										
5	Manufactured in-house									
6	Fixed cost	\$ 50,000								
7	Unit variable cost	\$ 125								
8										
9	Purchased from supplier									
10	Unit cost	\$ 175								
11										
12	Model									
13										
14	Demand volume	1500								
15										
16	Total manufacturing cost	\$ 237,500								
17	Total purchased cost	\$ 262,500								
18	Difference	\$ (25,000)								
19										
20	Decision	Manufacture								

**FIGURE 9.5** Two-Way Data Table

in cell D16. In the *Data Table* dialog box, the row input cell is B7 and the column input cell is B6. The result shows those combination of inputs for which manufacturing or outsourcing is more economical.

### SKILL-BUILDER EXERCISE 9.1

Develop a one-way data table for the *Airline Pricing Model* for airplane capacities ranging from 200 to 360 in increments of 20. Then develop a two-way data table for airplane capacities ranging from 200 to 360 in increments of 20, and fixed costs ranging from \$60,000 to \$120,000 in increments of \$10,000.

**SCENARIO MANAGER** The Excel *Scenario Manager* tool allows you to create **scenarios**—sets of values that are saved and can be substituted automatically on your worksheet (see Appendix 9.1B, *Using the Scenario Manager*). Scenarios are useful for conducting What-If analyses when you have more than two variables (which data tables cannot handle). The *Scenario Manager* can handle up to 32 variables. For example, suppose that the fixed and variable costs, as well as the demand volume are uncertain but that the supplier cost is fixed. Through discussions with key managers, you have determined best-case, worst-case, and most-likely scenarios for these inputs:



Spreadsheet Note

	Fixed Cost	Unit Variable Cost	Demand Volume
Best case	\$40,000	\$120	1,800
Worst case	\$60,000	\$140	1,000
Most likely case	\$55,000	\$125	1,500

The *Scenario Manager* evaluates the model for each of these situations and creates the summary report shown in Figure 9.6. This indicates that among these three scenarios, only the worst case results in a decision to outsource the part.

A	B	C	D	E	F	G
1						
Scenario Summary						
Current Values:		Best case		Worst case		Most likely case
Changing Cells:		\$B\$6	\$ 50,000	\$ 40,000	\$ 60,000	\$ 55,000
		\$B\$7	\$ 125	\$ 120	\$ 140	\$ 125
		\$B\$14	1500	1800	1000	1500
Result Cells:		\$B\$18	\$ (25,000)	\$ (59,000)	\$ 25,000	\$ (20,000)
		\$B\$20	Manufacture	Manufacture	Outsource	Manufacture
Notes: Current Values column represents values of changing cells at time Scenario Summary Report was created. Changing cells for each scenario are highlighted in gray.						

FIGURE 9.6 Scenario Summary for Outsourcing Model

## SKILL-BUILDER EXERCISE 9.2

Use the *Scenario Manager* to evaluate the following scenarios in the *Airline Pricing Model* spreadsheet.

	Fixed Cost	Unit Price
High	\$100,000	\$700
Low	\$80,000	\$500

**GOAL SEEK** If you know the result that you want from a formula, but are not sure what input value the formula needs to get that result, use the *Goal Seek* feature in Excel. *Goal Seek* works only with one variable input value. If you want to consider more than one input value or wish to maximize or minimize some objective, you must use the *Solver* add-in, which will be described later. For example, in the outsourcing decision model, you might be interested in finding the breakeven point. The breakeven point would be the value of demand volume for which total manufacturing cost would equal total purchased cost, or equivalently, for which the difference is 0. Therefore, you seek to find the value of demand volume in cell B14 that yields a value of 0 in cell B18. In the *Goal Seek* dialog box (see Appendix 9.1C, *Using Goal Seek*), enter B18 for the *Set cell*, enter 0 in the *To value* box, and B14 in the *By changing cell* box. The *Goal Seek* tool determines that the breakeven demand is 1,000.



Spreadsheet Note

## SKILL-BUILDER EXERCISE 9.3

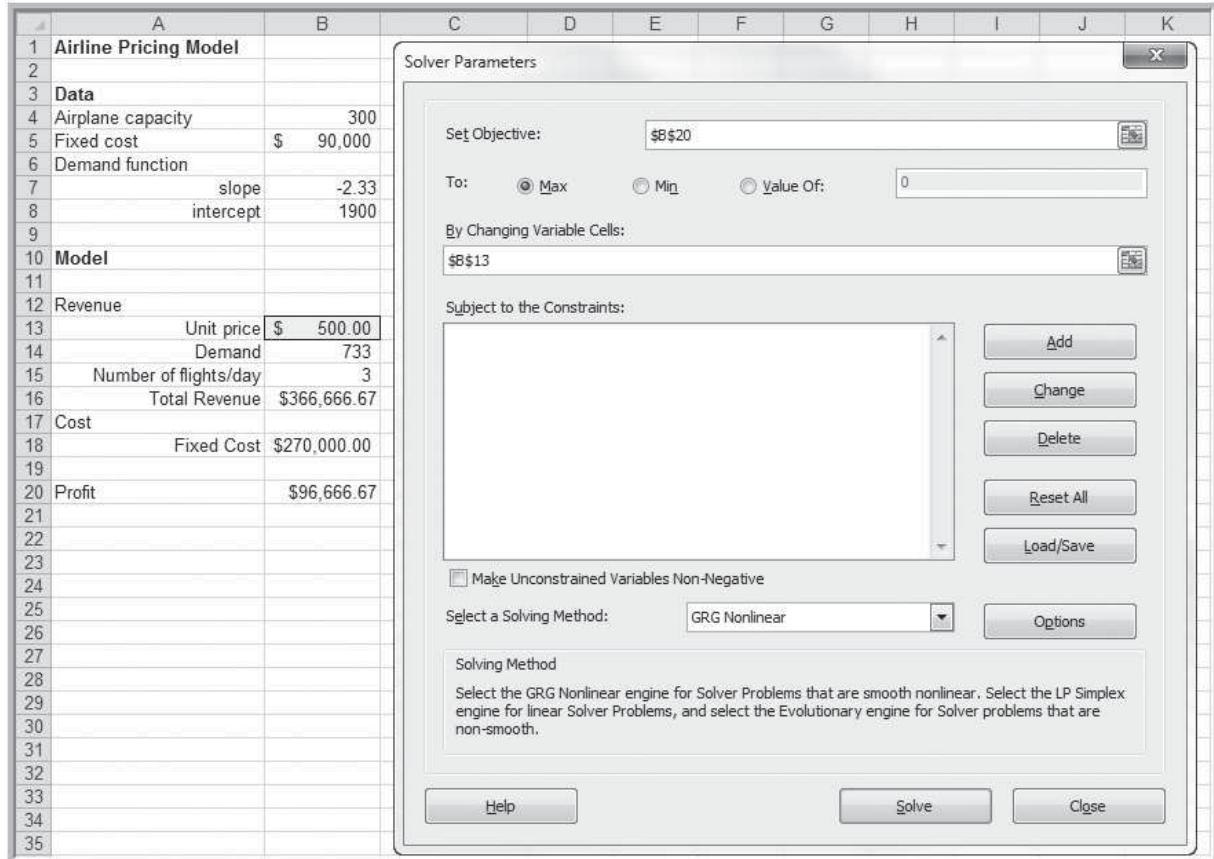
Apply the *Goal Seek* tool to find the unit price to yield a profit of \$60,000 in the *Airline Pricing Model* spreadsheet.

## Model Optimization

What-if analyses are useful approaches for descriptive models; however, the purpose of optimization models is to find the *best* solution. For some models, analytical solutions—closed-form mathematical expressions—can be obtained using such techniques as calculus. In most cases, however, an algorithm is needed to provide the solution. An **algorithm** is a systematic procedure that finds a solution to a problem. Researchers have developed algorithms to solve many types of optimization problems. However, we will not be concerned with the detailed mechanics of these algorithms; our focus will be on the use of the algorithms to solve the models we develop.

Excel includes an add-in called *Solver* that allows you to find optimal solutions to optimization problems formulated as spreadsheet models. *Solver* was developed and is maintained by Frontline Systems, Inc. ([www.solver.com](http://www.solver.com)). Frontline Systems also supports a more powerful version of *Solver*, *Premium Solver*, which is available with this book as part of the *Risk Solver Platform* add-in. We will use it exclusively in the last two chapters of this book. For now, we will use the standard version.

We will illustrate the use of *Solver* for the airline pricing model. *Solver* can be found in the *Analysis* group under the *Data* tab. (If *Solver* is not activated, choose *Options* from the *File* tab in Excel 2010 and select *Add-ins*. Click the *Go* button to manage *Excel Add-ins* and check the box for *Solver Add-in*.) When *Solver* is started, the *Solver Parameters* dialog



**FIGURE 9.7** Solver Parameters Dialog

box appears as shown in Figure 9.7. *Set Objective* is the cell in the spreadsheet that you wish to optimize; in this case the profit function in cell B20. Because we want to find the largest profit, we click the radio button for *Max*. *Changing Variable Cells* are another name for decision variables. In this example, we wish to find the unit price (cell B13) that results in the maximum profit. Ignore the *Subject to the Constraints* option for now. *Select a Solving Method* allows you do use different algorithms to solve the model; these will be explained in Chapters 13 and 14. For now, leave it as is with *GRG Nonlinear* chosen. Click the *Solve* button. *Solver* will display a *Solver Results* dialog. Click *OK*, and the solution will be shown in the spreadsheet:

Unit price = \$428.57 and Profit = \$115,714.28

The *Solver Results* dialog also provides the option to view several reports, but these are not important at this point. We will describe these and other features of *Solver* in Chapter 13.

Although the airline pricing model did not, most optimization models have **constraints**—limitations, requirements, or other restrictions that are imposed on any solution—such as “do not exceed the allowable budget” or “ensure that all demand is met.” For instance, a consumer products company manager would probably want to ensure that a specified level of customer service is achieved with the redesign of the distribution system. The presence of constraints makes modeling and solving optimization

problems more challenging, and we will address constrained optimization problems in the last two chapters of the book.

If possible, we would like to ensure that an algorithm such as the one *Solver* uses finds an optimal solution. However, some models are so complex that it is impossible to solve them optimally in a reasonable amount of computer time because of the extremely large number of computations that may be required or because they are so complex that an optimal solution cannot be guaranteed. In these cases, analysts use **heuristics**—solution procedures that generally find good solutions without guarantees of finding an optimal solution. (The term *heuristic* stems from a Greek word meaning “to discover.”) Researchers have developed powerful heuristics to obtain good solutions to extremely difficult optimization problems. Another practical reason for using heuristics is that a manager might be satisfied with a solution that is good but not necessarily optimal, particularly when:

- Inexact or limited data used to estimate uncontrollable quantities in models may contain more error than that of a nonoptimal solution
- The assumptions used in a model make it an inaccurate representation of the real problem, making having the “best” solution pointless
- Anything better than the current solution will suffice, so long as it can be obtained at a reasonable cost and in a reasonable amount of time

## TOOLS FOR MODEL BUILDING

Building decision models is more of an art than a science; however, there are many approaches that can facilitate the development of useful models.

### Logic and Business Principles

Building good decision models requires a solid understanding of basic business principles in all functional areas (such as accounting, finance, marketing, and operations), knowledge of business practice and research, and logical skills. For example, suppose that you wish to create a model to compute the profit associated with production and sale of a product. A fundamental business principle is:

$$\text{Profit} = \text{Revenue} - \text{Cost} \quad (9.1)$$

Using your knowledge and experience, you can expand *Revenue* and *Cost* terms as follows:

$$\text{Revenue} = (\text{Unit price})(\text{Quantity sold})$$

$$\text{Cost} = [\text{Fixed cost} + (\text{Unit cost})(\text{Quantity produced})]$$

Thinking more about this, you might realize that *Quantity sold* is related to both *Quantity produced* and the demand for the product. Specifically, the quantity sold must be equal to the smaller of the demand or the quantity produced:

$$\text{Quantity sold} = \text{Min}(\text{Quantity produced}, \text{Demand})$$

Therefore, the final model is:

$$\begin{aligned}\text{Profit} &= (\text{Unit price}) \text{Min}(\text{Quantity produced}, \text{Demand}) \\ &\quad - [\text{Fixed cost} + (\text{Unit cost})(\text{Quantity produced})]\end{aligned}$$

Many business ventures are evaluated on the basis of financial criteria such as *net present value (NPV)*. This will be used in several examples and problems in the remainder of this book; more complete discussions can be found in basic finance texts.

See Appendix 9.1D, *Net Present Value and the NPV Function* for details about implementing it on spreadsheets. The more you learn about fundamental theory in business, the better prepared you will be to develop good models. Another example of using logic in modeling follows.



Spreadsheet Note

**GASOLINE CONSUMPTION MODEL** Automobiles have different fuel economies (mpg), and commuters drive different distances to work or school. Suppose that a state Department of Transportation (DOT) is interested in measuring the average monthly fuel consumption of commuters in a certain city. The DOT might sample a group of commuters and collect information on the number of miles driven per day, number of driving days per month, and the fuel economy of their cars.

We can develop a model for the amount of gasoline consumed by first applying a bit of logic. If a commuter drives  $m$  miles per day and  $d$  days/month, then the total number of miles driven per month is  $m \times d$ . If a car gets  $f$  miles per gallon in fuel economy, then the number of gallons consumed per month must be  $(m \times d)/f$ . Notice that the dimensions of this expression are (miles/day)(days/month)(miles/gallon) = gallons/month. Consistency in dimensions is an important validation check for model accuracy.

### SKILL-BUILDER EXERCISE 9.4

Implement the gasoline consumption model on a spreadsheet. Survey five of your fellow students and obtain data on miles driven per day on a routine trip (work or school), number of days per month, and fuel economy of their vehicles. Use the *Scenario Manager* to define scenarios for each individual and create a Scenario Summary.

## Common Mathematical Functions

Understanding different functional relationships is instrumental in model building. For example, in the airline pricing model, we developed a linear function relating price and demand using two data points. Common types of mathematical functions used in models include:

**Linear:**  $y = mx + b$ . Linear functions show steady increases or decreases over the range of  $x$ .

**Logarithmic:**  $y = \ln(x)$ . Logarithmic functions are used when the rate of change in a variable increases or decreases quickly and then levels out, such as with diminishing returns to scale.

**Polynomial:**  $y = ax^2 + bx + c$  (second order—quadratic function),  
 $y = ax^3 + bx^2 + dx + e$  (third order—cubic function), etc. A second-order polynomial is parabolic in nature and has only one hill or valley; a third-order polynomial has one or two hills or valleys. Revenue models that incorporate price elasticity are often polynomial functions.

**Power:**  $y = ax^b$ . Power functions define phenomena that increase at a specific rate. Learning curves that express improving times in performing a task are often modeled with power functions having  $a > 0$  and  $b < 0$ .

**Exponential:**  $y = ab^x$ . Exponential functions have the property that  $y$  rises or falls at constantly increasing rates. For example, the perceived brightness of a light bulb grows at a decreasing rate as the wattage increases. In this case,  $a$  would be a positive number and  $b$  would be between 0 and 1. The exponential function is often defined as  $y = ae^x$ , where  $b = e$ , the base of natural logarithms (approximately 2.71828).

## Data Fitting

For many applications, functional relationships used in decision models are derived from the analysis of data. The Excel *Trendline* tool (described in Chapter 6) provides a convenient method for determining the best fitting functional relationship. Figure 9.8 shows a chart of historical data on crude oil prices on the first Friday of each month from January 2006 through June 2008 (data are in the Excel file *Crude Oil Prices*). Using the *Trendline* tool, we can try to fit the various functions to these data (here  $x$  represents the number of months starting with January 2006). The results are:

$$\text{Exponential: } y = 50.49e^{0.021x} \quad R^2 = 0.664$$

$$\text{Logarithmic: } y = 13.02\ln(x) + 39.60 \quad R^2 = 0.382$$

$$\text{Polynomial (second order): } y = 0.130x^2 - 2.399x + 68.01 \quad R^2 = 0.905$$

$$\text{Polynomial (second order): } y = 0.005x^3 - 0.111x^2 + 0.648x + 59.49 \quad R^2 = 0.928$$

$$\text{Power: } y = 45.96x^{0.169} \quad R^2 = 0.397$$

The best-fitting model is the third-order polynomial, shown in Figure 9.9.

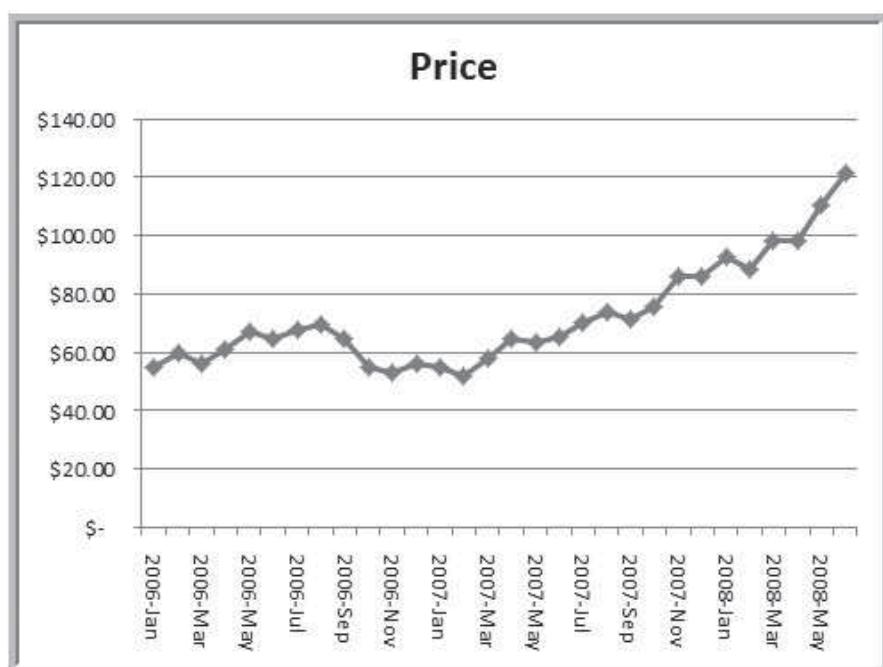
Of course, the proper model to use depends on the scope of the data. As the chart shows, crude oil prices were relatively stable until early 2007 and then began to increase rapidly. By including the early data, the long-term functional relationship might not adequately express the short-term trend. For example, fitting a model to only the data beginning with January 2007 yields the best-fitting models:

$$\text{Exponential: } y = 50.56e^{0.044x} \quad R^2 = 0.969$$

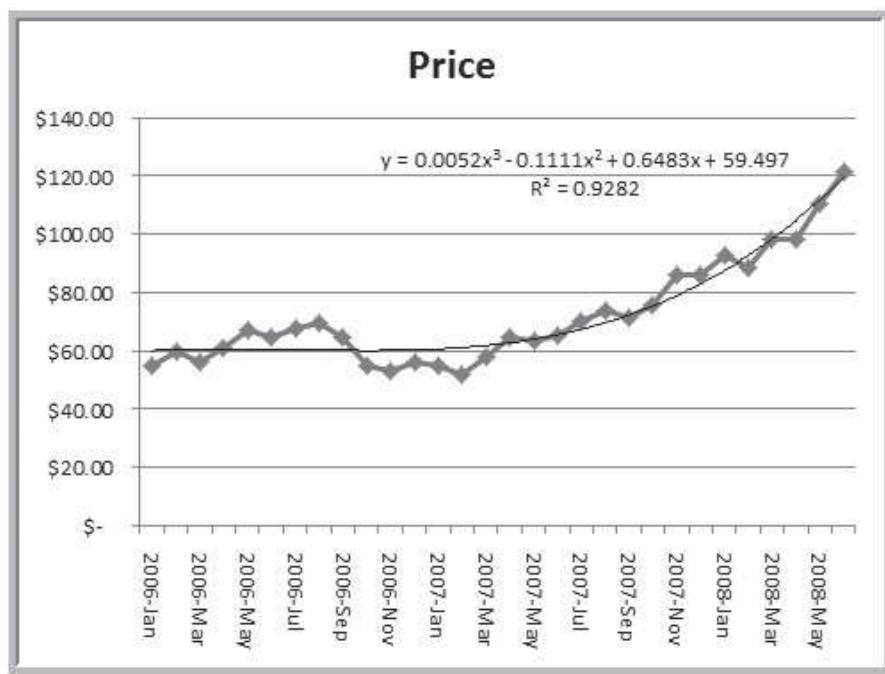
$$\text{Polynomial (second order): } y = 0.121x^2 + 1.232x + 53.48 \quad R^2 = 0.968$$

$$\text{Linear: } y = 3.548x + 45.76 \quad R^2 = 0.944$$

The difference in prediction can be significant. For example, to predict the price six months after the last data point ( $x = 36$ ) yields \$172.24 for the third-order polynomial



**FIGURE 9.8** Chart of Crude Oil Prices



**FIGURE 9.9** Polynomial Fit of Crude Oil Prices

fit with all the data, and \$246.45 for the exponential model with only the recent data. Thus, the analysis must be careful to select the proper amount of data for the analysis. The question now becomes one of choosing the best assumptions for the model. Is it reasonable to assume that prices would increase exponentially or perhaps at a slower rate, such as with the linear model fit? Or, would they level off and start falling? Clearly, factors other than historical trends would enter into this choice. As we now know, oil prices plunged in the latter half of 2008; thus, all predictive models are risky.

Data fitting is often combined with logical approaches in model building. For example, we might use regression to express some variable as a function of others based on empirical data, and then incorporate the regression equation into a larger model. The following example illustrates this.

**REVENUE MODEL** Suppose that a firm wishes to determine the best pricing for one of its products to maximize revenue over the next year. A market research study has collected data that estimate the expected annual sales for different levels of pricing as shown in Figure 9.10. Plotting these data on a scatter chart suggests that sales and price have a linear relationship. Adding a linear trendline confirms a strong correlation between the variables. Thus, the linear model,  $\text{Sales} = -2.9485 \times \text{Price} + 3,240.9$  is a good predictor of sales for any pricing decision between \$500 and \$1,000. Because revenue equals  $\text{price} \times \text{sales}$ , a model for total revenue is:

$$\begin{aligned}\text{Total revenue} &= \text{Price} \times \text{Sales} \\ &= \text{Price} \times (-2.9485 \text{ Price} + 3,240.9) \\ &= -2.9485 \text{ Price}^2 + 3,240.9 \times \text{Price}\end{aligned}$$

This revenue function is a second-order polynomial (quadratic).



**FIGURE 9.10** Price-Sales Data and Trendline

### SKILL-BUILDER EXERCISE 9.5

Implement the total revenue model on a spreadsheet and use *Solver* to find the price that maximizes total revenue.

### Spreadsheet Engineering

First and foremost, spreadsheets should be accurate. Spreadsheet errors can be disastrous. A large investment company once made a \$2.6 billion error. They notified holders of one mutual fund to expect a large dividend; fortunately, they caught the error before sending the checks. Industry surveys estimate that more than 90% of spreadsheets with more than 150 rows were incorrect by at least 5%. **Verification** is the process of ensuring that a model is accurate and free from logical errors. So how we do this? There are three basic approaches that can help:

**1. Improve the design and format of the spreadsheet itself.** After the inputs, outputs, and key relationships are well understood, you should sketch a logical design of the spreadsheet. For example, you might want the spreadsheet to resemble a financial statement to make it easier for managers to read. It is good practice to separate the model inputs from the model itself and to reference the input cells in the model formulas; that way, any changes in the inputs will be automatically reflected in the model. We have done this in the examples.

Another useful approach is to break complex formulas into smaller pieces. This reduces typographical errors, makes it easier to check your results, and also makes the spreadsheet easier to read for the user. For example, we could write one long formula for profit using the equation we developed earlier in this section. However, setting up the model in a form similar to a financial statement and breaking out the revenue and costs makes the calculations much clearer, especially to a manager who will be using the model.

**2. Improve the process used to develop a spreadsheet.** If you sketched out a conceptual design of the spreadsheet, work on each part individually before moving on

to the others to ensure that it is correct. As you enter formulas, check the results with simple numbers (such as 1) to determine if they make sense, or use inputs with known results. Be careful in using the *Copy* and *Paste* commands in Excel, particularly with respect to relative and absolute addresses. Use the Excel function wizard (the  $f_x$  button on the formula bar) to ensure that you are entering the correct values in the correct fields of the function. Use cell and range names to simplify formulas and make them more user-friendly. For example, suppose that the unit price is stored in cell B13 and quantity sold in cell B14. Suppose you wish to calculate revenue in cell C15. Instead of writing the formula  $=B13*B14$ , you could define the name of cell B13 in Excel as “UnitPrice” and the name of cell B14 as “QuantitySold.” Then in cell C15, you could simply write the formula  $=UnitPrice*QuantitySold$ . (In this book, however, we will use cell references so that you can more easily trace formulas in the examples.)

### 3. Inspect your results carefully and use appropriate tools available in Excel.

The *Data Validation* tool can signal errors if an input value is not the right type. For example, it does not make sense to input a Quantity Produced that is not a whole number. With this tool, you may define validation criteria for model inputs and pop up an error message if the wrong values or type of data are entered. The Excel *Auditing* tool also helps you to validate the logic of formulas by visually showing the relationships between input data and cell formulas. We encourage you to learn how to use these tools.

#### SKILL-BUILDER EXERCISE 9.6

Implement the model we developed earlier in this chapter,  $\text{Profit} = (\text{Unit price})[\text{Min}(\text{Quantity produced}, \text{Demand})] - [\text{Fixed cost} + (\text{Unit cost})(\text{Quantity produced})]$ , on a spreadsheet using sound spreadsheet design principles. Find the profit for the following values: unit price = \$40, unit cost = \$24, fixed cost = \$400,000, demand = 50,000, and quantity produced = 40,000.

## SPREADSHEET MODELING EXAMPLES

In this section, we present a few examples of decision models that are implemented on spreadsheets. Some of these will also be used in the next chapter when we discuss models that incorporate random variables.

### New Product Development

Many applications of spreadsheet models in business involve basic financial analysis. One example is the decision to launch a new product. In the pharmaceutical industry, for example, the process of research and development is a long and arduous process; total development expenses can approach one billion dollars. Suppose that Moore Pharmaceuticals has discovered a potential drug breakthrough in the laboratory and needs to decide whether to go forward to conduct clinical trials and seek FDA approval to market the drug. Total R&D costs will be expected to reach \$700 million, and the cost of clinical trials will be about \$150 million. The current market size is estimated to be two million people and is expected to grow at a rate of 3% each year. In the first year, Moore estimates gaining an 8% market share, which is anticipated to grow by 20% each year. It is difficult to estimate beyond five years as new competitors are expected to be entering the market. A monthly prescription is anticipated to generate a revenue of \$130 while incurring variable costs of \$40. A discount rate of 9% is assumed. The company needs to know how long it will take to recover its fixed expenses and the NPV over the first five years.

Figure 9.11 shows a spreadsheet model for this situation (Excel file *Moore Pharmaceuticals*). The model is based on a variety of known data, estimates, and assumptions.

	A	B	C	D	E	F
1	<b>Moore Pharmaceuticals</b>					
2						
3	<b>Data</b>					
4						
5	Market size	2,000,000				
6	Unit (monthly Rx) revenue	\$ 130.00				
7	Unit (monthly Rx) cost	\$ 40.00				
8	Discount rate	9%				
9						
10	<b>Project Costs</b>					
11	R&D	\$ 700,000,000				
12	Clinical Trials	\$ 150,000,000				
13	<b>Total Project Costs</b>	\$ 850,000,000				
14						
15	<b>Model</b>					
16						
17	Year	1	2	3	4	5
18	Market growth factor		3.00%	3.00%	3.00%	3.00%
19	Market size	2,000,000	2,060,000	2,121,800	2,185,454	2,251,018
20	Market share growth rate		20.00%	20.00%	20.00%	20.00%
21	Market share	8.00%	9.60%	11.52%	13.82%	16.59%
22	Sales	160,000	197,760	244,431	302,117	373,417
23						
24	Annual Revenue	\$ 249,600,000	\$ 308,505,600	\$ 381,312,922	\$ 471,302,771	\$ 582,530,225
25	Annual Costs	\$ 76,800,000	\$ 94,924,800	\$ 117,327,053	\$ 145,016,237	\$ 179,240,069
26	Profit	\$ 172,800,000	\$ 213,580,800	\$ 263,985,869	\$ 326,286,534	\$ 403,290,156
27	Cumulative Net Profit	\$ (677,200,000)	\$ (463,619,200)	\$ (199,633,331)	\$ 126,653,203	\$ 529,943,358
28						
29	<b>Net Present Value</b>	\$ 185,404,860				

	A	B	C	D	E	F
1	<b>Moore Pharmaceuticals</b>					
2						
3	<b>Data</b>					
4						
5	Market size	2000000				
6	Unit (monthly Rx) revenue	130				
7	Unit (monthly Rx) cost	40				
8	Discount rate	0.09				
9						
10	<b>Project Costs</b>					
11	R&D	700000000				
12	Clinical Trials	150000000				
13	<b>Total Project Costs</b>	=B11+B12				
14						
15	<b>Model</b>					
16						
17	Year	1	2	3	4	5
18	Market growth factor		0.03	0.03	0.03	0.03
19	Market size	=B5	=B19*(1+C18)	=C19*(1+D18)	=D19*(1+E18)	=E19*(1+F18)
20	Market share growth rate		0.2	0.2	0.2	0.2
21	Market share	0.08	=B21*(1+C20)	=C21*(1+D20)	=D21*(1+E20)	=E21*(1+F20)
22	Sales	=B19*B21	=C19*C21	=D19*D21	=E19*E21	=F19*F21
23						
24	Annual Revenue	=B22*\$B\$6*12	=C22*\$B\$6*12	=D22*\$B\$6*12	=E22*\$B\$6*12	=F22*\$B\$6*12
25	Annual Costs	=B22*\$B\$7*12	=C22*\$B\$7*12	=D22*\$B\$7*12	=E22*\$B\$7*12	=F22*\$B\$7*12
26	Profit	=B24-B25	=C24-C25	=D24-D25	=E24-E25	=F24-F25
27	Cumulative Net Profit	=B26-B13	=B27+C26	=C27+D26	=D27+E26	=E27+F26
28						
29	<b>Net Present Value</b>	=NPV(B8,B26:F26)-B13				

**FIGURE 9.11** Moore Pharmaceuticals Spreadsheet Model

If you examine the model closely, you will see that some of the inputs in the model are easily obtained from corporate accounting (e.g., discount rate, unit revenue, and cost), based on historical data (e.g., project costs), forecasts or judgmental estimates based on preliminary market research or previous experience (e.g., market size, market share, and yearly growth rates). The model itself is a straightforward application of accounting and financial logic; you should examine the Excel formulas to see how the model is built.

The assumptions used represent the “most likely” estimates, and the spreadsheet shows that the product will begin to be profitable by the fourth year. However, the model is based on some rather tenuous assumptions about the market size and market share growth rates. In reality, much of the data used in the model are uncertain, and the corporation would be remiss if it simply used the results of this one scenario. The real value of the model would be in analyzing a variety of scenarios that use ranges of these assumptions.

### SKILL-BUILDER EXERCISE 9.7

Use the *Goal Seek* tool to determine the value of the total fixed project costs that would result in a NPV of zero in the *Moore Pharmaceuticals* spreadsheet model.

### Single Period Purchase Decisions

Banana Republic, a division of Gap, Inc., was trying to build a name for itself in fashion circles as parent Gap shifted its product line to basics such as cropped pants, jeans, and khakis. In one recent holiday season, the company had bet that blue would be the top-selling color in stretch merino wool sweaters. They were wrong; as the company president noted, “The number 1 seller was moss green. We didn’t have enough.”<sup>1</sup>

This situation describes one of many practical situations in which a one-time purchase decision must be made in the face of uncertain demand. Department store buyers must purchase seasonal clothing well in advance of the buying season, and candy shops must decide on how many special holiday gift boxes to assemble. The general scenario is commonly known as the *newsvendor problem*: A street newsvendor sells daily newspapers and must make a decision of how many to purchase. Purchasing too few results in lost opportunity to increase profits, but purchasing too many will result in a loss since the excess must be discarded at the end of the day.

We will first develop a general model for this problem, then illustrate it with an example. Let us assume that each item costs \$C to purchase and is sold for \$R. At the end of the period, any unsold items can be disposed of at \$S each (the salvage value). Clearly, it makes sense to assume that  $R > C > S$ . Let  $D$  be the number of units demanded during the period and  $Q$  be the quantity purchased. Note that  $D$  is an uncontrollable input while  $Q$  is a decision variable. If demand is known, then the optimal decision is obvious: Choose  $Q = D$ . However, if  $D$  is not known in advance, we run the risk of over-purchasing or under-purchasing. If  $Q < D$ , then we lose the opportunity of realizing additional profit (since we assume that  $R > C$ ), and if  $Q > D$ , we incur a loss (because  $C > S$ ).

Notice that we cannot sell more than the minimum of the actual demand and the amount produced. Thus, the quantity sold at the regular price is the smaller of  $D$  and  $Q$ . Also, the surplus quantity is the larger of 0 and  $Q - D$ . The net profit is calculated as:

$$\text{Net profit} = R \times \text{Quantity Sold} + S \times \text{Surplus Quantity} - C \times Q$$

<sup>1</sup> Louise Lee, “Yes, We Have a New Banana,” *BusinessWeek*, May 31, 2004, 70–72.

To illustrate this model, let us suppose that a small candy store makes Valentine's Day gift boxes for \$12.00 and sells them for \$18.00. In the past, at least 40 boxes have been sold by Valentine's Day, but the actual amount is uncertain, and in the past, the owner has often run short or made too many. After the holiday, any unsold boxes are discounted 50% and are eventually sold. Figure 9.12 shows a spreadsheet model that computes the profit for any input values of the purchase quantity and demand (Excel file *Newsvendor Model*). An Excel data table can be used to evaluate the profit for various combinations of the inputs. This table shows, for example, that higher purchase quantities have the potential for higher profits, but are also at risk for yielding lower profits than smaller purchase quantities that provide safer and more stable profit expectations. In reality, demand would be a random variable that would best be modeled using a probability distribution. We will see how to do this in the next chapter.

### Overbooking Decisions

An important operations decision for service businesses such as hotels, airlines, and car rental companies is the number of reservations to accept to effectively fill capacity with the knowledge that some customers may not use their reservations nor tell the business. If a hotel, for example, holds rooms for customers who do not show up, they lose revenue opportunities. (Even if they charge a night's lodging as a guarantee, rooms held for additional days may go unused.) A common practice in these industries is to overbook reservations. When more

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Newsvendor Model		Demand		Purchase Quantity									
2			\$ 237.00		40	41	42	43	44	45	46	47	48	49
3	Data			40	\$ 240.00	\$ 237.00	\$ 234.00	\$ 231.00	\$ 228.00	\$ 225.00	\$ 222.00	\$ 219.00	\$ 216.00	\$ 213.00
4				41	\$ 240.00	\$ 246.00	\$ 243.00	\$ 240.00	\$ 237.00	\$ 234.00	\$ 231.00	\$ 228.00	\$ 225.00	\$ 222.00
5	Selling price	\$ 18.00		42	\$ 240.00	\$ 246.00	\$ 252.00	\$ 249.00	\$ 246.00	\$ 243.00	\$ 240.00	\$ 237.00	\$ 234.00	\$ 231.00
6	Cost	\$ 12.00		43	\$ 240.00	\$ 246.00	\$ 252.00	\$ 258.00	\$ 255.00	\$ 252.00	\$ 249.00	\$ 246.00	\$ 243.00	\$ 240.00
7	Discount price	\$ 9.00		44	\$ 240.00	\$ 246.00	\$ 252.00	\$ 258.00	\$ 264.00	\$ 261.00	\$ 258.00	\$ 255.00	\$ 252.00	\$ 249.00
8				45	\$ 240.00	\$ 246.00	\$ 252.00	\$ 258.00	\$ 264.00	\$ 270.00	\$ 267.00	\$ 264.00	\$ 261.00	\$ 258.00
9	Model			46	\$ 240.00	\$ 246.00	\$ 252.00	\$ 258.00	\$ 264.00	\$ 270.00	\$ 276.00	\$ 273.00	\$ 270.00	\$ 267.00
10				47	\$ 240.00	\$ 246.00	\$ 252.00	\$ 258.00	\$ 264.00	\$ 270.00	\$ 276.00	\$ 282.00	\$ 279.00	\$ 276.00
11	Demand	41		48	\$ 240.00	\$ 246.00	\$ 252.00	\$ 258.00	\$ 264.00	\$ 270.00	\$ 276.00	\$ 282.00	\$ 288.00	\$ 285.00
12	Purchase Quantity	44		49	\$ 240.00	\$ 246.00	\$ 252.00	\$ 258.00	\$ 264.00	\$ 270.00	\$ 276.00	\$ 282.00	\$ 288.00	\$ 294.00
13														
14	Quantity Sold	41												
15	Surplus Quantity	3												
16														
17	Profit	\$ 237.00												

A	B	
1	Newsvendor Model	
2		
3	Data	
4		
5	Selling price	18
6	Cost	12
7	Discount price	9
8		
9	Model	
10		
11	Demand	41
12	Purchase Quantity	44
13		
14	Quantity Sold	=MIN(B11,B12)
15	Surplus Quantity	=MAX(0,B12-B11)
16		
17	Profit	=B14*B5+B15*B7-B12*B6

FIGURE 9.12 Newsvendor Model

customers arrive than can be handled, the business usually incurs some cost to satisfy them (by putting them up at another hotel, or for most airlines, providing extra compensation such as ticket vouchers). Therefore, the decision becomes how much to overbook to balance the costs of overbooking against the lost revenue for underuse.

We will illustrate how a decision model can help in making this decision. Figure 9.13 shows a spreadsheet model (Excel file *Hotel Overbooking Model*) for a popular resort hotel that has 300 rooms and is usually fully booked. The hotel charges \$120 per room. Reservations may be cancelled by the 6:00 p.m. deadline with no penalty. The hotel has estimated that the average overbooking cost is \$100.

The logic of the model is straightforward. In the model section of the spreadsheet, cell B12 represents the decision variable of how many reservations to accept. Cell B13 represents the actual customer demand (the number of customers who want a reservation). The hotel cannot accept more reservations than its predetermined limit, so the number of reservations made in cell B13, therefore, is the smaller of the customer demand and the reservation limit. Cell B15 is the number of customers who decide to cancel their reservation. Therefore, the actual number of customers who arrive (cell B16) is the difference between the number of reservations made and the number of cancellations. If the actual number of customer arrivals exceeds the room capacity, overbooking occurs. This is modeled by the MAX function in cell B16. Net revenue is computed in cell B18. Using data tables and the *Scenario Manager*, we could easily analyze how the number of overbooked customers and net revenue would be influenced by changes in the reservation limit, customer demand, and cancellations.

As with the newsvendor model, the customer demand and the number of cancellations are in reality, random variables. We will show how to incorporate randomness into the model in the next chapter also.

## Project Management

Project management is concerned with scheduling the interrelated activities of a project. An important aspect of project management is identifying the expected completion time of the project. Activity times can be deterministic or probabilistic. We will address probabilistic times in the next chapter; for the purposes of model development, we will assume that activity times are constant.

	A	B
1	Hotel Overbooking Model	
2		
3	Data	
4		
5	Rooms available	300
6	Price	\$120
7	Overbooking cost	\$100
8		
9	Model	
10		
11	Reservation limit	300
12	Customer demand	290
13	Reservations made	290
14	Cancellations	15
15	Customer arrivals	275
16	Overbooked customers	0
17		
18	Net revenue	\$33,000

	A	B
1	Hotel Overbooking Model	
2		
3	Data	
4		
5	Rooms available	300
6	Price	120
7	Overbooking cost	100
8		
9	Model	
10		
11	Reservation limit	300
12	Customer demand	290
13	Reservations made	=MIN(B11,B12)
14	Cancellations	15
15	Customer arrivals	=B13-B14
16	Overbooked customers	=MAX(0,B15-B5)
17		
18	Net revenue	=MIN(B15,B5)*B6-B16*B7

FIGURE 9.13 Hotel Overbooking Model Spreadsheet

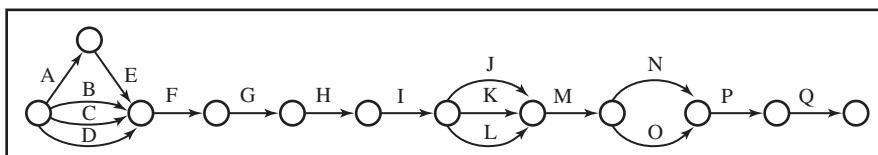
**TABLE 9.1** Activity and Time Estimate List

Activity	Predecessors	Activity Time
A	—	15
B	—	50
C	—	20
D	—	3
E	A	7
F	B,C,D,E	6
G	F	1
H	G	36
I	H	6
J	I	6
K	I	32
L	I	4
M	J,K,L	3
N	M	15
O	M	18
P	N,O	2
Q	P	5

As an example, Becker Consulting has been hired to assist in the evaluation of new software. The manager of the Information Systems department is responsible for coordinating all of the activities involving consultants and the company's resources. The activities shown in Table 9.1 have been defined for this project, which is depicted graphically in Figure 9.14.

Figure 9.15 shows a spreadsheet designed to calculate the project completion time (Excel file *Becker Consulting Project Management Model*). The model uses Excel MAX, MIN, and IF functions to implement the logic of calculating the project schedule and critical path. The project completion time depends on the specific time for each activity. Activities A, B, C, and D have no immediate predecessors and, therefore, have early start times of 0 (cells C5:C8). The early start time for each other activity is the maximum of the early finish times for the activity's immediate predecessor. Early finish times are computed as the early start time plus the activity time. The early finish time for the last activity, Q (cell D21, copied to cell D23), represents the earliest time the project can be completed, that is, the minimum project completion time.

To compute late start and late finish times, we set the late finish time of the terminal activity equal to the project completion time. The late start time is computed by subtracting the activity time from the late finish time. The late finish time for any other activity, say X, is defined as the minimum late start of all activities to which activity X is an immediate predecessor. Slack is computed as the difference between the late finish and

**FIGURE 9.14** Project Network Structure

	A	B	C	D	E	F	G	H
1	Becker Consulting Project Management Model							
2								
3		Activity	Early Start	Early Finish	Latest Start	Latest Finish		On Critical Path?
4	Activity	Time					Slack	
5	A	15.00	0.00	15.00	28.00	43.00	28.00	0
6	B	50.00	0.00	50.00	0.00	50.00	0.00	1
7	C	20.00	0.00	20.00	30.00	50.00	30.00	0
8	D	3.00	0.00	3.00	47.00	50.00	47.00	0
9	E	7.00	15.00	22.00	43.00	50.00	28.00	0
10	F	6.00	50.00	56.00	50.00	56.00	0.00	1
11	G	1.00	56.00	57.00	56.00	57.00	0.00	1
12	H	36.00	57.00	93.00	57.00	93.00	0.00	1
13	I	6.00	93.00	99.00	93.00	99.00	0.00	1
14	J	6.00	99.00	105.00	125.00	131.00	26.00	0
15	K	32.00	99.00	131.00	99.00	131.00	0.00	1
16	L	4.00	99.00	103.00	127.00	131.00	28.00	0
17	M	3.00	131.00	134.00	131.00	134.00	0.00	1
18	N	15.00	134.00	149.00	137.00	152.00	3.00	0
19	O	18.00	134.00	152.00	134.00	152.00	0.00	1
20	P	2.00	152.00	154.00	152.00	154.00	0.00	1
21	Q	5.00	154.00	159.00	154.00	159.00	0.00	1
22								
23			Project completion time	159.00				

	A	B	C	D	E	F	G	H
1	Decker Consulting							
2								
3		Activity	Early Start	Early Finish	Latest Start	Latest Finish		On Critical Path?
4	Activity	Time					Slack	
5	A	15	0	=C5+B5	=F5-B5	=E9	=F5-D5	=IF(G5<0.0001,1,0)
6	B	50	0	=C6+B6	=F6-B6	=E10	=F6-D6	=IF(G6<0.0001,1,0)
7	C	20	0	=C7+R7	=F7-R7	=F10	=F7-D7	=IF(G7<0.0001,1,0)
8	D	3	0	=C8+R8	=F8-R8	=F10	=F8-D8	=IF(G8<0.0001,1,0)
9	E	7	=D5	=C9+B9	=F9-B9	=E10	=F9-D9	=IF(G9<0.0001,1,0)
10	F	6	=MAX(D6,D7,D8,D9)	=C10+B10	=F10-B10	=E11	=F10-D10	=IF(G10<0.0001,1,0)
11	G	1	=D10	=C11+B11	=F11-B11	=E12	=F11-D11	=IF(G11<0.0001,1,0)
12	H	36	=D11	=C12+B12	=F12-B12	=E13	=F12-D12	=IF(G12<0.0001,1,0)
13	I	6	=D12	=C13+B13	=F13-B13	=MIN(E14,E15,E16)	=F13-D13	=IF(G13<0.0001,1,0)
14	J	6	=D13	=C14+B14	=F14-B14	=E17	=F14-D14	=IF(G14<0.0001,1,0)
15	K	32	=D13	=C15+B15	=F15-B15	=E17	=F15-D15	=IF(G15<0.0001,1,0)
16	L	4	=D13	=C16+B16	=F16-B16	=E17	=F16-D16	=IF(G16<0.0001,1,0)
17	M	3	=MAX(D14,D15,D16)	=C17+B17	=F17-B17	=MIN(E18,E19)	=F17-D17	=IF(G17<0.0001,1,0)
18	N	15	=D17	=C18+B18	=F18-B18	=E20	=F18-D18	=IF(G18<0.0001,1,0)
19	O	18	=U1/	=C19+B19	=F19-B19	=E20	=F19-D19	=IF(G19<0.0001,1,0)
20	P	2	=MAX(D18,D19)	=C20+B20	=F20-B20	=E21	=F20-D20	=IF(G20<0.0001,1,0)
21	Q	5	=D20	=C21+B21	=F21-B21	=D21	=F21-D21	=IF(G21<0.0001,1,0)
22			Project completion time		-D21			
23								

**FIGURE 9.15** Becker Consulting Project Management Spreadsheet

early finish. The critical path consists of activities with zero slack. Based on the expected activity times, the critical path consists of activities B-F-G-H-I-K-M-O-P-Q and has an expected duration of 159 days.

## MODEL ASSUMPTIONS, COMPLEXITY, AND REALISM

Models cannot capture every detail of the real problem, and managers must understand the limitations of models and their underlying assumptions. **Validity** refers to how well a model represents reality. One approach for judging the validity of a model is to identify and examine the assumptions made in a model to see how they agree with

one's perception of the real world; the closer the agreement, the higher the validity. A "perfect" model corresponds to the real world in every respect; unfortunately, no such model has ever existed, and never will exist in the future, because it is impossible to include every detail of real life in one model. To add more realism to a model generally requires more complexity and analysts have to know how to balance these.

To illustrate this, consider modeling a typical retirement plan. Suppose that an employee starts working after completing her MBA at age 25 at a starting salary of \$50,000. She expects an average salary increase of 4% each year. Her retirement plan requires that she contribute 8% of her salary, and her employer adds an additional 35% of her contribution. She anticipates an annual return of 8% on her retirement portfolio.

Figure 9.16 shows a simple spreadsheet model of her retirement investments through age 50. There are two validity issues with this model. One, of course, is whether

	A	B	C	D	E
1	<b>Retirement Plan Model</b>				
2					
3	<b>Data</b>				
4	Retirement contribution (% of salary)		8%		
5	Employer match		35%		
6	Annual salary increase		4%		
7	Annual return on investment		8%		
8					
9	<b>Model</b>			Employee	Employer
10		Age	Salary	Contribution	Contribution
11		25	\$50,000	\$4,000	\$1,400
12		26	\$ 52,000	\$4,160	\$1,456
13		27	\$ 54,080	\$4,326	\$1,514
14		28	\$ 56,243	\$4,499	\$1,575
15		29	\$ 58,493	\$4,679	\$1,638
16		30	\$ 60,833	\$4,867	\$1,703
17		31	\$ 63,266	\$5,061	\$1,771
18		32	\$ 65,797	\$5,264	\$1,842
19		33	\$ 68,428	\$5,474	\$1,916
20		34	\$ 71,166	\$5,693	\$1,993
21		35	\$ 74,012	\$5,921	\$2,072
22		36	\$ 76,973	\$6,158	\$2,155
23		37	\$ 80,052	\$6,404	\$2,241
24		38	\$ 83,254	\$6,660	\$2,331
25		39	\$ 86,584	\$6,927	\$2,424
26		40	\$ 90,047	\$7,204	\$2,521
27		41	\$ 93,649	\$7,492	\$2,622
28		42	\$ 97,395	\$7,792	\$2,727
29		43	\$ 101,291	\$8,103	\$2,836
30		44	\$ 105,342	\$8,427	\$2,950
31		45	\$ 109,556	\$8,764	\$3,068
32		46	\$ 113,938	\$9,115	\$3,190
33		47	\$ 118,496	\$9,480	\$3,318
34		48	\$ 123,236	\$9,859	\$3,451
35		49	\$ 128,165	\$10,253	\$3,589
36		50	\$ 133,292	\$10,663	\$3,732
					\$624,224

**FIGURE 9.16 Retirement Plan Model Spreadsheet**

	A	B	C	D	E
1	Retirement Plan Model				
2					
3	Data				
4	Retirement contribution (% of salary)	0.08			
5	Employer match	0.35			
6	Annual salary increase	0.04			
7	Annual return on investment	0.08			
8					
9	Model		Employee	Employer	
10		Age	Salary	Contribution	Contribution
11	25		50000	=B11*\$B\$4	=\$B\$5*C11
12			=B11*(1+\$B\$6)	=B12*\$B\$4	=\$B\$5*C12
13	26			=B12*(1+\$B\$6)	=E11*(1+\$B\$7) + C12+D12
14	27			=B13*\$B\$4	=\$B\$5*C13
15	28			=B13*(1+\$B\$6)	=E12*(1+\$B\$7) + C13+D13
16				=B14*\$B\$4	=\$B\$5*C14
17					=E13*(1+\$B\$7) + C14+D14

**FIGURE 9.16** (Continued)

the assumptions of the annual salary increase and return on investment are reasonable and whether they should be assumed to be the same each year. Assume the same rate of salary increases and investment returns each year simplifies the model, but detracts from the realism because these variables will clearly vary each year. A second validity issue is how the model calculates the return on investment. The model in Figure 9.16 assumes that the return on investment is applied to the previous year's balance, and not to the current year's contributions (examine the formula used in cell E12). An alternative would be to calculate the investment return based on the end of year balance, including current year contributions, using the formula  $=(E11+C12+D12)*(1+\$B$7)$  in cell E12, and copying it down the spreadsheet. This results in a balance of \$671,204 at age 50.

Neither of these assumptions are quite correct, since the contributions would normally be made on a monthly basis. To reflect this would require a larger and more complex spreadsheet model. Thus, building realistic models requires careful thought and creativity, and a good working knowledge of the capabilities of Excel.

### SKILL-BUILDER EXERCISE 9.8

Modify the *Retirement Plan Model* spreadsheet to spread out the contributions and investment returns on a monthly basis, and compare the balance at age 50 to the assumptions we used in the example.

## Basic Concepts Review Questions

- What is a model? When is a model called a decision model?
- What are optimization models?
- Describe how to use Excel data tables, *Scenario Manager*, and goal seek tools to analyze decision models.
- Explain the purpose of *Solver* and what type of decision model it is used for.
- How does model analysis help managers make a decision?
- Summarize the important knowledge that you need to successfully build good decision models.
- What is meant by verification? What is spreadsheet engineering?
- What is meant by *What-if Analysis* and *Sensitivity Analysis*?
- Discuss the inherent conflict between the realism and complexity of a model. Explain the meaning of the term validity in this connection.

## Problems and Applications

1. A supermarket has been experiencing long lines during peak periods of the day. The problem is noticeably worse on certain days of the week, and the peak periods are sometimes different according to the day of the week. There are usually enough workers on the job to open all cash registers. The problem is knowing when to call some of the workers stocking shelves up to the front to work the checkout counters. How might decision models help the supermarket? What data would be needed to develop these models?
2. Four key marketing decision variables are price ( $P$ ), advertising ( $A$ ), transportation ( $T$ ), and product quality ( $Q$ ). Consumer demand ( $D$ ) is influenced by these variables. The simplest model for describing demand in terms of these variables is:

$$D = k - pP + aA + tT + qQ$$

where  $k$ ,  $p$ ,  $a$ ,  $t$ , and  $q$  are constants. Discuss the assumptions of this model. Specifically, how does each variable affect demand? How do the variables influence each other? What limitations might this model have? How can it be improved?

3. *Total marketing effort* is a term used to describe the critical decision factors that affect demand: price, advertising, distribution, and product quality. Define the variable  $x$  to represent total marketing effort. A typical model that is used to predict demand as a function of total marketing effort is based on the power function:

$$D = ax^b$$

Suppose that  $a$  is a positive number. Different model forms result from varying the constant  $b$ . Sketch the graphs of this model for  $b = 0$ ,  $b = 1$ ,  $0 < b < 1$ ,  $b < 0$ , and  $b > 1$ . (We encourage you to use Excel to do this.) What does each model tell you about the relationship between demand and marketing effort? What assumptions are implied? Are they reasonable? How would you go about selecting the appropriate model?

4. A manufacturer is preparing to set the price on a new action game. Demand is thought to depend on the price and is represented by the model:

$$D = 2,000 - 3.5P$$

The accounting department estimates that the total costs can be represented by:

$$C = 5,000 + 4.1D$$

- a. Develop a model for the total profit and implement it on a spreadsheet.

- b. Develop a one-way data table to evaluate profit as a function of price (choose a price range that is reasonable and appropriate).
  - c. Use *Solver* to find the price that maximizes profit.
5. The Radio Shop sells two popular models of portable sport radios: model  $A$  and model  $B$ . The sales of these products are not independent of each other (in economics, we call these substitutable products, because if the price of one increases, sales of the other will increase). The store wishes to establish a pricing policy to maximize revenue from these products. A study of price and sales data shows the following relationships between the quantity sold ( $N$ ) and prices ( $P$ ) of each model:

$$N_A = 20 - 0.62P_A + 0.30P_B$$

$$N_B = 29 + 0.10P_A - 0.60P_B$$

- a. Construct a model for the total revenue and implement it on a spreadsheet.
- b. Develop a two-way data table to estimate the optimal prices for each product in order to maximize the total revenue.
- c. Use *Solver* to find the optimal prices.
6. A forest fire is burning down a narrow valley three miles wide at a speed of 40 feet per minute. The fire can be contained by cutting a firebreak through the forest across the valley. It takes 30 seconds for one person to clear one foot of the firebreak. The value of lost timber is \$4,000 per square mile. Each person hired is paid \$18 per hour, and it costs \$50 to transport and supply each person with the appropriate equipment.
  - a. Develop a model for determining how many people should be sent to contain the fire and for determining the best location for the firebreak (draw a picture first!).
  - b. Implement your model on a spreadsheet and find the optimal solution using *Solver*.
7. Each worksheet in the Excel file *LineFit Data* contains a set of data that describes a functional relationship between the dependent variable  $y$  and the independent variable  $x$ . Construct a line chart of each data set, and use the *Add Trendline* tool to determine the best-fitting functions to model these data sets.
8. Develop a spreadsheet model to determine how much a person or a couple can afford to spend on a house.<sup>2</sup> Lender guidelines suggest that the allowable monthly housing expenditure should be no more than 28% of monthly gross income. From this, you must subtract total nonmortgage housing expenses, which would include insurance and property taxes, and any other additional

<sup>2</sup> Adapted from Ralph R. Frasca, *Personal Finance*, 8<sup>th</sup> Edition, Boston: Prentice-Hall, 2009

expenses. This defines the affordable monthly mortgage payment. In addition, guidelines also suggest that total affordable monthly debt payments, including housing expenses, should not exceed 36% of gross monthly income. This is calculated by subtracting total nonmortgage housing expenses and any other installment debt, such as car loans, student loans, credit card debt, and so on, from 36% of total monthly gross income. The smaller of the affordable monthly mortgage payment and the total affordable monthly debt payments is the affordable monthly mortgage. To calculate the maximum that can be borrowed, find the monthly payment per \$1,000 mortgage based on the current interest rate and duration of the loan. Divide the affordable monthly mortgage amount by this monthly payment to find the affordable mortgage. Assuming a 20% downpayment, the maximum price of a house would be the affordable mortgage divided by 0.8.

Use the following data to test your model: total monthly gross income = \$6,500; nonmortgage housing expenses = \$350; monthly installment debt = \$500; monthly payment per \$1,000 mortgage = \$7.258.

9. Monthly rent at an apartment complex is \$500. Operating costs average \$15,000 per month regardless of the number of units rented. Construct a spreadsheet model to determine the profit if 35 units are rented. The manager has observed that the number of units rented during any given month varies between 30 and 40. Use your model to evaluate the profit for this range of unit rentals.
10. Think of any retailer that operates many stores throughout the country, such as Old Navy, Hallmark Cards, or Radio Shack, to name just a few. The retailer is often seeking to open new stores and needs to evaluate the profitability of a proposed location that would be leased for five years. An Excel model is provided in the *New Store Financial Model* spreadsheet. Use *Scenario Manager* to evaluate the cumulative discounted cash flow for the fifth year under the following scenarios:

	<b>Scenario 1</b>	<b>Scenario 2</b>	<b>Scenario 3</b>
Inflation Rate	1%	5%	3%
Cost of Merchandise (% of sales)	25%	30%	26%
Labor Cost	\$150,000	\$225,000	\$200,000
Other Expenses	\$300,000	\$350,000	\$325,000
First Year Sales Revenue	\$600,000	\$600,000	\$800,000
Sales growth year 2	15%	22%	25%
Sales growth year 3	10%	15%	18%

(Continued on next column)

Sales growth year 4	6%	11%	14%
Sales growth year 5	3%	5%	8%

11. A garage band wants to hold a concert. The expected crowd is 3,000. The average expenditure on concessions is \$15. Tickets sell for \$10 each, and the band's profit is 80% of the gate, along with concession sales, minus a fixed cost of \$10,000. Develop a spreadsheet model to find their expected profit. Define and run some reasonable scenarios using the *Scenario Manager* to evaluate profitability for variations in the estimates.
12. For a new product, sales volume in the first year is estimated to be 100,000 units and is projected to grow at a rate of 7% per year. The selling price is \$10, and will increase by \$0.50 each year. Per-unit variable costs are \$3, and annual fixed costs are \$200,000. Per-unit costs are expected to increase 5% per year. Fixed costs are expected to increase 10% per year. Develop a spreadsheet model to calculate the present value of profit over a three-year period, assuming a 7% discount rate.
13. MasterTech is a new software company that develops and markets productivity software for municipal government applications. In developing their income statement, the following formulas are used:

$$\text{Gross profit} = \text{Net sales} - \text{Cost of sales}$$

$$\text{Net operating profit} = \text{Gross profit} - \text{Administrative expenses} - \text{Selling expenses}$$

$$\text{Net income before taxes} = \text{Net operating profit} - \text{Interest expense}$$

$$\text{Net income} = \text{Net income before taxes} - \text{Taxes}$$

Net sales are expected to be \$900,000. Cost of sales is estimated to be \$540,000. Selling expenses has a fixed component that is estimated to be \$90,000, and a variable component that is estimated to be 7% of net sales. Administrative expenses are \$50,000. Interest expenses are \$10,000. The company is taxed at a 50% rate. Develop a spreadsheet model to calculate the net income. Design your spreadsheet using good spreadsheet engineering principles.

14. A local pharmacy orders 15 copies of a monthly magazine. Depending on the cover story, demand for the magazine varies. The pharmacy purchases the magazines for \$2.25 and sells them for \$5.00. Any magazines left over at the end of the month are donated to hospitals and other health care facilities. Investigate the financial implications of this policy if the demand is expected to vary between 5 and 15 copies each month.

**15.** Koehler Vision Associates (KVA) specializes in laser-assisted corrective eye surgery. Prospective patients make appointments for prescreening exams to determine their candidacy for the surgery, and if they qualify, the \$300 charge is applied as a deposit for the actual procedure. The weekly demand is 175, and about 15% of prospective patients fail to show up or cancel their exam at the last minute. Patients that do not show up do not pay the prescreening fee. KVA can handle 125 patients per week and is considering overbooking its appointments to reduce the lost revenue associated with cancellations. However, any patient who is overbooked may spread unfavorable comments about the company; thus, the overbooking cost is estimated to be \$125, the value of a referral. Develop a spreadsheet model for calculating net revenue, and use data tables to study how revenue is affected by changes in the number of appointments accepted and patient demand.

**16.** A stockbroker calls on potential clients from referrals. For each call, there is a 15% chance that the client will decide to invest with the firm. Sixty percent of those interested are not found to be qualified based on the brokerage firm's screening criteria. The remaining are qualified. Of these, half will invest an average of \$5,000, 25% will invest an average of \$20,000, 15% will invest an average of \$50,000, and the remainder will invest \$100,000. The commission schedule is as follows:

Transaction Amount	Commission
Up to \$25,000	\$60 + 0.5% of the amount
\$25,001 to \$50,000	\$85 + 0.4% of the amount
\$50,001 to \$100,000	\$135 + 0.3% of the amount

The broker keeps half the commission. Develop a spreadsheet to calculate the broker's commission based on the number of calls per month made. Use data tables to show how the commission is a function of the number of calls made.

**17.** The director of a nonprofit ballet company in a medium-sized U.S. city is planning its next fund-raising campaign. In recent years, the program has found the following percentages of donors and gift levels:

Gift Level	Amount	Average Number of Gifts
Benefactor	\$10,000	2
Philanthropist	\$5,000	5
Producer's Circle	\$1,000	20
Director's Circle	\$500	35
Principal	\$100	5% of solicitations
Soloist	\$50	7% of solicitations

Develop a spreadsheet model to calculate the total amount donated based on this information and the number of

prospective donors that are solicited at the \$100 level or below. Use a data table to show how the amount varies based on the number of solicitations.

**18.** Jennifer Bellin has been put in charge planning her company's annual leadership conference. The dates of the conference have been determined by her company's executive team. The table below contains information about the activities, predecessors, and activity times (in days):

	Activity	Predecessors	Activity Time
A	Develop conference theme		3
B	Determine attendees		3
C	Contract facility	A	7
D	Choose entertainment	A	10
E	Send announcement	B	5
F	Order gifts	B	5
G	Order materials	B	1
H	Plan schedule of sessions	C	40
I	Design printed materials	B,H	15
J	Schedule session rooms	C	1
K	Print directions	H	10
L	Develop travel memo	E	5
M	Write gift letter	F	5
N	Confirm catering	H	3
O	Communicate with speakers	H	3
P	Track RSVPs and assign roommates	L	30
Q	Print materials	I	3
R	Assign table numbers	P	1
S	Compile packets of materials	G	3
T	Submit audio-visual needs	O	1
U	Put together welcome letter	P	5
V	Confirm arrangements with hotel	P	3
W	Print badges	G,P	5

Develop a spreadsheet model for finding the project completion time and critical path.

**19.** The Hyde Park Surgery Center specializes in high-risk cardiovascular surgery. The center needs to forecast its profitability over the next three years to plan for capital growth projects. For the first year, the hospital anticipates serving 1,500 patients, which is expected to grow by 8% per year. Based on current reimbursement formulas, each patient provides an average billing of

\$150,000, which will grow by 3% each year. However, because of managed care, the center collects only 35% of billings. Variable costs for supplies and drugs are calculated to be 12% of billings. Fixed costs for salaries, utilities, and so on, will amount to \$20,000,000 in the first year and are assumed to increase by 6% per year. Develop a spreadsheet model to calculate the NPV of profit over the next three years. Use a discount rate of 7%. Define three reasonable scenarios that the center director might wish to evaluate and use the *Scenario Manager* to compare them.

20. The admissions director of an engineering college has \$500,000 in scholarships each year from an endowment to offer to high-achieving applicants. The value of each scholarship offered is \$25,000 (thus, 20 scholarships are offered). The benefactor who provided the money would like to see all of it used each year for new

students. However, not all students accept the money; some take offers from competing schools. If they wait until the end of the admissions deadline to decline the scholarship, it cannot be offered to someone else, as any other good students would already have committed to other programs. Consequently, the admissions director offers more money than available in anticipation that a percentage of offers will be declined. If more than 10 students accept the offers, the college is committed to honoring them, and the additional amount has to come out of the dean's budget. Based on prior history, the percentage of applicants that accept the offer is about 70%. Develop a spreadsheet model for this situation, and apply whatever analysis tools you deem appropriate to help the admissions director make a decision on how many scholarships to offer. Explain your results in a business memo to the director, Mr. P. Woolston.

## Case

### An Inventory Management Decision Model

Inventories represent a considerable investment for every organization; thus, it is important that they be managed well. Excess inventories can indicate poor financial and operational management. On the other hand, not having inventory when it is needed can also result in business failure. The two basic inventory decisions that managers face are *how much* to order or produce for additional inventory, and *when* to order or produce it to minimize total inventory cost, which consists of the cost of holding inventory and the cost of ordering it from the supplier.

**Holding costs**, or *carrying costs*, represent costs associated with maintaining inventory. These costs include interest incurred or the opportunity cost of having capital tied up in inventories; storage costs such as insurance, taxes, rental fees, utilities, and other maintenance costs of storage space; warehousing or storage operation costs, including handling, recordkeeping, information processing, and actual physical inventory expenses; and costs associated with deterioration, shrinkage, obsolescence, and damage. Total holding costs are dependent on how many items are stored and for how long they are stored. Therefore, holding costs are expressed in terms of *dollars associated with carrying one unit of inventory for one unit of time*.

**Ordering costs** represent costs associated with replenishing inventories. These costs are not dependent on how many items are ordered at a time, but on the number of orders that are prepared. Ordering costs include overhead, clerical work, data processing, and other expenses that are incurred in searching for supply sources, as well as costs associated with purchasing, expediting, transporting, receiving, and inspecting. It is typical to assume that

the ordering cost is constant and is expressed in terms of *dollars per order*.

For a manufacturing company that you are consulting for, managers are unsure about making inventory decisions associated with a key engine component. The annual demand is estimated to be 15,000 units and is assumed to be constant throughout the year. Each unit costs \$80. The company's accounting department estimates that its opportunity cost for holding this item in stock for one year is 18% of the unit value. Each order placed with the supplier costs \$220. The company's policy is to place a fixed order for  $Q$  units whenever the inventory level reaches a predetermined reorder point that provides sufficient stock to meet demand until the supplier's order can be shipped and received.

As a consultant, your task is to develop and implement a decision model to help them arrive at the best decision. As a guide, consider the following:

1. Define the data, uncontrollable inputs, and decision variables that influence total inventory cost.
2. Develop mathematical functions that compute the annual ordering cost and annual holding cost based on average inventory held throughout the year in order to arrive at a model for total cost.
3. Implement your model on a spreadsheet.
4. Use data tables to find an approximate order quantity that results in the smallest total cost.
5. Use *Solver* to verify your result.
6. Conduct what-if analyses to study the sensitivity of total cost to changes in the model parameters.
7. Explain your results and analysis in a memo to the vice president of operations.

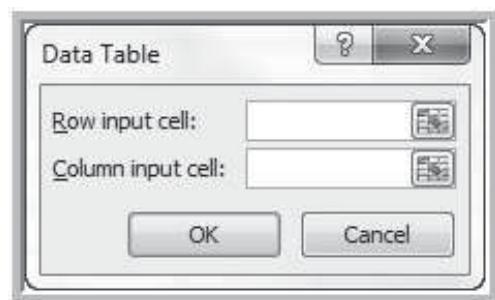
## Excel Notes

---

### A. Creating Data Tables

To create a one-way data table, first create a range of values for some input cell in your model that you wish to vary. The input values must be listed either down a column (column-oriented) or across a row (row-oriented). If the input values are column-oriented, enter the cell reference for the output variable in your model that you wish to evaluate in the row *above* the first value and one cell to the *right* of the column of input values. Reference any other output variable cells to the right of the first formula. If the input values are listed across a row, enter the cell reference of the output variable in the column to the *left* of the first value and one cell *below* the row of values. Type any additional output cell references below the first one. Next, select the range of cells that contains *both* the formulas and values you want to substitute. From the *Data* tab in Excel, select *Data Table* under the *What-If Analysis* menu. In the dialog box (see Figure 9A.1), if the input range is column-oriented, type the cell reference for the input cell in your model in the *Column input cell* box. If the input range is row-oriented, type the cell reference for the input cell in the *Row input cell* box.

To create a two-way data table, type a list of values for one input variable in a column, and a list of input values for the second input variable in a row, starting one row above and one column to the right of the column list. In the cell in the upper left-hand corner immediately above the column list and to the left of the row list, enter the cell reference of the output variable you wish to evaluate. Select the range of

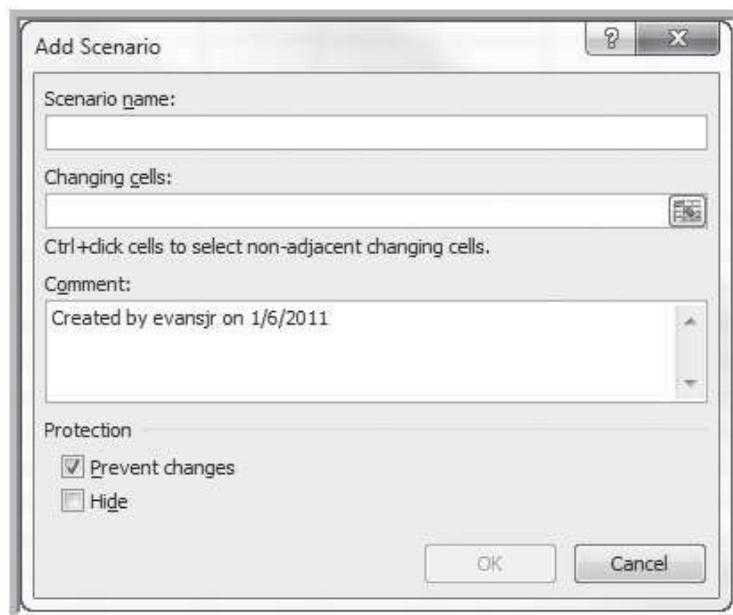


**FIGURE 9A.1** Data Table Dialog

cells that contains this cell reference and both the row and column of values. On the *What-If Analysis* menu, click *Data Table*. In the *Row input cell* of the dialog box, enter the reference for the input cell in the model for the input values in the row. In the *Column input cell* box, enter the reference for the input cell in the model for the input values in the column. Then click *OK*.

### B. Using the Scenario Manager

The Excel *Scenario Manager* is found under the *What-if Analysis* menu in the *Data Tools* group on the *Data* tab. When the tool is started, click the *Add* button to open the *Add Scenario* dialog and define a scenario (see Figure 9A.2). Enter the name of the scenario in the *Scenario name* box.

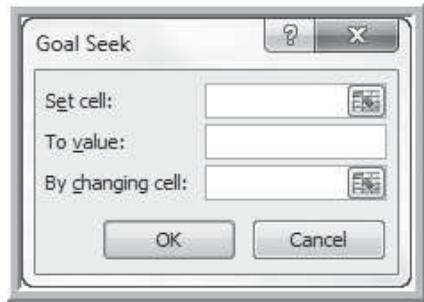


**FIGURE 9A.2** Add Scenario Dialog

In the *Changing cells* box, enter the references for the cells that you want to specify in your scenario. After all scenarios are added, they can be selected by clicking on the name of the scenario and then the *Show* button. Excel will change all values of the cells in your spreadsheet to correspond to those defined by the scenario in order for you to evaluate the results. You can also create a summary report on a new worksheet by clicking the *Summary* button on the *Scenario Manager* dialog.

## C. Using Goal Seek

On the *Data* tab, in the *Data Tools* group, click *What-If Analysis*, and then click *Goal Seek*. The dialog box shown in Figure 9A.3 will be shown. In the *Set cell* box, enter the reference for the cell that contains the formula that you want to resolve. In the *To value* box, type the formula result that you want. In the *By changing cell* box, enter the reference for the cell that contains the value that you want to adjust.



**FIGURE 9A.3** Goal Seek Dialog

## D. Net Present Value and the NPV Function

**Net present value** (also called **discounted cash flow**) measures the worth of a stream of cash flows, taking into account the time value of money. That is, a cash flow of  $F$  dollars  $t$  time periods in the future is worth  $F/(1 + i)^t$  dollars today, where  $i$  is the **discount rate**. The discount rate reflects the opportunity costs of spending funds now versus achieving a return through another investment, as well as the risks associated with not receiving returns until a later time. The sum of the present values of all cash flows over a stated time horizon is the NPV:

$$\text{NPV} = \sum_{t=0}^n \frac{F_t}{(1 + i)^t} \quad (9A.1)$$

where  $F_t$  = cash flow in period  $t$ . A positive NPV means that the investment will provide added value since the projected return exceeds the discount rate.

The Excel function  $\text{NPV}(rate, value1, value2, \dots)$  calculates the NPV of an investment by using a discount rate and a series of future payments (negative values) and income (positive values). *Rate* is the rate of discount over the length of one period ( $i$ ), and *value1, value2, ...* are 1 to 29 arguments representing the payments and income. The values must be equally spaced in time and are assumed to occur at the end of each period. The NPV investment begins one period before the date of the *value1* cash flow and ends with the last cash flow in the list. The NPV calculation is based on future cash flows. If the first cash flow (such as an initial investment or fixed cost) occurs at the beginning of the first period, then it must be added to the NPV result and *not* included in the function arguments.

# *Chapter 10*

# Decision Models with Uncertainty and Risk

- INTRODUCTION 325
- SPREADSHEET MODELS WITH RANDOM VARIABLES 325
  - Monte Carlo Simulation 326
- MONTE CARLO SIMULATION USING *CRYSTAL BALL* 327
  - Defining Uncertain Model Inputs 328
  - Running a Simulation 332
  - Saving *Crystal Ball* Runs 334
  - Analyzing Results 334
  - *Crystal Ball* Charts 339
  - *Crystal Ball* Reports and Data Extraction 342
  - *Crystal Ball* Functions and Tools 342
- APPLICATIONS OF MONTE CARLO SIMULATION AND *CRYSTAL BALL* FEATURES 343
  - Newsvendor Model: Fitting Input Distributions, *Decision Table* Tool, and Custom Distribution 343
  - Overbooking Model: *Crystal Ball* Functions 348
  - Cash Budgeting: Correlated Assumptions 349
  - New Product Introduction: *Tornado Chart* Tool 352
  - Project Management: Alternate Input Parameters and the *Bootstrap* Tool 353
- BASIC CONCEPTS REVIEW QUESTIONS 358
- PROBLEMS AND APPLICATIONS 359
- CASE: J&G BANK 362
- APPENDIX 10.1: *CRYSTAL BALL* NOTES 362
  - A. Customizing *Define Assumption* 362
  - B. Sensitivity Charts 363
  - C. Distribution Fitting with *Crystal Ball* 363
  - D. *Correlation Matrix* Tool 365
  - E. *Tornado* Charts 365
  - F. *Bootstrap* Tool 366

## INTRODUCTION

In the models we studied in Chapter 9, all the data—particularly the uncontrollable inputs—were assumed to be known and constant. In many situations, this assumption may be far from reality because uncontrollable inputs usually exhibit random behavior. Some examples would be customer demand, arrivals to ATM machines, and returns on investments. We often assume such variables to be constant in order to simplify the model and the analysis. However, many situations dictate that randomness be explicitly incorporated into our models. This is usually done by specifying probability distributions for the appropriate uncontrollable inputs. Models that include randomness are called *probabilistic*, or *stochastic*, models. These types of models help to evaluate risks associated with undesirable consequences and to find optimal decisions under uncertainty.

**Risk** is simply the probability of occurrence of an undesirable outcome. For example, we could answer such questions as: What is the probability that we will incur a financial loss? What is the probability that we will run out of inventory? What are the chances that a project will be completed on time? **Risk analysis** is an approach for developing “a comprehensive understanding and awareness of the risk associated with a particular variable of interest (be it a payoff measure, a cash flow profile, or a macroeconomic forecast).”<sup>1</sup> Hertz and Thomas present a simple scenario to illustrate the concept of risk analysis:

The executives of a food company must decide whether to launch a new packaged cereal. They have come to the conclusion that five factors are the determining variables: advertising and promotion expense, total cereal market, share of market for this product, operating costs, and new capital investment. On the basis of the “most likely” estimate for each of these variables, the picture looks very bright—a healthy 30% return, indicating a significantly positive expected net present value (NPV). This future, however, depends on each of the “most likely” estimates coming true in the actual case. If each of these “educated guesses” has, for example, a 60% chance of being correct, there is only an 8% chance that all five will be correct ( $0.60 \times 0.60 \times 0.60 \times 0.60 \times 0.60$ ) if the factors are assumed to be independent. So the “expected” return, or present value measure, is actually dependent on a rather unlikely coincidence. The decision maker needs to know a great deal more about the other values used to make each of the five estimates and about what he stands to gain or lose from various combinations of these values.<sup>2</sup>

Thus, risk analysis seeks to examine the impacts of uncertainty in the estimates and their potential interaction with one another on the output variable of interest. Hertz and Thomas also note that the challenge to risk analysts is to frame the output of risk analysis procedures in a manner that makes sense to the manager and provides clear insight into the problem, suggesting that simulation has many advantages.

In this chapter, we discuss how to build and analyze models involving uncertainty and risk. We will introduce Monte Carlo simulation with an Excel add-in, *Crystal Ball*, which is a powerful commercial package that is used by many of the *Fortune* 1000 companies to perform risk analyses.

## SPREADSHEET MODELS WITH RANDOM VARIABLES

Recall that in Chapter 3 we described how to sample randomly from probability distributions and to generate certain random variates using Excel tools and functions. We will use these techniques to show how to incorporate uncertainty into decision models.

<sup>1</sup> David B. Hertz and Howard Thomas, *Risk Analysis and Its Applications* (Chichester, UK: John Wiley & Sons, Ltd., 1983): 1.

<sup>2</sup> *Ibid.*, 24.

## Monte Carlo Simulation

Refer back to the outsourcing decision model we introduced in Chapter 9 (see Figure 9.2). Assume that demand is uncertain. We can model the demand as a random variable having some probability distribution. Suppose the manufacturer has enough data and information to assume that demand will be normally distributed with a mean of 1,000 and a standard deviation of 100. We could use the Excel function *NORMINV (probability, mean, standard\_deviation)* as described in Chapter 3 to generate random values of the demand by replacing the input in cell B14 of Figure 9.2 with the formula =ROUND(NORMINV(RAND(), 1000, 100),0). The ROUND function is used to ensure that the values will be whole numbers. Whenever the F9 key is pressed, the worksheet will be recalculated and the value of demand will change randomly.

We can use Monte Carlo sampling techniques that we introduced in Chapter 3 to analyze this model as a basis for a decision. We can randomly generate a demand and compute the difference and associated decision, and then repeat this for some number of trials. **Monte Carlo simulation** is the process of generating random values for uncertain inputs in a model, computing the output variables of interest, and repeating this process for many trials in order to understand the distribution of the output results.

Monte Carlo simulation can easily be accomplished for the outsourcing model using a data table as shown in Figure 10.1. Construct a data table by listing the number of trials down a column (here we used 20 trials), and referencing the cells associated with demand, the difference, and the decision in cells E2, F2, and G2, respectively (i.e., the formula in cell E2 is =B14; in cell F2, =B18; and in cell G2, =B20). Select the range of the table (D2:G22)—and here's the trick—in the *Column Input Cell* field in the *Data Table* dialog, enter any blank cell in the spreadsheet. This is done because the trial number does not relate to any parameter in the model; we simply want to repeat the spreadsheet

A	B	C	D	E	F	G
1	Outsourcing Decision Model			Demand	Difference	Decision
2			Trial	1090	\$ (4,500)	Manufacture
3	Data		1	952	\$ 2,400	Outsource
4			2	857	\$ 7,150	Outsource
5	Manufactured in-house		3	812	\$ 9,400	Outsource
6	Fixed cost	\$ 50,000	4	874	\$ 6,300	Outsource
7	Unit variable cost	\$ 125	5	860	\$ 7,000	Outsource
8			6	1037	\$ (1,850)	Manufacture
9	Purchased from supplier		7	888	\$ 5,600	Outsource
10	Unit cost	\$ 175	8	1023	\$ (1,150)	Manufacture
11			9	934	\$ 3,300	Outsource
12	Model		10	1054	\$ (2,700)	Manufacture
13			11	1096	\$ (4,800)	Manufacture
14	Demand volume	1090	12	911	\$ 4,450	Outsource
15			13	828	\$ 8,600	Outsource
16	Total manufacturing cost	\$ 186,250	14	1034	\$ (1,700)	Manufacture
17	Total purchased cost	\$ 190,750	15	997	\$ 150	Outsource
18	Difference	\$ (4,500)	16	1137	\$ (6,850)	Manufacture
19			17	904	\$ 4,800	Outsource
20	Decision	Manufacture	18	985	\$ 750	Outsource
21			19	970	\$ 1,500	Outsource
22			20	957	\$ 2,150	Outsource
23				Average	\$ 2,225	
24				% Manufacture		30%
25				% Outsource		70%

FIGURE 10.1 Monte Carlo Simulation of Outsourcing Decision Model

recalculation each time, knowing that the demand will change each time because of the use of the RAND function in the demand formula.

As you can see from the figure, each trial has a randomly generated demand. The data table substitutes these demands into cell B14 and finds the associated difference and decision. The average difference is \$240, and 70% of the trials resulted in outsourcing as the best decision. These results might suggest that, although the future demand is not known, the manufacturer's best choice might be to outsource. However, there is a risk that this may not be the best decision.

In addition, the small number of trials that we used in this example makes sampling error an important issue. We could easily obtain significantly different results if we repeat the simulation (by pressing the F9 key). For example, a few clicks yielded the following percentages for outsourcing as the best decision: 40%, 60%, 65%, 45%, 75%, 45%, and 35%. There is considerable variability in the results, but this can be reduced by using a larger number of trials.

### SKILL-BUILDER EXERCISE 10.1

Implement the Monte Carlo simulation for the outsourcing decision model but expand the data table to 50 trials. Repeat the simulation five times. What do you observe?

Although the use of a data table illustrates the value of incorporating uncertainty into a decision model, it is impractical to apply to more complex problems. For example, in the *Moore Pharmaceuticals* model in Chapter 9, many of the model parameters such as the initial market size, project costs, market size growth factors, and market share growth rates may all be uncertain. In addition, we need to be able to capture and save the results of each trial, and it would be useful to construct a histogram of the results in order to conduct further analyses. Fortunately, sophisticated software approaches that easily perform these functions are available. The remainder of this chapter is focused on learning to use *Crystal Ball* software to perform large-scale Monte Carlo simulation.

## MONTE CARLO SIMULATION USING CRYSTAL BALL

*Crystal Ball* is an Excel add-in that performs Monte Carlo simulation for risk analysis. We have also seen that *Crystal Ball* includes a forecasting module in Chapter 7; it also includes an optimization module that we will discuss in Chapter 14. Start *Crystal Ball* by selecting it from the Windows *Start* menu. Click the *Crystal Ball* tab in the Excel menu bar to display the menus.

To use *Crystal Ball*, you must perform the following steps:

1. Develop the spreadsheet model
2. Define *assumptions* for uncertain variables, that is, the probability distributions that describe the uncertainty
3. Define the *forecast cells*, that is, the output variables of interest
4. Set the number of trials and other run preferences
5. Run the simulation
6. Interpret the results

To illustrate this process, we will use the *Moore Pharmaceuticals* spreadsheet model that we introduced in Chapter 9 (Figure 9.11) as the basis for discussion. While the values used in the spreadsheet suggest that the new drug would become profitable by the fourth year, much of the data in this model are uncertain. Thus, we might

be interested in evaluating the risk associated with the project. Questions we might be interested in are:

- What is the risk that the NPV over the five years will not be positive?
- What are the chances that the product will show a cumulative net profit in the third year?
- What cumulative profit in the fifth year are we likely to realize with a probability of at least 0.90?

## Defining Uncertain Model Inputs

When model inputs are uncertain, we need to characterize them by some probability distribution. For many decision models, empirical data may be available, either in historical records or collected through special efforts. For example, maintenance records might provide data on machine failure rates and repair times, or observers might collect data on service times in a bank or post office. This provides a factual basis for choosing the appropriate probability distribution to model the input variable. We can identify an appropriate distribution by fitting historical data to a theoretical model (we will do this later in this chapter).

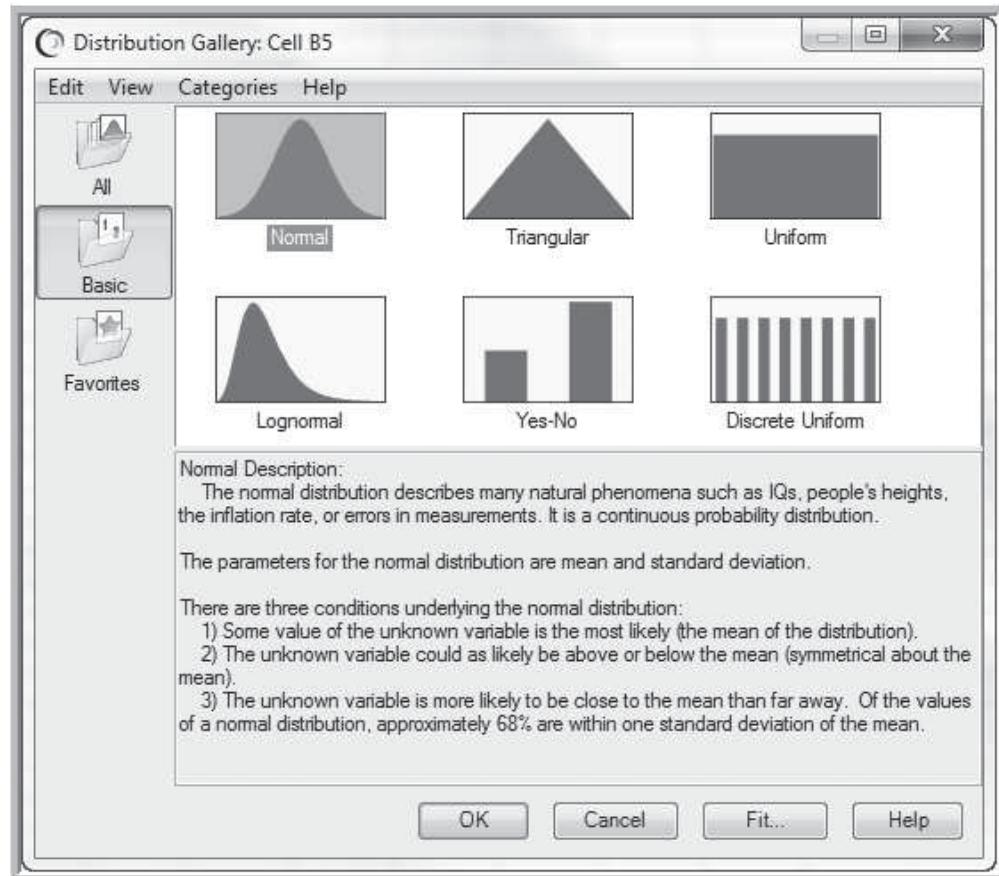
In other situations, historical data are not available, and we can draw upon the properties of common probability distributions and typical applications that we discussed in Chapter 3 to help choose a representative distribution that has the shape that would most reasonably represent the analyst's understanding about the uncertain variable. For example, a normal distribution is symmetric, with a peak in the middle. Exponential data are very positively skewed, with no negative values. A triangular distribution has a limited range and can be skewed in either direction.

Very often, uniform or triangular distributions are used in the absence of data. These distributions depend on simple parameters that one can easily identify based on managerial knowledge and judgment. For example, to define the uniform distribution, we need to know only the smallest and largest possible values that the variable might assume. For the triangular distribution, we also include the most likely value. In the construction industry, for instance, experienced foremen can easily tell you the fastest, most likely, and slowest times it would take to perform a task such as framing a house, taking into account possible weather and material delays, labor absences, and so on.

**ASSUMPTIONS** In *Crystal Ball*, uncertain inputs are called **assumptions**. Suppose that the project manager has identified the following distributions and parameters for these variables:

- Market size: normal with a mean of 2,000,000 units and a standard deviation of 400,000 units
- R&D costs: uniform between \$600,000,000 and \$800,000,000
- Clinical trial costs: lognormal with mean of \$150,000,000 and standard deviation \$30,000,000
- Annual market growth factor: triangular with minimum = 2%, maximum = 6%, and most likely = 3%
- Annual market share growth rate: triangular with minimum = 15%, maximum = 25%, and most likely = 20%

To define these assumptions in *Crystal Ball*, first select the cell corresponding to the uncertain input. Assumption cells must contain a value; they cannot be defined for formula, nonnumeric, or blank cells. From the *Define* group in the *Crystal Ball* tab, click *Define Assumption*. *Crystal Ball* displays a gallery of probability distributions from which to choose and prompts you for the parameters. For example, let us define the distribution for the market size. First, click on cell B5 and then select *Define Assumption*. *Crystal Ball*

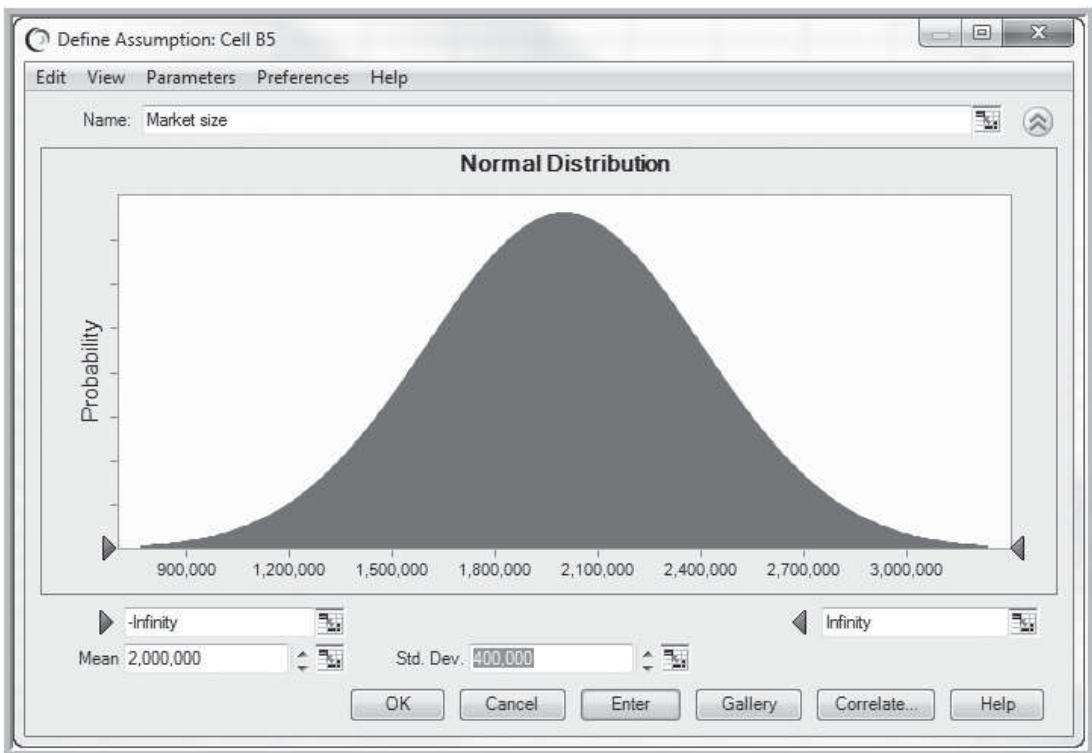


**FIGURE 10.2** Crystal Ball Distribution Gallery

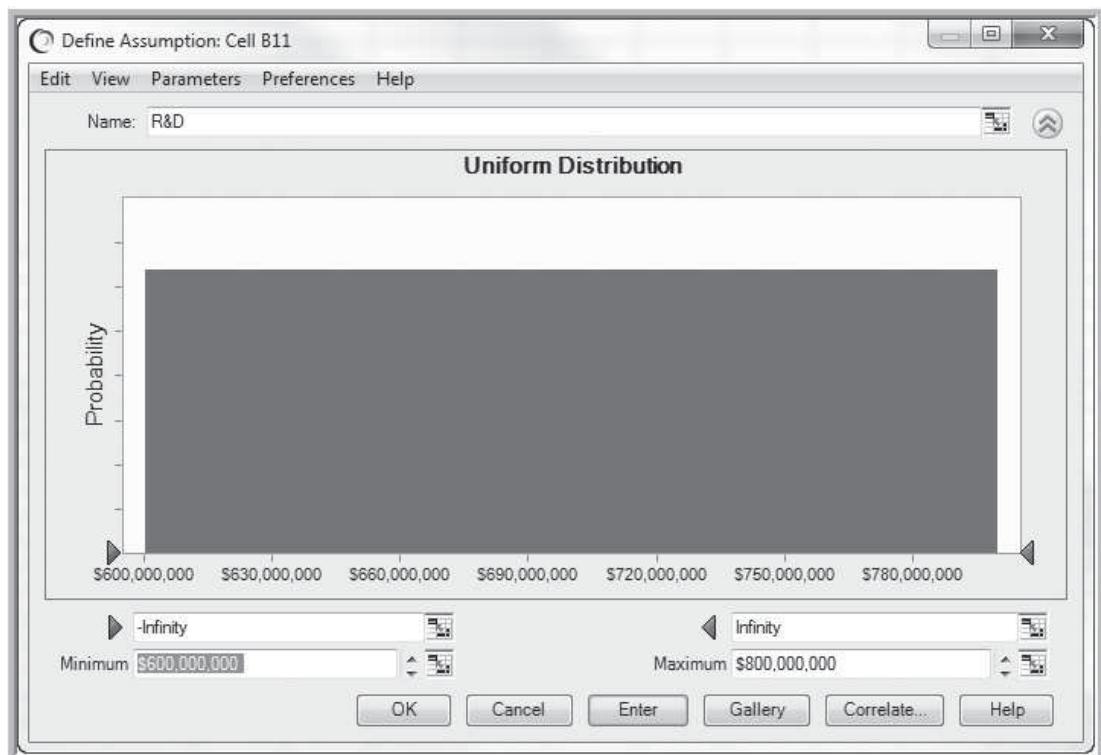
displays the distribution gallery, shown in Figure 10.2. You can select one of the folders on the left of the dialog box to display all distributions or previously defined favorites. Since we assume that this variable has a normal distribution, we click on the normal distribution then the *OK* button (or simply double-click the distribution). A dialog is then displayed, prompting you for the parameters associated with this distribution. We suggest that you first enter a clear, descriptive name for your assumptions in the top box. *Crystal Ball* will automatically use text in cells immediately to the left or above assumption cells, but these may not be the correct ones for your application.

*Crystal Ball* anticipates the default parameters based on the current values in the spreadsheet model. For example, with a normal distribution, the default mean is the assumption cell value, and the standard deviation is assumed to be 10% of the mean. Therefore, in our example, we need to change the standard deviation to 400,000. You may either enter the data in the appropriate boxes in the dialog, or click on the buttons to the right of the boxes to enter a cell reference. Clicking on *Enter* fixes these values and rescales the picture to allow you to see what the distribution looks like (this feature is quite useful for flexible families of distributions such as the triangular). Figure 10.3 shows the completed *Define Assumption* screen with the correct parameters for the market size. After an assumption is defined, the spreadsheet cell is shaded in green (default color) so that you may easily identify the assumptions.

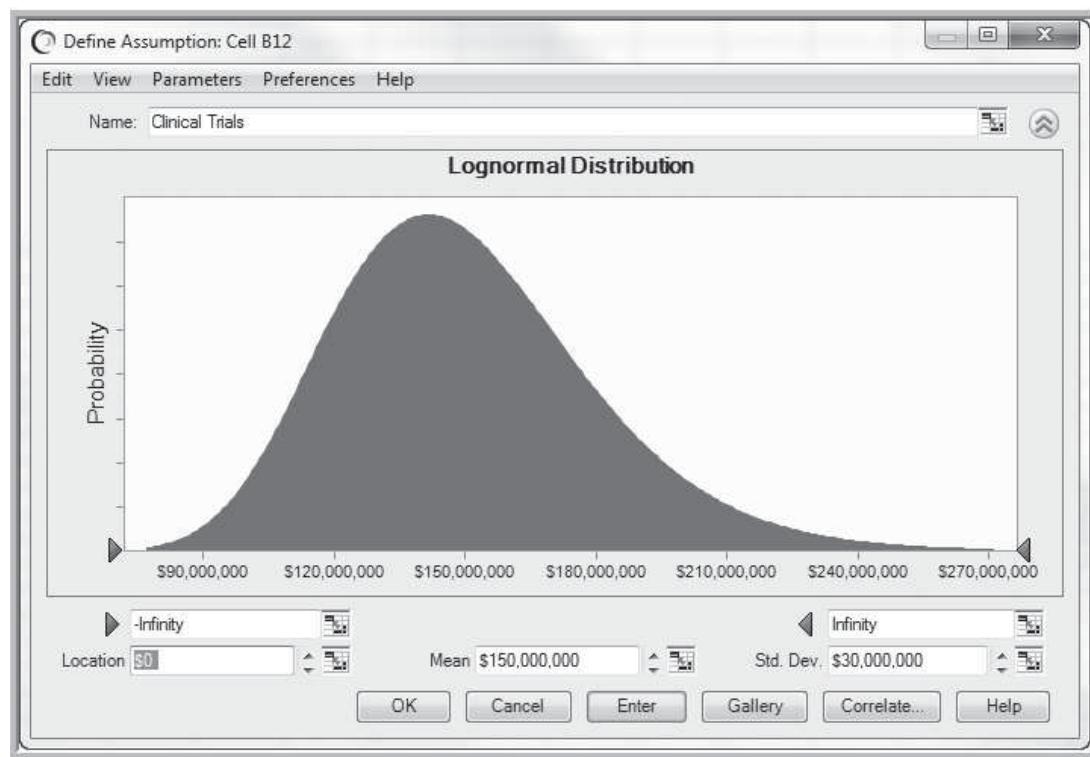
We repeat this process for each of the probabilistic assumptions in the model. Figure 10.4, for example, shows the uniform distribution for the R&D costs. Figures 10.5 and 10.6 show the dialogs for the clinical trials and the year 2 annual market growth factor assumptions, respectively. *Crystal Ball* also has various options for specifying



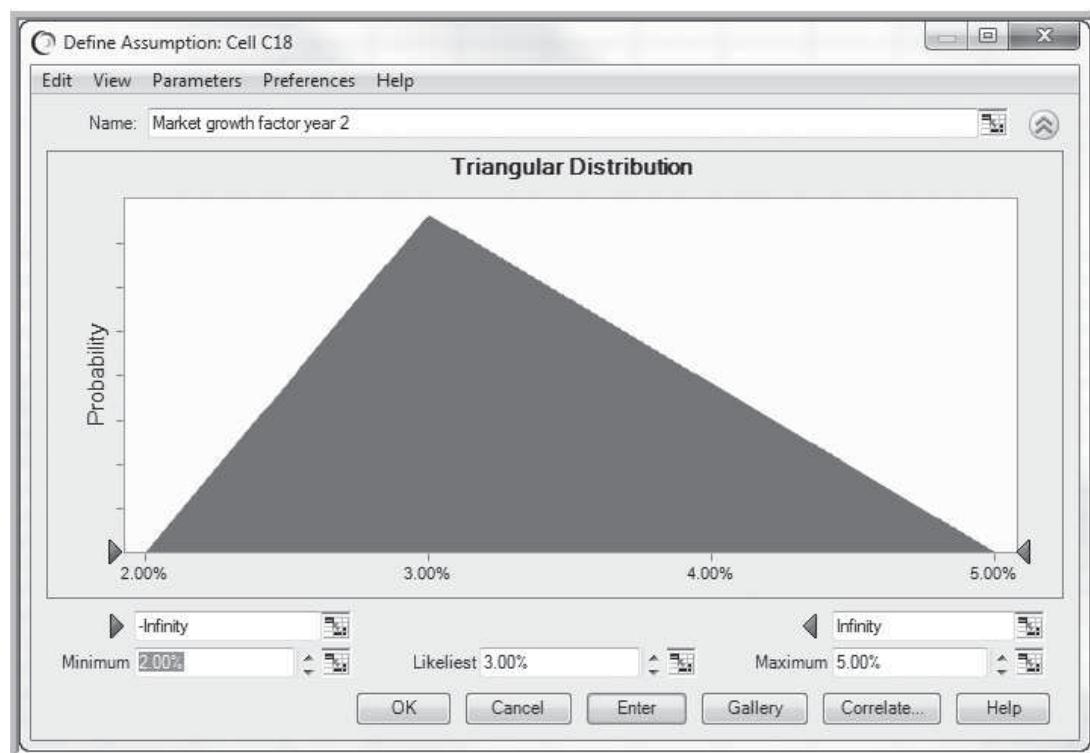
**FIGURE 10.3** Market Size Normal Distribution Assumption



**FIGURE 10.4** R&D Cost Assumption



**FIGURE 10.5** Clinical Trials Cost Assumption



**FIGURE 10.6** Year 2 Market Growth Factor Assumption



## Spreadsheet Note

input information and customizing the views of the assumptions (see Appendix 10.1A, *Customizing Define Assumption*).

If several cells have the same assumptions, you may use the *Copy* and *Paste* commands in the *Define* group within the *Crystal Ball* ribbon (do not use the standard Excel copy and paste commands). For example, you may copy the assumption for the year 2 market growth factor, and then paste it into the cells for years 3, 4, and 5. After you do this, you may need to edit the names of the assumptions in the dialogs.

**FORECASTS** The output variables in which we are interested are called **forecast cells**. In our example, a key forecast cell would be the NPV in cell B29. Click on *Define Forecast* from the *Define* group in the *Crystal Ball* ribbon, and enter a name for the forecast and unit of measure, if desired in the dialog. Figure 10.7 shows this for their NPV. After a forecast is defined, the spreadsheet cell is shaded in blue to indicate the forecast and distinguish it from the assumptions. We might also wish to have output information about the cumulative net profit each year (cells B27:F27). Instead of choosing them individually, highlight the entire range, and select *Define Forecast* (*Crystal Ball* will cycle through each of the cells, prompting for input). Figure 10.8 shows the completed spreadsheet in which the assumptions and forecasts have been highlighted by *Crystal Ball*.

## Running a Simulation

Prior to running a simulation, you need to set some specifications for the simulation. To do this, select the *Run Preferences* item from the *Run* group. The first dialog box has several tabs as shown in Figure 10.9. The first tab *Trials*, allows you to choose the number of times that *Crystal Ball* will generate assumptions for the assumption cells in the model and recalculate the entire spreadsheet. Because Monte Carlo simulation is essentially statistical sampling, the larger the number of trials you use, the more precise will be the result. Unless the model is extremely complex, a large number of trials will not unduly tax today's computers, so we recommend that you use at least the default value of 5,000 trials. You should use a larger number of trials as the number of assumption cells in your model increases so that *Crystal Ball* can generate representative samples from all distributions for assumptions. For our example, we will set the number of trials to 10,000.

In the *Sampling* tab, you can choose to use the same sequence of random numbers for generating the random variates in the assumption cells; this will guarantee that the same assumption values will be used each time you run the model. This is useful when you wish to change a controllable variable in your model and compare results for the same assumption values. If you check the box "*Use same sequence of random numbers*," you may specify an *Initial seed value*. As long as you use the same number, the assumptions generated will be the same for all simulations. Otherwise, every time you run the model you will get slightly different results because of sampling error.

*Crystal Ball* has two types of sampling methods: Monte Carlo and Latin Hypercube. Monte Carlo sampling selects random variates independently over the entire range of

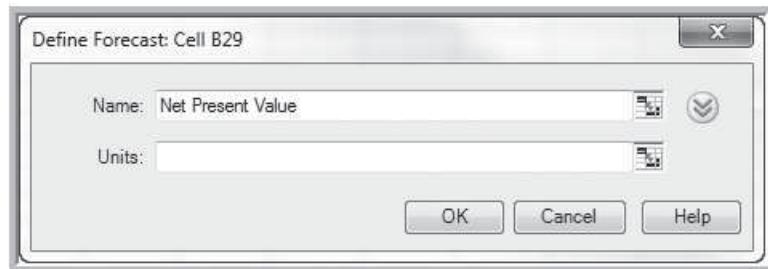
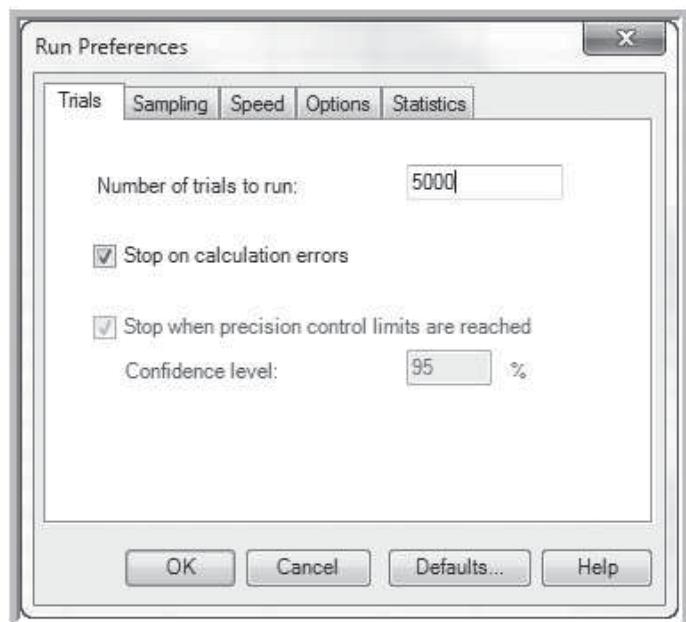


FIGURE 10.7 Define Forecast Dialog

A	B	C	D	E	F
1	Moore Pharmaceuticals				
2					
3	Data				
4					
5	Market size	2,000,000			
6	Unit (monthly Rx) revenue	\$ 130.00			
7	Unit (monthly Rx) cost	\$ 40.00			
8	Discount rate	9%			
9					
10	Project Costs				
11	R&D	\$ 700,000,000			
12	Clinical Trials	\$ 150,000,000			
13	Total Project Costs	\$ 850,000,000			
14					
15	Model				
16					
17	Year	1	2	3	4
18	Market growth factor		3.00%	3.00%	3.00%
19	Market size	2,000,000	2,060,000	2,121,800	2,185,454
20	Market share growth rate		20.00%	20.00%	20.00%
21	Market share	8.00%	9.60%	11.52%	13.82%
22	Sales	160,000	197,760	244,431	302,117
23					
24	Annual Revenue	\$ 249,600,000	\$ 308,505,600	\$ 381,312,922	\$ 471,302,771
25	Annual Costs	\$ 76,800,000	\$ 94,924,800	\$ 117,327,053	\$ 145,016,237
26	Profit	\$ 172,800,000	\$ 213,580,800	\$ 263,985,869	\$ 326,286,534
27	Cumulative Net Profit	\$ (677,200,000)	\$ (463,619,200)	\$ (199,633,331)	\$ 126,653,203
28					
29	Net Present Value	\$ 185,404,860			

**FIGURE 10.8** Moore Pharmaceuticals Spreadsheet with *Crystal Ball* Assumptions and Forecasts



**FIGURE 10.9** Run Preferences Dialog

possible values. With Latin Hypercube sampling, *Crystal Ball* divides each assumption's probability distribution into intervals of equal probability and generates an assumption value randomly within each interval. Latin Hypercube sampling results in a more even distribution of forecast values because it samples the entire range of the distribution in a more consistent manner, thus achieving more accurate forecast statistics (particularly the mean) for a fixed number of Monte Carlo trials. However, Monte Carlo sampling is more representative of reality and should be used if you are interested in evaluating the model performance under various "what-if" scenarios. We recommend leaving the default values in the *Speed*, *Options*, and *Statistics* tabs of the *Run Preferences* dialog.

The last step is to run the simulation by clicking the *Start* button in the *Run* group and watch *Crystal Ball* go to work! You may choose the *Step* run option if you wish to see the results of individual trials. This is a useful option in the early stages of model building and to debug your spreadsheets. You may stop and reset the simulation (which clears all statistical results) using the appropriate buttons.

### Saving Crystal Ball Runs

When you save your spreadsheet in Excel, any assumptions and forecasts that you defined for *Crystal Ball* are also saved. However, this does not save the results of a *Crystal Ball* simulation. To save a *Crystal Ball* simulation, select *Save or Restore* from the *Run* group and then click on *Save Results*. Doing so allows you to save any customized chart settings and other simulation results and recall them without rerunning the simulation. To retrieve a *Crystal Ball* simulation, choose *Restore Results* from the *Save or Restore* menu.

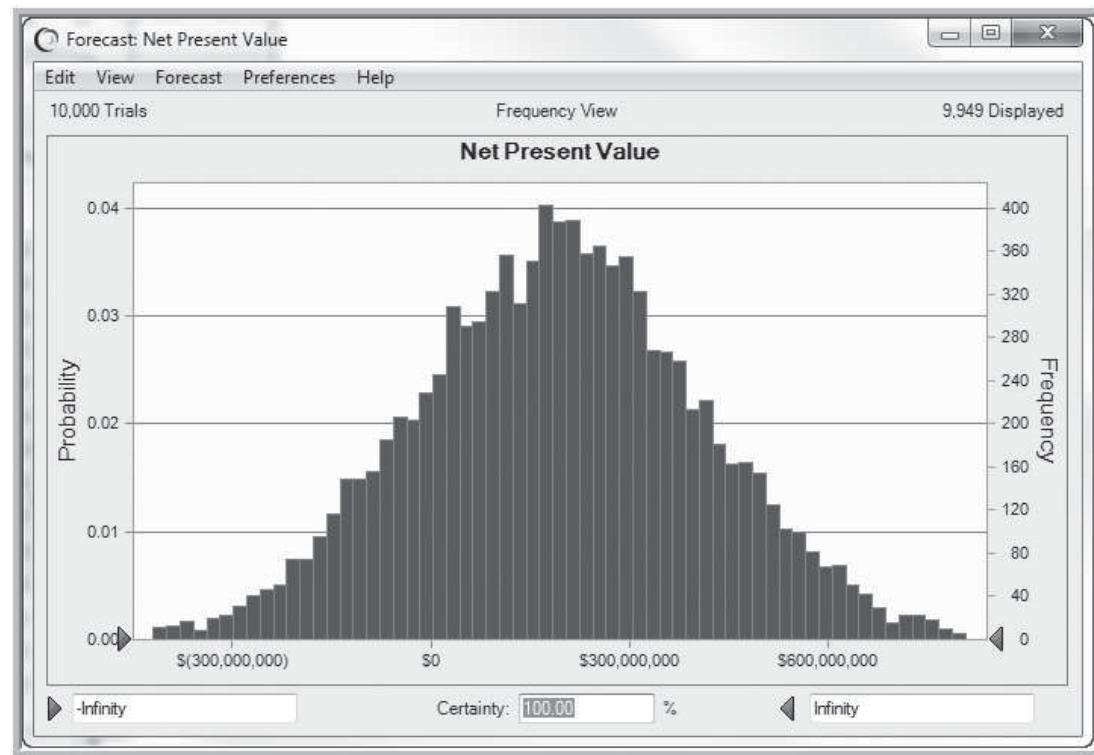
### Analyzing Results

The principal output results provided by *Crystal Ball* are the *forecast chart*, *percentiles summary*, and *statistics summary*. The forecast chart is automatically displayed when the simulation ends. Figure 10.10 shows the forecast chart for the five-year cumulative discounted cash flow. The forecast chart is simply a histogram of the outcome variable that includes all values within a default value of 2.6 standard deviations of the mean, which represents approximately 99% of the data. The actual number of values displayed is shown in the upper right corner of the chart.

Just below the horizontal axis at the extremes of the distribution are two small triangles called *endpoint grabbers*. The range values of the variable at these positions are given in the boxes at the bottom left and right corners of the chart; in Figure 10.10, this shows that the grabbers are positioned at minus and plus infinity. The percentage of data values between the grabbers is displayed in the *Certainty* box at the lower center of the chart. A **certainty level** is a probability interval that states the probability of the forecast falling within the specified range of the grabbers.

Questions involving risk can now be answered by manipulating the endpoint grabbers or by changing the range and certainty values in the boxes. Several options exist.

1. **You may move an endpoint grabber by clicking on the grabber and dragging it along the axis.** As you do, the distribution outside of the middle range changes color, the range value corresponding to the grabber changes to reflect its current position, and the certainty level changes to reflect the new percentage between the grabbers.
2. **You may type in specific values in the range boxes.** When you do, the grabbers automatically move to the appropriate positions and the certainty level changes to reflect the new percentage of values between the range values.
3. **You may specify a certainty level.** If the endpoint grabbers are free (as indicated by a dark gray color), the certainty range will be centered around the median. We caution you that a certainty range is *not* a confidence interval. It is simply a probability interval. Confidence intervals, as discussed in Chapter 4, depend on the sample size and relate to the sampling distribution of a statistic.



**FIGURE 10.10** Forecast Chart for Net Present Value

You may anchor an endpoint grabber by clicking on it. (When anchored, the grabber will be a lighter color; to free an anchored grabber, click on it again.) If a grabber is anchored and you specify a certainty level, the free grabber moves to a position corresponding to this level. You may also cross over the grabbers and move them to opposite ends to determine certainty levels for the tails of the distribution.

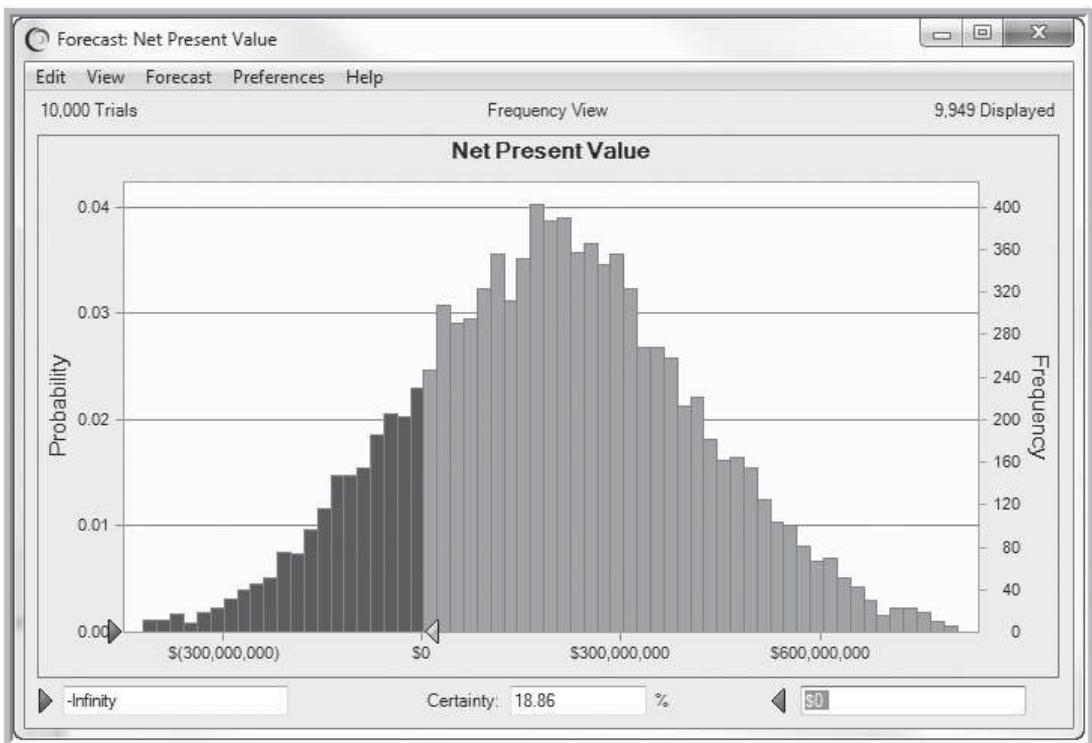
Let us use these capabilities to answer the risk analysis questions we posed earlier.

- **What is the risk that the NPV over the five years will not be positive?** If you enter the number 0 in the right range box of the forecast chart and press the enter key, the grabber will automatically move to that position, the portion of the histogram to the right of 0 will change color, and the certainty level will change to reflect the percentage of the distribution between the grabbers. This is illustrated in Figure 10.11, which shows an 81.14% chance of a positive NPV.

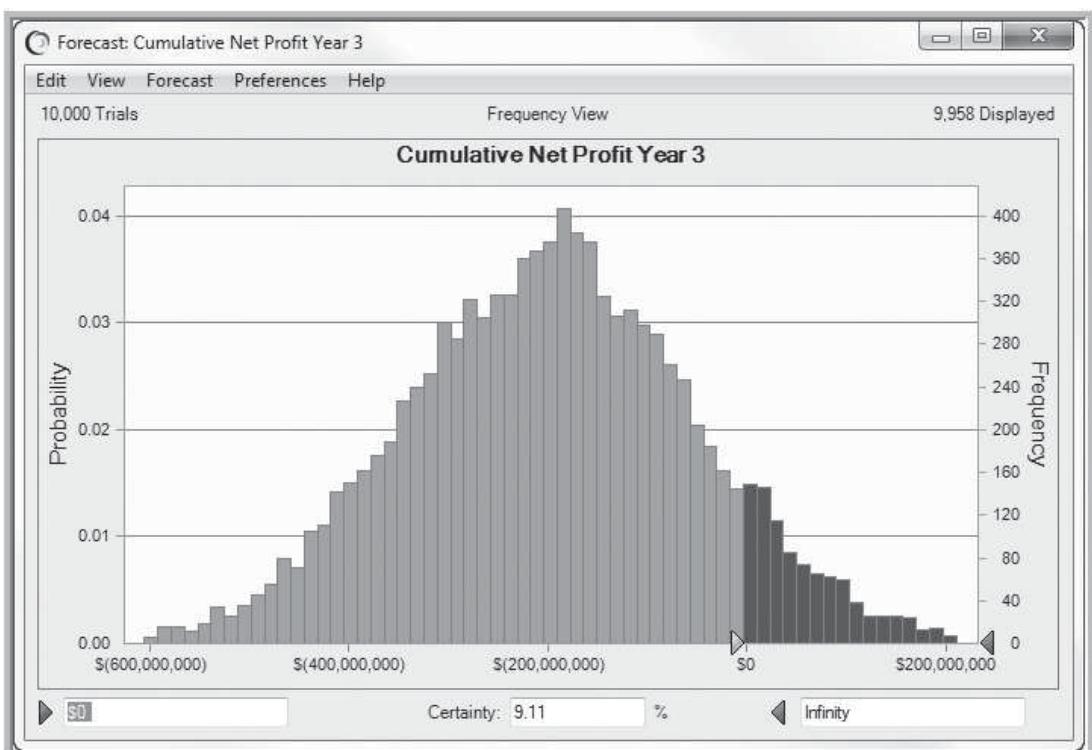
- **What are the chances that the product will show a cumulative net profit in the third year?** If you examine the forecast chart for the net profit in year 3, enter the value 0 into the left range box as illustrated in Figure 10.12. This shows that the probability of a positive cumulative net profit in the third year is only 9.11%.

- **What cumulative profit in the fifth year are we likely to realize with a probability of at least 0.90?** In the forecast chart for the cumulative net profit for year 5, first anchor the right grabber at infinity and then enter 90 in the *Certainty* box. This provides the value in the left range box for which there is a 0.9 probability of exceeding (see Figure 10.13). Therefore, we can expect a cumulative net profit of at least \$171,260,282 with 90% certainty.

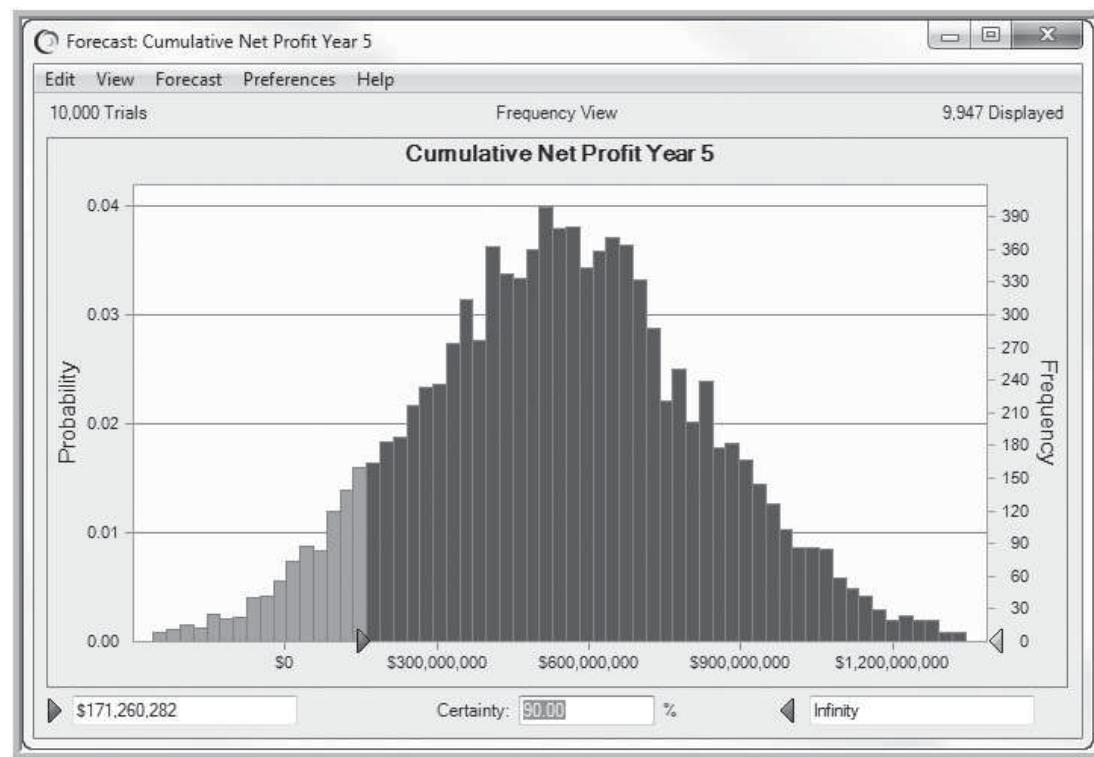
The forecast chart may be customized to change its appearance from the *Preferences* menu in the forecast chart. We encourage you to experiment with these options. You



**FIGURE 10.11** Probability of a Nonpositive Net Present Value



**FIGURE 10.12** Cumulative Net Profit in Year 3



**FIGURE 10.13** Cumulative Net Profit in Year 5

may also fit the forecast data to a distribution by selecting *Fit Probability Distribution* from the *Forecast* menu in the chart.

Percentiles and statistics summaries can be selected from the *View* menu in the forecast chart (selecting *Split View* first allows you to display the forecast chart and/or the percentiles and statistics in one window). The percentiles summary is shown in Figure 10.14. This is essentially the cumulative probability distribution of the forecast. For example, we see that the chance that the NPV will be less than \$304,117,491 is 70%.

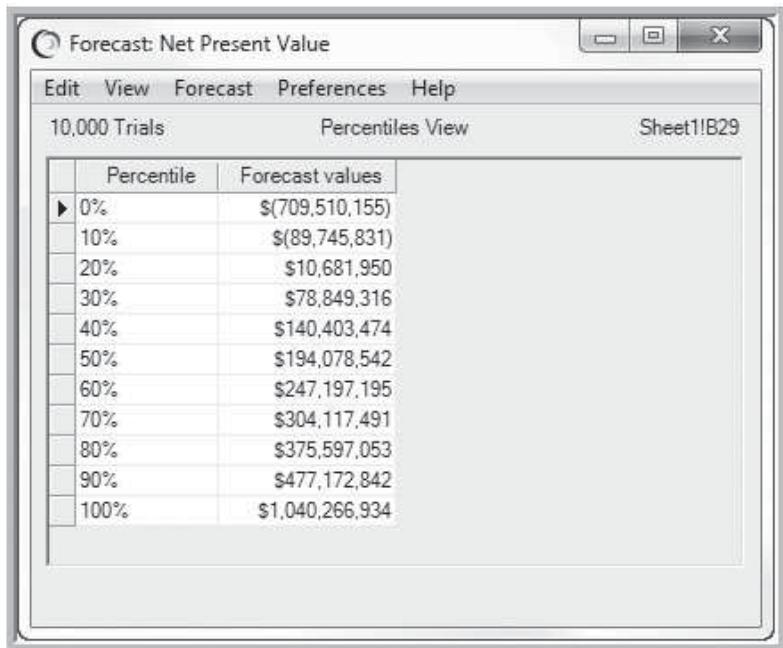
The statistics summary, shown in Figure 10.15, provides a summary of key descriptive statistical measures. The mean standard error is reported on the last line of the statistics report and defines the standard deviation for the sampling distribution of the mean as discussed in Chapter 3. We may use this to construct a confidence interval for the mean using the formula:

$$\bar{x} \pm z_{\alpha/2}(s/\sqrt{n})$$

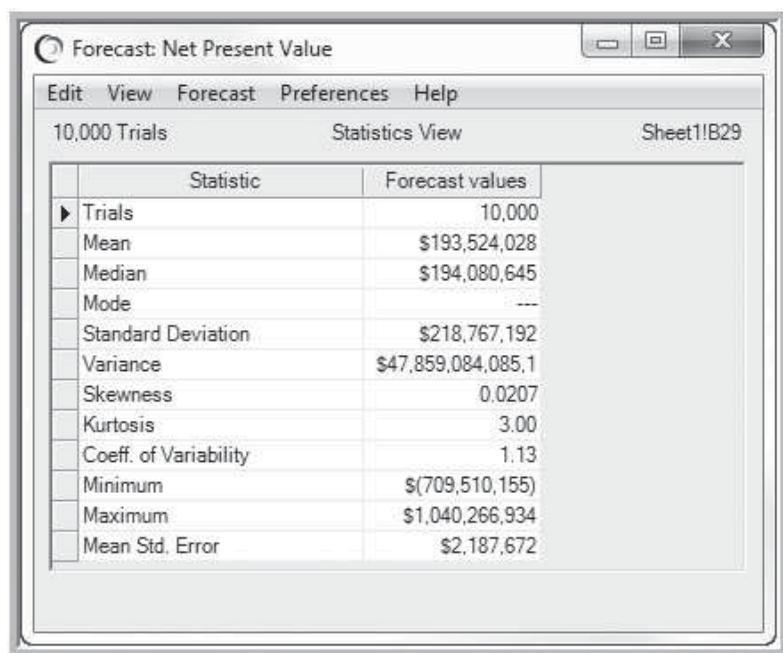
Because a *Crystal Ball* simulation will generally have a very large number of trials, we may use the standard normal *z*-value instead of the *t*-distribution. Thus, for the NPV results, a 95% confidence interval for the mean would be:

$$\$193,524,028 \pm 1.96(2,187,672) \text{ or } [\$189,236,191, \$197,811,865]$$

This means that if we ran the simulation again with different random inputs, we could expect the mean NPV to generally fall within this interval. To reduce the size of the confidence interval, we would need to run the simulation for a larger number of trials. For most risk analysis applications, however, the mean is less important than the actual distribution of outcomes.



**FIGURE 10.14** Percentiles Summary



**FIGURE 10.15** Statistics Summary

### SKILL-BUILDER EXERCISE 10.2

Open the *Moore Pharmaceuticals* spreadsheet and enter all *Crystal Ball* assumptions and forecasts. Use the *Copy* and *Paste* commands for the market growth factor and market share growth rate assumptions. Run the simulation for 10,000 trials. Find the mean and variance of the cumulative net profit for each year, and probabilities of a positive cumulative net profit for each year.

## Crystal Ball Charts

Crystal Ball also provides a variety of charts to help you analyze the results of a simulation. These can be selected from the *View Charts* option in the *Analyze* group. *Assumption Charts* show the sampled random variates superimposed over the theoretical distributions chosen as assumptions. They allow you to compare the effects of different settings, such as number of trials, on the simulated values. Larger samples generate smoother curves that conform more closely to the theoretical distributions. The *Forecast Charts* option allows you to open or close selected forecast charts.

An important reason for using simulation for risk analysis is the ability to conduct sensitivity analyses to understand the impacts of individual variables or their distributional assumptions on forecasts. A somewhat naïve way to investigate the impact of assumptions on forecast cells is to freeze, or hold, certain assumptions constant in the model and compare the results with a base case simulation. The *Freeze* command in the *Define* group allows you to temporarily disable certain assumptions from a simulation and conduct this type of sensitivity analysis.

The uncertainty in a forecast is the result of the combined effect of the uncertainties of all assumptions as well as the formulas used in the model. An assumption might have a high degree of uncertainty yet have little effect on the forecast because it is not weighted heavily in the model formulas. For instance, a spreadsheet formula for a forecast might be defined as:

$$0.9(\text{Assumption 1}) + 0.1(\text{Assumption 2})$$

In the model, the forecast is nine times as sensitive to changes in the value of Assumption 1 as it is to changes in the value of Assumption 2. Thus, even if Assumption 2 has a much higher degree of uncertainty, as specified by the variance of its probability distribution, it would have a relatively minor effect on the uncertainty of the forecast.

The *Sensitivity Chart* feature allows you to determine the influence that each assumption has individually on a forecast. The sensitivity chart displays the rankings of each assumption according to their impact on a forecast cell as a bar chart. See Appendix 10.1B, *Sensitivity Charts*. A sensitivity chart provides three benefits:

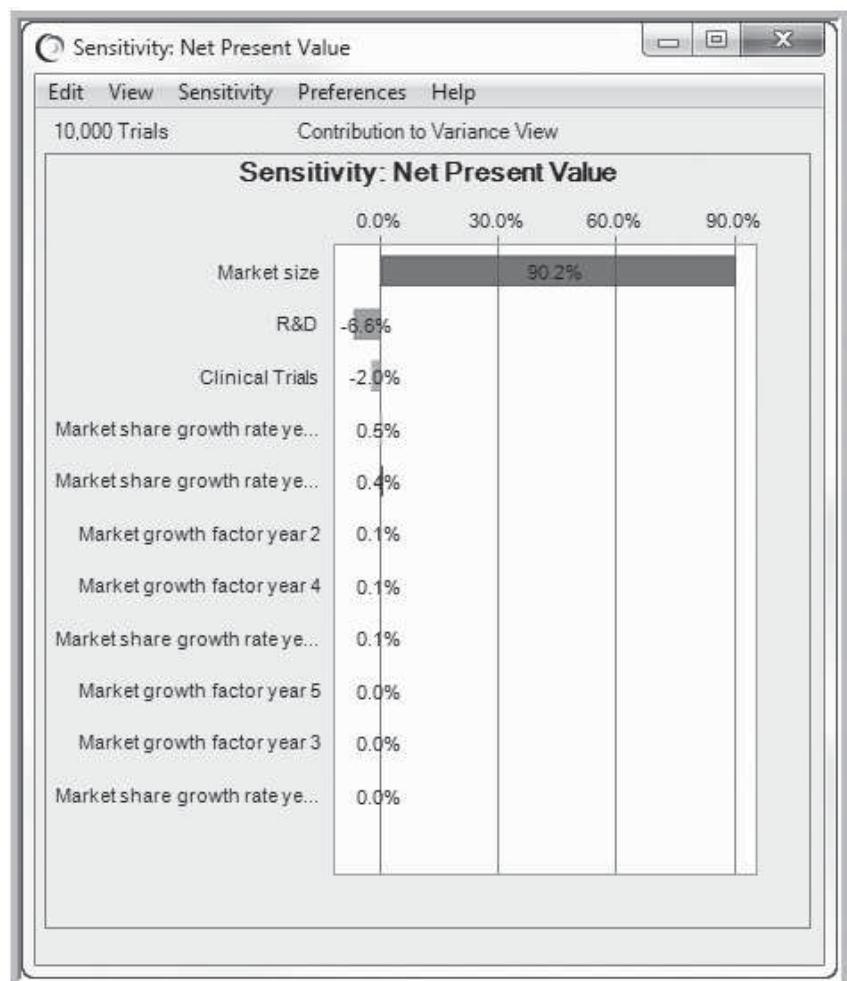
1. It tells which assumptions are influencing forecasts the most and which need better estimates.
2. It tells which assumptions are influencing forecasts the least and can be ignored or discarded altogether.
3. By understanding how assumptions affect your model, you can develop more realistic spreadsheet models and improve the accuracy of your results.

A *Contribution to Variance* sensitivity chart for the example is shown in Figure 10.16. The assumptions are ranked from top to bottom, beginning with the assumption having the highest sensitivity. Positive values indicate a direct relationship between the assumption and forecast, while negative values reflect an inverse relationship. The percentages represent the contribution that each assumption has on the variance of the forecast. For example, we see that the market size assumption accounts for about 90% of the variance; the R&D cost assumption accounts for about 6.6%, and the clinical trial cost assumption for about 2%. The other assumptions have a negligible effect. This means that if you want to reduce the variability in the forecast the most, you would need to obtain better information about the estimated market size and use an assumption that has a smaller variance.

If a simulation has multiple related forecasts, the *Overlay Charts* feature allows you to superimpose the frequency data from selected forecasts on one chart in order to compare differences and similarities that might not be apparent. You may select the forecasts



Spreadsheet Note

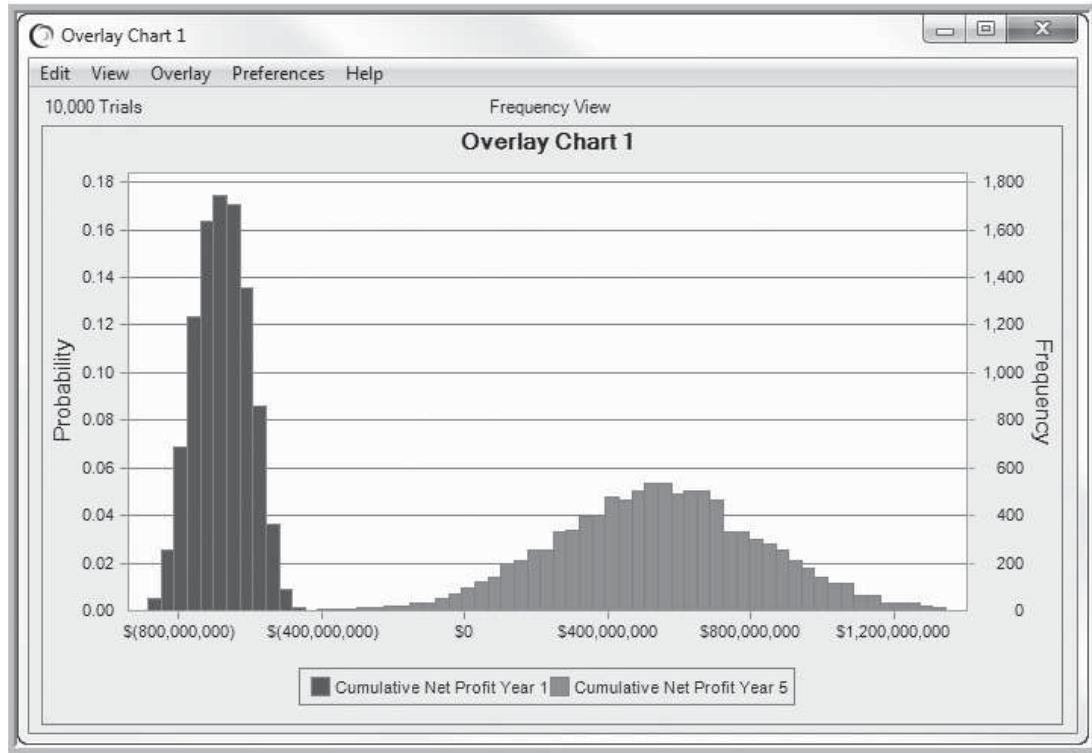


**FIGURE 10.16** Sensitivity Chart for Net Present Value

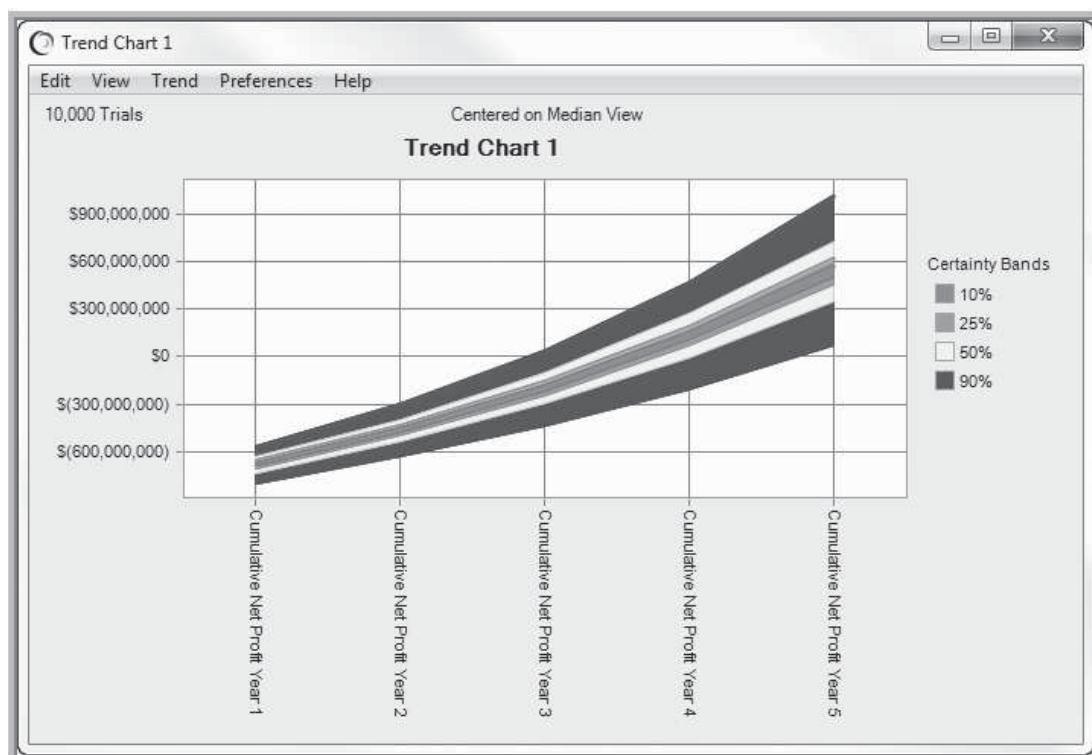
that you wish to display from the *Choose Forecasts* dialog that appears when creating a new chart. Figure 10.17 shows an overlay chart for the distributions of cumulative net profit for years 1 and 5. This chart makes it clear that the variance in year 5 is much larger than that in year 1. In the overlay chart, you may also view comparative statistics and percentiles from the *View* menu, and fit probability distributions to the forecasts from the *Overlay* menu.

If a simulation has multiple forecasts that are related to one another (such as over time), you can view the certainty ranges of all forecasts on a single chart, called a *Trend Chart*, which can be selected from the *View Charts* menu. Figure 10.18 shows a trend chart for the cumulative net profits over the five years. The trend chart displays certainty ranges in a series of patterned bands centered on the medians. For example, the band representing the 90% certainty range shows the range of values into which a forecast has a 90% chance of falling. From the trend chart in Figure 10.18, we see that although the median net profit increases over time, so does the variation, indicating that the uncertainty in forecasting the future also increases with time.

Finally, *Crystal Ball* can create *Scatter Charts*, which show correlations, dependencies, and other relationships between pairs of variables (forecasts and/or assumptions) plotted against each other. See the *Help* files for further information.



**FIGURE 10.17** Overlay Chart for Year 1 and Year 5 Cumulative Net Profit



**FIGURE 10.18** Trend Chart for Cumulative Net Profit Over Five Years

## **Crystal Ball Reports and Data Extraction**

The *Analyze* group in the menu bar has two other options: *Create Report* and *Extract Data*. *Create Report* allows you to build customized reports for all or a subset of assumptions and output information that we described. The (default) full report contains a summary of run preferences and run statistics, the most recent views of forecast charts that you may have analyzed, forecast statistics and percentiles, definitions of all assumptions, and any other charts that you may have created, such as sensitivity, overlay, etc. The report can be created in the current or new workbook from the *Options* tab of the *Create Report Preferences* dialog. You also have the option to display an image of the charts, or display them as Excel charts, which allows you to edit them using Excel chart commands. The report data may be edited, copied, and pasted into Word documents, or printed.

In addition, *Crystal Ball* allows you to extract selected data to an Excel worksheet for further analysis. In the dialog box that appears after clicking on *Extract Data* from the *Analyze* group, you may select various types of data to extract:

- Statistics
- Percentiles
- Chart Bins—the intervals in the forecast chart along with their probability and frequency of occurrence
- Sensitivity Data—sensitivity data for all pairs of assumptions and forecasts
- Trial Values—the generated assumption and forecast values for each simulation trial (be careful with this one, especially if you use 10,000 trials!)

## **Crystal Ball Functions and Tools**

A very useful feature of *Crystal Ball* is the ability to use its functions within Excel. A complete list may be found by clicking the *Insert Function* button ( $f_x$ ) on the Excel menu bar and selecting the *Crystal Ball* category. The random variate generation functions (CB.niform, CB.normal, CB.exponential, and so on) may be used in a spreadsheet model instead of defining an assumption in the usual way (however, this restricts the ability to use assumptions in sensitivity charts, output reports, and other features of *Crystal Ball*). These functions are often useful to embed within complex Excel formulas. We will see examples of using these functions in some of the applications later in this chapter.

*Crystal Ball* has a set of modeling and analysis tools to complement the basic simulation capability. These can be found in the *Tools* menu under the *Run* group on the toolbar. We will briefly summarize them here and illustrate several of them in applications that follow.

**BATCH FIT** The *Batch Fit* tool fits probability distributions to multiple data series. The advantage of this tool is that it eliminates the necessity to fit each distribution individually. The only requirement is that the data must be in adjacent rows or columns.

**CORRELATION MATRIX** In *Crystal Ball* each random input variable is assumed to be independent of the others; that is, random variates are sampled independently from each assumption distribution during the course of the simulation. In many situations, this is not realistic because assumptions would naturally be related to one another, and we would want to explicitly model such dependencies between variables. The *Correlation Matrix* tool allows you to define correlations between groups of assumptions in a model.

**TORNADO CHART** The *Tornado Chart* tool provides *a priori* sensitivity information about the impact of each model variable on a target forecast. The tornado chart differs from the sensitivity chart in that it tests each assumption independently while freezing the

other variables at their base values. This is useful in quickly prescreening variables in a model to determine which variables are the most important candidates to define as assumptions before building a simulation model. Those having little effect on a forecast might not need to be defined as assumptions and kept as constants in the model, thus simplifying it.

**BOOTSTRAP TOOL** The classical approach for confidence intervals assumes that the sampling distribution of the mean is normal. However, if the sampling distribution is not normally distributed, such a confidence interval is not valid. Also, if we wanted to develop a confidence interval for the median, standard deviation, or maximum forecast value, for example, we may not know the sampling distribution of these statistics. A statistical technique called **bootstrapping** analyzes sample statistics empirically by repeatedly sampling the data and creating distributions of the statistics. This approach allows you to estimate the sampling distribution of any statistic, even an unconventional one such as the minimum or maximum endpoint of a forecast. The *Bootstrap* tool does this.

**DECISION TABLE** The *Decision Table* tool runs multiple simulations to test different values for one or two decision variables. We will illustrate this tool for an example later in this chapter.

**SCENARIO ANALYSIS** The *Scenario Analysis* tool runs a simulation and then sorts and matches all the resulting values of a target forecast with their corresponding assumption values. This allows you to investigate which combination of assumption values gives you a particular result.

**TWO-DIMENSIONAL SIMULATION** The *Two-dimensional Simulation* tool allows you to distinguish between uncertainty in assumptions due to limited information or to data and variability—assumptions that change because they describe a population with different values. Theoretically, you can eliminate uncertainty by gathering more information; practically, it is usually impossible or cost prohibitive. Variability is inherent in the system, and you cannot eliminate it by gathering more information. Separating these two types of assumptions lets you more accurately characterize risk. The tool runs an outer loop to simulate the uncertainty values and then freezes them while it runs an inner loop to simulate variability. The process repeats for some small number of outer simulations, providing a portrait of how the forecast distribution varies due to the uncertainty.

## APPLICATIONS OF MONTE CARLO SIMULATION AND CRYSTAL BALL FEATURES

In this section, we present several additional examples of Monte Carlo simulation using *Crystal Ball*. These serve to illustrate the wide range of applications to which the approach may be applied and also various features of *Crystal Ball*.

### Newsvendor Model: Fitting Input Distributions, *Decision Table* Tool, and Custom Distribution

In Chapter 9, we developed the newsvendor model to analyze a single-period purchase decision. Here we will apply *Crystal Ball* to forecast the profitability of different purchase quantities when the future demand is uncertain. We will illustrate how to incorporate the *Crystal Ball* distribution fitting and *Decision Table* tools into the analysis.

Let us suppose that the store owner kept records for the past 20 years on the number of boxes sold, as shown below:

20 Years of Historical Candy Sales			
42	43	47	43
45	46	41	45
40	42	41	42
46	44	45	44
43	43	51	48

From what probability distribution could these data have been drawn? *Crystal Ball* provides a tool for identifying the best-fitting probability distribution for a set of data. The basis for fitting a data to a probability distribution is a statistical test of hypothesis called **goodness-of-fit**. Goodness-of-fit tests provide statistical evidence to test hypotheses about the *nature* of the distribution. Three methods commonly used are the chi-square test, which is similar in nature to the approach used in the test for independence of categorical variables; Anderson–Darling test; and Kolmogorov–Smirnov test. These approaches test the hypotheses:

**H<sub>0</sub>:** *the sample data come from a specified distribution (e.g., normal)*

**H<sub>1</sub>:** *the sample data do not come from the specified distribution*

As with any hypothesis test, you can disprove the null hypothesis but cannot statistically *prove* that data come from the specified distribution. However, if you cannot reject the null, then you at least have some measure of faith that the hypothesized distribution fits the data rather well. The chi-square goodness-of-fit test breaks down the hypothesized distribution into areas of equal probability and compares the data points within each area to the number that would be expected for that distribution. The Kolmogorov–Smirnov test compares the cumulative distribution of the data with the theoretical distribution and bases its conclusion on the largest vertical distance between them. The Anderson–Darling method is similar, but puts more weight on the differences between the tails of the distributions.

**FITTING A DISTRIBUTION WITH CRYSTAL BALL** Figure 10.19 shows the newsvendor model spreadsheet (Excel file *Newsvendor CB Model*) to which we have added the historical sales data. The distribution of sales seems to be some type of positively-skewed unimodal distribution. We can have *Crystal Ball* automatically define an assumption cell for demand based on the best-fitting probability distribution (see Appendix 10.1C, *Distribution Fitting with Crystal Ball*). First, click on cell B11, then *Define Assumption*, and then the *Fit* button. Proceed to fit the data as described in Appendix 10.1C. When the *Comparison Chart* pops up identifying the Negative Binomial distribution as the best fit, click the *Accept* button. This will display the *Define Assumption* dialog for cell B11 with the proper parameters for the chosen distribution automatically entered (see Figure 10.20). Click *OK* and the demand assumption will be defined. Next, define the profit in cell B17 as a forecast. You may now run *Crystal Ball*.

The forecast chart, shown in Figure 10.21, looks somewhat odd. However, recall that if demand exceeds the purchase quantity, then sales are limited to the number purchased; hence, the large spike at the right of the distribution. Although we could repeat the simulation for different values of purchase quantity, we will use the *Decision Table* tool to automatically run simulations for a range of values.

**USING THE DECISION TABLE TOOL** First, define the purchase quantity in cell B12 as a decision variable in *Crystal Ball*. Click on cell B12 and then on *Define Decision* in the



Spreadsheet Note

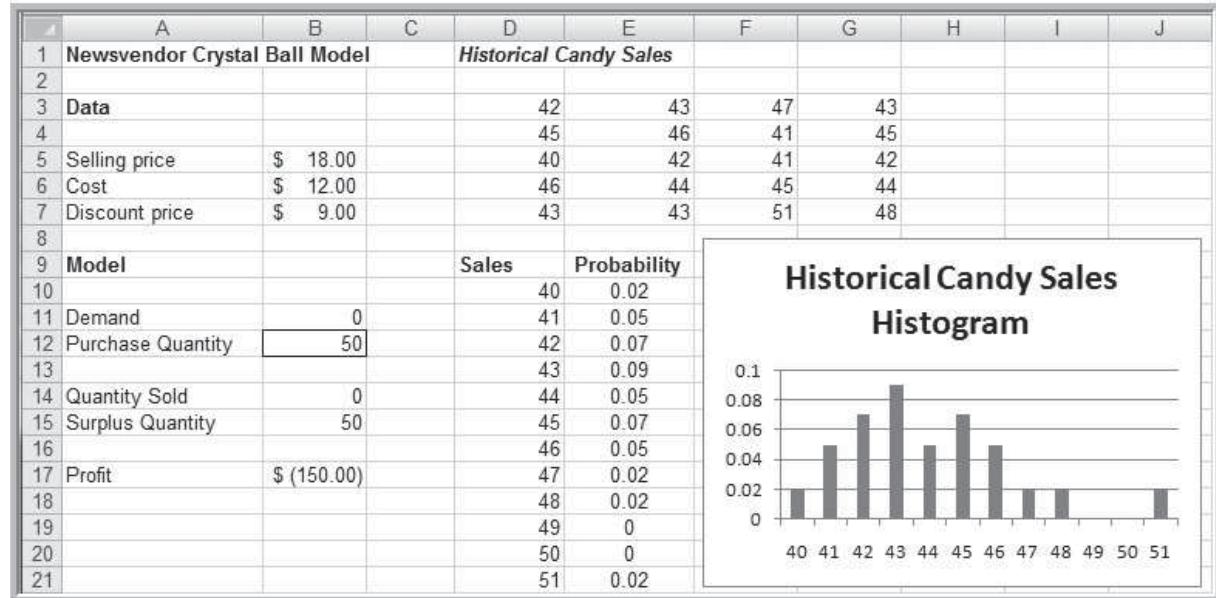


FIGURE 10.19 Newsvendor Crystal Ball Model Spreadsheet

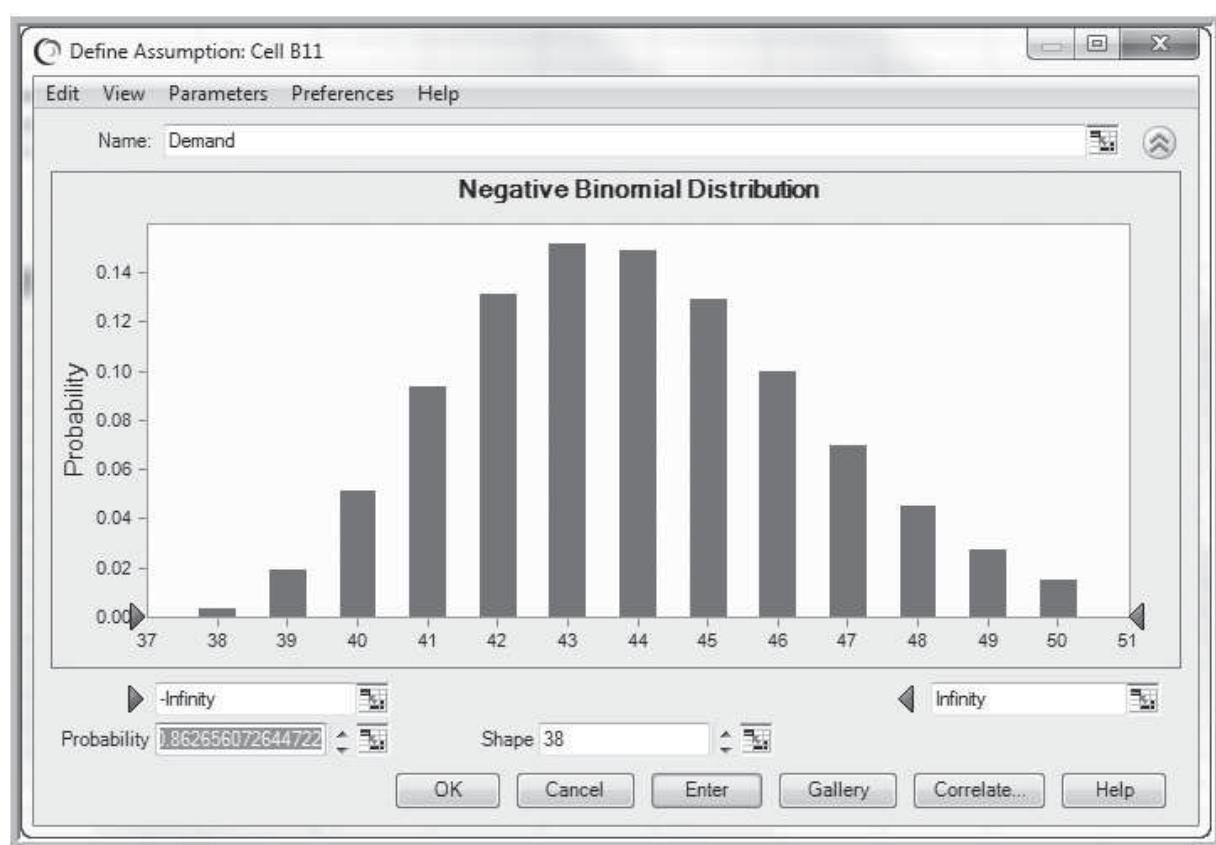
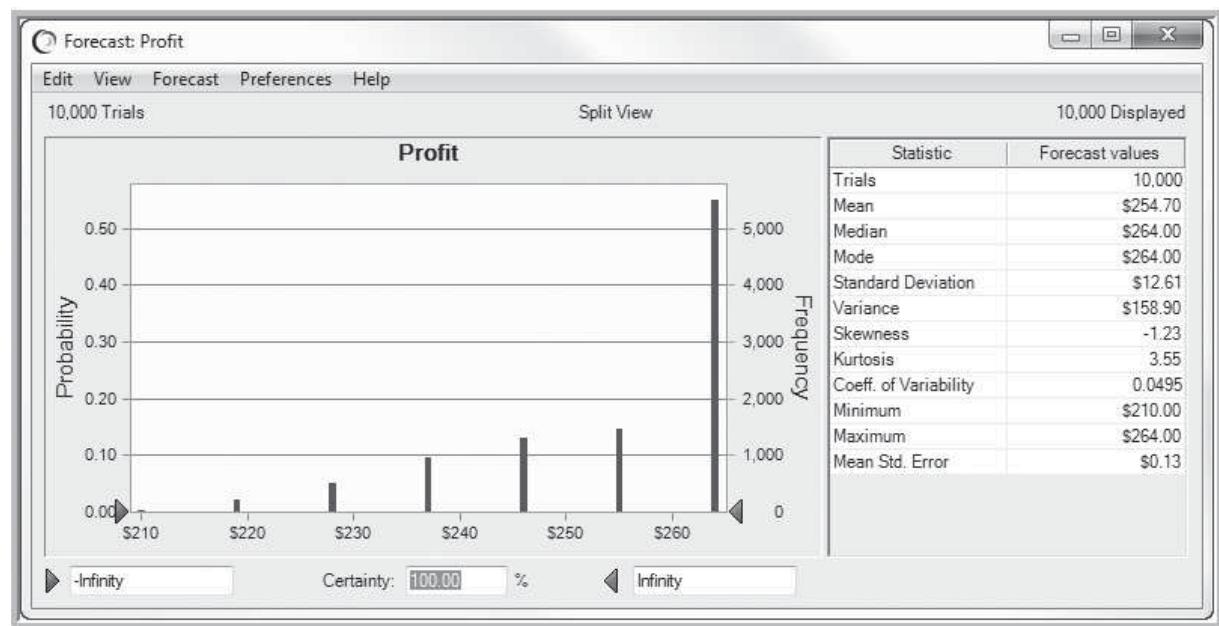
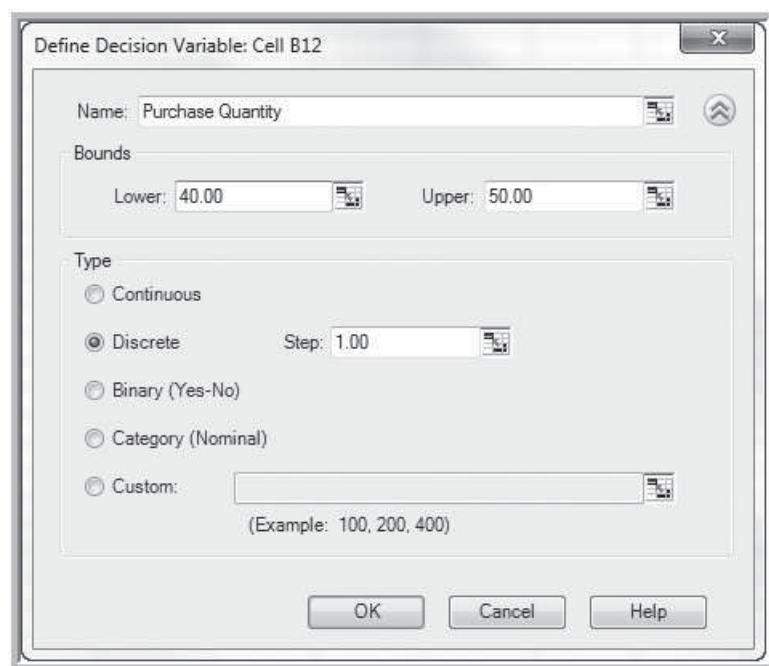


FIGURE 10.20 Best-Fitting Distributional Assumption



**FIGURE 10.21** Newsvendor Simulation Results for Purchase Quantity = 44



**FIGURE 10.22** Define Decision Variable Dialog

Define group. Enter 40 in the *Lower* bound box, and 50 in the *Upper* bound box. Because this should be a whole number, choose *Discrete* with a *Step* size of 1 under *Type* as shown in Figure 10.22.

Next, select the *Decision Table* tool from the *More Tools* menu in the *Tools* group. The first two dialogs that appear ask you to select the forecast and decision variable to evaluate; in this example, we have only one option for each to choose. In the third (*Options*)

A	B	C	D	E	F	G	H	I	J	K	L
	Trend Chart										
	Overlay Chart										
	Forecast Chart										
1	Purchase Quantity (40.00)										
2	\$239.76	\$245.06	\$249.58	\$252.95	\$254.97	\$255.67	\$255.21	\$253.86	\$251.75	\$249.23	\$246.49
3	1	2	3	4	5	6	7	8	9	10	11

**FIGURE 10.23** Decision Table Tool Results

dialog, set the number of trials for each simulation and click the *Run* button. The tool will run a simulation for each of the 11 decision variable values defined.

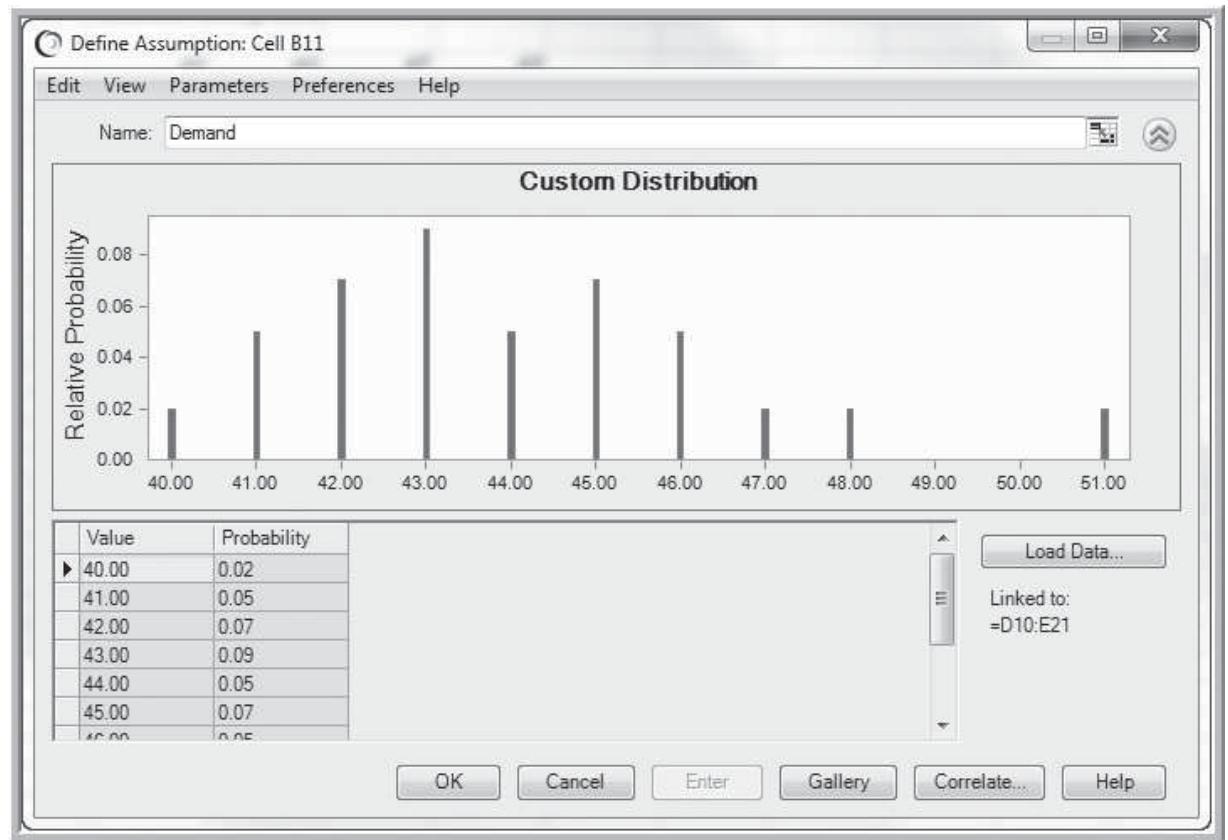
When the simulations are completed, a new worksheet will be displayed that shows the mean values of profit for each value of the decision variable (Figure 10.23). By selecting one of these cells and clicking on the *Forecast Charts* button, you can display the forecast chart for that simulation run. You may also select a range of cells and display trend or overlay charts. We see that a purchase quantity of 45 provides the largest average profit. The *Decision Table* tool is useful whenever you wish to identify the best values of a decision variable. You may also set up and run a two-dimensional decision table that evaluates all combinations of two decision variables.

### SKILL-BUILDER EXERCISE 10.3

Open the Excel File *Newsvendor CB Model* and work through the process of fitting a distribution to the data and using the *Decision Table* tool to evaluate each purchase quantity.

**THE CUSTOM DISTRIBUTION** An alternative approach to sampling from a fitted distribution is to sample from the empirical distribution of the data. We may use the custom distribution in *Crystal Ball* to define an assumption based on this probability distribution. The custom distribution can be used to model any type of discrete probability function that has a series of single values, discrete ranges, or continuous ranges. To define the demand in cell B11 as an assumption having this distribution, choose *Custom* from the *Distribution Gallery*. You may select the different options from the *Parameters* menu, with the default being *Weighted Values* (i.e., a discrete probability distribution). You can also create distributions that have continuous ranges associated with discrete probabilities. To do this, choose *Continuous Ranges* from the *Parameters* menu in the *Custom Distribution* dialog box. Other options are also available; see the *Help* files for further information.

With the custom distribution, you may either enter the value and probability for each outcome manually in the appropriate fields in the dialog or load the data from a worksheet, which is typically easier to do. To do this, click on the *Load* button. If it does not appear in the dialog, click the *More* button (the downward pointing arrows) to the right of the name field in the dialog; this will expand the dialog. *Crystal Ball* will prompt you for the location of the data; simply enter the cell references of the data range (in this case, the range D10:E21). The range of the data must be so that the value column is first, and the probabilities are to the right. The result is shown in Figure 10.24.



**FIGURE 10.24** Completed Custom Distribution Example

While sampling from empirical data is easy to do, it does have some drawbacks. First, the empirical data may not adequately represent the true underlying population because of sampling error. Second, using an empirical distribution precludes sampling values outside the range of the actual data. Therefore, it is usually advisable to fit a distribution and use it for the assumption.

#### SKILL-BUILDER EXERCISE 10.4

Use the historical candy sales data to define a custom distribution assumption for demand in the newsvendor model. Run the *Crystal Ball* simulation and compare your results with the previous exercise.

#### Overbooking Model: *Crystal Ball* Functions

In Chapter 9, we developed a simple model for overbooking decisions (*Hotel Overbooking Model*). In any realistic overbooking situation, the actual customer demand as well as the number of cancellations would be random variables. We will illustrate how a simulation model can help in making the best overbooking decision. We will also describe how to use *Crystal Ball* functions in the model rather than using the *Define Assumption* option.

Let us assume that the demand can be modeled by a Poisson distribution with a mean of 320; that is, an average of 320 customers call to request a reservation. In cell B13 (see Figure 10.25), we define the customer demand using the *Crystal Ball* function *CB*.

	A	B
1	Hotel Overbooking Model	
2		
3	Data	
4		
5	Rooms available	300
6	Price	\$120
7	Overbooking cost	\$100
8		
9	Model	
10		
11	Reservation limit	310
12	Customer demand	330
13	Reservations made	310
14	Cancellations	6
15	Customer arrivals	304
16	Overbooked customers	4
17		
18	Net revenue	\$35,600

	A	B
1	Hotel Overbooking Model	
2		
3	Data	
4		
5	Rooms available	300
6	Price	120
7	Overbooking cost	100
8		
9	Model	
10		
11	Reservation limit	310
12	Customer demand	=CB.Poisson(320)
13	Reservations made	=MIN(B11,B12)
14	Cancellations	=CB.Binomial(0.04,B13)
15	Customer arrivals	=B13-B14
16	Overbooked customers	=MAX(0,B15-B5)
17		
18	Net revenue	=MIN(B15,B5)*B6-B16*B7

**FIGURE 10.25** Crystal Ball Model for the Hotel Overbooking Example (*Hotel Overbooking CB Model*)

*Poisson(320)*. We will assume that each reservation has a constant probability  $p = 0.04$  of being cancelled; therefore, the number of cancellations (cell B15) can be modeled using a binomial distribution with  $n =$  number of reservations made and  $p =$  probability of cancellation. We may use the *Crystal Ball* function *CB.Binomial(0.04, B13)* to do this. Note that we are referencing cell B13 in this function. This is critical in this example, because the number of reservations made will change, depending on the customer demand in cell B12. When such functions are used, *Crystal Ball* has no defined assumption cells; however, whenever the spreadsheet is recalculated, the values in these cells will change, just as if we were using the Excel *RAND* function. We need only define cells B16 and B18 as forecast cells and run the simulation.

Figure 10.26 shows forecast charts for accepting 310 reservations. There is about a 15% chance of overbooking at least one customer. We may use the *Decision Table* tool to run simulations for a range of values of the number of reservations to accept and identify the best decision. The following exercise asks you to do this.

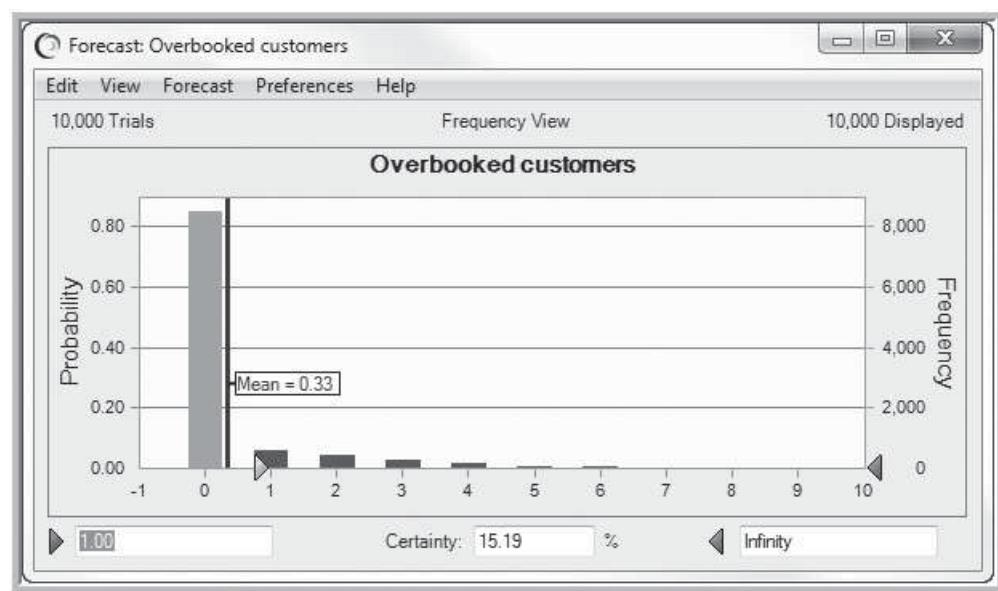
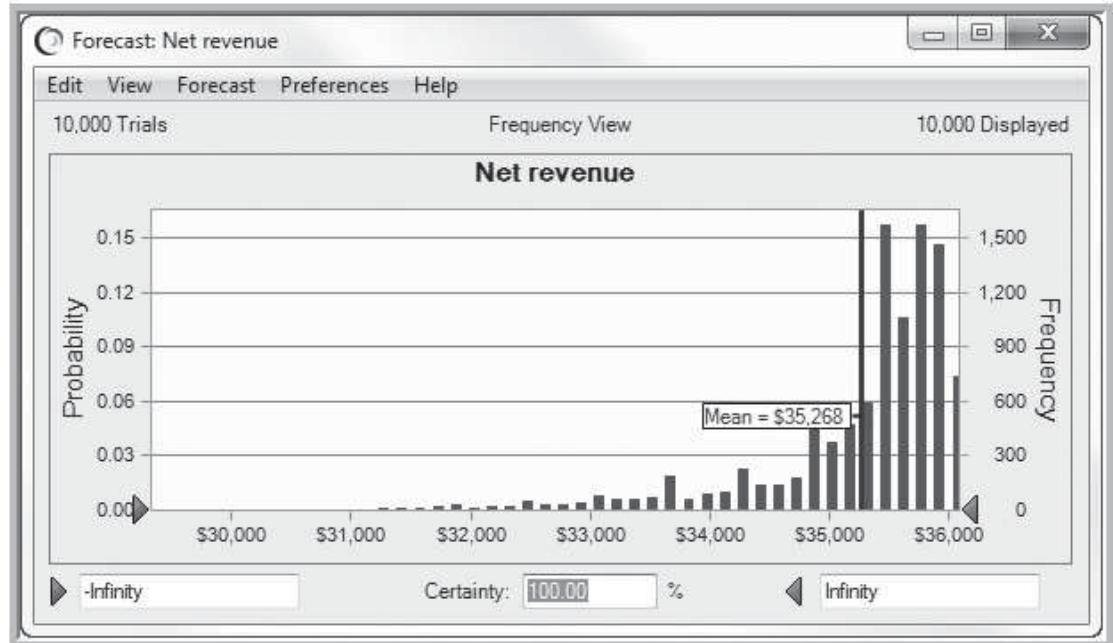
### SKILL-BUILDER EXERCISE 10.5

Open the Excel file *Hotel Overbooking CB Model*. Use the *Decision Table* tool to run simulations for accepting 300 to 320 reservations. What decision would you make?

### Cash Budgeting: Correlated Assumptions

Cash budgeting<sup>3</sup> is the process of projecting and summarizing a company's cash inflows and outflows expected during a planning horizon, usually 6 to 12 months. The cash budget also shows the monthly cash balances and any short-term borrowing used to cover cash shortfalls. Positive cash flows can increase cash, reduce outstanding loans, or be used elsewhere in the business; negative cash flows can reduce cash available or be

<sup>3</sup> Adapted from Douglas R. Emery, John D. Finnerty, and John D. Stowe, *Principles of Financial Management* (Upper Saddle River, NJ: Prentice Hall, 1998): 652–654.



**FIGURE 10.26** Hotel Overbooking Model Results

offset with additional borrowing. Most cash budgets are based on sales forecasts. With the inherent uncertainty in such forecasts, Monte Carlo simulation is an appropriate tool to analyze cash budgets.

Figure 10.27 shows an example of a cash budget spreadsheet (Excel file *Cash Budget*). The budget begins in April (thus, sales for April and subsequent months are uncertain). These are assumed to be normally distributed with a standard deviation of 10% of the mean. In addition, we assume that sales in adjacent months are correlated

A	B	C	D	E	F	G	H	I	J	K
1 Cash Budgeting										
2 Desired Minimum Balance	\$ 100,000									
3		February	March	April	May	June	July	August	September	October
4	Sales	\$ 400,000	\$ 500,000	\$ 600,000	\$ 700,000	\$ 800,000	\$ 800,000	\$ 700,000	\$ 600,000	\$ 500,000
5 Cash Receipts										
6 Collections (current)	20%			\$ 120,000	\$ 140,000	\$ 160,000	\$ 160,000	\$ 140,000	\$ 120,000	
7 Collections (previous month)	50%			\$ 250,000	\$ 300,000	\$ 350,000	\$ 400,000	\$ 400,000	\$ 350,000	
8 Collections (2nd month previous)	30%			\$ 120,000	\$ 150,000	\$ 180,000	\$ 210,000	\$ 240,000	\$ 240,000	
9 Total Cash Receipts				\$ 490,000	\$ 590,000	\$ 690,000	\$ 770,000	\$ 780,000	\$ 710,000	
10										
11 Cash Disbursements										
12 Purchases				\$ 420,000	\$ 480,000	\$ 480,000	\$ 420,000	\$ 360,000	\$ 300,000	
13 Wages and Salaries				\$ 72,000	\$ 84,000	\$ 96,000	\$ 96,000	\$ 84,000	\$ 72,000	
14 Rent				\$ 10,000	\$ 10,000	\$ 10,000	\$ 10,000	\$ 10,000	\$ 10,000	
15 Cash Operating Expenses				\$ 30,000	\$ 30,000	\$ 30,000	\$ 30,000	\$ 25,000	\$ 25,000	
16 Tax Installments				\$ 20,000			\$ 30,000			
17 Capital Expenditure						\$ 150,000				
18 Mortgage Payment					\$ 60,000					
19 Total Cash Disbursements				\$ 552,000	\$ 664,000	\$ 766,000	\$ 586,000	\$ 49,000	\$ 40,000	
20										
21 Ending Cash Balance										
22 Net Cash Flow				\$ (62,000)	\$ (74,000)	\$ (76,000)	\$ 184,000	\$ 301,000	\$ 303,000	
23 Beginning Cash Balance				\$ 150,000	\$ 100,000	\$ 100,000	\$ 100,000	\$ 122,000	\$ 423,000	
24 Available Balance				\$ 88,000	\$ 26,000	\$ 24,000	\$ 284,000	\$ 423,000	\$ 726,000	
25 Monthly Borrowing				\$ 12,000	\$ 74,000	\$ 76,000	\$ -	\$ -	\$ -	
26 Monthly Repayment				\$ -	\$ -	\$ -	\$ 162,000	\$ -	\$ -	
27 Ending Cash Balance				\$ 150,000	\$ 100,000	\$ 100,000	\$ 122,000	\$ 423,000	\$ 726,000	
28 Cumulative Loan Balance				\$ -	\$ 12,000	\$ 86,000	\$ 162,000	\$ -	\$ -	

**FIGURE 10.27** Cash Budget Model

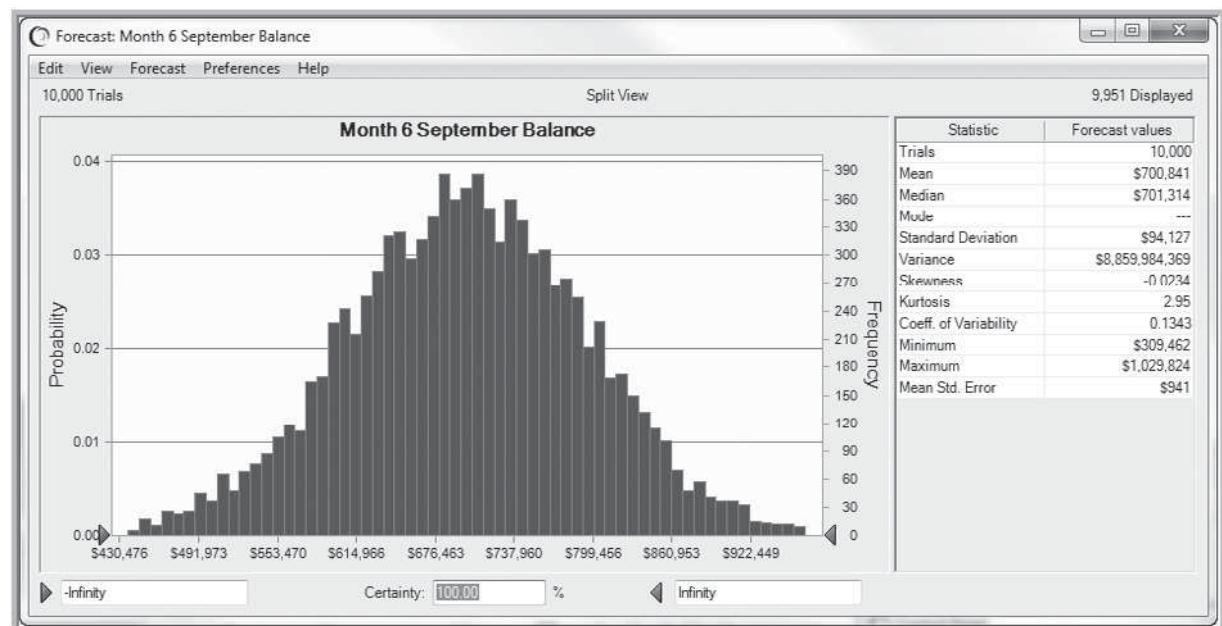
with one another, with a correlation coefficient of 0.6. On average, 20% of sales is collected in the month of sale, 50% in the month following the sale, and 30% in the second month following the sale. However, these figures are uncertain, so a uniform distribution is used to model the first two values (15% to 20% and 40% to 50%, respectively) with the assumption that all remaining revenues are collected in the second month following the sale. Purchases are 60% of sales and are paid for one month prior to the sale. Wages and salaries are 12% of sales and are paid in the same month as the sale. Rent of \$10,000 is paid each month. Additional cash operating expenses of \$30,000 per month will be incurred for April through July, decreasing to \$25,000 for August and September. Tax payments of \$20,000 and \$30,000 are expected in April and July, respectively. A capital expenditure of \$150,000 will occur in June, and the company has a mortgage payment of \$60,000 in May. The cash balance at the end of March is \$150,000, and managers want to maintain a minimum balance of \$100,000 at all times. The company will borrow the amounts necessary to ensure that the minimum balance is achieved. Any cash above the minimum will be used to pay off any loan balance until it is eliminated. The available cash balances in row 24 of the spreadsheet are the *Crystal Ball* forecast cells.

*Crystal Ball* allows you to specify correlation coefficients to define dependencies between assumptions. *Crystal Ball* uses the correlation coefficients to rearrange the generated random variates to produce the desired correlations. This can be done only after assumptions have been defined. For example, in the cash budget model, if the sales in April are high, then it would make sense that the sales in May would be high also. Thus, we might expect a positive correlation between these variables.

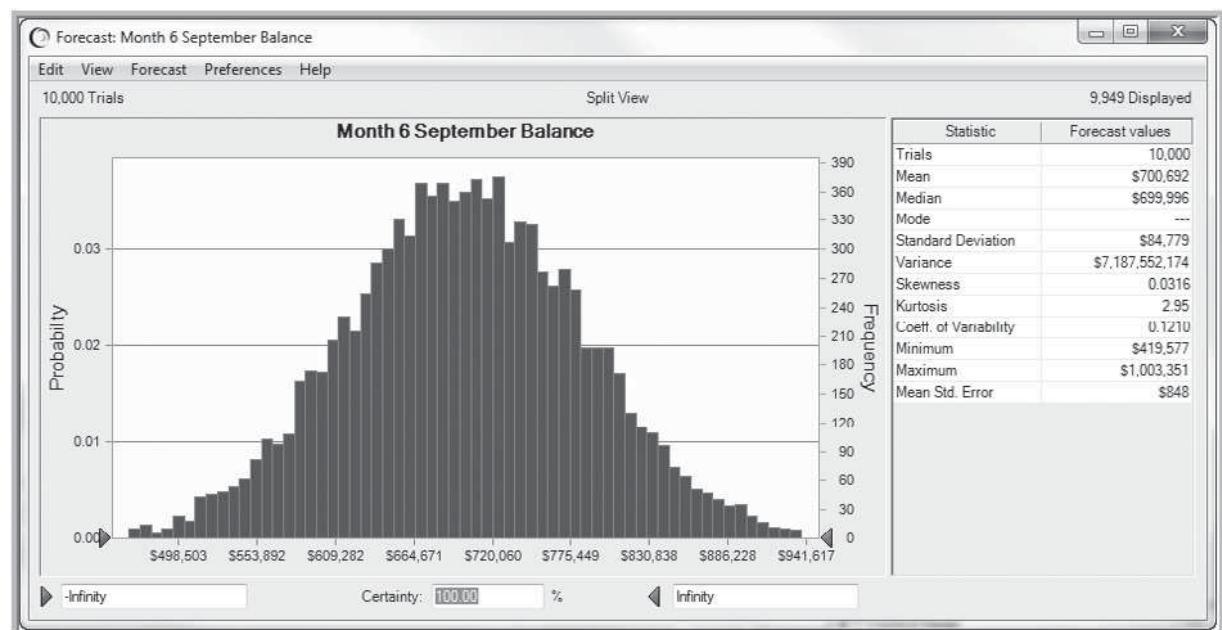
The *Crystal Ball Correlation Matrix* tool is used to specify the correlations between sales assumptions (see Appendix 10.1D, *Correlation Matrix Tool*). Let us assume that the correlation between sales in successive months is 0.6. Figures 10.28 and 10.29 show the forecast chart and statistics for the September cash balances without and with correlated assumptions. The key difference is that the variance of the forecast with correlated assumptions is smaller, thus providing more accuracy in the estimate.



Spreadsheet Note



**FIGURE 10.28** September Cash Balance Forecast Chart—Uncorrelated Assumptions

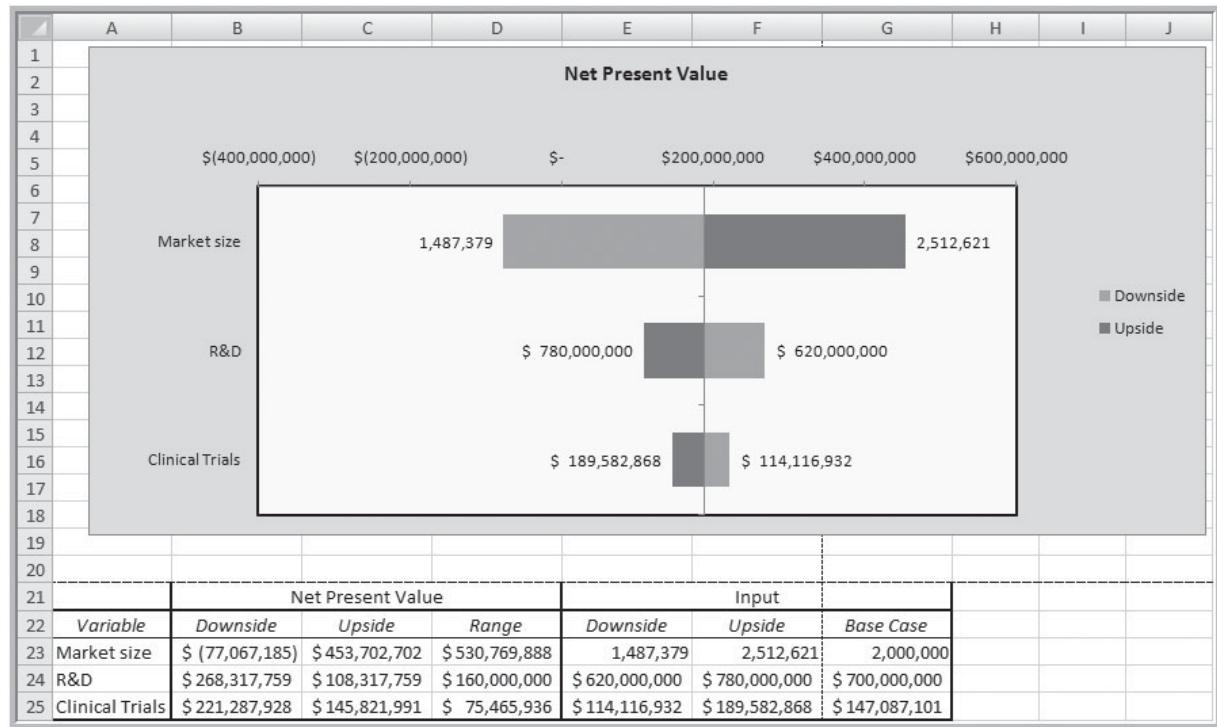


**FIGURE 10.29** September Cash Balance Forecast Chart—Correlated Assumptions

### New Product Introduction: **Tornado Chart Tool**

Crystal Ball has a tool that can analyze the sensitivity of each assumption or other model inputs to the forecast prior to conducting a simulation, the *Tornado Chart* tool (see Appendix 10.1E, *Tornado Charts*). The *Tornado Chart* tool is useful for measuring the sensitivity of variables that you have defined in Crystal Ball or quickly prescreening the variables in your model to determine which ones are good candidates to define as





**FIGURE 10.30** Tornado Chart for Moore Pharmaceuticals

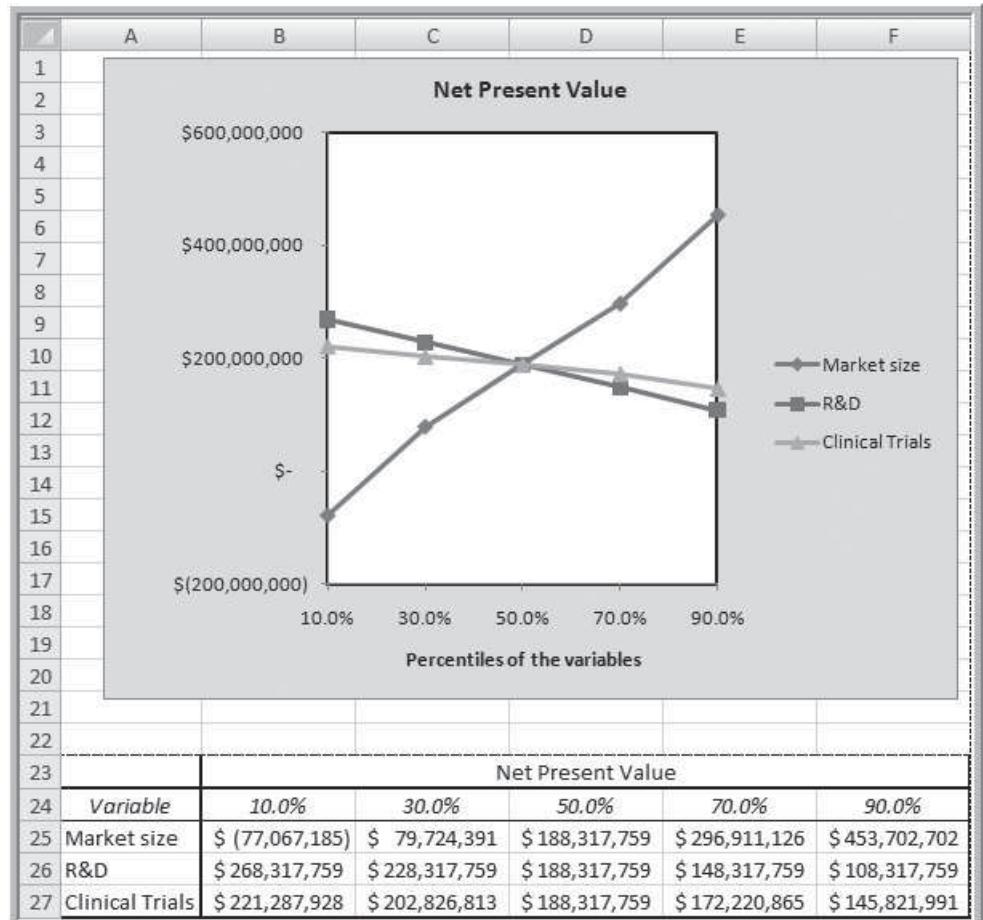
assumptions or decision variables. For example, suppose that we wanted to understand the impact of the market size, R&D cost, and clinical trials cost assumptions on the NPV forecast prior to running the simulation. Select these assumptions in the *Tornado Chart* tool dialog and run the tool. The tool creates two charts: a tornado chart and a spider chart. Figure 10.30 shows the tornado chart and data table that is created. In the data table, the tool examines values  $\pm 10\%$  away from the base case and evaluates the NPV forecast while holding all other variables constant. We see that the market size has the most impact while clinical trials cost has the least. The spider chart in Figure 10.31 shows these results as rates of change; the larger the slope, the higher the impact. These charts show similar insights as the sensitivity chart; however, running this before running the *Crystal Ball* model can help screen out assumptions that would have little impact on the forecast (such as the growth factors and growth rates as we saw earlier) before conducting an in-depth analysis. For models with large numbers of potential assumptions, this is a useful tool.

### SKILL-BUILDER EXERCISE 10.6

For the Moore Pharmaceuticals model, run the *Tornado Chart* tool for all assumptions and the NPV forecast. Compare the results to the sensitivity chart in Figure 10.16.

### Project Management: Alternate Input Parameters and the *Bootstrap* Tool

In Chapter 9, we developed a project management spreadsheet model for scheduling activities and computing the critical path. For many real projects, activity times are probabilistic, and we often assume that they have a beta or triangular distribution, especially



**FIGURE 10.31** Spider Chart for Moore Pharmaceuticals

when times are estimated judgmentally. Analytical methods, such as the Program Evaluation and Review Technique (PERT), allow us to determine probabilities of project completion times by assuming that the expected activity times define the critical path and invoking the central limit theorem to make an assumption of normality of the distribution of project completion time. However, this assumption may not always be valid. Simulation can provide a more realistic characterization of the project completion time and the associated risks.

We will illustrate this with the Becker Consulting project management model used in Chapter 9. Suppose that the Information Systems manager has determined the most likely time for each activity but, recognizing the uncertainty in the times to complete each task, has estimated the 10th and 90th percentiles for the activity times. Setting the parameters in this way is a common approach for estimating the distribution, since managers typically cannot estimate the absolute minimum or maximum times but can reasonably determine a time that might be met or exceeded 10% of the time. These are shown in Table 10.1. With only these estimates, a triangular distribution is an appropriate assumption. Note that the times for activities A, G, M, P, and Q are constant.

Figure 10.32 shows a spreadsheet model designed to simulate the project completion time when the activity times are uncertain (Excel file *Becker Consulting Project Management*

**TABLE 10.1** Uncertain Activity Time Data

Activity	Predecessors	10th Percentile	Most Likely	90th Percentile
A	Select steering committee	—	15	15
B	Develop requirements list	—	40	45
C	Develop system size estimates	—	10	14
D	Determine prospective vendors	—	2	3
E	Form evaluation team	A	5	7
F	Issue request for proposal	B,C,D,E	4	5
G	Bidders conference	F	1	1
H	Review submissions	G	25	30
I	Select vendor short list	H	3	5
J	Check vendor references	I	3	7
K	Vendor demonstrations	I	20	30
L	User site visit	I	3	4
M	Select vendor	J,K,L	3	3
N	Volume sensitive test	M	10	13
O	Negotiate contracts	M	10	14
P	Cost-benefit analysis	N,O	2	2
Q	Obtain board of directors approval	P	5	5

	A	B	C	D	E	F	G	H	I	J	K
1	Becker Consulting Project Management Simulation Model										
2		10th	Most	90th	Activity	Early	Early	Latest	Latest		
3	Activity	Percentile	Likely	Percentile	Time	Start	Finish	Start	Finish	Slack	On Critical Path?
5	A	15	15	15	15.00	0.00	15.00	27.32	42.32	27.32	0
6	B	40	45	60	49.32	0.00	49.32	0.00	49.32	0.00	1
7	C	10	14	30	19.19	0.00	19.19	30.12	49.32	30.12	0
8	D	2	3	5	3.43	0.00	3.43	45.88	49.32	45.88	0
9	E	5	7	9	7.00	15.00	22.00	42.32	49.32	27.32	0
10	F	4	5	8	5.86	49.32	55.18	49.32	55.18	0.00	1
11	G	1	1	1	1.00	55.18	56.18	55.18	56.18	0.00	1
12	H	25	30	50	36.49	56.18	92.67	56.18	92.67	0.00	1
13	I	3	5	10	6.29	92.67	98.96	92.67	98.96	0.00	1
14	J	3	7	10	6.57	98.96	105.53	124.54	131.11	25.58	0
15	K	20	30	45	32.15	98.96	131.11	98.96	131.11	0.00	1
16	L	3	4	5	4.00	98.96	102.96	127.11	131.11	28.15	0
17	M	3	3	3	3.00	131.11	134.11	131.11	134.11	0.00	1
18	N	10	13	20	14.72	134.11	148.83	137.71	152.43	3.60	0
19	O	10	14	28	18.32	134.11	152.43	134.11	152.43	0.00	1
20	P	2	2	2	2.00	152.43	154.43	152.43	154.43	0.00	1
21	Q	5	5	5	5.00	154.43	159.43	154.43	159.43	0.00	1
22											
23					Project completion time		159.43				

**FIGURE 10.32** Crystal Ball Model for Becker Consulting Example

*Simulation Model).* For those activity times that are not constant, we define the cell for the activity time as a *Crystal Ball* assumption using the triangular distribution. After selecting the triangular distribution in the *Crystal Ball* gallery, click on the *Parameters* menu in the dialog box. This provides a list of alternative ways to input the data. Select the *10th percentile, likeliest, and 90th percentile* option to input the activity time parameters.

In the analytical approach found in most textbooks, probabilities of completing the project within a certain time are computed assuming the following:

1. The distribution of project completion times is normal (by applying the central limit theorem).
2. The expected project completion time is the sum of the expected activity times along the critical path, which is found using the expected activity times.
3. The variance of the distribution is the sum of the variances of those activities along the critical path, which is found using the expected activity times. If more than one critical path exists, use the path with the largest variance.

In the *Define Assumption* dialog, selecting *Minimum*, *Likeliest*, *Maximum* from the *Parameters* menu will convert the percentile input data to the standard triangular distribution parameters. With these, we may use the formulas presented in Chapter 3 to compute the variance for each activity, as shown in Figure 10.33 and are included in columns M through Q of the *Becker Consulting Project Management Simulation Model* spreadsheet. The variance of the critical path is 281.88 (found by adding the variances of those activities with zero slack). Suppose that the company would like to complete the project within 150 days. With the normality assumption, the probability that the project will be completed within 150 days is found by computing the z-value:

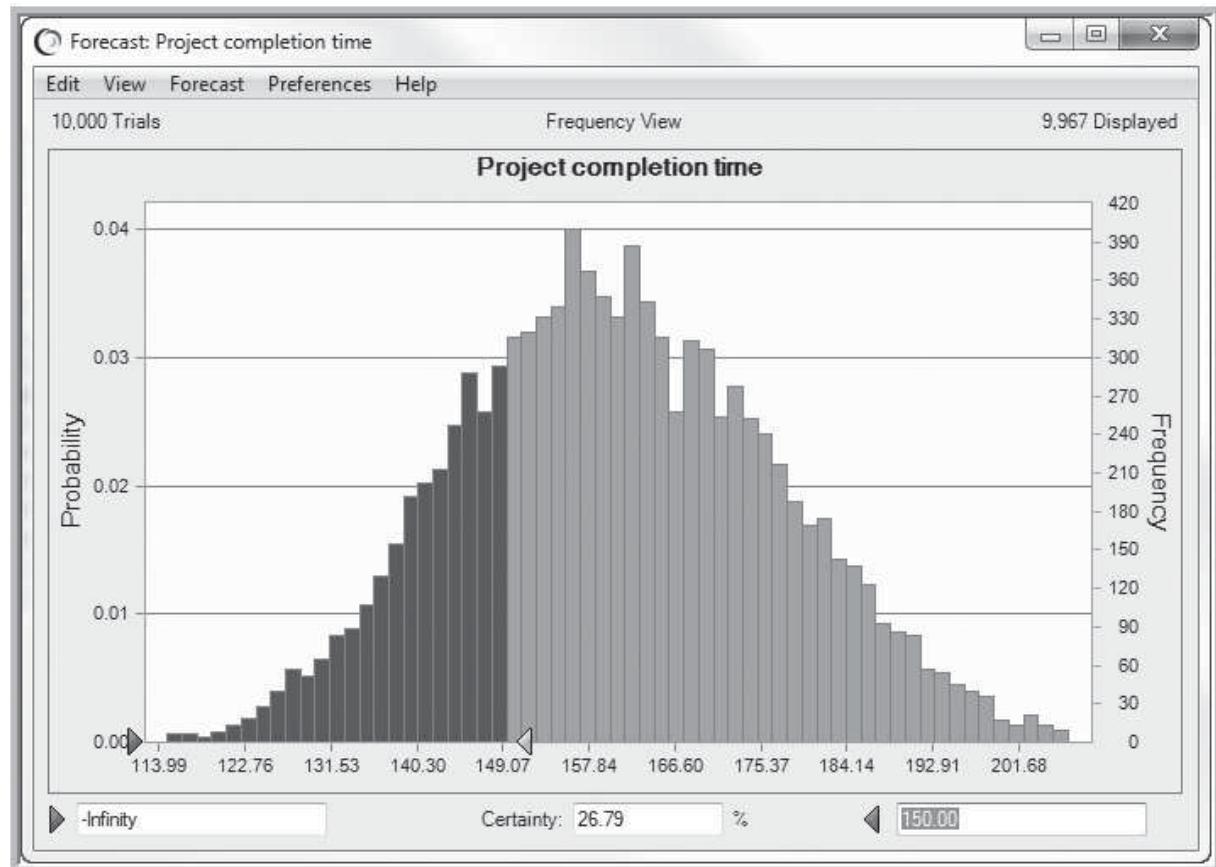
$$z = \frac{150 - 159.43}{16.79} = -0.562$$

This corresponds to a probability of approximately 0.29.

However, variations in actual activity times may yield different critical paths than the one resulting from expected times. This may change both the mean and variance of the actual project completion time, resulting in an inaccurate assessment of risk. Simulation can easily address these issues. Figure 10.34 shows the *Crystal Ball* forecast and statistics charts for 5,000 trials. The mean and variance are quite close to those predicted; in fact, fitting the normal distribution to the forecast in an overlay chart results in a very good fit. Whereas the analytical approach computed the probability of completing the project

<b>Critical Path Calculations</b>			<i>mean</i>	<i>variance</i>
<i>a</i>	<i>m</i>	<i>b</i>		
15	15	15	15.00	0.00
33.63	45	69.31	49.31	55.37
4.07	14	39.51	19.19	55.70
0.95	3	6.34	3.43	1.23
3.38	7	10.62	7.00	2.18
2.73	5	9.89	5.87	2.23
1	1	1	1.00	0.00
17.58	30	61.89	36.49	87.07
0.67	5	13.21	6.29	6.76
0.03	7	12.68	6.57	6.69
10.66	30	55.79	32.15	85.44
2.19	4	5.81	4.00	0.55
3	3	3	3.00	0.00
6.62	13	24.55	14.72	13.77
4.48	14	36.48	18.32	45.00
2	2	2	2.00	0.00
5	5	5	5.00	0.00
On critical path:			159.44	281.88
			standard deviation	16.79

**FIGURE 10.33** Analytical Critical Path Calculations



**FIGURE 10.34** Project Completion Time Forecast Chart

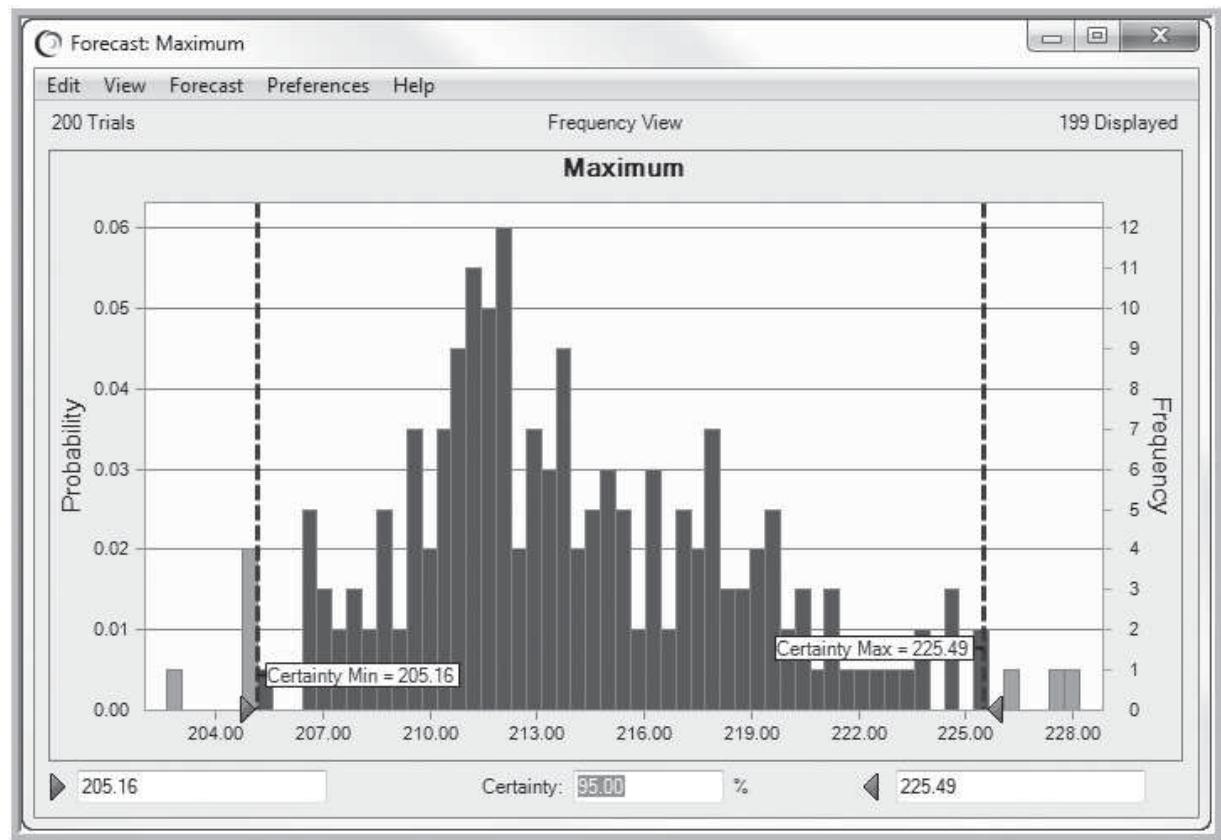
within 150 days as 0.29, analysis of the forecast chart shows this to be somewhat smaller, about 0.27. Note, however, that the structure of the project network is quite “linear,” resulting in few options for critical paths. For projects that have many more parallel paths, the results of the simulation and analytical models may differ significantly more.

In this example, the simulation results provide estimates of key statistics such as the mean, median, maximum value, and so on. If we run the model again, these values will change slightly because of the different random assumptions that would be generated. A logical question is, how accurate are these estimates? For example, suppose that Becker Consulting is concerned about the maximum project time. From the simulation (viewing the statistics associated with the forecast chart), an estimate of the maximum project time is about 219 days. To better understand the variability, we can use the *Bootstrap* tool in *Crystal Ball* (see Appendix 10.1F, *Bootstrap Tool*). The *Bootstrap* tool uses a simple technique that estimates the reliability or accuracy of any of the forecast statistics shown in the statistics view of the simulation results. Classical methods used in the past relied on mathematical formulas to describe the accuracy of sample statistics. In contrast, bootstrapping analyzes sample statistics empirically by repeatedly sampling the data and creating sampling distributions of the different statistics. The term *bootstrap* comes from the saying, “to pull oneself up by one’s own bootstraps,” since this method uses the distribution of statistics themselves to analyze the statistics’ accuracy.

We used 200 bootstrap samples with 1,000 trials per sample. Figure 10.35 shows the bootstrap forecast chart for the maximum value of the project time. In essence, this



Spreadsheet Note



**FIGURE 10.35** Bootstrap Forecast Chart for Maximum Project Completion Time

represents the sampling distribution of the statistic. Although the maximum project completion time has a mean of about 214 days, we see that this might reasonably vary from about 205 to 225 days.

#### SKILL-BUILDER EXERCISE 10.7

Apply the *Bootstrap* tool to the Moore Pharmaceuticals model to find the sampling distribution of the mean NPV.

### Basic Concepts Review Questions

1. What is meant by risk? What does risk analysis aim to do?
2. What is a Monte Carlo simulation? How is it helpful in the analysis of risk?
3. Explain the terms *assumption*, *forecast*, and *decision* as used in *Crystal Ball*.
4. How do we define uncertain model inputs in *Crystal Ball*?
5. Explain the types of charts that *Crystal Ball* provides for analyzing simulation results and how they may be used to provide useful information to a decision maker.
6. How can an analyst obtain the sampling distributions of standard descriptive statistical measures such as mean, median, variance, standard deviation, etc. in *Crystal Ball*?

## Problems and Applications

1. Financial analysts often use the following model to characterize changes in stock prices:

$$P_t = P_0 e^{(\mu - 0.5\sigma^2)t + \sigma Z\sqrt{t}}$$

where

$P_0$  = current stock price

$P_t$  = price at time  $t$

$\mu$  = mean (logarithmic) change of the stock price per unit time

$\sigma$  = (logarithmic) standard deviation of price change

$Z$  = standard normal random variable

This model assumes that the logarithm of a stock's price is a normally distributed random variable (see the discussion of the lognormal distribution and note that the first term of the exponent is the mean of the lognormal distribution). Using historical data, one can estimate values for  $\mu$  and  $\sigma$ . Suppose that the average daily change for a stock is \$0.003227, and the standard deviation is 0.026154. Develop a spreadsheet to simulate the price of the stock over the next 30 days, if the current price is \$53. Use the Excel function NORMSINV(RAND( )) to generate values for  $Z$ . Construct a chart showing the movement in the stock price.

2. The Miller–Orr model in finance addresses the problem of managing its cash position by purchasing or selling securities at a transaction cost in order to lower or raise its cash position. That is, the firm needs to have enough cash on hand to meet its obligations, but does not want to maintain too high a cash balance because it loses the opportunity for earning higher interest by investing in other securities. The Miller–Orr model assumes that the firm will maintain a minimum cash balance,  $m$ , a maximum cash balance,  $M$ , and an ideal level,  $R$ , called the return point. Cash is managed using a decision rule that states that whenever the cash balance falls to  $m$ ,  $R - m$  securities are sold to bring the balance up to the return point. When the cash balance rises to  $M$ ,  $M - R$  securities are purchased to reduce the cash balance back to the return point. Using some advanced mathematics, the return point and maximum cash balance levels are shown to be:

$$R = m + Z$$

$$M = R + 2Z$$

where

$$Z = \left( \frac{3C_0\sigma^2}{4r} \right)^{1/3}$$

$\sigma^2$  = variance of the daily cash flows

$r$  = average daily rate of return corresponding to the premium associated with securities

For example, if the premium is 4%,  $r = 0.04/365$ . To apply the model, note that we do not need to know

the actual demand for cash, only the daily variance. Essentially, the Miller–Orr model determines the decision rule that minimizes the expected costs of making the cash-security transactions and the expected opportunity costs of maintaining the cash balance based on the variance of the cash requirements. Suppose that the daily requirements are normally distributed with a mean of 0 and variance of \$60,000. Assume a transaction cost equal to \$35, with an interest rate premium of 4%, and a required minimum balance of \$7,500. Develop a spreadsheet implementation for this model. Apply Monte Carlo simulation to simulate the cash balance over the next year (365 days). Your simulation should apply the decision rule that if the cash balance for the current day is less than or equal to the minimum level, sell securities to bring the balance up to the return point. Otherwise, if the cash balance exceeds the upper limit, buy enough securities (i.e., subtract an amount of cash) to bring the balance back down to the return point. If neither of these conditions hold, there is no transaction and the balance for the next day is simply the current value plus the net requirement. Show the cash balance results on a line chart.

3. For the *Moore Pharmaceuticals* model, suppose that analysts have made the following assumptions:

R&D costs: Triangular(\$500, \$700, \$800) in millions of dollars

Clinical trials costs: Triangular(\$135, \$150, \$160) in millions of dollars

Market size: Normal(2,000,000, 250,000)

Market share in year 1: Uniform(6%, 10%)

All other data are considered constant. Develop and run a *Crystal Ball* model to forecast the NPV and cumulative net profit for each year. Summarize your results in a short memo to the R&D director.

4. For the *Outsourcing Decision Model*, suppose that the demand volume is lognormally distributed with a mean of 1,500 and a standard deviation of 500. What is the distribution of the cost differences between manufacturing in-house and purchasing? What decision would you recommend? Define both the cost difference and decision as forecast cells. Because *Crystal Ball* forecast cells must be numeric, replace the formula in cell B20 with =IF(B18 <= 0, 1, 0); that is, "1" represents manufacturing and "0" represents outsourcing.

5. For the *Airline Pricing Model* in Chapter 9, suppose that the fixed cost is triangular with a minimum of \$80,000, most likely value of \$90,000, and maximum value of \$95,000. Also assume that the values of the slope and intercept in the demand function are uniformly distributed plus or minus 5% around their current values.

Find the distribution of profit for a unit price of \$500. Use the *Decision Table* tool to find the best price between \$400 and \$600 in steps of \$25.

6. Using the generic profit model developed in the section *Logic and Business Principles* in Chapter 9, develop a financial simulation model for a new product proposal and construct a distribution of profits under the following assumptions: Price is fixed at \$1,000. Unit costs are unknown and follow the distribution.

Unit Cost	Probability
\$400	0.20
\$600	0.40
\$700	0.25
\$800	0.15

Demand is also variable and follows the following distribution:

Demand	Probability
120	0.25
140	0.50
160	0.25

Fixed costs are estimated to follow the following distribution:

Fixed Costs	Probability
\$45,000	0.20
\$50,000	0.50
\$55,000	0.30

Implement your model using *Crystal Ball* to determine the best production quantity to maximize the average profit. Would you conclude that this product is a good investment? (Data for this problem can be found in the *Problem 6* worksheet in the Excel file *Chapter 10 Problem Data*.)

7. The manager of the apartment complex in Problem 9 of Chapter 9 believes that the number of units rented during any given month has a triangular distribution with minimum 30, most likely 34, and maximum 40. Operating costs follow a normal distribution with mean \$15,000 and a standard deviation of \$300. Use *Crystal Ball* to estimate the 80%, 90%, and 95% confidence intervals for the profitability of this business.
- What is the probability that monthly profit will be positive?
  - What is the probability that monthly profit will exceed \$4,000?
  - Compare the 80%, 90%, and 95% certainty ranges.
  - What is the probability that profit will be between \$1,000 and \$3,000?

8. Develop a *Crystal Ball* model for the garage band in Problem 11 in Chapter 9 with the following assumptions. The expected crowd is normally distributed with a mean of 3,000 and a standard deviation of 400 (minimum of 0). The average expenditure on concessions is also normally distributed with mean \$15, standard deviation \$3, and minimum 0. Identify the mean profit, the minimum observed profit, maximum observed profit, and the probability of achieving a positive profit. Develop and interpret a confidence interval for the mean profit for a 5,000-trial simulation.

9. Develop a *Crystal Ball* model for a three-year financial analysis of total profit based on the following data and information. Sales volume in the first year is estimated to be 100,000 units and is projected to grow at a rate that is normally distributed with a mean of 7% per year and a standard deviation of 4%. The selling price is \$10, and the price increase is normally distributed with a mean of \$0.50 and standard deviation of \$0.05 each year. Per-unit variable costs are \$3, and annual fixed costs are \$200,000. Per-unit costs are expected to increase by an amount normally distributed with a mean of 5% per year and standard deviation of 2%. Fixed costs are expected to increase following a normal distribution with a mean of 10% per year and standard deviation of 3%. Based on 5,000 simulation trials, find the average three-year cumulative profit, and explain the percentile report. Generate and explain a trend chart showing net profit by year.

10. Develop a *Crystal Ball* model for MasterTech with the following assumptions (all other data are as described in Problem 13, Chapter 9): Net sales are uniformly distributed between \$600,000 and \$1,200,000. Cost of sales is normally distributed, with a mean of \$540,000 and a standard deviation of \$20,000. Selling expenses has a fixed component that is uniform between \$75,000 and \$110,000. Administrative expenses are normal, with a mean of \$50,000 and standard deviation of \$3,500. Develop a risk profile of net income using *Crystal Ball* and write a report to management.
11. Develop a *Crystal Ball* model for Koehler Vision Associates (KVA) in Problem 15 of Chapter 9 with the following assumptions: The weekly demand averages 175, but anywhere between 10% and 20% of prospective patients fail to show up or cancel their exam at the last minute. Determine the best level of overbooking to maximize the net profit (revenue less overbooking costs). Assume that the demand is uniform between 110 and 160 per week.
12. For the Hyde Park Surgery Center scenario described in Problem 19 in Chapter 9, suppose that the following assumptions are made: The number of patients served the first year is uniform between 1,300 and 1,700; the growth rate for subsequent years is triangular with parameters (5%, 8%, 9%), and the growth rate for year 2 is independent of the growth rate for year 3; average billing is normal with a mean of \$150,000 and a standard

deviation of \$10,000; and the annual increase in fixed costs is uniform between 5% and 7%, and independent of other years. Find the distribution of the NPV of profit over the three-year horizon and analyze the sensitivity and trend charts. Summarize your conclusions.

13. Review the MBA's retirement planning situation described in Chapter 9 (Figure 9.16). Modify the spreadsheet to include the assumptions that the annual raise each year is triangular with a minimum of 1%, a most likely value of 3%, and a maximum value of 5%, and that the investment return each year is triangular with a minimum of -8%, a most likely value of 5%, and a maximum value of 9%. Use *Crystal Ball* to find the distribution of the ending retirement fund balance under these assumptions. How do the results compare with the base case?
14. Refer back to the college admission director scenario (Problem 20 in Chapter 9). Develop a spreadsheet model and apply *Crystal Ball* tools to make a recommendation on how many scholarships to offer.
15. The data in the Excel file *Real Estate Sales* represent the annual sales of 100 branch offices of a national real estate company. Use the *Crystal Ball* distribution fitting procedure to find the best-fitting distribution for these data.
16. The data in the Excel file *Technical Support Data* represent a sample of the number of minutes spent resolving customer support issues for a computer manufacturer. Use the *Crystal Ball* distribution fitting procedure to find the best-fitting distribution for these data.
17. A sporting goods store orders ski jackets in the summer before the winter season. Each jacket costs \$80 and sells for \$200. Any not sold by March are discounted 75%. Demand depends on the weather. The distribution of demand is triangular (but must be whole numbers) with a minimum of 40, a maximum of 150, and a most likely value of 80. How many jackets should the retailer order? Conduct a simulation analysis using *Crystal Ball* to answer this question.
18. A plant manager is considering investing in a new \$30,000 machine. Use of the new machine is expected to generate a cash flow of about \$8,000 per year for each of the next five years. However, the cash flow is uncertain, and the manager estimates that the actual cash flow will be normally distributed with a mean of \$8,000 and a standard deviation of \$500. The discount rate is set at 8% and assumed to remain constant over the next five years. The company evaluates capital investments using NPV. How risky is this investment? Develop an appropriate simulation model and conduct experiments and statistical output analysis to answer this question.
19. Jennifer Bellin (Problem 18 in Chapter 9) believes that the activity times for her project are uncertain. The worksheet *Problem 19* in the Excel file *Chapter 10 Problem Data* contains information about the activities, predecessors, and activity times in the worksheet *Problem 19*. Assume a BetaPERT distribution for the activity times. How far in advance of the conference must Jennifer

begin the project activities to ensure a 95% chance of completing the project by the scheduled date?

20. A software development project consists of six activities. The activities and their predecessor relationships are given in the following table. The worksheet *Problem 20* in the Excel file *Chapter 10 Problem Data* provides a sample of times in weeks for activities A, B, E, and F from 40 past projects.

Activity	Activity Times	Predecessors
A Requirements analysis	See Excel file	None
B Programming	See Excel file	A
C Hardware	Constant 3 weeks	A
D User training	Constant 12 weeks	A
E Implementation	See Excel file	B,C
F Testing	See Excel file	E

Assume that activity start time is zero if there are no predecessor activities. Fit distributions to the sample data and use *Crystal Ball* to find the mean project completion time, minimum and maximum project completion times, skewness of the completion time distribution, and probability of completing the project in 14, 15, 16, or 17 weeks.

21. For the stock broker problem described in Problem 16 in Chapter 9, assume that among the qualified clients, half will invest between \$2,000 and \$10,000, 25% will invest between \$10,000 and \$25,000, 15% will invest between \$25,000 and \$50,000, and the remainder will invest between \$50,000 and \$100,000, each uniformly distributed. Using the same commission schedule, how many calls per month must the broker make each month to have at least a 75% chance of making at least \$5,000?
22. For the nonprofit ballet company in Problem 17 of Chapter 9, assume following percentages of donors and gift levels:

Gift Level	Amount	Number of Gifts
Benefactor	\$10,000	1–3
Philanthropist	\$5,000	3–7
Producer's Circle	\$1,000	16–25
Director's Circle	\$500	31–40
Principal	\$100	5–7% of solicitations
Soloist	\$50	5–7% of solicitations

The company has set a financial goal of \$150,000. How many prospective donors must they contact for donations at the \$100 level or below to have a 95% chance of meeting this goal? Assume that the number of gifts at each level follow a discrete uniform distribution or a uniform distribution for the percentage of solicitations at the \$100 and \$50 levels.

**23.** Arbino Mortgage Company obtains business from direct marketing letters signed by a fictitious loan officer named Jackson Smith. The percentages of interested responses from Jackson Smith letters are triangular with parameters 1%, 1.2%, and 2%. Of those responding, the percentage of respondents that actually close on a loan is also triangular with parameters 10%, 15%, and 18%. The average loan fee for a Jackson Smith loan is \$3,500 with a standard deviation of \$650. Other loan requests are obtained from two other sources: referrals and repeat customers, and unsolicited customers obtained from other advertising (billboard ads, Google searches, and so on). Fees for referrals and repeat customers average \$2,600 with a standard deviation of \$500 (these are less in order to provide an incentive for

future business), and unsolicited customers' loan fees are the same as Jackson Smith loans. The company has 15 loan officers. Each loan officer will close an average of one loan per month from referrals and repeat customers, and about one loan every four months from unsolicited customers (use judgment to define a reasonable uniform distribution around these averages). The company is moving to a new office and will double in size. This requires them to cover additional overhead expenses. The general manager wants to be 90% certain that the office will close at least \$600,000 in total loan fees each month from all sources. The principal question is how many Jackson Smith letters should be sent each month to ensure this. Develop a simulation model to help identify the best decision.

## Case

### J&G Bank

J&G Bank receives a large number of credit card applications each month, an average of 30,000 with a standard deviation of 4,000, normally distributed. Approximately 60% of them are approved, but this typically varies between 50% and 70%. Each customer charges a total of \$2,000, normally distributed, with a standard deviation of \$250, to his or her credit card each month. Approximately 85% pay off their balance in full, and the remaining incur finance charges. The average finance charge has recently varied from 3.0% to 4% per month. The bank also receives income from fees charged for late payments and annual fees associated with the credit cards. This is a percentage of total monthly charges, and has varied between 6.8% and 7.2%. It costs the bank \$20 per application, whether

it is approved or not. The monthly maintenance cost for credit card customers is normally distributed with a mean of \$10 and a standard deviation of \$1.50. Finally, losses due to charge-offs of customers' accounts range between 4.6% and 5.4%.

- a. Using average values for all uncertain inputs, develop a spreadsheet model to calculate the bank's total monthly profit.
- b. Use *Crystal Ball* to analyze the profitability of the credit card product and summarize your results to the manager of the credit card division. Use *Crystal Ball* tools to fully analyze your results and provide a complete and useful report to the manager.

## APPENDIX 10.1

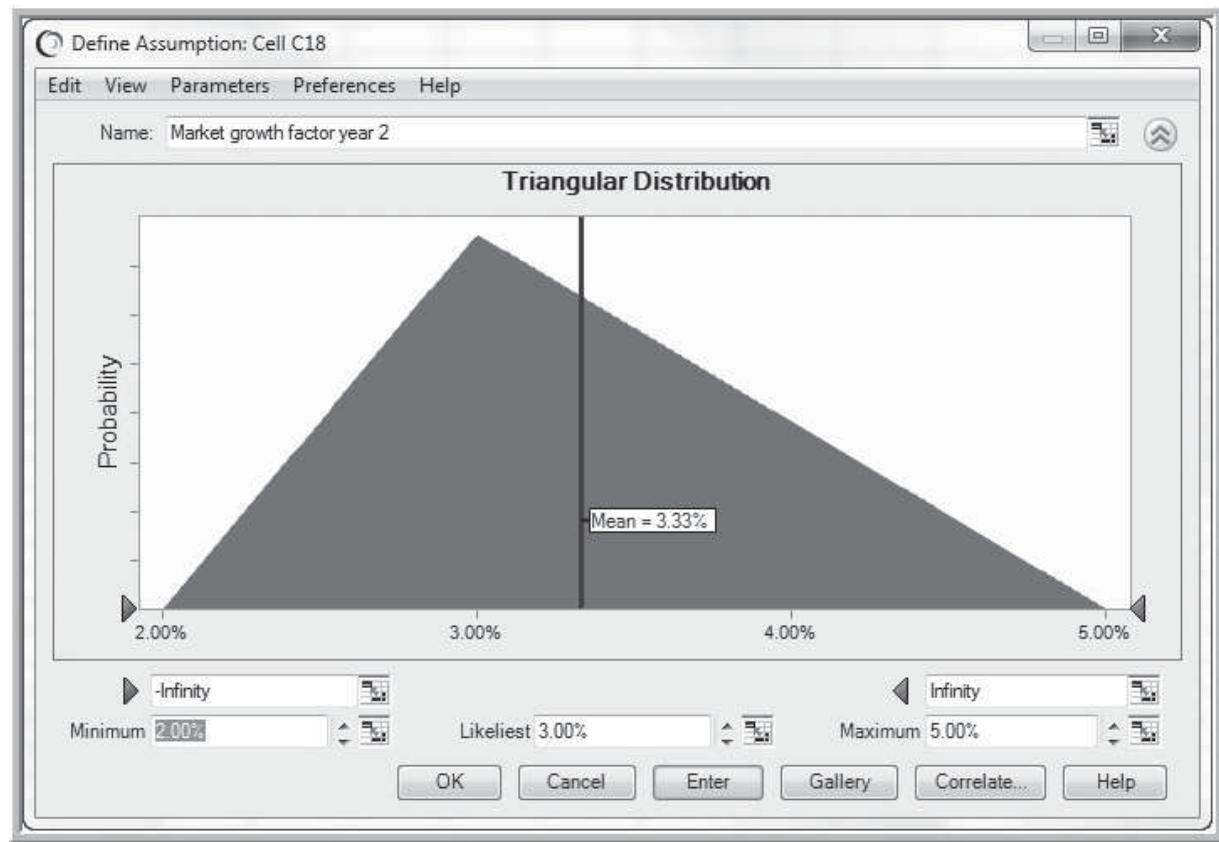
### Crystal Ball Notes

#### A. Customizing Define Assumption

For most continuous distributions, you have several options on how to specify the distribution in the *Parameters* menu in the *Define Assumption* dialog. For example, with the normal distribution, the default is to enter the mean and standard deviation; however, you can also define the distribution by its 10th and 90th percentiles, the mean and the 90th percentile, and several other ways. This option is useful when only percentile information is available or when specific parameters such as the mean and standard deviation are unknown. As a practical illustration, suppose that you are

interviewing a construction manager to identify the distribution of the time it takes to complete a task. Although a beta distribution is often appropriate in such applications, it would be very difficult to define the parameters for the beta distribution from judgmental information. However, it would be easy for the manager to estimate the 10th and 90th percentiles for task times.

*Crystal Ball* also allows you to customize the charts by adding gridlines, 3D effects, "marker lines" that show various statistical information, such as the mean, median, etc., as well as formatting the *x*-axis and other options. Marker lines are especially useful for distributions in which



**FIGURE 10A.1** Example of Marker Line

the mean or standard deviation are not input parameters. These options can be invoked by selecting *Chart* from the *Preferences* menu in the *Define Assumption* dialog and then clicking on the *Chart Type* tab. Figure 10A.1 shows a marker line for the mean in the market growth factor rate assumption for year 2.

## B. Sensitivity Charts

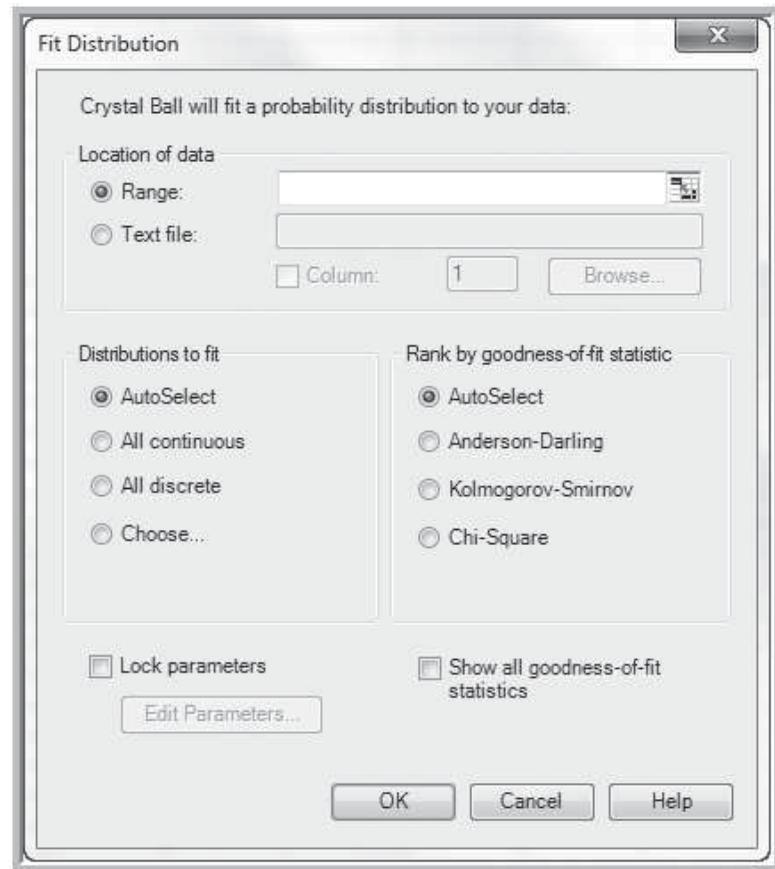
To open a sensitivity chart, select *Sensitivity Charts* from the *View Charts* menu in the *Analyze* group. Click the *New* button to create a new chart, or check the box corresponding to a previously defined chart and click the *Open* button. For a new chart, check the box corresponding to the forecast you wish to create a sensitivity chart for in the *Choose Forecast* dialog that appears. Two types of charts are available: *Rank Correlation View* and *Contribution to Variance View* (default). The *Contribution to Variance View* addresses the question “What percentage of the variance in the target forecast is due to a particular assumption?” For the *Rank Correlation View*, *Crystal Ball* computes rank correlation coefficients between each assumption and forecast. Rank correlation uses the ranking of assumption values rather than the actual numbers. These correlation coefficients provide a measure of the degree to which assumptions and forecasts change together.

Positive coefficients indicate that an increase in the assumption is associated with an increase in the forecast; negative coefficients imply the reverse. The larger the absolute value of the correlation coefficient, the stronger is the relationship. This chart may be displayed by selecting *Rank Correlation Chart* from the *View* menu. In addition, choosing *Sensitivity Data* from the *View* menu displays numerical results for both charts instead of graphical views.

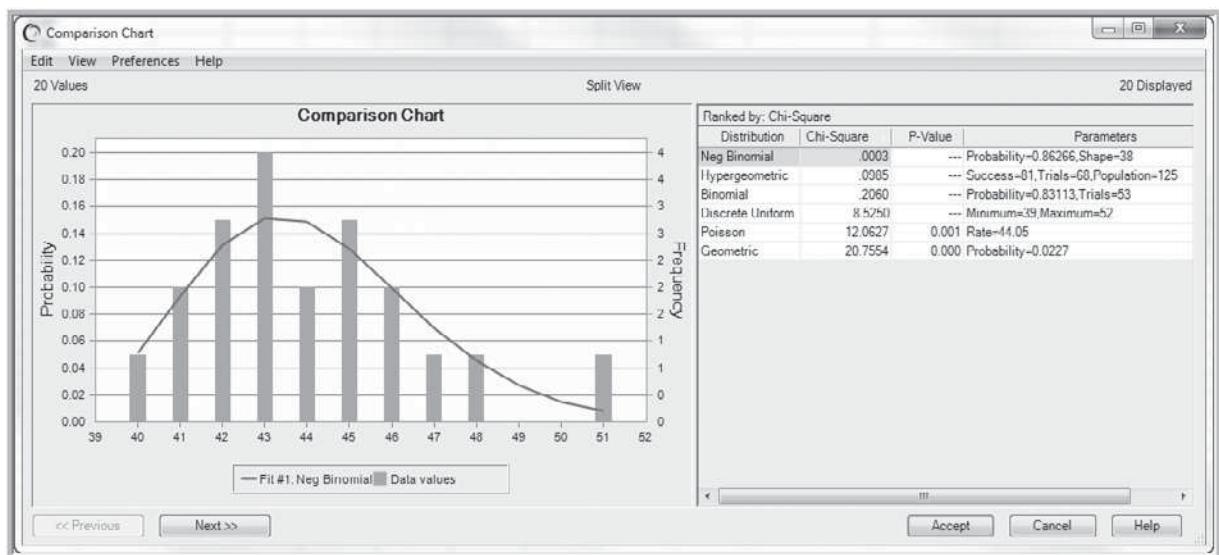
## C. Distribution Fitting with *Crystal Ball*

*Crystal Ball* provides a useful procedure for fitting a probability distribution to empirical data. Not only does it determine the closeness of each fit using one of the standard goodness-of-fit tests, but it also determines the set of parameters for each distribution that best describes the characteristics of the data.

To fit distributions to data, first select a data cell and click *Define Assumption* from the *Define* group in the menu bar. Click on the *Fit* button in the *Distribution Gallery* that pops up. In the next dialog, specify the range of the data, which distributions to fit, and the ranking method for choosing the best fit (see Figure 10A.2). Click on *OK*. The fitted distributions appear next in the *Comparison Chart* dialog (Figure 10A.3). The *Next* and *Previous* buttons beneath



**FIGURE 10A.2** Fit Distribution Dialog



**FIGURE 10A.3** Comparison Chart Dialog

the *Comparison Chart* will allow you to scroll through the fitted probability distributions. Each probability distribution is shown superimposed over the data. The parameters corresponding to the fitted distributions are also shown at the right.

## D. Correlation Matrix Tool

In *Crystal Ball*, you may enter correlations between any pair of assumptions by clicking on one of them, selecting *Define Assumption*, and then choosing the *Correlate* button in the *Define Assumption* dialog. This brings up the *Define Correlation* dialog, which allows you to choose the assumption to correlate with and enter the value of the correlation coefficient. The dialog displays a scatter chart of how the assumptions would be related to each other. Instead of manually entering the correlation coefficients this way, you can use the *Correlation Matrix* tool to define a matrix of correlations between assumptions in one simple step. This saves time and effort when building your spreadsheet model, especially for models with many correlated assumptions.

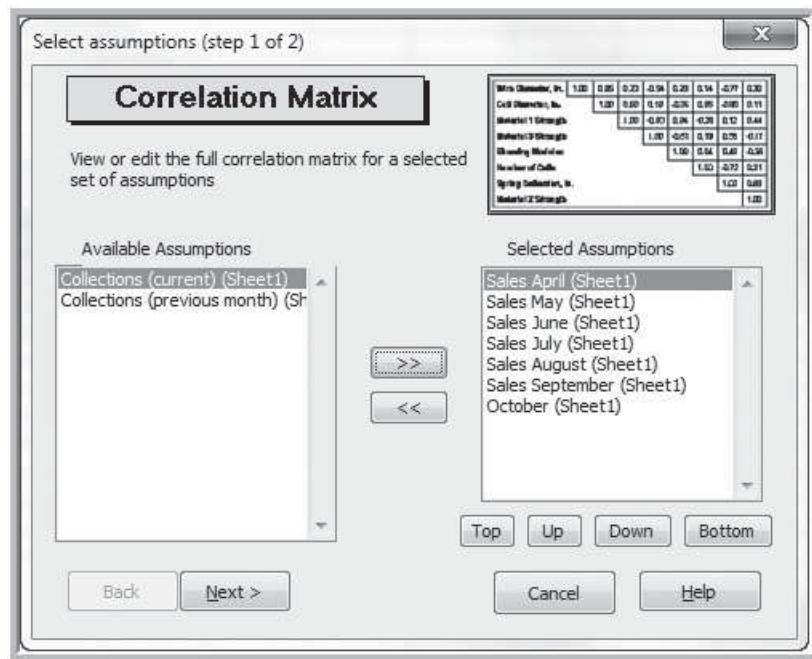
Choose the *Correlation Matrix* from the *Tools* menu in the *Run* group. This brings up a dialog box in which you select the assumptions to correlate. In the cash budget example, we wish to correlate sales, so choose all assumptions except the current and previous month's collections as shown in Figure 10A.4. The correlation matrix is either an upper or lower triangular matrix with 1's along the diagonal and correlation coefficients entered in the blank cells. The tool allows you to select the type of matrix

and whether to create in a new worksheet or the existing worksheet; we recommend using the existing worksheet simply to document the input data. Click *Start* in the next dialog to create the correlation matrix, which is shown in Figure 10A.5. Click the *Load the Matrix* button to enter the correlations into your model. If you enter inconsistent correlations, *Crystal Ball* tries to adjust the correlations so they don't conflict.

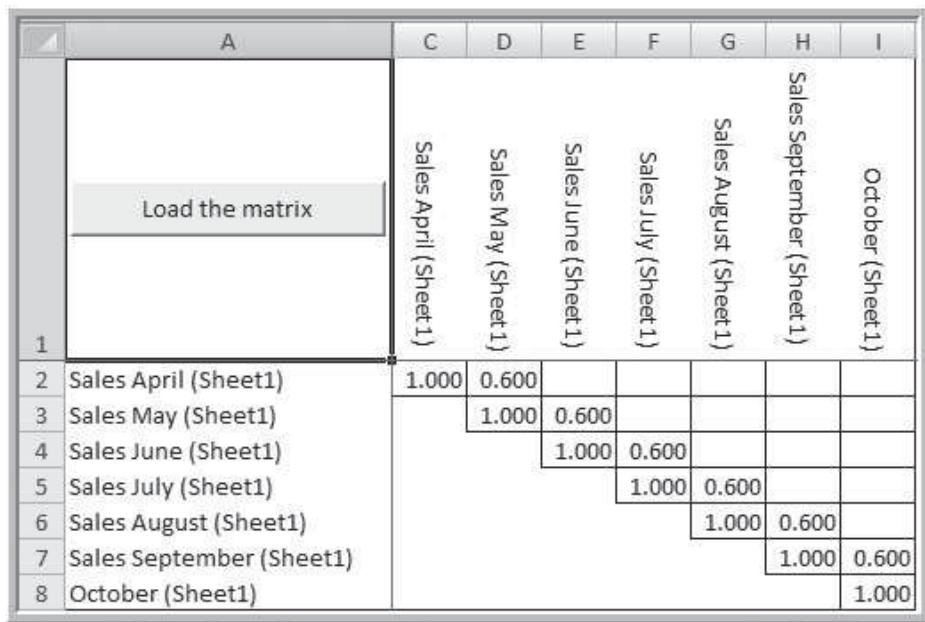
## E. Tornado Charts

Select *Tornado Charts* from the *Tools* menu in the *Run* group. The first dialog box that opens asks you to select the forecast you wish to analyze. In the next box, click the appropriate button to add assumptions or other variables defined in your model. Use the *Add Range* button to add variables that are not defined as assumptions or decisions. The default values in the third box can be left alone. Click on *Start*.

The *Tornado Chart* tool tests the range of each variable at percentiles you specify and then calculates the value of the forecast at each point. The tornado chart illustrates the swing between the maximum and minimum forecast values for each variable, placing the variable that causes the largest swing at the top and the variable that causes the smallest swing at the bottom. The top variables have the most effect on the forecast, and the bottom variables have the least effect on the forecast. While analyzing one variable, the tool freezes the other variables at their base values. This measures the effect each variable has on the forecast cell while removing the effects of the other variables.



**FIGURE 10A.4** Selecting Assumptions to Correlate



**FIGURE 10A.5** Crystal Ball Correlation Matrix

The bars next to each variable represent the forecast value range across the variable tested. Next to the bars are the values of the variables that produced the greatest swing in the forecast values. The bar colors indicate the direction of the relationship between the variables and the forecast. For variables that have a positive effect on the forecast, the upside of the variable (dark shading in Figure 10.30, actually shown in blue) is to the right of the base case and the downside of the variable (light shading in Figure 10.30, actually shown in red) is to the left side of the base case. For variables that have a reverse relationship with the forecast, the bars are reversed.

## F. Bootstrap Tool

Choose *Bootstrap* from the *Tools* menu in the *Run* group. The first dialog box asks you to select the forecast you wish to analyze. In the second box, you must choose one of two alternative methods:

1. *One-simulation method*, which simulates the model data once and repeatedly samples with replacement

2. *Multiple-simulation method*, which repeatedly simulates the model and creates a sampling distribution from each simulation; this method is more accurate but is slower

The *Bootstrap* tool constructs sampling distributions for the following statistics:

- Mean
- Median
- Standard deviation
- Variance
- Skewness
- Kurtosis
- Coefficient of variability
- Mean standard error
- Range minimum (multiple-simulation method only)
- Range maximum (multiple-simulation method only)
- Range width (multiple-simulation method only)

In the third dialog box, set the number of samples and number of trials per sample and click the *Start* button.

## *Chapter 11*

# Decisions, Uncertainty, and Risk

- INTRODUCTION 368
- DECISION MAKING UNDER CERTAINTY 368
  - Decisions Involving a Single Alternative 369
  - Decisions Involving Non-mutually Exclusive Alternatives 369
  - Decisions Involving Mutually Exclusive Alternatives 370
- DECISIONS INVOLVING UNCERTAINTY AND RISK 371
  - Making Decisions with Uncertain Information 371
  - Decision Strategies for a Minimize Objective 372
  - Decision Strategies for a Maximize Objective 374
  - Risk and Variability 375
- EXPECTED VALUE DECISION MAKING 377
  - Analysis of Portfolio Risk 378
  - The “Flaw of Averages” 380
- DECISION TREES 381
  - A Pharmaceutical R&D Model 381
  - Decision Trees and Risk 382
  - Sensitivity Analysis in Decision Trees 384
- THE VALUE OF INFORMATION 384
  - Decisions with Sample Information 386
  - Conditional Probabilities and Bayes’s Rule 387
- UTILITY AND DECISION MAKING 389
  - Exponential Utility Functions 393
- BASIC CONCEPTS REVIEW QUESTIONS 394
- PROBLEMS AND APPLICATIONS 395
- CASE: THE SANDWICH DECISION 399
- APPENDIX 11.1: EXCEL, PHSTAT, AND TREEPLAN NOTES 399
  - A. Using the IRR Function 399
  - B. Using the *Expected Monetary Value* Tool 399
  - C. Constructing Decision Trees in Excel 400

## INTRODUCTION

A world of difference exists between building, analyzing, and solving decision models and making decisions. Decision models such as Monte Carlo simulations can provide insight into the impacts of potential decisions or, in the case of prescriptive models, recommendations as to the best course of action to take. However, people make decisions, and their decisions often have significant economic or human resource consequences that cannot always be predicted accurately. Therefore, understanding the philosophy of decision making and how to deal with uncertainty and risk is vitally important to being a good manager.

Managers make many kinds of decisions. Some of these decisions are repetitive, perhaps on an annual or quarterly basis. These might include selecting new employees from a pool of applicants or selecting specific business improvement projects from a pool of proposals. These types of decisions have little financial impact and would not be considered very risky. Other decisions such as deciding where to locate a plant or warehouse or determining whether to pursue a major investment opportunity commit the firm to spending large sums of money. These usually are one-time decisions and generally involve a higher level of risk because of uncertain data and imperfect information. In this chapter, we present various approaches for evaluating and making decisions.

## DECISION MAKING UNDER CERTAINTY

The first steps in making a decision are to understand the alternative decisions that can be made and to identify a rational criterion (or multiple criteria) by which to evaluate the impact of the decision. **Decision alternatives** represent the choices that a decision maker can make. These alternatives might be a simple set of decisions, such as locating a factory among five potential sites or choosing one of three corporate health plan options. Other situations require a more complex sequence of decisions. For example, in new product introduction, a marketing manager might have to decide on whether to test-market a new product, and then, based on the test-market results, decide to conduct further tests, begin a full-scale marketing effort, or drop the product from further consideration. In either case, the manager must list the options that are available.

Generating viable alternatives might involve some prescreening (perhaps using optimization models). For instance, a company might develop, solve, and perform sensitivity analysis on an optimization model to generate potential plant location sites based on total distribution costs. However, making the final decision would involve many qualitative factors such as labor supply, tax incentives, environmental regulations, and future uncertainties. Managers must ensure that they have considered all possible options so that the “best” one will be included in the list. This often requires a good deal of creativity to define unusual options that might not normally be considered. Managers must put aside the tendency to jump to a quick solution and consider creative alternatives.

After the alternatives are defined, we must select the criteria on which to evaluate them. Decision criteria might be to maximize discounted net profits, customer satisfaction, or social benefits, or to minimize costs, environmental impact, or some measures of loss. Decision criteria should be aligned with an organization’s strategic objectives.

Many decisions involve making a choice among one or more alternatives under *certainty*, that is, when the consequences of the decision can be reasonably predicted and uncertainty and risk are not considered important. In this type of situation, financial analysis and optimization models are often used. We may classify typical decisions under certainty as:

- Decisions involving a single alternative
- Decisions involving non-mutually exclusive alternatives
- Decisions involving mutually exclusive alternatives

## Decisions Involving a Single Alternative

Decisions involving a single alternative, such as whether to outsource a product, purchase a new technology, or invest in a research and development project, generally are straightforward to make. The choice is either to accept or to reject the alternative based on some criterion. For example, in the outsourcing decision model in Chapter 9, the decision could be made by finding the breakeven volume and determining whether the anticipated volume is larger or smaller than the breakeven point.

In most cases, the decision criterion is based on financial considerations. Three common criteria used to evaluate capital investments are the payback period, net present value (NPV), and internal rate of return (IRR). The **payback period** is the number of time periods (usually in years) before the cash inflows of a proposed project equal the amount of the initial investment. The required payback period is a policy decision; if a project's payback period meets the requirement, then it is accepted. The payback criterion is simple, but it does not consider all future cash flows or the time value of money.

NPV was introduced in Chapter 9 and supports the financial goal of wealth maximization. NPV represents the dollar amount of change in the value of the firm as a result of choosing a decision. A positive NPV, therefore, should result in a decision to accept the alternative; one with a negative NPV should be rejected.

The IRR is the discount rate that makes the total present value of all cash flows sum to 0:

$$\sum_{t=0}^n \frac{F_t}{(1 + \text{IRR})^t} = 0 \quad (11.1)$$

IRR is often used to compare a project against a predetermined *hurdle rate*, a rate of return required by management to accept a project, which is often based on the rate of return required by shareholders. If IRR is greater than the hurdle rate, the project should be accepted; otherwise, it should be rejected. The IRR can be computed in Excel (see Appendix 11.1A, *Using the IRR Function*).

To illustrate this, consider the model for Moore Pharmaceuticals in Chapter 9 (Figure 9.11). The model calculated the net profit for each of five years and the NPV. The company's hurdle rate has been determined to be 10%. While the model showed a positive NPV of over \$185 million, it is not clear whether this project will meet the required hurdle rate. Using the data in Figure 9.11, the IRR is found to be 15.96%, and therefore, the company should continue to develop the drug.



Spreadsheet Note

### SKILL-BUILDER EXERCISE 11.1

An investment requires an initial cash outlay of \$100,000 and additional outlays of \$50,000 at the end of each of the first three years. This investment is expected to result in incomes of \$40,000 at the end of the first year, \$70,000 at the end of the second year, \$90,000 at the end of the third year, and \$120,000 at the end of the fourth year. Calculate the IRR using the IRR function.

## Decisions Involving Non-mutually Exclusive Alternatives

If several non-mutually exclusive alternatives are being considered, a ranking criterion provides a basis for evaluation and selection. Alternatives are ranked in order of desirability, and the highest ranked alternatives are chosen in order, provided that sufficient resources are available to support the decisions. For example, in selecting among R&D projects, there is usually a limited budget for initial investment or other resource constraints such as professional staff. If the NPV of each alternative is computed, we would select the projects in order of the largest NPV until the budget or other resources are exhausted.

Other criteria for such decisions are return on investment (ROI) or more general benefit/cost ratios. ROI is computed as:

$$ROI = \frac{\text{Annual Revenue} - \text{Annual Costs}}{\text{Initial Investment}} \quad (11.2)$$

Benefit/cost analysis is based on examining the ratios of expected benefits to expected costs. Benefit/cost analysis is often used in evaluating social projects where benefits generally cannot be quantified, for example, municipal projects such as parks and urban renewal. Ratios greater than 1.0 generally indicate that a proposal should be adopted if sufficient resources exist to support it.

### Decisions Involving Mutually Exclusive Alternatives

Many decisions involve a choice among several mutually exclusive alternatives, that is, if one is selected, none of the others may be selected. Some examples would be decisions about purchasing automobiles, choosing colleges, selecting mortgage instruments, investing money, introducing new products, locating plants, and choosing suppliers, to name just a few. Decision criteria might be to maximize discounted net profits or social benefits, or to minimize costs or some measure of loss.

When only one alternative can be selected from among a small set of alternatives, the best choice can usually be identified by evaluating each alternative using the criterion chosen. For example, in analyzing the newsvendor model in Chapter 10, we used a *Crystal Ball* decision table to evaluate the average profit for each purchase quantity and then chose the best. When the set of alternatives is very large or infinite, then an optimization model and a solution technique such as *Solver* can help identify the best choice. The airline pricing model we developed in Chapter 9 is one example. Chapters 13 and 14 will deal with more complex optimization models.

For decisions involving multiple criteria, simple scoring models are often used. A **scoring model** is a quantitative assessment of a decision alternative's value based on a set of attributes. Usually, each attribute is characterized by several levels, and a score is assigned to each level that reflects the relative benefits of that level. For example, in evaluating new product ideas, some attributes might include product development time, the competitive environment, and ROI. A simple scoring model for these attributes might be the one shown in Figure 11.1. For each decision alternative, a score would be

LEVEL	SCORE
1. Product development time	
a. Less than 6 months	+5
b. 6 months to 1 year	+3
c. 1–2 years	0
d. 2–3 years	-3
e. more than 2 years	-5
2. Competitive environment	
a. None	+5
b. Few minor competitors	+3
c. Many minor competitors	0
d. Few major competitors	-3
e. Many major competitors	-5
3. Return on investment	
a. 30% or more	+5
b. 25–30%	+3
c. 15–25%	0
d. 10–15%	-3
e. below 15%	-5

FIGURE 11.1 Scoring Model Example

assigned to each attribute (which might be weighted to reflect different priorities), and the overall score would be used as a basis for selection.

## DECISIONS INVOLVING UNCERTAINTY AND RISK

*Uncertainty* is imperfect knowledge of what will happen; *risk* is associated with the consequences of what actually happens. Even though uncertainty may exist, there may be no risk. For example, the change in the stock price of Apple on the next day of trading is uncertain. This uncertainty has no impact if you don't own Apple stock. However, if you do, then you bear the risk associated with the possibility of losing money. Thus, risk is an outcome of uncertainty.

The importance of risk in business has long been recognized. The renowned management writer, Peter Drucker, observed in 1974:

To try to eliminate risk in business enterprise is futile. Risk is inherent in the commitment of present resources to future expectations. Indeed, economic progress can be defined as the ability to take greater risks. The attempt to eliminate risks, even the attempt to minimize them, can only make them irrational and unbearable. It can only result in the greatest risk of all: rigidity.<sup>1</sup>

Consideration of risk is a vital element of decision making. For instance, you would probably not choose an investment simply on the basis of its expected return because, typically, higher returns are associated with higher risk. Therefore, you have to make a trade-off between the benefits of greater rewards and the risks of potential losses. We can see this in the *Crystal Ball* results for the Moore Pharmaceuticals model in Chapter 10. When uncertainty exists in the model parameters, even though the average NPV is positive, there is a 19% chance that the NPV will be negative (Figure 10.11). This might make the decision maker think twice about developing the drug.

Decisions involving uncertainty and risk have been studied for many years. A large body of knowledge has been developed that helps explain the philosophy associated with making decisions and provides techniques for incorporating uncertainty and risk in making decisions.

### Making Decisions with Uncertain Information

Many decisions involve a choice from among a small set of decisions with uncertain consequences. We may characterize such decisions by defining three things:

1. The decision alternatives
2. The outcomes that may occur once a decision is made
3. The payoff associated with each decision and outcome

Outcomes, often called **events**, may be quantitative or qualitative. For instance, in selecting the size of a new factory, the future demand for the product would be uncertain. The demand might be expressed quantitatively in sales units or dollars. On the other hand, suppose that you are planning a spring break vacation to Florida in January; you might define uncertain weather-related outcomes qualitatively: sunny and warm, sunny and cold, rainy and warm, or rainy and cold, etc. The payoff is a measure of the value of making a decision and having a particular outcome occur. This might be a

---

<sup>1</sup> P.F. Drucker, *The Manager and the Management Sciences in Management: Tasks, Responsibilities, Practices* (London: Harper and Row, 1974).

simple estimate made judgmentally or a value computed from a complex spreadsheet model. Payoffs are often summarized in a **payoff table**, a matrix whose rows correspond to decisions and whose columns correspond to events. The following example illustrates these concepts.

Many young families face the decision of choosing a mortgage instrument. Suppose the Durr family is considering purchasing a new home and would like to finance \$150,000. Three mortgage options are available: a one-year adjusted-rate mortgage (ARM) at a low interest rate, a three-year ARM at a slightly higher rate, and a 30-year fixed mortgage at the highest rate. However, both ARMs are sensitive to interest rate changes, and the rates may change, resulting in either higher or lower interest charges; thus, the potential changes in interest rates are the uncertain outcomes. As the family anticipates staying in the home for at least five years, they are interested in the total interest costs they might incur; these represent the payoffs associated with their choice and the future change in interest rates and can easily be calculated using a spreadsheet. The payoff table is given below:

Decision	Outcome		
	Rates Rise	Rates Stable	Rates Fall
1-year ARM	\$61,134	\$46,443	\$40,161
3-year ARM	\$56,901	\$51,075	\$46,721
30-year fixed	\$54,658	\$54,658	\$54,658

Clearly, no decision is best for all outcome scenarios. The best decision clearly depends on what outcome occurs. If rates rise, then the 30-year fixed would be the best decision. If rates remain stable or fall, then the one-year ARM is best. Of course, you cannot predict the outcome with certainty, so the question is how to choose one of the options. Not everyone views risk in the same fashion. Most individuals will weigh their potential losses against potential gains. For example, if they choose the one-year ARM mortgage instead of the fixed-rate mortgage, they risk losing money if rates rise; however, they would clearly save a lot if rates remain stable or fall. Would the potential savings be worth the risk?

Evaluating risk should take into account both the magnitude of potential gains and losses and their probabilities of occurrence, if this can be assessed. For instance, suppose that you are offered a chance to win a \$40,000 car in a charity raffle for \$100 in which only 1,000 tickets are sold. Although the probability of losing is 0.999, most individuals would not view this to be very risky because the loss of only \$100 would not be viewed as catastrophic (ignoring the charitable issues!).

### Decision Strategies for a Minimize Objective

We will discuss several quantitative approaches that model different risk behaviors for making decisions involving uncertainty.

**AVERAGE PAYOFF STRATEGY** Because the future events are unpredictable, we might simply assume that each one is as likely to occur as the others. This approach was proposed by the French mathematician Laplace, who stated the *principle of insufficient reason*: If there is no reason for one state of nature to be more likely than another, treat them as equally likely. Under this assumption, which is called the **Laplace or average payoff strategy**, we evaluate each decision by simply averaging the payoffs. We then

select the decision with the best average payoff. For the mortgage selection problem, we have the following:

Decision	Outcome			
	Rates Rise	Rates Stable	Rates Fall	Average Payoff
1-year ARM	\$61,134	\$46,443	\$40,161	\$49,246
3-year ARM	\$56,901	\$51,075	\$46,721	\$51,566
30-year fixed	\$54,658	\$54,658	\$54,658	\$54,658

Based on this criterion, we choose the decision having the smallest average payoff, or the one-year ARM.

**AGGRESSIVE STRATEGY** An aggressive decision maker might seek the option that holds the promise of minimizing the potential loss. This type of decision maker would first ask the question, “What is the best that could result from each decision?” and then choose the decision that corresponds to the “best of the best.” For our example, this is summarized below:

Decision	Outcome			
	Rates Rise	Rates Stable	Rates Fall	Best Payoff
1-year ARM	\$61,134	\$46,443	\$40,161	\$40,161
3-year ARM	\$56,901	\$51,075	\$46,721	\$46,721
30-year fixed	\$54,658	\$54,658	\$54,658	\$54,658

Because our goal is to minimize costs, we would choose the one-year ARM. For a minimization objective, this strategy is also often called a **minimin strategy**, that is, we choose the decision that minimizes the minimum payoff. Aggressive decision makers are often called speculators, particularly in financial arenas because they increase their exposure to risk in hopes of increasing their return.

**CONSERVATIVE STRATEGY** A conservative decision maker, on the other hand, might take a more pessimistic attitude and ask, “What is the worst thing that might result from my decision?” and then select the decision that represents the “best of the worst.” For the mortgage decision problem, the largest costs for each option are as follows:

Decision	Outcome			
	Rates Rise	Rates Stable	Rates Fall	Worst Payoff
1-year ARM	\$61,134	\$46,443	\$40,161	\$61,134
3-year ARM	\$56,901	\$51,075	\$46,721	\$56,901
30-year fixed	\$54,658	\$54,658	\$54,658	\$54,658

In this case, we want to choose the decision that has the smallest worst payoff, or the 30-year fixed mortgage. Thus, no matter what the future holds, a cost of \$54,658 is guaranteed. Such a strategy is also known as a **minimax strategy** because we seek the decision that corresponds to the minimum value of the largest cost. Conservative decision makers are often called **risk averse** and are willing to forgo potential returns in order to reduce their exposure to risk.

**OPPORTUNITY LOSS STRATEGY** A fourth approach that underlies decision choices for many individuals is to consider the *opportunity loss* associated with a decision. Opportunity loss represents the “regret” that people often feel after making a nonoptimal decision (“I should have bought that stock years ago!”). In our example, suppose we chose the 30-year fixed mortgage and later find out that the interest rates rose. We could not have done any better by selecting a different decision; in this case, the opportunity loss is zero. However, if the interest rates remained stable, the best decision *would have been* to choose the one-year ARM. By choosing the 30-year fixed instrument, the investor lost a total of  $\$54,658 - \$46,443 = \$8215$ . This represents the opportunity loss associated with making the wrong decision. If the rates fell, the best decision would have been the one-year ARM also, and choosing the 30-year fixed mortgage would result in a larger opportunity loss of  $\$54,658 - \$40,161 = \$14,497$ .

In general, the opportunity loss associated with any decision and event is the difference between the *best* decision for that particular outcome and the payoff for the decision that was chosen. *Opportunity losses can only be nonnegative values!* If you get a negative number, then you made a mistake.

Once opportunity losses are computed, the decision strategy is similar to a conservative strategy. The decision maker would select the decision that minimizes the largest opportunity loss. For these reasons, this is also called a **minimax regret strategy**. This is summarized in the opportunity loss table below:

Decision	Outcome			Max Opportunity Loss
	Rates Rise	Rates Stable	Rates Fall	
1-year ARM	\$6,476	\$—	\$—	\$6,476
3-year ARM	\$2,243	\$4,632	\$6,560	\$6,560
30-year fixed	\$—	\$8,215	\$14,497	\$14,497

Using this strategy, we would choose the one-year ARM. This ensures that, no matter what outcome occurs, we will never be further than \$6,476 away from the least cost we could have incurred. *Different criteria, different decisions.* Which criterion best reflects your personal values?

### SKILL-BUILDER EXERCISE 11.2

Develop a spreadsheet for the Durr family mortgage example and use it to compute the information needed to implement the average payoff, aggressive, conservative, and opportunity loss strategies.

## Decision Strategies for a Maximize Objective

When the objective is to maximize the payoff, we can still apply the average payoff, aggressive, conservative, and opportunity loss strategies, but we must make some key changes in the analysis.

- For the average payoff strategy, compute the averages in the same fashion, but choose the decision with the *largest* average payoff.
- For the aggressive strategy, the best payoff for each decision would be the *largest* value among all outcomes, and we would choose the decision corresponding to the largest of these, thus called a **maximax strategy**.
- For the conservative strategy, the worst payoff for each decision would be the *smallest* value among all outcomes and we would choose the decision corresponding to the largest of these, thus called a **maximin strategy**.

- For the opportunity loss strategy, we need to be careful in calculating the opportunity losses. With a maximize objective, the decision with the largest value for a particular event has an opportunity loss of zero. The opportunity losses associated with other decisions is the difference between their payoff and the largest value. The actual decision is the same as when payoffs are costs: choose the decision that minimizes the maximum opportunity loss.

Table 11.1 summarizes the decision rules for both minimize and maximize objectives.

## Risk and Variability

A serious drawback of the average payoff strategy is that it neglects to consider the actual outcomes that can occur. For any decision (with the trivial exception of equal payoffs), the average outcome will *never occur*. For instance, choosing the one-year ARM will never result in the average payoff of \$49,246; the actual cost to the Durr family will be \$61,134, \$46,443, or \$40,161. Choosing the three-year ARM results in an average cost of \$51,566, but the possible outcomes vary only between \$46,721 and \$56,901. Therefore, while the averages are fairly similar, note that the one-year ARM has a larger variation in the possible outcomes.

In financial investment analysis, risk is often measured by the standard deviation. For example, *Fortune* magazine evaluates mutual fund risk using the standard deviation because it measures the tendency of a fund's monthly returns to vary from their long-term average. (As *Fortune* stated in one of its issues, "...standard deviation tells you what to expect in the way of dips and rolls. It tells you how scared you'll be.") For

**TABLE 11.1 Summary of Decision Strategies under Uncertainty**

Strategy Objective	Average Payoff Strategy	Aggressive Strategy	Conservative Strategy	Opportunity Loss Strategy
Minimize objective	Choose the decision with the smallest average payoff	Find the smallest payoff for each decision among all outcomes, and choose the decision with the smallest of these ( <i>minimum</i> )	Find the largest payoff for each decision among all outcomes, and choose the decision with the smallest of these ( <i>minimax</i> )	For each outcome, compute the opportunity loss for each decision as the difference between its payoff and the <i>smallest</i> payoff for that outcome. Find the maximum opportunity loss for each decision, and choose the decision with the smallest opportunity loss ( <i>minimax regret</i> ).
Maximize objective	Choose the decision with the largest average payoff	Find the largest payoff for each decision among all outcomes, and choose the decision with the largest of these ( <i>maximax</i> )	Find the smallest payoff for each decision among all outcomes, and choose the decision with the largest of these ( <i>maximin</i> )	For each outcome, compute the opportunity loss for each decision as the difference between its payoff and the <i>largest</i> payoff for that outcome. Find the maximum opportunity loss for each decision, and choose the decision with the smallest opportunity loss ( <i>minimax regret</i> ).

example, a mutual fund's return might have averaged 11% with a standard deviation of 10%. Thus, about two-thirds of the time the annualized monthly return was between 1% and 21%. By contrast, another fund's average return might be 14%, but have a standard deviation of 20%. Its returns would have fallen in a range of -6% to 34%, and therefore, it is more risky.

The statistical measure of coefficient of variation, which is the ratio of the standard deviation to the mean, provides a relative measure of risk to return. The smaller the coefficient of variation, the smaller the relative risk is for the return provided. The reciprocal of the coefficient of variation, called **return to risk**, is often used because it is easier to interpret. That is, if the objective is to maximize return, a higher return-to-risk ratio is often considered better. A related measure in finance is the **Sharpe ratio**, which is the ratio of a fund's excess returns (annualized total returns minus Treasury bill returns) to its standard deviation. *Fortune* noted that, for example, although both the American Century Equity-Income fund and Fidelity Equity-Income fund had three-year returns of about 21% per year, the Sharpe ratio for the American Century fund was 1.43 compared with 0.98 for Fidelity. If several investment opportunities have the same mean but different variances, a rational (risk-averse) investor will select the one that has the smallest variance.<sup>2</sup> This approach to formalizing risk is the basis for modern portfolio theory, which seeks to construct minimum-variance portfolios. As *Fortune* noted, "It's not that risk is always bad. . . . It's just that when you take chances with your money, you want to be paid for it."

Standard deviations may not tell the complete story about risk, however. It is also important to consider the skewness and kurtosis of the distribution of outcomes. For example, both a negatively and positively skewed distribution may have the same standard deviation, but clearly if the objective is to achieve high return, the negatively skewed distribution will have higher probabilities of larger returns. Higher kurtosis values indicate that there is more area in the tails of the distribution than for distributions with lower kurtosis values. This indicates a greater potential for extreme and possibly catastrophic outcomes.

Applying these concepts to the mortgage example, we may compute the standard deviation of the outcomes associated with each decision:

Decision	Standard Deviation
1-year ARM	\$10,763.80
3-year ARM	\$5,107.71
30-year fixed	\$—

Based solely on the standard deviation, the 30-year fixed mortgage has no risk at all, while the one-year ARM appears to be the riskiest. Although based only on three data points, the three-year ARM is fairly symmetric about the mean, while the one-year ARM is positively skewed—most of the variation around the average is driven by the upside potential (i.e., lower costs), not by the downside risk of higher costs. Although none of the formal decision strategies chose the three-year ARM, viewing risk from this perspective might lead to this decision. For instance, a conservative decision maker who is willing to tolerate a moderate amount of risk might choose the three-year ARM over the 30-year fixed because the downside risk is relatively small (and is smaller than the risk of the one-year ARM) and the upside potential is much larger. The larger upside potential associated with the one-year ARM might even make this decision attractive.

---

<sup>2</sup> David G. Luenberger, *Investment Science* (New York: Oxford University Press, 1998).

Thus, it is important to understand that making decisions under uncertainty cannot be done using only simple rules, but careful evaluation of risk versus rewards. This is why top executives make the big bucks!

## EXPECTED VALUE DECISION MAKING

The average payoff strategy is not appropriate for one-time decisions because, as we noted, average payoffs don't occur and we must consider risk. However, for decisions that occur on a repeated basis, an average payoff strategy can be used. For example, real estate development, day trading, and pharmaceutical research projects all fall into this scenario. Drug development is a good example. The cost of research and development projects in the pharmaceutical industry is generally in the hundreds of millions of dollars and could approach \$1 billion. Many projects never make it to clinical trials or might not get approved by the Food and Drug Administration. Statistics indicate that seven of ten products fail to return the cost of the company's capital. However, large firms can absorb such losses because the return from one or two blockbuster drugs can easily offset these losses. On an average basis, drug companies make a net profit from these decisions. This leads to the notion of expected value decision making. The newsvendor model that we discussed in Chapter 10 is a classic example of expected value decision making. We based the optimal purchase quantity on the average profit obtained by Monte Carlo simulation.

For decisions with a finite set of outcomes, we will assume that we know or can estimate the probabilities. If historical data on past occurrences of events are available, then we can estimate the probabilities objectively. If not, managers will often be able to estimate the likelihood of events from their experience and good judgment. The **expected monetary value (EMV) approach** selects the decision based on the best expected payoff. Define  $V(D_i, S_j)$  to be the payoff associated with choosing decision  $i$  and having outcome  $S_j$  to occur subsequently. Suppose that  $P(S_j)$  = the probability that outcome  $S_j$  occurs, and  $n$  = the number of outcomes. The expected payoff for each decision alternative  $D_i$  is:

$$E(D_i) = \sum_{j=1}^n P(S_j)V(D_i, S_j) \quad (11.3)$$

To illustrate this, let us consider a simplified version of the typical revenue management process that airlines use. At any date prior to a scheduled flight, airlines must make a decision as to whether to reduce ticket prices to stimulate demand for unfilled seats. If the airline does not discount the fare, empty seats might not be sold and the airline will lose revenue. If the airline discounts the remaining seats too early (and could have sold them at the higher fare), they would lose profit. The decision depends on the probability  $p$  of selling a full-fare ticket if they choose not to discount the price. Because an airline makes hundreds or thousands of such decisions each day, the expected value approach is appropriate.

Assume that only two fares are available: full and discount. Suppose that a full-fare ticket is \$560, the discount fare is \$400, and  $p = 0.75$ . For simplification, assume that if the price is reduced, then any remaining seats would be sold at that price. The expected value of not discounting the price is  $0.25(0) + 0.75(\$560) = \$420$ . Because this is higher than the discounted price, the airline should not discount at this time. In reality, airlines constantly update the probability  $p$  based on information they collect and analyze in a database. When the value of  $p$  drops below the breakeven point:  $\$400 = p(\$560)$  or  $p = 0.714$ , then it is beneficial to discount. It can also work in reverse; if demand is such that the probability that a higher-fare ticket would be sold, then the price may be adjusted upward. This is why published fares constantly change and why

you may receive last-minute discount offers or may pay higher prices if you wait too long to book a reservation. Other industries such as cruise lines use similar decision strategies.



#### Spreadsheet Note

*PHStat* provides tools for computing EMV and other decision information (see Appendix 11.1B, *Using the Expected Monetary Value Tool*). A completed example is shown in Figure 11.2 (Excel file *Day Trading Example*) for a situation in which a day trader wants to decide on investing \$100 in either a Dow index fund or a NASDAQ index fund. The probabilities and anticipated payoffs are shown in the Probabilities & Payoffs Table. We see that the best expected value decision would be to invest in the NASDAQ fund, which has a higher EMV as well as a better return-to-risk ratio, which indicates a higher comparative risk, despite a larger standard deviation. Of course, this analysis assumes that the probabilities and payoffs will be constant, at least over the near term.

### Analysis of Portfolio Risk

Concepts of expected value decisions and risk may be applied to the analysis of portfolios. A **portfolio** is simply a collection of assets, such as stocks, bonds, or other investments, that are managed as a group. In constructing a portfolio, one usually seeks to maximize expected return while minimizing risk. The expected return of a two-asset portfolio  $P$  is computed as:

$$E[P] = w_X E[X] + w_Y E[Y] \quad (11.4)$$

where  $w_X$  = fraction of portfolio for asset  $X$  and  $w_Y$  = fraction of portfolio for asset  $Y$ .

Risk depends on the correlation among assets in a portfolio. This can be seen by examining the formula for the standard deviation of a two-asset portfolio:

$$\sigma_P = \sqrt{w_X^2 \sigma_X^2 + w_Y^2 \sigma_Y^2 + 2w_X w_Y \sigma_{XY}} \quad (11.5)$$

Here,  $\sigma_{XY}$  is the covariance between  $X$  and  $Y$ . Note that if the covariance is zero, the variance of the portfolio is simply a weighted sum of the individual variances. If

A	B	C	D
1 Day Trading Decisions			
2			
3 Probabilities & Payoffs Table:			
4			
5 Dow Up/NASDAQ Up	0.42	\$ 4.00	\$ 5.00
6 Dow Up/NASDAQ Unchanged	0.04	\$ 3.00	\$ -
7 Dow Up/NASDAQ Down	0.10	\$ 1.00	\$ (3.00)
8 Dow Unchanged/NASDAQ Up	0.25	\$ -	\$ 4.00
9 Dow Unchanged/NASDAQ Unchanged	0.02	\$ -	\$ -
10 Dow Unchanged/NASDAQ Down	0.05	\$ -	\$ (2.00)
11 Dow Down/NASDAQ Up	0.05	\$ (2.00)	\$ 1.00
12 Dow Down/NASDAQ Unchanged	0.03	\$ (3.00)	\$ -
13 Dow Down/NASDAQ Down	0.04	\$ (6.00)	\$ (8.00)
14			
15 Statistics for:	Dow Index Fund	NASDAQ Index Fund	
16      Expected Monetary Value	1.47	2.43	
17      Variance	6.9291	12.3051	
18      Standard Deviation	2.63231837	3.507862597	
19      Coefficient of Variation	1.790692769	1.443564855	
20      Return to Risk Ratio	0.558443088	0.692729528	

**FIGURE 11.2**  
Expected Monetary  
Value Results for Day  
Trading Example

the covariance is negative, the standard deviation of the portfolio is smaller, while if the covariance is positive, the standard deviation of the portfolio is larger. Thus, if two investments (such as two technology stocks) are positively correlated, the overall risk increases, for if the stock price of one falls, the other would generally fall also. However, if two investments are negatively correlated, when one increases, the other decreases, reducing the overall risk. This is why financial experts advise diversification.

The *PHStat* tool *Covariance and Portfolio Analysis*, found under the *Decision Making* menu, can be used to perform these calculations for a two-asset portfolio. In the dialog that appears, enter the number of outcomes and check the box “Portfolio Management Analysis.” *PHStat* creates a template in which you enter the probabilities and outcomes similar to those in the *Expected Monetary Value* tool. Figure 11.3 shows the output for equal proportions of the index funds in the day trading example. Note that the expected return is \$1.95, the covariance is positive, and the portfolio risk is 2.83. Using the tool, it is easy to change the weighting and experiment with different values. For instance, a portfolio consisting of only the Dow index fund has a portfolio risk of 2.63 while yielding an expected return of \$1.47, and a portfolio consisting of only the NASDAQ index fund has a portfolio risk of 3.51 while yielding an expected return of \$2.43. You may also use *Solver* (see Chapter 9) to find the best weights to minimize the portfolio risk.

	A	B	C	D
1	Day Trading Portfolio			
2				
3	Probabilities & Outcomes:	Probability	Dow Index Fund	NASDAQ Index Fund
4		0.42	\$ 4.00	\$ 5.00
5		0.04	\$ 3.00	\$ -
6		0.10	\$ 1.00	\$ (3.00)
7		0.25	\$ -	\$ 4.00
8		0.02	\$ -	\$ -
9		0.05	\$ -	\$ (2.00)
10		0.05	\$ (2.00)	\$ 1.00
11		0.03	\$ (3.00)	\$ -
12		0.04	\$ (6.00)	\$ (8.00)
13				
14	Weight Assigned to X	0.5		
15				
16	Statistics			
17	E(X)	1.47		
18	E(Y)	2.43		
19	Variance(X)	6.9291		
20	Standard Deviation(X)	2.63231837		
21	Variance(Y)	12.3051		
22	Standard Deviation(Y)	3.5078626		
23	Covariance(XY)	6.3479		
24	Variance(X+Y)	31.93		
25	Standard Deviation(X+Y)	5.65066368		
26				
27	Portfolio Management			
28	Weight Assigned to X	0.5		
29	Weight Assigned to Y	0.5		
30	Portfolio Expected Return	1.95		
31	Portfolio Risk	2.82533184		

**FIGURE 11.3** *PHStat* Output for Portfolio Risk Analysis

## SKILL-BUILDER EXERCISE 11.3

Use Solver to find the best weights to minimize risk for the day trading portfolio example.

### The “Flaw of Averages”

One might think that any decision model with probabilistic inputs can be easily analyzed by simply using expected values for the inputs. Let’s see what happens if we do this for the newsvendor model. If we find the average of the historical candy sales, we obtain 44.05, or rounded to a whole number, 44. Using this value for demand and purchase quantity, the model predicts a profit of \$264 (see Figure 11.4). However, if we construct a data table to evaluate the profit for each of the historical values (also shown in Figure 11.4), we see that the average profit is only \$255. Dr. Sam Savage, a strong proponent of spreadsheet modeling, coined the term “the flaw of averages” to describe this phenomenon. Basically what this says is that the evaluation of a model output using the average value of the input is not necessarily equal to the average value of the outputs when evaluated with each of the input values, or in mathematical terms,  $f(E[X]) \neq E[f(X)]$ , where  $f$  is a function,  $X$  is a random variable, and  $E[]$  signifies expected value. The reason this occurs in the newsvendor example is because the quantity sold is limited by the smaller of the demand and purchase quantity, so even when demand exceeds the purchase quantity, the profit is limited. Using averages in models can conceal risk, and this is a common error among users of quantitative models.

## SKILL-BUILDER EXERCISE 11.4

Illustrate the “flaw of averages” using the overbooking model developed in Chapter 10.

A	B	C	D	E
1 Newsvendor Model				<i>Historical Candy Sales</i>
2				\$ 264.00
3 Data			42	\$ 246.00
4			45	\$ 264.00
5 Selling price	\$ 18.00		40	\$ 228.00
6 Cost	\$ 12.00		46	\$ 264.00
7 Discount price	\$ 9.00		43	\$ 255.00
8			43	\$ 255.00
9 Model			46	\$ 264.00
10			42	\$ 246.00
11 Demand	44		44	\$ 264.00
12 Purchase Quantity	44		43	\$ 255.00
13			47	\$ 264.00
14 Quantity Sold	44		41	\$ 237.00
15 Surplus Quantity	0		41	\$ 237.00
16			45	\$ 264.00
17 Profit	\$ 264.00		51	\$ 264.00
18			43	\$ 255.00
19			45	\$ 264.00
20			42	\$ 246.00
21			44	\$ 264.00
22			48	\$ 264.00
23			Average	\$ 255.00

FIGURE 11.4 Newsvendor Model with Average Demand

## DECISION TREES

A useful approach to structuring a decision problem involving uncertainty is to use a graphical model called a **decision tree**. Decision trees consist of a set of **nodes** and **branches**. Nodes are points in time at which events take place. The event can be a selection of a decision from among several alternatives, represented by a **decision node**, or an outcome over which the decision maker has no control, an **event node**. Event nodes are conventionally depicted by circles, while decision nodes are expressed by squares. Many decision makers find decision trees useful because *sequences* of decisions and outcomes over time can be modeled easily.

### A Pharmaceutical R&D Model

To illustrate the application of decision analysis techniques using decision trees, we will consider the R&D process for a new drug (you might recall the basic financial model we developed for the Moore Pharmaceuticals example in Chapter 9). Suppose that the company has spent \$300 million to date in research expenses. The next decision is whether or not to proceed with clinical trials. The cost of clinical trials is estimated to be \$250 million, and the probability of a successful outcome is 0.3. After clinical trials are completed, the company may seek approval from the Food and Drug Administration. This is estimated to cost \$25 million and there is a 60% chance of approval. The market potential has been identified as large, medium, or small with the following characteristics:

Market Potential Expected Revenues (millions of \$)		
		Probability
Large	\$4,500	0.6
Medium	\$2,200	0.3
Small	\$1,500	0.1

This book provides an Excel add-in called *TreePlan*, which allows you to construct decision trees and perform calculations within an Excel worksheet (see Appendix 11.1C, *Constructing Decision Trees in Excel*). The software contains further documentation and examples of how to use *TreePlan*, and we encourage you to experiment with it.

A decision tree for this situation constructed in *TreePlan* is shown in Figure 11.5 (Excel file *Drug Development Decision Tree*). In *TreePlan*, there are two ways of specifying the terminal values (payoffs). First, a value for the terminal value at the end of a path can simply be entered in the appropriate cell. A second method is to enter values or formulas for partial cash flows associated with each branch, as we have done here. *TreePlan* will accumulate the cash flows at the end of the tree. This approach is particularly useful when we wish to examine the sensitivity of the decision to specific values associated with decisions or events.

A decision tree is evaluated by “rolling back” the tree from right to left. When we encounter an event node, we compute the expected value of all events that emanate from the node since each branch will have an associated probability. For example, the rollback value of the top-right event node in Figure 11.5 is found by taking the expected value of the payoffs associated with market potential (note that payoffs in parentheses are negative values):

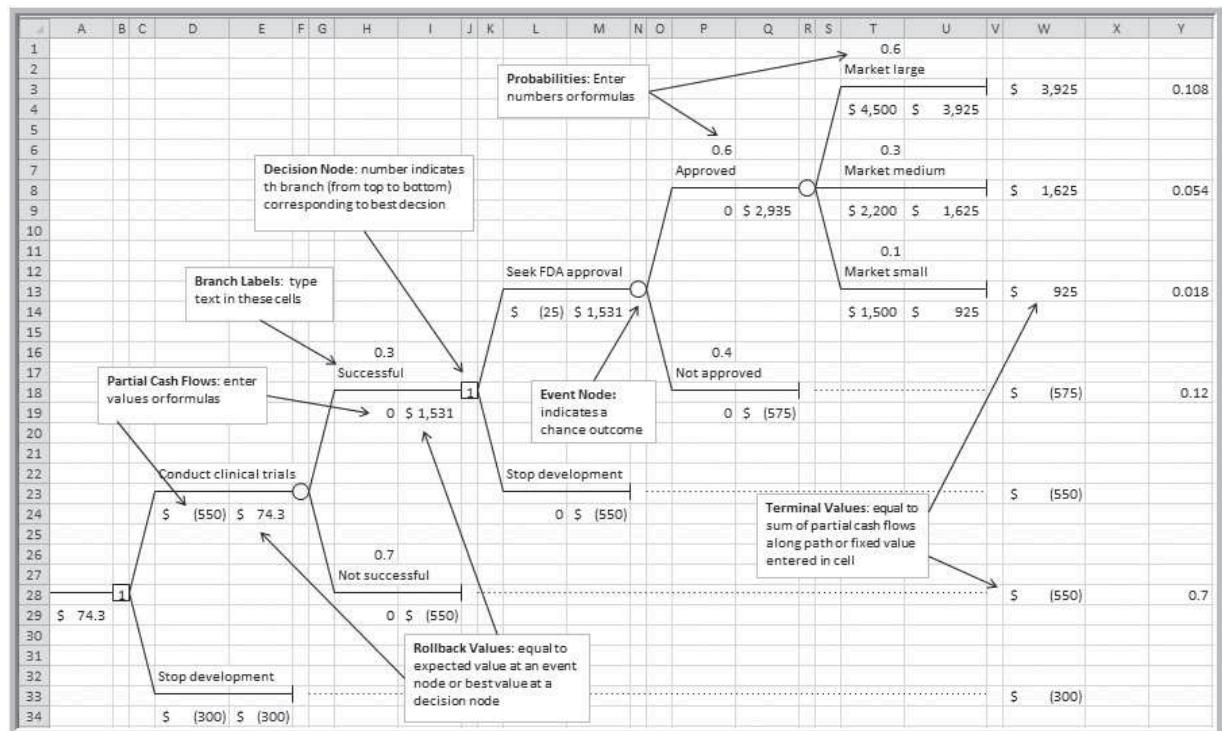
$$\$3,925 \times 0.6 + 1,625 \times 0.3 + \$925 \times 0.1 = \$2,935$$

Likewise, the expected value corresponding to the branch “Seek FDA approval” is computed as:

$$\$2,935 \times 0.6 + (\$575) \times 0.4 = \$1,531$$



Spreadsheet Note



**FIGURE 11.5** Drug Development Decision Tree Model

When we encounter a decision node (for instance, whether or not to seek FDA approval), we take the best of the expected values of all nodes that follow. The best decision here is to seek approval because the rollback value is \$1,531 versus a negative \$550 if the company decides to stop development. *TreePlan* designates the best choice by entering the number of the branch in the decision node square.

A **decision strategy** is a specification of an initial decision and subsequent decisions to make after knowing what events occur. For example, the best strategy is to conduct clinical trials, and if successful, seek FDA approval and market the drug. The expected net revenue is calculated as \$74.3 million. Another strategy would be to conduct clinical trials, and even if successful, stop development.

### Decision Trees and Risk

The decision tree approach is an example of expected value decision making, or essentially, the average payoff strategy in which the outcomes are weighted by their probabilities of occurrence. Thus, in the drug development example, if the company's portfolio of drug development projects has similar characteristics, then pursuing further development is justified on an expected value basis. However, this approach does not explicitly consider risk.

From a classical decision analysis perspective, we may summarize the company's decision as the following payoff table:

	Unsuccessful Clinical Trials	Successful Clinical Trials; No FDA approval	Successful Trials and Approval; Large Market	Successful Trials and Approval; Medium Market	Successful Trials and Approval; Small Market
Develop drug	\$(-550)	\$(-575)	\$3,925	\$1,625	\$925
Stop development	\$(-300)	\$(-300)	\$(-300)	\$(-300)	\$(-300)

If we apply the aggressive, conservative, and opportunity loss decision strategies to these data (note that the payoffs are profits as opposed to costs, so it is important to use the correct rule as discussed earlier in the chapter), we obtain:

Aggressive strategy (maximax)	
Maximum	
Develop drug	\$3,925
Stop development	\$(300)

The decision that maximizes the maximum payoff is to develop the drug.

Conservative strategy (maximin)	
Minimum	
Develop drug	\$(575)
Stop development	\$(300)

The decision that maximizes the minimum payoff is to stop development.

#### Opportunity loss:

Unsuccessful Clinical Trials	Successful Clinical Trials; No FDA Approval	Successful Trials and Approval; Large Market	Successful Trials and Approval; Medium Market	Successful Trials and Approval; Small Market	Maximum
Develop drug	\$250	\$275	\$—	\$—	\$275
Stop development	\$—	\$—	\$4,225	\$1,925	\$1,225

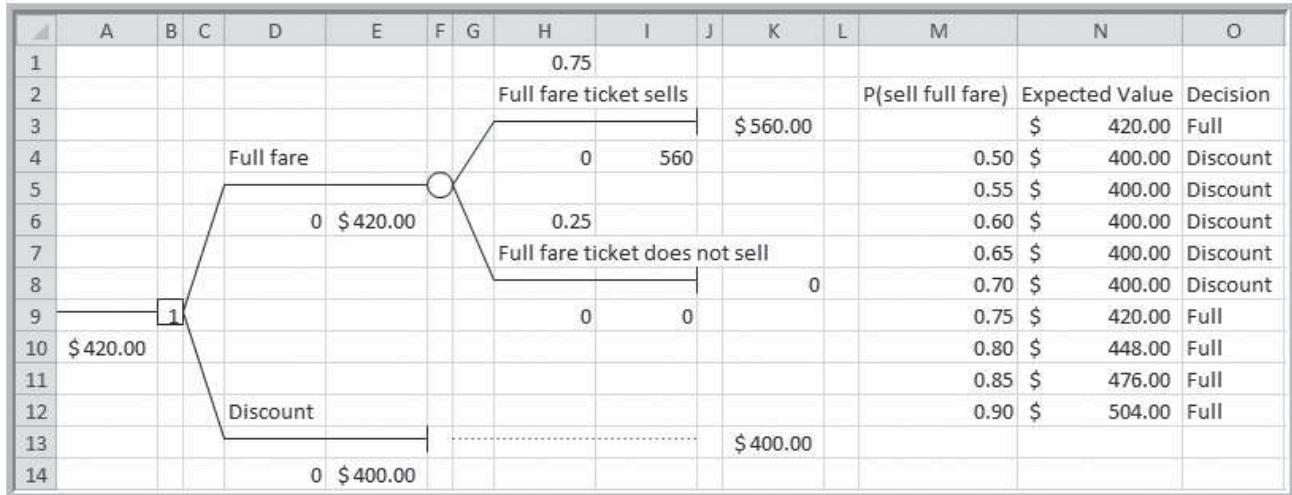
The decision that minimizes the maximum opportunity loss is to develop the drug.

As we noted, however, we must evaluate risk by considering both the magnitude of the payoffs and their chances of occurrence. The aggressive, conservative, and opportunity loss rules do not consider the probabilities of the outcomes.

Each decision strategy has an associated payoff distribution, called a **risk profile**. Risk profiles show the possible payoff values that can occur and their probabilities. For example, consider the strategy of pursuing development. The possible outcomes that can occur and their probabilities are:

Terminal Outcome	Net Revenue	Probability
Market large	\$3,925	0.108
Market medium	\$1,625	0.054
Market small	\$925	0.018
FDA not approved	(\$575)	0.120
Clinical trials not successful	(\$550)	0.700

The probabilities are computed by multiplying the probabilities on the event branches along the path to the terminal outcome. For example, the probability of getting to “Market large” is  $0.3 \times 0.6 \times 0.6 = 0.108$ . Thus, we see that the probability that the drug will not reach the market is  $1 - (0.108 + 0.054 + 0.018) = 0.82$ , and the company will incur a loss of over \$500 million. On the other hand, if they decide not to pursue clinical trials, the loss would only be \$300 million, the cost of research to date. If this



**FIGURE 11.6** Revenue Management Decision Tree and Data Table

were a one-time decision, what decision would you make if you were a top executive of this company?

### Sensitivity Analysis in Decision Trees

We may use Excel data tables to investigate the sensitivity of the optimal decision to changes in probabilities or payoff values. We will illustrate this using the airline revenue management scenario we discussed earlier in this chapter.

Figure 11.6 shows the decision tree (Excel file *Revenue Management Decision Tree*) for deciding whether or not to discount the fare with a data table for varying the probability of success with two output columns: one providing the expected value from cell A10 in the tree and the second column providing the best decision. The formula in cell O3 is = IF(B9 = 1, "Full", "Discount"). However, we must first modify the worksheet prior to constructing the data table so that probabilities will always sum to 1. To do this, enter the formula = 1 - H1 in cell H6 corresponding to the probability of not selling the full-fare ticket. From the results, we see that if the probability of selling the full-fare ticket is 0.7 or less, then the best decision is to discount the price. Two-way data tables may also be used in a similar fashion to study simultaneous changes in model parameters.

#### SKILL-BUILDER EXERCISE 11.5

Set up a decision tree for the airline revenue management example shown in Figure 11.6 using *TreePlan* and find the breakeven probability to within 0.01 using a data table.

### THE VALUE OF INFORMATION

When we deal with uncertain outcomes, it is logical to try to obtain better information about their likelihood of occurrence before making a decision. The **value of information** represents the improvement in the expected return that can be achieved if the decision maker is able to acquire—before making a decision—additional information about the future event that will take place. **Sample information** is the result of conducting some type of experiment, such as a market research study or interviewing an expert.

Sample information is always imperfect. In the ideal case, we would like to have **perfect information** that tells us with certainty what outcome will occur. Although this will never occur, it is useful to know the value of perfect information because it provides an upper bound on the value of any sample information.

Often, sample information comes at a cost. Thus, it is useful to know how much we should be willing to pay for it. The **Expected Value of Sample Information (EVSI)** is the EMV with sample information (assumed at no cost) minus the EMV without sample information; it represents the most you should be willing to pay for the sample information. The **Expected Value of Perfect Information (EVPI)** is the EMV with perfect information (assumed at no cost) minus the EMV without any information; again, it represents the most you should be willing to pay for perfect information. Clearly,  $\text{EVSI} < \text{EVPI}$ .

The *PHStat Expected Monetary Value* tool computes an Opportunity Loss Table and EVPI, shown in Figure 11.7, for the day trading example. This can also be done independently using the *PHStat* tool *Opportunity Loss* from the *Decision Making* options. The **expected opportunity loss**, shown in row 35 of Figure 11.4, represents the average additional amount the investor would have achieved by making the right decision instead of a wrong one. To find the expected opportunity loss, we create an opportunity loss table as we discussed earlier in this chapter. For example, if the event “Dow Up/NASDAQ Up” occurs, the best decision would have been to invest in the NASDAQ index fund and no opportunity loss would be incurred; if the Dow fund was chosen, the opportunity loss would be \$1. Once the opportunity loss table is constructed, the expected opportunity loss for each action is found by weighting the values by the event probabilities. We see that the NASDAQ index fund has the smallest expected opportunity loss. *It will always be true that the decision having the best expected value will also have the minimum expected opportunity loss.*

The minimum expected opportunity loss is the EVPI. For example, if we know with certainty that both the Dow and NASDAQ will rise, then the optimal decision would be to choose the NASDAQ index fund and get a return of \$5 instead of only \$4. Likewise, if we know that the second event will occur, we should choose the Dow index fund, and so on. By weighting these best outcomes by their probabilities of occurrence, we can compute the expected return under the assumption of having perfect information:

$$\begin{aligned}\text{Expected Return with Perfect Information} = & 0.42(\$5) + 0.04(\$3) + 0.1(\$1) + 0.25(\$4) \\ & + 0.02(\$0) + 0.05(\$0) + 0.05(\$1) \\ & + 0.03(\$0) + 0.04(-\$6) = \$3.13\end{aligned}$$

A	B	C	D	E
22	Opportunity Loss Table:			
23		Optimum	Optimum	Alternatives
24		Action	Profit	Dow Index Fund      NASDAQ Index Fund
25	Dow Up/NASDAQ Up	NASDAQ Index Fund	5	1      0
26	Dow Up/NASDAQ Unchanged	Dow Index Fund	3	0      3
27	Dow Up/NASDAQ Down	Dow Index Fund	1	0      4
28	Dow Unchanged/NASDAQ Up	NASDAQ Index Fund	4	4      0
29	Dow Unchanged/NASDAQ Unchanged	Dow Index Fund	0	0      0
30	Dow Unchanged/NASDAQ Down	Dow Index Fund	0	0      2
31	Dow Down/NASDAQ Up	NASDAQ Index Fund	1	3      0
32	Dow Down/NASDAQ Unchanged	NASDAQ Index Fund	0	3      0
33	Dow Down/NASDAQ Down	Dow Index Fund	-6	0      2
34			Dow Index Fund	NASDAQ Index Fund
35		Expected Opportunity Loss	1.66	0.7
36				EVPI

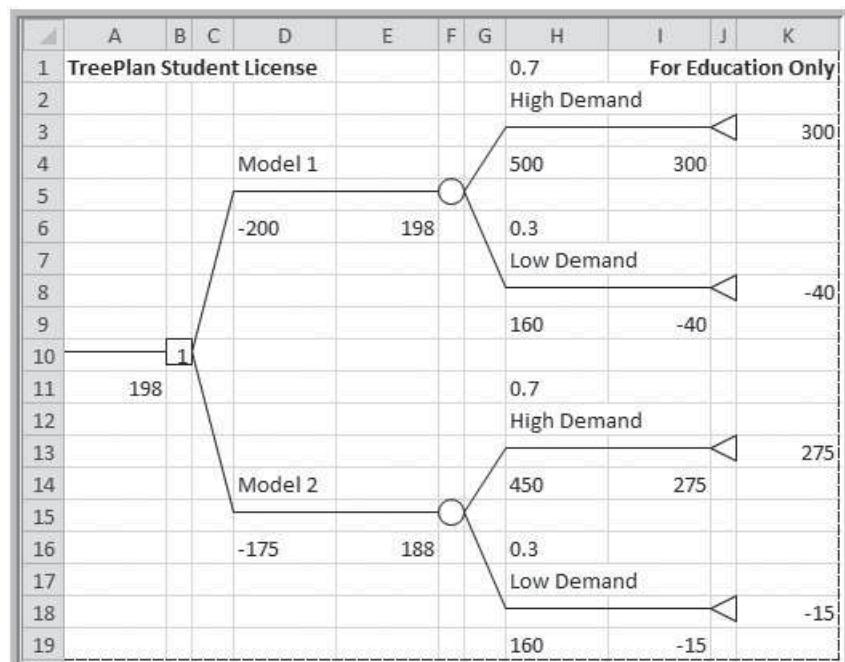
**FIGURE 11.7** Opportunity Loss Table for Day Trading Example

Because the expected value without having the perfect information is only \$2.43, we would have increased our average return by  $\$3.13 - \$2.43 = \$0.70$ . This is the EVPI. We would never want to pay more than \$0.70 for any information about the future event, no matter how good.

### Decisions with Sample Information

Suppose that a company is developing a new touch screen cell phone. Historically, 70% of their new phones have resulted in high consumer demand, while 30% have resulted in low consumer demand. The company has the decision of choosing between two alternative models with different features that require different amounts of investment and also have different sales potential. Figure 11.8 shows a completed decision tree in which all cash flows are in thousands of dollars. For example, model 1 requires an initial investment for development of \$200,000, and model 2 requires an investment of \$175,000. If demand is high for model 1, the company will gain \$500,000 in revenue with a net profit of \$300,000; it will receive only \$160,000 if demand is low, resulting in a net profit of -\$40,000. Based on the probabilities of demand, the expected profit is \$198,000. For model 2, we see that the expected profit is only \$188,000. Therefore, the best decision is to select model 1. Clearly, there is risk in either decision, but on an expected value basis, model 1 is the best decision.

Now suppose that the firm conducts a market research study to obtain sample information and to better understand the nature of consumer demand. Analysis of past market research studies, conducted prior to introducing similar products, has found that 90% of all products that resulted in high consumer demand had previously received a high survey response, while only 20% of all products with ultimately low consumer demand had previously received a high survey response. These probabilities show that the market research is not always accurate and can lead to a false indication of the true market potential. However, we should expect that a high survey response would increase the historical probability of high demand, while a low survey response



**FIGURE 11.8** Decision Tree for Cell Phone Options

would increase the historical probability of a low demand. Thus, we need to compute the conditional probabilities:

$$\begin{aligned} P(\text{High demand} \mid \text{High survey response}) \\ P(\text{High demand} \mid \text{Low survey response}) \\ P(\text{Low demand} \mid \text{High survey response}) \\ P(\text{Low demand} \mid \text{Low survey response}) \end{aligned}$$

This can be accomplished using Bayes's rule.

### Conditional Probabilities and Bayes's Rule

Bayes's rule extends the concept of conditional probability to revise historical probabilities based on sample information. Define the events:

$$\begin{aligned} A_1 &= \text{High consumer demand} \\ A_2 &= \text{Low consumer demand} \\ B_1 &= \text{High survey response} \\ B_2 &= \text{Low survey response} \end{aligned}$$

We need to compute  $P(A_i \mid B_j)$  for each  $i$  and  $j$ .

Using these definitions and the information presented, we have:

$$\begin{aligned} P(A_1) &= 0.7 \\ P(A_2) &= 0.3 \\ P(B_1 \mid A_1) &= 0.9 \\ P(B_1 \mid A_2) &= 0.2 \end{aligned}$$

This implies that  $P(B_2 \mid A_1) = 1 - P(B_1 \mid A_1) = 0.1$  and  $P(B_2 \mid A_2) = 1 - P(B_1 \mid A_2) = 0.8$  because the probabilities  $P(B_1 \mid A_i) + P(B_2 \mid A_i)$  must add to one for each  $A_i$ .

It is important to carefully distinguish between  $P(A \mid B)$  and  $P(B \mid A)$ . As stated, *among all products that resulted in high consumer demand, 90% received a high market survey response*. Thus, the probability of a high survey response *given* high consumer demand is 0.90, and not the other way around.

Suppose that  $A_1, A_2, \dots, A_k$  is a set of mutually exclusive and collectively exhaustive events, and we seek the probability that some event  $A_i$  occurs given that another event  $B$  has occurred. We will use the conditional probability formula introduced in Chapter 3 to establish Bayes's rule:

$$P(A_i \mid B) = P(A_i \text{ and } B) / P(B)$$

In Chapter 3, we observed that  $P(A_i \text{ and } B) = P(B \mid A_i) \times P(A_i)$ . For example,  $P(A_1 \text{ and } B_1) = (0.9)(0.7) = 0.63$ . Thus, we have the joint and marginal probabilities:

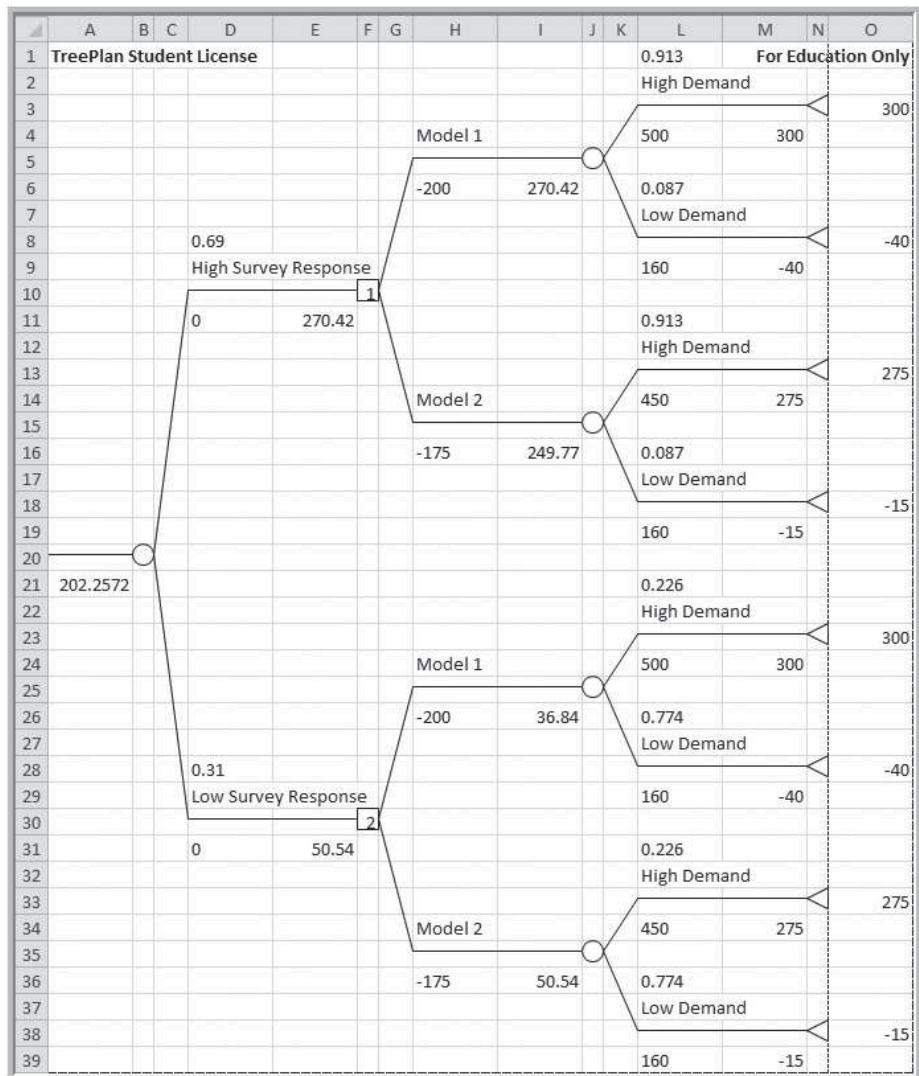
High Consumer Demand ( $A_1$ )	Low Consumer Demand ( $A_2$ )	Marginal Probabilities
High survey response ( $B_1$ )	0.06	0.69
Low survey response ( $B_2$ )	0.24	0.31
Marginal probabilities	0.7	0.3

The marginal probabilities state that there is a 69% chance that the survey will return a high demand response, and there is a 31% chance that the survey will result in a

low demand response. From this table, we may compute the conditional probabilities  $P(A_i|B_j)$  as  $P(A_i \text{ and } B_j)/P(B_j)$ . For example,  $P(A_1|B_1) = 0.63/0.69 = 0.913$ , and  $P(A_1|B_2) = 0.07/0.31 = 0.226$ . Thus,  $P(A_2|B_1) = 1 - 0.913 = 0.087$ , and  $P(A_2|B_2) = 1 - 0.226 = 0.774$ . Although 70% of all previous new models historically had high demand, knowing that the marketing report is favorable increases the likelihood to 91.3%, and if the marketing report is unfavorable, then the probability of low demand increases to 77%.

A formal statement of Bayes's rule can be derived as follows. Substituting the formula for the joint probability into the numerator of the conditional probability formula, we obtain:

$$P(A_i|B) = \frac{P(B|A_i) \times P(A_i)}{P(B)} \quad (11.6)$$



**FIGURE 11.9** Decision Tree with Sample Market Survey

If we substitute the general formula for computing marginal probabilities for  $P(B)$  in the denominator, we may find the probability of event  $A_i$  given event  $B$  as:

$$P(A_i | B) = \frac{P(B | A_i) \times P(A_i)}{P(B | A_1) \times P(A_1) + P(B | A_2) \times P(A_2) + \cdots + P(B | A_k) \times P(A_k)} \quad (11.7)$$

This is a statement of Bayes's rule. Note that this simply expresses the calculations we performed using the joint and marginal probabilities in the example into one succinct formula.

Using Bayes's rule, the probability that a new phone will have high consumer demand given that the marketing survey has a high response is:

$$P(A_1 | B_1) = \frac{P(B_1 | A_1) \times P(A_1)}{P(B_1 | A_1) \times P(A_1) + P(B_1 | A_2) \times P(A_2)}$$

$$P(A_1 | B_1) = \frac{(0.9)(0.7)}{(0.9)(0.7) + (0.2)(0.3)} = 0.913$$

Figure 11.9 shows a decision tree that incorporates the market survey information. If the survey response is high, then select model 1; if the response is low, then select model 2. Note that the expected value (which includes the probabilities of obtaining the survey responses) is \$202,257 thousand. Comparing this to Figure 11.8, we see that the sample information increases the expected value by  $\$202,257 - \$198,000 = \$4,257$ . This is the value of EVSI. So we should not pay more than \$4,257 to conduct the market survey.

## UTILITY AND DECISION MAKING

A typical charity raffle involves selling one thousand \$100 tickets to win a \$40,000 automobile. The probability of winning is only 0.001, and the expected payoff is  $(-\$100)(.999) + (\$40,000)(.001) = -\$59.90$ . From a pure economic standpoint, this would be a poor gamble. Nevertheless, many people would take this chance because the financial risk is low (and it's for charity!). On the other hand, if only 10 tickets were sold at \$5,000 with a chance to win \$100,000, even though the expected value would be  $(-\$5000)(0.9) + (\$100,000)(0.1) = \$5500$ , most people would *not* take the chance because of the higher monetary risk involved.

An approach for assessing risk attitudes quantitatively is called **utility theory**. This approach quantifies a decision maker's relative preferences for particular outcomes. We can determine an individual's utility function by posing a series of decision scenarios. This is best illustrated with an example; we will use a personal investment problem to do this.

Suppose that you have \$10,000 to invest and are expecting to buy a new car in a year, so you can tie the money up for only 12 months. You are considering three options: a bank CD paying 4%, a bond mutual fund, and a stock fund. Both the bond and stock funds are sensitive to changing interest rates. If rates remain the same over the coming year, the share price of the bond fund is expected to remain the same, and you expect to earn \$840. The stock fund would return about \$600 in dividends and capital gains. However, if interest rates rise, you can anticipate losing about \$500 from the bond fund after taking into account the drop in share price, and likewise expect to lose \$900 from the stock fund. If interest rates fall, however, the yield from the bond fund would be \$1,000 and the stock fund would net \$1,700. Table 11.2 summarizes the payoff table for this decision problem.

The decision could result in a variety of payoffs, ranging from a profit of \$1,700 to a loss of \$900. The first step in determining a utility function is to rank-order the payoffs

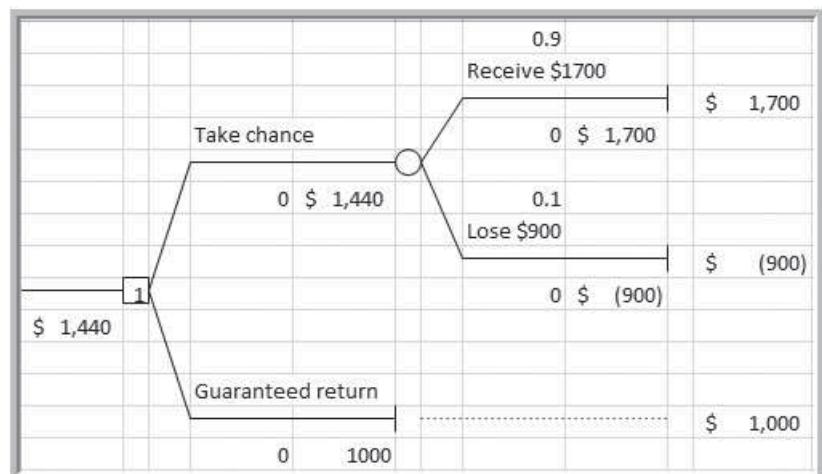
**TABLE 11.2 Investment Return Payoff Table**

Decision/Event	Rates Rise	Rates Stable	Rates Fall
Bank CD	\$400	\$400	\$400
Bond fund	-\$500	\$840	\$1,000
Stock fund	-\$900	\$600	\$1,700

from highest to lowest. We arbitrarily assign a utility of 1.0 to the highest payoff and a utility of zero to the lowest:

Payoff, $x$	Utility, $U(x)$
\$1,700	1.0
\$1,000	
\$840	
\$600	
\$400	
-\$500	
-\$900	0.0

Next, for each payoff between the highest and lowest, we present you with the following situation: Suppose you have the opportunity of achieving a *guaranteed return* of  $x$ , or taking a chance of receiving \$1,700 (the highest payoff) with probability  $p$  and losing \$900 (the lowest payoff) with probability  $1 - p$ . What value of  $p$  would make you indifferent to these two choices? Let us start with  $x = \$1,000$ . This is illustrated in the simple decision tree in Figure 11.10. Because this is a relatively high value, you decide that  $p$  would have to be at least 0.9 to take this risk. This represents the utility of a payoff of \$1,000, denoted as  $U(\$1,000)$ . We use the term **certainty equivalent** to represent the amount that a decision maker feels is equivalent to an uncertain gamble. For example, \$1,000 is this decision maker's certainty equivalent for the uncertain situation of receiving \$1,700 with probability 0.9 or -\$900 with probability 0.1.

**FIGURE 11.10** Decision Tree Lottery for Determining the Utility of \$1,000

We repeat this process for each payoff, the probabilities  $p$  that you select for each scenario from your utility function. Suppose this process results in the following:

Payoff, $x$	Utility, $U(x)$
\$1,700	1.0
\$1,000	0.90
\$840	0.85
\$600	0.80
\$400	0.75
-\$500	0.35
-\$900	0.0

If we compute the expected value of each of the gambles for the chosen values of  $p$ , we see that they are higher than the corresponding payoffs. For example, for the payoff of \$1,000 and the corresponding  $p = .9$ , the expected value of taking the gamble is:

$$0.9(\$1,700) + 0.1(-\$900) = \$1,440$$

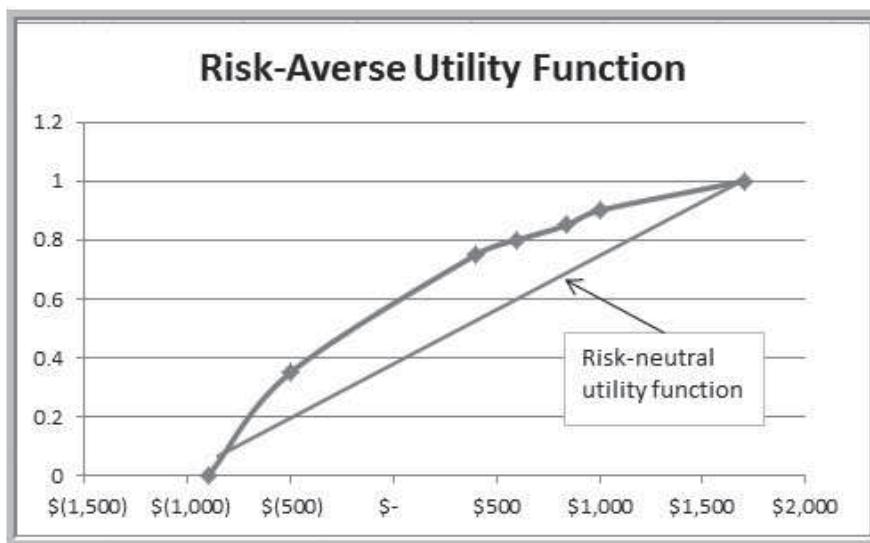
This is larger than accepting \$1,000 outright. We can interpret this to mean that you require a risk premium of  $\$1,440 - \$1,000 = \$440$  to feel comfortable enough to risk losing \$900 if you take the gamble. In general, the **risk premium** is the amount an individual is willing to forgo to avoid risk. This indicates that you are a *risk-averse individual*, that is, relatively conservative.

Another way of viewing this is to find the *break-even probability* at which you would be indifferent to receiving the guaranteed return and taking the gamble. This probability is found by solving the equation:

$$1,700p - 900(1 - p) = 1,000$$

resulting in  $p = 19/26 = 0.73$ . Because you require a higher probability of winning the gamble, it is clear that you are uncomfortable taking the risk.

If we graph the utility versus the payoffs, we can sketch a utility function as shown in Figure 11.11. This utility function is generally *concave downward*. This type of curve is



**FIGURE 11.11** Example of a Risk-Averse Utility Function

characteristic of risk-averse individuals. Such decision makers avoid risk, choosing conservative strategies and those with high return-to-risk values. Thus, a gamble must have a higher expected value than a given payoff to be preferable, or equivalently, a higher probability of winning than the breakeven value.

Other individuals might be risk taking. What would their utility functions look like? As you might suspect, they are *concave upward*. These individuals would take a gamble that offers higher rewards even if the expected value is less than a certain payoff. An example of a utility function for a risk-taking individual in this situation would be as follows:

Payoff, $x$	Utility, $U(x)$
\$1,700	1.0
\$1,000	0.6
\$840	0.55
\$600	0.45
\$400	0.40
-\$500	0.1
-\$900	0.0

For the payoff of \$1,000, this individual would be indifferent between receiving \$1,000 and taking a chance at \$1,700 with probability 0.6 and losing \$900 with probability 0.4. The expected value of this gamble is:

$$0.6(\$1,700) + 0.4(-\$900) = \$660$$

Since this is considerably less than \$1,000, the individual is taking a larger risk to try to receive \$1,700. Note that the probability of winning is less than the breakeven value. Risk takers generally prefer more aggressive strategies.

### SKILL-BUILDER EXERCISE 11.6

Generate an Excel chart for the utility function of a risk-taking individual using the data above.

Finally, some individuals are risk neutral; they prefer neither taking risks nor avoiding them. Their utility function would be linear and would correspond to the breakeven probabilities for each gamble. For example, a payoff of \$600 would be equivalent to the gamble if:

$$\$600 = p(\$1,700) + (1 - p)(-\$900)$$

Solving for  $p$  we obtain  $p = 15/26$ , or .58, which represents the utility of this payoff. The decision of accepting \$600 outright or taking the gamble could be made by flipping a coin. These individuals tend to ignore risk measures and base their decisions on the average payoffs.

A utility function may be used instead of the actual monetary payoffs in a decision analysis by simply replacing the payoffs with their equivalent utilities and then computing expected values. The expected utilities and the corresponding optimal decision strategy then reflect the decision maker's preferences toward risk. For example,

if we use the average payoff strategy (because no probabilities of events are given) for the data in Table 11.2, the best decision would be to choose the stock fund. However, if we replace the payoffs in Table 11.2 with the (risk-averse) utilities that we defined, and again use the average payoff strategy, the best decision would be to choose the bank CD as opposed to the stock fund, as shown in the table below.

Decision/Event	Rates Rise	Rates Stable	Rates Fall	Average Utility
Bank CD	0.75	0.75	0.75	0.75
Bond fund	0.35	0.85	0.9	0.70
Stock fund	0	0.80	1.0	0.60

If assessments of event probabilities are available, these can be used to compute the expected utility and identify the best decision.

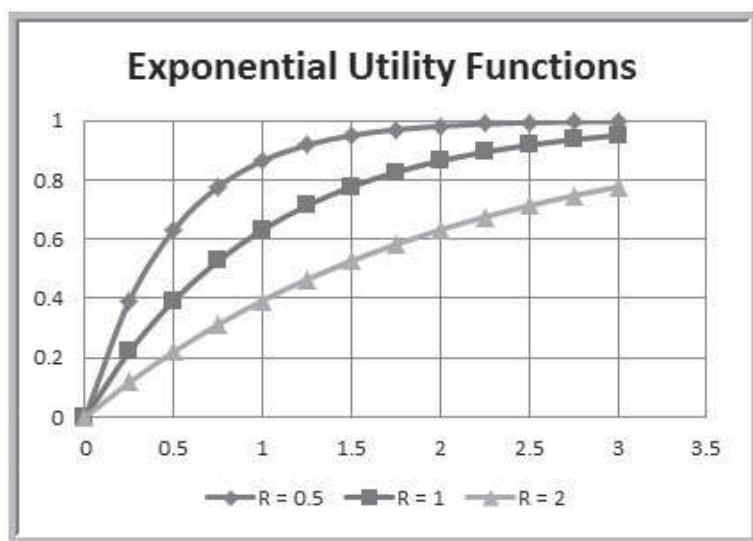
### Exponential Utility Functions

It can be rather difficult to compute a utility function, especially for situations involving a large number of payoffs. Because most decision makers typically are risk averse, we may use an exponential utility function to approximate the true utility function. The exponential utility function is:

$$U(x) = 1 - e^{-x/R} \quad (11.8)$$

where  $e$  is the base of the natural logarithm (2.71828 . . .) and  $R$  is a shape parameter. Figure 11.12 shows several examples of  $U(x)$  for different values of  $R$ . Notice that all of these functions are concave and that as  $R$  increases, the functions become flatter, indicating more tendency toward risk neutrality.

One approach to estimating a reasonable value of  $R$  is to find the maximum payoff  $\$R$  for which the decision maker is willing to take an equal chance on winning  $\$R$  or losing  $\$R/2$ . The smaller the value of  $R$ , the more risk averse is the individual. For



**FIGURE 11.12** Examples of Exponential Utility Functions

instance, would you take a bet on winning \$100 versus losing \$50? How about winning \$10,000 versus losing \$5000? Most people probably would not worry about taking the first gamble, but might definitely think twice about the second. Finding one's maximum comfort level establishes the utility function. For the example problem, suppose that  $R = \$400$ . The utility function would be  $U(x) = 1 - e^{-x/400}$ , resulting in the following utility values:

Payoff, $x$	Utility, $U(x)$
\$1,700	0.9857
\$1,000	0.9179
\$840	0.8775
\$600	0.7769
\$400	0.6321
-\$500	-2.4903
-\$900	-8.4877

Using these values in the payoff table, we find that the bank CD remains the best decision as shown in the table below.

Decision/Event	Rates Rise	Rates Stable	Rates Fall	Average Utility
Bank CD	0.6321	0.6321	0.6321	0.6321
Bond fund	-2.4903	0.8775	0.9179	-0.2316
Stock fund	-8.4877	0.7769	0.9857	-2.2417

### SKILL-BUILDER EXERCISE 11.7

Consider the investment return decision problem in Table 11.2. Find the decisions that would result from the aggressive, conservative, and opportunity loss strategies. Compare these decisions with those that would result from using the risk-averse utilities provided in the chapter instead of the payoff values, and the exponential utility function values that were calculated using  $R = \$400$ .

## Basic Concepts Review Questions

1. Describe the decisions involving single alternatives.
2. Describe the decisions involving non-mutually exclusive alternatives.
3. Describe the decisions involving mutually exclusive alternatives.
4. What are the characteristics of a scoring model?
5. Describe how to model decisions involving uncertainty.
6. Summarize the decision strategies that model different risk behaviors for making decisions involving uncertainty.
7. Explain how the standard deviation, coefficient of variation, skewness, and kurtosis of the distribution of outcomes provide information about risk.
8. Provide examples of when expected value decision making is appropriate and when it is not.
9. Explain the concepts of expected value of sample information and expected value of perfect information.
10. What are the implications of the “flaw of averages”?
11. Discuss the exponential utility function and its characteristics.
12. What is a risk profile and how can it be used in conjunction with the solution of a decision tree?
13. Explain the concept of utility theory and how utilities can be assessed.
14. What is a decision tree? How is Bayes’s rule useful in implementing a decision tree?

## Problems and Applications

Note: Data for selected problems can be found in the Excel file **Chapter 11 Problem Data** to facilitate your problem-solving efforts. Worksheet tabs correspond to the problem numbers.

1. A company has estimated that a proposed \$10,000 investment will generate \$3,250 for each of the next four years.
  - a. What is the payback period?
  - b. If the required rate of return is 9%, use internal rate of return to determine whether or not this proposal should be accepted.
2. An e-commerce firm is developing a new application. Financial analysts have estimated the expenses and revenues over the next five years:

Month	Development Expense	Operating Expense	Revenue
Initial investment	\$50,000.00	\$—	\$—
Year 1	\$10,000.00	\$10,000.00	\$5,000.00
Year 2	\$—	\$10,000.00	\$15,000.00
Year 3	\$—	\$10,000.00	\$25,000.00
Year 4	\$—	\$10,000.00	\$50,000.00
Year 5	\$—	\$10,000.00	\$50,000.00

The company's discount rate is 8%. Compute the NPV and IRR for net profit and make a recommendation on whether or not to pursue the project. Then, use a data table to evaluate the impact of changing the initial investment in increments of \$5,000 between \$30,000 and \$70,000. What might this mean with regard to the company's decision?

3. The finance committee at Olson, Inc. has 13 proposed new technology projects under consideration. Estimated initial investments, external revenues, and internal cost reductions are:

Project	Investment	Revenue Impact	Cost Reduction
A1	\$175,600.00	\$—	\$200,358.00
A2	\$126,512.00	\$422,580.00	\$(103,420.00)
A3	\$198,326.00	\$415,625.00	\$(226,413.00)
B4	\$421,618.00	\$—	\$486,312.00
B5	\$322,863.00	\$—	\$456,116.00
B6	\$398,810.00	\$—	\$508,213.00
B7	\$212,506.00	\$—	\$356,067.00
C8	\$813,620.00	\$416,283.00	\$386,229.00
C9	\$850,418.00	\$583,260.00	\$398,014.00
D10	\$522,615.00	\$916,426.00	\$(155,106.00)
D11	\$486,283.00	\$816,420.00	\$(103,210.00)
D12	\$683,407.00	\$758,420.00	\$(75,896.00)
D13	\$722,813.00	\$950,128.00	\$(120,063.00)

Note that a positive cost reduction increases revenue, while a negative cost reduction implies additional costs to the firm. Using ROI, determine which projects should be selected for further consideration by the committee. How would the committee rank the projects using a benefit/cost analysis?

4. The cost accountant of a large truck fleet is evaluating options for dealing with a large volume of flat tires. Currently, the company repairs tires on the open market by having the driver take the flat tire to the nearest tire dealer. Last year, this cost an average of \$30 per flat tire. The volume of flat tires experienced per year was 10,000 last year, and the expected rate of growth in flat tires is 10% per year. However, some feel that flat tire growth will be as low as 5%; others as high as 15%. A complicating factor is that the cost to repair a tire grows an average of 3% per year.

The company has two alternatives. A tire dealer has offered to fix all the company's flat tires for a fixed rate of \$36 per tire over a three-year period. The other alternative is for the company to go into the tire repair business for themselves. This option is expected to require an investment in equipment of \$200,000, with a salvage value of \$50,000 after three years. It would require an overhead expense of \$40,000 per year in the first year, \$45,000 the second year, and \$50,000 the third year. The variable cost for fixing a flat tire is \$12 per tire for the three-year period of analysis. Compare the net present costs using a discount rate of 8% over three years for each of these three options under conditions of tire growth rate ranging from 5% to 15% per year in increments of 1%. What is the best option under each scenario? What risk factors should be considered?

5. Your neighbor's son or daughter is starting to look at college choices and he has asked you to help make a good choice. You have suggested that a scoring model would be useful in helping to narrow down the choices. Develop a scoring model on a spreadsheet by defining a set of attributes and levels of those attributes that would be appropriate for evaluating a prospective college. Use your own retrospective experience in developing the model.
6. Slaggert Systems is considering becoming certified to the ISO 9000 series of quality standards. Becoming certified is expensive, but the company could lose a substantial amount of business if its major customers suddenly demand ISO certification and the company does not have it. At a management retreat, the senior executives of the firm developed the following payoff table, indicating the NPV of profits over the next five years. What decision should they make under the average payoff, aggressive, conservative, and opportunity loss decision strategies?

<b><i>Customer Response</i></b>		
	<b>Standards Required</b>	<b>Standards Not Required</b>
Become certified	\$600,000	\$475,000
Stay uncertified	\$325,000	\$525,000

7. The DoorCo Corporation is a leading manufacturer of garage doors. All doors are manufactured in their plant in Carmel, Indiana, and shipped to distribution centers or major customers. DoorCo recently acquired another manufacturer of garage doors, Wisconsin Door, and is considering moving its wood door operations to the Wisconsin plant. A key consideration in this decision is the transportation and production costs at the two plants and the new construction and relocation costs. Complicating matters is the fact that marketing is predicting a decline in the demand for wood doors. The company developed three scenarios:

1. Demand falls slightly, with no noticeable effect on production.
2. Demand and production decline 20%.
3. Demand and production decline 45%.

The table below shows the total costs under each decision and scenario.

	<b>Slight Decline</b>	<b>20% Decline</b>	<b>40% Decline</b>
Stay in Carmel	\$980,000	\$830,000	\$635,000
Move to Wisconsin	\$990,000	\$835,000	\$630,000

- a. What decision should DoorCo make using the average payoff, aggressive, conservative, and opportunity loss decision strategies discussed in this chapter?
- b. Suppose the probabilities of the three scenarios are estimated to be 0.15, 0.40, and 0.45, respectively. Construct a decision tree and compute the rollback values to find the best expected value decision.
8. Suppose that a car rental agency offers insurance for a week that will cost \$10 per day. A minor fender bender will cost \$1,500, while a major accident might cost \$15,000 in repairs. Without the insurance, you would be personally liable for any damages. What should you do? Clearly, there are two decision alternatives: take the insurance or do not take the insurance. The uncertain consequences, or events that might occur, are that you would not be involved in an accident, that you would be involved in a fender bender, or that you would be involved in a major accident. Assume that you researched insurance industry statistics and found out that the probability of major accident is 0.05%, and that the probability of a fender bender is 0.16%. What is the expected value decision? Would you choose this? Why or why not? What would be some alternate ways to evaluate risk?

9. An investor can invest in three highly speculative opportunities. The returns and standard deviations are given here.

	<b>Expected Return</b>	<b>Standard Deviation</b>
Investment A	\$50,000	\$25,000
Investment B	\$40,000	\$24,000
Investment C	\$30,000	\$10,000

Based on the return to risk, which of these is the best investment?

10. An information system consultant is bidding on a project that involves some uncertainty. Based on past experience, if all went well (probability 0.1), the project would cost \$1.2 million to complete. If moderate debugging were required (probability 0.7), the project would probably cost \$1.4 million. If major problems were encountered (probability 0.2), the project could cost \$1.8 million. Assume that the firm is bidding competitively, and the expectation of successfully gaining the job at a bid of \$2.2 million is 0, at \$2.1 million is 0.1, at \$2.0 million is 0.2, at \$1.9 million is 0.3, at \$1.8 million is 0.5, at \$1.7 million is 0.8, and at \$1.6 million is practically certain.
  - a. Calculate the expected monetary value for the given bids.
  - b. What is the best bidding decision?
  - c. What is the expected value of perfect information?
11. Mountain Ski Sports, a chain of ski equipment shops in Colorado, purchases skis from a manufacturer each summer for the coming winter season. The most popular intermediate model costs \$150 and sells for \$260. Any skis left over at the end of the winter are sold at the store's half-price sale (for \$130). Sales over the years are quite stable. Gathering data from all its stores, Mountain Ski Sports developed the following probability distribution for demand:

<b>Demand</b>	<b>Probability</b>
150	0.05
175	0.20
200	0.35
225	0.30
250	0.10

The manufacturer will take orders only for multiples of 20, so Mountain Ski is considering the following order sizes: 160, 180, 200, 220, and 240.

- a. Construct a payoff table for Mountain Ski's decision problem of how many pairs of skis to order. What is the best decision from an expected value basis?
- b. Find the expected value of perfect information.
- c. What is the expected demand? Is the optimal order quantity equal to the expected demand? Why?

- 12.** Bev's Bakery specializes in sourdough bread. Early each morning, Bev must decide how many loaves to bake for the day. Each loaf costs \$0.75 to make and sells for \$2.85. Bread left over at the end of the day can be sold the next day for \$1.00. Past data indicate that demand is distributed as follows:

Number of Loaves	Probability
15	0.05
16	0.05
17	0.10
18	0.10
19	0.20
20	0.35
21	0.10
22	0.05

- a. Construct a payoff table and determine the optimal quantity for Bev to bake each morning using expected values.
  - b. What is the optimal quantity for Bev to bake if the unsold loaves cannot be sold to the day-old store at the end of the day (so that unsold loaves are a total loss)?
- 13.** An investor is considering a two-asset portfolio. Stock A has an expected return of \$4.50 per share with a standard deviation of \$1.00, while stock B has an expected return of \$3.75 with a standard deviation of \$0.75. The covariance between the two stocks is  $-0.35$ . Find the portfolio risk if:
- a. The stocks are weighted equally in the portfolio
  - b. The amount of stock A is one-fourth as much as stock B
  - c. The amount of stock B is one-fourth as much as stock A
- 14.** A patient arrives at an emergency room complaining of abdominal pain. The ER physician must decide on whether to operate or to place the patient under observation for a non-appendix-related condition. If an appendectomy is performed immediately, the doctor runs the risk that the patient does not have appendicitis. If it is delayed and the patient does indeed have appendicitis, the appendix might perforate, leading to a more severe case and possible complications. However, the patient might recover without the operation.
- a. Construct a decision tree for the doctor's dilemma.
  - b. How might payoffs be determined?
  - c. Would utility be a better measure of payoff than actual costs? If so, how might utilities be derived for each path in the tree?
- 15.** Midwestern Hardware must decide how many snow shovels to order for the coming snow season. Each shovel costs \$15.00 and is sold for \$29.95. No inventory is carried from one snow season to the next. Shovels unsold after February are sold at a discount price of \$10.00. Past data

indicate that sales are highly dependent on the severity of the winter season. Past seasons have been classified as mild or harsh, and the following distribution of regular price demand has been tabulated:

<i>Mild Winter</i>		<i>Harsh Winter</i>	
No. of Shovels	Probability	No. of Shovels	Probability
250	0.5	1,500	0.2
300	0.4	2,500	0.4
350	0.1	3,000	0.4

Shovels must be ordered from the manufacturer in lots of 200. Construct a decision tree to illustrate the components of the decision model, and find the optimal quantity for Midwestern to order if the forecast calls for a 70% chance of a harsh winter.

- 16.** Perform a sensitivity analysis of the Midwestern Hardware scenario (Problem 15). Find the optimal order quantity and optimal expected profit for probabilities of a harsh winter ranging from 0.2 to 0.8 in increments of 0.2. Plot optimal expected profit as a function of the probability of a harsh winter.
- 17.** Dean Kuroff started a business of rehabbing old homes. He recently purchased a circa-1800 Victorian mansion and converted it into a three-family residence. Recently, one of his tenants complained that the refrigerator was not working properly. Since Dean's cash flow was not extensive, he was not excited about purchasing a new refrigerator. He is considering two other options: purchase a used refrigerator or repair the current unit. He can purchase a new one for \$400, and it will easily last three years. If he repairs the current one, he estimates a repair cost of \$150, but he also believes that there is only a 30% chance that it will last a full three years and he will end up purchasing a new one anyway. If he buys a used refrigerator for \$200, he estimates that there is a 0.6 probability that it will last at least three years. If it breaks down, he will still have the option of repairing it for \$150 or buying a new one. Develop a decision tree for this situation and determine Dean's optimal strategy.
- 18.** Many automobile dealers advertise lease options for new cars. Suppose that you are considering three alternatives:
1. Purchase the car outright with cash.
  2. Purchase the car with 20% down and a 48-month loan.
  3. Lease the car.
- Select an automobile whose leasing contract is advertised in a local paper. Using current interest rates and advertised leasing arrangements, perform a decision analysis of these options. Make, but clearly define, any assumptions that may be required.
- 19.** Drilling decisions by oil and gas operators involve intensive capital expenditures made in an environment characterized by limited information and high risk.

A well site is dry, wet, or gushing. Historically, 50% of all wells have been dry, 30% wet, and 20% gushing. The value (net of drilling costs) for each type of well is as follows:

Dry	-\$80,000
Wet	\$100,000
Gushing	\$200,000

Wildcat operators often investigate oil prospects in areas where deposits are thought to exist by making geological and geophysical examinations of the area before obtaining a lease and drilling permit. This often includes recording shock waves from detonations by a seismograph and using a magnetometer to measure the intensity of the Earth's magnetic effect to detect rock formations below the surface. The cost of doing such studies is approximately \$15,000. Of course, one may choose to drill in a location based on "gut feel" and avoid the cost of the study. The geological and geophysical examination classify an area into one of three categories: no structure (NS), which is a bad sign; open structure (OS), which is an "OK" sign; and closed structure (CS), which is hopeful. Historically, 40% of the tests have resulted in NS, 35% resulted in OS, and 25% resulted in CS readings. After the result of the test is known, the company may decide not to drill. The following table shows probabilities that the well will actually be dry, wet, or gushing based on the classification provided by the examination (in essence, the examination cannot accurately predict the actual event):

	<b>Dry</b>	<b>Wet</b>	<b>Gushing</b>
NS	0.73	0.22	0.05
OS	0.43	0.34	0.23
CS	0.23	0.372	0.398

- a. Construct a decision tree of this problem that includes the decision of whether or not to perform the geological tests.
  - b. What is the optimal decision under expected value when no experimentation is conducted?
  - c. Find the overall optimal strategy by rolling back the tree.
20. Hahn Engineering is planning on bidding on a job and often competes against a major competitor, Sweigart and Associates (S&A), as well as other firms. Historically, S&A has bid for the same jobs 80% of the time. If S&A bids on a job, the probability that Hahn Engineering will win it is 0.30. If S&A does not bid on a job, the probability that Hahn will win the bid is 0.60. Apply Bayes's rule

to find the probability that Hahn Engineering will win the bid. If they do, what is the probability that S&A did bid on it?

21. MJ Logistics has decided to build a new warehouse to support its supply chain activities. They have the option of building either a large warehouse or a small one. Construction costs for the large facility are \$8 million versus \$5 million for the small facility. The profit (excluding construction cost) depends on the volume of work the company expects to contract for in the future. This is summarized in the table below (in millions of dollars):

	<b>High Volume</b>	<b>Low Volume</b>
Large Warehouse	\$35	\$20
Small Warehouse	\$15	\$9

The company believes that there is a 60% chance that the volume of demand will be high.

- a. Construct a decision tree to identify the best choice.
  - b. Suppose that the company engages some economic experts to provide their opinion about the future economic conditions. Historically, their upside predictions have been 75% accurate, while their downside predictions have been 90% accurate. Determine the best strategy if their predictions suggest that the economy will improve or will deteriorate. What is the EVSI? What is EVPI?
22. Consider the car rental insurance scenario in Problem 8. Use the approach described in this chapter to develop your personal utility function for the payoffs associated with this decision. Determine the decision that would result using the utilities instead of the payoffs. Is the decision consistent with your choice?
23. A college football team is trailing 14–0 late in the game. The team is getting close to making a touchdown. If they can score now, hold the opponent, and score one more time, they can tie or win the game. The coach is wondering whether to go for an extra-point kick or a two-point conversion now, and what to do if they can score again.
- a. Develop a decision tree for the coach's decision. Develop a utility function to represent the final score for each path in the tree.
  - b. Estimate probabilities for successful kicks or two-point conversions. (You might want to do this by some group brainstorming or by calling on experts, such as your school's coach or a sports journalist.) Using the probabilities and utilities from part (a), determine the optimal strategy.
  - c. Perform a sensitivity analysis on the probabilities to evaluate alternative strategies (such as when the starting kicker is injured).

# Case

## The Sandwich Decision

A national restaurant chain has developed a new specialty sandwich. Initially, it faces two possible decisions: introduce the sandwich nationally at a cost of \$200,000 or evaluate it in a regional test market at a cost of \$30,000. If it introduces the sandwich nationally, the chain might find either a high or low response to the idea. Probabilities of these events are estimated to be 0.6 and 0.4, respectively. With a high response, gross revenues of \$700,000 (at NPV) are expected; with a low response, the figure is \$150,000. If it starts with a regional marketing strategy, it might find a low response or a high response at the regional level with probabilities 0.3 and 0.7, respectively. This may or may not reflect the national market potential. In any case, the chain next needs to decide whether to remain regional, market nationally, or drop the product. If the regional response is high and it remains regional, the expected revenue is \$200,000. If it markets nationally (at an additional cost of \$200,000), the probability of a high national response is 0.9 with revenues of \$700,000

(\$150,000 if the national response is low). If the regional response is low and it remains regional, the expected revenue is \$100,000. If it markets nationally (at an additional cost of \$200,000), the probability of a high national response is 0.05 with revenues of \$700,000 (\$150,000 if the national response is low).

- a. Using *TreePlan*, construct a decision tree and determine the optimal strategy.
- b. Conduct sensitivity analyses for the probability estimates using both one- and two-way data tables as appropriate.
- c. Develop the risk profile associated with the optimal strategy.
- d. Evaluate the risk associated with this decision, considering that it is a one-time decision.
- e. Summarize all your results, including your recommendation and justification for it, in a formal report to the executive in charge of making this decision.

## APPENDIX 11.1

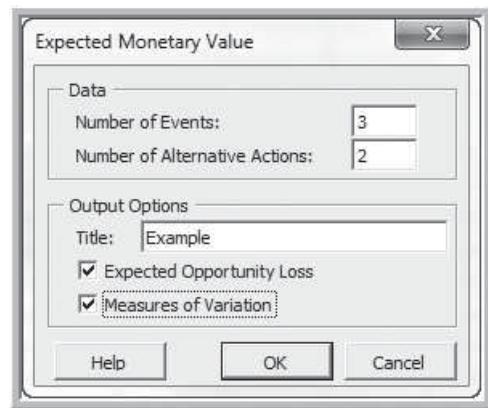
### Excel, PHStat, and TreePlan Notes

#### A. Using the IRR Function

The Excel function for internal rate of return is  $\text{IRR}(\text{values}, \text{guess})$ . *Values* represents the series of cash flows (at least one of which must be positive and one of which must be negative). *Guess* is a number believed close to the value of IRR that is used to facilitate the mathematical algorithm used to find the solution. Occasionally, the function might not converge to a solution; in those cases, you should try a different value for *guess*. In most cases, the value of *guess* can be omitted from the function.

#### B. Using the Expected Monetary Value Tool

Click *PHStat* from the Excel Add-Ins tab and select *Decision Making* and then *Expected Monetary Value* from the menu. The dialog box is shown in Figure 11A.1; you need only specify the number of actions (alternatives) and events. *PHStat* creates a worksheet in which you must enter your data (see Figure 11A.2). Note that the events correspond to rows and decisions to columns, which is the opposite of



**FIGURE 11A.1** *PHStat* Dialog for EMV

the way it is presented in this text and most other books. You may customize the worksheet to change the row and column labels in the Probabilities & Payoffs Table for your

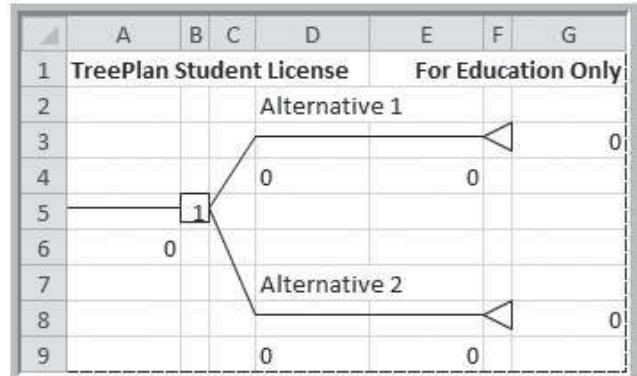
	A	B	C	D
1	Example			
2				
3	Probabilities & Payoffs Table:			
4		P	A1	A2
5	E1			
6	E2			
7	E3			

**FIGURE 11A.2** PHStat EMV Template

specific problem. After you enter the data, the expected values and other statistical information are automatically computed. The tool also computes the opportunity loss table and the expected opportunity loss for each decision alternative. EVPI is the expected value of perfect information, and is discussed in the text.

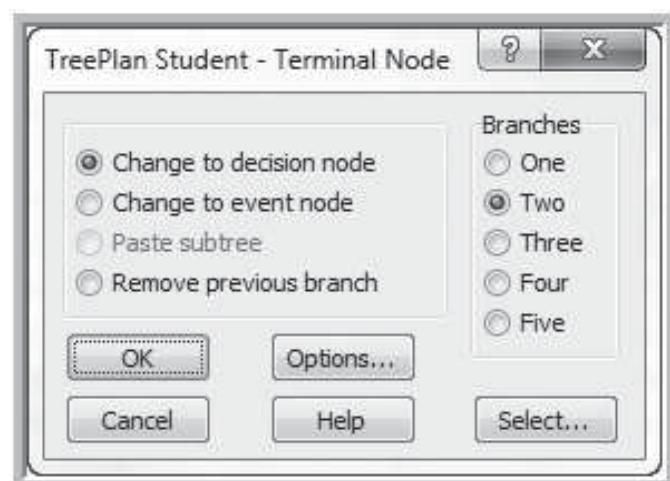
### C. Constructing Decision Trees in Excel

To use *TreePlan* within Excel, first select the upper left corner of the worksheet where you wish to draw the tree. Note that *TreePlan* writes over existing values in the spreadsheet; therefore, begin your tree to the *right* of the area where your data are stored and do not subsequently add or delete rows or columns in the tree-diagram area. Double-click the *TreePlan XLA* file (the actual name of the file depends on the latest version accompanying this text) and allow Excel to enable macros. *Decision Tree* (*Ctrl + Shift + T*) will appear in the *Add-Ins* tab of Excel; click on this to start *TreePlan*. *TreePlan* then prompts you with a dialog box with three options: *New Tree*, *Cancel*, and *Help*; choose *New Tree* to begin a new tree. *TreePlan* will then draw a default initial decision tree with its upper left corner at the selected cell as shown in Figure 11A.3.

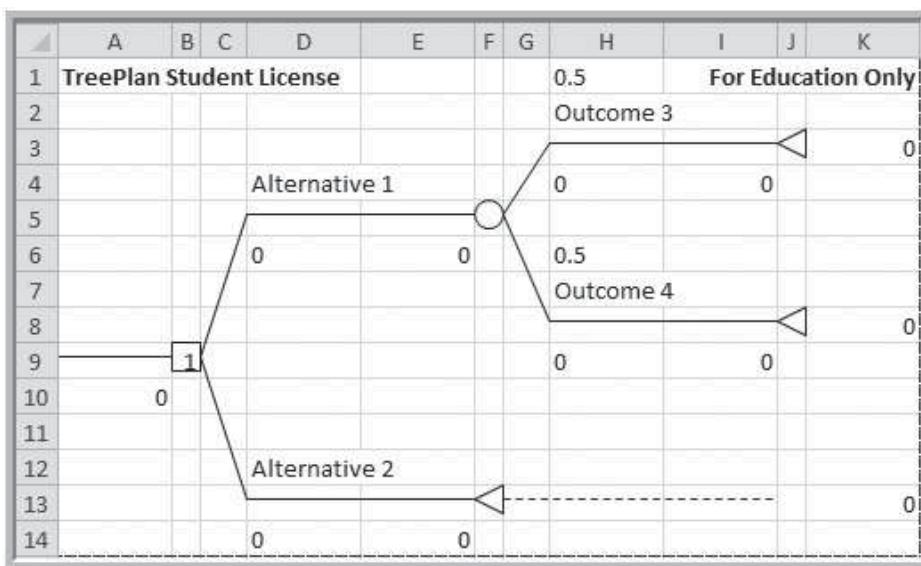


**FIGURE 11A.3** *TreePlan* Initial Decision Tree Structure

Expand a tree by adding or modifying branches or nodes in the default tree. To change the branch labels or probabilities, click on the cell containing the label or probability and type the new label or probability. To modify the structure of the tree (e.g., add or delete branches or nodes in the tree), select the node or branch in the tree to modify and select *Decision Tree* from the *Add-Ins* tab or simply press the key combination *Ctrl + Shift + T*. *TreePlan* will then present a dialog showing the available commands. For example, to add an event node to the top branch of the tree in Figure 11A.3, select the terminal node at the end of that branch (F3) and press *Ctrl + Shift + T*. *TreePlan* then presents the dialog shown in Figure 11A.4. To add an event node to the branch, change the selected terminal node to an event node by selecting *Change to event node* in the dialog box, select the number of branches (two), and press *OK*. *TreePlan* then redraws the tree with a chance node in place of the terminal node, as shown in Figure 11A.5. The dialog boxes presented by *TreePlan* vary depending on what is selected. For instance, if you select



**FIGURE 11A.4** *TreePlan* Dialog



**FIGURE 11A.5** Expanded Decision Tree

an event node, a different dialog box appears, allowing you options to add a branch, insert another event, change it to a decision, and so on. When building large trees, the *Copy subtree* option allows you to copy a selected node and everything to the right of it. You can then select a terminal node and choose *Paste subtree*.

The *Options* button in a *TreePlan* dialog box allows you to use expected values (default) or an exponential utility function and to select whether the objective is to maximize (default) or minimize. Exponential utility functions convert cash flows into “utilities,” or scaled payoffs for individuals with risk-averse attitudes.

## *Chapter 12*

# Queues and Process Simulation Modeling

- INTRODUCTION 402
- QUEUES AND QUEUING SYSTEMS 403
  - Basic Concepts of Queueing Systems 403
  - Customer Characteristics 404
  - Service Characteristics 405
  - Queue Characteristics 405
  - System Configuration 405
  - Performance Measures 406
- ANALYTICAL QUEUING MODELS 406
  - Single-Server Model 407
  - Little's Law 408
- PROCESS SIMULATION CONCEPTS 409
- PROCESS SIMULATION WITH *SIMQUICK* 410
  - Getting Started with *SimQuick* 411
  - A Queuing Simulation Model 412
  - Queues in Series with Blocking 417
  - Grocery Store Checkout Model with Resources 418
  - Manufacturing Inspection Model with Decision Points 421
  - Pull System Supply Chain with Exit Schedules 424
  - Other *SimQuick* Features and Commercial Simulation Software 426
- CONTINUOUS SIMULATION MODELING 427
- BASIC CONCEPTS REVIEW QUESTIONS 430
- PROBLEMS AND APPLICATIONS 431
- CASE: PRODUCTION/INVENTORY PLANNING 434

## **INTRODUCTION**

Many production and service operations involve the flow of some type of entity through a system over time. The entity might be a customer, a physical object, or a piece of information. Some common examples are people being served at a driver's license bureau, jobs being processed in

a factory, messages moving through a communication system, and calls being processed at a call center. Most of these situations involve *queuing*, or waiting for service to occur. Decisions regarding such operations often involve determining the best design configuration or operating policies of the system to reduce customer wait time or to improve the efficiency of the operation.

Modeling these situations is considerably more difficult than modeling other types of problems, primarily because the sequence of events over time must be explicitly taken into account. While some simple situations are amenable to analytical models, most practical systems are modeled and analyzed using process simulation. **Process simulation** is an approach for modeling the logical sequence of events as they take place. In this chapter, we introduce basic concepts of queuing systems and process simulation modeling using a simple Excel-based software package, *SimQuick*.<sup>1</sup>

## QUEUES AND QUEUING SYSTEMS

Waiting lines occur in many important business operations, as well as in everyday life. Most service systems, such as fast-food restaurants, banks, gasoline stations, and technical support telephone hotlines, involve customer waiting. In these systems, customers arrive at random times, and service times are rarely predictable. Managers of these systems would be interested in how long customers have to wait, the length of waiting lines, use of the servers, and other measures of performance. The important issue in designing such systems involves the trade-off between customer waiting and system cost, usually determined by the number of servers. A design that balances the average demand with the average service rate will cause unacceptable delays. The decision is difficult because the marginal return for increasing service capacity declines. For example, a system that can handle 99% of expected demand will cost much more than a system designed to handle 90% of expected demand. In this section, we discuss the basic components of waiting line models and illustrate a process-driven simulation model for a simple case.

### Basic Concepts of Queuing Systems

The analysis of waiting lines, called *queuing theory*, applies to any situation in which customers arrive to a system, wait, and receive service. Queuing theory had its origins in 1908, with a Danish telephone engineer, A.K. Erlang, who began to study congestion in the telephone service of the Copenhagen Telephone Company. Erlang developed mathematical formulas to predict waiting times and line lengths. Over the years, queuing theory has found numerous applications in telecommunications and computer systems and has expanded to many other service systems. The objectives of queuing theory are to improve customer service and reduce operating costs. As consumers began to differentiate firms by their quality of service, reducing waiting times has become an obsession with many firms. Many restaurant and department store chains take waiting seriously—some have dedicated staffs who study ways to speed up service.

All queuing systems have three elements in common:

1. **Customers waiting for service.** Customers need not be people but can be machines awaiting repair, airplanes waiting to take off, subassemblies waiting for a machine, computer programs waiting for processing, or telephone calls awaiting a customer service representative.
2. **Servers providing the service.** Again, servers need not be only people, such as clerks, customer service representatives, or repairpersons; servers may be airport runways, machine tools, repair bays, ATMs, or computers.

---

<sup>1</sup> David Hartvigsen, *SimQuick, Process Simulation with Excel*, 2nd Ed. (Upper Saddle River, NJ: Pearson Prentice Hall, 2004).

3. A **waiting line or queue**. The queue is the set of customers waiting for service. In many cases, a queue is a physical line, as you experience in a bank or grocery store. In other situations, a queue may not even be visible or in one location, as with computer jobs waiting for processing or telephone calls waiting for an open line.

To understand the operation of a queuing system, we need to describe the characteristics of the customer, server, and queue, and how the system is configured.

## Customer Characteristics

Customers arrive to the system according to some *arrival process*, which can be deterministic or probabilistic. Examples of deterministic arrivals would be parts feeding from an automated machine to an assembly line or patients arriving at appointed times to a medical facility. Most arrival processes, such as people arriving at a supermarket, are probabilistic. We can describe a probabilistic arrival process by a probability distribution representing the number of arrivals during a specific time interval, or by a distribution that represents the time between successive arrivals.

Many models assume that arrivals are governed by a **Poisson process**. This means that

1. Customers arrive one at a time, independently of each other and at random.
2. Past arrivals do not influence future arrivals; that is, the probability that a customer arrives at any point in time does not depend on when other customers arrived (sometimes we say that the system has *no memory*).
3. The probability of an arrival does not vary over time (the arrival rate is **stationary**).

One way to validate these assumptions is to collect empirical data about the pattern of arrivals. We can observe and record the actual times of individual arrivals in order to determine the probability distribution and check if the arrival rate is constant over time. We can also observe if customers arrive individually or in groups and whether they exhibit any special behavior, such as not entering the system if the line is perceived as too long.

If the arrival pattern is described by a Poisson process, then the Poisson probability distribution with mean arrival rate  $\lambda$  (customers per unit time) can be used to describe the probability that a particular number of customers arrives during a specified time interval. For example, an arrival rate of two customers per minute means that on the average, customers arrive every half minute, or every 30 seconds. Thus, an equivalent way of expressing arrivals is to state the **mean interarrival time** between successive customers. If  $\lambda$  is the mean arrival rate, the mean interarrival time,  $t$ , is simply  $1/\lambda$ . One useful result is that if the number of arrivals follows a Poisson process with mean  $\lambda$ , then the time between arrivals has an exponential distribution with a mean rate  $1/\lambda$ . This fact will be very useful when simulating queuing systems later in this chapter.

Sometimes, the arrival rate is not stationary (that is, customers arrive at different rates at different times). For instance, the demand for service at a quick-service restaurant is typically low in the mid-morning and mid-afternoon and is peak during the breakfast, lunch, and dinner hours. Individual customers may also arrive singly and independently (telephone calls to a mail order company) or in groups (a pallet-load of parts arriving at a machine center, or patrons at a movie theater).

The **calling population** is the set of potential customers. In many applications, the calling population is assumed to be infinite; that is, an unlimited number of possible customers can arrive to the system. This would be the case with telephone calls to a mail order company or shoppers at a supermarket. In other situations, the calling population is finite. One example would be a factory in which failed machines await repair.

Once in line, customers may not always stay in the same order as they arrived. It is common for customers to **renege**, or leave a queue, before being served if they get tired of waiting. In queuing systems with multiple queues, customers may **jockey**, or switch lines if they perceive another to be moving faster. Some customers may arrive at the system, determine that the line is too long, and decide not to join the queue. This behavior is called **balking**.

## Service Characteristics

Service occurs according to some service process. The time it takes to serve a customer may be deterministic or probabilistic. In the probabilistic case, the service time is described by some probability distribution. In many queuing models, we make the assumption that service times follow an exponential distribution with a mean service rate  $\mu$ , the average number of customers served per unit of time. Thus, the average service time is  $1/\mu$ . One reason that the exponential distribution describes many realistic service phenomena is that it has a useful property—the probability of small service times is large. For example, the probability that  $t$  exceeds the mean is only 0.368. This means that we see a large number of short service times and a few long ones. Think of your own experience in grocery stores. Most customers' service times are relatively short; however, every once in a while you see a shopper with a large number of groceries.

The exponential distribution, however, does not seem to be as common in modeling service processes as the Poisson is in modeling arrival processes. Analysts have found that many queuing systems have service time distributions that are not exponential, and may be constant, normal, or some other probability distribution.

Other service characteristics include nonstationary service times (taking orders and serving dinner might be longer than for breakfast), and service times that depend on the type of customer (patients at an emergency room). The service process may include one or several servers. The service characteristics of multiple servers may be identical or different. In some systems, certain servers may only service specific types of customers. In many systems, such as restaurants and department stores, managers vary the number of servers to adjust to busy or slack periods.

## Queue Characteristics

The order in which customers are served is defined by the **queue discipline**. The most common queue discipline is first come, first served (FCFS). In some situations, a queue may be structured as last come, first served; just think of the inbox on a clerk's desk. At an emergency room, patients are usually serviced according to a priority determined by triage, with the more critical patients served first.

## System Configuration

The customers, servers, and queues in a queuing system can be arranged in various ways. Three common queuing configurations are as follows:

1. One or more parallel servers fed by a single queue—this is the typical configuration used by many banks and airline ticket counters.
2. Several parallel servers fed by their own queues—most supermarkets and discount retailers use this type of system.
3. A combination of several queues in series—this structure is common when multiple processing operations exist, such as in manufacturing facilities.

## Performance Measures

A queuing model provides measures of system performance, which typically fall into one of these categories:

1. The quality of the service provided to the customer
2. The efficiency of the service operation and the cost of providing the service

Various numerical measures of performance can be used to evaluate the quality of the service provided to the customer. These include the following:

- Waiting time in the queue
- Time in the system (waiting time plus service time)
- Completion by a deadline

The efficiency of the service operation can be evaluated by computing measures such as the following:

- Average queue length
- Average number of customers in the system (queue plus in service)
- Throughput—the rate at which customers are served
- Server use—percentage of time servers are busy
- Percentage of customers who balk or renege

Usually, we can use these measures to compute an operating cost in order to compare alternative system configurations. The most common measures of queuing system performance, called the **operating characteristics** of the queuing system, and the symbols used to denote them are shown here:

$$L_q = \text{average number in the queue}$$

$$L = \text{average number in the system}$$

$$W_q = \text{average waiting time in the queue}$$

$$W = \text{average time in the system}$$

$$P_0 = \text{probability that the system is empty}$$

A key objective of waiting line studies is to minimize the total expected cost of the waiting line system. The total system cost consists of service costs and waiting costs. As the level of service increases (e.g., as the number of checkout counters in a grocery store increases), the cost of service increases. Simultaneously, customer waiting time will decrease and, consequently, expected waiting cost will decrease. Waiting costs are difficult to measure because they depend on customers' perceptions of waiting. It is difficult to quantify how much revenue is lost because of long lines. However, managerial judgment, tempered by experience, can provide estimates of waiting costs. If waiting occurs in a work situation, such as workers waiting in line at a copy machine, the cost of waiting should reflect the cost of productive resources lost while waiting. Although we usually cannot eliminate customer waiting completely without prohibitively high costs, we can minimize the total expected system cost by balancing service and waiting costs.

## ANALYTICAL QUEUING MODELS

Many analytical models have been developed for predicting the characteristics of waiting line systems. Analytical models of queuing behavior depend on some key assumptions about the arrival and service processes, the most common being Poisson arrivals and exponential services. In applying these models, as well as for simulation purposes,

the unit of time that we use in modeling both arrival and service processes can be arbitrary. For example, a mean arrival rate of two customers per minute is equivalent to 120 customers per hour. We must be careful, however, to express the arrival rate and service rate in the same time units.

Except for some special cases, queuing models in general are rather difficult to formulate and solve even when the distribution of arrivals and departures is known. We present analytical results for the simplest case, the single-server model.

### Single-Server Model

The most basic queuing model assumes Poisson arrivals with mean arrival rate  $\lambda$ , exponential service times with mean service rate  $\mu$ , a single server, and an FCFS queue discipline. For this model, the operating characteristics are:

$$\text{Average number in the queue} = L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} \quad (12.1)$$

$$\text{Average number in the system} = L = \frac{\lambda}{\mu - \lambda} \quad (12.2)$$

$$\text{Average waiting time in the queue} = W_q = \frac{\lambda}{\mu(\mu - \lambda)} \quad (12.3)$$

$$\text{Average time in the system} = W = \frac{1}{\mu - \lambda} \quad (12.4)$$

$$\text{Probability that the system is empty} = P_0 = 1 - \lambda/\mu \quad (12.5)$$

Note that these formulas are valid only if  $\lambda < \mu$ . If  $\lambda \geq \mu$  (that is, the rate of arrivals is at least as great as the service rate), the numerical results become nonsensical. In practice, this means that the queue will never “average out” but will grow indefinitely (we will discuss this further in the section on simulation). It should be obvious that when  $\lambda > \mu$ , the server will not be able to keep up with the demand. However, it may seem a little strange that this will occur even when  $\lambda = \mu$ . You would think that an equal arrival rate and service rate should result in a “balanced” system. This *would* be true in the deterministic case when both arrival and service rates are constant. However, when *any* variation exists in the arrival or service pattern, the queue will eventually build up indefinitely. The reason is that individual arrival times and service times vary in an unpredictable fashion even though their averages may be constant. As a result, there will be periods of time in which demand is low and the server is idle. This time is lost forever, and the server will not be able to make up for periods of heavy demand at other times. This also explains why queues form when  $\lambda < \mu$ .

To illustrate the use of this model, suppose that customers arrive at an airline ticket counter at a rate of two customers per minute and can be served at a rate of three customers per minute. Note that the average time between arrivals is 1/2 minute per customer and the average service time is 1/3 minute per customer. Using the queuing formulas, we have:

$$\text{Average number in the queue} = L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{2^2}{3(3 - 2)} = 1.33 \text{ customers}$$

$$\text{Average number in the system} = L = \frac{\lambda}{\mu - \lambda} = \frac{2}{3 - 2} = 2.00 \text{ customers}$$

$$\text{Average waiting time in the queue} = W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{2}{3(3 - 2)} = 0.67 \text{ minutes}$$

$$\text{Average time in the system} = W = \frac{1}{\mu - \lambda} = \frac{1}{3 - 2} = 1.00 \text{ minute}$$

$$\text{Probability that the system is empty} = P_0 = 1 - \lambda/\mu = 1 - \frac{2}{3} = 0.33$$

These results indicate that on the average, 1.33 customers will be waiting in the queue. In other words, if we took photographs of the waiting line at random times, we would find an average of 1.33 customers waiting. If we include any customers in service, the average number of customers in the system is 2. Each customer can expect to wait an average of 0.67 minutes in the queue and spend an average of one minute in the system. About one-third of the time, we would expect to see the system empty and the server idle.

### SKILL-BUILDER EXERCISE 12.1

Develop an Excel worksheet for computing the operating characteristics for the single-server queuing model. Apply your model to verify the calculations for the airline ticket counter example with an arrival rate of two customers per minute and a service rate of three customers per minute.

The analytical formulas provide long-term expected values for the operating characteristics; they do not describe short-term dynamic behavior of system performance. In a real waiting line system, we typically see large fluctuations around the averages, and in systems in which the system begins empty, it may take a very long time to reach these expected performance levels. Simulation provides information about the dynamic behavior of waiting lines that analytical models cannot. In addition, simulation is not constrained by the restrictive assumptions necessary to obtain a mathematical solution. Thus, simulation has some important advantages over analytical approaches.

Analytical queuing models provide steady-state values of operating characteristics. By **steady state**, we mean that the probability distribution of the operating characteristics does not vary with time. This means that no matter when we observe the system, we would expect to see the same average values of queue lengths, waiting times, and so on. However, this usually does not happen in practice, even if the average arrival rate and average service rate are constant over time. To understand this, think of an amusement park that opens at 10:00 a.m. When it opens, there are no customers in the system and hence no queues at any of the rides. As customers arrive, it will take some time for queues to build up. For the first hour or so, the lines and waiting times for popular rides grow longer and eventually level off. This is called the **transient period**. Thus, if we are interested in how long it takes to reach steady state or understand the dynamic behavior during the transient period, we must resort to other methods of analysis, such as simulation.

### Little's Law

MIT Professor John D.C. Little has made many contributions to the field of management science. He is most famous for recognizing a simple yet powerful relationship among operating characteristics in queuing systems. Little's Law, as it has become known, is very simple:

$$\text{For any steady-state queuing system, } L = \lambda W \tag{12.6}$$

This states that the average number of customers in a system is equal to the mean arrival rate times the average time in the system. An intuitive explanation of this result can be

seen in the following way. Suppose that you arrive at a queue and spend  $W$  minutes in the system (waiting plus service). During this time, more customers will arrive at a rate  $\lambda$ . Thus, when you complete service, a total of  $\lambda W$  customers will have arrived after you. This is precisely the number of customers that remain in the system when you leave, or  $L$ . Using similar arguments, we can also show that for any steady-state queuing system,  $L_q = \lambda W_q$ . This is similar to the first result and states that the average length of the queue equals the mean arrival rate times the average waiting time.

These results provide an alternative way of computing operating characteristics instead of using the formulas provided earlier. For example, if  $L$  is known, then we may compute  $W$  by  $L/\lambda$ . Also,  $W_q$  can be computed as  $L_q/\lambda$ . Two other general relationships that are useful are:

$$L = L_q + \lambda/\mu \quad (12.7)$$

and

$$W = W_q + 1/\mu \quad (12.8)$$

The first relationship states that the average number in the system is equal to the average queue length plus  $\lambda/\mu$ . This makes sense if you recall that the probability that the system is empty is  $P_0 = 1 - \lambda/\mu$ . Thus,  $\lambda/\mu$  is the probability that at least one customer is in the system. If there is at least one customer in the system, then the server must be busy. The term  $\lambda/\mu$  simply represents the expected number of customers in service.

The second relationship states that the average time in the system is equal to the average waiting time plus the average service time. This makes sense because the time spent in the system for any customer consists of the waiting time plus the time in service.

## PROCESS SIMULATION CONCEPTS

Process simulation is used routinely in business to address complex operational problems. Building a process simulation model involves first describing how the process operates, normally using some type of graphical flowchart that describes all process steps and logical decisions that route entities to different locations. Second, all key inputs such as how long it takes to perform each step of the process and resources needed to perform process tasks must be identified. Typically, the activity times in a process are uncertain and described by probability distributions. The intent is for the model to duplicate the real process so that “what-if” questions can easily be evaluated without having to make time-consuming or costly changes to the real process. Once the model is developed, the simulation process repeatedly samples from the probability distributions of the input variables to drive the flow of entities.

To understand the logic behind a process simulation model, we will use a single-server queue. Consider the sequence of activities that each customer undergoes:

1. Customer arrives.
2. Customer waits for service if the server is busy.
3. Customer receives service.
4. Customer leaves the system.

In order to compute the waiting time, we need to know the time a customer arrived and the time service began; the waiting time is the difference. Similarly, to compute the server idle time, we need to know if the arrival time of the next customer is greater than the time at which the current customer completes service. If so, the idle time is the difference. To find the number in the queue, we note that when a customer arrives, then

**TABLE 12.1** Manual Process Simulation of a Single-Server Queue

Customer	Arrival Time	Service Time	Start Time	End Time	Waiting Time	Server Idle Time
1	3.2	3.7	3.2	6.9	0.0	3.2
2	10.5	3.5	10.5	14.0	0.0	3.6
3	12.8	4.3	14.0	18.3	1.2	0.0
4	14.5	3.0	18.3	21.3	3.8	0.0
5	17.2	2.8	21.3	24.1	4.1	0.0
6	19.7	4.2	24.1	28.3	4.4	0.0
7	28.7	2.8	28.7	31.5	0.0	0.4
8	29.6	1.3	31.5	32.8	1.9	0.0
9	32.7	2.1	32.8	34.9	0.1	0.0
10	36.9	4.8	36.9	41.7	0.0	2.0

all prior customers who have not completed service by that time must still be waiting. We can make three other observations:

1. If a customer arrives and the server is idle, then service can begin immediately upon arrival.
2. If the server is busy when a customer arrives, then the customer cannot begin service until the previous customer has completed service.
3. The time that a customer completes service equals the time service begins plus the actual service time.

These observations provide all the information we need to run a small manual simulation. Table 12.1 shows such a simulation. We assume that the system opens at time 0 and that the arrival times and service times have been generated by some random mechanism and are known. We can use the logic above to complete the last four columns. For example, the first customer arrives at time 3.2 (the server is idle from time 0 until this event). Because the queue is empty, customer 1 immediately begins service and ends at time 6.9. The server is idle until the next customer arrives at time 10.5 and completes service at time 14.0. Customer 3 arrives at time 12.8. Because customer 2 is still in service, customer 3 must wait until time 14.0 to begin service, incurring a waiting time of 1.2. You should verify the calculations for the remaining customers in this simulation.

This logic is not difficult to implement in an Excel spreadsheet; see the Skill-Builder exercise that follows. However, modeling and simulation of queuing systems and other types of process simulation models is facilitated by special software applications, one of which we introduce in the next section.

### SKILL-BUILDER EXERCISE 12.2

Build an Excel model for implementing the process simulation of a single-server queue using the first three columns of data in Table 12.1 (that is, your model should calculate the start time, end time, waiting time, and server idle time).

## PROCESS SIMULATION WITH *SIMQUICK*

Many different commercial software packages are available for process simulation. Although these are very powerful, they can take considerable time to learn and master. We will use an Excel-based academic package, *SimQuick*, that is quite similar in nature to more sophisticated commercial software to learn how to perform process simulation.

## Getting Started with *SimQuick*

*SimQuick* allows you to run simple simulations in an Excel spreadsheet. To launch *SimQuick*, simply double-click the Excel file *SimQuick-v2* on the Companion Website accompanying this book. *SimQuick* contains hidden Excel macros, so please allow macros to run in Excel. You will normally receive a security warning message if macros are disabled; just click on the button *Enable Content*. Or, click *Options* from the *File* tab, click *Trust Center*, and then the *Trust Center Settings* button. Choose *Macro Settings* and change the macro security.

Figure 12.1 shows the *SimQuick* control panel. The control panel has several buttons that are used for entering information in the model. *SimQuick* uses three types of building blocks for simulation models: *objects*, *elements*, and *statistical distributions*. Objects represent the entities that move in a process (customers, parts, messages, and so on). Elements are stationary in a process and consist of five types:

1. **Entrances**—where objects enter a process
2. **Buffers**—places where objects can be stored (inventory storage, queues of people or parts, and so on)
3. **Work Stations**—places where work is performed on objects (machines, service personnel, and so on)
4. **Decision Points**—where an object goes in one of two or more directions (outcomes of processing activities, routings for further processing, and so on)
5. **Exits**—places where objects leave a process according to a specified schedule

Simulation models are created by specifying elements and their properties

Statistical distributions are limited to one of the following:

- Normal:  $\text{Nor}(\text{mean}, \text{standard deviation})$
- Exponential:  $\text{Exp}(\text{mean})$
- Uniform:  $\text{Uni}(\text{lower}, \text{upper})$
- Constant
- Discrete:  $\text{Dis}(i)$ , where  $i$  is the reference to table  $i$  of the worksheet *Discrete Distributions* (click the *Other Features* button to find this)

To save a *SimQuick* model, save the Excel worksheet under a different name.

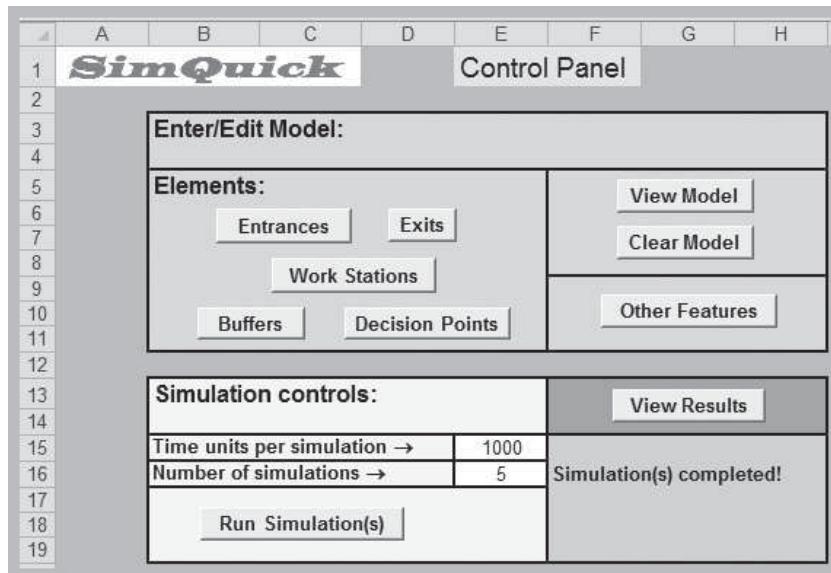


FIGURE 12.1 *SimQuick* Control Panel

## A Queuing Simulation Model

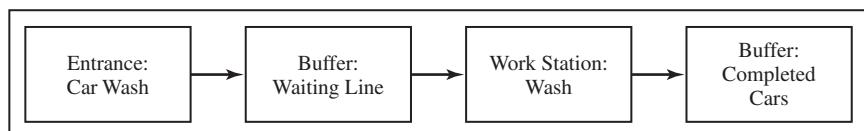
To illustrate simulation modeling and analysis for queues, we will begin with a simple single-server system and compare the results with the analytical solution we presented earlier. Dan O'Callahan operates a car wash and is in charge of finance, accounting, marketing, and analysis; his son is in charge of production. During the "lunch hour," which Dan defines as the period from 11 A.M. to 1 P.M., customers arrive randomly at an average of 15 cars per hour (or one car every four minutes). A car takes an average of three minutes to wash (or 20 cars per hour), but this fluctuates quite a bit due to variations in hand-prepping. Dan doesn't understand how a line could possibly pile up when his son can work faster than the rate at which cars arrive. Although customers complain a bit, they do not leave if they have to wait. Dan is particularly interested in understanding the waiting time, the number waiting, and how long his son is actually busy before considering improving his facility.

The first step is to draw a flowchart that represents the flow of objects through the process using the five-element structures in *SimQuick*. This is shown in Figure 12.2. Note that a buffer is used to represent cars leaving the system and will allow us to count the number of customers served. The *Exit* element is used only when objects leave the model according to a specified schedule instead of when they are ready to leave.

To build a *SimQuick* model, click on the *Entrances* button in the control panel. This brings up a new worksheet that prompts you for information. Fill out one table (working from left to right) for each entrance in the model. In the Name cell, enter the name of the Entrance block. Next, enter the time between arrivals and number of objects per arrival; this may be one of the statistical distributions, or a custom schedule, which may be accessed from the *Other Features* button. In many queuing models, we assume that the number of arrivals in a fixed time period ( $\lambda$ ) is Poisson distributed. If this is the case, then it is true that the time between arrivals ( $1/\lambda$ ) has an exponential distribution. Therefore, we specify the time between arrivals to be  $\text{Exp}(4)$ , using minutes for simulated time. Finally, specify where the objects go after entering the system; this is the next block in the process map that we labeled Waiting Line. Figure 12.3 shows the completed worksheet. Click on the *Return to Control Panel* button.

Next, click on *Buffers* to input information for the waiting line. In the table that appears, enter the name of the buffer, its capacity—that is, the maximum number of objects that can be stored in the buffer, use the word *Unlimited* to represent an infinite capacity—the initial number of objects in the buffer at the start of the simulation, and the output destinations of this process block. You have the option to move objects in different group sizes (think of a pallet of machined parts that would be transferred to a new work station as a group). Figure 12.4 shows the completed *Buffers* worksheet.

The next block in the process flow map is the work station labeled Wash. Click on the *Work Stations* button in the Control Panel. Enter the name and working time using one of the statistical distributions. In this example, we will assume that car washing time is exponentially distributed with a mean of  $1/\mu = 3$  minutes (note that  $\mu$  represents the number of completed services per unit time in the analytical solution). The



**FIGURE 12.2** Process Flow Map for Car Wash System

A	B	C	D	E	F	G	H	I
1 Entrances	Return to Control Panel		Choices for "Time bet. arr." and "Num. ob. per arr.": Nor(m,s), Exp(m), Uni(a,b), Constant, Dis(i), Cus(i)					
2								
3								
4 Examples								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								

**1**

**Name →**  
Time between arrivals → Exp(4)  
Num. objects per arrival → 1

Output destination(s) ↓

Waiting Line

**2**

**Name →**  
Time between arrivals →  
Num. objects per arrival →

Output destination(s) ↓

**3**

**Name →**  
Time between arrivals →  
Num. objects per arrival →

Output destination(s) ↓

**FIGURE 12.3 Entrances Worksheet**

A	B	C
1 Buffers	Return to Control Panel	
2		
3		
4 Examples		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		
21		
22		
23		

**1**

**Name →**  
Capacity → Unlimited  
Initial # objects → 0

Output destination(s) ↓

Waiting Line

Output group size ↓

Wash      1

**FIGURE 12.4 Buffers Worksheet**

Resource fields are not needed in this example; we will discuss them later. Figure 12.5 shows the completed *Work Stations* worksheet.

Next, click on the *Buffers* button again and enter the data for the last block in the process flow map, Completed Cars. Because this is the last block in the process flow map, there is no output destination. Return to the Control Panel and enter the time units per

A	B	C	D	E	F	G	H	I	J
1	Work Stations	Return to Control Panel			Choices for "Working time": Nor(m,s), Exp(m), Uni(a,b), Constant, Dis(l)				
2									
3	Examples								
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									
20									
21									
22									

**FIGURE 12.5** Work Stations Worksheet

simulation and number of simulations. The time units can be any time interval, even fractions, but must be used consistently in all statistical distributions. Because we used minutes in the time between arrivals and service times, we should specify the simulation duration as minutes. Because we are simulating the system from 11:00 a.m. to 1:00 p.m., enter 120 as the time units per simulation. The number of simulations is the number of replications to run the simulation (analogous to the number of trials in *Crystal Ball*). For reasonable statistical results, let us run the simulation 30 times. As with risk analysis simulations, as the number of uncertain inputs increases, the simulation should be run for a larger number of trials.

The complete model is available on the Companion Website as the Excel file *SimQuick Car Wash Simulation*. You may click on the *View Model* button for a summary of your inputs. To run the model, click on the *Run Simulation(s)* button (you may press the Esc key at any time to abort). When the simulation has completed, click on the *View Results* button. *SimQuick* creates a spreadsheet with statistical information for each model element, including overall means and detailed statistics for each run. Figure 12.6 shows a portion of the simulation results. For the Car Wash Entrance element, we have information on two things. First, we know the number of objects that entered and could not enter (this is applicable for models in which the following buffer had a finite capacity and might have blocked incoming objects). Second, we know the service level, which is defined as the percentage of objects entering the process divided by the total number that tried to enter.

For the Wash Work Station, which represents the server in the queuing system, the statistics are defined as follows:

- Final status: status of the work station when the simulation ends
- Final inventory (int. buff.), Mean inventory (int. buff.), and Mean cycle time (int. buff.): Work Stations have small internal buffers with enough room to hold one object after it has completed processing. In some models, it might not be able to pass along an object to the next buffer if it is full or another work station if it is working. In such a case, the work station is called *blocked*. These statistics provide information on the levels in this internal buffer.
- Work cycles started: the number of times the work station has started processing
- Fraction time working: use of the work station
- Fraction time blocked: fraction of time that the work station was waiting to pass an object on to the next element

A		B	C	D	E	F	G	H
1	<b>Simulation Results</b>		Return to Control Panel					
3	Element types	Element names	Statistics	Overall means	Simulation Numbers			
4					1	2	3	4
5								
6	Entrance(s)	Car Wash	Objects entering process	29.33	19	27	30	31
7			Objects unable to enter	0.00	0	0	0	0
8			Service level	1.00	1.00	1.00	1.00	1.00
9								
10	Work Station(s)	Wash	Final status	NA	Working	Working	Working	Working
11			Final inventory (int. buff.)	0.00	0	0	0	0
12			Mean inventory (int. buff.)	0.00	0.00	0.00	0.00	0.00
13			Mean cycle time (int. buff.)	0.00	0.00	0.00	0.00	0.00
14			Work cycles started	27.60	19	27	27	31
15			Fraction time working	0.65	0.37	0.45	0.54	0.85
16			Fraction time blocked	0.00	0.00	0.00	0.00	0.00
17								
18	Buffer(s)	Waiting Line	Objects leaving	27.60	19	27	27	31
19			Final inventory	1.73	0	0	3	0
20			Minimum inventory	0.00	0	0	0	0
21			Maximum inventory	4.63	3	2	4	6
22			Mean inventory	1.23	0.27	0.20	0.61	1.43
23			Mean cycle time	4.89	1.72	0.89	2.71	5.54
24								
25		Completed Cars	Objects leaving	0.00	0	0	0	0
26			Final inventory	26.73	18	26	26	30
27			Minimum inventory	0.00	0	0	0	0
28			Maximum inventory	26.73	18	26	26	30
29			Mean inventory	13.23	10.91	14.13	12.58	14.99
30			Mean cycle time	Infinite	Infinite	Infinite	Infinite	Infinite

**FIGURE 12.6** Portion of *SimQuick* Car Wash Simulation Results

We see that the mean fraction of time the car wash was busy is 0.65. However, note that the variability over different simulation runs is quite high, ranging from 0.37 to 0.85 over the first four runs.

The buffer statistics provide information about the waiting line or the buffer representing completed cars. The results summarized are as follows:

- Objects leaving: number of objects that left the buffer
- Final inventory: “Inventory” refers to the number of objects in the buffer. Final inventory is the number remaining at the end of the simulation
- Minimum inventory, maximum inventory, mean inventory: statistics on the number of objects during the simulation
- Mean cycle time: mean time that an object spends in the buffer

In this example, the waiting line itself had a mean number of cars of 1.23 over the 30 simulations. Again, this number varied considerably over the different simulation runs. The mean cycle time is the average time in the queue.

Using the results in the spreadsheet, it would be rather straightforward to perform additional statistical analyses, such as computing the minimum and maximum values of individual statistics, standard deviations, and histograms to better understand the variability in the 30 simulated runs.

How do the simulation results compare with the analytical results? Table 12.2 shows some comparisons. The simulated averages appear to be significantly different from the

**TABLE 12.2** Analytical Results versus Simulation Statistics

	Analytical Results	Simulation Means (30 Runs)
Average number in the queue	$L_q = 2.25$ customers	1.23 customers
Average waiting time in the queue	$W_q = 0.15$ hours = 9 minutes	4.89 minutes
Probability that the system is empty	$P_0 = 0.25$	$1 - 0.65 = 0.35$

analytical results. Recall that the analytical results provide steady-state averages for the behavior of the queuing system. Each of our simulated runs was for only a two-hour period, and each simulation began with an empty system. Early arrivals to the system would not expect to wait very long, but when a short-term random surge of customers arrives, the queue begins to build up. As the length of the simulation increases, the number in the queue averaged over all customers begins to level off, reaching steady state. However, during the lunch period, the car wash would probably not run long enough to reach steady state; therefore, the analytical results will never present an accurate picture of the system behavior. Thus, the differences are not surprising. However, from a practical perspective, the simulation provided information about the system behavior during this short time period that the analytical formulas could not. Dan might have made some poor decisions based on the analytical results alone.

Table 12.3 shows how these simulation statistics change as the number of time units per simulation increases. Note that as the simulation run time increases, the statistical results tend to converge toward the steady-state analytical results. Therefore, if you are interested in obtaining steady-state results from a simulation, then a long run time must be chosen.

### SKILL-BUILDER EXERCISE 12.3

Conduct a sensitivity analysis of the car wash simulation as the arrival rate and service rate vary. How sensitive are the results to these parameters?

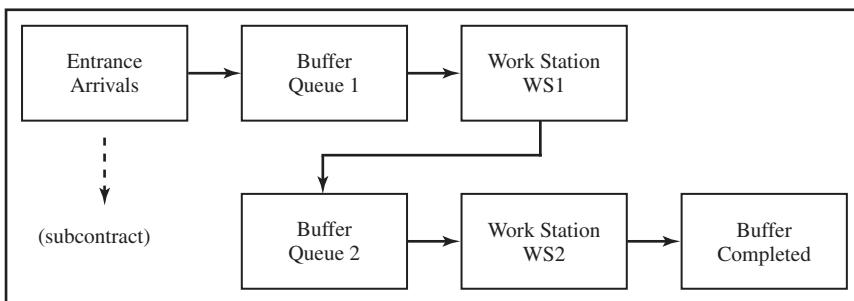
In the following sections, we present a variety of additional examples of process simulation models. These serve not only to illustrate the variety of problems to which simulation can be applied but also to demonstrate some of the additional features available in *SimQuick*.

**TABLE 12.3** Mean Queuing Statistics as a Function of Simulation Run Time

Element Type	Name	Statistics	Simulation Run Time					Analytical Result
			120	480	1,000	5,000	10,000	
Work station	Wash	Fraction time working	0.65	0.74	0.74	0.75	0.76	$1 - P_0 = 0.75$
Buffer	Waiting line	Mean inventory	1.23	1.97	2.49	2.12	2.44	$L_q = 2.25$
		Mean cycle time	4.89	7.94	0.97	8.48	9.71	$W_q = 9$

## Queues in Series with Blocking

In many manufacturing processes, the output from one work station is sent directly to the queue of another work station. Because of space limitations, these queues may have limited capacity. If they fill up, the work stations in front of them become blocked, meaning that they cannot process a new unit because they are unable to transfer the unit to the next queue. A flow process map for such a situation is shown in Figure 12.7. Assume that orders enter the system with a time between arrivals that is exponentially distributed with a mean of 0.4 minutes. The processing time at work station 1 is exponential with a mean of 0.25 minutes, and at work station 2, it is exponential with a mean of 0.5 minutes. The queue for work station 1 has a capacity of 4, while the queue for work station 2 has a capacity of 2. If an arriving order cannot enter the production process because the queue for work station 1 is full, then it is subcontracted to another manufacturer. To model these capacities in *SimQuick*, enter the capacities in the buffer tables as shown in Figure 12.8.



**FIGURE 12.7** Flow Process Map for Serial Queues and Work Stations

A	B	C	D	E	F
1	<b>Buffers</b>	<a href="#">Return to Control Panel</a>			
2					
3	<a href="#">Examples</a>				
4					
5					
6					
7	1			2	
8	<b>Name →</b>	Queue1		<b>Name →</b>	Queue2
9	Capacity →	4		Capacity →	2
10	Initial # objects →	0		Initial # objects →	0
11	Output destination(s) ↓	Output group size ↓		Output destination(s) ↓	Output group size ↓
12	WS1	1		WS2	1
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					

**FIGURE 12.8** Queue Capacities in Buffer Tables

A	B	C	D	E	F
Simulation Results			Return to Control Panel		
Element types	Element names	Statistics	Overall means	Simulation Numbers	
				1	2
6	Entrance(s)	Arrivals	Objects entering process	2696.73	2644 2670
7			Objects unable to enter	918.43	974 975
8			Service level	0.75	0.73 0.73
9					
10	Work Station(s)	WS1	Final status	NA	Working Not Working
11			Final inventory (int. buff.)	0.30	0 0
12			Mean inventory (int. buff.)	0.42	0.43 0.44
13			Mean cycle time (int. buff.)	0.23	0.23 0.24
14			Work cycles started	2694.47	2640 2670
15			Fraction time working	0.47	0.47 0.47
16			Fraction time blocked	0.42	0.43 0.44
17					
18		WS2	Final status	NA	Not Working Working
19			Final inventory (int. buff.)	0.00	0 0
20			Mean inventory (int. buff.)	0.00	0.00 0.00
21			Mean cycle time (int. buff.)	0.00	0.00 0.00
22			Work cycles started	2692.23	2639 2668
23			Fraction time working	0.93	0.93 0.94
24			Fraction time blocked	0.00	0.00 0.00
25					
26	Buffer(s)	Queue1	Objects leaving	2694.47	2640 2670
27			Final inventory	2.27	4 0
28			Minimum inventory	0.00	0 0
29			Maximum inventory	4.00	4 4
30			Mean inventory	2.15	2.20 2.21
31			Mean cycle time	1.15	1.20 1.19
32					
33		Queue2	Objects leaving	2692.23	2639 2668
34			Final inventory	1.30	0 2
35			Minimum inventory	0.00	0 0
36			Maximum inventory	2.00	2 2
37			Mean inventory	1.49	1.51 1.54
38			Mean cycle time	0.80	0.83 0.83
39					
40		Completed	Objects leaving	0.00	0 0
41			Final inventory	2691.30	2639 2667
42			Minimum inventory	0.00	0 0
43			Maximum inventory	2691.30	2639 2667
44			Mean inventory	1340.78	1317.63 1337.81
45			Mean cycle time	Infinite	Infinite Infinite

**FIGURE 12.9** Portion of Simulation Results for Serial Queue Model

Figure 12.9 shows the *SimQuick* results for 30 replications of a 24-hour (1,440 minutes) simulation in the Excel model file *SimQuick Serial Queues*. In the Entrance statistics, we see that the service level is 75%, or equivalently, 25% of the arriving objects are unable to enter the system and hence are subcontracted. In the Work Station statistics, we see that work station 1 is blocked 42% of the time because the following queue is full.

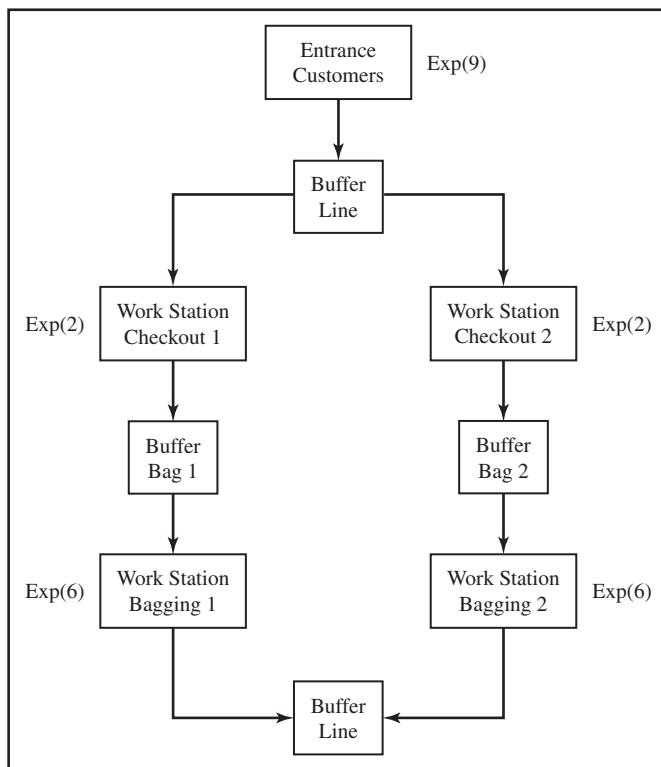
### Grocery Store Checkout Model with Resources

A small grocery store has two checkout lines and a floating bagger, who bags groceries at one of the two lines as needed. Customers enter the queue for the checkout lines, arriving on average every nine minutes (exponentially distributed). Checkout time is exponential with a mean of two minutes. If line 1 is open, a customer will go there first. If line 1 is occupied and line 2 is free, the customer will go to line 2 instead. If both lines are occupied, the customer will wait in the queue until one of them finishes service. The bagger can only work one line at a time and takes an average of six minutes to bag, again exponentially distributed. If a checker finishes scanning all of a customer's

items and the bagger is working at the other line, we will assume that the customer and checker will wait for the bagger. In such cases, the checker is blocked from processing another customer who might be waiting in line.

The bagger is an example of a **resource** in *SimQuick*. Resources are defined and assigned to work stations. A work station cannot start processing unless all resources that are assigned to it are available. If more than one work station competes for a limited resource, the work station with the higher priority gets it. The priority is determined by the number of the table in the *SimQuick* model (the lower the number, the higher the priority).

The process flow map of the grocery store operation and the distribution of processing times are shown in Figure 12.10. In the *SimQuick* Excel model *SimQuick Grocery Base Case*, the bagger is defined as a resource by clicking the button *Other Features* in the control panel and then choosing *Resources*. In the Resources worksheet, enter the name of the resource (Bagger) and the number of resources available (1) as shown in Figure 12.11. For



**FIGURE 12.10** Process Flow Map of Grocery Store Operation

A	B	C	D	E
<b>Resources</b>	<a href="#">Return to Control Panel</a>		<a href="#">Return to Other Features</a>	
1				
2				
3				
4				
5				
6				
7				
8				
9				
	Name ↓	Number available ↓		
1	Bagger	1		
2				

**FIGURE 12.11**  
*SimQuick Resources Worksheet*

the work stations associated with the bagging processes, the resource Bagger is assigned to each of them (see Figure 12.12). Because only one bagger is available, the resource can be used by only one of the bagging processes at any one time.

Figure 12.13 shows the *SimQuick* results for 30 runs of 600 time units each. The key statistic for the work stations is the fraction time blocked, which represents the time the

	L	M	N	O	P	Q	R	S	I
5									
6									
7		3							
8			Name → Working time	Bagging 1 Exp(6)					
9	Output destination(s) ↓	# of output objects ↓	Resource name(s) ↓	Resource # units needed ↓					
10	Finish	1	Bagger	1					
11									
12									
13									
14									
15									
16									
17									
18									
19									
20									
21									
22									

	L	M	N	O	P	Q	R	S	I
5									
6									
7		4							
8			Name → Working time	Bagging 2 Exp(6)					
9	Output destination(s) ↓	# of output objects ↓	Resource name(s) ↓	Resource # units needed ↓					
10	Finish	1	Bagger	1					
11									
12									
13									
14									
15									
16									
17									
18									
19									
20									
21									
22									

**FIGURE 12.12**

Assigning Resources  
to Work Stations

	A	B	C	D	E	F	
1	<b>Simulation Results</b>			Return to Control Panel			
2							
3	<b>Element types</b>	<b>Element names</b>	<b>Statistics</b>	<b>Overall means</b>	<b>Simulation Numbers</b>		
4					1	2	
5							
6	Entrance(s)	Customers	Objects entering process	69.00	66	76	
7			Objects unable to enter	0.00	0	0	
8			Service level	1.00	1.00	1.00	
9							
10	Work Station(s)	Checkout 1	Final status	NA	Not Working	Not Working	
11			Final inventory (int. buff.)	0.17	0	0	
12			Mean inventory (int. buff.)	0.19	0.12	0.46	
13			Mean cycle time (int. buff.)	2.46	1.28	5.64	
14			Work cycles started	48.57	55	49	
15			Fraction time working	0.17	0.17	0.13	
16			Fraction time blocked	0.19	0.12	0.46	
17							
18		Checkout 2	Final status	NA	Not Working	Not Working	
19			Final inventory (int. buff.)	0.13	0	0	
20			Mean inventory (int. buff.)	0.11	0.04	0.36	
21			Mean cycle time (int. buff.)	3.03	2.14	8.02	
22			Work cycles started	20.10	11	27	
23			Fraction time working	0.07	0.04	0.10	
24			Fraction time blocked	0.11	0.04	0.36	
25							
26		Bagging 1	Final status	NA	Working	Not Working	
27			Final inventory (int. buff.)	0.00	0	0	
28			Mean inventory (int. buff.)	0.00	0.00	0.00	
29			Mean cycle time (int. buff.)	0.00	0.00	0.00	
30			Work cycles started	47.93	55	48	
31			Fraction time working	0.47	0.64	0.53	
32			Fraction time blocked	0.00	0.00	0.00	
33							
34		Bagging 2	Final status	NA	Not Working	Working	
35			Final inventory (int. buff.)	0.00	0	0	
36			Mean inventory (int. buff.)	0.00	0.00	0.00	
37			Mean cycle time (int. buff.)	0.00	0.00	0.00	
38			Work cycles started	19.57	11	27	
39			Fraction time working	0.19	0.07	0.29	
40			Fraction time blocked	0.00	0.00	0.00	

**FIGURE 12.13** Portion of *SimQuick* Results for Grocery Store Simulation Model

A	B	C	D	E	F
42	Buffer(s)	Line	Objects leaving	68.67	66
43			Final inventory	0.33	0
44			Minimum inventory	0.00	0
45			Maximum inventory	2.60	2
46			Mean inventory	0.16	0.03
47			Mean cycle time	1.39	0.29
48					
49		Bag 1	Objects leaving	47.93	55
50			Final inventory	0.37	0
51			Minimum inventory	0.00	0
52			Maximum inventory	1.00	1
53			Mean inventory	0.41	0.37
54			Mean cycle time	5.13	4.06
55					
56		Bag 2	Objects leaving	19.57	11
57			Final inventory	0.30	0
58			Minimum inventory	0.00	0
59			Maximum inventory	1.00	1
60			Mean inventory	0.26	0.07
61			Mean cycle time	7.60	3.86
62					
63		Finish	Objects leaving	0.00	0
64			Final inventory	66.73	65
65			Minimum inventory	0.00	0
66			Maximum inventory	66.73	65
67			Mean inventory	32.02	31.30
68			Mean cycle time	Infinite	Infinite
69					
70					
71					
72	<b>Resource(s)</b>				
73	Bagger	Mean number in use	0.66	0.71	0.81

**FIGURE 12.13 (Continued)**

checker must wait for a bagger after servicing a customer and before he or she can service another waiting customer. For Checkout lines 1 and 2, these values are, respectively, 0.19 and 0.11. Note that the fraction of time working is higher for Checkout 1 because that line has a higher priority for entering customers. In the buffer statistics, there are an average of 0.16 customers waiting (mean inventory) for checkout, with an average waiting time of 1.39 (mean cycle time). However, for the bag buffers, the mean waiting times are 5.13 and 7.60, respectively, which is probably an unacceptable service level for a grocery store.

One way of possibly improving the process is to assign the bagger to line 1 exclusively and have the checker in line 2 also bag groceries. To model this (see Excel file *SimQuick Grocery Enhanced*), define the checkout clerk in the second line as a resource, Checker2, in the Resources worksheet. The bagger need not be defined as a resource in this model because the two work stations are no longer competing for it. Then, for the work stations Checkout 2 and Bagging 2, assign Checker 2 as a required resource. In this fashion, Checker 2 may work on only one process at a time. If you examine the results, you will find that the fraction of time blocked has decreased substantially to 0.07 for line 1 and 0 for line 2 and that the mean waiting times for both the entrance queues and bagging queues have also decreased. The mean waiting time for entering customers is now only 0.27, and for the two bagging operations, it has decreased to 2.48 and 1.12, respectively.

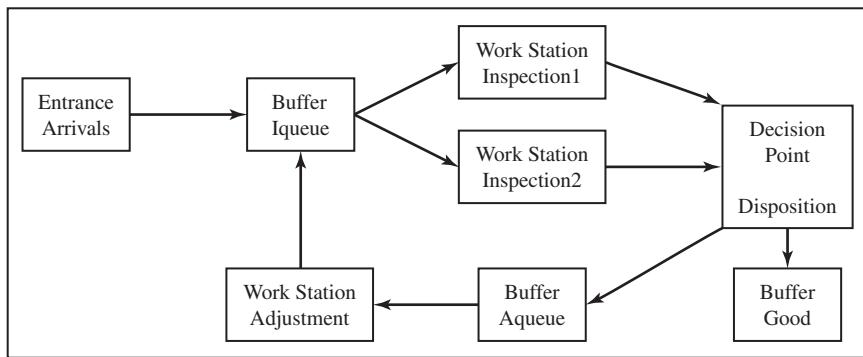
## Manufacturing Inspection Model with Decision Points

A Decision Point in a *SimQuick* model is a point in a process flow map where the routes for some objects are chosen randomly. Suppose that the output from an electronics

manufacturing operation is inspected and that a percentage of the items require adjustment and are reinspected. Assume that units arrive to the inspection process queue (Iqueue) at a constant rate of one every five minutes. The inspection process is normal with a mean of nine minutes and a standard deviation of two minutes, and two inspectors are working. Fifteen percent of the units require adjustment and are sent to the queue (Aqueue) of an adjustment process, which takes an average of 30 minutes, with a standard deviation of five minutes, normally distributed. The remaining 85% of the units are classified as good and sent on to packaging and distribution.

Figure 12.14 shows the process flow map for this situation. To define a decision point, click on the *Decision Points* button in the control panel. In the *SimQuick* table, enter the name of the output destinations and the percentages of objects that are routed to these destinations. This is shown in Figure 12.15.

Figure 12.16 shows the results of 30 simulations for 400 time units (see Excel file *SimQuick Inspection Adjustment Model*). We see that the inspections have high utilizations,



**FIGURE 12.14** Process Flow Map for Inspection and Adjustment Process

A	B	C	D	E	F			
1	<b>Decision Points</b>		Return to Control					
2	(Note: Percents columns must sum to 100)							
3	<a href="#">Examples</a>							
4								
5								
6								
7	1		2					
8	<b>Name → Disposition</b>		<b>Name →</b>					
9	Output destinations ↓		Output destinations ↓		Percents ↓			
10	Aqueue 15							
11	Good 85							
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								

**FIGURE 12.15** SimQuick Decision Points Table

A	B	C	D	E	F	G	
Simulation Results			Return to Control Panel				
Element types	Element names	Statistics	Overall means	Simulation Numbers			
				1	2	3	
6	Entrance(s)	Arrivals	Objects entering process	80.00	80	80	80
7			Objects unable to enter	0.00	0	0	0
8			Service level	1.00	1.00	1.00	1.00
10	Work Station(s)	Inspection1	Final status	NA	Working	Working	Working
11			Final inventory (int. buff.)	0.00	0	0	0
12			Mean inventory (int. buff.)	0.00	0.00	0.00	0.00
13			Mean cycle time (int. buff.)	0.00	0.00	0.00	0.00
14			Work cycles started	43.83	45	44	43
15			Fraction time working	0.98	0.97	0.98	0.98
16			Fraction time blocked	0.00	0.00	0.00	0.00
18		Inspection2	Final status	NA	Working	Working	Working
19			Final inventory (int. buff.)	0.00	0	0	0
20			Mean inventory (int. buff.)	0.00	0.00	0.00	0.00
21			Mean cycle time (int. buff.)	0.00	0.00	0.00	0.00
22			Work cycles started	43.60	44	46	44
23			Fraction time working	0.97	0.96	0.97	0.98
24			Fraction time blocked	0.00	0.00	0.00	0.00
26		Adjustment	Final status	NA	Working	Working	Working
27			Final inventory (int. buff.)	0.00	0	0	0
28			Mean inventory (int. buff.)	0.00	0.00	0.00	0.00
29			Mean cycle time (int. buff.)	0.00	0.00	0.00	0.00
30			Work cycles started	11.57	11	13	13
31			Fraction time working	0.83	0.73	0.91	0.93
32			Fraction time blocked	0.00	0.00	0.00	0.00

A	B	C	D	E	F	G	
34	Buffer(s)	Iqueue	Objects leaving	87.43	89	90	87
35			Final inventory	3.23	1	2	5
36			Minimum inventory	0.00	0	0	0
37			Maximum inventory	5.10	3	5	7
38			Mean inventory	1.80	0.64	1.68	2.52
39			Mean cycle time	8.25	2.88	7.45	11.61
41		Aqueue	Objects leaving	11.57	11	13	13
42			Final inventory	2.60	1	1	5
43			Minimum inventory	0.00	0	0	0
44			Maximum inventory	3.97	3	3	5
45			Mean inventory	1.40	0.50	1.22	1.58
46			Mean cycle time	46.84	18.34	37.52	48.56
48		Good	Objects leaving	0.00	0	0	0
49			Final inventory	71.27	75	74	67
50			Minimum inventory	0.00	0	0	0
51			Maximum inventory	71.27	75	74	67
52			Mean inventory	34.53	36.69	34.74	33.50
53			Mean cycle time	Infinite	Infinite	Infinite	Infinite
55	Decision Point(s)	Disposition	Objects leaving	85.43	87	88	85
56			Final inventory (int. buff.)	0.00	0	0	0
57			Mean inventory (int. buff.)	0.00	0.00	0.00	0.00
58			Mean cycle time (int. buff.)	0.00	0.00	0.00	0.00

FIGURE 12.16 Portion of SimQuick Results for Inspection Adjustment Model

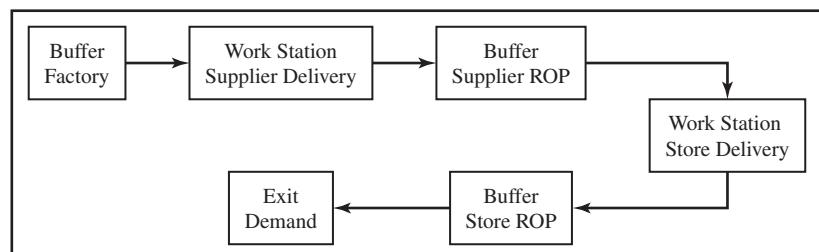
98% and 97%. The adjustor is working 83% of the time. The mean waiting time in the inspection queue is 8.25 and an average of 1.8 units wait. In the adjustment queue, the mean waiting time is 46.84, and on average, 1.4 units are waiting for adjustment. This process appears to be running adequately. This model might be used to investigate the effect of quality improvements in the manufacturing process or the impact of cross-training inspectors to perform adjustments.

### Pull System Supply Chain with Exit Schedules

A supply chain for a particular model of high-definition TV consists of a manufacturer, wholesale supplier, and retail store. Customer demand pulls inventory from the system at a rate that is exponentially distributed with a mean of one unit per day. The retail store uses a reorder point (ROP) inventory system in which six units are ordered whenever the inventory drops to three units. The number of units ordered is called the **lot size** or **order quantity**, and the inventory level that triggers an order is called the **reorder point**. Delivery time from the supplier to the store is normal with a mean of three days and standard deviation of one day. At the supplier's warehouse, inventory is ordered from the manufacturer in lot sizes of 10 with a ROP of 15. Lead time for delivery from the manufacturer is normal with a mean of two days and a standard deviation of 0.5 days.

A *SimQuick* model of this system (Excel file *SimQuick Pull Supply Chain*) is shown in Figure 12.17. This model introduces a new element, *Exit*. An *Exit* pulls objects according to a specified “time between departures.” Exits are modeled by clicking on the appropriate button in the control panel. You must also specify the number of objects per departure. In this model, the *Exit* element, Demand, has a time between departures that is  $\text{Exp}(1)$ , and one object per departure. The buffers model the ROP processes. The Store ROP buffer has a capacity and initial number of objects equal to 3, the ROP at the retail store, and output group size of 1, the demand for each customer. Similarly, the Supplier ROP buffer has a capacity and initial number of objects equal to its ROP, 15, and an output group size of 6 corresponding to the order quantity it sends to the retail store. The Factory buffer is given an arbitrarily large capacity of 5,000, which also equals the initial number of objects. The output group size from the factory is the order quantity of the supplier, 10. Each time Supplier Delivery obtains one object from Factory, the number of objects in the Factory drops by 10. The Store Delivery work station sends six objects at a time (the order quantity) to the Store ROP buffer with a working time defined as  $\text{Nor}(3, 1)$ , and the Supplier Delivery work station sends 10 objects at a time to the Supplier ROP buffer with a working time that is  $\text{Nor}(2, 0.5)$ .

Store inventory is modeled by the Store ROP buffer and the internal buffer of the Store Delivery work station. The internal buffer of a work station holds objects completed after one working cycle. Work Stations are blocked from working on a new object as long as it has objects in its internal buffer. The *Exit* element pulls objects from Store ROP, which in turn, pulls objects from the internal buffer of Store Delivery. When the number



**FIGURE 12.17** Process Flow Map of Pull Supply Chain

of objects in the internal buffer falls to zero, the inventory is entirely contained in the Store ROP buffer, and the amount of inventory is less than or equal to the ROP, 3, which is the capacity of the Store ROP buffer. At this time, Store Delivery becomes unblocked and pulls an object from Supplier ROP, which corresponds to placing an order. When Store Delivery is finished, it deposits the order size of 6 units into its internal buffer. Store Delivery is then blocked until it can pass all inventory to the following buffer.

Figure 12.18 shows the mean results for 30 runs of a 365-day simulation (Excel file *SimQuick Pull Supply Chain*). Of particular interest is the service level for the *Exit*

A	B	C	D	E	F	G	
1	<b>Simulation Results</b>		Return to Control Panel				
3	Element types	Element names	Statistics	Overall means	Simulation Numbers		
4					1	2	3
6	Work Station(s)	Store Delivery	Final status	NA	Not Working	Not Working	Working
7			Final inventory (int. buff.)	1.63	3	2	0
8			Mean inventory (int. buff.)	1.41	1.41	1.38	1.20
9			Mean cycle time (int. buff.)	1.60	1.63	1.54	1.38
10			Work cycles started	54.23	53	55	54
11			Fraction time working	0.45	0.44	0.46	0.49
12			Fraction time blocked	0.55	0.56	0.54	0.51
14		Supplier Delivery	Final status	NA	Not Working	Not Working	Working
15			Final inventory (int. buff.)	4.67	2	10	0
16			Mean inventory (int. buff.)	4.57	4.53	4.57	4.60
17			Mean cycle time (int. buff.)	5.15	5.20	5.05	5.24
18			Work cycles started	33.10	32	34	33
19			Fraction time working	0.18	0.17	0.18	0.17
20			Fraction time blocked	0.82	0.83	0.82	0.83
22	Buffer(s)	Factory	Objects leaving	331.00	320	340	330
23			Final inventory	4669.00	4680	4660	4670
24			Minimum inventory	4669.00	4680	4660	4670
25			Maximum inventory	5000.00	5000	5000	5000
26			Mean inventory	4829.74	4832.54	4823.50	4828.73
27			Mean cycle time	5335.02	5512.12	5178.17	5340.86
29		Supplier ROP	Objects leaving	325.40	318	330	324
30			Final inventory	14.60	15	15	11
31			Minimum inventory	8.93	9	9	9
32			Maximum inventory	15.00	15	15	15
33			Mean inventory	14.62	14.63	14.61	14.62
34			Mean cycle time	16.43	16.79	16.16	16.47
36		Store ROP	Objects leaving	322.03	315	328	321
37			Final inventory	2.33	3	3	0
38			Minimum inventory	0.00	0	0	0
39			Maximum inventory	3.00	3	3	3
40			Mean inventory	2.38	2.43	2.35	2.22
41			Mean cycle time	2.71	2.81	2.62	2.52
43	Exit(s)	Demand	Objects leaving process	322.03	315	328	321
44			Object departures missed	44.60	46	44	69
45			Service level	0.88	0.87	0.88	0.82

**FIGURE 12.18** Portion of Simulation Results for Pull Supply Chain

element, which is 0.88. This means that only 88% of customer demand can be satisfied (the store is out of stock 12% of the time). Service level is primarily influenced by the ROP, suggesting that a higher ROP should be used at the retail store. Other information that we can obtain from the simulation results include the following:

- **Mean number of orders placed.** This is found from the number of work cycles started at the work stations. For example, the supplier placed an average of 33 orders, while the store placed about 54 orders.
- **Mean inventory level.** This is found by adding the mean inventory in the internal buffer plus the mean inventory in the corresponding buffer. Thus, for the supplier, the mean amount of inventory held is  $4.57 + 14.62 = 20.19$  units. At the store, we have a mean inventory level of  $1.41 + 2.38 = 3.79$  units.

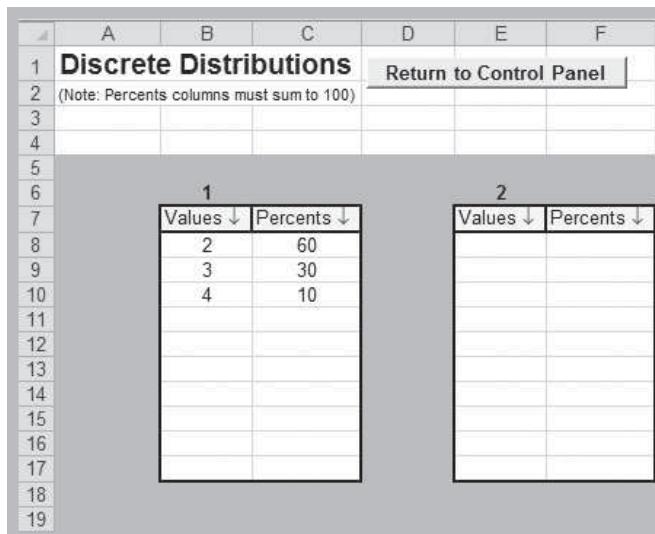
### Other **SimQuick** Features and Commercial Simulation Software

Two other features of *SimQuick* that we have not illustrated in the examples are discrete distributions and custom schedules. Discrete distributions can be defined by clicking on the *Other Features* button and selecting *Discrete Distributions*. For example, suppose that a more realistic assumption for the working time for the car wash in our first example is:

Time	Probability
2	.6
3	.3
4	.1

Car wash times are usually constant but depend on the features that customers order (for instance, underspray, wax, and so on), which would change the amount of time it takes to service the car. In the Discrete Distribution table, enter the values for time and the probabilities expressed as percentages as shown in Figure 12.19. Then in the Work Station table, specify *Dis(1)* as the working time. In general, *Dis(*i*)* refers to the distribution you specify in table *i* in the Discrete Distributions worksheet.

A Custom Schedule allows a modeler to specify the exact times of arrivals at Entrances and departures at Exits rather than using some probability distribution that



**FIGURE 12.19** *Discrete Distributions Worksheet in SimQuick*

	A	B	C	D	E	F
1	<b>Custom Schedules</b>		Return to Control Panel		Return to Other Features	
2						
3						
4	A Custom Schedule can have at most 1,000 arrivals/departures					
5						
6						
7	1	Times ↓	Quantity arriving/departing ↓	2	Times ↓	Quantity arriving/departing ↓
8						
9						

**FIGURE 12.20** Custom Schedules Input Table

generates random values. From the *Other Features* menu, select *Custom Schedules*. In the table, enter the times and the quantity that arrive or depart in the appropriate columns as shown in Figure 12.20. Then use  $\text{Cus}(i)$ , where  $i$  represents the table number of the custom schedule, in the fields for *Time between arrivals* and *Num. objects per arrival* in an entrance element, or in the fields *Time between departures* and *Num. Objects per departure* in an exit element.

*SimQuick* can model a variety of additional applications, such as manufacturing cells, assembly/disassembly processes, job shops, quality control policies, project management, and more complex inventory situations. For further information about *SimQuick* and additional examples, we recommend that you consult the book by David Hartvigsen, *SimQuick: Process Simulation with Excel*, 2nd edition, Prentice-Hall, 2004.

Although *SimQuick* is a convenient software application for learning the basic concepts of systems simulation, it is limited in its modeling capabilities. More powerful commercial software packages are available that can model and simulate virtually any situation. These include GPSS, ProcessModel, Arena, Extend, and many others. Information about these packages can be found by searching online.

## CONTINUOUS SIMULATION MODELING

Many models contain variables that change continuously over time. One example would be a model of an oil refinery. The amount of oil moving between various stages of production is clearly a continuous variable. In other models, changes in variables occur gradually (though discretely) over an extended time period; however, for all intents and purposes, they may be treated as continuous. An example would be the amount of inventory at a warehouse in a production-distribution system over several years. As customer demand is fulfilled, inventory is depleted, leading to factory orders to replenish the stock. As orders are received from suppliers, the inventory increases. Over time, particularly if orders are relatively small and frequent, as we see in just-in-time environments, the inventory level can be represented by a smooth, continuous function.

Continuous variables are often called *state variables*. A continuous simulation model defines equations for relationships among state variables so that the dynamic behavior of the system over time can be studied. To simulate continuous systems, we will decompose time into small increments. The defining equations are used to determine how the state variables change during an increment of time. A specific type of continuous simulation is called **system dynamics**, which dates back to the early 1960s when it was created by Jay Forrester of MIT. System dynamics focuses on the structure and behavior of systems that are composed of interactions among variables and feedback loops. A system dynamics model usually takes the form of an influence diagram that shows the relationships and interactions among a set of variables.

To gain an understanding of system dynamics and how continuous simulation models work, let us develop a model for the cost of medical care and implement it using

general Excel features. Doctors and hospitals charge more for services, citing the rising cost of research, equipment, and insurance rates. Insurance companies cite rising court awards in malpractice suits as the basis for increasing their rates. Lawyers stress the need to force professionals to provide their patients with the best care possible and use the courts as a means to enforce patient rights. The medical cost system has received focused attention from those paying for medical care and from government officials.

Let us suppose that we are interested in how medical rates (MEDRATE) are influenced by other factors, specifically:

1. The demand for medical service (DEMAND)
2. Insurance rates (INSRATE)
3. Population levels (POPLVL)
4. Medical-related lawsuits (MEDSUIT)
5. Avoidance of risk by doctors (RISK)

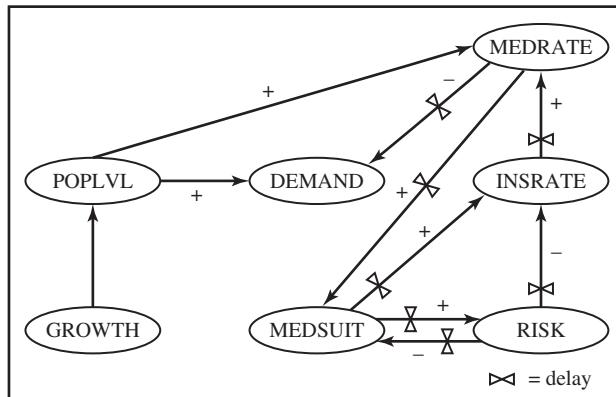
Figure 12.21 shows an influence diagram of how these factors might relate to one another. For example, rates rise as the demand for medical service increases and as insurance rates rise. The demand is influenced by the population level and its growth rate. Also, increasing rates have a negative influence on demand, meaning that as rates rise, the demand will decrease. Insurance rates increase as medical lawsuits increase and drop as doctors avoid taking risks. At the same time, lawsuits increase as medical rates increase but also decline with risk avoidance. Some of these influences do not occur immediately, as noted by the “delay” factors in the figure. It might take about one year before some variables actually influence others.

We may express these relationships quantitatively through a set of equations that describes how each variable changes from one year to the next (that is, year  $t - 1$  to year  $t$ ). At time  $t = 0$ , we index all variables to 1.0. We will assume that the population level grows each year by a value,  $GROWTH(t)$ , that is normally distributed with a mean of 0.05 and a standard deviation of 0.03. This is expressed by the equation:

$$POPLVL(t) = POPLVL(t - 1) + GROWTH(t)$$

The demand for medical services increases with the population and decreases with the rate of increase in the cost of medical service, lagged by one year. Thus, demand is computed by the formula:

$$DEMAND(t) = POPLVL(t) - [MEDRATE(t - 1) - MEDRATE(t - 2)]$$



**FIGURE 12.21** Influence Diagram for the Cost of Medical Services

The cost of medical services increases with the change in population level and a portion (80%) of the increase in insurance rates, lagged by one year:

$$\begin{aligned} \text{MEDRATE}(t) &= \text{MEDRATE}(t - 1) + \text{POPLVL}(t) - \text{POPLVL}(t - 1) \\ &\quad + .8 \times [\text{INSRATE}(t - 1) - \text{INSRATE}(t - 2)] \end{aligned}$$

Insurance rates increase by a fraction (10%) of the previous year's level of lawsuits and decrease with any increases in doctors' adoption of safer practices to avoid risk:

$$\begin{aligned} \text{INSRATE}(t) &= \text{INSRATE}(t - 1) + .10 \times \text{MEDSUIT}(t - 1) \\ &\quad - [\text{RISK}(t - 1) - \text{RISK}(t - 2)] \end{aligned}$$

Increase in lawsuits is proportional to the increased costs of medical service and inversely proportional to risk avoidance, both lagged by one year:

$$\text{MEDSUIT}(t) = \text{MEDSUIT}(t - 1) + [\text{MEDRATE}(t - 1) - 1]/\text{RISK}(t - 1)$$

Finally, the avoidance of risk increases as a proportion (10%) of the increase in the level of lawsuits, based on the previous year:

$$\text{RISK}(t) = \text{RISK}(t - 1) + .10 \times [\text{MEDSUIT}(t - 1) - 1]$$

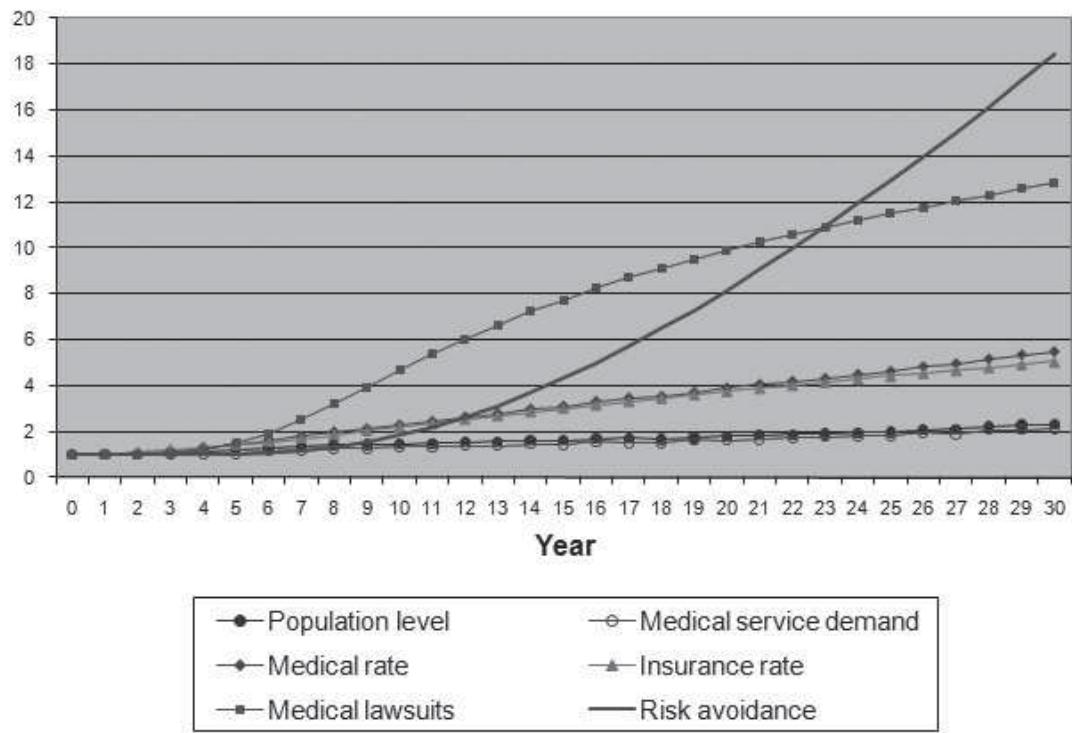
Figure 12.22 shows a portion of (first 10 time periods) spreadsheet model for simulating this system (Excel file *Continuous Simulation Model*). Population growth for each year is modeled using the *Crystal Ball* function CB.Normal(0.05,0.03). (Make sure that *Crystal Ball* is loaded to have access to this function in Excel.) The remainder of the simulation model is deterministic because none of the other variables are assumed to be uncertain. Figure 12.23 shows a graph of each of the variables over the 30-year period of the simulation. Based on our assumptions, the population has increased by almost 350%. However, the demand for medical services has not quite reached that level, damped by a fivefold increase in the cost of medical services. Insurance rates have increased five times, and lawsuits have increased 13 times (a compounded rate of 9% per year), while risk avoidance practices have increased an average of more than 10% per year.

System dynamics has been applied to the analysis of material and information flows in logistics systems, sales and marketing problems, social organizations, ecology, and many other fields. System dynamics was quite popular among researchers and practitioners until the early 1970s. The concept was brought back to the attention of

	A	B	C	D	E	F	G	H
1	Time	Population growth	Population level	Med. Service demand	Medical rate	Insurance rate	Medical lawsuits	Risk avoidance
2	period							
3	0			1	1	1	1	1
4	1	0.005	1.005	1.005	1.005	1	1	1
5	2	0.014	1.019	1.014	1.019	1.1	1.005	1
6	3	0.051	1.071	1.057	1.151	1.201	1.025	1.001
7	4	0.081	1.151	1.020	1.312	1.302	1.176	1.003
8	5	0.063	1.215	1.054	1.457	1.418	1.486	1.021
9	6	0.079	1.294	1.149	1.628	1.549	1.934	1.069
10	7	0.053	1.347	1.175	1.785	1.693	2.521	1.163
11	8	0.076	1.423	1.265	1.977	1.852	3.197	1.315
12	9	0.011	1.434	1.242	2.116	2.020	3.940	1.534
13	10	0.030	1.464	1.326	2.280	2.194	4.667	1.828

**FIGURE 12.22** Portion of Spreadsheet for Continuous Simulation of Medical Rates

## Medical Rate Simulation Results



**FIGURE 12.23** Dynamic Behavior of Variables in Medical Rate Simulation

business in the 1990s by Peter Senge through his book *The Fifth Discipline* (New York: Currency, 1994), which explores the role and importance of systems thinking in modern organizations and has formed the basis for many contemporary concepts in supply chain management.

### Basic Concepts Review Questions

1. What is a queuing system?
2. What is an arrival process? Describe the most common arrival process used in practice.
3. Describe the standard assumptions for the serving characteristics in a queuing system.
4. What is meant by operating characteristics? What do they measure?
5. Explain the following terms:
  - a. Calling population
  - b. Reneging
  - c. Jockeying
  - d. Balking
  - e. Queue discipline
6. What is the purpose of process simulation? Name the tool which may be used to perform process simulation in Excel.
7. List and define the major types of performance measures used in queuing analysis.
8. Why do queues build up when the mean arrival rate equals the mean service rate?
9. Explain the difference between steady-state and transient behavior in queuing systems.
10. What is Little's Law, and why is it important?
11. Explain the key features of SimQuick and how they are used in simulation models.
12. How does continuous simulation modeling differ from process simulation? Provide some practical examples different from the text where continuous simulation models might be used.

## Problems and Applications

1. A grocery store checkout counter has a service rate of 45 per hour, exponentially distributed, with Poisson arrivals at the rate of 30 per hour. Find the operating characteristics of this system.
2. Suppose students arrive during the office hours of a college professor according to a Poisson process with mean arrival rate  $\lambda$ . The professor, who deals with his students one by one, has an exponentially distributed service time of ten per hour. Analyze how the average waiting time is expected to change as the arrival rate  $\lambda$  varies from five to eight students per hour.
3. A hair cutting saloon has a single barber who can serve a customer in an average of 20 minutes. Customers arrive in the saloon at an average rate of two per hour for getting a haircut this service.
  - a. How long will a customer expect to wait before being served?
- b. On the average how many customers will be waiting?
- c. What is the probability that the barber will be idle?
4. A single teller serves the customers in a bank during the afternoon hours of 2 to 3 pm. The teller can serve a customer in an average of four minutes. Suppose that the arrival rates of customers are 10 customers per hour.
  - a. Find the operating characteristics of the single teller queuing system.
  - b. If the arrival rate of the customers goes up in the future, how high should the average arrival rate be before a second teller is added?
5. A fast food store has a single server. Customers arrive at the store at the rate of six per hour according to a Poisson process. Serving times average four minutes according to an exponential distribution. What are the operating characteristics for this system?
6. Complete the following table for a queuing simulation.

Customer	Arrival Time	Service Time	Start Time	End Time	Waiting Time	Server Idle Time
1	1.1	0.6				
2	2.2	1.1				
3	2.9	3.3				
4	3.7	1.0				
5	4.9	0.8				
6	6.6	4.2				
7	9.4	0.8				
8	10.6	1.3				
9	11.1	0.1				
10	13.3	3.8				

7. People arrive at a busy airport news kiosk at a rate that is uniformly distributed between five and ten seconds. Most people buy only one paper, which takes an average of six seconds, but may be as little as three or as much as nine seconds, and 20% buy two papers, which takes between six and twelve seconds to buy. Develop a *SimQuick* model of this process and interpret the results.
8. People arrive at a self-service cafeteria on average every 15 seconds, exponentially distributed. There are two counters, entrées and drinks. Thirty percent of customers want entrées only, 60% go first to entrées then to drinks, and the remaining get only a drink. Service time at each counter is exponential with a mean of 15 seconds. Develop a *SimQuick* model of this system and interpret the results.
9. A machine shop has a large number of machines that fail regularly. One repairperson is available to fix them. Each machine fails on average every three hours, with time between failures being exponential. Repair time (in minutes) has the distribution:

Time	Probability
15	0.1
30	0.2
45	0.3
50	0.4

Develop and run a *SimQuick* model to determine the average time that machines spend waiting for repair and the average percentage of time the repairperson is busy.
10. A small factory has two serial work stations: mold/trim and assemble/package. Jobs to the factory arrive at an exponentially distributed rate of one every 10 hours. Time at the mold/trim station is exponential,

with a mean of seven hours. Each job proceeds next to the assemble/package station and requires an average of five hours, again exponentially distributed. Using *SimQuick*, estimate the average time in the system, average waiting time at mold/trim, and average waiting time at assembly/packaging.

11. Modify the factory simulation in Problem 10 to include two mold/trim servers (each with the original time distribution). How do your results change?
12. A gasoline station near an outlet mall has two self-service pumps and one full-serve pump that services only handicapped customers, which is 10% of the calling population. Customers arrive at an exponential rate of 1.6 customers per minute. Service times have discrete distributions as follows:

<b><i>Self-Serve</i></b>	
<b>Time (minutes)</b>	<b>Probability</b>
2.5	0.1
3.3	0.5
4.0	0.3
5.0	0.1

<b><i>Full-Serve</i></b>	
<b>Time (minutes)</b>	<b>Probability</b>
3.0	0.2
4.2	0.3
5.3	0.4
6.0	0.06
7.0	0.04

Customers will not wait if four or more cars are waiting in either line. Simulate this system for an eight-hour period. What percentage of customers is lost? How might you improve this system?

13. A small manufacturing facility produces custom-molded plastic parts. The process consists of two sequential operations: mold/trim and assemble/package. Jobs arrive in batches at a mean rate of five jobs every two hours. For any job, the mold/trim process is exponential, with a mean of 15 minutes, and the assemble/package operation is also exponential, with a mean of 30 minutes. The mold/trim work station has room to store only four jobs awaiting processing. The assemble/package station has room to store only two waiting jobs. Jobs are transported from the first to the second work station in 10 minutes. If the mold/trim work station is full when a job arrives, the job is subcontracted. If the assemble/package work station is full when a job

is completed at the mold/trim operation, the mold/trim operation cannot process any more jobs until space is freed up in the assemble/package area. Develop a *SimQuick* model to simulate this process, and explain the results.

14. A bank has an inside teller operation with two windows and a drive-through. Teller 2 works both inside and is responsible for the drive-through, while Teller 1 works only inside. Cars arrive to the drive-through uniformly every 5 to 15 minutes. Inside customers arrive uniformly every 10 to 20 minutes. Transaction times at the drive-through are exponential with a mean of three minutes. Times for inside transactions are exponential with a mean of four minutes. The drive-through has a capacity for only four cars. Develop a *SimQuick* model, and run it for 600 minutes. Explain the results.
15. Prescriptions at a hospital pharmacy arrive at an average rate of 22 per hour, with exponentially distributed times between arrivals. One pharmacist can fill a prescription at a rate of 12 per hour, or one every five minutes, exponentially distributed. Develop a *SimQuick* model and use it to conduct experiments to determine the optimal number of pharmacists to have.
16. A single clerk fills two types of orders. The first type arrives randomly, with a time between arrivals of 6 and 24 minutes; the second type of order has a time between arrivals of 45 and 75 minutes. It takes the clerk anywhere between 5 and 15 minutes to fill an order. Develop a *SimQuick* model to simulate this system and determine the average number of orders waiting, their average waiting time, and the percentage of time the clerk is busy.
17. Computer monitors arrive to a final inspection station with a mean time between arrivals of 5.5 minutes (exponentially distributed). Two inspectors test the units, which takes an average of 10 minutes (again exponential). However, only 90% of the units pass the test; the remaining 10% are routed to an adjustment station with a single worker. Adjustment takes an average of 30 minutes, after which the units are routed back to the inspection station. Construct a *SimQuick* model of this process and determine the average waiting time at each station, number of units in each queue, and percentage use of the workers.
18. A retail store uses a ROP inventory system for a popular MP3 player that orders 40 units whenever the ROP of 25 is reached. The delivery time from the supplier has a mean of five days, with a standard deviation of 0.25 days. The time between customer demands is exponential with a mean of two hours. The store is open 10 hours per day, seven days a week. Suppose that the order cost is \$50 per order, and the cost of holding items in inventory is \$0.40 per day. The store manager wants to minimize the total supply chain costs. Develop a *SimQuick* model for this supply chain, run

the simulation for a two-month period, and experiment with the model to find an order quantity that minimizes total cost (keeping the ROP fixed at 25).

19. An emergency room consists of four stations. Arrivals to the emergency room occur at a rate of four patients per hour and are assumed to follow an exponential distribution. Incoming patients are initially screened to determine their level of severity. Past data indicate that 5% of incoming patients require hospital admission and leave the emergency room. Thirty percent of incoming patients require ambulatory care, after which they are released. Twenty percent of incoming patients are sent to the X-ray unit, and the last 45% are sent to the laboratory unit. Of those going to the X-ray unit, 30% require admission to the hospital system, 10% are sent to the laboratory unit for additional testing, and 60% have no need of additional care and are thus released. Of patients entering the laboratory unit, 10% require hospitalization and 90% are released. Current facilities are capable of keeping up with average traffic, although there is some concern that the existing laboratory facilities can become a bottleneck on particularly busy nights, especially as the community grows. The system is shown in Figure 12.24 and consists of the following activities, durations, and flows:

Activity	Duration	Routing
Arrivals	4/hour, exponential	Initial desk
Front desk	0.05 hour (constant)	0.30 ambulatory 0.20 X-ray 0.45 lab 0.05 hospital
Ambulatory care	Normal (0.25 hour, 0.1)	Released
X-ray	Normal (0.25 hour, 0.05)	0.1 lab 0.3 hospital 0.6 released
Laboratory testing	Normal (0.5 hour, 0.1)	0.1 hospital 0.9 released

Develop a *SimQuick* model for this system to evaluate performance at each of the stations.

20. Consider the continuous system dynamics model of medical rates in this chapter. A proposal has been made to improve the system by limiting medical rate and/or insurance rate increases to a maximum of 5% per year. Modify the spreadsheet to simulate each of the following scenarios and discuss the results:
- Limit medical rate increases to 5% per year only
  - Limit insurance rate increases to 5% per year only
  - Limit both medical rate and insurance rate increases to 5% per year

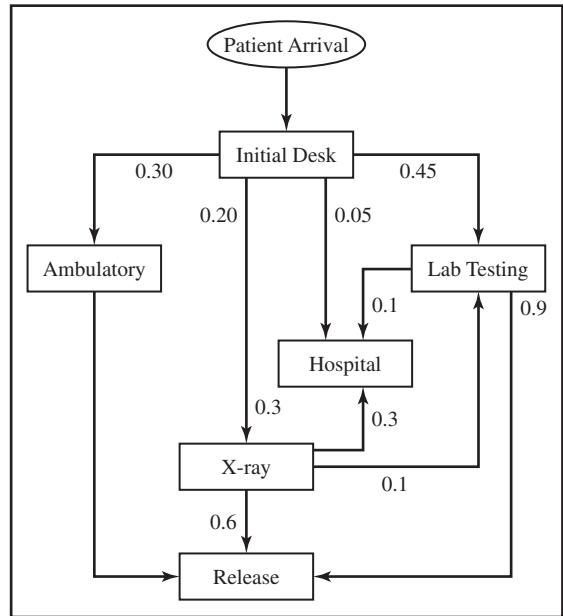


FIGURE 12.24 Emergency Room System

21. The “cobweb” model in economics assumes that the quantity demanded of a particular product in a specified time period depends on the price in that period. The quantity supplied depends on the price in the preceding period. Also, the market is assumed to be cleared at the end of each period. These assumptions can be expressed in the following equations:

$$S(t) = c + dP(t - 1) + v(t)$$

$$D(t) = a - bP(t) + w(t)$$

$$P(t) = \frac{a - c - dP(t - 1) - v(t) + w(t)}{b} + u(t)$$

where  $P(t)$  is the price in period  $t$ ,  $D(t)$  is the demand in period  $t$ , and  $S(t)$  is the quantity supplied in period  $t$ . The variables  $u(t)$ ,  $v(t)$ , and  $w(t)$  are random variables with mean zero and some variance.

- Draw an influence diagram.
- Suppose that  $a = 10,000$ ,  $b = 2$ ,  $c = 0$ ,  $d = 0.1$ ,  $u(t)$  is normal with variance 1,  $v(t)$  is normal with variance 0.5,  $w(t)$  is normal with variance 0.2, and  $P(0) = 4,738$ . Simulate this model for 50 time periods. (Note that prices are not allowed to be less than zero.)
- Examine the effect on  $P(t)$  of increasing the variance of  $v(t)$  from 0.5 to 10 and from 10 to 15.
- Compare the results from part (b) with the assumptions that  $u$ ,  $v$ , and  $w$  are fixed at zero.

## Case

### Production/Inventory Planning

A manufacturing plant supplies various engine components to manufacturers of motorcycles on a just-in-time basis. Planned production capacity for one component is 100 units per shift. Because of fluctuations in customers' assembly operations, demand fluctuates and is historically between 80 and 130 units per day. To maintain sufficient inventory to meet its just-in-time commitments, Tracway's management is considering a policy to run a second shift if inventory falls to 50 or below. For the annual budget planning process, managers need to know how many additional shifts will be needed.

The fundamental equation that governs this process each day is:

$$\text{Ending inventory} = \text{Beginning inventory} + \text{Production} - \text{Demand}$$

Develop an Excel spreadsheet model to simulate 260 working days (one year) and count the number of additional shifts that are required. Use this variable as a forecast cell for a *Crystal Ball* simulation of the model to find a distribution of the number of shifts required over the next year. Then use the *Crystal Ball Bootstrap* tool to find the sampling distribution of the maximum number of additional shifts required and compute a 95% confidence interval for this statistic. Summarize your findings in a report to the plant manager.

## *Chapter 13*

# Linear Optimization

- INTRODUCTION 435
- BUILDING LINEAR OPTIMIZATION MODELS 436
  - Characteristics of Linear Optimization Models 439
- IMPLEMENTING LINEAR OPTIMIZATION MODELS ON SPREADSHEETS 440
  - Excel Functions to Avoid in Modeling Linear Programs 441
- SOLVING LINEAR OPTIMIZATION MODELS 442
  - Solving the SSC Model Using Standard Solver 442
  - Solving the SSC Model Using *Premium Solver* 444
  - *Solver* Outcomes and Solution Messages 446
  - Interpreting *Solver* Reports 446
  - How *Solver* Creates Names in Reports 451
  - Difficulties with *Solver* 451
- APPLICATIONS OF LINEAR OPTIMIZATION 451
  - Process Selection 453
  - Blending 454
  - Portfolio Investment 456
  - Transportation Problem 457
  - Interpreting Reduced Costs 461
  - Multiperiod Production Planning 461
  - Multiperiod Financial Planning 463
  - A Model with Bounded Variables 464
  - A Production/Marketing Allocation Model 469
- HOW SOLVER WORKS 473
- BASIC CONCEPTS REVIEW QUESTIONS 474
- PROBLEMS AND APPLICATIONS 474
- CASE: HALLER'S PUB & BREWERY 481

## **INTRODUCTION**

Throughout this book, we have explored the role of data and analysis tools in managerial decisions. While many decisions involve only a limited number of alternatives and can be addressed using statistical analysis, simple decision models, or simulation, others have a very large or even an infinite

number of possibilities. We introduced **optimization**—the process of selecting values of decision variables that *minimize* or *maximize* some quantity of interest—in Chapter 9. Recall that we developed a simple optimization model for setting the best price to maximize revenue:

$$\text{Maximize revenue} = -2.794 \times \text{Price}^2 + 3149 \times \text{Price}$$

We showed that we could estimate the optimal price rather easily by developing a simple data table in a spreadsheet that evaluates this function for various values of the price, or use Excel's Solver to find the best price exactly.

Optimization models have been used extensively in operations and supply chains, finance, marketing, and other disciplines for more than 50 years to help managers allocate resources more efficiently and make lower-cost or more-profitable decisions. Optimization is a very broad and complex topic; in this chapter, we introduce you to the most common class of optimization models—linear optimization models. In the next chapter, we will discuss more complex types of optimization models, called integer and nonlinear optimization models.

## BUILDING LINEAR OPTIMIZATION MODELS

Developing any optimization model consists of four basic steps:

1. Define the decision variables.
2. Identify the objective function.
3. Identify all appropriate constraints.
4. Write the objective function and constraints as mathematical expressions.

**Decision variables** are the unknown values that the model seeks to determine. Depending on the application, decision variables might be the quantities of different products to produce, amount of money spent on R&D projects, the amount to ship from a warehouse to a customer, the amount of shelf space to devote to a product, and so on. The quantity we seek to minimize or maximize is called the **objective function**; for example, we might wish to maximize profit or revenue, or minimize cost or some measure of risk. Any set of decision variable values that maximizes or minimizes (in generic terms, *optimizes*) the objective function is called an **optimal solution**.

Most practical optimization problems consist of many decision variables and numerous **constraints**—limitations or requirements that decision variables must satisfy. Some examples of constraints are as follows:

- The amount of material used to produce a set of products cannot exceed the available amount of 850 square feet.
- The amount of money spent on research and development projects cannot exceed the assigned budget of \$300,000.
- Contractual requirements specify that at least 500 units of product must be produced.
- A mixture of fertilizer must contain exactly 30% nitrogen.
- We cannot produce a negative amount of product (this is called a *nonnegativity constraint*).

The presence of constraints along with a large number of variables usually makes identifying an optimal solution considerably more difficult and necessitates the use of powerful software tools.

The essence of building an optimization model is to translate constraints and the objective function into mathematical expressions. Constraints are generally expressed mathematically as algebraic inequalities or equations. Note that the phrase “cannot exceed” specifies a “ $\leq$ ” inequality, “at least” specifies a “ $\geq$ ” inequality, and “must contain exactly” specifies an “ $=$ ” relationship. All constraints in optimization

models must be one of these three forms. Thus, for the examples previously provided, we would write:

- Amount of material used  $\leq$  850 square feet
- Amount spent on research and development  $\leq$  \$300,000
- Number of units of product produced  $\geq$  500
- Amount of nitrogen in mixture/total amount in mixture = 0.30
- Amount of product produced  $\geq$  0

The left-hand side of each of these expressions is called a **constraint function**. A constraint function is a function of the decision variables in the problem. For example, suppose that in the first case, we are producing three products. Further assume that the material requirements of these three products are 3.0, 3.5, and 2.3 square feet per unit, respectively. If  $A$ ,  $B$ , and  $C$  represent the number of units of each product to produce, then  $3.0A$  represents the amount of material used to produce  $A$  units of product A,  $3.5B$  represents the amount of material used to produce  $B$  units of product B, and  $2.3C$  represents the amount of material used to produce  $C$  units of product C. Note that dimensions of these terms are (square feet/unit)(units) = square feet. Hence, “amount of material used” can be expressed mathematically as the constraint function  $3.0A + 3.5B + 2.3C$ . Therefore, the constraint that limits the amount of material that can be used is written as:

$$3.0A + 3.5B + 2.3C \leq 850$$

As another example, if two ingredients contain 20% and 33% nitrogen, respectively, then the fraction of nitrogen in a mixture of  $x$  pounds of the first ingredient and  $y$  pounds of the second ingredient is expressed by the constraint function:

$$(0.20x + 0.33y)/(x + y)$$

If the fraction of nitrogen in the mixture must be 0.30, then we would have:

$$(0.20x + 0.33y)/(x + y) = 0.3$$

This can be rewritten as:

$$(0.20x + 0.33y) = 0.3(x + y)$$

and simplified as:

$$-0.1x + 0.03y = 0$$

In a similar fashion, we must translate the objective function into a mathematical expression involving the decision variables. To see this entire process in action, let us examine a typical decision scenario:

Sklenka Ski Company (SSC) is a small manufacturer of two types of popular all-terrain snow skis, the Jordanelle and Deercrest models. The manufacturing process consists of two principal departments: fabrication and finishing. The fabrication department has 12 skilled workers, each of whom works 7 hours per day. The finishing department has three workers, who also work a 7-hour shift. Each pair of Jordanelle skis requires 3.5 labor hours in the fabricating department and one labor hour in finishing. The Deercrest model requires four labor hours in fabricating and 1.5 labor hours in finishing. The company operates five days per week. SSC makes a net profit of \$50 on the Jordanelle model, and \$65 on the Deercrest model. In anticipation of the next ski sale season, SSC must plan its production of these two models. Because

of the popularity of its products and limited production capacity, its products are in high demand and SSC can sell all it can produce each season. The company anticipates selling at least twice as many Deercrest models as Jordanelle models. The company wants to determine how many of each model should be produced on a daily basis to maximize net profit.

**Step 1: Define the decision variables.** SSC wishes to determine how many of each model skis to produce. Thus, we may define

$Jordanelle$  = number of pairs of Jordanelle skis produced/day

$Deercrest$  = number of pairs of Deercrest skis produced/day

We usually represent decision variables by short, descriptive names, abbreviations, or subscripted letters such as  $X_1$  and  $X_2$ . For many mathematical formulations involving many variables, subscripted letters are often more convenient; however, in spreadsheet models, we recommend using more descriptive names to make the models and solutions easier to understand. Also, it is very important to clearly specify the dimensions of the variables, for example, "pairs/day" rather than simply "Jordanelle skis."

**Step 2: Identify the objective function.** The problem states that SSC wishes to maximize profit. In some problems, the objective is not explicitly stated, and you must use logic and business experience to identify the appropriate objective.

**Step 3: Identify the constraints.** From the information provided, we see that labor hours are limited in both the fabrication department and finishing department. Therefore, we have the constraints:

Fabrication: Total labor used in fabrication cannot exceed  
the amount of labor available.

Finishing: Total labor used in finishing cannot exceed  
the amount of labor available.

In addition, the problem notes that the company anticipates selling at least twice as many Deercrest models as Jordanelle models. Thus, we need a constraint that states:

The number of pairs of Deercrest skis must be at least twice  
the number of Jordanelle skis.

Finally, we must ensure that negative values of the decision variables cannot occur. Nonnegativity constraints are assumed in nearly all optimization models.

**Step 4: Write the objective function and constraints as mathematical expressions**  
Because SSC makes a net profit of \$50 on the Jordanelle model, and \$65 on the Deercrest model, the objective function is:

Maximize total profit = 50  $Jordanelle$  + 65  $Deercrest$

For the constraints, we will use the approach described earlier in this chapter. First, consider the fabrication and finishing constraints. Write these as:

Fabrication: Total labor used in fabrication  
 $\leq$  the amount of labor available

Finishing: Total labor used in finishing  
 $\leq$  the amount of labor available

Now translate both the constraint functions on the left and the limitations on the right into mathematical or numerical terms. Note that the amount of labor available in fabrication is (12 workers) (7 hours/day) = 84 hours/day, while in finishing we have (3 workers) (7 hours/day) = 21 hours/day. Because each pair of Jordanelle skis requires 3.5 labor hours and Deercrest skis require 4 labor hours in the fabricating department, the total labor used in fabrication is  $3.5 \text{ Jordanelle} + 4 \text{ Deercrest}$ . Note that the dimensions of these terms are (hours/pair of skis) (number of pairs of skis produced) = hours. Similarly, for the finishing department, the total labor used is  $1 \text{ Jordanelle} + 1.5 \text{ Deercrest}$ . Therefore, the appropriate constraints are:

$$\text{Fabrication: } 3.5 \text{ Jordanelle} + 4 \text{ Deercrest} \leq 84$$

$$\text{Finishing: } 1 \text{ Jordanelle} + 1.5 \text{ Deercrest} \leq 21$$

For the market mixture constraint, "Number of pairs of Deercrest skis must be at least twice the number of pairs of Jordanelle skis," we have:

$$\text{Deercrest} \geq 2 \text{ Jordanelle}$$

It is customary to write all the variables on the left-hand side of the constraint. Thus, an alternative expression for this constraint is:

$$\text{Deercrest} - 2 \text{ Jordanelle} \geq 0$$

The difference between the number of pairs of Deercrest skis and twice the number of pairs of Jordanelle skis can be thought of as the excess number of pairs of Deercrest skis produced over the minimum market mixture requirement. Finally, nonnegativity constraints are written as:

$$\text{Deercrest} \geq 0$$

$$\text{Jordanelle} \geq 0$$

The complete optimization model is:

$$\text{Maximize total profit} = 50 \text{ Jordanelle} + 65 \text{ Deercrest}$$

$$3.5 \text{ Jordanelle} + 4 \text{ Deercrest} \leq 84$$

$$1 \text{ Jordanelle} + 1.5 \text{ Deercrest} \leq 21$$

$$\text{Deercrest} - 2 \text{ Jordanelle} \geq 0$$

$$\text{Deercrest} \geq 0$$

$$\text{Jordanelle} \geq 0$$

## Characteristics of Linear Optimization Models

A *linear* optimization model (often called a **linear program**, or LP) has two basic properties. First, the objective function and all constraints are *linear functions* of the decision variables. This means that each function is simply a sum of terms, each of which is some constant multiplied by a decision variable. The SSC model has this property. Recall the constraint example that we developed earlier for the nitrogen requirement. Notice that the constraint function on the left-hand side of the constraint:

$$(0.20x + 0.33y)/(x + y) = 0.3$$

as originally written is not linear. However, we were able to convert it to a linear form using simple algebra. This is advantageous, as special, highly efficient solution algorithms are used for linear optimization problems.

The second property of a linear optimization problem is that all variables are *continuous*, meaning that they may assume any real value (typically, nonnegative). Of course, this assumption may not be realistic for a practical business problem (you cannot produce half a refrigerator!). However, because this assumption simplifies the solution method and analysis, we often apply it in many situations where the solution would not be seriously affected. In the next chapter, we will discuss situations where it is necessary to force variables to be whole numbers (integers). For all examples and problems in this chapter, we will assume continuity of the variables.

## IMPLEMENTING LINEAR OPTIMIZATION MODELS ON SPREADSHEETS

We will learn how to solve optimization models using an Excel tool called *Solver*. To facilitate the use of *Solver*, we suggest the following spreadsheet engineering guidelines for designing spreadsheet models for optimization problems:

- **Put the objective function coefficients, constraint coefficients, and right-hand values in a logical format in the spreadsheet.** For example, you might assign the decision variables to columns and the constraints to rows, much like the mathematical formulation of the model, and input the model parameters in a matrix. If you have many more variables than constraints, it might make sense to use rows for the variables and columns for the constraints.
- **Define a set of cells (either rows or columns) for the values of the decision variables.** In some models, it may be necessary to define a matrix to represent the decision variables. The names of the decision variables should be listed directly above the decision variable cells. Use shading or other formatting to distinguish these cells.
- **Define separate cells for the objective function and each constraint function (the left-hand side of a constraint).** Use descriptive labels directly above these cells.

We will illustrate these principles for the Sklenka Ski example. Figure 13.1 shows a spreadsheet model for the product mix example (Excel file *Sklenka Skis*). The *Data* portion of the spreadsheet provides the objective function coefficients, constraint coefficients, and right-hand sides of the model. Such data should be kept separate from the actual model so that if any data are changed, the model will automatically be updated. In the *Model* section, the number of each product to make is given in cells B14 and C14. Also in the *Model* section are calculations for the constraint functions:

$$3.5 \text{ Jordanelle} + 4 \text{ Deercrest} \quad (\text{hours used in fabrication, cell D15})$$

$$1 \text{ Jordanelle} + 1.5 \text{ Deercrest} \quad (\text{hours used in finishing, cell D16})$$

$$\text{Deercrest} - 2 \text{ Jordanelle} \quad (\text{market mixture, cell D19})$$

and the objective function,  $50 \text{ Jordanelle} + 65 \text{ Deercrest}$  (cell D22).

To show the correspondence between the mathematical model and the spreadsheet model more clearly, we will write the model in terms of the spreadsheet cells:

$$\text{Maximize profit} = \text{D22} = \text{B9*B14} + \text{C9*C14}$$

subject to the constraints:

$$\text{D15} = \text{B6*B14} + \text{C6*C14} \leq \text{D6} \quad (\text{fabrication})$$

$$\text{D16} = \text{B7*B14} + \text{C7*C14} \leq \text{D7} \quad (\text{finishing})$$

A	B	C	D
1 Sklenka Skis			
2			
3 Data			
4			
5 Department	Jordanelle	Deercrest	Limitation (min.)
6 Fabrication	3.5	4	84
7 Finishing	1	1.5	21
8			
9 Profit/unit	\$ 50.00	\$ 65.00	
10			
11			
12 Model			
13			
14 Quantity Produced	Jordanelle	Deercrest	Hours Used
15 Fabrication	0	0	0
16 Finishing	0	0	0
17			
18			Excess Deercrest
19 Market mixture			0
20			
21			Total Profit
22 Profit Contribution	\$ -	\$ -	\$ -

A	B	C	D
1 Sklenka Skis			
2			
3 Data			
4			
5 Department	Jordanelle	Deercrest	Limitation (min.)
6 Fabrication	3.5	4	84
7 Finishing	1	1.5	21
8			
9 Profit/unit	50	65	
10			
11			
12 Model			
13			
14 Quantity Produced	Jordanelle	Deercrest	Hours Used
15 Fabrication	=B6*\$B\$14	=C6*\$C\$14	=B15+C15
16 Finishing	=B7*\$B\$14	=C7*\$C\$14	=B16+C16
17			
18			Excess Deercrest
19 Market mixture			=C14-2*B14
20			
21			Total Profit
22 Profit Contribution	=B9*\$B\$14	=C9*\$C\$14	=B22+C22

**FIGURE 13.1** Sklenka Skis Model Spreadsheet Implementation

$$D19 = C14 - 2*B14 \geq 0 \quad (\text{market mixture})$$

$$B14 \geq 0, C14 \geq 0 \quad (\text{nonnegativity})$$

Observe how the constraint functions and right-hand-side values are stored in separate cells within the spreadsheet.

In Excel, the pair-wise sum of products of terms can easily be computed using the SUMPRODUCT function. This often simplifies the model-building process, particularly when many variables are involved. For example, the objective function formula could have been written as:

$$B9*B14 + C9*C14 = \text{SUMPRODUCT}(B9:C9, B14:C14)$$

Similarly, the labor limitation constraints could have been expressed as:

$$B6*B14 + C6*C14 = \text{SUMPRODUCT}(B6:C6, B14:C14)$$

$$B7*B14 + C7*C14 = \text{SUMPRODUCT}(B7:C7, B14:C14)$$

## Excel Functions to Avoid in Modeling Linear Programs

Several common functions in Excel can cause difficulties when attempting to solve linear programs using *Solver* because they are discontinuous (or “nonsmooth”) and no longer satisfy the conditions of a linear model. For instance, in the formula IF(A12 < 45, 0, 1), the cell value jumps from 0 to 1 when the value of cell A12 crosses 45. In such situations, the correct solution may not be identified. Common Excel functions to avoid are ABS, MIN, MAX, INT, ROUND, IF, and COUNT. While these are useful in

general modeling tasks with spreadsheets, you should avoid them in linear optimization models.

## SOLVING LINEAR OPTIMIZATION MODELS

To solve an optimization problem, we seek values of the decision variables that maximize or minimize the objective function and also satisfy all constraints. Any solution that satisfies all constraints of a problem is called a **feasible solution**. Finding an optimal solution among the infinite number of possible feasible solutions to a given problem is not an easy task. A simple approach is to try to manipulate the decision variables in the spreadsheet models to find the best solution possible; however, for many problems, it might be very difficult to find a feasible solution, let alone an optimal solution. You might try to find the best solution possible for the Sklenka Ski problem by using the spreadsheet model. With a little experimentation and perhaps a bit of luck, you might be able to zero in on the optimal solution or something close to it. However, to guarantee finding an optimal solution, some type of systematic mathematical solution procedure is necessary. Fortunately, such a procedure is provided by the Excel *Solver* tool, which we discuss next.

*Solver* is an add-in packaged with Excel that was developed by Frontline Systems Inc. ([www.solver.com](http://www.solver.com)) and can be used to solve many different types of optimization problems. *Premium Solver*, which is part of *Risk Solver Platform* that accompanies this book, is an improved alternative to the standard Excel-supplied *Solver*. *Premium Solver* has better functionality, numerical accuracy, reporting, and user interface. We will show how to solve the SSC model using both the standard and premium versions; however, we highly recommend using the premium version, and we will use it in the remainder of this chapter.

### Solving the SSC Model Using Standard Solver

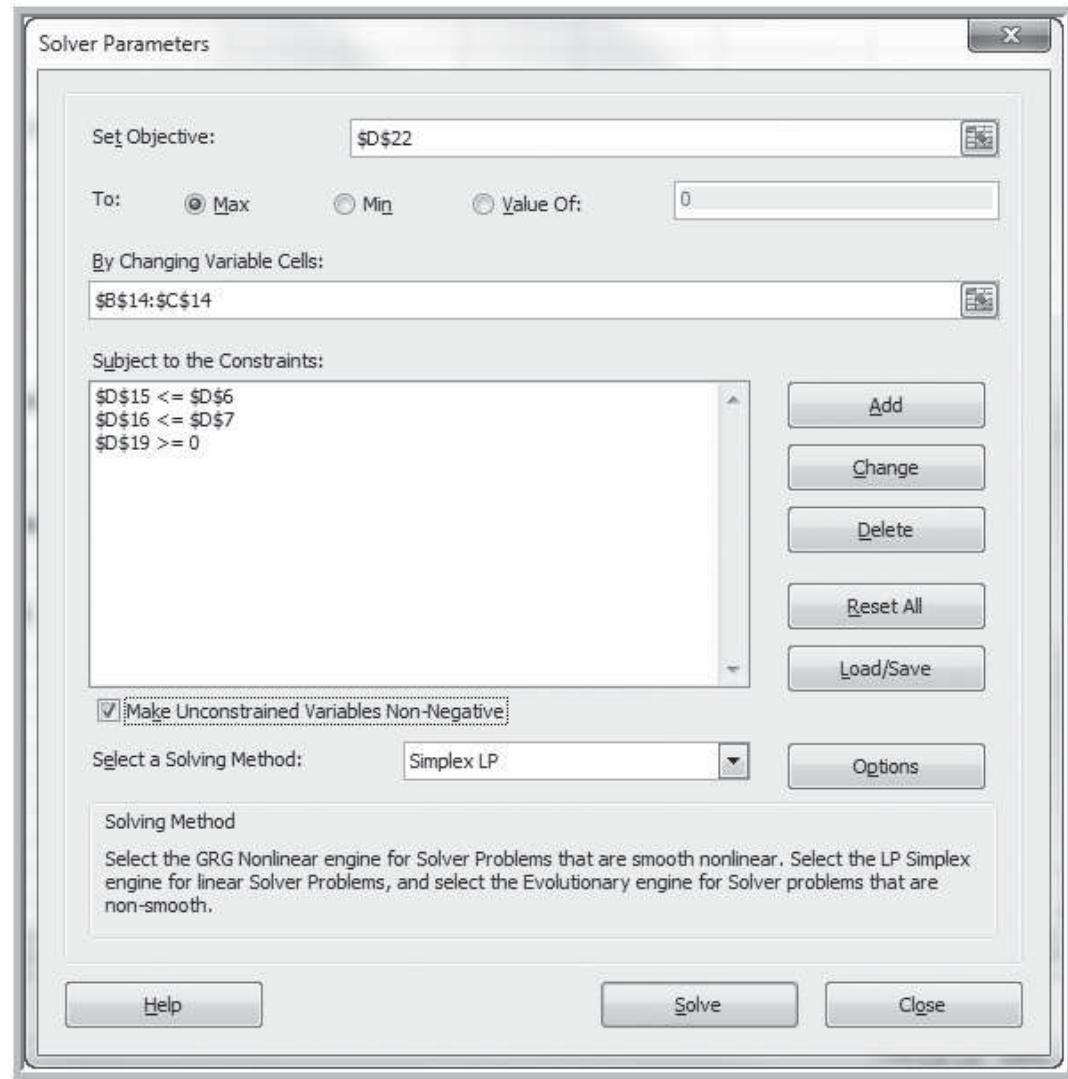
The standard *Solver* can be found in the *Analysis* group under the *Data* tab in Excel 2010 (see Chapter 9 for installation directions). Figure 13.2 shows the completed *Solver Parameters* dialog for this example. The objective function cell in the spreadsheet (D22) is defined in the *Set Objective* field. Either enter the cell reference or click within the field and then in the cell in the spreadsheet. Click the appropriate radio button for *Max* or *Min*. Decision variables (cells B14 and C14) are entered in the field called *By Changing Variable Cells*; click within this field and highlight the range corresponding to the decision variables in your spreadsheet.

To enter a constraint, click the *Add* button. A new dialog, *Add Constraint*, appears (see Figure 13.3). In the left field, *Cell Reference*, enter the cell that contains the constraint function (left-hand side of the constraint). For example, the constraint function for the fabrication constraint is in cell D15. Make sure that you select the correct type of constraint ( $\leq$ ,  $\geq$ , or  $=$ ) in the drop down box in the middle of the dialog. The other options are discussed in the next chapter. In the right field, called *Constraint*, enter the numerical value of the right-hand side of the constraint or the cell reference corresponding to it. For the fabrication constraint, this is cell D6. Figure 13.3 shows the completed dialog for the fabrication constraint. To add other constraints, click the *Add* button.

You may also define a group of constraints that all have the same algebraic form (either all  $\leq$ , all  $\geq$ , or all  $=$ ) and enter them together. For example, the department resource limitation constraints are expressed within the spreadsheet model as:

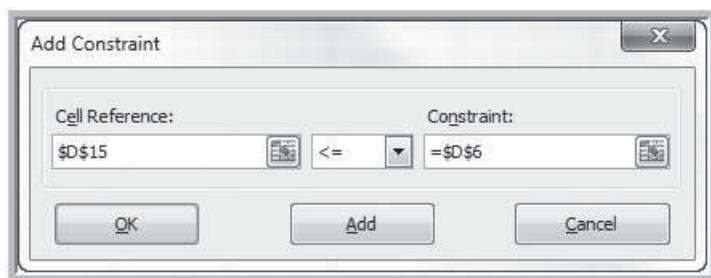
$$D15 \leq D6$$

$$D16 \leq D7$$



**FIGURE 13.2** Solver Parameters Dialog

Because both constraints are  $\leq$  types, we could define them as a group by entering the range D15:D16 in the *Cell Reference* field and D6:D7 in the *Constraint* field to simplify the input process. When all constraints are added, click *OK* to return to the *Solver Parameters* dialog box. You may add, change, or delete these as necessary by clicking the



**FIGURE 13.3** Add Constraint Dialog

appropriate buttons. You need not enter nonnegativity constraints explicitly. Just check the box in the dialog *Make Unconstrained Variables Non-Negative*.

For linear optimization problems it is very important to select the correct solving method. The standard Excel *Solver* provides three options for the solving method:

1. *Standard GRG Nonlinear*—used for solving nonlinear optimization problems
2. *Standard LP Simplex*—used for solving linear and linear integer optimization problems
3. *Standard Evolutionary*—used for solving complex nonlinear and nonlinear integer problems

In the field labeled *Select a Solving Method*, choose *Simplex LP*. Then click the *Solve* button to solve the problem. The *Solver Results* dialog will appear, as shown in Figure 13.4, with the message “Solver found a solution.” If a solution could not be found, *Solver* will notify you with a message to this effect. This generally means that you have an error in your model or you have included conflicting constraints that no solution can satisfy. In such cases, you will need to reexamine your model.

*Solver* generates three reports as listed in Figure 13.4: Answer, Sensitivity, and Limits. To add them to your Excel workbook, click on the ones you want, and then click *OK*. The optimal solution will be shown in the spreadsheet as in Figure 13.5. The maximum profit is \$945, obtained by producing 5.25 pairs of Jordanelle skis and 10.5 pairs of Deercrest skis per day (remember that linear models allow fractional values for the decision variables!). If you save your spreadsheet after setting up a *Solver* model, the *Solver* model will also be saved.

### Solving the SSC Model Using Premium Solver

*Premium Solver* has a different user interface than the standard *Solver*. After installing *Risk Solver Platform*, *Premium Solver* will be found under the *Add-Ins* tab in the Excel ribbon. Figure 13.6 shows the *Premium Solver* dialog. First click on *Objective* and then click the *Add* button. The *Add Objective* dialog appears, prompting you for the cell reference for the objective function and the type of objective (min or max) similar

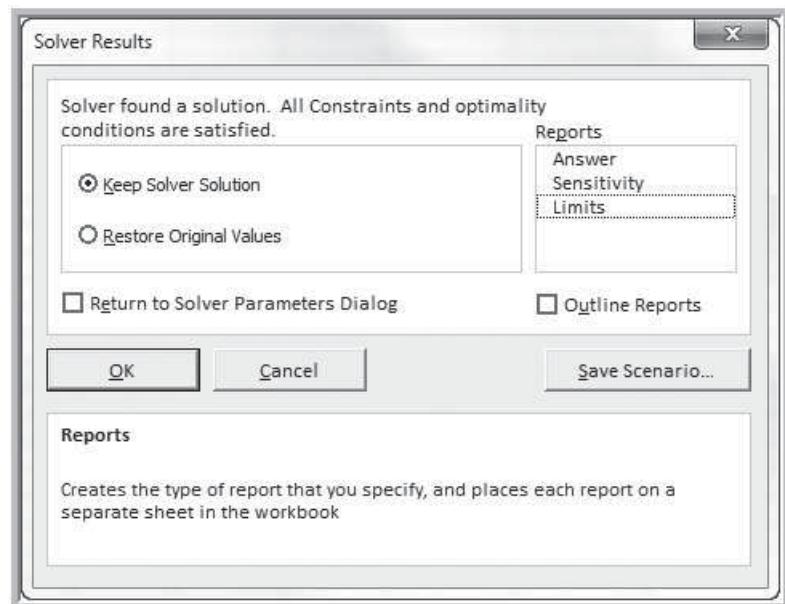


FIGURE 13.4 Solver Results Dialog

A	B	C	D
1 Sklenka Skis			
2			
3 Data			
4			
5 Department	Jordanelle	Deercrest	Limitation (min.)
6 Fabrication	3.5	4	84
7 Finishing	1	1.5	21
8			
9 Profit/unit	\$ 50.00	\$ 65.00	
10			
11			
12 Model			
13	Jordanelle	Deercrest	
14 Quantity Produced	5.25	10.5	Hours Used
15 Fabrication	18.375	42	60.375
16 Finishing	5.25	15.75	21
17			
18			Excess Deercrest
19 Market mixture			0
20			
21			Total Profit
22 Profit Contribution	\$ 262.50	\$ 682.50	\$ 945.00

**FIGURE 13.5** Optimal Solution to the SSC Model



**FIGURE 13.6** Premium Solver Parameters Dialog

to the top portion of the standard *Solver Parameters* dialog. Next, highlight *Normal* under the *Variables* list and click *Add*; this will bring up an *Add Variable Cells* dialog. Enter the range of the decisions variables in the *Cell Reference* field. Next, highlight *Normal* under the *Constraints* list and click the *Add* button; this brings up the *Add Constraint* dialog just like in the standard version. Add the constraints in the same fashion as in the standard *Solver*. Check the box *Make Unconstrained Variables Non-Negative*. The premium version provides the same solving method options as the standard version (except that *Standard LP Simplex* is called *Standard LP/Quadratic*), so select this for linear optimization. Quadratic optimization problems have a special structure that we will not address in this book. The premium version also has three additional advanced solving methods. The *Solver Results* dialog is the same as in the standard version.

## **Solver Outcomes and Solution Messages**

Solving a linear optimization model can result in four possible outcomes:

1. Unique optimal solution
2. Alternate optimal solutions
3. Unboundedness
4. Infeasibility

When a model has a *unique optimal solution*, it means that there is exactly one solution that will result in the maximum (or minimum) objective. The solution to the SSC model is unique. If a model has *alternate optimal solutions*, the objective is maximized (or minimized) by more than one combination of decision variables, all of which have the same objective function value. *Solver* does not tell you when alternate solutions exist and only reports one of the many possible alternate optimal solutions. However, you can use the sensitivity report information to identify the existence of alternate optimal solutions. When any of the Allowable Increase or Allowable Decrease values for changing cells are zero, then alternate optimal solutions exist, although *Solver* does not provide an easy way to find them.

A problem is *unbounded* if the objective can be increased or decreased without bound (i.e., to infinity or negative infinity) while the solution remains feasible. A model is unbounded if *Solver* reports “The Set Cell values do not converge.” This generally indicates an incorrect model, usually when some constraint or set of constraints have been left out.

Finally, an *infeasible* model is one for which no feasible solution exists; that is, when there is no solution that satisfies all constraints together. When a problem is infeasible, *Solver* will report “Solver could not find a feasible solution.” Infeasible problems *can* occur in practice, for example, when a demand requirement is higher than available capacity, or when managers in different departments have conflicting requirements or limitations. In such cases, the model must be reexamined and modified. Sometimes infeasibility or unboundedness is simply a result of a misplaced decimal or other error in the model or spreadsheet implementation, so accuracy checks should be made.

## **Interpreting Solver Reports**

The Answer Report, shown in Figure 13.7 (all reports in this section were generated using *Premium Solver*), provides basic information about the solution, including the values of the optimal objective function (in the *Objective Cell* section) and decision variables (in the *Decision Variable Cells* section). In the *Constraints* section, *Cell Value* refers to the value of the constraint function using the optimal values of the decision variables. In other words, we used 60.375 minutes in the fabrication department and 21 minutes

A	B	C	D	E	F	G
11						
12	Objective Cell (Max)					
13	Cell	Name	Original Value	Final Value		
14	\$D\$22	Profit Contribution Total Profit	0	945		
15						
16						
17	Decision Variable Cells					
18	Cell	Name	Original Value	Final Value		
19	\$B\$14	Quantity Produced Jordanelle	0	5.25		
20	\$C\$14	Quantity Produced Deercrest	0	10.5		
21						
22	Constraints					
23	Cell	Name	Cell Value	Formula	Status	Slack
24	\$D\$15	Fabrication Hours Used	60.375	\$D\$15<=\$D\$6	Not Binding	23.625
25	\$D\$16	Finishing Hours Used	21	\$D\$16<=\$D\$7	Binding	0
26	\$D\$19	Market mixture Excess Deercrest	0	\$D\$19>=0	Binding	0

**FIGURE 13.7** Solver Answer Report

in the finishing department by producing 5.25 pairs of Jordanelle skis and 10.5 pairs of Deercrest skis. The *Status* column tells whether each constraint is binding or not binding. A **binding constraint** is one for which the *Cell Value* is equal to the right-hand side of the value of the constraint. In this example, the constraint for fabrication is not binding, while the constraints for finishing and market mixture are binding. This means that there is excess time that is not used in fabrication; this value is shown in the *Slack* column as 23.626 hours. For finishing, we used all the time available, and hence, the slack value is zero. Because we produced exactly twice the number of Deercrest skis as Jordanelle skis, the market mixture constraint is binding. It would have been not binding if we had produced more than twice the number of Deercrest skis as Jordanelle.

In general, the **slack** is the difference between the right- and left-hand sides of a constraint. Examine the fabrication constraint:

$$3.5 \text{ Jordanelle} + 4 \text{ Deercrest} \leq 84$$

We interpret this as:

$$\text{Number of fabrication hours used} \leq \text{Hours available}$$

Note that if the amount used is strictly less than the availability, we have slack, which represents the amount unused; thus,

$$\begin{aligned} &\text{Number of fabrication hours used} + \text{Number of fabrication hours unused} \\ &= \text{Hours available} \end{aligned}$$

or

$$\begin{aligned} \text{Slack} &= \text{Number of hours unused} \\ &= \text{Hours Available} - \text{Number of fabrication hours used} \\ &= 84 - (3.5 \times 5.25 + 4 \times 10.5) = 23.625 \end{aligned}$$

Slack variables are always nonnegative, so for  $\geq$  constraints, slack represents the difference between the left-hand side of the constraint function and the right-hand side of the requirement. The slack on a binding constraint will always be zero.

A	B	C	D	E	F	G	H
4							
5	Objective Cell (Max)						
6	Cell	Name	Final Value				
7	\$D\$22	Profit Contribution Total Profit	945				
8							
9	Decision Variable Cells						
10	Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
11	\$B\$14	Quantity Produced Jordanelle	5.25	0	50	1E+30	6.6666668
12	\$C\$14	Quantity Produced Deercrest	10.5	0	65	10.0000002	90.00000013
13							
14							
15	Constraints						
16	Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
17	\$D\$15	Fabrication Hours Used	60.375	0	84	1E+30	23.625
18	\$D\$16	Finishing Hours Used	21	45	21	8.217391304	21
19	\$D\$19	Market mixture Excess Deercrest	0	-2.5	0	14	42
20							

**FIGURE 13.8** Solver Sensitivity Report

The Sensitivity Report (Figure 13.8) provides a variety of useful information for managerial interpretation of the solution. Specifically, it allows us to understand how the optimal objective value and optimal decision variables are affected by changes in the objective function coefficients, the impact of forced changes in certain decision variables, or the impact of changes in the constraint resource limitations or requirements. In the *Decision Variable Cells* section, the final value for each decision variable is given, along with its reduced cost, objective coefficient, and allowable increase and decrease. The **reduced cost** tells *how much the objective coefficient needs to be reduced in order for a nonnegative variable that is zero in the optimal solution to become positive*. If a variable is positive in the optimal solution, as it is for both variables in this example, its reduced cost is always zero. We will see an example later in this chapter that will help you to understand reduced costs.

The Allowable Increase and Allowable Decrease values tell how much an individual objective function coefficient can change before the optimal values of the decision variables will change (a value listed as “1E + 30” is interpreted as infinity). For example, if the unit profit for Deercrest skis either increases by more than 10 or decreases by more than 90, then the optimal values of the decision variables will change (as long as the other objective coefficient stays the same). For instance, if we increase the unit profit by 11 (to 76) and re-solve the model, the new optimal solution will be to produce 14 pairs of Deercrest skis and no Jordanelle skis. However, any increase less than 10 will keep the current solution optimal. For Jordanelle skis, we can increase the unit profit as much as we wish without affecting the current optimal solution; however, a decrease of at least 6.67 will cause the solution to change.

Note that if the objective coefficient of any one variable that has positive value in the current solution changes but stays within the range specified by the Allowable Increase and Allowable Decrease, the optimal decision variables will stay the same; however, the objective function value will change. For example, if the unit profit of Jordanelle skis were changed to \$46 (a decrease of 4, within the allowable increase), then we are guaranteed that the optimal solution will still be to produce 5.25 pairs of Jordanelle and 10.5 pairs of Deercrest. However, each of the 5.25 pairs of Jordanelle skis produced and sold would realize \$4 less profit—a total decrease of  $5.25(\$4) = \$21$ . Thus, the new value of the objective function would be  $\$945 - \$21 = \$924$ . If an objective coefficient

changes beyond the Allowable Increase or Allowable Decrease, then we must re-solve the problem with the new value to find the new optimal solution and profit.

The range within which the objective function coefficients will not change the optimal solution provides a manager with some confidence about the stability of the solution in the face of uncertainty. If the allowable ranges are large, then reasonable errors in estimating the coefficients will have no effect on the optimal policy (although they will affect the value of the objective function). Tight ranges suggest that more effort might be spent in ensuring that accurate data or estimates are used in the model.

The *Constraints* section of the Sensitivity Report lists the final value of the constraint function (the left-hand side), the shadow price, the original constraint right-hand side, and an Allowable Increase and Allowable Decrease. The **shadow price** tells *how much the value of the objective function will change as the right-hand side of a constraint is increased by 1*. Whenever a constraint has positive slack, the shadow price is zero. For example, in the fabrication constraint, we are not using all of the available hours in the optimal solution. Thus, having one more hour available will not help us improve the solution. However, if a constraint is binding, then any change in the right-hand side will cause the optimal values of the decision variables as well as the objective function value to change.

Let us illustrate this with the finishing constraint. The shadow price of 45 states that if an additional hour of finishing time is available, then the total profit will change by \$45. To see this, change the limitation of the number of finishing hours available to 22 and re-solve the problem. The new solution is to produce 5.5 pairs of Jordanelle and 11.0 pairs of Deercrest, yielding a profit of \$990. We see that the total profit increases by \$45 as predicted. Thus, the shadow price represents the economic value of having an additional unit of a resource.

The shadow price is a valid predictor of the change in the objective function value for each unit of increase in the constraint right-hand side up to the value of the Allowable Increase. Thus, if up to about 8.2 additional hours of finishing time were available, profit would increase by \$45 for each additional hour (but we would have to resolve the problem to actually find the optimal values of the decision variables). Similarly, the negative of the shadow price predicts the change in the objective function value for each unit the constraint right-hand side is decreased, up to the value of the Allowable Decrease. For example, if one person was ill or injured, resulting in only 14 hours of finishing time available, then profit would decrease by  $7(\$45) = \$315$ . This can be predicted because a decrease of 7 hours is within the Allowable Decrease of 21. Beyond these ranges, the shadow price does not predict what will happen, and the problem must be re-solved.

Another way of understanding the shadow price is to break down the impact of a change in the right-hand side of the value. How was the extra hour of finishing time used? After solving the model with 22 hours of finishing time, we see that we were able to produce an additional 0.25 pairs of Jordanelle and 0.5 pairs of Deercrest skis as compared to the original solution. Therefore, the profit increased by  $0.25(\$50) + 0.5(65) = \$12.50 + 32.50 = \$45$ . In essence, a change in a binding constraint causes a reallocation of how the resources are used.

Interpreting the shadow price associated with the market mixture constraint is a bit more difficult. If you examine the constraint,  $Deercrest - 2 Jordanelle \geq 0$ , closely, an increase in the right-hand side from 0 to 1 results in a change of the constraint to:

$$(Deercrest - 1) - 2 Jordanelle \geq 0$$

This means that the number of pairs of Deercrest skis produced would be one short of the requirement that it be at least twice the number of Jordanelle skis. If the problem is re-solved with this constraint, we find the new optimal solution to be 4.875 Jordanelle, 10.75 Deercrest, and profit = 942.50. The profit changed by the value of the shadow price and we see that  $2 \times Jordanelle = 9.75$ , one short of the requirement.

A	B	C	D	E	F	G	H	I	J
4									
5									
6	<b>Objective</b>								
7	<b>Cell</b>	<b>Name</b>	<b>Value</b>						
8	\$D\$22	Profit Contribution Total Profit	\$ 945.00						
9									
10	<b>Decision Variable</b>								
11	<b>Cell</b>	<b>Name</b>	<b>Value</b>	<b>Lower Limit</b>	<b>Objective Result</b>	<b>Upper Limit</b>	<b>Objective Result</b>		
12	\$B\$14	Quantity Produced Jordanelle	5.25	0	\$ 682.50	5.25	\$ 945.00		
13	\$C\$14	Quantity Produced Deercrest	10.5	10.5	\$ 945.00	10.5	\$ 945.00		

**FIGURE 13.9** Solver Limits Report

Why are shadow prices useful to a manager? They provide guidance on how to reallocate resources or change values over which the manager may have control. In linear optimization models, the parameters of some constraints cannot be controlled. For instance, the amount of time available for production or physical limitations on machine capacities would clearly be uncontrollable. Other constraints represent policy decisions, which, in essence, are arbitrary. Although it is correct to state that having an additional hour of finishing time will improve profit by \$45, does this necessarily mean that the company should spend up to this amount for additional hours? This depends on whether the relevant costs have been included in the objective function coefficients. If the cost of labor *has not* been included in the objective function unit profit coefficients, then the company will benefit by paying less than \$45 for additional hours. However, if the cost of labor *has* been included in the profit calculations, the company should be willing to pay up to an *additional* \$45 over and above the labor costs that have already been included in the unit profit calculations.

The Limits Report (Figure 13.9) shows the lower limit and upper limit that each variable can assume while satisfying all constraints and holding all of the other variables constant. Generally, this report provides little useful information for decision making and can be effectively ignored.

### SKILL-BUILDER EXERCISE 13.1

Make the following changes to the Sklenka Ski Company model, re-solve using *Solver*, and answer the following questions:

- Increase the unit profit on Jordanelle skis by \$10. What happens to the solution? Could you have predicted this from the Sensitivity Report (Figure 13.8)?
- Decrease the unit profit on Jordanelle skis by \$10. What happens to the solution? Could you have predicted this from the Sensitivity Report (Figure 13.8)?
- Increase the number of finishing hours available by 10. What happens to the solution? Could you have predicted this from the Sensitivity Report (Figure 13.8)?
- Decrease the number of finishing hours available by 10. What happens to the solution? Could you have predicted this from the Sensitivity Report (Figure 13.8)?
- Change the unit profit for Deercrest skis to \$75. What solution do you get? Do alternate optimal solutions exist? Verify that producing 0 pairs of Jordanelle and 14 pairs of Deercrest skis is an alternate optimal solution for this scenario.
- Change the finishing and fabrication constraints to be  $\geq$  instead of  $\leq$  type of constraints. What happens? Why did this occur?
- Change the finishing and fabrication constraints to  $=$  instead of  $\leq$  type of constraints. What happens? Why did this occur?

## How Solver Creates Names in Reports

How you design your spreadsheet model will affect on how *Solver* creates the names used in the output reports. Poor spreadsheet design can make it difficult or confusing to interpret the Answer and Sensitivity reports. Thus, it is important to understand how to do this properly.

*Solver* assigns names to target cells, changing cells, and constraint function cells by concatenating the text in the first cell containing text to the left of the cell with the first cell containing text above it. For example, in the SSC model, the target cell is D22. The first cell containing text to the left of D15 is “Profit Contribution” in A22, and the first cell containing text above D22 is “Total Profit” in cell D21. Concatenating these text strings yields the target cell name “Profit Contribution Total Profit,” which is found in the *Solver* reports. The constraint functions are calculated in cells D15 and D16. Note that their names are “Fabrication Hours Used” and “Finishing Hours Used.” Similarly, the changing cells in B14 and C14 have the names “Quantity Produced Jordanelle” and “Quantity Produced Deercrest.” These names make it easy to interpret the information in the Answer and Sensitivity reports. We encourage you to examine each of the target cells, changing variable cells, and constraint function cells in your models carefully so that names are properly established.

## Difficulties with *Solver*

A poorly scaled model—one in which the parameters of the objective and constraint functions differ by several orders of magnitude (as we have in the transportation example where costs are in tens and supplies/demands in thousands) may cause round-off errors in internal computations or error messages such as “The conditions for Assume Linear Model are not satisfied.” This does not happen often; if it does, you should consult the Frontline Systems’ Web site for additional information. Usually, all you need to do is to keep the solution that *Solver* found and run *Solver* again starting from that solution. Experts often suggest that the values of the coefficients in the objective function and constraints, as well as the right-hand sides, should not differ from each other by a factor of more than 1,000 or 10,000. *Solver* has an option called *Use Automatic Scaling*; it can be accessed by clicking the *Options* button in the *Solver Parameters* dialog, especially if *Solver* gives an error message that linearity is not satisfied.

## APPLICATIONS OF LINEAR OPTIMIZATION

Linear optimization models are the most ubiquitous of optimization models used in organizations today. Applications abound in operations, finance, marketing, engineering, and many other disciplines. Table 13.1 summarizes some common types of generic linear optimization models. We already saw an example of a product mix model with the Sklenka Ski problem. This list represents but a very small sample of the many practical types of linear optimization models that are used.

Building optimization models is more of an art than a science, as there often are several ways of formulating a particular problem. Learning how to build optimization models requires logical thought but can be facilitated by studying examples of different models and observing their characteristics.

The most challenging aspect of model formulation is identifying constraints. Understanding the different types of constraints can help in proper identification and modeling. Constraints generally fall into one of the following categories:

- **Simple Bounds.** Simple bounds constrain the value of a single variable. You can recognize simple bounds in problem statements like “no more than \$10,000 may be invested in stock XYZ” or “we must produce at least 350 units of product Y to

**TABLE 13.1** Generic Examples of Linear Optimization Models

Type of Model	Decision Variables	Objective Function	Typical Constraints
Product mix	Quantities of product to produce and sell	Maximize contribution to profit	Resource limitations (e.g., production time, labor, material); minimum sales requirements; maximum sales potential
Process selection	Quantities of product to make using alternative processes	Minimize cost	Demand requirements; resource limitations
Blending	Quantity of materials to mix to produce one unit of output	Minimize cost	Specifications on acceptable mixture
Portfolio selection	Proportions to invest in different financial instruments	Maximize future return or minimize risk exposure	Limit on available funds; sector requirements or restrictions; proportional relationships on investment mix
Transportation	Amount to ship between sources of supply and destinations	Minimize total transportation cost	Limited availability at sources; required demands met at destinations
Multiperiod production planning	Quantities of product to produce in each of several time periods; amount of inventory to hold between periods	Minimize total production and inventory costs	Limited production rates; material balance equations
Multiperiod financial management	Amounts to invest in short-term instruments	Maximize cash on hand	Cash balance equations; required cash obligations
Production/marketing	Allocation of advertising expenditures; production quantities	Maximize profit	Budget limitation; production limitations; demand requirements

meet customer commitments this month.” The mathematical forms for these examples are:

$$XYZ \leq 10,000$$

$$Y \geq 350$$

- **Limitations.** Limitations usually involve the allocation of scarce resources. Problem statements such as “the amount of material used in production cannot exceed the amount available in inventory,” “minutes used in assembly cannot exceed the available labor hours,” or “the amount shipped from the Austin plant in July cannot exceed the plant’s capacity” are typical of these types of constraints.
- **Requirements.** Requirements involve the specification of minimum levels of performance. Such statements as “enough cash must be available in February to meet financial obligations,” “production must be sufficient to meet promised customer orders,” or “the marketing plan should ensure that at least 400 customers are contacted each month” are some examples.
- **Proportional Relationships.** Proportional relationships are often found in problems involving mixtures or blends of materials or strategies. Examples include “the amount invested in aggressive growth stocks cannot be more than twice the amount invested in equity-income funds,” or “the octane rating of gasoline obtained from mixing different crude blends must be at least 89.”
- **Balance Constraints.** Balance constraints essentially state that “input = output” and ensure that the flow of material or money is accounted for at locations or between time periods. Examples include “production in June plus any available inventory must equal June’s demand plus inventory held to July,” “the total amount shipped to a distribution center from all plants must equal the amount shipped from the distribution center to all customers,” or “the total amount of money invested or saved in March must equal the amount of money available at the end of February.”

Constraints in linear optimization models are generally some combination of constraints from these categories. Problem data or verbal clues in a problem statement often help you identify the appropriate constraint. In some situations, all constraints may not be explicitly stated, but are required for the model to represent the real problem accurately. An example of implicit constraints is nonnegativity of the decision variables.

In the following sections, we present examples of different types of linear optimization applications. Each of these models has different characteristics, and by studying how they are developed, you will improve your ability to model other problems. We encourage you to use the four-step process that we illustrated with the Sklenka Ski problem; however, to conserve space in this book, we will go directly to the mathematical model instead of first conceptualizing the constraints and objective functions in verbal terms. We will also use these models to illustrate specific issues associated with formulation, implementation on spreadsheets, and using *Solver*.

## Process Selection

Process selection models generally involve choosing among different types of processes to produce a good. Make-or-buy decisions are examples of process selection models whereby one must choose whether to make one or more products in-house or subcontract them out to another firm. The following example illustrates these concepts.

Camm Textiles has a mill that produces three types of fabrics on a make-to-order basis. The mill operates on a  $24 \times 7$  basis. The key decision facing the plant manager is on what type of loom to process each fabric during the coming quarter (13 weeks). Two types of looms are used: dobbie and regular. Dobbie looms can be used to make all fabrics and are the only looms that can weave certain fabrics, such as plaids. Demands, variable costs for each fabric, and production rates on the looms are given in Table 13.2. The mill has 15 regular looms and 3 dobbie looms. After weaving, fabrics are sent to the finishing department and then sold. Any fabrics that cannot be woven in the mill because of limited capacity will be purchased from an external supplier, finished at the mill, and sold at the selling price. In addition to determining which looms to process the fabrics, the manager also needs to determine which fabrics to buy externally.

To formulate a linear programming model, define:

$$D_i = \text{number of yards of fabric } i \text{ to produce on dobbie looms}, i = 1, \dots, 3$$

$$R_i = \text{number of yards of fabric } i \text{ to produce on regular looms}, i = 1, \dots, 3$$

$$P_i = \text{number of yards of fabric } i \text{ to purchase from an outside supplier}, i = 1, \dots, 3$$

Note that we are using *subscripted variables* to simplify their definition, rather than defining nine individual variables with unique names. The objective function is to minimize total cost:

$$\text{Min } 0.65D_1 + 0.61D_2 + 0.50D_3 + 0.61R_2 + 0.50R_3 + 0.85P_1 + 0.75P_2 + 0.65P_3$$

**TABLE 13.2** Textile Production Data

Fabric	Demand (Yards)	Dobbie Loom Capacity (Yards/Hour)	Regular Loom Capacity (Yards/Hour)	Mill Cost (\$/Yard)	Outsourcing Cost (\$/Yard)
1	45,000	4.7	0.0	\$0.65	\$0.85
2	76,500	5.2	5.2	\$0.61	\$0.75
3	10,000	4.4	4.4	\$0.50	\$0.65

Constraints to ensure meeting production requirements are:

$$\begin{aligned}D_1 + P_1 &= 45,000 \\D_2 + R_2 + P_2 &= 76,500 \\D_3 + R_3 + P_3 &= 10,000\end{aligned}$$

To specify the constraints on loom capacity, we must convert yards per hour into hours per yard. For example, for fabric 1 on a dobbie loom, 4.7 yards/hour = 0.213 hours/yard. Therefore, the term  $0.213D_1$  represents the total time required to produce  $D_1$  yards of fabric 1 on a dobbie loom. The total capacity for dobbie looms is (24 hours/day) (7 days/week) (13 weeks) (3 looms) = 6,552 hours. Thus, the constraint on available production time on dobbie looms is:

$$0.213D_1 + 0.192D_2 + 0.227D_3 \leq 6,552$$

For regular looms we have:

$$0.192R_2 + 0.227R_3 \leq 32,760$$

Finally, all variables must be nonnegative. The following exercise asks you to implement and solve this model using Excel.

### SKILL-BUILDER EXERCISE 13.2

Implement the Camm Textiles process selection model on a spreadsheet and solve it using *Solver*.

## Blending

Blending problems involve mixing several raw materials that have different characteristics to make a product that meets certain specifications. Dietary planning, gasoline and oil refining, coal and fertilizer production, and the production of many other types of bulk commodities involve blending. We typically see proportional constraints in blending problems.

To illustrate this type of model, consider the BG Seed Company, which specializes in food products for birds and other household pets. In developing a new birdseed mix, company nutritionists have specified that the mixture must contain at least 13% protein and 15% fat, and no more than 14% fiber. The percentages of each of these nutrients in eight types of ingredients that can be used in the mix are given in Table 13.3 along with the wholesale cost per pound. What is the minimum cost mixture that meets the stated nutritional requirements?

In this example, the decisions are the amount of each ingredient to include in a given quantity—for example, one pound—of mix. Define  $X_i$  = number of pounds of ingredient  $i$  to include in one pound of the mix, for  $i = 1, \dots, 8$ . The use of subscripted variables like this simplifies modeling large problems. The objective is to minimize total cost:

$$\text{Minimize } 0.22X_1 + 0.19X_2 + 0.10X_3 + 0.10X_4 + 0.07X_5 + 0.05X_6 + 0.26X_7 + 0.11X_8$$

To ensure that the mix contains the appropriate proportion of ingredients, observe that multiplying the number of pounds of each ingredient by the percentage of nutrient in that ingredient (a dimensionless quantity) specifies the number of pounds of nutrient provided. For example,  $0.169X_1$  represents the number of pounds of protein

**TABLE 13.3** Birdseed Nutrition Data

Ingredient	Protein %	Fat %	Fiber %	Cost/lb
Sunflower seeds	16.9	26	29	\$0.22
White millet	12	4.1	8.3	\$0.19
Kibble corn	8.5	3.8	2.7	\$0.10
Oats	15.4	6.3	2.4	\$0.10
Cracked corn	8.5	3.8	2.7	\$0.07
Wheat	12	1.7	2.3	\$0.05
Safflower	18	17.9	28.8	\$0.26
Canary grass seed	11.9	4	10.9	\$0.11

in sunflower seeds. Therefore, the total number of pounds of protein provided by all ingredients is:

$$0.169X_1 + 0.12X_2 + 0.085X_3 + 0.154X_4 + 0.085X_5 + 0.12X_6 + 0.18X_7 + 0.119X_8$$

Because the total number of pounds of ingredients that are mixed together equals  $X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8$ , the proportion of protein in the mix is:

$$(0.169X_1 + 0.12X_2 + 0.085X_3 + 0.154X_4 + 0.085X_5 + 0.12X_6 + 0.18X_7 + 0.119X_8) / (X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8)$$

This proportion must be at least 0.13. However, we wish to determine the best amount of ingredients to include in *one pound* of mix; therefore, we add the constraint:

$$X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 = 1$$

Now we can substitute 1 for the denominator in the proportion of protein, yielding the constraint:

$$0.169X_1 + 0.12X_2 + 0.085X_3 + 0.154X_4 + 0.085X_5 + 0.12X_6 + 0.18X_7 + 0.119X_8 \geq 0.13$$

This ensures that at least 13% of the mixture will be protein. In a similar fashion, the constraints for the fat and fiber requirements are:

$$0.26X_1 + 0.014X_2 + 0.038X_3 + 0.063X_4 + 0.038X_5 + 0.017X_6 + 0.179X_7 + 0.04X_8 \geq 0.15$$

$$0.29X_1 + 0.083X_2 + 0.027X_3 + 0.024X_4 + 0.027X_5 + 0.023X_6 + 0.288X_7 + 0.109X_8 \leq 0.14$$

Finally, we have nonnegativity constraints:

$$X_i \geq 0, \quad \text{for } i = 1, 2, \dots, 8$$

The complete model is:

$$\text{Minimize } 0.22X_1 + 0.19X_2 + 0.10X_3 + 0.10X_4 + 0.07X_5 + 0.05X_6 + 0.26X_7 + 0.11X_8$$

$$X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 = 1$$

$$0.169X_1 + 0.12X_2 + 0.085X_3 + 0.154X_4 + 0.085X_5 + 0.12X_6 + 0.18X_7 + 0.119X_8 \geq 0.13$$

$$\begin{aligned}
0.26X_1 + 0.041X_2 + 0.038X_3 + 0.063X_4 + 0.038X_5 + 0.017X_6 \\
+ 0.179X_7 + 0.04X_8 \geq 0.15 \\
0.29X_1 + 0.083X_2 + 0.027X_3 + 0.024X_4 + 0.027X_5 + 0.023X_6 \\
+ 0.288X_7 + 0.109X_8 \leq 0.14 \\
X_i \geq 0, \quad \text{for } i = 1, 2, \dots, 8
\end{aligned}$$

We ask you to implement and solve this model in the following exercise.

### SKILL-BUILDER EXERCISE 13.3

Implement the BG Seed Company blending model on a spreadsheet and solve it using *Solver*. Explain the *Solver* results from a practical perspective.

### Portfolio Investment

Many types of financial investment problems are modeled and solved using linear optimization. Such problems have the basic characteristics of blending models. Innis Investments has a client that has acquired \$500,000 from an inheritance. Innis Investments manages six mutual funds:

Fund	Expected Annual Return	Risk Measure
1. Innis Low-Priced Stock Fund	8.13%	10.57
2. Innis Multinational Fund	9.02%	13.22
3. Innis Mid-Cap Stock Fund	7.56%	14.02
4. Innis Mortgage Fund	3.62%	2.39
5. Innis Income Equity Fund	7.79%	9.30
6. Innis Balanced Fund	4.40%	7.61

Innis Investments uses a proprietary algorithm to establish a measure of risk for its funds based on the historical volatility of the investments. The higher the volatility, the greater the risk. Innis Investments recommends that no more than \$200,000 be invested in any individual fund, that at least \$50,000 be invested in each of the multinational and balanced funds, and that at least 40% be invested in the Income Equity and Balanced funds. The client would like to have an average return of at least 5% but would like to minimize risk. What portfolio would achieve this?

Let  $X_1$  through  $X_6$  represent the amount invested in funds 1 through 6, respectively. The total risk would be measured by the weighted standard deviation of the portfolio, where the weights are the proportion of the total investment in any fund ( $X_j/500,000$ ). Thus, the objective function is:

$$\begin{aligned}
\text{Minimize Total Risk} = & (10.57X_1 + 13.22X_2 + 14.02X_3 + 2.39X_4 + 9.30X_5 \\
& + 7.61X_6)/500,000
\end{aligned}$$

The first constraint ensures that \$500,000 is invested:

$$X_1 + X_2 + X_3 + X_4 + X_5 + X_6 = 500,000$$

The next constraint ensures that the weighted return is at least 5%:

$$(8.13X_1 + 9.02X_2 + 7.56X_3 + 3.62X_4 + 7.79X_5 + 4.40X_6)/500,000 \geq 5.00$$

The next constraint ensures that at least 40% be invested in the Income Equity and Balanced funds:

$$X_5 + X_6 \geq 0.4(500,000)$$

The following specify that at least \$50,000 be invested in each of the multinational and balanced funds:

$$X_2 \geq 50,000$$

$$X_6 \geq 50,000$$

Finally, we restrict each investment to a maximum of \$200,000 and include nonnegativity:

$$X_j \leq 200,000 \quad \text{for } j = 1, \dots, 6$$

$$X_j \geq 0 \quad \text{for } j = 1, \dots, 6$$

The following exercise asks you to solve this model.

#### SKILL-BUILDER EXERCISE 13.4

Implement the Innis Investments portfolio investment model on a spreadsheet and solve it using *Solver*.

### Transportation Problem

Many practical models in supply chain optimization stem from a very simple model called the transportation problem. This involves determining how much to ship from a set of sources of supply (factories, warehouses, etc.) to a set of demand locations (warehouses, customers, etc.) at minimum cost. We illustrate this with the following scenario.

General Appliance Corporation (GAC) produces refrigerators at two plants: Marietta, Georgia, and Minneapolis, Minnesota. They ship them to major distribution centers in Cleveland, Baltimore, Chicago, and Phoenix. The Accounting, Production, and Marketing departments have provided the information in Table 13.4, which shows the unit cost of shipping between any plant and distribution center, plant capacities over the next planning period, and distribution center demands. GAC's supply chain manager faces the problem of determining how much to ship between each plant and distribution center to minimize the total transportation cost, not exceed available capacity, and meet customer demand.

To develop a linear optimization model, we first define the decision variables as the amount to ship between each plant and distribution center. In this model, we

**TABLE 13.4** Cost, Capacity, and Demand Data

Plant	Distribution Center				Capacity
	Cleveland	Baltimore	Chicago	Phoenix	
Marietta	\$12.60	\$14.35	\$11.52	\$17.58	1,200
Minneapolis	\$9.75	\$16.26	\$8.11	\$17.92	800
Demand	150	350	500	1,000	

will use *double-subscripted variables* to simplify the formulation. Define  $X_{ij}$  = amount shipped from plant  $i$  to distribution center  $j$ , where  $i = 1$  represents Marietta,  $i = 2$  represents Minneapolis,  $j = 1$  represents Cleveland, and so on. Using the unit cost data in Table 13.4, the total cost of shipping is equal to the unit cost times amount shipped, summed over all combinations of plants and distribution centers. Therefore, the objective function is to minimize total cost:

$$\begin{aligned} \text{Minimize } & 12.60X_{11} + 14.35X_{12} + 11.52X_{13} + 17.58X_{14} + 9.75X_{21} \\ & + 16.26X_{22} + 8.11X_{23} + 17.92X_{24} \end{aligned}$$

Because capacity is limited, the amount shipped from each plant cannot exceed its capacity. The total amount shipped from Marietta, for example, is  $X_{11} + X_{12} + X_{13} + X_{14}$ . Therefore, we have the constraint:

$$X_{11} + X_{12} + X_{13} + X_{14} \leq 1,200$$

Similarly, the capacity limitation at Minneapolis leads to the constraint:

$$X_{21} + X_{22} + X_{23} + X_{24} \leq 800$$

Next, we must ensure that the demand at each distribution center is met. This means that the total amount shipped to any distribution center from both plants must equal the demand. For instance, at Cleveland, we must have:

$$X_{11} + X_{21} = 150$$

For the remaining three distribution centers, the constraints are:

$$X_{12} + X_{22} = 350$$

$$X_{13} + X_{23} = 500$$

$$X_{14} + X_{24} = 1,000$$

Last, we need nonnegativity,  $X_{ij} \geq 0$ , for all  $i$  and  $j$ . The complete model is:

$$\begin{aligned} \text{Minimize } & 12.60X_{11} + 14.35X_{12} + 11.52X_{13} + 17.58X_{14} + 9.75X_{21} \\ & + 16.26X_{22} + 8.11X_{23} + 17.92X_{24} \\ X_{11} + X_{12} + X_{13} + X_{14} & \leq 1,200 \\ X_{21} + X_{22} + X_{23} + X_{24} & \leq 800 \\ X_{11} + X_{21} & = 150 \\ X_{12} + X_{22} & = 350 \\ X_{13} + X_{23} & = 500 \\ X_{14} + X_{24} & = 1,000 \\ X_{ij} & \geq 0, \quad \text{for all } i \text{ and } j \end{aligned}$$

Figure 13.10 shows a spreadsheet implementation for the GAC transportation problem (Excel file *Transportation Model*). In the *Model* section, the decision variables are stored in the plant-distribution center matrix. The objective function is computed in cell A18 as:

$$\begin{aligned} \text{Total cost} = & B6 \times B13 + C6 \times C13 + D6 \times D13 + E6 \times E13 + B7 \times B14 \\ & + C7 \times C14 + D7 \times D14 + E7 \times E14 \end{aligned}$$

A	B	C	D	E	F
1 Transportation Model					
2					
3 Data					
4					
5 Plant	Cleveland	Baltimore	Chicago	Phoenix	Distribution Center
6 Marietta	\$ 12.60	\$ 14.35	\$ 11.52	\$ 17.58	Capacity
7 Minneapolis	\$ 9.75	\$ 16.26	\$ 8.11	\$ 17.92	1200
8 Demand	150	350	500	1000	800
9					
10 Model					
11					
12 Plant	Cleveland	Baltimore	Chicago	Phoenix	Distribution Center
13 Marietta	0	0	0	0	Total shipped
14 Minneapolis	0	0	0	0	0
15 Demand met	0	0	0	0	0
16					
17 Total cost					
18 \$	-				

**FIGURE 13.10** *Transportation Model Spreadsheet*

The SUMPRODUCT function is particularly useful for such large expressions; we could write the total cost as:

$$\text{SUMPRODUCT}(\text{B6:E7}; \text{B13:E14})$$

Note that the SUMPRODUCT function applies to a matrix of values as long as the dimensions are the same.

To ensure that we do not exceed the capacity of any plant, the total shipped from each plant (cells F13:F14) cannot be greater than the plant capacities (cells F6:F7). For example,

$$\begin{aligned}\text{Total shipped from Marietta (cell F13)} &= \text{B13} + \text{C13} + \text{D13} + \text{E13} \\ &= \text{SUM}(\text{B13:E13}) \leq \text{F6}\end{aligned}$$

The constraint for Minneapolis is similar. Can you write it?

To ensure that demands are met, the total shipped to each distribution center (cells B15:E15) must equal or exceed the demands (cells B8:E8). Thus, for Cleveland,

$$\begin{aligned}\text{Total shipped to Cleveland (cell B15)} &= \text{B13} + \text{B14} \\ &= \text{SUM}(\text{B13:B14}) = \text{B8}\end{aligned}$$

As you become more proficient in using spreadsheets, you should consider creating range names for the decision variables and constraint functions. This allows you to locate and manipulate elements of the model more easily. For example, in the transportation model, you might define the range B13:E14 as *Decisions* and the range B6:E7 as *Cost*. The total cost can then be computed easily as SUMPRODUCT(*Decisions*, *Cost*). In this book, however, we will stick with using cell references in all formulas to keep it simple.

*A very important note on formatting the Sensitivity Report:* Depending on how cells in your spreadsheet model are formatted, the Sensitivity Report produced by *Solver* may not reflect the accurate values of reduced costs or shadow prices because an insufficient number

of decimal places may be displayed. For example, Figure 13.11 shows the Sensitivity Report created by *Solver*. Note that the data in columns headed Reduced Cost and Shadow Price are formatted as whole numbers. The correct values are shown in Figure 13.12 (obtained by simply formatting the data to have two decimal places). Thus, we *highly recommend* that after you save the Sensitivity Report to your workbook, you select the reduced cost and shadow price ranges and format them to have at least two or three decimal places.

A	B	C	D	E	F	G	H
5							
6	Adjustable Cells						
7	Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
9	\$B\$13	Marietta Cleveland	0	3	12.6	1E+30	3.19
10	\$C\$13	Marietta Baltimore	350	0	14.35	1.57	1E+30
11	\$D\$13	Marietta Chicago	0	4	11.52	1E+30	3.75
12	\$E\$13	Marietta Phoenix	850	0	17.58	0.34	1.57
13	\$B\$14	Minneapolis Cleveland	150	0	9.75	3.19	1E+30
14	\$C\$14	Minneapolis Baltimore	0	2	16.26	1E+30	1.57
15	\$D\$14	Minneapolis Chicago	500	0	8.11	3.75	1E+30
16	\$E\$14	Minneapolis Phoenix	150	0	17.92	1.57	0.34
17							
18	Constraints						
19	Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
21	\$F\$13	Marietta Total shipped	1200	0	1200	150	0
22	\$F\$14	Minneapolis Total shipped	800	0	800	1E+30	0
23	\$B\$15	Demand met Cleveland	150	10	150	0	150
24	\$C\$15	Demand met Baltimore	350	15	350	0	150
25	\$D\$15	Demand met Chicago	500	8	500	0	500
26	\$E\$15	Demand met Phoenix	1000	18	1000	0	150

**FIGURE 13.11** Original Sensitivity Report for GAC Transportation Model

A	B	C	D	E	F	G	H
5							
6	Adjustable Cells						
7	Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
9	\$B\$13	Marietta Cleveland	0	3.19	12.6	1E+30	3.19
10	\$C\$13	Marietta Baltimore	350	0.00	14.35	1.57	1E+30
11	\$D\$13	Marietta Chicago	0	3.75	11.52	1E+30	3.75
12	\$E\$13	Marietta Phoenix	850	0.00	17.58	0.34	1.57
13	\$B\$14	Minneapolis Cleveland	150	0.00	9.75	3.19	1E+30
14	\$C\$14	Minneapolis Baltimore	0	1.57	16.26	1E+30	1.57
15	\$D\$14	Minneapolis Chicago	500	0.00	8.11	3.75	1E+30
16	\$E\$14	Minneapolis Phoenix	150	0.00	17.92	1.57	0.34
17							
18	Constraints						
19	Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
21	\$F\$13	Marietta Total shipped	1200	-0.34	1200	150	0
22	\$F\$14	Minneapolis Total shipped	800	0.00	800	1E+30	0
23	\$B\$15	Demand met Cleveland	150	9.75	150	0	150
24	\$C\$15	Demand met Baltimore	350	14.69	350	0	150
25	\$D\$15	Demand met Chicago	500	8.11	500	0	500
26	\$E\$15	Demand met Phoenix	1000	17.92	1000	0	150

**FIGURE 13.12** Accurate Sensitivity Report for GAC Transportation Model

## Interpreting Reduced Costs

The transportation model is a good example to use to discuss the interpretation of reduced costs. First, note that the reduced costs are zero for all variables that are positive in the solution. Now examine the reduced cost, 3.19, associated with shipping from Marietta to Cleveland. A question to ask is “why does the optimal solution ship nothing between these cities?” The answer is simple: It is not economical to do so! In other words, it costs too much to ship from Marietta to Cleveland; the demand can be met less expensively by shipping from Minneapolis. The next logical question to ask is “what would the unit shipping cost have to be to make it attractive to ship from Marietta instead of Minneapolis?” The answer is given by the reduced cost. If the unit cost is reduced by at least \$3.19, then the optimal solution will change and would include a positive value for the Marietta-Cleveland variable. Again, this is only true if all other data are held constant.

## Multiperiod Production Planning

Many linear optimization problems involve planning over multiple time periods. Kristin’s Kreations is a home-based company that makes hand-painted jewelry boxes for teenage girls. Forecasts of sales for the next year are 150 in the autumn, 400 in the winter, and 50 in the spring. Plain jewelry boxes are purchased from a supplier for \$20. The cost of capital is estimated to be 24% per year (or 6% per quarter); thus, the holding cost per item is  $0.06(\$20) = \$1.20$  per quarter. Kristin hires art students part-time to craft her designs during the autumn, and they earn \$5.50 per hour. Because of the high demand for part-time help during the winter holiday season, labor rates are higher in the winter, and workers earn \$7.00 per hour. In the spring, labor is more difficult to keep, and the owner must pay \$6.25 per hour to retain qualified help. Each jewelry box takes 2 hours to complete. How should production be planned over the three quarters to minimize the combined production and inventory holding costs?

The principal decision variables are the number of jewelry boxes to produce during each of the three quarters. While it might seem obvious to simply produce to the anticipated level of sales, it may be advantageous to produce more during some quarter and carry the items in inventory, thereby letting lower labor rates offset the carrying costs. Therefore, we must also define decision variables for the number of units to hold in inventory at the end of each quarter. The decision variables are:

$P_A$  = amount to produce in autumn

$P_W$  = amount to produce in winter

$P_S$  = amount to produce in spring

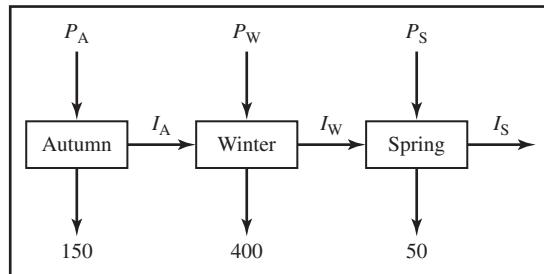
$I_A$  = inventory held at the end of autumn

$I_W$  = inventory held at the end of winter

$I_S$  = inventory held at the end of spring

The production cost per unit is computed by multiplying the labor rate by the number of hours required to produce one. Thus, the unit cost in the autumn is  $(\$5.50)(2) = \$11.00$ ; in the winter,  $(\$7.00)(2) = \$14.00$ ; and in the spring,  $(\$6.25)(2) = \$12.50$ . The objective function is to minimize the total cost of production and inventory. (Because the cost of the boxes themselves is constant, it is not relevant to the problem we are addressing.) The objective function is therefore:

$$\text{Minimize } 11P_A + 14P_W + 12.50P_S + 1.20I_A + 1.20I_W + 1.20I_S$$



**FIGURE 13.13** Material Balance Constraint Structure

The only explicit constraint is that demand must be satisfied. Note that both the production in a quarter as well as the inventory held from the *previous* time quarter can be used to satisfy demand. In addition, any amount in excess of the demand is held to the next quarter. Therefore, the constraints take the form of *inventory balance equations* that essentially say, “what is available in any time period must be accounted for somewhere.” More formally,

$$\begin{aligned} \text{Production + Inventory from the previous quarter} &= \text{Demand} \\ &\quad + \text{Inventory held to the next quarter} \end{aligned}$$

This can be represented visually using the diagram in Figure 13.13. For each quarter, the sum of the variables coming in must equal the sum of the variables going out. Drawing such a figure is very useful for any type of multiple time period planning model. This results in the constraint set:

$$\begin{aligned} P_A + 0 &= 150 + I_A \\ P_W + I_A &= 400 + I_W \\ P_S + I_W &= 50 + I_S \end{aligned}$$

Moving all variables to the left side results in the model:

$$\text{Minimize } 11P_A + 14P_W + 12.50P_S + 1.20I_A + 1.20I_W + 1.20I_S$$

Subject to

$$\begin{aligned} P_A - I_A &= 150 \\ P_W + I_A - I_W &= 400 \\ P_S + I_W - I_S &= 50 \\ P_i &\geq 0, \quad \text{for all } i \\ I_j &\geq 0, \quad \text{for all } j \end{aligned}$$

As we have noted, developing models is more of an art than a science; consequently, there is often more than one way to model a particular problem. Using the ideas presented in this example, we may construct an alternative model involving only the production variables. We simply have to make sure that demand is satisfied. We can do this by ensuring that the cumulative production in each quarter is at least as great as the cumulative demand. This is expressed by the following constraints:

$$\begin{aligned} P_A &\geq 150 \\ P_A + P_W &\geq 550 \end{aligned}$$

$$P_A + P_W + P_S \geq 600$$

$$P_A, P_W, P_S \geq 0$$

The differences between the left- and right-hand sides of these constraints are the ending inventories for each period (and we need to keep track of these amounts because inventory has a cost associated with it). Thus, we use the following objective function:

$$\begin{aligned} \text{Minimize } & 11P_A + 14P_W + 12.50P_S + 1.20(P_A - 150) + 1.20(P_A + P_W - 550) \\ & + 1.20(P_A + P_W + P_S - 600) \end{aligned}$$

Of course, this function can be simplified algebraically by combining like terms. Although these two models look very different, they are mathematically equivalent and will produce the same solution. The following exercise asks you to verify this.

### SKILL-BUILDER EXERCISE 13.5

Implement both the original and alternative models developed for Kristin's Kreations multiperiod planning model to verify that they result in the same optimal solution. In the alternative model, how can you determine the values of the inventory variables?

## Multiperiod Financial Planning

Financial planning often occurs over an extended time horizon and can be formulated as multiperiod optimization models. For example, a financial manager at D.A. Branch & Sons must ensure that funds are available to pay company expenditures in the future, but would also like to maximize investment income. Three short-term investment options are available over the next six months: *A*, a one-month CD that pays 0.25%, available each month; *B*, a three-month CD that pays 1.00%, available at the beginning of the first four months; and *C*, a six-month CD that pays 2.3%, available in the first month. The net expenditures for the next six months are forecast as \$50,000, (\$12,000), \$23,000, (\$20,000), \$41,000, (\$13,000). Amounts in parentheses indicate a net inflow of cash. The company must maintain a cash balance of at least \$10,000 at the end of each month. The company currently has \$200,000 in cash.

At the beginning of each month, the manager must decide how much to invest in each alternative that may be available. Define:

$A_i$  = amount (\$) to invest in a one-month CD at the start of month *i*

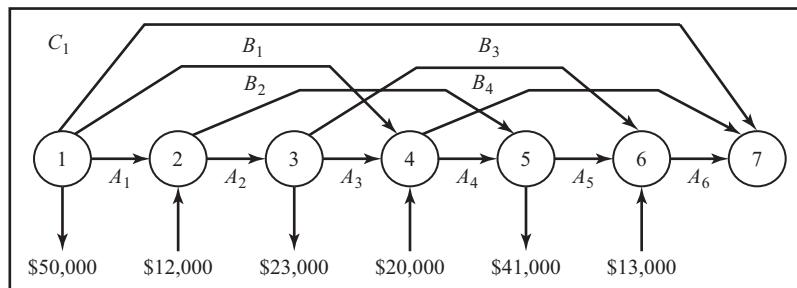
$B_i$  = amount (\$) to invest in a three-month CD at the start of month *i*

$C_i$  = amount (\$) to invest in a six-month CD at the start of month *i*

Because the time horizons on these alternatives vary, it is helpful to draw a picture to represent the investments and returns for each year as shown in Figure 13.14. Each circle represents the beginning of a month. Arrows represent the investments and cash flows. For example, investing in a three-month CD at the start of month 1 ( $B_1$ ) matures at the beginning of month 4. It is reasonable to assume that all funds available would be invested.

From Figure 13.14, we see that investments  $A_6$ ,  $B_4$ , and  $C_1$  will mature at the end of month 6; that is, at the beginning of month 7. To maximize the amount of cash on hand at the end of the planning period, we have the objective function:

$$\text{Maximize } 1.0025A_6 + 1.00B_4 + 1.023C_1$$



**FIGURE 13.14** Cash Balance Constraint Structure

The only constraints necessary are minimum cash balance equations. For each month, the net cash available, which is equal to the cash in less cash out, must be at least \$100,000. These follow directly from Figure 13.14. The complete model is:

$$\text{Maximize } 1.0025A_6 + 1.00B_4 + 1.023C_1$$

Subject to

$$200,000 - (A_1 + B_1 + C_1 + 50,000) \geq 10,000 \quad (\text{Month 1})$$

$$1.0025A_1 + 12,000 - (A_2 + B_2) \geq 10,000 \quad (\text{Month 2})$$

$$1.0025A_2 - (A_3 + B_3 + 23,000) \geq 10,000 \quad (\text{Month 3})$$

$$1.0025A_3 + 1.00B_1 + 20,000 - (A_4 + B_4) \geq 10,000 \quad (\text{Month 4})$$

$$1.0025A_4 + 1.00B_2 - (A_5 + 41,000) \geq 10,000 \quad (\text{Month 5})$$

$$1.0025A_5 + 1.00B_3 + 13,000 - A_6 \geq 10,000 \quad (\text{Month 6})$$

$$A_i, B_i, C_i \geq 0, \quad \text{for all } i$$

### SKILL-BUILDER EXERCISE 13.6

Implement the D.A. Branch & Sons multiperiod investment model on a spreadsheet and solve it using *Solver*.

### A Model with Bounded Variables

Product mix problems involve determining the best mix of products to produce with limited resources and other requirements (the Sklenka Ski Company problem is a simple example). Suppose that J&M Manufacturing makes four models of gas grills, *A*, *B*, *C*, and *D*. Each grill must flow through five departments, stamping, painting, assembly, inspection, and packaging. Table 13.5 shows the relevant data. Production rates are shown in units/hour. (Grill *A* uses imported parts and does not require painting.)

J&M wants to determine how many grills to make to maximize monthly profit. To formulate this as a linear optimization model, let:

$A, B, C$ , and  $D$  = number of units of model *A*, *B*, *C*, and *D* to produce, respectively

The objective function is to maximize the total net profit:

$$\begin{aligned} \text{Maximize } & (250 - 210)A + (300 - 240)B + (400 - 300)C + (650 - 520)D \\ & = 40A + 60B + 100C + 130D \end{aligned}$$

**TABLE 13.5** J&M Manufacturing Data

Grill Model	Selling Price/Unit	Variable Cost/Unit	Minimum Monthly Sales Requirements	Maximum Monthly Sales Potential	
A	\$250	\$210	0	4,000	
B	\$300	\$240	0	3,000	
C	\$400	\$300	500	2,000	
D	\$650	\$520	500	1,000	
Department	A	B	C	D	Hours Available
Stamping	40	30	10	10	320
Painting		20	10	10	320
Assembly	25	15	15	12	320
Inspection	20	20	25	15	320
Packaging	50	40	40	30	320

The constraints include limitations on the amount of production hours available in each department, the minimum sales requirements, and maximum sales potential limits. Here is an example of where you must carefully look at the dimensions of the data. The production rates are given in units/hour, so if you multiply these values by the number of units produced, you will have an expression that makes no sense. Therefore, you must divide the decision variables by units per hour or equivalently, convert these data to hours/unit, and then multiply by the decision variables:

$$\begin{aligned}
 A/40 + B/30 + C/10 + D/10 &\leq 320 && (\text{Stamping}) \\
 B/20 + C/10 + D/10 &\leq 320 && (\text{Painting}) \\
 A/25 + B/15 + C/15 + D/12 &\leq 320 && (\text{Assembly}) \\
 A/20 + B/20 + C/25 + D/15 &\leq 320 && (\text{Inspection}) \\
 A/50 + B/40 + C/40 + D/30 &\leq 320 && (\text{Packaging})
 \end{aligned}$$

The sales constraints are simple upper and lower bounds on the variables:

$$\begin{aligned}
 A &\geq 0 \\
 B &\geq 0 \\
 C &\geq 500 \\
 D &\geq 500 \\
 A &\leq 4,000 \\
 B &\leq 3,000 \\
 C &\leq 2,000 \\
 D &\leq 1,000
 \end{aligned}$$

Nonnegativity constraints are implied by the lower bounds on the variables and, therefore, do not need to be explicitly stated. Figure 13.15 shows a spreadsheet implementation (Excel file *J&M Manufacturing*) with the optimal solution and the *Solver* model used to find it.

*Solver* handles simple lower bounds (e.g.,  $C \geq 500$ ) and upper bounds (e.g.,  $D \leq 1,000$ ) quite differently from ordinary constraints in the Sensitivity Report. To see this, look at the Answer and Sensitivity reports for the J&M Manufacturing model in Figures 13.16 and 13.17 (the spreadsheet model is available on the Companion Website). In the Answer Report, all constraints are listed along with their status. For example,

	A	B	C	D	E	F	G	H	I	J
1	J&M Manufacturing									
2										
3	Data									
4	Grill model	Selling price	Variable cost	Min Sales	Max Sales					
5	A	\$ 250.00	\$ 210.00	0	4000					
6	B	\$ 300.00	\$ 240.00	0	3000					
7	C	\$ 400.00	\$ 300.00	500	2000					
8	D	\$ 650.00	\$ 520.00	500	1000					
9										
10	Production rates (hours/unit)	A	B	C	D	Hours Available				
11	Stamping		40	30	10	10	320			
12	Painting			20	10	10	320			
13	Assembly		25	15	15	12	320			
14	Inspection		20	20	25	15	320			
15	Packaging		50	40	40	30	320			
16										
17	Model									
18	Department	A	B	C	D	Hours Used				
19	Stamping	96.429	0.000	123.571	100.000	320.000				
20	Painting		0.000	123.571	100.000	223.571				
21	Assembly	154.286	0.000	82.381	83.333	320.000				
22	Inspection	192.857	0.000	49.429	66.667	308.952				
23	Packaging	77.143	0.000	30.893	33.333	141.369				
24										
25	Number produced	3857.14	0.00	1235.71	1000.00					
26	Net profit/unit	\$ 40.00	\$ 60.00	\$ 100.00	\$ 130.00	Total Profit				
27	Profit contribution	\$ 154,285.71	\$ -	\$ 123,571.43	\$ 130,000.00	\$ 407,857.14				

FIGURE 13.15 Spreadsheet Implementation and Solver Model for J&M Manufacturing

	A	B	C	D	E	F	G
11							
12	Objective Cell (Max)						
13	Cell	Name	Original Value	Final Value			
14	\$F\$27	Profit contribution Total Profit	0	407857.1429			
15							
16							
17	Decision Variable Cells						
18	Cell	Name	Original Value	Final Value			
19	\$B\$25	Number produced A	0	3857.142857			
20	\$C\$25	Number produced B	0	0			
21	\$D\$25	Number produced C	0	1235.714286			
22	\$E\$25	Number produced D	0	1000			
23							
24	Constraints						
25	Cell	Name	Cell Value	Formula	Status	Slack	
26	\$F\$19	Stamping Hours Used	320.000	\$F\$19<=\$F\$11	Binding	0	
27	\$F\$20	Painting Hours Used	223.571	\$F\$20<=\$F\$12	Not Binding	96.42857143	
28	\$F\$21	Assembly Hours Used	320.000	\$F\$21<=\$F\$13	Binding	0	
29	\$F\$22	Inspection Hours Used	308.952	\$F\$22<=\$F\$14	Not Binding	11.04761905	
30	\$F\$23	Packaging Hours Used	141.369	\$F\$23<=\$F\$15	Not Binding	178.6309524	
31	\$B\$25	Number produced A	3857.142857	\$B\$25<=\$E\$5	Not Binding	142.8571429	
32	\$C\$25	Number produced B	0	\$C\$25<=\$E\$6	Not Binding	3000	
33	\$D\$25	Number produced C	1235.714286	\$D\$25<=\$E\$7	Not Binding	764.2857143	
34	\$E\$25	Number produced D	1000	\$E\$25<=\$E\$8	Binding	0	
35	\$B\$25	Number produced A	3857.142857	\$B\$25>=\$D\$5	Not Binding	3857.142857	
36	\$C\$25	Number produced B	0	\$C\$25>=\$D\$6	Binding	0	
37	\$D\$25	Number produced C	1235.714286	\$D\$25>=\$D\$7	Not Binding	735.7142857	
38	\$E\$25	Number produced D	1000	\$E\$25>=\$D\$8	Not Binding	500	

FIGURE 13.16 J&M Manufacturing Solver Answer Report

	A	B	C	D	E	F	G	H
4								
5		Objective Cell (Max)						
6	Cell	Name	Final Value					
7	\$F\$27	Profit contribution Total Profit	407857.1429					
8								
9	Decision Variable Cells							
10	Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease	
11	\$B\$25	Number produced A	3857.142857	0	40	20.00000004	1.000000042	
12	\$C\$25	Number produced B	0	-1.904761905	60	1.904761905	1E+30	
13	\$D\$25	Number produced C	1235.714286	0	100	13.33333389	33.33333339	
14	\$E\$25	Number produced D	1000	19.28571429	130	1E+30	19.28571429	
15								
16								
17	Constraints							
18	Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease	
19	\$F\$19	Stamping Hours Used	320.000	571.429	320	44.58333333	5	
20	\$F\$20	Painting Hours Used	223.571	0.000	320	1E+30	96.42857143	
21	\$F\$21	Assembly Hours Used	320.000	642.857	320	3.333333333	71.33333333	
22	\$F\$22	Inspection Hours Used	308.952	0.000	320	1E+30	11.04761905	
23	\$F\$23	Packaging Hours Used	141.369	0.000	320	1E+30	178.6309524	
24								

**FIGURE 13.17** J&M Manufacturing Solver Sensitivity Report

we see that the upper bound on  $D$  and lower bound on  $B$  are binding. However, none of the bound constraints appear in the Constraints section of the Sensitivity Report.

In *Solver*, lower and upper bounds are treated in a manner similar to nonnegativity constraints, which also do not appear explicitly as constraints in the model. *Solver* does this to increase the efficiency of the solution procedure used; for large models this can represent significant savings in computer-processing time. However, it makes it more difficult to interpret the sensitivity information, because we no longer have the shadow prices and allowable increases and decreases associated with these constraints. Actually, this isn't quite true; the shadow prices are there, but in a different form.

First, let us interpret the reduced costs. Recall that in an ordinary model with only nonnegativity constraints and no other simple bounds, the reduced cost tells how much the objective coefficient needs to be reduced in order for a variable to become positive in an optimal solution. For product  $B$ , we have the lower bound constraint  $B \geq 0$ . Note that the optimal solution specifies that we produce only the minimum amount required. Why? It is simply not economical to produce more because the profit contribution of  $B$  is too low relative to the other products. How much more would the profit on  $B$  have to be in order for it to be economical to produce anything other than the minimum amount required? As we saw earlier for the transportation problem, the answer is given by the reduced cost. The unit profit on  $B$  would have to be reduced by at least -\$1.905 (that is, increased by at least +\$1.905). If a nonzero lower bound constraint is binding, the interpretation is similar; the reduced cost is the amount the unit profit would have to be reduced in order to produce more than the minimum amount.

For product  $D$ , the reduced cost is \$19.29. Note that  $D$  is at its upper bound, 1,000. We want to produce as much of  $B$  as possible because it generates a large profit. How much would the unit profit have to be lowered before it is no longer economical to produce the maximum amount? Again, the answer is the reduced cost.

Now, let's ask these questions in a different way. For product  $B$ , what would the effect be of increasing the right-hand side value of the bound constraint,  $B \geq 0$  by one unit? If we increase the right-hand side of a lower bound constraint by 1, we are essentially forcing the solution to produce more than the minimum requirement. How would

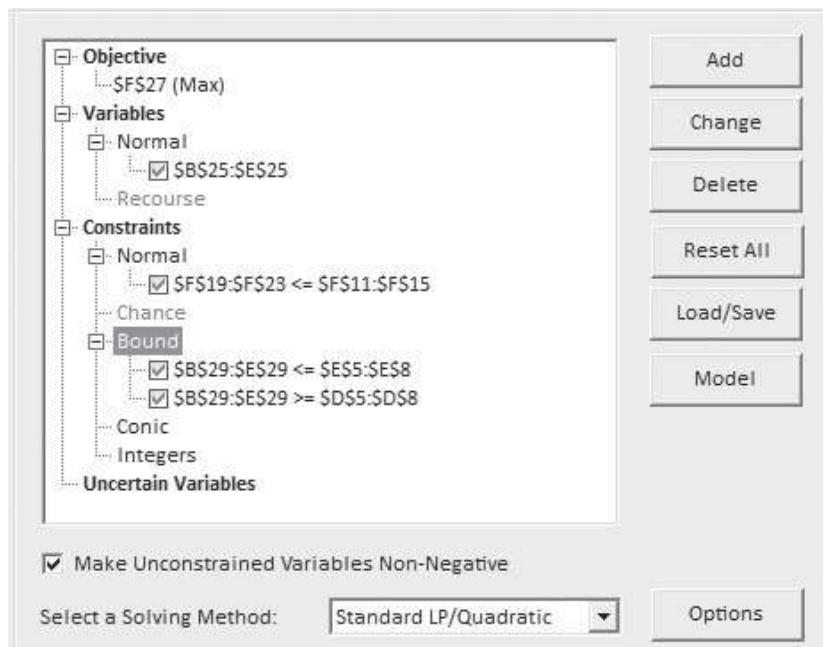
	A	B	C	D	E	F
24						
25 Number produced		0	0	0	0	
26 Net profit/unit	=B5-C5	=B6-C6	=B7-C7	=B8-C8	Total Profit	
27 Profit contribution	=B25*B26	=C25*C26	=D25*D26	=E25*E26	=SUM(B27:E27)	
28						
29 Auxiliary variable	=B25	=C25	=D25	=E25		

**FIGURE 13.18** Auxiliary Variable Cells in J&M Manufacturing Model

the objective function change if we do this? It would have to decrease because we would lose money by producing an extra unit of a non-profitable product. How much? The answer again is the reduced cost! Producing an additional unit of product *B* will result in a profit reduction of \$1.905. Similarly, increasing the right-hand side of the constraint  $D \leq 1,000$  by one will increase the profit by \$19.29. Thus, *the reduced cost associated with a bounded variable is the same as the shadow price of the bound constraint*. However, we no longer have the allowable range over which we can change the constraint values. (*Important:* The Allowable Increase and Allowable Decrease values in the Sensitivity Report refer to the objective coefficients, not the reduced costs.)

Interpreting reduced costs as shadow prices for bounded variables can be a bit confusing. Fortunately, there is a neat little trick that you can use to eliminate this issue. In your spreadsheet model, define a new set of cells for any decision variables that have upper or lower bound constraints by referencing (not copying) the original changing cells. This is shown in row 29 of Figure 13.18.

In the *Solver* model, use these auxiliary variable cells—not the changing variable cells as defined—to define the bound constraints. Thus, the *Solver* constraints would be as shown in Figure 13.19. The Sensitivity Report for this model is shown in Figure 13.20. We now see that the *Constraints* section has rows corresponding to the bound constraints and that the shadow prices are the same as the reduced costs in the previous sensitivity report. Moreover, we now know the allowable increases and decreases for each shadow



**FIGURE 13.19** J&M Manufacturing *Solver* Model with Auxiliary Variables

A	B	C	D	E	F	G	H
4							
5	Objective Cell (Max)						
6	Cell	Name	Final Value				
7	\$F\$27	Profit contribution Total Profit	407857.1429				
8							
9	Decision Variable Cells						
10			Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
11	Cell	Name					
12	\$B\$25	Number produced A	3857.1	0.0	40	20.00000004	1.000000053
13	\$C\$25	Number produced B	0.0	-1.9	60	1.904761905	1E+30
14	\$D\$25	Number produced C	1235.7	0.0	100	13.33333403	33.33333339
15	\$E\$25	Number produced D	1000.0	0.0	130	1E+30	19.28571439
16							
17	Constraints						
18			Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
19	Cell	Name					
20	\$B\$29	Auxiliary variable A	3857.14	0.00	4000	1E+30	142.8571429
21	\$C\$29	Auxiliary variable B	0.00	0.00	3000	1E+30	3000
22	\$D\$29	Auxiliary variable C	1235.71	0.00	2000	1E+30	764.2857143
23	\$E\$29	Auxiliary variable D	1000.00	19.29	1000	895.6521739	200
24	\$B\$29	Auxiliary variable A	3857.14	0.00	0	3857.142857	1E+30
25	\$C\$29	Auxiliary variable B	0.00	0.00	0	0	1E+30
26	\$D\$29	Auxiliary variable C	1235.71	0.00	500	735.7142857	1E+30
27	\$E\$29	Auxiliary variable D	1000.00	0.00	500	500	1E+30
28	\$F\$19	Stamping Hours Used	320.000	571.429	320	44.58333333	5
29	\$F\$20	Painting Hours Used	223.571	0.000	320	1E+30	96.42857143
30	\$F\$21	Assembly Hours Used	320.000	642.857	320	3.333333333	71.33333333
31	\$F\$22	Inspection Hours Used	308.952	0.000	320	1E+30	11.04761905
32	\$F\$23	Packaging Hours Used	141.369	0.000	320	1E+30	178.6309524

**FIGURE 13.20** J&M Manufacturing Sensitivity Report with Auxiliary Variables

price, which we did not have in the previous sensitivity report. Thus, we recommend that you use this approach unless solution efficiency is an important issue.

### A Production/Marketing Allocation Model

Many problems involve allocation of marketing effort, such as advertising dollars. The following is an example of combining elements of a product mix model with marketing budget allocation decisions based on demand elasticity.<sup>1</sup> This example also illustrates some important issues of properly interpreting sensitivity results and the influence that modeling approaches can have.

A small winery, Walker Wines, buys grapes from local growers and blends the pressings to make two types of wine: Shiraz and merlot. It costs \$1.60 to purchase the grapes needed to make a bottle of Shiraz, and \$1.40 to purchase the grapes needed to make a bottle of merlot. The contract requires that they provide at least 40% but not more than 70% Shiraz. Based on market research related to it, it is estimated that the base demand for Shiraz is 1,000 bottles, but increases by 5 bottles for each \$1 spent on advertising while the base demand for merlot is 2,000 bottles and increases by 8 bottles for each \$1 spent on advertising. Production should not exceed demand. Shiraz sells to retail stores for \$6.25 per bottle while merlot is sold for \$5.25 per bottle. Walker Wines has \$50,000 available to purchase grapes and advertise its products, with an objective of maximizing profit contribution.

<sup>1</sup> Adapted from an example in Roger D. Eck, *Operations Research for Business* (Belmont, CA: Wadsworth, 1976), 129–131.

To formulate this model, let:

$S$  = number of bottles of Shiraz produced

$M$  = number of bottles of merlot produced

$A_s$  = dollar amount spent on advertising Shiraz

$A_m$  = dollar amount spent on advertising merlot

The objective is to maximize profit (revenue minus costs) =  $(\$6.25S + \$5.25M) - (\$1.60S + \$1.40M + A_s + A_m) = 4.65S + 4.85M - A_s - A_m$

Constraints are defined as follows:

1. Budget cannot be exceeded:

$$\$1.60S + \$1.40M + A_s + A_m \leq \$50,000$$

2. Contractual requirements must be met:

$$0.4 \leq S/(S + M) \leq 0.7$$

or, expressed in linear form:

$$0.6S - 0.4M \geq 0 \text{ and } 0.3S - 0.7M \leq 0$$

3. Production must not exceed demand:

$$S \leq 1,000 + 5A_s$$

$$M \leq 2,000 + 8A_m$$

4. Nonnegativity

Figure 13.21 shows a spreadsheet implementation of this model (Excel file *Walker Wines*) along with the *Solver* solution.

Figure 13.22 shows the *Solver* Sensitivity Report. A variety of practical questions can be posed around the sensitivity report. For example, suppose that the accountant noticed a small error in computing the profit contribution for Shiraz. The cost of Shiraz grapes should have been \$1.65 instead of \$1.60. How will this affect the solution?

In the model formulation, you can see that a \$0.05 increase in cost results in a drop in the unit profit of Shiraz from \$4.65 to \$4.60. In the Sensitivity Report, the change in the profit coefficient is within the allowable decrease of 0.05328, thus concluding that no change in the optimal solution will result. However, this is *not* the correct interpretation! If the model is re-solved using the new cost parameter, the solution changes dramatically as shown in Figure 13.23.

Why did this happen? One crucial assumption in interpreting sensitivity analysis information for changes in model parameters is that *all other model parameters are held constant*. It is easy to fall into a trap of ignoring this assumption and blindly crunching through the numbers. This is particularly true when using spreadsheet models. In this case, the unit cost is also reflected in the binding budget constraint. When we change the cost parameter, the constraint also changes. This violates the assumption. The change causes the budget constraint to become infeasible, and the solution must be adjusted to maintain feasibility.

This example points out the importance of fully understanding the mathematical model when analyzing sensitivity information. One suggestion to ensure that sensitivity analysis information is interpreted properly in spreadsheet models is to use Excel's formula auditing capability. If you select the cost of Shiraz (cell C5) and apply the "Trace Dependents" command from the *Formula Auditing* menu, you will see that the unit cost influences both the unit profit (cell C20) and the budget constraint function (cell C29).

A	B	C	D	E	F
<b>1 Walker Wines Product Mix</b>					
2					
<b>3 Data</b>					
		Shiraz	Merlot		
5	Cost/bottle	\$ 1.60	\$ 1.40		
6	Price/bottle	\$ 6.25	\$ 5.25		
7					
8	Base demand	1,000.00	2,000.00		
9	Increase/\$1 Adv.	5	8		
10	Min. percent requirement	40%			
11	Max. percent limitation	70%			
12					
13	Total Budget	\$ 50,000.00			
14					
15	<b>Model</b>				
16					
17	Total profit				
18	\$ 124,775.84				
19		Shiraz	Merlot	Total	
20	Unit profit	\$ 4.65	\$ 3.85		
21	Advertising dollars	\$ 3,912.37	\$ 851.53	\$ 4,763.90	
22	Demand	20,561.86	8,812.23	29,374.09	
23	Quantity produced	20,561.86	8,812.23	29,374.09	
24					
25	Min. percent requirement	8812.227074	>=	0	
26	Max. percent limitation	0	<=	0	
27					
28			Used	Unused	
29	Budget	\$ 36,811.35	\$ 13,188.65	\$ 50,000.00	\$ -

A	B	C	D	E	F
<b>1 Walker Wines Product Mix</b>					
2					
<b>3 Data</b>					
		Shiraz	Merlot		
5	Cost/bottle	1.6	1.4		
6	Price/bottle	6.25	5.25		
7					
8	Base demand	1000	2000		
9	Increase/\$1 Adv.	5	8		
10	Min. percent requirement	0.4			
11	Max. percent limitation	0.7			
12					
13	Total Budget	50000			
14					
15	<b>Model</b>				
16					
17	Total profit				
18	=C20*C23)+(D20*D23)-C21-D21				
19		Shiraz	Merlot	Total	
20	Unit profit	=C6-C5	=D6-D5		
21	Advertising dollars	0	0	=SUM(C21:D21)	
22	Demand	=C8+(C9*C21)	=D8+(D9*D21)	=SUM(C22:D22)	
23	Quantity produced	0	0	=SUM(C23:D23)	
24					
25	Min. percent requirement	=C10*C23-C10*D23	>=	0	
26	Max. percent limitation	=C11*C23-C11*D23	<=	0	
27					
28			Used	Unused	
29	Budget	=C21+(C23*C5)	=D21+(D23*D5)	=SUM(C29:D29)	=C13-E29

**FIGURE 13.21** Walker Wines Spreadsheet Model

	A	B	C	D	E	F	G	H
4								
5		Objective Cell (Max)						
6	Cell	Name	Final Value					
7	\$B\$18	Total profit	124775.837					
8								
9	Decision Variable Cells							
10	Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease	
11	\$C\$21	Advertising dollars Shiraz	\$ 3,912.37	\$ -	-1	3.771791052	0.266394356	
12	\$D\$21	Advertising dollars Merlot	\$ 851.53	\$ -	-1	0.36111235	112.8666705	
13	\$C\$23	Quantity produced Shiraz	20,561.86	0.00	4.65	1E+30	0.053278871	
14	\$D\$23	Quantity produced Merlot	8,812.23	0.00	3.85	0.045139044	14.10833381	
15								
16								
17	Constraints							
18	Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease	
19	\$C\$23	Quantity produced Shiraz	20,561.86	0.69	1000	1E+30	252250	
20	\$C\$25	Min. percent requirement Shiraz	8812.227074	0	0	8812.227074	1E+30	
21	\$C\$26	Max. percent limitation Shiraz	0	0.047307132	0	1E+30	9256.880734	
22	\$D\$23	Quantity produced Merlot	8,812.23	0.43	2000	1E+30	403600	
23	\$E\$29	Budget Used	\$ 50,000.00	\$ 2.46	50000	1E+30	50450	
24								

FIGURE 13.22 Walker Wines Solver Sensitivity Report

	A	B	C	D	E	F
1	Walker Wines Product Mix					
2						
3	Data					
4			Shiraz	Merlot		
5	Cost/bottle		\$ 1.65	\$ 1.40		
6	Price/bottle		\$ 6.25	\$ 5.25		
7						
8	Base demand		1,000.00	2,000.00		
9	Increase/\$1 Adv.		5	8		
10	Min. percent requirement		40%			
11	Max. percent limitation		70%			
12						
13	Total Budget		\$ 50,000.00			
14						
15	Model					
16						
17	Total profit					
18		\$ 122,231.12				
19			Shiraz	Merlot	Total	
20	Unit profit		\$ 4.60	\$ 3.85		
21	Advertising dollars		\$ 2,238.67	\$ 2,036.25	\$ 4,274.92	
22	Demand		12,193.35	18,290.03	30,483.38	
23	Quantity produced		12,193.35	18,290.03	30,483.38	
24						
25	Min. percent requirement		0	>=	0	
26	Max. percent limitation		-9145.01511	<=	0	
27						
28				Used	Unused	
29	Budget		\$ 22,357.70	\$ 27,642.30	\$ 50,000.00	\$ -

FIGURE 13.23 Walker Wines Solver Solution After Cost Increase

## HOW SOLVER WORKS

Solver uses a mathematical algorithm called the “simplex method.” This was developed in 1947 by the late Dr. George Dantzig. The simplex method characterizes feasible solutions algebraically by solving systems of linear equations (obtained by adding slack variables to the constraints as we described earlier in this chapter). It moves systematically from one solution to another to improve the objective function until an optimal solution is found (or until the problem is deemed infeasible or unbounded). Because of the linearity of the constraints and objective function, the simplex method is guaranteed to find an optimal solution if one exists, and usually does so quickly and efficiently.

To gain some intuition into the logic of *Solver*, consider the following example. Crebo Manufacturing produces four types of structural support fittings—plugs, rails, rivets, and clips—that are machined on two CNC machining centers. The machining centers have a capacity of 280,000 minutes per year. The gross margin per unit and machining requirements are provided below:

Product	Plugs	Rails	Rivets	Clips
Gross margin/unit	\$0.30	\$1.30	\$0.75	\$1.20
Minutes/unit	1	2.5	1.5	2

How many of each product should be made to maximize gross profit margin?

To formulate this as a linear optimization model, define  $X_1$ ,  $X_2$ ,  $X_3$ , and  $X_4$  to be the number of plugs, rails, rivets, and clips to produce. The problem is to:

$$\text{Maximize Gross Margin} = 0.3X_1 + 1.3X_2 + 0.75X_3 + 1.2X_4$$

subject to the constraint that limits the machining capacity and nonnegativity of the variables:

$$1X_1 + 2.5X_2 + 1.5X_3 + 2X_4 \leq 280,000 \text{ and } X_1, X_2, X_3, X_4 \geq 0$$

To solve this problem, your first thought might be to choose the variable with the highest marginal profit. Because  $X_2$  has the highest marginal profit, you might try producing as many rails as possible. Since each rail requires 2.5 minutes, the maximum number that can be produced is  $280,000/2.5 = 112,000$ , for a total profit of  $1.3(112,000) = \$145,600$ . However, notice that each rail uses a lot more machining time than the other products. The best solution isn’t necessarily the one with the highest marginal profit, but the one that provides the highest *total* profit. Therefore, more profit might be realized by producing a proportionately larger quantity of a different product having a smaller marginal profit. This is the key insight. What the simplex method does is essentially evaluate the impact of constraints in terms of their contribution to the objective function for each variable. For the simple case of only one constraint, the optimal (maximum) solution is found by simply choosing the variable with the highest ratio of the objective coefficient to the constraint coefficient; in this example, the gross margin/unit to the minutes per unit of machining capacity used. The highest ratio occurs for  $X_4$ ,  $1.2/2 = 0.6$ . This can be interpreted as the marginal profit per unit of resource consumed. If we produce the maximum number of clips,  $280,000/2 = 140,000$ , the total profit is  $1.2(140,000) = \$168,000$ , considerably more than the profit for rails alone. The mathematics gets complicated with more constraints and requires multiple iterations to systematically improve the solution, but that’s the basic idea!

The simplex method allows many real business problems involving thousands or even millions of variables, and often hundreds or thousands of constraints, to be solved in reasonable computational time, and is the basis for advanced optimization algorithms involving integer variables that we describe in the next chapter.

## Basic Concepts Review Questions

1. What is mean by optimization? What are some common scenarios where optimization models are used?
2. Describe the four main components of an optimization model. What are its constraints?
3. List the important guidelines to follow for modeling optimization problems on spreadsheets.
4. Why do some common Excel functions cause difficulties when attempting to solve linear programs in Excel?
5. What is a linear program?
6. For each of the examples in this chapter, classify the constraints into the following categories. Are there any other types of constraints that do not fall into these categories?
  - a. Simple bounds
  - b. Limitations
  - c. Requirements
  - d. Proportional relationships
  - e. Balance constraints
7. Discuss the implementation of optimization models on spreadsheets. Name some useful Excel tools which are helpful in this respect.
8. What is the transportation problem?

## Problems and Applications

Note: Data for most of these problems can be found in the Excel file **Chapter 13 Problem Data** to facilitate model development. Tab names correspond to the problem scenario names.

1. Korey is a business student at State U. She has just completed a course in decision models, which had a midterm exam, a final exam, individual assignments, and class participation. She earned an 86% on the midterm, 94% on the final, 93% on the individual assignments, and 85% on participation. The benevolent instructor is allowing his students to determine their own weights for each of the four grade components; of course, with some restrictions:
  - The participation grade can be no more than 15% of the total.
  - The midterm grade must count at least twice as much as the individual assignment score.
  - The final exam grade must count at least three times as much as the individual assignment score.
  - Each of the four components must count for at least 10% of the course grade.
  - The weights must sum to 1.0 and be nonnegative.
    - a. Develop a model that will yield a valid set of weights to maximize Korey's score for the course.
    - b. Implement your model on a spreadsheet and find a good solution using only your intuition.
    - c. Find an optimal solution using *Solver*.
2. Burger Office Equipment produces two types of desks: standard and deluxe. Deluxe desks have oak tops and more expensive hardware and require additional time for finishing and polishing. Standard desks require 80 square feet of pine and 10 hours of labor, while deluxe desks require 60 square feet of pine, 18 square feet of oak, and 16 hours of labor. For the next week, the company has 5,000 square feet of pine, 750 square feet of oak, and 400 hours of labor available. Standard desks net a profit of \$150, while deluxe desks net a profit of \$320. All desks can be sold to national chains such as Staples or Office Depot.

	Control Valve	Metering Pump	Hydraulic Cylinder
Assembly time (min)	45	20	30
Testing time (min)	20	15	25
Profit/unit	\$372	\$174	\$288
Maximum sales	20	50	45
Minimum sales	5	12	22

A total of 3,150 minutes of assembly time and 2,100 minutes of testing time are available next week.
  - a. Develop a linear optimization model to determine how many pieces of equipment the company should make next week to maximize profit contribution.

- b. Implement your model on a spreadsheet and find an optimal solution.
  - c. Explain the sensitivity information for the objective coefficients. What happens if the profit on hydraulic cylinders is decreased by \$10?
  - d. Due to scheduled maintenance, the assembly time is expected to be only 3,000 minutes. How will this affect the solution?
  - e. A worker in the testing department has to take a personal leave because of a death in the family and
- will miss three days (24 hours). How will this affect the optimal solution?
- f. Use the auxiliary variable technique to handle the bound constraints and generate all shadow prices.
4. Fruity Juices, Inc. produces five different flavors of fruit juice: apple, cherry, pomegranate, orange, and pineapple. Each batch of product requires processing in three departments (blending, straining, and bottling). The relevant data (per 1,000-gallon batches) are shown below.

	Time required in minutes/batch					
	Apple	Cherry	Pomegranate	Orange	Pineapple	Minutes Avail.
Blend	23	22	18	19	19	5,000
Strain	22	40	20	34	22	3,000
Bottle	10	10	10	10	10	5,000

	Profit and Sales Potential					
	Apple	Cherry	Pomegranate	Orange	Pineapple	
Profit (\$/1000 gal.)	\$800	\$320	\$1,120	\$1,440	\$800	
Max Sales (000)	20	30	50	50	20	
Min Sales (000)	10	15	20	40	10	

- a. Formulate a linear program to find the amount of each product to produce.
  - b. Implement your model on a spreadsheet and find an optimal solution with *Solver*.
  - c. What effect would an increase of capacity in the straining department have on profit?
5. Rosenberg Land Development (RLD) is a developer of condominium properties in the Southwest United States. RLD has recently acquired a 40.625 acre site outside of Phoenix, Arizona. Zoning restrictions allow at most 8 units per acre. Three types of condominiums are planned: one, two, and three bedroom units. The average construction costs for each type of unit are \$450,000, \$600,000, and \$750,000. These units will generate a net profit of 10%. The company has equity and loans totaling \$180 million dollars for this project. From prior development projects, senior managers have determined that there must be a minimum of 15% one-bedroom units, 25% two-bedroom units, and 25% three-bedroom units.
- a. Develop a linear optimization model to determine how many of each type of unit the developer should build.
  - b. Implement your model on a spreadsheet and find an optimal solution.
  - c. Explain the value of increasing the budget for the project.
6. Marketing managers have various media alternatives, such as radio, TV, magazines, etc., in which to advertise and must determine which to use, the number of insertions in each, and the timing of insertions to maximize advertising effectiveness within a limited budget. Suppose that three media options are available to Kerman Services Corporation: radio, TV, and magazine. The following table provides some information about costs, exposure values, and bounds on the permissible number of ads in each medium desired by the firm. The exposure value is a measure of the number of people exposed to the advertisement and is derived from market research studies, and the client's objective is to maximize the total exposure value. The company would like to achieve a total exposure value of at least 90,000.
- |          | Medium Cost/ad | Exposure Value/ad | Min Units | Max Units |
|----------|----------------|-------------------|-----------|-----------|
| Radio    | \$500          | 2,000             | 0         | 15        |
| TV       | \$2,000        | 4,000             | 10        |           |
| Magazine | \$200          | 2,700             | 6         | 12        |
- How many of each type of ad should be placed in order to minimize the cost of achieving the minimum required total exposure? Use the auxiliary variable approach to model this problem, and write a short memo to the marketing manager explaining the solution and sensitivity information.
7. Klein Industries manufactures three types of portable air compressors: small, medium, and large, which have unit profits of \$20.50, \$34.00, and \$42.00, respectively. The projected monthly sales are:
- |         | Small  | Medium | Large |
|---------|--------|--------|-------|
| Minimum | 14,000 | 6,200  | 2,600 |
| Maximum | 21,000 | 12,500 | 4,200 |

The production process consists of three primary activities: bending and forming, welding, and painting. The amount of time in minutes needed to process each product in each department is shown below:

	<b>Small</b>	<b>Medium</b>	<b>Large</b>	<b>Available Time</b>
Bending/ forming	0.4	0.7	0.8	23,400
Welding	0.6	1.0	1.2	23,400
Painting	1.4	2.6	3.1	46,800

How many of each type of air compressor should the company produce to maximize profit?

- a. Formulate and solve a linear optimization model using the auxiliary variable cells method and write a short memo to the production manager explaining the sensitivity information.
- b. Solve the model without the auxiliary variables and explain the relationship between the reduced costs and the shadow prices found in part a.
8. Beverly Ann Cosmetics has created two new perfumes: Summer Passion and Ocean Breeze. It costs \$5.25 to purchase the fragrance needed for each bottle of Summer Passion, and \$4.70 for each bottle of Ocean Breeze. The marketing department has stated that at least 30% but no more than 70% of the product mix be Summer Passion; the forecasted monthly demand is 7,000 bottles, and is estimated to increase by 8 bottles for each \$1 spent on advertising. For Ocean Breeze, the demand is forecast to be 12,000 bottles, and is expected to increase by 15 bottles for each \$1 spent on advertising. Summer Passion sells for \$42.00 per bottle, and Lavender Breeze for \$30.00 per bottle. A monthly budget of \$100,000 is available for both advertising and purchase of the fragrances. Develop and solve a linear optimization model to determine how much of each type of perfume should be produced to maximize the net profit.
9. A recent MBA graduate, Dara, has gained control over custodial accounts that her parents had established. Currently, her money is invested in four funds but she has identified several other funds as options for investment. She has \$100,000 to invest with the following restrictions:
  - Keep at least \$5,000 in savings.
  - Invest at least 14% in the money market fund.
  - Invest at least 16% in international funds.
  - Keep 35% of funds in current holdings.
  - Do not allocate more than 20% of funds to any one investment except for the money market and savings account.
  - Allocate at least 30% into new investments.

		<b>Average Return</b>	<b>Expenses</b>
1	Large cap blend	17.2%	0.93% (current holding)
2	Small cap growth	20.4%	0.56% (current holding)
3	Green fund	26.3%	0.70% (current holding)
4	Growth and income	15.6%	0.92% (current holding)
5	Multi-cap growth	19.8%	0.92%
6	Mid-cap index	22.1%	0.22%
7	Multi-cap core	27.9%	0.98%
8	Small cap international	35.0%	0.54%
9	Emerging international	36.1%	1.17%
10	Money market fund	4.75%	0
11	Savings account	1.0%	0

Develop a linear optimization model to determine the best investment strategy.

10. Holcomb Candles, Inc. manufactures decorative candles and has contracted with a national retailer to supply a set of special holiday candles to its 8,500 stores. These include large jars, small jars, large pillars, small pillars, and a package of four votive candles. In negotiating the contract for the display, the manufacturer and retailer agreed that 8 feet would be designated for the display in each store, but that at least 2 feet be dedicated to large jars and large pillars, and at least one foot to the votive candle packages. At least as many jars as pillars must be provided. The manufacturer has obtained 200,000 pounds of wax, 250,000 feet of wick, and 100,000 ounces of holiday fragrance. The amount of materials and display size required for each product is shown in the table below:

	<b>Large Jar</b>	<b>Small Jar</b>	<b>Large Pillar</b>	<b>Small Pillar</b>	<b>Votive Pack</b>
Wax	0.5	0.25	0.5	0.25	0.3125
Fragrance	0.24	0.12	0.24	0.12	0.15
Wick	0.43	0.22	0.58	0.33	0.8
Display feet	0.48	0.24	0.23	0.23	0.26
Profit/unit	\$0.25	\$0.20	\$0.24	\$0.21	\$0.16

How many of each product should be made to maximize the profit? Interpret the shadow prices in the sensitivity report.

11. Jaycee's department store chain is planning to open a new store. It needs to decide how to allocate the 100,000 square feet of available floor space among seven departments. Data on expected performance of each department per month, in terms of square feet (sf), are shown below.

Department	Investment/sf	Risk as a % of \$ Invested	Minimum sf	Maximum sf	Expected Profit per sf
Electronics	\$100	24	6,000	30,000	\$12.00
Furniture	50	12	10,000	30,000	6.00
Men's clothing	30	5	2,000	5,000	2.00
Clothing	600	10	3,000	40,000	30.00
Jewelry	900	14	1,000	10,000	20.00
Books	50	2	1,000	5,000	1.00
Appliances	400	3	12,000	40,000	13.00

The company has gathered \$20 million to invest in floor stock. The risk column is a measure of risk associated with investment in floor stock based on past data from other stores and accounts for outdated inventory, pilferage, breakage, etc. For instance, electronics loses 24% of its total investment, furniture loses 12% of its total investment, etc.

- a. Develop a linear optimization model to maximize profit.
- b. If the chain obtains another \$1 million of investment capital for stock, what would the new solution be?

12. Roberto's Honey Farm in Chile makes five types of honey: cream, filtered, pasteurized, mélange (a mixture of several types), and strained, which are sold in 1 or 0.5 kg glass containers, 1 kg and 0.75 kg plastic containers, or in bulk. Key data are shown in the following table.

Selling Prices (Chilean Pesos)				
	0.75 kg Plastic	1 kg Plastic	0.5 kg Glass	1 kg Glass
Cream	744	880	760	990
Filtered	635	744	678	840
Pasteurized	696	821	711	930
Mélange	669	787	683	890
Strained	683	804	697	910

Minimum Demand			
	0.75 kg Plastic	1 kg Plastic	0.5 kg Glass
Cream	300	250	350
Filtered	250	240	300
Pasteurized	230	230	350
Mélange	350	300	250
Strained	360	350	250

Maximum Demand			
	0.75 kg Plastic	1 kg Plastic	0.5 kg Glass
Cream	550	350	470
Filtered	400	380	440
Pasteurized	360	390	490
Mélange	530	410	390
Strained	480	420	380

Package Costs (Chilean pesos)			
0.75 kg Plastic	1 kg Plastic	0.5 kg Glass	1 kg Glass
91	112	276	351

Harvesting and production costs for each product per kilogram in pesos are:

Cream: 322  
Filtered: 255  
Pasteurized: 305  
Mélange: 272  
Strained: 287

Develop a linear optimization model to maximize profit if a total of 10,000 kg of honey is available.

13. Sandford Tile Company makes ceramic and porcelain tile for residential and commercial use. They produce three different grades of tile (for walls, residential flooring, and commercial flooring), each of which requires different amounts of materials and production time, and generates different contribution to profit. The information below shows the percentage of materials needed for each grade and the profit per square foot.

	Grade I	Grade II	Grade III
Profit/square foot	\$2.50	\$4.00	\$5.00
Clay	50%	30%	25%
Silica	5%	15%	10%
Sand	20%	15%	15%
Feldspar	25%	40%	50%

Each week, Sanford Tile receives raw material shipments and the operations manager must schedule the plant to efficiently use the materials to maximize profitability. Currently, inventory consists of 6,000 pounds of clay, 3,000 pounds of silica, 5,000 pounds of sand, and 8,000 pounds of feldspar. Because demand varies for the different grades, marketing estimates that at most 8,000 square feet of Grade III tile should be produced, and that at least 1,500 square feet of Grade I tiles are required. Each square foot of tile weighs approximately two pounds.

- a. Develop a linear optimization model to determine how many of each grade of tile the company should make next week to maximize profit contribution.
- b. Implement your model on a spreadsheet and find an optimal solution.
- c. Explain the sensitivity information for the objective coefficients. What happens if the profit on Grade I is increased by \$0.05?
- d. If an additional 500 pounds of feldspar is available, how will the optimal solution be affected?

e. Suppose that 1,000 pounds of clay are found to be of inferior quality. What should the company do?

f. Use the auxiliary variable cells technique to handle the bound constraints and generate all shadow prices.

14. Janette Douglas is coordinating a bake sale for a non-profit organization. The organization has acquired \$2,200 in donations to hold the sale. The table below shows the amounts and costs of ingredients used per batch of each baked good:

Ingredient	Brownies	Cupcakes	Peanut Butter Cups	Shortbread Cookies	Cost/Unit
Butter (cups)	0.67	0.33	1	0.75	\$1.44
Flour (cups)	1.5	1.5	1.25	2	\$0.09
Sugar (cups)	1.75	1	2	0.25	\$0.16
Vanilla (tsp)	2	0.5	0	0	\$0.06
Eggs	3	2	1	0	\$0.12
Walnuts (cups)	2	0	0	0	\$0.31
Milk (cups)	0.5	1	2	0	\$0.05
Chocolate (oz)	8	2.5	9	0	\$0.10
Baking soda (tsp)	2	1	0	0	\$0.07
Frosting (cups)	0.5	1.5	0	1	\$2.74
Peanut butter (cups)	0	0	2.5	0	\$2.04

One batch results in 10 brownies, 12 cupcakes, 8 peanut butter cups, and 12 shortbread cookies. Each batch of brownies can be sold for \$6.00, cupcakes for \$10.00, peanut butter cups for \$12.00, and shortbread cookies for \$7.50. The organization anticipates that a total of at least 4,000 baked goods must be made. For adequate variety, at least 30 batches of each baked good are required, except for the popular brownies, which require at least 100 batches. In addition, no more than 40 batches of shortbread cookies should be made. How can the organization best use its budget and make the largest amount of money?

15. The International Chef, Inc. markets three blends of oriental tea: premium, Duke Grey, and breakfast. The firm uses tea leaves from India, China, and new domestic California sources.

Tea Leaves (%)			
Quality	Indian	Chinese	California
Premium	40	20	40
Duke Grey	30	50	20
Breakfast	40	40	20

Net profit per pound for each blend is \$0.50 for premium, \$0.30 for Duke Grey, and \$0.20 for breakfast. The firm's regular weekly supplies are 20,000 pounds of Indian tea leaves, 22,000 pounds of Chinese tea leaves, and 16,000 pounds of California tea leaves. Develop and solve a linear optimization model to determine the optimal mix to maximize profit, and write a short memo to

the president, Kathy Chung, explaining the sensitivity information in language that she can understand.

16. Young Energy operates a power plant that includes a coal-fired boiler to produce steam to drive a generator. The company can purchase different types of coals and blend them to meet the requirements for burning in the boiler. The table below shows the characteristics of the different types of coals:

Type	BTU/lb.	% Ash	% Moisture	Cost (\$/lb)
A	11,500	13%	10%	\$2.49
B	11,800	10%	8%	\$3.04
C	12,200	12%	8%	\$2.99
D	12,100	12%	8%	\$2.61

The required BTU/lb must be at least 11,900. In addition, the ash content can be at most 12.2% and the moisture content at most 9.4%. Develop and solve a linear optimization model to find the best coal blend for Young Energy. Explain how the company might reduce its costs by changing the blending restrictions.

17. The Hansel Corporation, located in Bangalore, India, makes plastics materials that are mixed with various additives and reinforcing materials before being melted, extruded, and cut into small pellets for sale to other manufacturers. Four grades of plastic are made, each of which might include up to four different additives. The table below shows the number of pounds of additive per pound of each grade of final product, the weekly availability of the additives, and cost and profitability information.

	<b>Grade 1</b>	<b>Grade 2</b>	<b>Grade 3</b>	<b>Grade 4</b>	<b>Availability</b>
Additive A	0.40	0.37	0.34	0.90	100,000
Additive B	0.30	0.33	0.33		90,000
Additive C	0.20	0.25	0.33		40,000
Additive D	0.10	0.05		0.10	10,000
Profit/lb	\$2.00	\$1.70	\$1.50	\$2.80	

Because of marketing considerations, the total amount of grades 1 and 2 should not exceed 65% of the total of all grades produced, and at least 25% of the total product mix should be grade 4.

- a. How much of each grade should be produced to maximize profit? Develop and solve a linear optimization model.

b. A labor strike in India leads to a shortage of 20,000 units of additive C. What should the production manager do?

c. Management is considering raising the price on grade 2 to \$2.00 per pound. How will the solution be changed?

18. Liquid Gold, Inc. transports radioactive waste from nuclear power plants to disposal sites around and about the country. Each plant has an amount of material that must be moved each period. Each site has a limited capacity per period. The cost of transporting between sites is given here (some combinations of plants and storage sites are not to be used, and no figure is given). Develop and solve a transportation model for this problem.

Cost to Site							
Plant	Material	S1	S2	S3	S4	Site	Capacity
P1	20,876	\$105	\$86	—	\$23	S1	285,922
P2	50,870	\$86	\$58	\$41	—	S2	308,578
P3	38,652	\$93	\$46	\$65	\$38	S3	111,955
P4	28,951	\$116	\$27	\$94	—	S4	208,555
P5	87,423	\$88	\$56	\$82	\$89		
P6	76,190	\$111	\$36	\$72	—		
P7	58,237	\$169	\$65	\$48	—		

19. The Children's Theater Company is a nonprofit corporation managed by Shannon Board. The theater performs in two venues: Kristin Marie Hall and the Lauren Elizabeth Theater. For the upcoming season, seven shows have been chosen. The question Shannon faces is how many performances of each of the seven shows should be scheduled. A financial analysis has estimated revenues for each performance of the seven shows, and Shannon has set the minimum number of performances of each show based upon union agreements with Actor's Equity Association and the popularity of the shows in other markets:

Show	Revenue	Cost	Minimum Number of Performances
1	\$2,217	\$968	32
2	\$2,330	\$1,568	13
3	\$1,993	\$755	23
4	\$3,364	\$1,148	34
5	\$2,868	\$1,180	35
6	\$3,851	\$1,541	16
7	\$1,836	\$1,359	21

Kristin Marie Hall is available for 60 performances during the season, while Lauren Elizabeth Theater is available for 150 performances. Shows 3 and 7 must be performed in Kristin Marie Hall, while the other shows are performed in the Lauren Elizabeth Theater. The company wants to achieve revenues of at least \$550,000 while minimizing its production costs. Develop and solve a linear optimization model to determine the best

way to schedule the shows. Is it possible to achieve revenues of \$600,000? What is the highest amount of revenue that can be achieved?

20. The Kelly Company is a producer of canned vegetables. The company's food processing plants are located in Sacramento and Dallas. Distribution centers are located in Denver, Cincinnati, Phoenix, and Atlanta. The supply chain manager wants to minimize the cost of shipping the goods from the processing plants to the distribution centers. Relevant data are given below. Develop and solve a linear optimization model to minimize the cost of shipping while meeting capacity and demand requirements. Evaluate the sensitivity report and provide a full explanation and any recommendations to the supply chain manager.

Per-Case Cost of Shipping				
	Denver	Phoenix	Cincinnati	Atlanta
<b>Sacramento</b>	\$1.75	\$1.25	\$3.25	\$5.50
<b>Dallas</b>	\$1.90	\$1.35	\$2.75	\$2.55
Weekly Capacity (cases)				
<b>Sacramento</b>	3,000			
<b>Dallas</b>	5,000			
Weekly Demand (cases)				
<b>Denver</b>	1,200			
<b>Phoenix</b>	1,000			
<b>Cincinnati</b>	1,900			
<b>Atlanta</b>	2,500			

21. Shafer Office Supplies has four distribution centers located in Atlanta, Cincinnati, Chicago, and Salt Lake City, and ships to 12 retail stores located in Portland, San Jose, Las Vegas, Tucson, Colorado Springs, Kansas City, St. Paul, Austin, Jackson, Montgomery, Cleveland, and Pittsburgh. The company wants to minimize the transportation cost of shipping one of its higher-volume products, boxes of standard copy paper. The per-unit shipping cost from each distribution center to each retail location and the amounts currently in inventory and ordered at each retail location are shown in the table below. Develop and solve an optimization model

	<b>Seattle</b>	<b>San Francisco</b>	<b>Las Vegas</b>	<b>Tucson</b>	<b>Denver</b>	<b>Charlotte</b>
<b>Atlanta</b>	\$2.15	\$2.10	\$1.75	\$1.50	\$1.20	\$0.65
<b>Lexington</b>	\$1.95	\$2.00	\$1.70	\$1.53	\$1.10	\$0.55
<b>Milwaukee</b>	\$1.70	\$1.85	\$1.50	\$1.41	\$0.95	\$0.40
<b>Salt Lake City</b>	\$0.60	\$0.55	\$0.35	\$0.60	\$0.40	\$0.95
<b>Demand</b>	5,000	16,000	4,200	3,700	4,500	7,500

	<b>Fayetteville</b>	<b>Birmingham</b>	<b>Orlando</b>	<b>Cleveland</b>	<b>Philadelphia</b>	<b>Supply</b>
<b>Atlanta</b>	\$0.80	\$0.35	\$0.15	\$0.60	\$0.50	40,000
<b>Lexington</b>	\$1.05	\$0.60	\$0.50	\$0.25	\$0.30	35,000
<b>Milwaukee</b>	\$0.95	\$0.70	\$0.70	\$0.35	\$0.40	15,000
<b>Salt Lake City</b>	\$1.10	\$1.35	\$1.60	\$1.60	\$1.70	16,000
<b>Demand</b>	9,000	3,300	12,000	9,500	16,000	

22. Mirza Manufacturing makes four electronic products, each of which is comprised of three main materials: magnet, wiring, and casing. The products are shipped to three distribution centers in North America, Europe, and Asia. Marketing has specified that no location should receive more than the maximum demand and should receive at least the minimum demand. The material costs/unit are: magnet—\$0.59, wire—\$0.29, and casing—\$0.31. The table below shows the number of units of each material required in each unit of end product and the production cost per unit.

<b>Product</b>	<b>Production Cost/Unit</b>	<b>Magnets</b>	<b>Wire</b>	<b>Casing</b>
A	\$0.25	4	2	2
B	\$0.35	3	1	3
C	\$0.15	2	2	1
D	\$0.10	8	3	2

Additional information is provided below.

<b>Min Demand</b>			
<b>Product</b>	<b>NA</b>	<b>EU</b>	<b>Asia</b>
A	850	900	100
B	700	200	500
C	1,100	800	600
D	1,500	3,500	2,000

- to minimize the total transportation cost and answer the following questions. Use the sensitivity report to answer as appropriate to answer questions c and d.
- What is the minimum monthly cost of shipping?
  - Which plants will operate at capacity in this solution?
  - Suppose that 500 units of extra supply are available (and that the cost of this extra capacity is a sunk cost). To which plant should this extra supply go, and why?
  - Suppose that the cost of shipping from Atlanta to Birmingham increased to \$0.45 per unit. What would happen to the optimal solution?

<b>Max Demand</b>			
<b>Product</b>	<b>NA</b>	<b>EU</b>	<b>Asia</b>
A	2,550	2,700	300
B	2,100	600	1,500
C	3,300	2,400	1,800
D	4,500	10,500	6,000

<b>Packaging and Shipping Cost/Unit</b>			
<b>Product</b>	<b>NA</b>	<b>EU</b>	<b>Asia</b>
A	\$0.20	\$0.25	\$0.35
B	\$0.18	\$0.22	\$0.30
C	\$0.18	\$0.22	\$0.30
D	\$0.17	\$0.20	\$0.25

<b>Unit Sales Revenue</b>			
<b>Product</b>	<b>NA</b>	<b>EU</b>	<b>Asia</b>
A	\$4.00	\$4.50	\$4.55
B	\$3.70	\$3.90	\$3.95
C	\$2.70	\$2.90	\$2.40
D	\$6.80	\$6.50	\$6.90

<b>Available Raw Material</b>			
	<b>Magnet</b>	<b>Wire</b>	<b>Casing</b>
Magnet	120,000		
Wire		50,000	
Casing		40,000	

Develop an appropriate linear optimization model to maximize net profit.

- 23.** Raturi and Rao, Inc. produces four industrial chemicals with variable production costs of \$9.00, \$6.75, \$5.25, and \$7.50 per pound, respectively. Because of increasing supplier costs, the variable cost of each of the products will increase by 6% at the beginning of month three. Demand forecasts are shown in the table below. There are currently 100 pounds of each product on hand, and company wants to maintain an inventory of 100 pounds of each product at the end of every month. The four products share a common process that operates two shifts of 8 hours each per day, seven days per week. Processing requirements are 0.06 hours/lb for product 1, 0.05 hours/lb for product 2, 0.2 hours/lb for product 3, and 0.11 hours/lb for product 4. The per-pound cost of holding inventory each month is estimated to be 12% of the cost of the product. Develop an optimization model to meet demand and minimize the total cost. Implement your model on a spreadsheet and find an optimal solution with *Solver*.

**Product Demand**

Product	Month 1	Month 2	Month 3
1	1,000	800	1,000
2	1,000	900	500
3	600	600	500
4	0	200	500

- 24.** An international graduate student will receive a \$28,000 foundation scholarship and reduced tuition. He must pay \$1,500 in tuition for each of the autumn, winter, and spring quarters, and \$500 in the summer. Payments are due on the first day of September, December, March, and May, respectively. Living expenses are estimated to be \$1,500 per month, payable on the first day of the month. The foundation will pay him \$18,000 on August 1, and the remainder on May 1. To earn as much interest as possible, the student wishes to invest the money. Three types of investments are available at

his bank: a 3-month CD, earning 0.75% (net 3-month rate); a 6-month CD, earning 1.9%; and a 12-month CD, earning 4.2%. Develop a linear optimization model to determine how he can best invest the money and meet his financial obligations.

- 25.** Jason Wright is a part-time MBA student who would like to optimize his financial decisions. Currently, he has \$16,000 in his savings account. Based on an analysis of his take-home pay, expected bonuses, and anticipated tax refund, he has estimated his income for each month over the next year. In addition, he has estimated his monthly expenses, which vary because of scheduled payments for insurance, utilities, tuition and books, and so on. The table below summarizes his estimates:

Month	Income	Expenses
1	\$3,400	\$3,360
2	\$3,400	\$2,900
3	\$3,400	\$6,600
4	\$9,500	\$2,750
5	\$3,400	\$2,800
6	\$5,000	\$6,800
7	\$4,600	\$3,200
8	\$3,400	\$3,600
9	\$3,400	\$6,550
10	\$3,400	\$2,800
11	\$3,400	\$2,900
12	\$5,000	\$6,650

Jason has identified several short-term investment opportunities:

- A 3-month CD yielding 0.60% at maturity
- A 6-month CD yielding 1.42% at maturity
- An 11-month CD yielding 3.08% at maturity
- His savings account yields 0.0375% per month

To ensure enough cash for emergencies, he would like to maintain at least \$2,000 in the savings account. Jason's objective is to maximize his cash balance at the end of the year. Develop a linear optimization model to find the best investment strategy.

## Case

### Haller's Pub & Brewery

Jeremy Haller of Haller's Pub & Brewery has compiled data describing the amount of different ingredients and labor resources needed to brew six different types of beers that the brewery makes. He also gathered financial information and estimated demand over a 26-week forecast horizon. These data are provided in the *Haller's Pub* tab of the Excel file *Chapter 13 Problem Data*. The profits for each batch of each type of beer are:

Light Ale: \$3,925.78

Golden Ale: \$4,062.75

Freedom Wheat: \$3,732.34

Berry Wheat: \$3,704.49

Dark Ale: \$3,905.79

Hearty Stout: \$3,490.22

These values incorporate fixed overhead costs of \$7,500 per batch. Use the data to validate the profit figures and develop a linear optimization model to maximize profit. Write a report to Mr. Haller explaining the results and sensitivity analysis information in language that he (a nonquantitative manager) can understand.

## *Chapter 14*

# Integer, Nonlinear, and Advanced Optimization Methods

- INTRODUCTION 482
- INTEGER OPTIMIZATION MODELS 483
  - A Cutting Stock Problem 483
  - Solving Integer Optimization Models 484
- INTEGER OPTIMIZATION MODELS WITH BINARY VARIABLES 487
  - Project Selection 487
  - Site Location Model 488
  - Computer Configuration 491
  - A Supply Chain Facility Location Model 494
- MIXED INTEGER OPTIMIZATION MODELS 495
  - Plant Location Model 495
  - A Model with Fixed Costs 497
- NONLINEAR OPTIMIZATION 499
  - Hotel Pricing 499
  - Solving Nonlinear Optimization Models 501
  - Markowitz Portfolio Model 503
- EVOLUTIONARY SOLVER FOR NONSMOOTH OPTIMIZATION 506
  - Rectilinear Location Model 508
  - Job Sequencing 509
  - Risk Analysis and Optimization 512
- COMBINING OPTIMIZATION AND SIMULATION 515
  - A Portfolio Allocation Model 515
  - Using *OptQuest* 516
- BASIC CONCEPTS REVIEW QUESTIONS 524
- PROBLEMS AND APPLICATIONS 524
- CASE: TINDALL BOOKSTORES 530

## **INTRODUCTION**

This chapter extends the concepts developed in Chapter 13 to integer and nonlinear optimization models. In an **integer linear optimization model (integer program)**, some or all of the variables are restricted to being *whole numbers*. If only a subset of variables is restricted to being integer

while others are continuous, we call this a **mixed integer linear optimization model**. A special type of integer problem is one in which variables can only be 0 or 1. These *binary variables* help us to model logical “yes or no” decisions. Integer linear optimization models are generally more difficult to solve than pure linear optimization models. In a **nonlinear optimization model (nonlinear program)**, the objective function and/or constraint functions are *nonlinear functions* of the decision variables; that is, terms cannot be written as a constant times a variable. Some examples of nonlinear terms are  $3x^2$ ,  $4/y$ , and  $6xy$ . Nonlinear optimization models are considerably more difficult to solve than linear or integer models. Both integer and nonlinear models have many important applications in areas such as scheduling, supply chains, and portfolio management.

## INTEGER OPTIMIZATION MODELS

An *integer (linear) optimization model* is a linear model in which some or all of the decision variables are restricted to integer (whole-number) values. For many practical applications, we need not worry about forcing the decision variables to be integers. For example, in deciding on the optimal number of cases of diapers to be produced next month, we could use a linear model, since rounding a value like 5,621.63 would have little impact on the results. However, in a production planning decision involving low-volume, high-cost items such as airplanes, an optimal value of 10.42 would make little sense, and a difference of one unit (rounded up or down) could have significant economic consequences. Decision variables that we force to be integers are called *general integer variables*. An example for which general integer variables are needed follows.

### A Cutting Stock Problem

The paper industry uses integer optimization to find the best mix of cutting patterns to meet demand for various sizes of paper rolls. In a similar fashion, sheet steel producers cut strips of different sizes from rolled coils of thin steel. Suppose that a company makes standard 110-inch-wide rolls of thin sheet metal, and slits them into smaller rolls to meet customer orders for widths of 12, 15, and 30 inches. The demands for these widths vary from week to week.

From a 110-inch roll, there are many different ways to slit 12-, 15-, and 30-inch pieces. A *cutting pattern* is a configuration of the number of smaller rolls of each type that are cut from the raw stock. Of course, one would want to use as much of the roll as possible to avoid costly scrap. For example, one could cut seven 15-inch rolls, leaving a 5-inch piece of scrap. Finding good cutting patterns for a large set of end products is in itself a challenging problem. Suppose that the company has proposed the following cutting patterns:

Size of End Item				
Pattern	12"	15"	30"	Scrap
1	0	7	0	5"
2	0	1	3	5"
3	1	0	3	8"
4	9	0	0	2"
5	2	1	2	11"
6	7	1	0	11"

Demands this week are 500 12-inch rolls, 715 15-inch rolls, and 630 30-inch rolls. The problem is to develop a model that will determine how many 110-inch rolls to cut into each of the six patterns in order to meet demand and scrap.

Define  $X_i$  to be the number of 110-inch rolls to cut using cutting pattern  $i$ , for  $i = 1, \dots, 6$ . Note that  $X_i$  needs to be a whole number because each roll that is cut generates a different

number of end items. Thus,  $X_i$  will be modeled using general integer variables. Because the objective is to minimize scrap, the objective function is:

$$\text{Min } 5X_1 + 5X_2 + 8X_3 + 2X_4 + 11X_5 + 11X_6$$

The only constraints are that end item demand must be met; that is, we must produce at least 500 12-inch rolls, 715 15-inch rolls, and 630 30-inch rolls. The number of end item rolls produced is found by multiplying the number of end item rolls produced by each cutting pattern by the number of 110-inch rolls cut using that pattern. Therefore, the constraints are:

$$0X_1 + 0X_2 + 1X_3 + 9X_4 + 2X_5 + 7X_6 \geq 500 \quad (\text{12-inch rolls})$$

$$7X_1 + 1X_2 + 0X_3 + 0X_4 + 1X_5 + 1X_6 \geq 715 \quad (\text{15-inch rolls})$$

$$0X_1 + 3X_2 + 3X_3 + 0X_4 + 2X_5 + 0X_6 \geq 630 \quad (\text{30-inch rolls})$$

Finally, we include nonnegativity and integer restrictions:

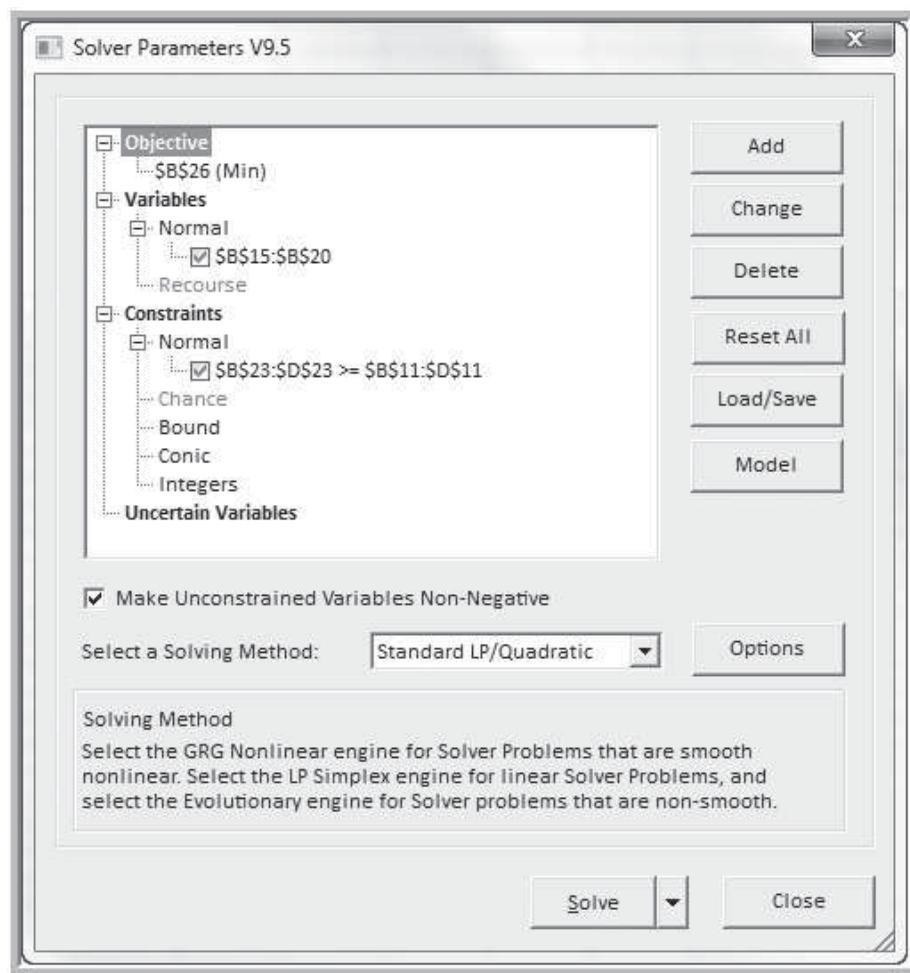
$$X_i \geq 0 \text{ and integer}$$

### Solving Integer Optimization Models

Integer optimization models are set up in the same manner as linear models in a spreadsheet. Figure 14.1 shows the cutting stock model implementation on a spreadsheet (Excel file *Cutting Stock Model*) and the solution that results if we ignore the integer

	A	B	C	D	E
1	<b>Cutting Stock Problem</b>				
2					
3	<b>Data</b>				
4	Pattern	12-in rolls	15-in rolls	30-in rolls	Scrap
5		1	0	7	0
6		2	0	1	3
7		3	1	0	3
8		4	9	0	0
9		5	2	1	2
10		6	7	1	0
11	Demand	500	715	630	
12					
13	<b>Model</b>				
14		No. of rolls			
15	Pattern 1	72.14			
16	Pattern 2	210.00			
17	Pattern 3	0.00			
18	Pattern 4	55.56			
19	Pattern 5	0.00			
20	Pattern 6	0.00			
21					
22		12-in rolls	15-in rolls	30-in rolls	
23	Number produced	500	715	630	
24					
25		Total			
26	Scrap	1521.8254			
27					

**FIGURE 14.1** Cutting Stock Model Spreadsheet and Linear Optimization Solution

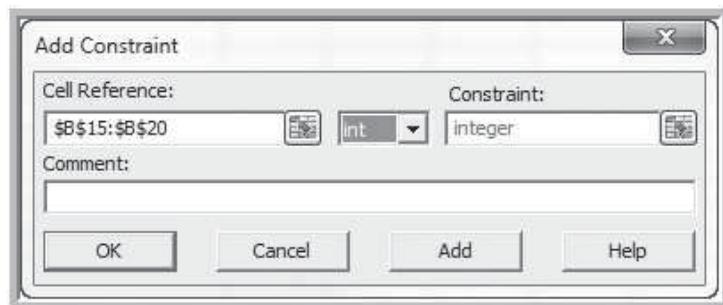


**FIGURE 14.2** Solver Model for Cutting Stock Problem

restrictions on the variables (that is, the solution to the linear optimization model). The *Solver* model is shown in Figure 14.2. Note that the optimal solution results in fractional values of the variables.

To enforce integer restrictions on variables using *Solver*, click on *Integers* under the *Constraints* list and then click the *Add* button. In the *Add Constraint* dialog, enter the variable range in the *Cell Reference* field and choose *int* from the drop-down box as shown in Figure 14.3. Figure 14.4 shows the resulting solution. Notice that the minimum value of the objective function (total scrap) is larger than the linear optimization solution. This is expected because we have added an additional constraint (the integer restrictions). Whenever you add a constraint to a model, the value of the objective function can never improve and usually worsens. Also note that simply rounding the linear solution would not have provided the optimal integer solution.

If the optimal linear solution had turned out to have all integer values, then it clearly would have solved the integer model. In fact, the algorithm used to solve integer optimization models begins by solving the associated linear model without the integer restrictions and proceeds to enforce the integer restrictions using a



**FIGURE 14.3** Enforcing Integer Restrictions in *Solver*

A	B	C	D	E
1 <b>Cutting Stock Problem</b>				
2				
3 <b>Data</b>				
4 Pattern	12-in rolls	15-in rolls	30-in rolls	Scrap
5	1	0	7	0
6	2	0	1	3
7	3	1	0	3
8	4	9	0	0
9	5	2	1	2
10	6	7	1	0
11 Demand	500	715	630	
12				
13 <b>Model</b>				
14	No. of rolls			
15 Pattern 1	73.00			
16 Pattern 2	210.00			
17 Pattern 3	0.00			
18 Pattern 4	56.00			
19 Pattern 5	0.00			
20 Pattern 6	0.00			
21				
22	12-in rolls	15-in rolls	30-in rolls	
23 Number produced	504	721	630	
24				
25 Total				
26 Scrap	1527			

**FIGURE 14.4** Optimal Integer Solution to the Cutting Stock Problem

systematic search process that involves solving a series of modified linear optimization problems.

#### SKILL-BUILDER EXERCISE 14.1

Find other feasible cutting patterns for the cutting stock model and include them in the model. Does adding the additional patterns improve the solution?

Because integer models are discontinuous by their very nature, sensitivity information cannot be generated in the same manner as for linear models, and therefore, no sensitivity report is provided by *Solver*; only the Answer Report is available. To investigate changes in model parameters, it is necessary to re-solve the model.

## INTEGER OPTIMIZATION MODELS WITH BINARY VARIABLES

Many optimization models require *binary variables*, which are variables that are restricted to being either 0 or 1. Mathematically, a binary variable  $x$  is simply a general integer variable that is restricted to being between 0 and 1:

$$0 \leq x \leq 1 \text{ and integer}$$

However, we usually just write this as:

$$x = 0 \text{ or } 1$$

Binary variables enable us to model logical decisions in optimization models. For example, binary variables can be used to model decisions such as whether ( $x = 1$ ) or not ( $x = 0$ ) to place a facility at a certain location, whether or not to run a production line, or whether or not to invest in a certain stock. One common example we present next is project selection, in which a subset of potential projects must be selected with limited resource constraints. Capital budgeting problems in finance have a similar structure.

### Project Selection

Hahn Engineering's research and development group has identified five potential new engineering and development projects; however, the firm is constrained by its available budget and human resources. Each project is expected to generate a return (given by the net present value) but requires a fixed amount of cash and personnel. Because the resources are limited, all projects cannot be selected. Projects cannot be partially completed; thus, either the project must be undertaken completely or not at all. The data are given in Table 14.1. If a project is selected, it generates the full value of the expected return and requires the full amount of cash and personnel shown in Table 14.1. For example, if we select projects 1 and 3, the total return is  $\$180,000 + \$150,000 = \$330,000$ , and these projects require cash totaling  $\$55,000 + \$24,000 = \$79,000$  and  $5 + 2 = 7$  personnel.

To model this situation, we define the decision variables to be binary, corresponding to either not selecting or selecting each project, respectively. Define  $X_i = 1$  if project  $i$  is selected, and 0 otherwise. By multiplying these binary variables by the expected returns, the objective function is:

$$\text{Maximize } \$180,000X_1 + \$220,000X_2 + \$150,000X_3 + \$140,000X_4 + \$200,000X_5$$

Because cash and personnel are limited, we have the constraints:

$$\begin{aligned} \$55,000X_1 + \$83,000X_2 + \$24,000X_3 + \$49,000X_4 \\ + \$61,000X_5 &\leq \$150,000 \quad (\text{cash limitation}) \\ 5X_1 + 3X_2 + 2X_3 + 5X_4 + 3X_5 &\leq 12 \quad (\text{personnel limitation}) \end{aligned}$$

**TABLE 14.1** Project Selection Data

	Project 1	Project 2	Project 3	Project 4	Project 5	Available Resources
Expected return (NPV)	\$180,000	\$220,000	\$150,000	\$140,000	\$200,000	
Cash requirements	\$55,000	\$83,000	\$24,000	\$49,000	\$61,000	\$150,000
Personnel requirements	5	3	2	5	3	12

A	B	C	D	E	F	G
1 Project Selection Model						
2						
3 Data						
4	Project 1	Project 2	Project 3	Project 4	Project 5	Available
5 Expected Return (NPV)	\$ 180,000	\$ 220,000	\$ 150,000	\$ 140,000	\$ 200,000	Resources
6 Cash requirements	\$ 55,000	\$ 83,000	\$ 24,000	\$ 49,000	\$ 61,000	\$ 150,000
7 Personnel requirements	5	3	2	5	3	12
8						
9 Model						
10						
11 Project selection decisions	1	0	1	0	1	Total
12 Cash Used	\$ 55,000	\$ -	\$ 24,000	\$ -	\$ 61,000	\$ 140,000
13 Personnel Used	5	0	2	0	3	10
14 Return	\$ 180,000	\$ -	\$ 150,000	\$ -	\$ 200,000	\$ 530,000

**FIGURE 14.5 Project Selection Model Spreadsheet**

Note that if projects 1 and 3 are selected, then  $X_1 = 1$  and  $X_3 = 1$  and the objective and constraint functions equal:

$$\begin{aligned} \text{Return} &= \$180,000(1) + \$220,000(0) + \$150,000(1) + \$140,000(0) \\ &\quad + \$200,000(0) = \$330,000 \end{aligned}$$

$$\begin{aligned} \text{Cash Required} &= \$55,000(1) + \$83,000(0) + \$24,000(1) + \$49,000(0) \\ &\quad + \$61,000(0) = \$79,000 \end{aligned}$$

$$\text{Personnel Required} = 5(1) + 3(0) + 2(1) + 5(0) + 3(0) = 7$$

This model is easy to implement on a spreadsheet, as shown in Figure 14.5 (Excel file *Project Selection Model*). The decision variables are defined in cells B11:F11. The objective function, computed in cell G14, is the total return, which can be expressed as the sum of the product of the return from each project and the binary decision variable:

$$\text{Total Return} = B5 \times B11 + C5 \times C11 + D5 \times D11 + E5 \times E11 + F5 \times F11$$

These constraints can be written as:

$$\text{Cash Used} = B6 \times B11 + C6 \times C11 + D6 \times D11 + E6 \times E11 + F6 \times F11 \leq G6$$

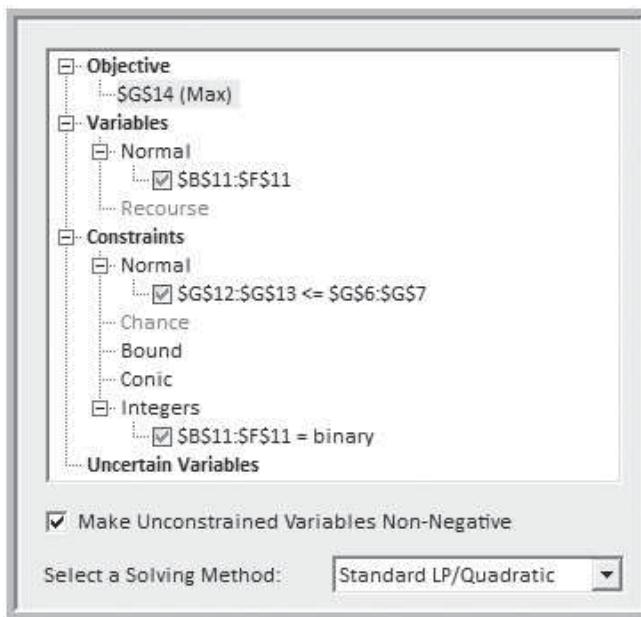
$$\text{Personnel Used} = B7 \times B11 + C7 \times C11 + D7 \times D11 + E7 \times E11 + F7 \times F11 \leq G7$$

The left-hand sides of these functions can be found in cells G12 and G13.

The *Solver* model is shown in Figure 14.6. To invoke the binary constraints on the variables, use the same process as defining integer variables, but choose *bin* from the drop-down box in the *Add Constraint* dialog. The resulting constraint is  $B11:F11 = \text{binary}$  as shown in the *Solver* model. The optimal solution, shown in Figure 14.5, is to select projects 1, 3, and 5 for a total return of \$530,000.

### Site Location Model

Integer optimization models have wide applications in locating facilities. The following is an example of a “covering problem,” one in which we seek to choose a subset of locations that serve, or cover, all locations in a service area. Suppose that an unincorporated township wishes to find the best locations for fire stations. Assume that the township is divided into smaller districts or neighborhoods, and that transportation studies have estimated the response time for emergency vehicles to travel between each pair of districts. The township wants to locate the fire stations so that all districts



**FIGURE 14.6** Solver Model for Project Selection

can be reached within an 8-minute response time. The table below shows the estimated response time in minutes between each pair of districts:

From/To	1	2	3	4	5	6	7
1	0	2	10	6	12	5	8
2	2	0	6	9	11	7	10
3	10	6	0	5	5	12	6
4	6	9	5	0	9	4	3
5	12	11	5	9	0	10	8
6	5	7	12	4	10	0	6
7	8	10	6	3	8	6	0

Define  $X_j = 1$  if a fire station is located in district  $j$ , and 0 if not. The objective is to minimize the number of fire stations that need to be built:

$$\text{Min } X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7$$

Each district must be reachable within 8 minutes by some fire station. Thus, from the table, for example, we see that in order to be able to respond to district 1 in 8 minutes or less, a station must be located in either district 1, 2, 4, 6, or 7. Therefore, we must have the constraint:

$$X_1 + X_2 + X_4 + X_6 + X_7 \geq 1$$

Similar constraints may be formulated for each of the other districts:

$$\begin{aligned}
 X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 &\geq 1 && (\text{District 2}) \\
 X_1 + X_2 + X_3 + X_6 &\geq 1 && (\text{District 3}) \\
 X_2 + X_3 + X_4 + X_5 + X_7 &\geq 1 && (\text{District 4}) \\
 X_1 + X_3 + X_4 + X_6 + X_7 &\geq 1 && (\text{District 5}) \\
 X_1 + X_2 + X_4 + X_6 + X_7 &\geq 1 && (\text{District 6}) \\
 X_1 + X_3 + X_4 + X_5 + X_6 + X_7 &\geq 1 && (\text{District 7})
 \end{aligned}$$

	A	B	C	D	E	F	G	H	I	J
1	<b>Fire Station Location Model</b>									
2										
3	<b>Data</b>									
4										
5	Response time		8							
6										
7	<b>Response Times</b>									
8	From/To	1	2	3	4	5	6	7		
9	1	0	2	10	6	12	5	8		
10	2	2	0	6	9	11	7	10		
11	3	10	6	0	5	5	12	6		
12	4	6	9	5	0	9	4	3		
13	5	12	11	5	9	0	10	8		
14	6	5	7	12	4	10	0	6		
15	7	8	10	6	3	8	6	0		
16										
17	<b>Model</b>									
18										
19	From/To	1	2	3	4	5	6	7	Covered?	Requirement
20	1	1	1	0	1	0	1	1	1	1
21	2	1	1	1	0	0	1	0	1	1
22	3	0	1	1	1	1	0	1	2	1
23	4	1	0	1	1	0	1	1	2	1
24	5	0	0	1	0	1	0	1	2	1
25	6	1	1	0	1	0	1	1	1	1
26	7	1	0	1	1	1	1	1	2	1
27								Total		
28	Location	0	0	1	0	0	0	1	2	

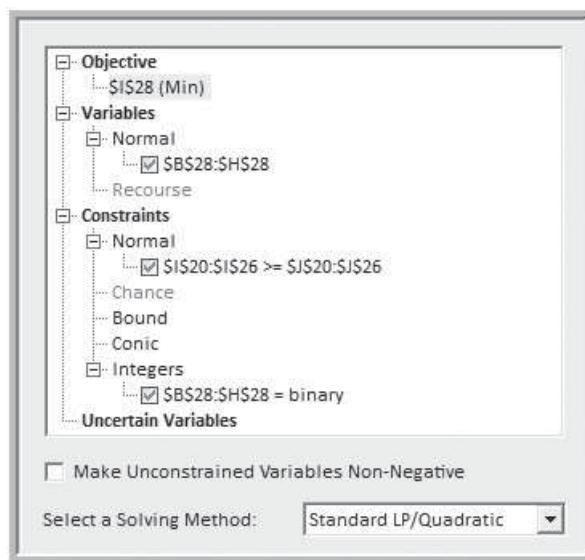
**FIGURE 14.7** Fire Station Location Model Spreadsheet

Figure 14.7 shows a spreadsheet model for this problem (Excel file *Fire Station Location Model*). To develop the constraints in the model, we construct a matrix by converting all response times that are within 8 minutes to 1s, and those that exceed 8 minutes to 0s. Then the constraint functions for each district are simply the SUMPRODUCT of the decision variables and the rows of this matrix, making the *Solver* model, shown in Figure 14.8, easy to define. For this example, the solution is to site fire stations in districts 3 and 7.

As noted, sensitivity analysis can only be conducted for integer optimization by re-solving the model. Suppose that the township's Board of Trustees wants to better understand the trade-offs between the response time and minimum number of fire stations needed. We could change the value of the response time in cell C5 and resolve the model, obtaining the results:

Response Time	Min. # of Sites
5	3
6	2
7	2
8	2
9	1
10	1

These results show the maximum response time can be reduced to six minutes while still using only two fire stations (the model solution yields districts 1 and 3). This would



**FIGURE 14.8** Solver Model for Fire Station Location

clearly be a better alternative. Also, if the response time is increased by only one minute from its original target, the township could save the cost of building a second facility. Of course, such decisions need to be evaluated carefully.

### SKILL-BUILDER EXERCISE 14.2

In the fire station location model, suppose that the township wants to be able to reach each district from at least two alternative fire stations. Modify the model to find a solution.

The next example illustrates a more complicated example involving the use of binary variables to configure a personal computer.

### Computer Configuration

Consumers and business customers have the opportunity to select the features and options of many products when ordering. For example, companies like Dell allow customers to choose the configuration of computers. Suppose that a customer wants to buy a new notebook computer with a limited budget of \$1,300. Many options are generally available. For example, suppose that the base price of a notebook is \$500 and includes one-year warranty, 1 GB RAM, a 160 GB hard drive; CD-ROM/DVD read-only optical drive, and a four-cell lithium ion battery. The following options are available, along with price changes:

- Processor: slower (subtract \$179), faster (add \$100), fastest (add \$300)
- Warranty: two year (add \$129), three year (add \$269)
- Memory: 2 GB (add \$50), 4 GB (add \$500)
- Hard drive: 80 GB (subtract \$29), 250 GB (add \$39), 320 GB (add \$79)
- Optical drive: CD-ROM only (subtract \$39); CD/DVD burner (add \$79), CD/DVD dual-layer burner (add \$179)
- Battery: eight-cell lithium ion (add \$59)
- Enhanced video and photo software (add \$79)

If the customer wants to configure the top-of-the-line system, it would require an additional \$1,465 over the base price, exceeding the budget. Thus, the customer must choose the options carefully.

We can build a model for doing this by defining binary variables corresponding to each possible choice. Let:

$X_{p1} = 1$  if slower processor is selected;  $X_{p2} = 1$  if faster processor is selected, and

$X_{p3} = 1$  if fastest processor is selected

$X_{w1} = 1$  if two-year warranty is chosen;  $X_{w2} = 1$  if three-year warranty is chosen

$X_{m1} = 1$  if 2 GB memory is chosen;  $X_{m2} = 1$  if 4 GB is chosen

$X_{d1} = 1$  if 80 GB hard drive is chosen;  $X_{d2} = 1$  if 250 GB hard drive is chosen; and

$X_{d3} = 1$  if 320 GB hard drive is chosen.

$X_{c1} = 1$  if CD only is chosen;  $X_{c2} = 1$  if CD/RW is chosen, and

$X_{c3} = 1$  if DVD/RW is chosen

$X_{c1} = 1$  if CD-ROM only is chosen;  $X_{c2} = 1$  if CD/DVD burner is chosen, and

$X_{c3} = 1$  if CD/DVD dual layer burner is chosen

$X_b = 1$  if eight-cell battery is chosen

$X_s = 1$  if enhanced video and photo software is chosen

One way of specifying an appropriate objective function is to develop preferences for each option. This can be done using the notion of utility that we described in Chapter 10. For example, suppose that the customer is intending to purchase the machine for music and photo storage; in this case, a large hard drive might be the most important feature. If the computer is to be used for video editing, then a fast processor, large hard drive, and large RAM memory would also be most important. For business travel, then a more minimal configuration with a longer warranty might be preferred. Suppose that the consumer has ranked the options and assigned a utility to each as summarized as follows:

Option	Variable	Utility
Slower processor	$X_{p1}$	0.20
Faster processor	$X_{p2}$	0.70
Fastest processor	$X_{p3}$	0.90
2-year warranty	$X_{w1}$	0.50
3-year warranty	$X_{w2}$	0.55
2 GB memory	$X_{m1}$	0.70
4 GB memory	$X_{m2}$	0.80
80 GB hard drive	$X_{d1}$	0.10
160 GB hard drive	$X_{d2}$	0.30
320 GB hard drive	$X_{d3}$	1.00
CD-ROM only	$X_{c1}$	0.00
CD/DVD burner	$X_{c2}$	0.95
CD/DVD dual layer	$X_{c3}$	0.45
8-cell battery	$X_b$	0.15
Video and photo software	$X_s$	0.85

The objective function would be to maximize utility:

$$\begin{aligned} \text{Maximize } & 0.20X_{p1} + 0.70X_{p2} + 0.90X_{p3} + 0.50X_{w1} + 0.55X_{w2} + 0.70X_{m1} + 0.80X_{m2} \\ & + 0.10X_{d1} + 0.30X_{d2} + 1.0X_{d3} + 0.0X_{c1} + 0.95X_{c2} + 0.45X_{c3} + 0.15X_b + 0.85X_s \end{aligned}$$

The first constraint is that the budget not be exceeded:

$$\begin{aligned} -179X_{p1} + 100X_{p2} + 300X_{p3} + 129X_{w1} + 269X_{w2} + 50X_{m1} + 500X_{m2} \\ -29X_{d1} + 39X_{d2} + 79X_{d3} - 39X_{c1} + 79X_{c2} + 179X_{c3} + 59X_b + 79X_s \leq 800 \end{aligned}$$

Because we defined the variables as choices over and above the base configuration, the budget constraint reflects the amount available over the base price. Next, for each group of options, at most one can be chosen:

$$\begin{aligned} X_{p1} + X_{p2} + X_{p3} &\leq 1 \quad (\text{processor}) \\ X_{w1} + X_{w2} &\leq 1 \quad (\text{warranty}) \\ X_{m1} + X_{m2} &\leq 1 \quad (\text{memory}) \\ X_{d1} + X_{d2} + X_{d3} &\leq 1 \quad (\text{hard drive}) \\ X_{c1} + X_{c2} + X_{c3} &\leq 1 \quad (\text{modular bay}) \\ X_b &\leq 1 \quad (\text{battery}) \\ X_s &\leq 1 \quad (\text{software}) \end{aligned}$$

Finally, the variables must be binary:

$$X_{ij} = 0 \text{ or } 1 \text{ for each } i \text{ and } j$$

Binary variables allow us to model a wide variety of logical constraints. For example, in the computer configuration example, there are often technical restrictions or recommendations that require a specific option if another one is chosen. For example, suppose that choosing a dual-layer CD/DVD burner requires at least 2 GB of memory. This means that if  $X_{c3} = 1$  then either  $X_{m1}$  or  $X_{m2}$  must be equal to 1. We can model this with the constraint:

$$X_{m1} + X_{m2} \geq X_{c3}$$

Note that if  $X_{c3} = 1$ , then we must have  $X_{m1} + X_{m2} \geq 1$ , which will force either  $X_{m1}$  or  $X_{m2}$  to be 1. However, if  $X_{c3} = 0$ , then  $X_{m1}$  or  $X_{m2}$  can assume any value.

As another example, suppose that if we choose the slowest processor, then we cannot choose the dual-layer CD/DVD burner. Thus, if  $X_{p1} = 1$ , then  $X_{c3}$  must be 0. This can be modeled by the constraint:

$$X_{c3} \leq 1 - X_{p1}$$

As a third example, suppose that the customer wants to ensure that if the enhanced video and photo software is chosen, then both the 4 GB memory and 320 GB hard drive should be chosen. In other words, if  $X_s = 1$ , we want to force  $X_{d3} = 1$  and  $X_{m2} = 1$ . We can model this in one of two ways. First, use two constraints:

$$\begin{aligned} X_{d3} &\geq X_s \\ X_{m2} &\geq X_s \end{aligned}$$

An alternative way of doing this is to use one constraint that is the sum of both of these:

$$X_{d3} + X_{m2} \geq 2X_s$$

Table 14.2 summarizes common types of logical conditions and how they can be modeled using binary variables.

**TABLE 14.2** Modeling Logical Conditions Using Binary Variables

Logical Condition	Constraint Model Form
If A, then B	$B \geq A$ or $B - A \geq 0$
If not A, then B	$B \geq 1 - A$ or $A + B \geq 1$
If A, then not B	$B \leq 1 - A$ or $B + A \leq 1$
At most one of A and B	$A + B \leq 1$
If A, then B and C	$(B \geq A \text{ and } C \geq A) \text{ or } B + C \geq 2A$
If A and B, then C	$C \geq A + B - 1 \text{ or } A + B - C \leq 1$

**SKILL-BUILDER EXERCISE 14.3**

Implement and solve the computer configuration model on a spreadsheet.

**A Supply Chain Facility Location Model**

In 1993, Procter & Gamble began an effort entitled Strengthening Global Effectiveness (SGE) to streamline work processes, drive out non-value-added costs, and eliminate duplication.<sup>1</sup> A principal component of SGE was the North American Product Supply Study, designed to reexamine and reengineer P&G's product-sourcing and distribution system for its North American operations, with an emphasis on plant consolidation. Prior to the study, the North American supply chain consisted of hundreds of suppliers, more than 50 product categories, more than 60 plants, 15 distribution centers (DCs), and more than 1,000 customers. The need to consolidate plants was driven by the move to global brands and common packaging and the need to reduce manufacturing expense, improve speed to market, avoid major capital investments, and deliver better consumer value.

One of the key submodels in the overall optimization effort was an integer optimization model to identify optimal distribution center (DC) locations in the supply chain and to assign customers to the DCs. Customers were aggregated into 150 zones. P&G had a policy of single sourcing—that is, each customer should be served by only one DC. The optimization model used in the analysis was:

$$\begin{aligned} \text{Min } & \sum C_{ij}X_{ij} \\ \sum X_{ij} &= 1, \text{ for every } j \text{ (summed over } i) \\ \sum Y_i &= k \text{ (summed over } i) \\ X_{ij} &\leq Y_i, \text{ for every } i \text{ and } j \end{aligned}$$

In this model,  $X_{ij} = 1$  if customer zone  $j$  is assigned to DC  $i$ , and 0 if not, and  $Y_i = 1$  if DC  $i$  is chosen from among a set of  $k$  potential locations. In the objective function,  $C_{ij}$  is the total cost of satisfying the demand in customer zone  $j$  from DC  $i$ . The first constraint ensures that each customer zone is assigned to exactly one DC. The next constraint limits the number of DCs to be selected from among the candidate set. The parameter  $k$  was varied by the analysis team to examine the effects of choosing different numbers of locations. The final constraint ensures that customer zone  $j$  cannot be assigned to DC  $i$  unless DC  $i$  is selected in the supply chain.

<sup>1</sup> Jeffrey D. Camm, Thomas E. Chorman, Franz A. Dill, James R. Evans, Dennis J. Sweeney, and Glenn W. Wegryn, "Blending OR/MS, Judgment, and GIS: Restructuring P&G's Supply Chain," *Interfaces*, 27, no. 1 (January–February, 1997), 128–142.

This model was used in conjunction with a simple transportation model for each of 30 product categories. Product-strategy teams used these models to specify plant locations and capacity options, and optimize the flow of product from plants to DCs and customers. In reconfiguring the supply chain, P&G realized annual cost savings of over \$250 million.

In the next example, we will develop a different optimization model without the single sourcing restriction for a similar type of supply chain design problem.

## MIXED INTEGER OPTIMIZATION MODELS

Many practical applications of optimization involve a combination of continuous variables and binary variables. This provides flexibility to model many different types of complex decision problems.

### Plant Location Model

Suppose that in the transportation model example discussed in Chapter 13, demand forecasts exceed the existing capacity and the company is considering adding a new plant from among two choices: Fayetteville, Arkansas, or Chico, California. Both plants would have a capacity of 1,500 units but only one can be built. Table 14.3 shows the revised data.

The company now faces two decisions. It must decide which plant to build, and then how to best ship the product from the plant to the distribution centers. Of course, one approach would be to solve two separate transportation models, one that includes the Fayetteville plant, and the other that includes the Chico plant. However, we will demonstrate how to answer both questions simultaneously, as this provides the most efficient approach, especially if the number of alternatives and combinations is larger than for this example.

Define a binary variable for the decision of which plant to build:  $Y_1 = 1$  if the Fayetteville plant is built, and  $Y_2 = 1$  if the Chico plant is built. The objective function now includes terms for the proposed plant locations:

$$\begin{aligned} \text{Minimize } & 12.60X_{11} + 14.35X_{12} + 11.52X_{13} + 17.58X_{14} + 9.75X_{21} + 16.26X_{22} \\ & + 8.11X_{23} + 17.92X_{24} + 10.41X_{31} + 11.54X_{32} + 9.87X_{33} + 11.64X_{34} \\ & + 13.88X_{41} + 16.95X_{42} + 12.51X_{43} + 8.32X_{44} \end{aligned}$$

Capacity constraints for the Marietta and Minneapolis plants remain as before. However, for Fayetteville and Chico, we can only allow shipping from those locations if a plant is built there. In other words, if we do not build a plant in Fayetteville ( $Y_1 = 0$ ), for example, then we must ensure that the amount shipped from Fayetteville

**TABLE 14.3** Plant Location Data

Plant	Distribution Center				Capacity
	Cleveland	Baltimore	Chicago	Phoenix	
Marietta	\$12.60	\$14.35	\$11.52	\$17.58	1,200
Minneapolis	\$9.75	\$16.26	\$8.11	\$17.92	800
Fayetteville	\$10.41	\$11.54	\$9.87	\$11.64	1,500
Chico	\$13.88	\$16.95	\$12.51	\$8.32	1,500
Demand	300	500	700	1,800	

to any distribution center must be 0, or  $X_{3j} = 0$  for  $j = 1$  to 4. To do this, we multiply the capacity by the binary variable corresponding to the location:

$$\begin{aligned} X_{11} + X_{12} + X_{13} + X_{14} &\leq 1,200 \\ X_{21} + X_{22} + X_{23} + X_{24} &\leq 800 \\ X_{31} + X_{32} + X_{33} + X_{34} &\leq 1,500Y_1 \\ X_{41} + X_{42} + X_{43} + X_{44} &\leq 1,500Y_2 \end{aligned}$$

Note that if the binary variable is 0, then the right-hand side of the constraint is 0, forcing all shipment variables to be 0 also. If, however, a particular  $Y$  variable is 1, then shipping up to the plant capacity is allowed. The demand constraints are the same as before, except that additional variables corresponding to the possible plant locations are added and new demand values are used:

$$\begin{aligned} X_{11} + X_{21} + X_{31} + X_{41} &= 300 \\ X_{12} + X_{22} + X_{32} + X_{42} &= 500 \\ X_{13} + X_{23} + X_{33} + X_{43} &= 700 \\ X_{14} + X_{24} + X_{34} + X_{44} &= 1,800 \end{aligned}$$

To guarantee that only one new plant is built, we must have:

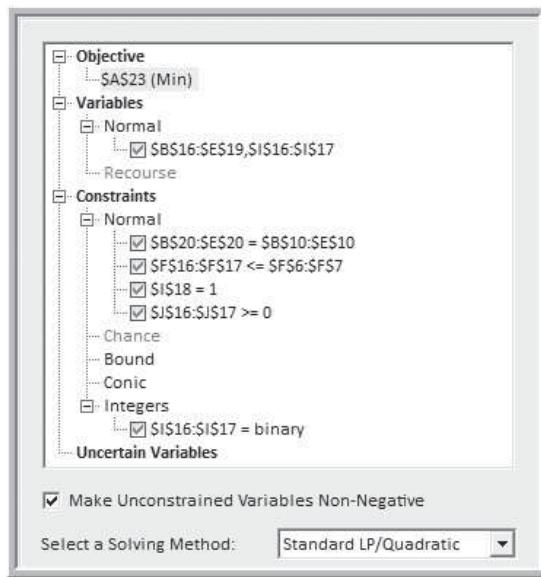
$$Y_1 + Y_2 = 1$$

Finally, we have nonnegativity for the continuous variables.  $X_{ij} \geq 0$ , for all  $i$  and  $j$ .

Figure 14.9 shows the spreadsheet model (Excel file *Plant Location Model*) and optimal solution. Note that in addition to the continuous variables  $X_{ij}$ , in the range B16:E19, we defined binary variables  $Y_i$  in cells I16 and I17. Cells J16 and J17 represent the constraint functions  $1,500Y_1 - X_{31} - X_{32} - X_{33} - X_{34}$  and  $1500Y_2 - X_{41} - X_{42} - X_{43} - X_{44}$ ,

	A	B	C	D	E	F	G	H	I	J
1	Plant Location Model									
2										
3	Data									
4										
5										
6	Plant	Cleveland	Baltimore	Chicago	Phoenix	Capacity				
7	Marietta	\$ 12.60	\$ 14.35	\$ 11.52	\$ 17.58		1200			
8	Minneapolis	\$ 9.75	\$ 16.26	\$ 8.11	\$ 17.92		800			
9	Fayetteville	\$ 10.41	\$ 11.54	\$ 9.87	\$ 11.64		1500			
10	Chico	\$ 13.88	\$ 16.95	\$ 12.51	\$ 8.32		1500			
11	Demand	300	500	700	1800					
12	Model									
13										
14	Amount Shipped									
15										
16	Plant	Cleveland	Baltimore	Chicago	Phoenix	Total shipped		New Plant Chosen	Surplus Capacity	
17	Marietta	200	500	0	300	1000	Fayetteville	0	0	
18	Minneapolis	100	0	700	0	800	Chico	1	0	
19	Fayetteville	0	0	0	0	0	Total	1		
20	Chico	0	0	0	1500	1500				
21	Demand met	300	500	700	1800					
22	Total cost									
23	\$	34,101								

FIGURE 14.9 Plant Location Model Spreadsheet



**FIGURE 14.10** Solver Model for Plant Location

respectively. These are restricted to be  $\geq 0$  to enforce the capacity constraints at the potential locations in the *Solver* model (Figure 14.10). You should closely examine the other constraints in the *Solver* model to verify that they are correct. The solution specifies selecting the Chico location. Models of this type are commonly used in supply chain design and other facility location applications.

## A Model with Fixed Costs

Many business problems involve fixed costs; they are either incurred in full, or not at all. Binary variables can be used to model such problems. To illustrate this, consider the multiperiod production-inventory planning model for Kristin's Kreations that we developed in Chapter 13. Suppose that Kristin must rent some equipment to produce her products, which costs \$65 for three months. The equipment can be rented or returned each quarter, so if nothing is produced in a quarter, it makes no sense to incur the rental cost.

The fixed costs can be incorporated into the model by defining an additional set of variables:

$$Y_A = 1 \text{ if production occurs during the autumn, and } 0 \text{ if not}$$

$$Y_W = 1 \text{ if production occurs during the winter, and } 0 \text{ if not}$$

$$Y_S = 1 \text{ if production occurs during the spring, and } 0 \text{ if not}$$

Then the objective function becomes:

$$\text{Minimize } 11P_A + 14P_W + 12.50P_S + 1.20I_A + 1.20I_W + 1.20I_S + 65(Y_A + Y_W + Y_S)$$

The basic material balance equations are the same:

$$P_A - I_A = 150$$

$$P_W + I_A - I_W = 400$$

$$P_S + I_W - I_S = 50$$

However, we must ensure that whenever a production variable,  $P$ , is positive, that the corresponding  $Y$  variable is equal to 1; and conversely, if the  $Y$  variable is 0 (you don't

rent the equipment), then the corresponding production variable must also be 0. This can be accomplished with the following constraints:

$$P_A \leq 600Y_A$$

$$P_W \leq 600Y_W$$

$$P_S \leq 600Y_S$$

Note that if any  $Y$  is 0 in a solution, then  $P$  is forced to be 0, and if  $P$  is positive, then  $Y$  must be 1. Because we don't know how much the value of any production variable will be, we use 600, which is the sum of the demands over the time horizon to multiply by  $Y$ . So when  $Y$  is 1, any amount up to 600 units can be produced. Actually any large number can be used, so long as it doesn't restrict the possible values of  $P$ . Generally, the smallest value should be used for efficiency. Finally,  $P_A$ ,  $P_W$ , and  $P_S$  must be nonnegative, and  $Y_A$ ,  $Y_W$ , and  $Y_S$  are binary. Figure 14.11 shows a spreadsheet implementation for this model with the optimal solution (Excel file *Kristin's Kreations Fixed Cost Model*).

	A	B	C	D
<b>1 Kristin's Kreations Fixed Cost Model</b>				
2				
3 Cost		Quarter 1	Quarter 2	Quarter 3
4 Production	\$ 11.00	\$ 14.00	\$ 12.50	
5 Inventory	\$ 1.20	\$ 1.20	\$ 1.20	
6 Demand	150	400	50	
7 Fixed cost	\$ 65.00	\$ 65.00	\$ 65.00	
8				
9		Quarter 1	Quarter 2	Quarter 3
10 Production	600	0	0	
11 Inventory	450	50	0	
12 Binary	1	0	0	
13				
14 Binary constraints	600	0	0	
15 Net production	150	400	50	
16				
17		Cost		
18 Total	\$ 7,265.00			

	A	B	C	D
<b>1 Kristin's Kreations</b>				
2				
3 Cost		Quarter 1	Quarter 2	Quarter 3
4 Production	11		14	12.5
5 Inventory	1.2		1.2	1.2
6 Demand	150		400	50
7 Fixed cost	65		65	65
8				
9		Quarter 1	Quarter 2	Quarter 3
10 Production	600		0	0
11 Inventory	450		50	0
12 Binary	1		0	0
13				
14 Binary constraints	=600*B12		=600*C12	=600*D12
15 Net production	=B10-B11		=C10-C11+B11	=D10-D11+C11
16				
17		Cost		
18 Total	=SUMPRODUCT(B4:D5,B10:D11) + 65*(B12+C12+D12)			

**FIGURE 14.11** Spreadsheet Model for Kristin's Kreations Fixed Cost Problem

You might observe that this model does not preclude feasible solutions in which a production variable is 0 while its corresponding  $Y$  variable is 1. This implies that we incur the fixed cost even though no production is incurred during that time period. While such a solution is feasible, it can never be optimal, as a lower cost could be obtained by setting the  $Y$  variable to 0 without affecting the value of the production variable, and the solution algorithm will always ensure this.

## NONLINEAR OPTIMIZATION

In many situations, the relationship among variables in a model is not linear. Whenever either the objective function or a constraint is not linear, the model becomes a *nonlinear optimization problem*, requiring different solution techniques. Nonlinear models do not have a common structure as do linear models, making it more difficult to develop appropriate models. We present two examples of nonlinear optimization models in business.

### Hotel Pricing

The Marquis Hotel is considering a major remodeling effort and needs to determine the best combination of rates and room sizes to maximize revenues. Currently, the hotel has 450 rooms with the following history:

Room Type	Rate	Daily Avg. No. Sold	Revenue
Standard	\$85	250	\$21,250
Gold	\$98	100	\$9,800
Platinum	\$139	50	\$6,950
Total revenue \$38,000			

Each market segment has its own price/demand elasticity. Estimates are:

Room Type	Price Elasticity of Demand
Standard	-1.5
Gold	-2.0
Platinum	-1.0

This means, for example, that a 1% decrease in the price of a standard room will increase the number of rooms sold by 1.5%. Similarly, a 1% increase in the price will decrease the number of rooms sold by 1.5%. For any pricing structure (in \$), the projected number of rooms of a given type sold (we will allow continuous values for this example) can be found using the formula:

$$(\text{Historical average number of rooms sold}) + (\text{Elasticity})(\text{New price} - \text{Current price})(\text{Historical average number of rooms sold}) / (\text{Current price})$$

The hotel owners want to keep the price of a standard room between \$70 and \$90; a gold room between \$90 and \$110; and a platinum room between \$120 and \$149. Define  $S$  = price of a standard room,  $G$  = price of a gold room, and  $P$  = price of a platinum room. Thus, for standard rooms, the projected number of rooms sold is

$250 - 1.5(S - 85)(250)/85 = 625 - 4.41176S$ . The objective is to set the room prices to maximize total revenue. Total revenue would equal the price times the projected number of rooms sold, summed over all three types of rooms. Therefore, total revenue would be:

$$\begin{aligned}\text{Total revenue} &= S(625 - 4.41176S) + G(300 - 2.04082G) + P(100 - 0.35971P) \\ &= 625S + 300G + 100P - 4.41176S^2 - 2.04082G^2 - 0.35971P^2\end{aligned}$$

To keep prices within the stated ranges, we need constraints:

$$70 \leq S \leq 90$$

$$90 \leq G \leq 110$$

$$120 \leq P \leq 149$$

Finally, although the rooms may be renovated, there are no plans to expand beyond the current 450-room capacity. Thus, the projected number of total rooms sold cannot exceed 450:

$$(625 - 4.41176S) + (300 - 2.04082G) + (100 - 0.35971P) \leq 450$$

or simplified as:

$$1025 - 4.41176S - 2.04082G - 0.35971P \leq 450$$

The complete model is:

$$\text{Maximize } 625S + 300G + 100P - 4.41176S^2 - 2.04082G^2 - 0.35971P^2$$

$$70 \leq S \leq 90$$

$$90 \leq G \leq 110$$

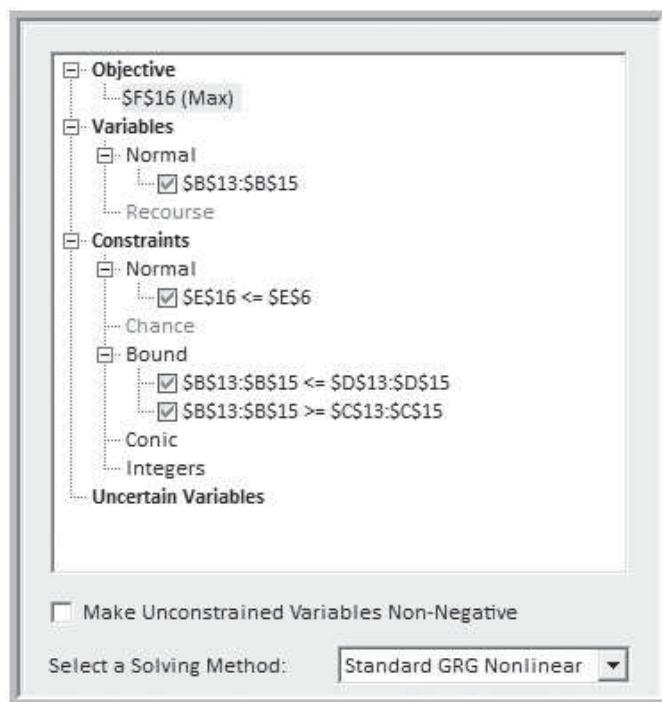
$$120 \leq P \leq 149$$

$$1025 - 4.41176S - 2.04082G - 0.35971P \leq 450$$

Figure 14.12 shows a spreadsheet model (Excel file *Hotel Pricing Model*) for this example showing the optimal solution. The decision variables, the new prices to charge,

	A	B	C	D	E	F
1	Marquis Hotel					
2						
3	Data					
4		Current	Average		Total Room	
5	Room type	Rate	Daily Sold	Elasticity	Capacity	
6	Standard	\$ 85.00	250	-1.5	450	
7	Gold	\$ 98.00	100	-2		
8	Platinum	\$ 139.00	50	-1		
9						
10	Model				Projected	
11					Rooms	Projected
12	Room type	New Price	Price Range	Sold	Revenue	
13	Standard	\$ 76.87	\$ 70.00 \$ 90.00	286	\$ 21,974.39	
14	Gold	\$ 90.00	\$ 90.00 \$ 110.00	116	\$ 10,469.39	
15	Platinum	\$ 145.04	\$ 120.00 \$ 149.00	48	\$ 6,936.87	
16			Totals	450	\$ 39,380.65	

FIGURE 14.12 Hotel Pricing Model Spreadsheet and Optimal Solution



**FIGURE 14.13** Solver Model for Hotel Pricing Example

are given in cells B13:B15. The projected numbers of rooms sold are computed in cells E13:E15 using the preceding formula. By multiplying the number of rooms sold by the new price for each room type, the projected revenue is calculated, as given in cells F13:F15. The total revenue in cell F16 represents the objective function. The constraints—(1) the new price must fall within the allowable price range and (2) the total projected number of rooms sold must not exceed 450—can be expressed within the spreadsheet model as:

$$\begin{aligned} B13:B15 &\geq C13:C15 \\ B13:B15 &\leq D13:D15 \end{aligned}$$

and

$$E16 \leq E6$$

Note that it is easier to formulate this model more as a financial spreadsheet than to enter the analytical formulas as they were developed. The optimal prices predict a demand for all 450 rooms with a total revenue of \$39,380.65.

### Solving Nonlinear Optimization Models

Nonlinear optimization models are formulated with *Solver* in the same fashion as linear or integer models. In *Premium Solver*, you should select *Standard GRG Nonlinear* as the solution procedure. “GRG” stands for “generalized reduced gradient,” which is the name of the algorithm used to solve nonlinear models in *Solver*. Figure 14.13 shows the premium version *Solver Parameters* dialog box for the hotel pricing model. The optimal solution was shown in Figure 14.11.

The information contained in the Answer Report (Figure 14.14) is the same as for linear models. Because nonlinear models are continuous (as long as they do not contain any integer variables), *Solver* produces a sensitivity report. However, for nonlinear models, the Sensitivity Report (Figure 14.15) is quite different. In the *Adjustable Cells*

	A	B	C	D	E	F	G
11							
12			Target Cell (Max)				
13	Cell	Name	Original Value	Final Value			
14	\$F\$16	Totals Revenue	0	39380.65104			
15							
16			Adjustable Cells				
17	Cell	Name	Original Value	Final Value			
18	\$B\$13	Standard New Price	\$ -	\$ 76.87			
19	\$B\$14	Gold New Price	\$ -	\$ 90.00			
20	\$B\$15	Platinum New Price	\$ -	\$ 145.04			
21							
22			Constraints				
23	Cell	Name	Cell Value	Formula	Status	Slack	
24	\$E\$16	Totals Sold	450.0000004	\$E\$16<= \$E\$6	Binding	0	
25	\$B\$13	Standard New Price	\$ 76.87	\$B\$13>= \$C\$13	Not Binding	6.87476046	
26	\$B\$14	Gold New Price	\$ 90.00	\$B\$14>= \$C\$14	Binding	0	
27	\$B\$15	Platinum New Price	\$ 145.04	\$B\$15>= \$C\$15	Not Binding	25.0414271	
28	\$B\$13	Standard New Price	\$ 76.87	\$B\$13<= \$D\$13	Not Binding	13.1252395	
29	\$B\$14	Gold New Price	\$ 90.00	\$B\$14<= \$D\$14	Not Binding	20	
30	\$B\$15	Platinum New Price	\$ 145.04	\$B\$15<= \$D\$15	Not Binding	3.95857289	
31							

**FIGURE 14.14** Hotel Pricing Example Answer Report

	A	B	C	D	E
4					
5			Target Cell (Max)		
6	Cell	Name	Final Value		
7	\$F\$16	Totals Revenue	39380.65104		
8					
9			Adjustable Cells		
10				Final	Reduced
11	Cell	Name	Value	Gradient	
12	\$B\$13	Standard New Price	\$ 76.87	\$ -	
13	\$B\$14	Gold New Price	\$ 90.00	\$ (42.69)	
14	\$B\$15	Platinum New Price	\$ 145.04	\$ -	
15					
16			Constraints		
17				Final	Lagrange
18	Cell	Name	Value	Multiplier	
19	\$E\$16	Totals Sold	450.0000004	12.08293216	

**FIGURE 14.15** Hotel Pricing Example Sensitivity Report

section, the *Reduced Gradient* is analogous to the *Reduced Cost* in linear models. For this problem, however, the objective function coefficient of each price depends on many parameters, and therefore, the reduced gradient is more difficult to interpret in relation to the problem data. *Lagrange Multipliers* in the *Constraints* section are similar to shadow prices for linear models. For nonlinear models, the Lagrange multipliers give the *approximate* rate of change in the objective function as the right-hand side of a binding constraint is increased by one unit. Thus, for the hotel pricing problem, if the number of available rooms is increased by 1 to 451, the total revenue would increase by *approximately* \$12.08. For linear models, shadow prices give the *exact* rate of change within the Allowable Increase and Decrease limits. Thus, you should be somewhat cautious when interpreting these values and will need to re-solve the models to find the true effect of changes to constraints.

Many nonlinear problems are notoriously difficult to solve. *Solver* cannot guarantee finding the absolute best solution (called a *global optimal solution*) for all problems. A *local optimum solution* is one for which all points close by are no better than the solution (think of the analogy of being at the top of a hill when the highest peak is on another mountain). The solution found often depends much on the starting solution in your spreadsheet. For complex problems, it is wise to run *Solver* from different starting points. You should also look carefully at the *Solver* results dialog box when the model has completed running. If it indicates “Solver has found a solution. All constraints and optimality conditions are satisfied,” then at least a local optimal solution has been found. If you get the message: “Solver has converged to the current solution. All constraints are satisfied.” then you should run *Solver* again from the current solution to try to find a better solution.

## Markowitz Portfolio Model

The Markowitz portfolio model<sup>2</sup> is a classic optimization model in finance that seeks to minimize the risk of a portfolio of stocks subject to a constraint on the portfolio’s expected return. For example, suppose an investor is considering three stocks. The expected return for stock 1 is 10%; for stock 2, 12%; and for stock 3, 7%; and she would like an expected return of at least 10%. Clearly one option is to invest everything in stock 1; however, this may not be a good idea as the risk might be too high. Recall from Chapter 10 that we can measure risk by the standard deviation, or equivalently, the variance. Research has found the variance–covariance matrix of the individual stocks to be:

	<b>Stock 1</b>	<b>Stock 2</b>	<b>Stock 3</b>
<b>Stock 1</b>	0.025	0.015	-0.002
<b>Stock 2</b>		0.030	0.005
<b>Stock 3</b>			0.004

Thus, the decision variables are the percentage of each stock to allocate to the portfolio. (You might be familiar with the term “asset allocation model” that many financial investment companies suggest to their clients; for example, “maintain 60% equities, 30% bonds, and 10% cash.”) Define  $x_j$  to be the fraction of the portfolio to invest in stock  $j$ .

The objective function is to minimize the risk of the portfolio as measured by its variance. Because stock prices are correlated with one another, the variance of the portfolio must reflect not only variances of the stocks in the portfolio but also the covariance between stocks. The variance of a portfolio is the weighted sum of the variances and covariances:

$$\text{Variance of Portfolio} = \sum_{i=1}^k s_i^2 x_i^2 + \sum_{i=1}^k \sum_{j>1} 2s_{ij}x_i x_j$$

where

$s_i^2$  = the sample variance in the return of stock  $i$

$s_{ij}$  = the sample covariance between stocks  $i$  and  $j$

Using the preceding data, the objective function is:

$$\begin{aligned} \text{Minimize Variance} &= 0.025x_1^2 + 0.030x_2^2 + 0.004x_3^2 + 2(0.015)x_1 x_2 \\ &\quad + 2(-0.002)x_1 x_3 + 2(0.005)x_2 x_3 \end{aligned}$$

---

<sup>2</sup>H.M. Markowitz, *Portfolio Selection, Efficient Diversification of Investments* (New York: John Wiley & Sons, 1959).

The constraints must first ensure that we invest 100% of our budget. Because the variables are defined as fractions, we must have:

$$x_1 + x_2 + x_3 = 1$$

Second, the portfolio must have an expected return of at least 10%. The return on a portfolio is simply the weighted sum of the returns of the stocks in the portfolio. This results in the constraint:

$$10x_1 + 12x_2 + 7x_3 \geq 10$$

Finally, we cannot invest negative amounts:

$$x_1, x_2, x_3 \geq 0$$

The complete model is:

$$\begin{aligned} \text{Minimize Variance} &= 0.025x_1^2 + 0.030x_2^2 + 0.004x_3^2 + 0.03x_1x_2 \\ &\quad - 0.004x_1x_3 + 0.010x_2x_3 \\ &\quad x_1 + x_2 + x_3 = 1 \\ &\quad 10x_1 + 12x_2 + 7x_3 \geq 10 \\ &\quad x_1, x_2, x_3 \geq 0 \end{aligned}$$

Figure 14.16 shows a spreadsheet model for this example (Excel file *Markowitz Model*). The decision variables (fraction of each stock in the portfolio) are entered in cells B14:B16. The expected return and variance of the portfolio are computed in cells B20 and C20. The variance of the optimal portfolio is 0.012.

The *Solver Sensitivity Report* is shown in Figure 14.17. As we noted, for nonlinear models, the Lagrange multipliers are only approximate indicators of shadow prices. For this solution, the Lagrange multiplier predicts that the minimum variance will increase by 63.2% if the target return is increased from 10% to 11%. If you re-solve the model, you will find that the minimum variance increases to 0.020, a 66.67% increase.

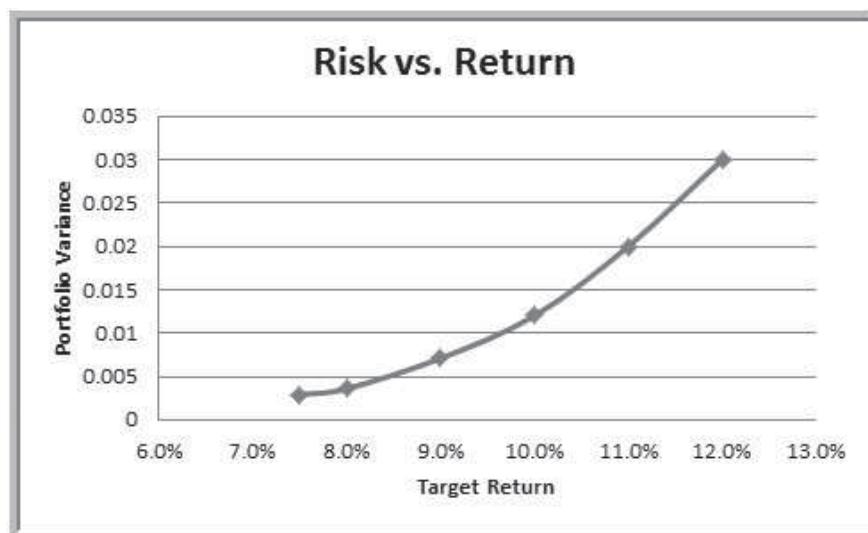
A	B	C	D	E	F	G
1	Markowitz Model					
2						
3	Data					
4		Expected			Variance-Covariance Matrix	
5		Return		Stock 1	Stock 2	Stock 3
6	Stock 1	10%	Stock 1	0.025	0.015	-0.002
7	Stock 2	12%	Stock 2		0.03	0.005
8	Stock 3	7%	Stock 3			0.004
9	Target Return	10%				
10						
11	Model				Variance Calculations	
12					Squared Terms	Cross-Products
13		Allocation				
14	Stock 1	0.25		0.001579256	0.003387	
15	Stock 2	0.45		0.006053361	-0.000301067	
16	Stock 3	0.30		0.000358718	0.001345191	
17	Total	1				
18						
19						
20		Return	Variance			
21	Portfolio	10.0%	0.012			

FIGURE 14.16 *Markowitz Model Spreadsheet*

A	B	C	D	E
4				
5	Objective Cell (Min)			
6	Cell	Name	Final Value	
7	\$C\$21	Portfolio Variance	0.01242246	
8				
9	Decision Variable Cells			
10			Final	Reduced
11	Cell	Name	Value	Gradient
12	\$B\$14	Stock 1 Allocation	0.25	0.00
13	\$B\$15	Stock 2 Allocation	0.45	0.00
14	\$B\$16	Stock 3 Allocation	0.30	0.00
15				
16	Constraints			
17			Final	Lagrange
18	Cell	Name	Value	Multiplier
19	\$B\$17	Total Allocation	1	-0.038363637
20	\$B\$21	Portfolio Return	10.0%	63.2%

**FIGURE 14.17** Markowitz Model Sensitivity Report

Optimization models can and should provide decision makers with valuable insight beyond simply finding an optimal solution. Using spreadsheet models and *Solver*, it is easy to systematically vary a parameter of a model and investigate its impact on the solution. For example, we might be interested in understanding the relationship between the minimum risk and the target return. By changing the target return, and re-solving the model, we obtain the chart shown in Figure 14.18 (advanced users of Excel can program a macro to automate this process). This clearly shows that the minimum variance increases at a faster rate as the target return increases, indicating that the investor faces a risk premium in seeking higher returns.



**FIGURE 14.18** Risk vs. Return Profile for Markowitz Portfolio Example

An alternative modeling approach would be to maximize the return subject to a constraint on risk. For example, suppose the investor wants to maximize expected return subject to a risk (variance) no greater than 1%. This form of the model would be:

$$\begin{aligned}
 & \text{Maximize } 10x_1 + 12x_2 + 7x_3 \\
 & x_1 + x_2 + x_3 = 1 \\
 & 0.025x_1^2 + 0.030x_2^2 + 0.004x_3^2 + 0.03x_1x_2 - 0.004x_1x_3 + 0.010x_2x_3 \leq 0.01 \\
 & x_1, x_2, x_3 \geq 0
 \end{aligned}$$

In this case, we would have a linear objective function and a mixture of linear and non-linear constraints.

#### SKILL-BUILDER EXERCISE 14.4

For the Markowitz portfolio example, solve models for required returns from 7% to 12% and develop a stacked bar chart showing the proportion of each stock in the portfolio. How would you explain this to an investor?

### **EVOLUTIONARY SOLVER FOR NONSMOOTH OPTIMIZATION**

In the fixed-cost problem for Kristin's Kreations, we needed binary variables in order to incorporate the fixed costs into the objective function and model the logical conditions that ensure that the fixed cost will only be incurred if the production in a period is positive. We did this for an important reason—to preserve linearity of the constraints so that we could use *Solver*'s linear optimization method to find an optimal solution. However, it might have been simpler to use an IF function in the spreadsheet to model the fixed costs. For example, we could express the objective function as:

$$\begin{aligned}
 & \text{Minimize } 11P_A + 14P_W + 12.50P_S + 1.20I_A + 1.20I_W + 1.20I_S + \text{IF}(P_A > 0, 65, 0) \\
 & \quad + \text{IF}(P_W > 0, 65, 0) + \text{IF}(P_S > 0, 65, 0)
 \end{aligned}$$

The only constraints needed are the material balance constraints:

$$\begin{aligned}
 P_A - I_A &= 150 \\
 P_W + I_A - I_W &= 400 \\
 P_S + I_W - I_S &= 50
 \end{aligned}$$

In this way, there is no need for the binary variables and the additional constraints that involve them. Doing so, however, results in a “nonsmooth” model that violates the linearity conditions required for the linear optimization solution method used by *Solver*. As we noted in Chapter 13, such Excel functions as IF, ABS, MIN, and MAX lead to nonsmooth models.

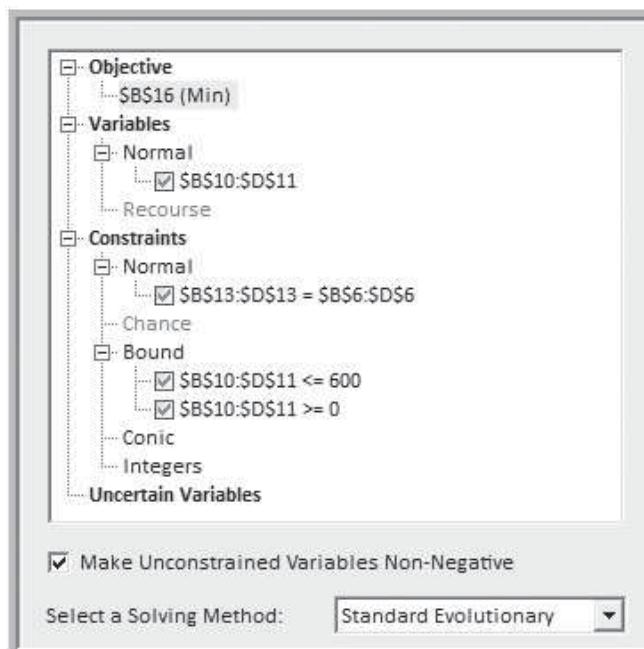
Problems that are nonsmooth or involve both nonlinear functions and integer variables are usually difficult to solve using conventional techniques. To overcome these limitations, new approaches called *metaheuristics* have been developed by researchers. These approaches have some exotic names, including genetic algorithms, neural networks, and tabu search. Such approaches use heuristics—intelligent rules for systematically searching among solutions—that remember the best solutions they find, then

modifying or combining them in attempting to find better solutions. *Solver's Standard Evolutionary* algorithm uses such an approach. The evolutionary algorithm is available in both the standard Excel-supplied *Solver* as well as in the premium version. We will illustrate *Evolutionary Solver* first using the new model for Kristin's Kreations, and then with some other applications.

Figure 14.19 shows the modified spreadsheet for Kristin's Kreations (Excel file *Kristin's Kreations Evolutionary Solver Model*); the objective function in cell B16 is =SUM PRODUCT(B4:D5,B10:D11)+IF(B10>0,B7,0)+IF(C10>0,C7,0)+IF(D10>0,D7,0). Figure 14.20 shows the *Solver* model. The *Evolutionary Solver* algorithm requires that all variables

	A	B	C	D
1	Kristin's Kreations Evolutionary Solver Model			
2				
3	Cost	Quarter 1	Quarter 2	Quarter 3
4	Production	\$ 11.00	\$ 14.00	\$ 12.50
5	Inventory	\$ 1.20	\$ 1.20	\$ 1.20
6	Demand	150	400	50
7	Fixed cost	\$ 65.00	\$ 65.00	\$ 65.00
8				
9		Quarter 1	Quarter 2	Quarter 3
10	Production	550	0	50
11	Inventory	400	0	0
12				
13	Net production	150	400	50
14				
15		Cost		
16	Total	\$ 7,285.00		

**FIGURE 14.19** Modified Spreadsheet for Kristin's Kreations



**FIGURE 14.20** Evolutionary Solver Model for Kristin's Kreations

have simple upper and lower bounds to restrict the search space to a manageable region. Thus, we set upper bounds of 600 (the total demand) and lower bounds of zero for each of them. *Evolutionary Solver* finds the same optimal solution as the integer optimization problem we solved earlier.

### Rectilinear Location Model

Edwards Manufacturing is studying where to locate a tool bin on the factory floor. The locations of five production cells are expressed as  $x$ - and  $y$ -coordinates on a rectangular grid of the factory layout. The daily demand for tools (measured as the number of trips to the tool bin) at each production cell is also known. The relevant data are:

Cell	X-coordinate	Y-coordinate	Demand
Fabrication	1	4	12
Paint	1	2	24
Subassembly 1	2.5	2	13
Subassembly 2	3	5	7
Assembly	4	4	17

Because of the nature of the equipment layout in the factory and for safety reasons, workers must travel along marked horizontal and vertical aisles to access the tool bin. Thus, the distance from a cell to the tool bin cannot be measured as a straight line; rather it must be measured as rectilinear distance. Using rectilinear distance measure, the distance between coordinates  $(x, y)$  and  $(a, b)$  is absolute value of  $(x - a)$  plus the absolute value of  $(y - b)$ . The optimal location should minimize the total weighted distance between the tool bin and all production cells, where the weights are the daily number of trips to the tool bin.

To formulate an optimization model for the best location, define  $(X, Y)$  as the location coordinates of the tool bin. The weighted distance between the tool bin and each cell is expressed by the objective function:

$$\text{Minimize } 12(|X - 1| + |Y - 4|) + 24(|X - 1| + |Y - 2|) + 13(|X - 2.5| + |Y - 2|) + 7(|X - 3| + |Y - 5|) + 17(|X - 4| + |Y - 4|)$$

The absolute value functions used in this objective function create a nonsmooth model. Thus, *Evolutionary Solver* is an appropriate solution technique.

Figure 14.21 shows a spreadsheet model for the Edwards Manufacturing example (Excel file *Edwards Manufacturing Model*). The upper bounds are chosen as the maximum coordinate values and the lower bounds are zero. The *Solver* model is shown in Figure 14.22.

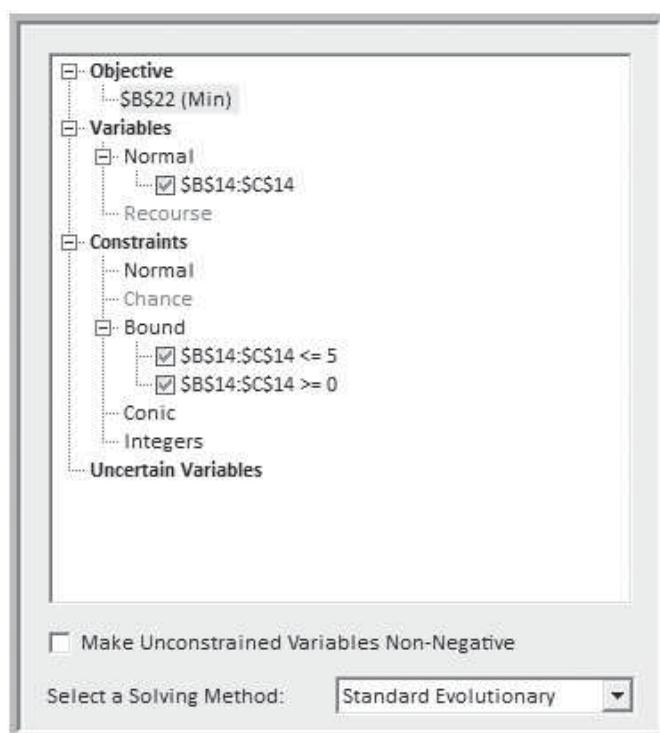
The results obtained by *Evolutionary Solver* depend heavily on the starting values of the decision variables and the amount of time devoted to the search. Different starting values can produce different results for complicated models, and increasing the maximum search time may improve the solution. Thus, for complex problems, it is wise to run the procedure from different starting points. The maximum search time and other parameters can be changed from the *Options* button in the dialog; however, this is usually only necessary for advanced users.

#### SKILL-BUILDER EXERCISE 14.5

Suppose that the tool bin location must be an integer coordinate. Modify the *Evolutionary Solver* model to find the best location.

	A	B	C	D
1	Edwards Manufacturing			
2				
3	Data			
4				
5	Cell	x-coordinate	y-coordinate	Demand
6	Fabrication	1	4	12
7	Paint	1	2	24
8	Subassembly 1	2.5	2	13
9	Subassembly 2	3	5	7
10	Assembly	4	4	17
11	Maximum	4	5	
12				
13	Model			
14	Tool bin location	2.499997179	2.489551412	
15				
16	Cell	Weighted Distance		
17	Fabrication	36.1253492		
18	Paint	47.7491662		
19	Subassembly 1	6.364205031		
20	Subassembly 2	21.07315986		
21	Assembly	51.17767394		
22	Total	162.4895542		

**FIGURE 14.21** Edwards Manufacturing Spreadsheet



**FIGURE 14.22** Evolutionary Solver Model for Edwards Manufacturing

## Job Sequencing

A rather unique application of Excel modeling and *Evolutionary Solver* is for job sequencing problems. Job sequencing problems involve finding an optimal sequence, or order, by which to process a set of jobs. For example, suppose that a custom manufacturing

company has ten jobs waiting to be processed. Each job  $i$  has an estimated processing time ( $P_i$ ) and a due date ( $D_i$ ) that was requested by the customer, as shown in the table below:

Job	1	2	3	4	5	6	7	8	9	10
Time	8	7	6	4	10	8	10	5	9	5
Due date	20	27	39	28	23	40	25	35	29	30

For any job sequence, we may compute the completion time by successively adding the processing times. We may then compare the completion times with the requested due dates to determine if the job is either completed early or late. *Lateness* ( $L_i$ ) is the difference between the completion time and the due date (either positive or negative). *Tardiness* ( $T_i$ ) is the amount of time by which the completion time exceeds the due date; thus tardiness is zero if a job is completed early. Thus, for job  $i$ ,

$$L_i = C_i - D_i$$

$$T_i = \text{Max}\{0, L_i\}$$

Researchers have shown that sequencing jobs in order of shortest processing time (SPT) first will minimize the average completion time for all jobs. Sequencing by earliest due date (EDD) first will minimize the maximum number of tardy jobs. However, the manager might be interested in minimizing other criteria, such as the average tardiness, total tardiness, or total lateness. It is possible to develop integer optimization models for such problems, but there are quite complicated and can take a very long time to solve.

To develop a spreadsheet model for this problem, we use the Excel function INDEX to identify the processing times and due data for the job assigned to a particular sequence. Figure 14.23 shows the model (Excel file, *Job Sequencing Model*). A particular

	A	B	C	D	E	F	G	H	I	J	K
1	Job Sequencing										
2											
3	Data										
4	Job	1	2	3	4	5	6	7	8	9	10
5	Time	8	7	6	4	10	8	10	5	9	5
6	Due date	26	27	39	28	23	40	25	35	29	30
7											
8	Model										
9	Sequence	1	2	3	4	5	6	7	8	9	10
10	Job Assigned	5	7	1	2	4	9	10	8	3	6
11	Processing time	10	10	8	7	4	9	5	5	6	8
12	Completion time	10	20	28	35	39	48	53	58	64	72
13	Due Date	23	25	26	27	28	29	30	35	39	40
14	Lateness	-13	-5	2	8	11	19	23	23	25	32
15	Tardiness	0	0	2	8	11	19	23	23	25	32
16											
17	Average Completion Time	42.7									
18	Maximum Number Tardy	8									
19	Total Lateness	125									
20	Average Lateness	12.5									
21	Variance of Lateness	188.85									
22	Total Tardiness	143									
23	Average Tardiness	14.3									
24	Variance of Tardiness	121.21									

FIGURE 14.23 Spreadsheet Model for Job Sequencing

job sequence (the decision variables) is given in row 12; for this example, we show the sequence for the EDD rule. In rows 13 and 15 we use the INDEX function to identify the processing time and due date associated with a specific job. For example, the formula in cell B13 is =INDEX(\$B\$5:\$K\$7,2,B12). This function references the value in the second row and column of the range B5:K7 corresponding to the job assigned to cell B12; in this case, job 5. Likewise, the formula in cell B15, =INDEX(\$B\$5:\$K\$7,3,B12), finds the due date associated with job 5.

Any sequence of integers in the decision variable range is called a *permutation*. Our goal is to find a permutation that optimizes the chosen criteria. Solver has an option to define decision variables as a permutation; this is called an *alldifferent* constraint. To do this, click on *Integers* in the *Solver Parameters* dialog, choose the range of the decision variables, and then choose *dif* from the drop-down box as shown in Figure 14.24. The final model, shown in Figure 14.25, is quite simple: minimize the chosen objective

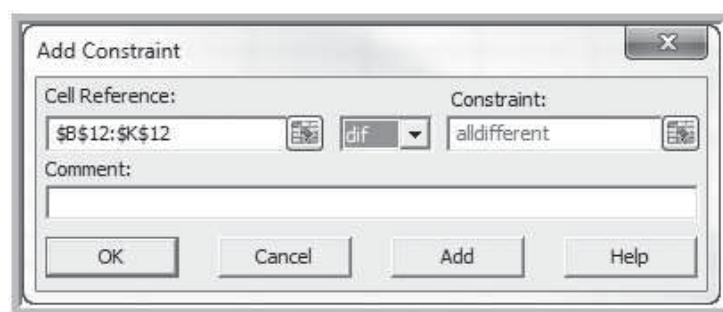


FIGURE 14.24 Solver *alldifferent* Constraint Definition

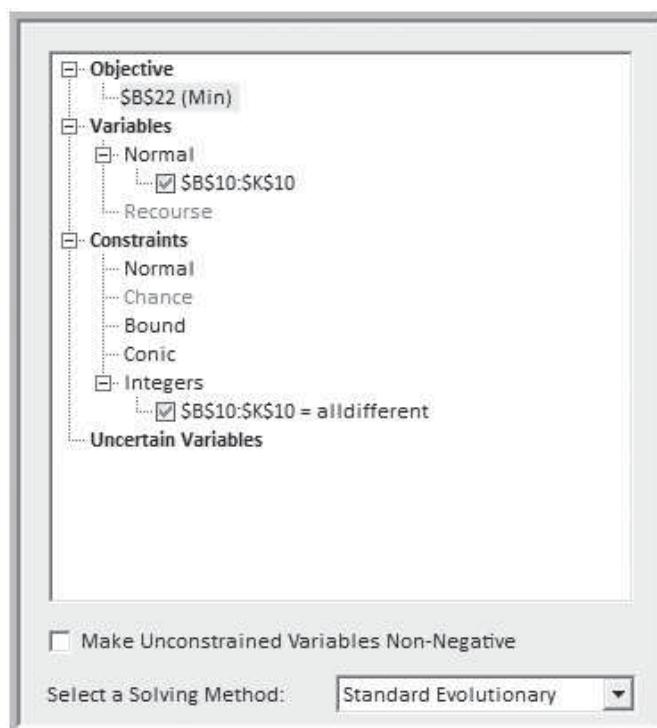


FIGURE 14.25 Solver Model for Job Sequencing to Minimize Total Tardiness

A	B	C	D	E	F	G	H	I	J	K
1 Job Sequencing										
2										
3 Data										
4	Job	1	2	3	4	5	6	7	8	9
5	Time	8	7	6	4	10	8	10	5	9
6	Due date	26	27	39	28	23	40	25	35	29
7										
8 Model										
9	Sequence	1	2	3	4	5	6	7	8	9
10	Job Assigned	2	5	1	4	10	8	3	6	9
11	Processing time	7	10	8	4	5	5	6	8	9
12	Completion time	7	17	25	29	34	39	45	53	62
13	Due Date	27	23	26	28	30	35	39	40	29
14	Lateness	-20	-6	-1	1	4	4	6	13	33
15	Tardiness	0	0	0	1	4	4	6	13	33
16										
17	Average Completion Time	38.3								
18	Maximum Number Tardy	7								
19	Total Lateness	81								
20	Average Lateness	8.1								
21	Variance of Lateness	331.69								
22	Total Tardiness	108								
23	Average Tardiness	10.8								
24	Variance of Tardiness	236.96								

**FIGURE 14.26** Evolutionary Solver Minimum Total Tardiness Solution

cell—in this case, total tardiness—and ensure that the decision variables are a valid permutation of the job numbers. Figure 14.26 shows the *Solver* solution.

### SKILL-BUILDER EXERCISE 14.6

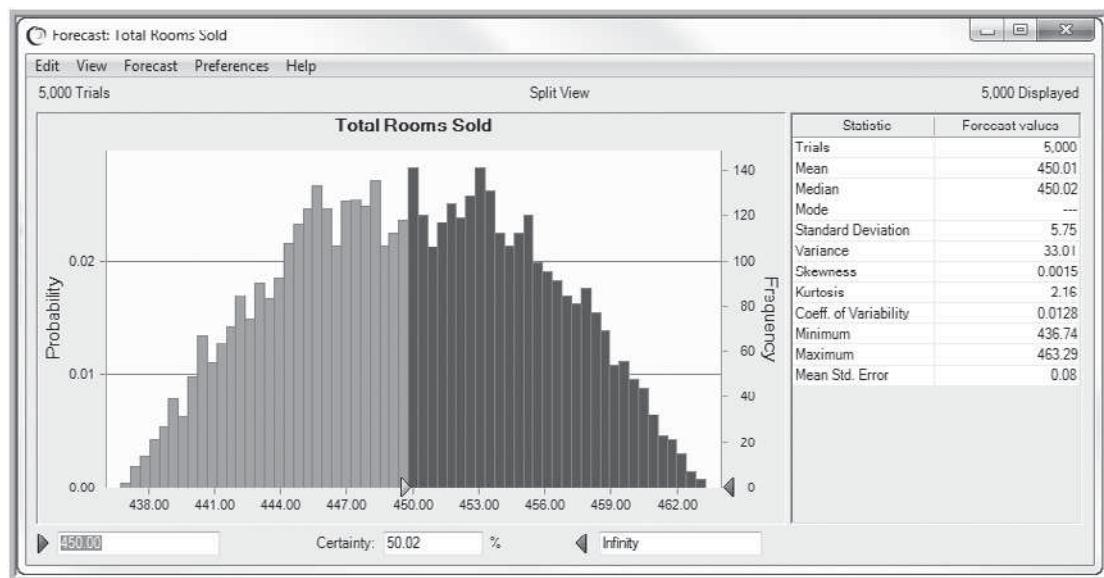
Use the *Job Sequencing Model* spreadsheet to find optimal sequences that minimize the variance of both lateness and tardiness.

## RISK ANALYSIS AND OPTIMIZATION

It is rare that any optimization model is completely deterministic; in most cases, some of the data will be uncertain. This implies that inherent risk exists in using the optimal solution obtained from a model. Using the capabilities of risk analysis software such as *Crystal Ball*, these risks can be better understood and mitigated. To illustrate this, we will use the hotel pricing problem.

In this problem, the price–demand elasticities of demand are only estimates and most likely are quite uncertain. Because we probably will not know anything about their distributions, let us conservatively assume that the true values might vary from the estimates by plus or minus 25%. Thus, we model the elasticities by uniform distributions. Using the optimal prices identified by *Solver* earlier in this chapter, let us see what happens to the forecast of the number of rooms sold under this assumption using *Crystal Ball*.

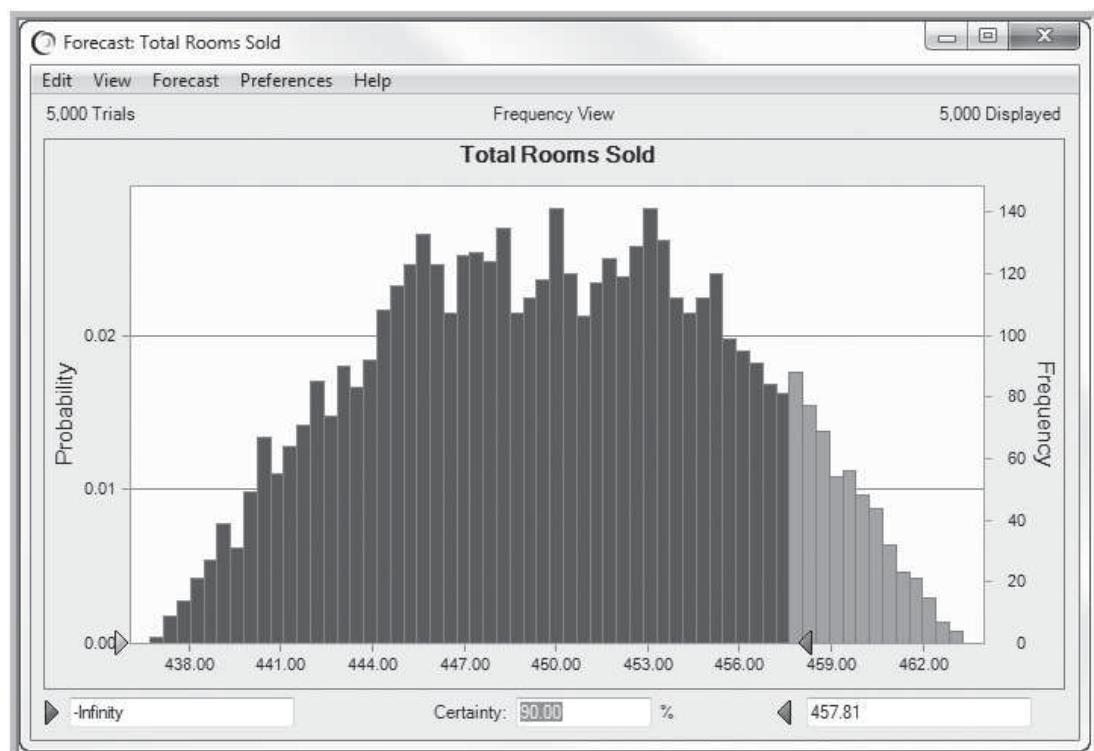
In the spreadsheet model, select cells D6:D8 as assumption cells with uniform distributions having minimum and maximum values equal to 75% and 125% of the estimated values, respectively. The item of total rooms sold (E16) is defined as a forecast cell. The model was replicated 5,000 times, creating the report in Figure 14.27. We see that the mean number of rooms sold under these prices is 450, which should be expected, since the mean values of the elasticities were used to derive the optimal prices. However, because of the uncertainty associated with the elasticities, the probability that *more* than 450 rooms will be sold (demanded) is approximately 0.5! This suggests that if the



**FIGURE 14.27** Crystal Ball Results for Hotel Pricing Example

assumptions of the uncertain elasticities are true, the hotel might anticipate that demand will exceed its room capacity about half the time, resulting in many unhappy customers.

We could use these results, however, to identify the appropriate hotel capacity to ensure, for example, only a 10% chance exists that demand will exceed capacity. Figure 14.28 shows the forecast chart when the certainty level is set at 90% and the left



**FIGURE 14.28** Forecast Chart for a 10% Risk of Exceeding Capacity

grabber is anchored. We could interpret this as stating that if the hotel capacity were about 457 or 458 rooms, then demand will exceed capacity at most 10% of the time. So if we shift the capacity constraint down by 7 rooms to 443 and find the optimal prices associated with this constraint, we would expect demand to exceed 450 at most 10% of the time. The *Solver* results for this case are shown in Figure 14.29, and Figure 14.30 shows the results of a *Crystal Ball* run confirming that with these prices, demand will exceed 450 less than 10% of the time.

	A	B	C	D	E	F
1	Marquis Hotel					
2						
3	Data					
4		Current	Average		Total Room	
5	Room type	Rate	Daily Sold	Elasticity	Capacity	
6	Standard	\$ 85.00	250	-1.5	443	
7	Gold	\$ 98.00	100	-2		
8	Platinum	\$ 139.00	50	-1		
9						
10	Model				Projected	
11					Rooms	Projected
12	Room type	New Price	Price Range	Sold		Revenue
13	Standard	\$ 78.34	\$ 70.00 \$ 90.00	279	\$ 21,886.69	
14	Gold	\$ 90.00	\$ 90.00 \$ 110.00	116	\$ 10,469.39	
15	Platinum	\$ 146.51	\$ 120.00 \$ 149.00	47	\$ 6,929.72	
16				Totals	443	\$ 39,285.80

FIGURE 14.29 Solver Solution for 443-Room Capacity

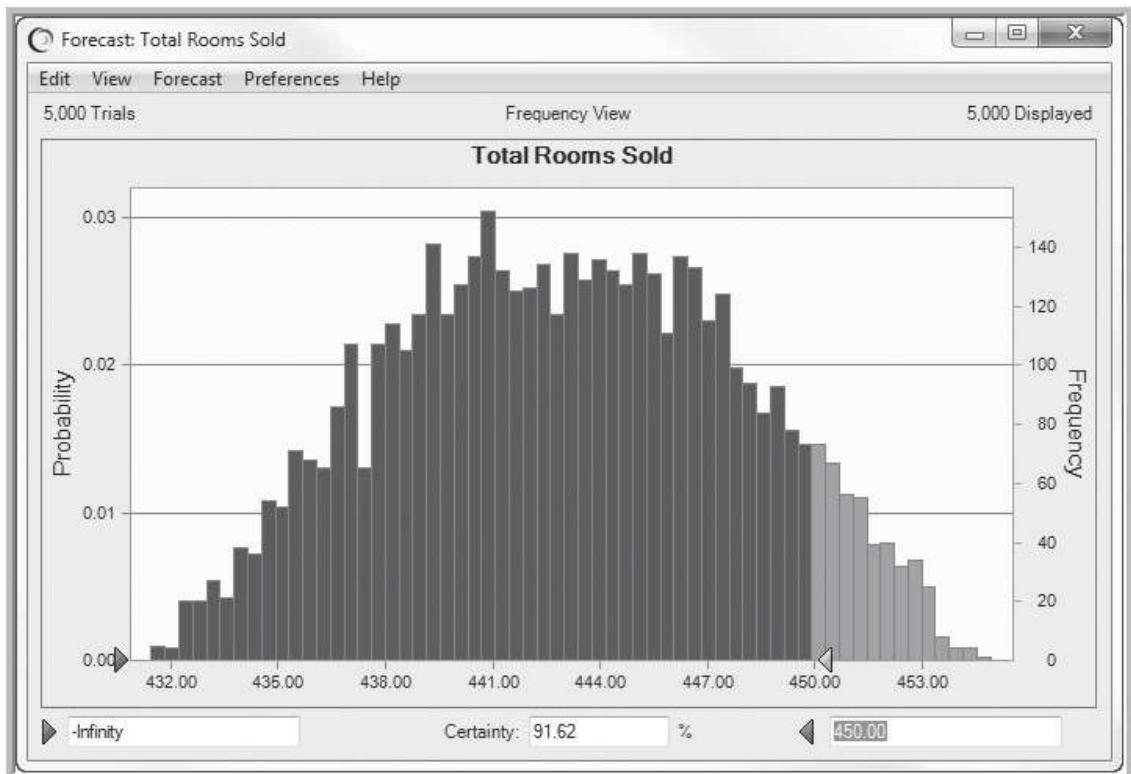


FIGURE 14.30 Crystal Ball Confirmation Run Results

## COMBINING OPTIMIZATION AND SIMULATION

To find an optimal set of decision variables for any simulation-based model, you generally need to search in a heuristic or ad hoc fashion. This usually involves running a simulation for an initial set of variables, analyzing the results, changing one or more variables, rerunning the simulation, and repeating this process until a satisfactory solution is obtained. This process can be very tedious and time-consuming, and often how to adjust the variables from one iteration to the next is not clear. While the *Crystal Ball Decision Table* tool is available, its functionality is somewhat limited.

*OptQuest* enhances the analysis capabilities of *Crystal Ball* by automatically searching for optimal solutions within *Crystal Ball* simulation model spreadsheets. Within *OptQuest*, you describe your optimization problem and search for values of decision variables that maximize or minimize a predefined objective. Additionally, *OptQuest* is designed to find solutions that satisfy a wide variety of constraints or a set of goals that you may define.

### A Portfolio Allocation Model

We will use a portfolio allocation model to illustrate the steps of setting up and running an optimization problem using *Crystal Ball* and *OptQuest*. An investor has \$100,000 to invest in four assets. The expected annual returns and minimum and maximum amounts with which the investor will be comfortable allocating to each investment follow:

Investment	Annual Return	Minimum	Maximum
1. Life insurance	5%	\$2,500	\$5,000
2. Bond mutual funds	7%	\$30,000	none
3. Stock mutual funds	11%	\$15,000	none
4. Savings account	4%	none	none

The major source of uncertainty in this problem is the annual return of each asset. In addition, the decision maker faces other risks, for example, unanticipated changes in inflation or industrial production, the spread between high- and low-grade bonds, and the spread between long- and short-term interest rates. One approach to incorporating such risk factors in a decision model is arbitrate pricing theory (APT).<sup>3</sup> APT provides estimates of the sensitivity of a particular asset to these types of risk factors. Let us assume that the risk factors per dollar allocated to each asset have been determined as follows:

Investment	Risk Factor/Dollar Invested
1. Life insurance	-0.5
2. Bond mutual funds	1.8
3. Stock mutual funds	2.1
4. Savings account	-0.3

The investor may specify a target level for the weighted risk factor, leading to a constraint that limits the risk to the desired level. For example, suppose that our investor will tolerate a weighted risk per dollar invested of at most 1.0. Thus, the weighted risk for a \$100,000 total investment will be limited to 100,000. If our investor allocates \$5,000 in life insurance, \$50,000 in bond mutual funds, \$15,000 in stock mutual funds, and

<sup>3</sup> See Schniederjans, M., T. Zorn, and R. Johnson, "Allocating Total Wealth: A Goal Programming Approach," *Computers and Operations Research*, 20, no. 7(1993): 679–685.

\$30,000 in a savings account (which fall within the minimum and maximum amounts specified), the total expected annual return would be:

$$0.05(\$5,000) + 0.07(\$50,000) + 0.11(\$15,000) + 0.04(\$30,000) = \$6,600$$

However, the total weighted risk associated with this solution is:

$$-0.5(5,000) + 1.8(50,000) + 2.1(15,000) - 0.3(30,000) = 110,000$$

Because this is greater than the limit of 100,000, this solution could not be chosen.

The decision problem, then, is to determine how much to invest in each asset to maximize the total expected annual return, remain within the minimum and maximum limits for each investment, and meet the limitation on the weighted risk.

### Using OptQuest

The basic process for using *OptQuest* is described as follows:

1. Create a *Crystal Ball* model of the decision problem.
2. Define the decision variables within *Crystal Ball*.
3. Run *OptQuest* from the *Crystal Ball Tools* group.
4. Specify objectives, decision variables, constraints, and other options as appropriate.
5. Solve the optimization problem.

**CREATE THE CRYSTAL BALL SPREADSHEET MODEL** An important task in using *OptQuest* is to create a useful spreadsheet model. A spreadsheet for this problem is shown in Figure 14.31 (Excel file *Portfolio Allocation Model*). Problem data are specified in rows 4 through 8. On the bottom half of the spreadsheet, we specify the model outputs, namely, the values of the decision variables, objective function, and constraints (the total weighted risk and total amount invested). You can see that this particular solution is not feasible because the total weighted risk exceeds the limit of 100,000.

Now that the basic model is developed, we define the assumptions and forecast cells in *Crystal Ball*. We will assume that the annual returns for life insurance and mutual

	A	B	C	D	E
1	Portfolio Allocation Model				
2		Annual			Risk factor
3	Investment	return	Minimum	Maximum	per dollar
4	Life Insurance	5.0%	\$ 2,500.00	\$ 5,000.00	-0.5
5	Bond mutual funds	7.0%	\$ 30,000.00	none	1.8
6	Stock mutual funds	11.0%	\$ 15,000.00	none	2.1
7	Savings Account	4.0%	none	none	-0.3
8	Total amount available	\$100,000		Limit	100,000
9					
10		Amount			Total weighted
11	Decision variables	invested			risk
12	Life Insurance	\$ 5,000.00			146,000.00
13	Bond mutual funds	\$ 50,000.00			
14	Stock mutual funds	\$ 30,000.00			Total expected
15	Savings Account	\$ 15,000.00			return
16	Total amount invested	\$ 100,000.00			\$ 7,650.00

FIGURE 14.31 Portfolio Allocation Model Spreadsheet

funds are uncertain, but that the rate for the savings account is constant. We will make the following assumptions in the *Crystal Ball* model:

- Cell B4: uniform distribution with minimum 4% and maximum 6%
- Cell B5: normal distribution with mean 7% and standard deviation 1%
- Cell B6: lognormal distribution with mean 11% and standard deviation 4%

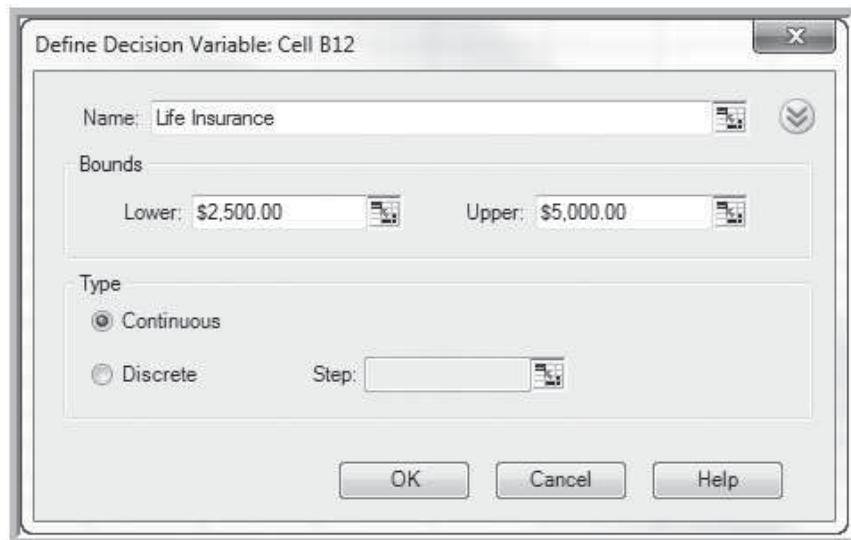
We define the forecast cell to be the total expected return, cell E16. As would be the case with any *Crystal Ball* application, you would select *Run Preferences* from the *Run* menu and choose appropriate settings. For this example, we set the number of trials per simulation to 2,000.

**DEFINE DECISION VARIABLES** The next step is to identify the decision variables in the model. This is accomplished using the *Define Decision* option in the *Define* group. Position the cursor on cell B12 and click *Define Decision*. Set the minimum and maximum values according to the problem data (i.e., columns C and D in the spreadsheet), as shown in Figure 14.32.

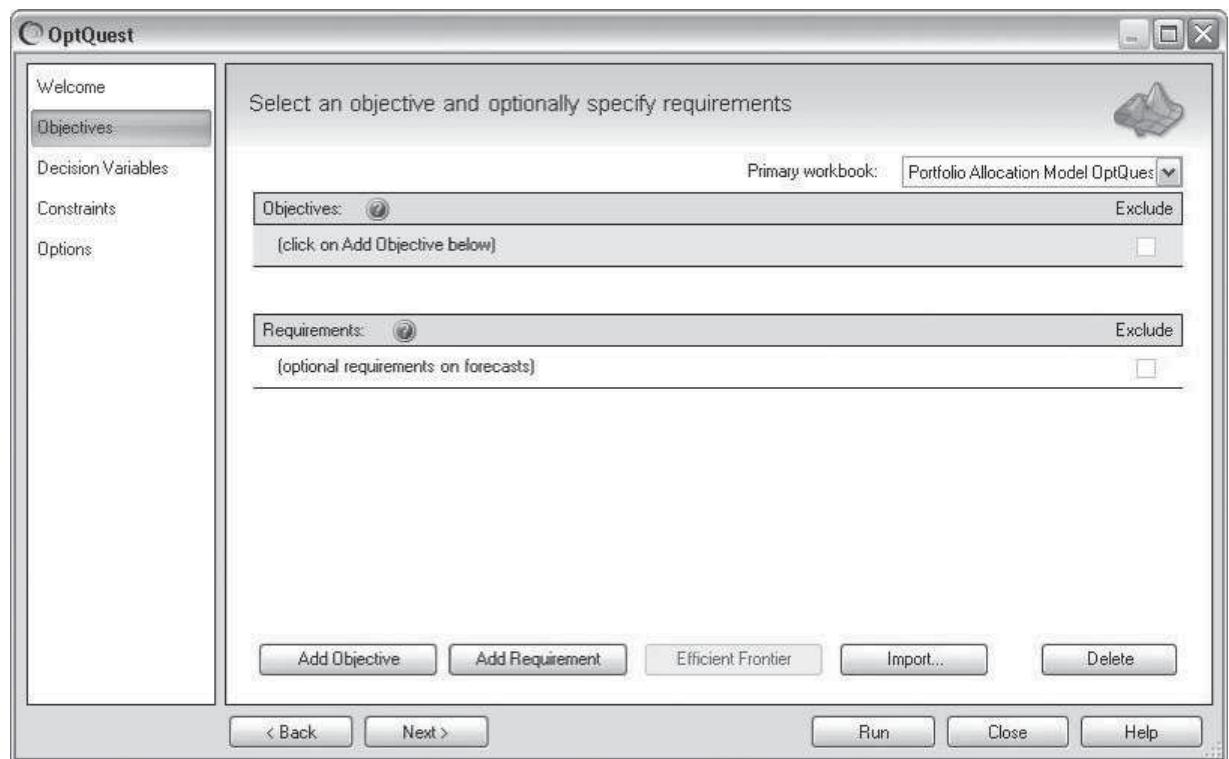
Next, we repeat the process of defining decision variables for cells B13, B14, and B15. When the maximum limit is “none,” you may use a value of \$100,000 because this is the total amount available. You are now ready to run *OptQuest* by clicking on the *Opt-Quest* button in the *Crystal Ball* ribbon.

**CREATE A NEW OPTIMIZATION FILE AND SPECIFY OBJECTIVES** In the *Welcome* screen in *OptQuest*, click *Next* to get started. This will step you through the process of setting up your optimization model. *OptQuest* will display a dialog to specify your objectives. Click the *Add Objective* button. For the portfolio example, the window will display the objective “*Maximize the mean of Total expected return*” which you may customize as appropriate. This is shown in Figure 14.33. We will discuss adding requirements after completing this example.

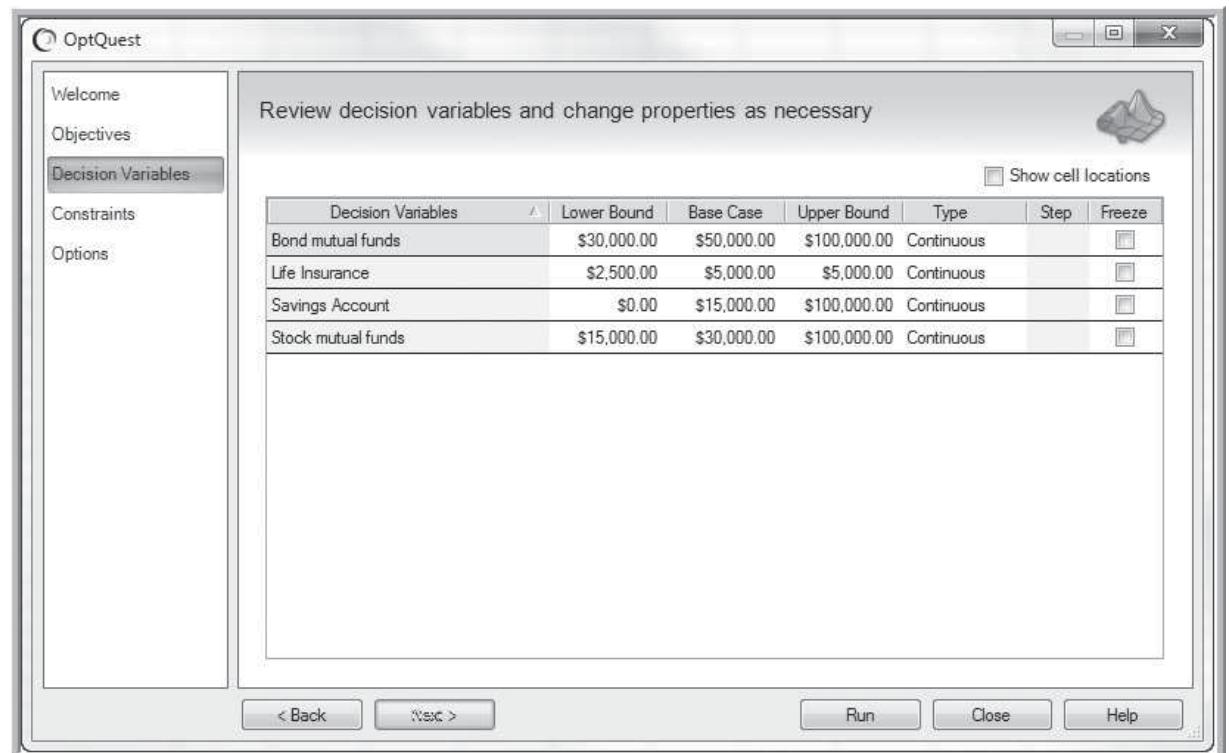
**SPECIFY DECISION VARIABLES** After clicking *Next* (or choosing *Decision Variables* from the left panel), the *Decision Variables* screen is displayed (Figure 14.34). The decision variables defined in your *Crystal Ball* model will be shown. You may unselect (freeze) any of them as appropriate. The *Base Case* values are the values currently in the spreadsheet. The *Type* column indicates whether a variable is discrete or continuous. The variable



**FIGURE 14.32** Define Decision Variable Dialog



**FIGURE 14.33** OptQuest Objectives Input Screen



**FIGURE 14.34** OptQuest Decision Variables Input Screen

type can be changed in this window or in the *Define Decision* dialog of *Crystal Ball*. A step size is associated with discrete variables. A variable of the type Discrete\_2, for example, has a step size of 2. Therefore, if the lower and upper bounds for this variable are 0 and 7, respectively, the only feasible values are 0, 2, 4, and 6. Any values or type may be changed by clicking on the appropriate value or name. Click *Next* to continue.

**SPECIFY CONSTRAINTS** The *Constraints* screen displayed allows you to specify any constraints. A constraint is any limitation or requirement that restricts the possible solutions to the problem. In our example, we have two constraints. The first constraint limits the total weighted risk to 100,000, and the second ensures that we do not allocate more than \$100,000 in total to all assets. Constraints are entered in a manner similar to the way they are defined in *Solver*; that is, by specifying the cells corresponding to the left and right hand side of the constraint. In this example, the risk constraint is expressed as  $E12 \leq E8$ , and the allocation constraint as  $B16 \leq B8$ . This is shown in Figure 14.35.

**SPECIFY OPTQUEST RUN OPTIONS** Next, the *Options* window allows you to select various options for running the simulation (see Figure 14.36). *Optimization control* allows you to specify the total time that the system is allowed to search for the best values for the optimization variables. You may enter either the total number of minutes or number of simulations. Performance will depend on the speed of your microprocessor; however, you are able to choose any time limit you desire. Selecting a very long time limit does not present a problem because you are always able to terminate the search in the control panel. You may click the *Run Preferences* button to change the number of trials and other *Crystal Ball* preferences (in this example, we set the number of trials to 2,500). Additionally, you will be given the option to extend the search and carry the optimization process farther once the

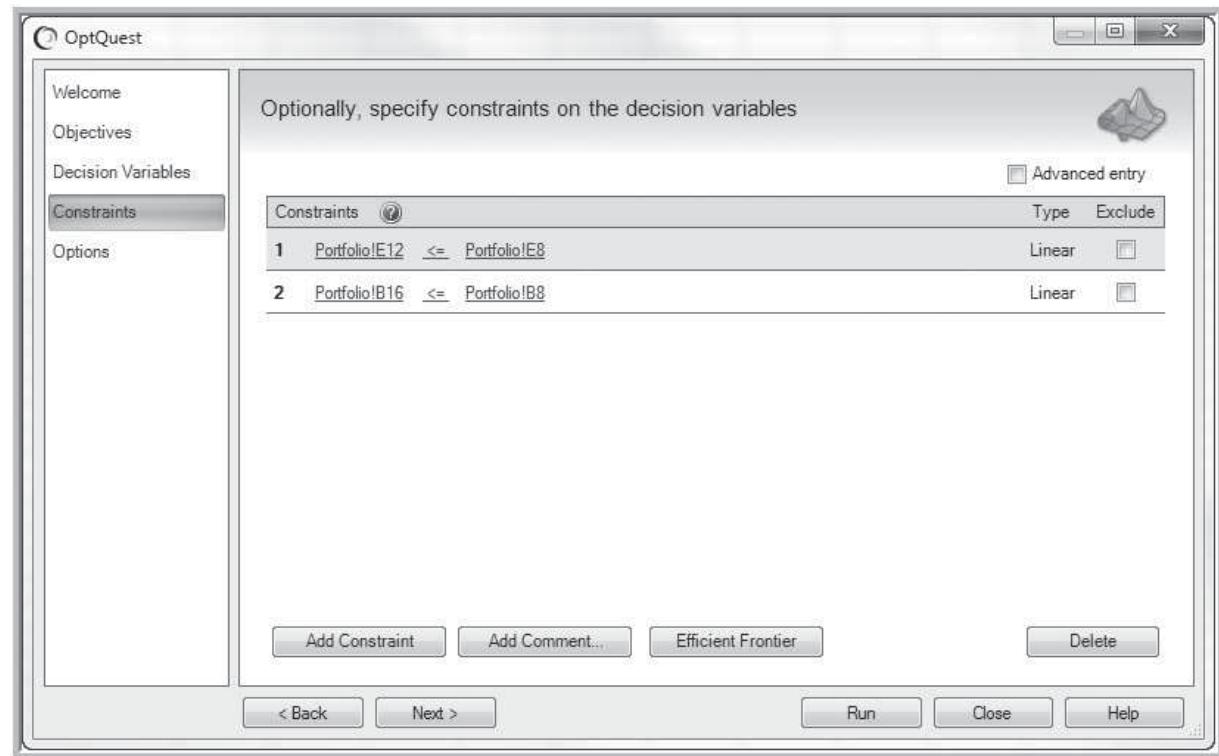


FIGURE 14.35 OptQuest Constraints Input Screen

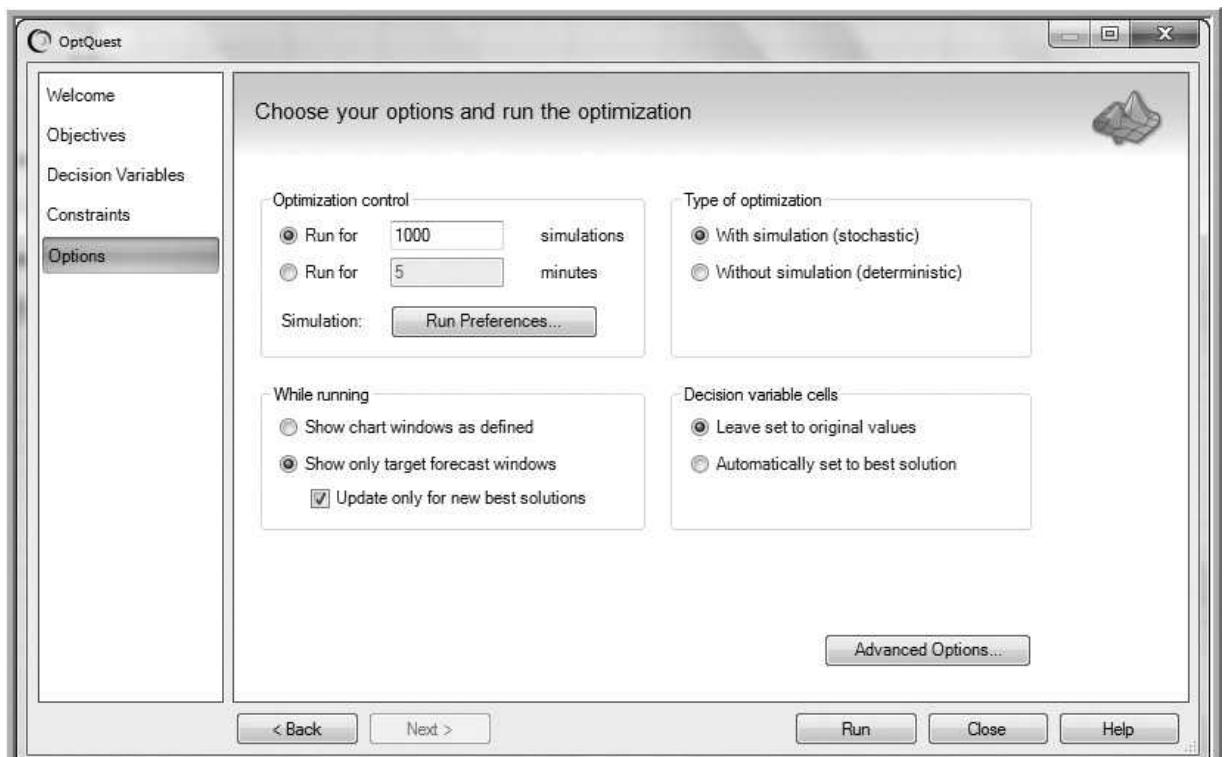


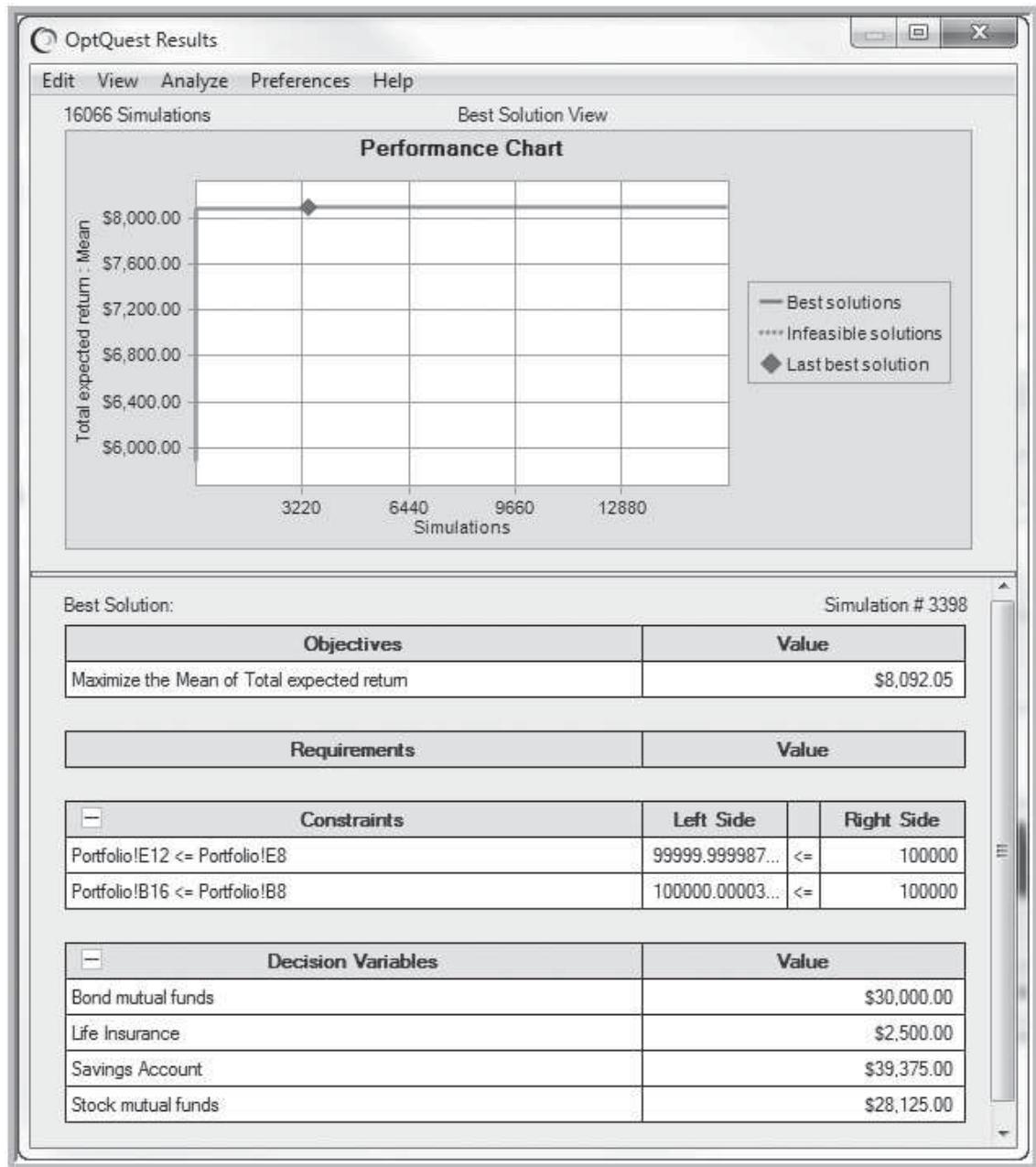
FIGURE 14.36 OptQuest Options Input screen

selected time has expired. Under *Type of optimization*, you can set the optimization type as stochastic—that is, with *Crystal Ball* assumptions or deterministic without assumptions. You may also specify what windows are displayed while the simulation is running and whether to set the decision variable cells to the best solution upon completion.

**SOLVE THE OPTIMIZATION PROBLEM** The optimization process is initiated by clicking the *Run* button. As the simulation is running, you will see a *Performance Chart* that shows a plot of the value of the objective as a function of the number of simulations evaluated, and details of the best solutions generated during the search. Figure 14.37 shows the *OptQuest Results* screen upon completion of the optimization. From the Edit menu, you may copy the best solution to your spreadsheet.

**INTERPRETING OPTQUEST RESULTS** You should note that the “best” *OptQuest* solution identified may not be the true optimal solution to the problem, but will, it is hoped, be close to the actual optimal solution. The accuracy of the results depends on the time limit you select for searching, the number of decision variables, the number of trials per simulation, and the complexity of the problem. With more decision variables, you need a larger number of trials.

After solving an optimization problem with *OptQuest*, you probably would want to examine the *Crystal Ball* simulation using the optimal values of the decision variables in order to assess the risks associated with the recommended solution. Figure 14.38 shows the *Crystal Ball* forecast chart associated with the best solution. Although the mean value was optimized, we see that a high amount of variability exists in the actual return because of the uncertainty in the returns of the individual investments. Also note that the mean value for the simulation is different than that found in the *OptQuest* solution because we used a larger number of trials. With *OptQuest*, we used only 1,000 trials in solution, and resulted in a mean annual return of 10.4%.

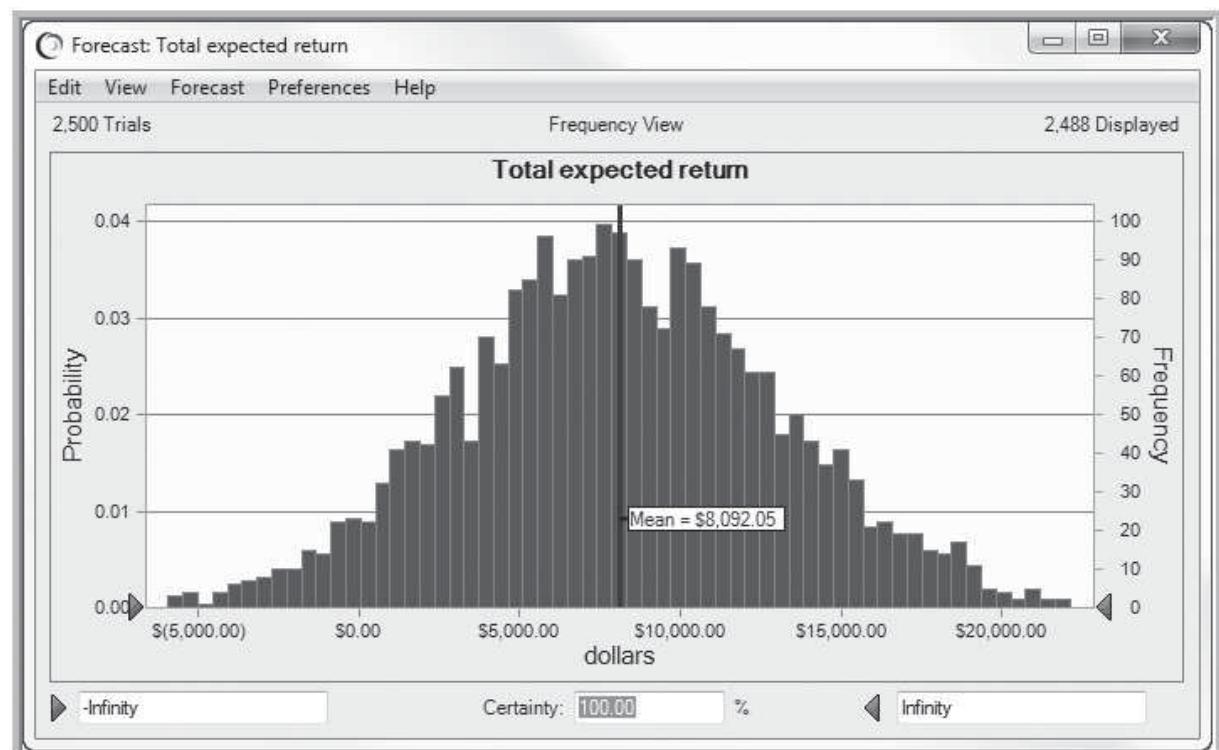


**FIGURE 14.37** OptQuest Results Summary

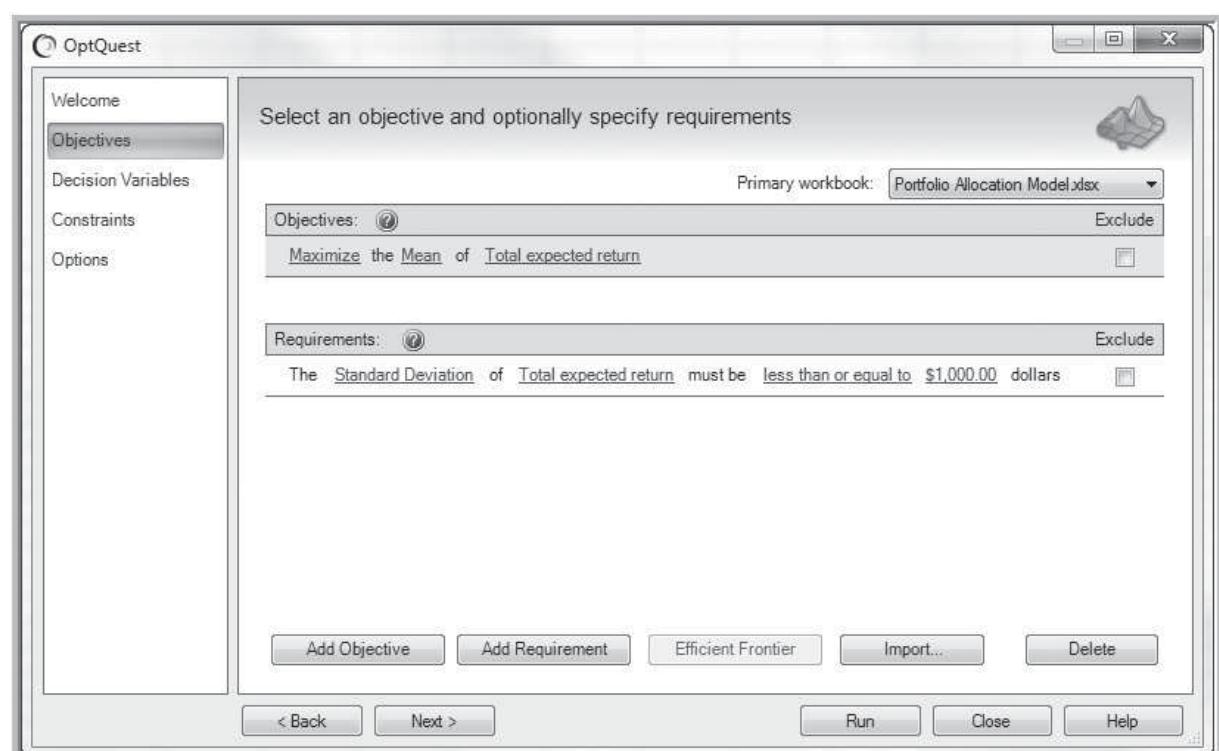
trials, the search would have uncovered fewer solutions in the time allotted. Thus, a trade-off exists between precision and time.

**ADDING A REQUIREMENT** A *requirement* is a forecast statistic that is restricted to fall within a specified lower and upper bound. The forecast statistic may be one of the following:

- Mean
- Median
- Mode
- Standard deviation
- Variance



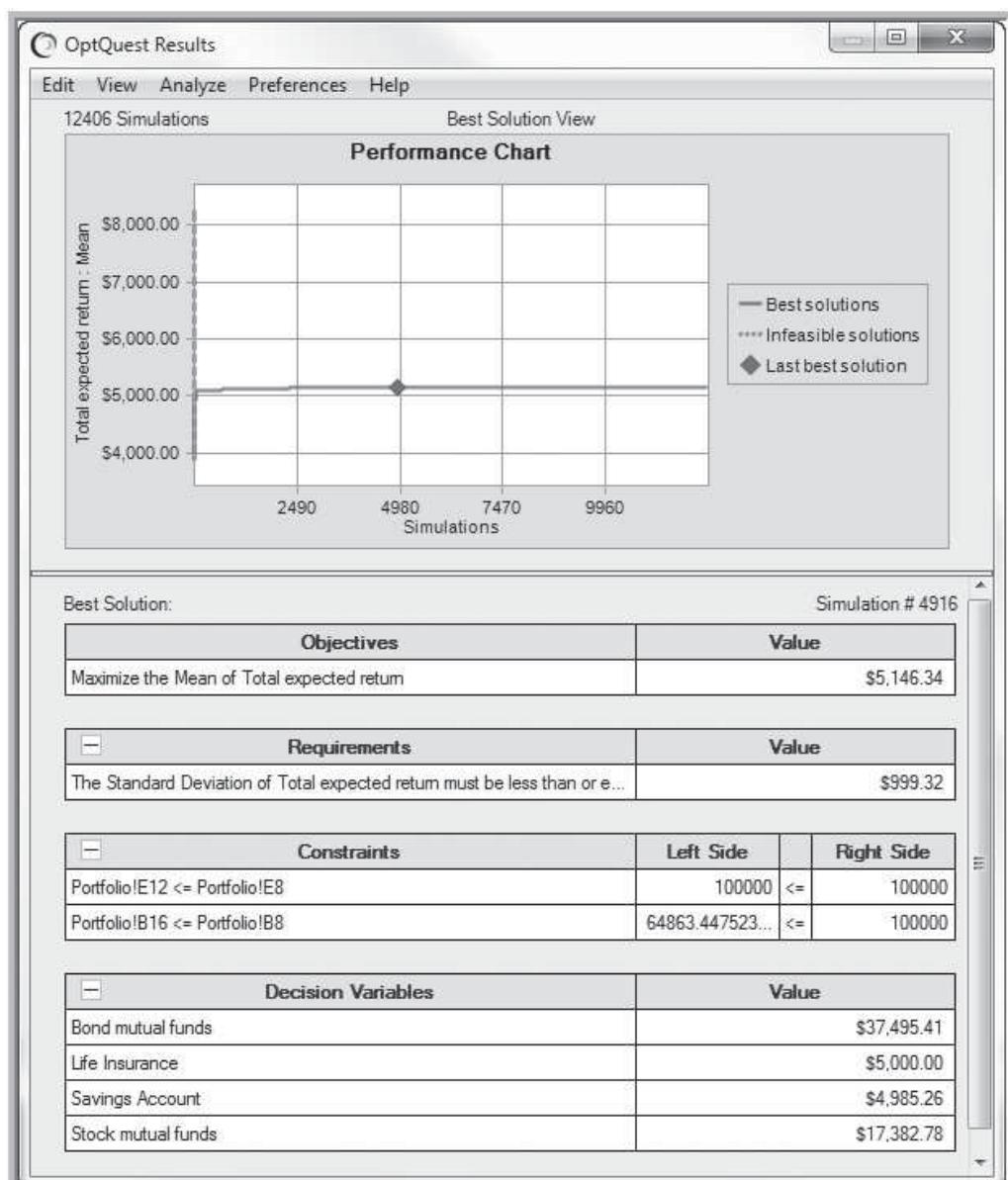
**FIGURE 14.38** Crystal Ball Simulation Results for Best OptQuest Solution



**FIGURE 14.39** Forecast Selection Window with Standard Deviation Requirement

- Percentile (as specified by the user)
- Skewness
- Kurtosis
- Coefficient of variation
- Range (minimum, maximum, and width)
- Standard error

For example, to reduce the uncertainty of returns in the portfolio while also attempting to maximize the expected return, we might want to restrict the standard deviation to be no greater than 1,000. To add such a requirement in *OptQuest*, click the *Add Requirement* button in the *Objectives* input screen. You will need to change the default requirement to reflect the correct values. This is shown in Figure 14.39. You may now run the new model. The results (starting from the solution from the previous optimization run) are shown in Figure 14.40. The best solution among those with standard deviations less than or equal to 1,000 is identified.



**FIGURE 14.40**  
*OptQuest Results with Standard Deviation Requirement*

## SKILL-BUILDER EXERCISE 14.7

Set up and run the *OptQuest* model for the portfolio allocation problem with the requirement that the standard deviation be no greater than 700.

### Basic Concepts Review Questions

1. What is an integer program?
2. What is a mixed linear optimization model?
3. Why can't Sensitivity Reports be produced for integer models in *Solver* in the usual manner?
4. How can binary variables be used to model logical conditions? Provide several examples.
5. In both the plant location and fixed-cost models presented in this chapter, we used "linking" constraints that established a relationship between the binary variables and other continuous variables in the model.

- Why are such constraints necessary in any mixed integer model?
6. What is a local optimum solution? What should you do to improve your chances of getting a global optimum solution?
7. What are *metaheuristics*? What are the tools in Excel that can implement such approaches?
8. How can *Crystal Ball* and *OptQuest* enhance the analysis of optimization problems?

### Problems and Applications

Note: Data for most of these problems are provided in the Excel file **Chapter 14 Problem Data** to facilitate model building. Worksheet tabs correspond to problem numbers.

1. Thermal transfer printing transfers ink from a ribbon onto paper through a combination of heat and pressure. Different types of printers use different sizes of ribbons. A ribbon manufacturer has forecasted demands for seven different ribbon sizes:

Size	Forecast (Rolls)
60 mm	1,620
83 mm	520
102 mm	840
110 mm	2,640
120 mm	500
130 mm	740
165 mm	680

The rolls from which ribbons are cut are 880 mm in length. Scrap is valued at \$0.07 per mm. Generate 10 different cutting patterns so that each size can be cut from at least one pattern. Use your data to construct and solve an optimization model for finding the number of patterns to cut to meet demand and minimize trim loss.

2. For the cutting stock example in the chapter, suppose that the scrap generated from each pattern costs \$2 per inch. The company also incurs a fixed cost for setting up the machine to cut different patterns. These costs are:

Pattern	Fixed Cost
1	\$70
2	\$40
3	\$40
4	\$90
5	\$50
6	\$80

Modify the model developed in this chapter to include the scrap and set up costs, and solve the model with *Solver*.

3. The personnel director of a company that recently absorbed another firm and is now downsizing and must relocate five information systems analysts from recently closed locations. Unfortunately, there are only three positions available for five people. Salaries are

fairly uniform among this group (those with higher pay were already given the opportunity to begin anew). Moving expenses will be used as the means of determining who will be sent where. Estimated moving expenses are:

<b>Moving Cost To</b>			
<b>Analyst</b>	<b>Gary</b>	<b>Salt Lake City</b>	<b>Fresno</b>
Arlene	\$8,500	\$6,000	\$5,000
Bobby	\$5,000	\$8,000	\$12,000
Charlene	\$9,500	\$14,000	\$17,000
Douglas	\$4,000	\$8,000	\$13,000
Emory	\$7,000	\$3,500	\$4,500

Model this as an integer optimization model to minimize cost and determine which analysts to relocate to the three locations.

4. A medical device company is allocating next year's budget among its divisions. As a result, the R&D Division needs to determine which research and development projects to fund. Each project requires various software and hardware and consulting expenses, along with internal human resources. A budget allocation of \$1,300,000 has been approved, and 35 engineers are available to work on the projects. The R&D group has determined that at most one of projects 1 and 2 should be pursued, and that if project 4 is chosen, then project 2 must also be chosen. Develop a model to select the best projects within the budget.

<b>Project</b>	<b>NPV</b>	<b>Internal Engineers</b>	<b>Additional Costs</b>
1	\$600,000	9	\$196,000
2	680,000	12	400,000
3	550,000	7	70,000
4	400,000	4	180,000
5	350,000	8	225,000
6	725,000	10	275,000
7	340,000	8	130,000

5. A software support division of a major corporation has eight projects that can be performed. Each project requires different amounts of development time and testing time. In the coming planning period, 1,150 hours of development time and 900 hours of testing time are available, based on the skill mix of the staff. The internal transfer price (revenue to the support division) and the times required for each project are shown

in the table. Which projects should be selected to maximize revenue?

<b>Project</b>	<b>Development Time</b>	<b>Testing Time</b>	<b>Transfer Price</b>
1	80	67	\$23,520
2	248	208	\$72,912
3	41	34	\$12,054
4	10	92	\$32,340
5	240	202	\$70,560
6	195	164	\$57,232
7	269	226	\$79,184
8	110	92	\$32,340

6. Soapbox is a local band that plays classic and contemporary rock. The band members charge \$600 for a three-hour gig. They would like to play at least 30 gigs per year but need to determine the best way to promote themselves. The most they are willing to spend on promotion is \$2,500. The possible promotion options are as follows:

- Playing free gigs
- Making a demo CD
- Hiring an agent
- Handing out fliers
- Creating a Web site

Each free gig costs them about \$250 for travel and equipment, but generates about 3 paying gigs. A high-quality studio demo CD should help the band book about 20 gigs, but will cost about \$1,000. A demo CD made on home recording equipment will cost only \$400 but may result in only 10 bookings. A good agent will get the band about 15 gigs, but will charge \$1,500. The band can create a Web site for \$400 and would expect to generate 6 gigs from this exposure. They also estimate that they may book 1 gig for every 500 fliers they hand out, which would cost \$0.08 each. They don't want to play more than 10 free gigs or send out more than 2,500 fliers. Develop and solve an optimization model to find the best promotion strategy to maximize their revenue.

7. Premier Paints supplies to major contractors. One of their contracts for a specialty paint requires them to supply 750, 500, 400, and 950 gallons over the next 4 months. To produce this paint requires a shutdown and cleaning of one of their manufacturing departments at a cost of \$1,000. The entire contract requirement can be produced during the first month in one production run; however, the inventory that must be held until delivery costs \$0.75 per gallon per month. If the paint is produced in other months, then the cleaning costs are incurred during each month of

production. Formulate and solve an integer optimization model to determine the best monthly production schedule to meet delivery contracts and minimize total costs.

- 8.** Dannenfelser Design works with clients in three major project categories: architecture, interior design, and combined. Each type of project requires an estimated number of hours for different categories of employees as shown in the table below.

	Architecture	Interior Design	Combined	Hourly Rate
Principal	15	5	18	\$150
Sr. Designer	25	35	40	\$110
Draftsman	40	30	60	\$75
Administrator	5	5	8	\$50

In the coming planning period, 184 hours of Principal time, 414 hours of Sr. Designer time, 588 hours of drafter time, and 72 hours of administrator time are available. Profit per project averages \$1,290 for architecture, \$1,110 for interior design, and \$1,878 for combined projects. The firm would like to work on at least

one of each type of project for exposure among clients. Assuming that the firm has more demand than they can possibly handle, find the best mix of projects to maximize profit.

- 9.** Anya is a part-time MBA student who works full-time and is constantly on the run. She recognized the challenge of eating a balanced diet and wants to minimize cost while meeting some basic nutritional requirements. Based on some research, she found that a very active woman should consume 2,250 calories per day. According to one author's guidelines, the following daily nutritional requirements are recommended:

Source	Recommended Intake (grams)
Fat	maximum 75
Carbohydrates	maximum 225
Fiber	maximum 30
Protein	at least 168.75

She chose a sample of meals that could be obtained from healthier quick-service restaurants around town as well as some that could be purchased at the grocery store.

Food	Cost/Serving	Calories	Fat	Carbs	Fiber	Protein
Turkey sandwich	\$4.69	530	14	73	4	28
Baked potato soup	\$3.39	260	16	23	1	6
Whole grain chicken sandwich	\$6.39	750	28	83	10	44
Bacon turkey sandwich	\$5.99	770	28	84	5	47
Southwestern refrigerated chicken wrap	\$3.69	220	8	29	15	21
Sesame chicken refrigerated chicken wrap	\$3.69	250	10	26	15	26
Yogurt	\$0.75	110	2	19	0	5
Raisin bran with skim milk	\$0.40	270	1	58	8	12
Cereal bar	\$0.43	110	2	22	0	1
1 cup broccoli	\$0.50	25	0.3	4.6	2.6	2.6
1 cup carrots	\$0.50	55	0.25	13	3.8	1.3
1 scoop protein powder	\$1.29	120	4	5	0	17

Anya does not want to eat the same entrée (first six foods) more than once each day but does not mind eating breakfast or side items (last five foods) twice a day and protein powder-based drinks up to four times a day for convenience. Develop an integer linear optimization model to find the number of servings of each food choice in a daily diet to minimize cost and meet the nutritional targets.

- 10.** Brewer Services contracts with outsourcing partners to handle various customer service functions. The customer service department is open Monday through Friday from 8 A.M. to 5 P.M. Calls vary over the course of a typical day. Based on a study of call volumes provided by

one of the firm's partners, the minimum number of staff needed each hour of the day are:

Hour	Minimum Staff Required
8–9	5
9–10	12
10–11	16
11–noon	12
Noon–1	11
1–2	16
2–3	17
3–4	18
4–5	14

Mr. Brewer wants to hire some permanent employees and staff the remaining requirements using part-time employees who work four-hour shifts (four consecutive hours starting as early as 8 A.M. or as late as 1 P.M.). Develop and solve an integer optimization model to answer the following questions.

- a. Suppose Mr. Brewer uses five permanent employees. What is the minimum number of part-time employees he will need for each four-hour shift to ensure meeting the staffing requirements?
  - b. What if he uses 11 permanent employees?
  - c. Investigate other possible numbers of permanent employees between 5 and 15, compare the differences, and make a recommendation.
11. Josh Steele manages a professional choir in a major city. His marketing plan is focused on generating additional local demand for concerts and increasing ticket revenue, and also gaining attention at the national level to build awareness of the ensemble across the country. He has \$20,000 to spend on media advertising. The goal of the advertisement campaign is to generate as much local recognition as possible while reaching at least 4,000 units of national exposure. He has set a limit of 100 total ads. Additional information is shown below.

Media	Price	Local Exposure	National Exposure	Limit
FM radio spot	\$80.00	110	40	30
AM radio spot	\$65.00	55	20	30
Cityscape ad	\$250.00	80	5	24
MetroWeekly ad	\$225.00	65	8	24
Hometown paper ad	\$500.00	400	70	10
Neighborhood paper ad	\$300.00	220	40	10
Downtown magazine ad	\$55.00	35	0	15
Choir journal ad	\$350.00	10	75	12
Professional organization magazine ad	\$300.00	20	65	12

The last column sets limits on the number of ads to ensure that the advertising markets do not become saturated.

- a. Find the optimal number of ads of each type to run to meet the choir's goals by developing and solving an integer optimization model.
  - b. What if he decides to use no more than six different types of ads? Modify the model in part (a) to answer this question.
12. Chris Corry has a company-sponsored retirement plan at a major brokerage firm. He has the following funds available:

Fund	Risk	Type	Return
1	High	Stock	11.98%
2	High	Stock	13.18%
3	High	Stock	9.40%
4	High	Stock	7.71%
5	High	Stock	8.35%
6	High	Stock	16.38%
7	Medium	Blend	4.10%
8	Medium	Blend	12.52%
9	Medium	Blend	8.62%
10	Medium	Blend	11.14%
11	Medium	Blend	8.78%
12	Low	Blend	9.44%
13	Low	Blend	8.38%
14	Low	Bond	7.65%
15	Low	Bond	6.90%
16	Low	Bond	5.53%
17	Low	Bond	6.30%

His financial advisor has suggested that at most 40% of the portfolio should be composed of high-risk funds. At least 25% should be invested in bond funds, and at most 40% can be invested in any single fund. At least six funds should be selected, and if a fund is selected, it should be funded with at least 5% of the total contribution.

Develop and solve an integer optimization model to determine which funds should be selected and what percentage of his total investment should be allocated to each fund.

13. Mark Haynes is interested in buying a new car. He decided on a particular model, which has lots of options from which to choose. The base price of the car is \$16,510 and he allotted a budget of \$19,250 to purchase it. The table below shows the possible options over and above the base model that he could choose, their cost, and the utility that he assigned to each option:

Option	Utility	Cost
Slower engine	-0.10	(\$300.00)
Faster engine	0.40	\$600.00
Fastest engine	0.50	\$1,000.00
No warranty	0.00	(\$250.00)
3-year warranty	0.30	\$450.00
5-year warranty	0.70	\$750.00
Automatic transmission	0.80	\$800.00
15-inch wheels	-0.15	(\$150.00)
16-inch wheels	0.25	\$300.00
Alloy wheels	0.35	\$500.00
AM/FM radio	0.10	\$200.00
AM/FM/CD	0.30	\$300.00
AM FM/CD/DVD	0.50	\$400.00
AM/BM/CD/DVD 6 speaker	0.65	\$750.00
Sunroof	0.25	\$50.00
Moonroof	0.40	\$150.00
2-wheel disc brakes	-0.10	(\$250.00)
4-wheel disc brakes with ABS	0.35	\$250.00

All of these are options to replace the base model configuration. Thus, for example, he may choose one of the three engine options or keep the base model engine, choose one of these three warranty options or keep the base model warranty, choose automatic transmission or keep the base model manual transmission, and so on. In addition, if the automatic transmission is chosen, then a three- or five-year warranty must be chosen, and if the fastest engine is chosen, then either the 16-inch or alloy wheels must be chosen.

- a. Develop and solve an integer optimization model to maximize his utility and stay within budget.
  - b. If he has an additional \$500 to spend, how would his choices change?
  - c. In a magazine, Mark found a car stereo system for \$300. If he decides to replace the base model radio with this one, how would the model and his decisions change?
14. The Spurling Group is considering using magazine outlets to advertise their online Web site. The company has identified seven publishers. Each publisher breaks down their subscriber base into a number of groups based on demographics and location. These data are shown in the table below:

Publisher	Groups	Subscribers/Group	Cost/Group
A	5	460,000	\$1,560
B	10	50,000	\$290
C	4	225,000	\$1,200
D	20	24,000	\$130
E	5	1,120,000	\$2,500
F	1	1,700,000	\$7,000
G	2	406,000	\$1,700

The company has set a budget of \$25,000 for advertising and wants to maximize the number of subscribers exposed to their ads. However, publishers B and D are competitors and only one of these may be chosen. A similar situation exists with publishers C and G. Formulate and solve an integer optimization model to determine which publishers to select and how many groups to purchase for each publisher.

15. A young entrepreneur has invented a new air-adjustable basketball shoe with pump, similar to those advertised widely by more expensive brand names. He contacted a supplier of Victor basketball shoes, a little-known brand with low advertising. This supplier would provide shoes at the nominal price of \$6 per pair of shoes. He needs to know the best price at which to sell these shoes. As a business student with strong economics training, he remembered that the volume sold is affected by the

product's price—the higher the price, the lower the volume. He asked his friends and acquaintances what they would pay for a premium pair of basketball shoes that were a "little off-brand." Based on this information, he developed the formula:

$$\text{Volume} = 1,000 - 20 \text{ Price}$$

There are some minor expenses involved, including a \$50 fee for selling shoes in the neighborhood (a fixed cost), as well as his purchase price of \$6 per shoe. Develop an appropriate objective function and find the optimal price level using *Solver*.

16. The entrepreneur in the previous problem did very well selling Victor shoes. His shoe supplier told him of a new product, Top Notch, that was entering the market. This shoe would be a product substitute for Victors, so that the higher the price of either shoe, the greater the demand for the other. He interviewed more potential clients to determine price response and cross elasticities. This yielded the following relationships:

$$\text{Volume of Victors} = 1,000 - 20P_v + 1P_n$$

$$\text{Volume of Top Notch} = 800 + 2P_v - 18P_n$$

where  $P_v$  = price of Victors and  $P_n$  = price of Top Notch. Develop a model to maximize the total revenue and find the optimal prices using *Solver*.

17. The Hal Chase Investment Planning Agency is in business to help investors optimize their return from investment. Hal deals with three investment mediums: a stock fund, a bond fund, and his own Sports and Casino Investment Plan (SCIP). The stock fund is a mutual fund investing in openly traded stocks. The bond fund focuses on the bond market, which has a more stable but lower expected return. SCIP is a high-risk scheme, often resulting in heavy losses but occasionally coming through with spectacular gains. Average returns, their variances, and covariances are:

	Stock	Bond	SCIP
Average return	0.148	0.060	0.152
Variance	0.014697	0.000155	0.160791
Covariance with stock		0.000468	-0.002222
Covariance with bond			-0.000227

Develop and solve a portfolio optimization model for this situation for a target return of 12%.

18. Stout Investments wishes to design a minimum variance portfolio of index funds. The funds selected for

consideration and their variance–covariance matrix and average returns are given below:

	Bond	S&P 500	Small Cap	Mid Cap	Large Cap	Emerging Market	Commodity
Bond	0.002%						
S&P 500	-0.001%	0.020%					
Small cap	-0.001%	0.027%	0.047%				
Mid cap	-0.001%	0.024%	0.039%	0.033%			
Large cap	-0.001%	0.019%	0.027%	0.023%	0.027%		
Emerging market	0.000%	0.032%	0.050%	0.043%	0.041%	0.085%	
Commodity	0.000%	0.000%	0.005%	0.005%	0.009%	0.015%	0.054%
Average weekly return	0.044%	0.118%	0.256%	0.226%	0.242%	0.447%	0.053%

Stout Investments would like to achieve an average weekly return of 0.19%, or roughly a 10% annual return.

- a. Formulate and solve a Markowitz portfolio optimization model for this situation.
- b. Suppose the company wants to restrict the percentage of investments in each fund as follows:

Bond: between 10% and 50%

S&P 500: between 30% and 50%

Small cap: no more than 20%

Mid cap: no more than 20%

Large cap: no more than 20%

Emerging market: no more than 10%

Commodity: no more than 20%

How would the optimal portfolio change? Compare the solutions obtained using the GRG and Evolutionary algorithms provided by *Solver*.

19. Tejeda Investment Management, LLC manages 401K retirement plans. A client has asked them to recommend a portfolio. In discussing options with the client, 15 mutual funds were selected. The variance–covariance matrix and other relevant information are provided in the *Problem 19* tab of the *Chapter 14 Problem Data* Excel file (using the quotation symbols for these funds). The client would like to achieve a minimum return of at least 10%. The company also recommends the following:

- At least five funds be selected.
  - No more than three funds in the portfolio should have a Morningstar rating of three or less.
  - At most two funds should have an above average risk rating.
  - If a fund is selected, then at least 5% but no more than 25% should be invested in that fund.
- a. Formulate and solve an integer nonlinear optimization model to minimize the portfolio variance and meet the recommended restrictions. Use *Evolutionary Solver*.

- b. Suppose the client wants to maximize the expected return with the restriction that the portfolio

variance be no higher than 15%. How does the solution change?

20. Many manufacturing situations, such as the production of large and complex items as aircraft or machines, exhibit a learning effect in which the production time per unit decreases as more units are produced. This is often modeled by a power curve  $y = ax^{-b}$ , where  $a$  and  $b$  are constants. Suppose that data on production times for the first 10 units produced were collected from a new project:

Unit	Production Hours
1	3,161
2	2,720
3	2,615
4	2,278
5	2,028
6	2,193
7	2,249
8	2,268
9	1,994
10	2,000

- a. Develop a model for estimating the power curve to minimize the sum of the squared deviations of the errors. Use nonlinear optimization to find the parameters.
- b. Modify your model in part (a) to minimize the sum of the absolute deviations of errors. Use *Evolutionary Solver* to find the best parameters.

21. An IT support group at a university has seven projects to complete. The time in days and project deadlines are shown below.

Project	1	2	3	4	5	6	7
Time	4	9	12	16	9	15	8
Deadline	12	24	60	28	24	36	48

- a. Sequence the projects to minimize the average lateness.
  - b. Sequence the projects to minimize the average tardiness.
  - c. Compare these solutions to the SPT and EDD rules.
22. The *traveling salesperson problem* involves finding an optimal route (called a *tour*) that visits each of  $n$  cities exactly once and returns to the start. For example, suppose the distances between medical offices for a pharmaceutical representative are:

From	To							
	1	2	3	4	5	6	7	8
1	999	19	57	51	49	4	12	92
2	19	999	51	10	53	25	80	53
3	57	51	999	49	18	30	6	47
4	51	10	49	999	50	11	91	38
5	49	53	18	50	999	68	62	9
6	4	25	30	11	68	999	48	94
7	12	80	6	91	62	48	999	9
8	92	53	47	38	9	94	9	999

Note that the distance from one location to itself is an arbitrarily high number, 999. An example of a tour is 1-4-2-7-8-3-6-5-1. The total distance traveled would be  $51 + 10 + 80 + 9 + 47 + 30 + 68 + 49 = 344$ . The objective is to find the minimum distance tour. Set up and solve this problem using *Evolutionary Solver*. (Hint: Use the INDEX function to find the distance between locations and the *alldifferent* constraint in *Solver*. However, ensure that your solution goes back to the starting location.)

23. For the project selection example, suppose that the returns in the objective function are normally distributed with means as given by the expected returns and standard deviation equal to 10% of the mean. However, also assume that each project has a success rate modeled as a Bernoulli distribution. That is, the return will be realized only if the project is successful (use the "Yes-No" distribution in *Crystal Ball*). Success rates for the five projects are 0.80, 0.70, 0.90, 0.40, and 0.60, respectively. Modify the spreadsheet model, and use *OptQuest* to find a solution that maximizes the mean return.

## Case

### Tindall Bookstores

Tindall Bookstores<sup>4</sup> is a major national retail chain with stores located principally in shopping malls. For many years, the company has published a Christmas catalog that was sent to current customers on file. This strategy generated additional mail-order business while also attracting customers to the stores. However the cost-effectiveness of this strategy was never determined. In 2008, John Harris, vice president of marketing, conducted a major study on the effectiveness of direct-mail delivery of Tindall's Christmas catalog. The results were favorable: Patrons who were catalog recipients spent more, on average, than did comparable nonrecipients. These revenue gains more than compensated for the costs of production, handling, and mailing, which had been substantially reduced by cooperative allowances from suppliers.

With the continuing interest in direct mail as a vehicle for delivering holiday catalogs, Harris continued to investigate how new customers could most effectively be reached. One of these ideas involved purchasing mailing lists of magazine subscribers through a list broker. In order to determine which magazines might be more appropriate, a mail questionnaire was administered to a sample of current customers to ascertain which magazines they regularly read. Ten magazines were selected for the survey. The assumption behind this strategy is that subscribers of magazines that a high proportion of current customers read would be viable targets for future purchases at

Tindall stores. The question is which magazine lists should be purchased to maximize reaching of potential customers in the presence of a limited budget for the purchase of lists.

Data from the customer survey have begun to trickle in. The information about the 10 magazines a customer subscribes to is provided on the returned questionnaire. Harris has asked you to develop a prototype model, which later can be used to decide which lists to purchase. So far only 53 surveys have been returned. To keep the prototype model manageable Harris has instructed you to go ahead with the model development using the data from the 53 returned surveys. These data are shown in Table 14.4. The costs of the first ten lists are given below and your budget is \$3,000.

**Data for Tindall Bookstores Survey**

List	1	2	3	4	5	6	7	8	9	10
Cost(000)	\$1	\$1	\$1	\$1.5	\$1.5	\$1.5	\$1	\$1.2	\$5.5	\$1.1

What magazines should be chosen to maximize overall exposure? Conduct a budget sensitivity analysis on the Tindall magazine list selection problem. Solve the problem for a variety of budgets and graph percentage of total reach (number reached 53) versus budget amount. In your opinion, when is an increment in budget no longer warranted (based on this limited data)?

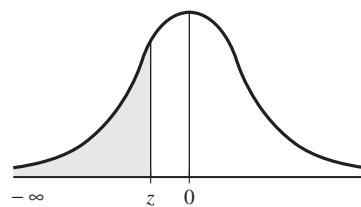
<sup>4</sup>This case is based on the case by the same name developed by James R. Evans of the University of Cincinnati and was sponsored by the Direct Marketing Policy Center.

**TABLE 14.4 Survey Results**

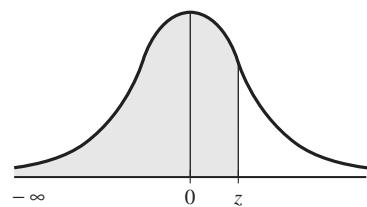
Customer	Magazines	Customer	Magazines
1	10	28	4, 7
2	1, 4	29	6
3	1	30	3, 4, 5, 10
4	5, 6	31	4
5	5	32	8
6	10	33	1, 3, 10
7	2, 9	34	4, 5
8	5, 8	35	1, 5, 6
9	1, 5, 10	36	1, 3
10	4, 6, 8, 10	37	3, 5, 8
11	6	38	3
12	3	39	2, 7
13	5	40	2, 7
14	2, 6	41	7
15	8	42	4, 5, 6
16	6	43	NONE
17	4, 5	44	5, 10
18	7	45	1, 2
19	5, 6	46	7
50	2, 8	47	1, 5, 10
21	7, 9	48	3
22	6	49	1, 3, 4
23	3, 6, 10	50	NONE
24	NONE	51	2, 6
25	5, 8	52	NONE
26	3, 10	53	2, 5, 8, 9, 10
27	2, 8		

## *Appendix*

- TABLE A.1 THE CUMULATIVE STANDARD NORMAL DISTRIBUTION 534
- TABLE A.2 CRITICAL VALUES OF  $t$  536
- TABLE A.3 CRITICAL VALUES OF  $\chi^2$  539
- TABLE A.4 CRITICAL VALUES OF  $F$  540
- TABLE A.5 CRITICAL VALUES OF THE STUDENTIZED RANGE Q 543

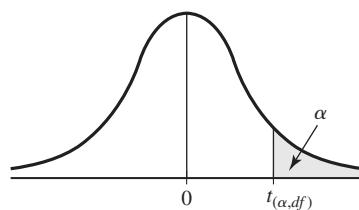
**TABLE A.1** The Cumulative Standard Normal Distribution

<b><i>z</i></b>	<b>.00</b>	<b>.01</b>	<b>.02</b>	<b>.03</b>	<b>.04</b>	<b>.05</b>	<b>.06</b>	<b>.07</b>	<b>.08</b>	<b>.09</b>
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00103	.00100
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2388	.2358	.2327	.2296	.2266	.2236	.2006	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2482	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

**TABLE A.1** (Continued)

<b><i>z</i></b>	<b>.00</b>	<b>.01</b>	<b>.02</b>	<b>.03</b>	<b>.04</b>	<b>.05</b>	<b>.06</b>	<b>.07</b>	<b>.08</b>	<b>.09</b>
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7518	.7549
0.7	.7580	.7612	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9089	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99897	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99997	.99997	.99997

Entry represents area under the cumulative standardized normal distribution from  $-\infty$  to  $z$ .

**TABLE A.2 Critical Values of  $t$** 

Degrees of Freedom	Upper Tail Areas					
	.25	.10	.05	.025	.01	.005
1	1.0000	3.0777	6.3138	12.7062	31.8207	63.6574
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0322
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.6864	1.3232	1.7207	2.0796	2.5177	2.8314
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6848	1.3178	1.7109	2.0639	2.4922	2.7969
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500
31	0.6825	1.3095	1.6955	2.0395	2.4528	2.7440
32	0.6822	1.3086	1.6939	2.0369	2.4487	2.7385
33	0.6820	1.3077	1.6924	2.0345	2.4448	2.7333
34	0.6818	1.3070	1.6909	2.0322	2.4411	2.7284
35	0.6816	1.3062	1.6896	2.0301	2.4377	2.7238

**TABLE A.2** (Continued)

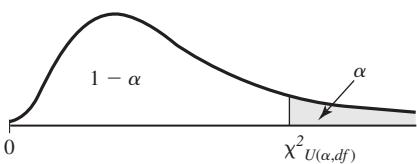
Degrees of Freedom	Upper Tail Areas					
	.25	.10	.05	.025	.01	.005
36	0.6814	1.3055	1.6883	2.0281	2.4345	2.7195
37	0.6812	1.3049	1.6871	2.0262	2.4314	2.7154
38	0.6810	1.3042	1.6860	2.0244	2.4286	2.7116
39	0.6808	1.3036	1.6849	2.0227	2.4258	2.7079
40	0.6807	1.3031	1.6839	2.0211	2.4233	2.7045
41	0.6805	1.3025	1.6829	2.0195	2.4208	2.7012
42	0.6804	1.3020	1.6820	2.0181	2.4185	2.6981
43	0.6802	1.3016	1.6811	2.0167	2.4163	2.6951
44	0.6801	1.3011	1.6802	2.0154	2.4141	2.6923
45	0.6800	1.3006	1.6794	2.0141	2.4121	2.6896
46	0.6799	1.3002	1.6787	2.0129	2.4102	2.6870
47	0.6797	1.2998	1.6779	2.0117	2.4083	2.6846
48	0.6796	1.2994	1.6772	2.0106	2.4066	2.6822
49	0.6795	1.2991	1.6766	2.0096	2.4049	2.6800
50	0.6794	1.2987	1.6759	2.0086	2.4033	2.6778
51	0.6793	1.2984	1.6753	2.0076	2.4017	2.6757
52	0.6792	1.2980	1.6747	2.0066	2.4002	2.6737
53	0.6791	1.2977	1.6741	2.0057	2.3988	2.6718
54	0.6791	1.2974	1.6736	2.0049	2.3974	2.6700
55	0.6790	1.2971	1.6730	2.0040	2.3961	2.6682
56	0.6789	1.2969	1.6725	2.0032	2.3948	2.6665
57	0.6788	1.2966	1.6720	2.0025	2.3936	2.6649
58	0.6787	1.2963	1.6716	2.0017	2.3924	2.6633
59	0.6787	1.2961	1.6711	2.0010	2.3912	2.6618
60	0.6786	1.2958	1.6706	2.0003	2.3901	2.6603
61	0.6785	1.2956	1.6702	1.9996	2.3890	2.6589
62	0.6785	1.2954	1.6698	1.9990	2.3880	2.6575
63	0.6784	1.2951	1.6694	1.9983	2.3870	2.6561
64	0.6783	1.2949	1.6690	1.9977	2.3860	2.6549
65	0.6783	1.2947	1.6686	1.9971	2.3851	2.6536
66	0.6782	1.2945	1.6683	1.9966	2.3842	2.6524
67	0.6782	1.2943	1.6679	1.9960	2.3833	2.6512
68	0.6781	1.2941	1.6676	1.9955	2.3824	2.6501
69	0.6781	1.2939	1.6672	1.9949	2.3816	2.6490
70	0.6780	1.2938	1.6669	1.9944	2.3808	2.6479
71	0.6780	1.2936	1.6666	1.9939	2.3800	2.6469
72	0.6779	1.2934	1.6663	1.9935	2.3793	2.6459
73	0.6779	1.2933	1.6660	1.9930	2.3785	2.6449
74	0.6778	1.2931	1.6657	1.9925	2.3778	2.6439
75	0.6778	1.2929	1.6654	1.9921	2.3771	2.6430
76	0.6777	1.2928	1.6652	1.9917	2.3764	2.6421
77	0.6777	1.2926	1.6649	1.9913	2.3758	2.6412
78	0.6776	1.2925	1.6646	1.9908	2.3751	2.6403
79	0.6776	1.2924	1.6644	1.9905	2.3745	2.6395
80	0.6776	1.2922	1.6641	1.9901	2.3739	2.6387

(Continued)

**TABLE A.2** (*Continued*)

Degrees of Freedom	Upper Tail Areas					
	.25	.10	.05	.025	.01	.005
81	0.6775	1.2921	1.6639	1.9897	2.3733	2.6379
82	0.6775	1.2920	1.6636	1.9893	2.3727	2.6371
83	0.6775	1.2918	1.6634	1.9890	2.3721	2.6364
84	0.6774	1.2917	1.6632	1.9886	2.3716	2.6356
85	0.6774	1.2916	1.6630	1.9883	2.3710	2.6349
86	0.6774	1.2915	1.6628	1.9879	2.3705	2.6342
87	0.6773	1.2914	1.6626	1.9876	2.3700	2.6335
88	0.6773	1.2912	1.6624	1.9873	2.3695	2.6329
89	0.6773	1.2911	1.6622	1.9870	2.3690	2.6322
90	0.6772	1.2910	1.6620	1.9867	2.3685	2.6316
91	0.6772	1.2909	1.6618	1.9864	2.3680	2.6309
92	0.6772	1.2908	1.6616	1.9861	2.3676	2.6303
93	0.6771	1.2907	1.6614	1.9858	2.3671	2.6297
94	0.6771	1.2906	1.6612	1.9855	2.3667	2.6291
95	0.6771	1.2905	1.6611	1.9853	2.3662	2.6286
96	0.6771	1.2904	1.6609	1.9850	2.3658	2.6280
97	0.6770	1.2903	1.6607	1.9847	2.3654	2.6275
98	0.6770	1.2902	1.6606	1.9845	2.3650	2.6269
99	0.6770	1.2902	1.6604	1.9842	2.3646	2.6264
100	0.6770	1.2901	1.6602	1.9840	2.3642	2.6259
110	0.6767	1.2893	1.6588	1.9818	2.3607	2.6213
120	0.6765	1.2886	1.6577	1.9799	2.3578	2.6174
$\infty$	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758

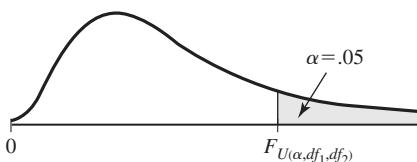
For a particular number of degrees of freedom, entry represents the critical value of  $t$  corresponding to a specified upper tail area ( $\alpha$ ).

**TABLE A.3** Critical Values of  $\chi^2$ 

Degrees of Freedom	Upper Tail Areas ( $\alpha$ )											
	.995	.99	.975	.95	.90	.75	.25	.10	.05	.025	.01	.005
1												
2	0.010	0.020	0.051	0.103	0.211	0.575	2.773	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.213	4.108	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.923	5.385	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	2.675	6.626	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	3.455	7.841	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	4.255	9.037	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	5.071	10.219	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	5.899	11.389	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	6.737	12.549	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	7.584	13.701	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	8.438	14.845	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	9.299	15.984	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	10.165	17.117	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	11.037	18.245	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	11.912	19.369	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	12.792	20.489	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	13.675	21.605	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	14.562	22.718	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	15.452	23.828	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	16.344	24.935	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	17.240	26.039	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	18.137	27.141	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	19.037	28.241	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	19.939	29.339	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	20.843	30.435	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	21.749	31.528	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	22.657	32.620	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	23.567	33.711	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	24.478	34.800	40.256	43.773	46.979	50.892	53.672

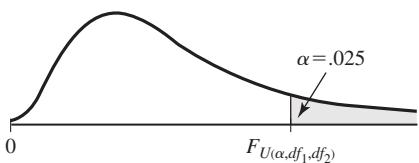
For a particular number of degrees of freedom, entry represents the critical value of corresponding  $\chi^2$  to a specified upper tail area ( $\alpha$ ).

For larger values of degrees of freedom (df) the expression  $Z = \sqrt{2\chi^2} - \sqrt{2(df) - 1}$  may be used, and the resulting upper tail area can be obtained from the table of the standard normal distribution (Table A.1a).

**Table A.4 Critical Values of  $F$** 


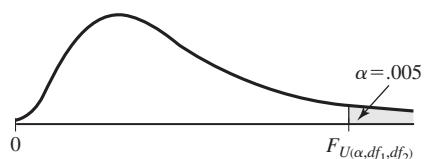
		Numerator $df_1$																		
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	$\infty$
Denominator $df_2$		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	$\infty$
1		161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2		18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3		10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4		7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5		6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6		5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7		5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8		5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9		5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.32	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10		4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11		4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12		4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13		4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14		4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15		4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16		4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17		4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18		4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19		4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20		4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21		4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22		4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23		4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24		4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25		4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26		4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27		4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28		4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29		4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30		4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40		4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60		4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120		3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
$\infty$		3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

**LE A.4 (Continued)**



Denominator df <sub>2</sub>	Numerator df <sub>1</sub>																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	$\infty$
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.36	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.37	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.68	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
$\infty$	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

(Continued)

**LE A.4 (Continued)**

		Numerator df <sub>1</sub>																		
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
Denominator df <sub>2</sub>		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1		16211	20000	21615	22500	23056	23437	23715	23925	24091	24224	24426	24630	24836	24940	25044	25148	25253	25359	25465
2		198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5	199.5
3		55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69	43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	41.83
4		31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.32
5		22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	12.14
6		18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.88
7		16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.65	7.53	7.42	7.31	7.19	7.08
8		14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	5.95
9		13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.19
10		12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	4.64
11		12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.24	5.05	4.86	4.76	4.65	4.55	4.44	4.34	4.23
12		11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.90
13		11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76	3.65
14		11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55	3.44
15		10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37	3.26
16		10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.10	3.92	3.73	3.64	3.54	3.44	3.33	3.22	3.11
17		10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10	2.98
18		10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.99	2.87
19		10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89	2.78
20		9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.69
21		9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	3.60	3.43	3.24	3.15	3.05	2.95	2.84	2.73	2.61
22		9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	3.54	3.36	3.18	3.08	2.98	2.88	2.77	2.66	2.55
23		9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60	2.48
24		9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.43
25		9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	3.37	3.20	3.01	2.92	2.82	2.72	2.61	2.50	2.38
26		9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49	3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45	2.33
27		9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45	3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	2.29
28		9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41	3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	2.25
29		9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38	3.21	3.04	2.86	2.76	2.66	2.56	2.45	2.33	2.21
30		9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.18
40		8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	2.95	2.78	2.60	2.50	2.40	2.30	2.18	2.06	1.93
60		8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74	2.57	2.39	2.29	2.19	2.08	1.96	1.83	1.69
120		8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54	2.37	2.19	2.09	1.98	1.87	1.75	1.61	1.43
∞		7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.36	2.19	2.00	1.90	1.79	1.67	1.53	1.36	1.00

a particular combination of numerator and denominator degrees of freedom, entry represents the critical values of  $F$  corresponding to a specified upper tail area ( $\alpha$ ).

**Table A.5 Critical Values<sup>a</sup> of the Studentized Range Q**

		Upper 5% Points ( $\alpha = 0.05$ )																		
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		18.00	27.00	32.80	37.10	40.40	43.10	45.40	47.40	49.10	50.60	52.00	53.20	54.30	55.40	56.30	57.20	58.00	58.80	59.60
2		6.09	8.30	9.80	10.90	11.70	12.40	13.00	13.50	14.00	14.40	14.70	15.10	15.40	15.70	15.90	16.10	16.40	16.60	16.80
3		4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46	9.72	9.95	10.15	10.35	10.52	10.69	10.84	10.98	11.11	11.24
4		3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83	8.03	8.21	8.37	8.52	8.66	8.79	8.91	9.03	9.13	9.23
5		3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17	7.32	7.47	7.60	7.72	7.83	7.93	8.03	8.12	8.21
6		3.46	4.34	4.90	5.31	5.63	5.89	6.12	6.32	6.49	6.65	6.79	6.92	7.03	7.14	7.24	7.34	7.43	7.51	7.59
7		3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30	6.43	6.55	6.66	6.76	6.85	6.94	7.02	7.09	7.17
8		3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05	6.18	6.29	6.39	6.48	6.57	6.65	6.73	6.80	6.87
9		3.20	3.95	4.42	4.76	5.02	5.24	5.43	5.60	5.74	5.87	5.98	6.09	6.19	6.28	6.36	6.44	6.51	6.58	6.64
10		3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72	5.83	5.93	6.03	6.11	6.20	6.27	6.34	6.40	6.47
11		3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61	5.71	5.81	5.90	5.99	6.06	6.14	6.20	6.26	6.33
12		3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.40	5.51	5.62	5.71	5.80	5.88	5.95	6.03	6.09	6.15	6.21
13		3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43	5.53	5.63	5.71	5.79	5.86	5.93	6.00	6.05	6.11
14		3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46	5.55	5.64	5.72	5.79	5.85	5.92	5.97	6.03
15		3.01	3.67	4.08	4.37	4.60	4.78	4.94	5.08	5.20	5.31	5.40	5.49	5.58	5.65	5.72	5.79	5.85	5.90	5.96
16		3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26	5.35	5.44	5.52	5.59	5.66	5.72	5.79	5.84	5.90
17		2.98	3.63	4.02	4.30	4.52	4.71	4.86	4.99	5.11	5.21	5.31	5.39	5.47	5.55	5.61	5.68	5.74	5.79	5.84
18		2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17	5.27	5.35	5.43	5.50	5.57	5.63	5.69	5.74	5.79
19		2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14	5.23	5.32	5.39	5.46	5.53	5.59	5.65	5.70	5.75
20		2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11	5.20	5.28	5.36	5.43	5.49	5.55	5.61	5.66	5.71
21		2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01	5.10	5.18	5.25	5.32	5.38	5.44	5.50	5.54	5.59
22		2.89	3.49	3.84	4.10	4.30	4.46	4.60	4.72	4.83	4.92	5.00	5.08	5.15	5.21	5.27	5.33	5.38	5.43	5.48
23		2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.74	4.82	4.91	4.98	5.05	5.11	5.16	5.22	5.27	5.31	5.36
24		2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73	4.81	4.88	4.94	5.00	5.06	5.11	5.16	5.20	5.24
25		2.80	3.36	3.69	3.92	4.10	4.24	4.36	4.48	4.56	4.64	4.72	4.78	4.84	4.90	4.95	5.00	5.05	5.09	5.13
26		2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55	4.62	4.68	4.74	4.80	4.85	4.89	4.93	4.97	5.01

(Continued)

**TABLE A.5 (Continued)**

		Upper 1% Points ( $\alpha = 0.01$ )																	
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
90.00	135.00	164.00	186.00	202.00	216.00	227.00	237.00	246.00	253.00	260.0	266.00	272.00	277.00	282.00	286.00	290.00	294.00	298.00	
14.00	19.00	22.30	24.70	26.60	28.20	29.50	30.70	31.70	32.60	33.40	34.10	34.80	35.40	36.00	36.50	37.00	37.50	37.90	
8.26	10.60	12.20	13.30	14.20	15.00	15.60	16.20	16.70	17.10	17.50	17.90	18.20	18.50	18.80	19.10	19.30	19.50	19.80	
6.51	8.12	9.17	9.96	10.60	11.10	11.50	11.90	12.30	12.60	12.80	13.10	13.30	13.50	13.70	13.90	14.10	14.20	14.40	
5.70	6.97	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48	10.70	10.89	11.08	11.24	11.40	11.55	11.68	11.81	11.93	
5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30	9.49	9.65	9.81	9.95	10.08	10.21	10.32	10.43	10.54	
4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55	8.71	8.86	9.00	9.12	9.24	9.35	9.46	9.55	9.65	
4.74	5.63	6.20	6.63	6.96	7.24	7.47	7.68	7.87	8.03	8.18	8.31	8.44	8.55	8.66	8.76	8.85	8.94	9.03	
4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.32	7.49	7.65	7.78	7.91	8.03	8.13	8.23	8.32	8.41	8.49	8.57	
4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36	7.48	7.60	7.71	7.81	7.91	7.99	8.07	8.15	8.22	
4.39	5.14	5.62	5.97	6.26	6.48	6.67	6.84	6.99	7.13	7.25	7.36	7.46	7.56	7.65	7.73	7.81	7.88	7.95	
4.32	5.04	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94	7.06	7.17	7.26	7.36	7.44	7.52	7.59	7.66	7.73	
4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79	6.90	7.01	7.10	7.19	7.27	7.34	7.42	7.48	7.55	
4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66	6.77	6.87	6.96	7.05	7.12	7.20	7.27	7.33	7.39	
4.17	4.83	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55	6.66	6.76	6.84	6.93	7.00	7.07	7.14	7.20	7.26	
4.13	4.78	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46	6.56	6.66	6.74	6.82	6.90	6.97	7.03	7.09	7.15	
4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38	6.48	6.57	6.66	6.73	6.80	6.87	6.94	7.00	7.05	
4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31	6.41	6.50	6.58	6.65	6.72	6.79	6.85	6.91	6.96	
4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25	6.34	6.43	6.51	6.58	6.65	6.72	6.78	6.84	6.89	
4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19	6.29	6.37	6.45	6.52	6.59	6.65	6.71	6.76	6.82	
3.96	4.54	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02	6.11	6.19	6.26	6.33	6.39	6.45	6.51	6.56	6.61	
3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85	5.93	6.01	6.08	6.14	6.20	6.26	6.31	6.36	6.41	
3.82	4.37	4.70	4.93	5.11	5.27	5.39	5.50	5.60	5.69	5.77	5.84	5.90	5.96	6.02	6.07	6.12	6.17	6.21	
3.76	4.28	4.60	4.82	4.99	5.13	5.25	5.36	5.45	5.53	5.60	5.67	5.73	5.79	5.84	5.89	5.93	5.98	6.02	
3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.38	5.44	5.51	5.56	5.61	5.66	5.71	5.75	5.79	5.83	
3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23	5.29	5.35	5.40	5.45	5.49	5.54	5.57	5.61	5.65	

Range /  $S \sim Q_{1-\alpha} : \nu : \eta$  •  $\eta$  is the size of the sample from which the range is obtained, and  $\nu$  is the number of degrees of freedom of  $S$ .

Source: H. L. Harter and D. S. Clemm, "The Probability Integrals of the Range and of the Studentized Range—Probability Integral, Percentage Points, and Moments of the Range," Wright Air Development Technical Report 58-484, Vol. I, 1959.

# INDEX

## A

Additive seasonality, 266

Adjusted R-square, 205, 233

Adjusted-rate mortgage, 372

Albuquerque Microelectronics Operation (AMO), 262–263

Algorithms, 127, 302–304, 440, 456, 473, 485, 495, 506

Algorithms, 302

Alternate optimal solutions, 446

Alternative hypothesis, 163

American Society for Quality (ASQ), 81

Analysis

of control charts, 280–284

data, 74–78, 273, 342

of decision models, 299–304

probability, 118

process capability, 288–290

risk, 325

sensitivity, 299, 339, 368

single-factor, 195

Tukey-Kramer multiple comparison procedure and, 184–185

using *Crystal Ball*, 342

what-if, 299–302

Analysis of variance (ANOVA), 205

assumptions of, 184

of entire model, 212

Excel and, 182–183

*PHStat* and, 185

regression as, 205, 231–233

single-factor, 195

sum of squares of errors, 232

theory of, 192

total sum of squares, 232

Tukey-Kramer multiple comparison procedure, 184–185

*Analysis Toolpak*, 43, 56–57

Analytical queuing models

Little's Law in, 408–409

long-term expected values provided by, 408

single-server, 407–408

steady-state values provided by, 408

transient period in, 408

Anderson-Darling test, 344

ANOVA. *See* Analysis of variance

Area chart, 48

Arithmetic mean(s), 64, 83

one-sample tests for, 169–170, 193

with paired samples, 179

sampling distribution of, 133–134, 275

in simulation statistics, 416

standard error of, 133

test differences between two populations, 193–194

two-sample tests for, 177–178, 191–192

ARM. *See* Adjusted-rate mortgage

Arrival process, 404–405

Arrival process, 404

ASQ. *See* American Society for Quality (ASQ)

Attributes

control chart for, 284–288

definition, 274

*Auditing* tool, 309

Autocorrelation, 209

Autoregressive forecasting models, 250–252

Autoregressive integrated moving average (ARIMA) model, 261

Average payoff strategy, 372–373

## B

Backward elimination, in stepwise regression, 217

Balance constraints, 452, 506

Balking, 405

Banana Republic, 311

Bar charts, 45–46

*Batch Fit* tool, 342

Bayes's rule, 387–389

Becker Consulting project management model, 354–356

Benefit/cost analysis, 370

Bernoulli distribution, 99

Bernoulli random variable, 99

Best-fitting models, 306

Best-subsets regression, 217–218, 236

Beta distribution, 111

Beta risk, 203

Beta value, 202

BG Seed Company, 454

Biased estimators, 136

Bimodal distributions, 67

Binary variables, 492–493

computer configuration using, 492

fixed costs modeled with, 497–499

integer (linear) optimization model, 487–495

modeling logical conditions using, 493–494

in plant location model, 496

project selection and, 487–488, 506

*Solver* and, 488

Binding constraint, 447

Binomial distribution, 99–100, 118–119, 282

Blending models, 456

Boeing Company, business data and analysis capabilities of, 28–29

*Bootstrap* tool, 343

multiple-simulation method of, 366

one-simulation method of, 366

in project management, 353–358

sampling distributions created by, 366

Bootstrapping, 343

Bounded variables model, 464–469

Box plot, 73

Box-and-whisker plots, 73

Breakeven probability, 391

Bubble chart, 49

Buffers, in *SimQuick*, 411–414, 424

*Business Conditions Digest*, 240

*Business Cycle Indicators*, 240

Business environment

in Boeing Company, 29

in data, 28–30

metrics and measurement in, 32

## C

Calling population, 404

Camm Textiles, 453

Carrying costs, 321, 461

Cash budgeting application, of Monte Carlo simulation, 349–352

Categorical independent variables with more than two levels, 223–225

regression analysis with, 220–223

Categorical (nominal) data, 33, 58

Causal forecasting methods. *See* Explanatory/causal forecasting methods

Causal variables, regression analysis with, 255–257

*CB Predictor*

*Data Attributes* dialog, 259, 262, 270

illustration of, 257–259

*Input Data* dialog, 259, 268–269

*Method Gallery* dialog, 259, 270

*Options* dialog, 271

*View Autocorrelations*, 259

Center line, hugging of, 281

Central limit theorem, 133

Charts. *See also* Control charts

area, 48

bar, 45–46

bubble, 49

clustered column, 45

column, 45–46

*Crystal Ball*, 339–341

data display with, 45–50

doughnut, 49

line, 47

pie, 47

radar, 49

spider, 353–354

stacked column, 45

Chebyshev's theorem, 65

Chi-square

distribution, 144, 174, 187

goodness-of-fit test, 344

statistic, 187

test for independence, 186–188, 195

test of variance, 174

values, 539

CK. *See* Coefficient of kurtosis (CK)

Cluster sampling, 126

Clustered column chart, 45

Coding of variables, 220

Coefficient of determination, 204

Coefficient of kurtosis (CK), 68, 84

Coefficient of multiple determination, 212

Coefficient of skewness (CS), 67

Coefficient of variation (CV), 66–67

Column charts, 45–46

bar *v.*, 46

clustered, 45–46

stacked, 45–46

Commercial simulation software, 426–427

Common causes of variation, 273–274

Complement of an event, 92

Computer configuration

integer (linear) optimization model

and, 491–494

using binary variables, 492

Concave downward utility function, 391

Concave upward utility function, 392

Conditional probability, 92–94, 387–389

Confidence band, 233

Confidence coefficient, 166

Confidence intervals, 233

“119% confident,” 137

common, 138

definition, 137

differences between means, 138

independent samples, 149, 155

**Confidence (continued)**  
differences between means, paired samples, 149  
differences between proportions, 150, 157–158  
estimates, 134  
formulas, 157–158  
for independent samples with equal variances, 156  
for independent samples with unequal variances, 155  
for mean with known population standard deviation, 138–140  
for mean with unknown population standard deviation, 140–142  
for paired samples, 156–157  
for a population total, 145–146  
prediction intervals, 148–149  
for a proportion, 142–143  
and sample size, 146–148  
theory underlying, 153–154  
use in decision making, 146  
for variance and standard deviation, 143–145

**Conservative decision making strategy**, 373

**Constraint function**, 437

**Constraints**, 303–304, 436, 451–452  
balance, 452  
binding, 447, 449, 502  
categories, 451–452  
function, 437, 459, 470, 488, 490, 496  
in linear optimization models, 453  
logical, binary variables modeling, 484–485

**Consumer Price Index**, 240

**Contingency table**, 72

**Continuous data**, 31

**Continuous metrics**, 31

**Continuous probability distribution**, 102–112  
beta distribution, 111  
exponential distribution, 109–110  
extreme value distribution, 112  
gamma distributions, 111  
geometric distribution, 111  
hypergeometric distribution, 111  
logistic distribution, 111  
lognormal distribution, 111  
negative binomial distribution, 111  
normal distribution, 105–108  
Pareto distribution, 112  
triangular distribution, 108–109  
uniform distribution, 104–105  
Weibull distributions, 111

**Continuous random variables**, 94

**Continuous simulation model**, 427–430

**Control charts**, 274  
analysis of, 280–284  
for attributes, 284–288  
center line of, 280–281  
control limits in, 280  
factors, 278  
out-of-control conditions, 281–284  
*p*-chart, 284–286  
process average, 281  
*R*-chart, 280  
for surgery infection rate using average sample size, 287  
trends, 281  
using Excel, 285–287  
using PHStat, 286–287  
*x*-chart, 280

**Control limits**, 274–275  
hugging of, 282–284

**Convenience sampling**, 125

**Copenhagen Telephone Company**, 403

**CORREL function**, 87

**Correlation**, 69  
coefficient, 69–70, 85, 204  
examples, 70–71  
Excel tool of, 86–87, 235  
multiple linear regression and, 212–214

**Correlation Matrix tool**, 342, 351, 365

**Costs**  
carrying, 321, 461  
fixed, 497–499  
holding, 321  
ordering, 321  
reduced, 448  
reduced, interpretation of, 461

**Covariance**, 85

**Covariance and Portfolio Analysis tool**, 379

**Covering problem**, 488

**Cp statistic**, 218, 290

**Critical path**, 315

**Critical value**, 167, 169–171, 173–175, 178, 181–183, 209, 536–544

**Cross-tabulation table**, 72

**Crude oil decision model**, 306–307

**Crystal Ball suite**, 43, 104, 257  
assumption charts, 339  
assumptions, 328–332  
assumptions in, 517  
*Batch Fit* tool, 342  
*Bootstrap* tool, 343, 366  
for cash budgeting, 349–352  
*Correlation Matrix* tool, 342, 365  
custom distribution in, 347–348  
*Decision Table* tool, 343–347  
*Define Assumption* dialog, 344, 356, 362–363  
defining uncertain model inputs, 328–332  
distribution fitting with, 363–365  
fitting a distribution with, 344  
forecast cells, 332  
forecast chart, 520  
forecast charts, 339  
*Freeze* command, 339  
function *CB.Normal(0.29, 0.27)*, 429  
modeling and analysis tools, 342–343  
Monte Carlo simulation. *See Monte Carlo simulation*

for new product introduction, 352–353

**NewsVendor Model**, 344–348

for overbooking decisions, 348–349

overlay charts, 339–341

for project management, 353–358

random variate generation functions, 342

reports and data extraction, 342

risk analysis and optimization, 513–514

running a simulation, 332–334

saving runs of, 334–338

scatter charts, 340

*Scenario Analysis* tool, 343

sensitivity chart, 339–340, 363

spreadsheet model, 516–517

steps to start, 327

*Tornado Chart* tool, 342–343, 365–366

trend chart, 340–341

*Two-dimensional Simulation* tool, 343

uncertain activity time data, 355

**Crystal Ball version35.25.26.25**, 261

**CS. See Coefficient of skewness (CS)**

**Cumulative distribution function**, 97–98

**Cumulative relative frequency**, 61

**Cumulative standard normal distribution**, 534–535

**Curvilinear regression model**, 226

**Customer-focused outcomes**, 30

**Cutting pattern**, 483

**Cutting stock problem**  
integer optimization model for, 483–484  
optimal linear solution to, 485–486  
*Solver* add-in for, 485–486

**CV. See Coefficient of variation (CV)**

**Cycles**, in control charts, 281

**Cyclical effects**, on time series, 240, 242–243

**D**

**Dantzig, George**, 473

**Data**, 296  
analysis using PivotTables, 74–78  
in business environment, 28–30  
categorical (nominal), 33  
continuous, 31  
discrete, 31  
display with Excel charts, 45–50  
files available on companion website, 32  
fitting, in decision models, 306–308  
interval, 34–35  
numerical, 58–62  
ordinal, 34  
percentiles, 62  
profiles, 62–63  
ratio, 35  
sources and types, 30–35  
types of, 33  
variables, 274

**Data analysis**  
*Crystal Ball*, 342  
in quality control, 273

**Data tables**, 299–301  
creation of, 322  
one-way, 299–300  
two-way, 300

**Data Validation tool**, 309

**Deciles**, 63

**Decision alternatives**, 368

**Decision making**  
aggressive strategy of, 373  
analysis of portfolio risk, 378–379  
average payoff strategy of, 372–373  
under certainty, 368–371  
confidence intervals for, 146  
conservative strategy of, 373  
decision maker's certainty equivalent, 390  
decision trees in, 381–384  
expected value, 377–380  
flaw of averages, 380  
involving uncertainty and risks, 371–377  
for a maximize objective, 374–375  
for a minimize objective, 372–374  
mutually exclusive alternatives, 370–371  
non-mutually exclusive alternatives, 369–370  
opportunity loss strategy of, 374  
under risk and variability, 375–377  
with sample information, 386–387  
with single alternative, 369  
with uncertain information, 371–372  
utility and, 389–394

**Decision models**, 29, 296–299  
analysis of, 299–304  
business principles behind, 304–305

data fitting in, 306–308  
decision makers' intuition and, 298  
descriptive, 298  
forms of, 298  
for gasoline consumption, 305  
logic principles behind, 304–305  
mathematical functions used in, 305  
mathematical model, 297  
Monte Carlo simulation for, 326–327  
for new product development, 309–311  
optimization model, 298, 302–304  
for outsourcing, 296–297  
for overbook reservations, 312–313  
prescriptive, 298  
for pricing, 298–299  
for project management, 313–315  
revenue model, 307  
single period purchase decisions,  
    311–312  
spreadsheet engineering and, 308–309  
on spreadsheets, 296, 309–315  
tools for, 304–309

Decision node, 381  
Decision points, in *SimQuick*, 411, 421–424  
Decision rules, in hypothesis testing,  
    166–169

Decision strategy, 382  
*Decision Table* tool, 343–347, 349, 515

Decision tree  
    branches, 381  
    decision node, 381  
    defined, 381  
    for determining utility, 390  
    drug development model and, 377  
    event node, 381  
    expanded, 401  
    expected value decisions and, 377–380  
    nodes, 381  
    for pharmaceutical R&D model,  
        381–382  
    and risk, 382–384  
    sensitivity analysis in, 384  
    *TreePlan* tool for, 381–382, 399–401  
    utility and, 390

Decision variables, 296, 487  
Defects per million opportunities (dpmo),  
    36  
Define Assumption, customization of,  
    328–329, 332  
Degrees of freedom (*df*), 141–142, 144, 170,  
    174, 179, 181, 187, 536–538  
Delphi method, 239  
Deming, W. Edwards, 273  
Department of Commerce, 240  
Department of Energy (DOE), 262  
Descriptive decision models, 298  
Descriptive statistics, 37, 56–57, 288  
    for categorical data, 71–72  
    correlation and, 85  
    Excel tool of, 68, 85–88, 289  
    frequency distributions and, 57–63  
    functions and tools, 57  
    histograms and, 57–63  
    measures of association and, 69–71  
    measures of dispersion and, 64–67  
    measures of location and, 63–64  
    measures of shape and, 67–68  
    for numerical data, 63–71  
    *PHStat* tool of, 73  
    theory and computation, 83–85  
Design specifications, variation and,  
    288–290  
Deterministic activity time, 313, 404–405  
Deterministic arrivals, 404

*df*. See Degrees of freedom  
DIRECTV, 263  
Discount rate, 323  
Discounted cash flow, 323  
Discrete data, 31  
Discrete metric, 31  
Discrete probability distribution, 97–101  
Discrete random variable, 94, 118  
Discrete uniform distribution, 105  
Dispersion, defined, 64  
Distributions. *See* Probability distributions  
DMAIC (define, measure, analyze,  
    improve, and control), 36  
DOE. *See* Department of Energy  
Dot-scale diagram, 73  
Double exponential smoothing, 265–266  
Double moving average, 265  
Doughnut chart, 49  
dpmo. *See* Defects per million  
    opportunities (dpmo)  
Dummy variables, 220  
Durbin–Watson statistic, 209, 257

**E**

Econometric models. *See* Explanatory /  
    causal forecasting methods  
Empirical probability distribution, 95–96  
Empirical rules, 65  
EMV. *See* Expected monetary value (EMV)  
Endpoint grabbers, 334  
Entrances, in *SimQuick*, 411  
Erlang distribution, 111  
Error metrics, 244  
Estimation, 134  
    interval estimate, 137  
    point estimates, 134–136  
    unbiased estimates, 136  
Event, 261, 371  
    mutually exclusive, 92  
    nodes, 381  
    quantitative *v.* qualitative, 371  
Evolutionary algorithm, 507  
*Evolutionary Solver*, 506–512  
Excel, 29  
    add-ins, 43–44  
    *Analysis Toolpak* add-in, 43–44, 56–57,  
        59, 129, 135, 168, 201, 242  
    *Auditing* tool, 309  
    basic probability distribution support,  
        96  
    basic skills, 38  
    *Beverage Sales*, curvilinear regression of,  
        226–227  
    cash flows in, 323  
    categorical variables with more than  
        two levels, 223–225  
    *CB Predictor*, 242  
    cell references, 39–40  
    *Chart Wizard*, 62  
    column and row width, 43  
    copying formulas, 38–39  
    *Correlation* tool, 235  
    data display with charts, 45–50  
    *Data Table* tool, 301, 322  
    *Data Validation* tool, 309  
    *Descriptive Statistics* tool, 68–69  
    discount rate, 323  
    displaying formulas in worksheets, 43  
    *Expected Monetary Value* tool, 399–400  
    exponential smoothing, 267–268  
    filling a range with a series of numbers,  
        43  
    function for internal rate of return, 399  
    functions, 40–42

functions to avoid in modeling linear  
    programs, 441–442  
Goal Seek tool, 302, 323  
Histogram tool, 62  
measure of dispersion, 65  
moving averages method, 267  
net present value, 323  
normal probabilities, 105–108  
NORMINV function, 326  
one-way data table, 322  
Paste Special, 42–43  
for process capability analysis, 289  
*Project Selection Model* spreadsheet,  
    487–488  
RAND function, 327, 349  
random sampling, 158–159  
Rank and Percentile tool, 62  
Regression tool, 233–234  
ribbon, 38  
ROUND function, 326  
*Scenario Manager* tool, 301, 322–323  
Science and Engineering Jobs worksheet,  
    40  
simple linear regression in, 203–206  
*Solver* add-in, 302–304  
split screen, 42  
standard deviation, 65  
statistical functions and tools, 57  
SUMPRODUCT function, 441, 490  
*Surgery Infections* file, 287  
*Syringe Samples* file, 276–277  
*TreePlan* add-in, 381, 400–401  
Trendline tool, 233, 306  
VLOOKUP function, 129, 159  
Exits, in *SimQuick*, 411, 424–426  
Expanded decision tree, 401  
Expected monetary value (EMV), 377  
    best expected payoff in, 377  
    decision tree as example of, 382–384  
EVPI and, 385  
flaw of averages in, 380  
opportunity loss and, 385  
*PHStat* tool for, 385  
portfolio risk analysis and, 378–379  
Expected opportunity loss, 385  
Expected value  
    of a discrete random variable, 118  
    of perfect information (EVPI), 385  
    of a random variable, 98  
    of sample information (EVSI), 385  
Experiment, defined, 90  
Explanatory/causal forecasting methods,  
    238  
Exponential distribution, 120  
Exponential functions, 305  
Exponential smoothing, 241–242  
    double, 248, 255, 259, 265–266  
    forecasting with, 267–268  
    models, 246–247  
Exponential utility function, 393–394  
Extreme value distribution, 112

**F**

*F* values, 540–542  
Factor, 183  
FCFS. *See* First come, first served (FCFS)  
Feasible solutions, 442  
FedStats, 31  
Financial outcomes, 30  
Finite population correction (FPC), 138  
First come, first served (FCFS), 405  
First-order autoregressive model, 250  
Fixed costs, 497–499  
Flaw of averages, 380

Forecasting, 198  
advanced models of, 249–257  
autoregressive models, 249–252  
with *Crystal Ball* suite of applications, 332  
double exponential smoothing model, 248  
double moving average model, 248  
explanatory/causal forecasting methods, 238  
of gross domestic product (GDP), 239–240  
Holt–Winters additive model, 266–267  
Holt–Winters multiplicative model, 267  
indicators and indexes for, 239–240  
judgmental, 238–240  
models for time series with a linear trend, 248–249  
Monte Carlo simulation and, 326–327  
practice of forecasting, 262–263  
regression-based, 248–249  
seasonal additive model, 266  
seasonal multiplicative model, 266  
for stationary time series, 242–249  
statistical forecasting methods, 240–242  
using *CB Predictor*, 257–262

Forrester, Jay, 427

Forward selection, in stepwise regression, 217

FPC. *See* Finite population correction (FPC)

Fractiles, 63

Fraction defective chart, 284

Fraction nonconforming chart, 284

Frame, 124

F-ratio, 233

Frequency distribution, 57–59, 64

creation of, 85

cumulative relative, 61

in Excel, 59

relative, 58

Frontline Systems, Inc., 302

F-statistic, 212

F-test, 205

## G

Gamma distributions, 111

GDP. *See* Gross Domestic Product (GDP)

General Appliance Corporation (GAC), 457

General integer variables, 483–484

General integer variables, 483

Generalized reduced gradient (GRG), 501

Genetic algorithms, 506

Geometric distribution, 111

*Goal Seek* tool, 302

Goodness-of-fit tests, 344

Graphs, 45, 49

box-and-whisker plots, 73

stem-and-leaf displays, 73

GRG. *See* Generalized reduced gradient

Grocery store checkout model with resources, 418–421

Gross Domestic Product (GDP), 239

## H

HATCO, Inc., 191

Heuristics, 304

Histograms, 59–60

  bimodal distributions, 67

Monte Carlo simulation and, 327

unimodal distributions, 67

Historical analogy, 239

Holding costs, 321

Holt, C.C., 255

Holt–Winters multiplicative method, 259

*Home Market Value* data, 74

Homoscedasticity, 209

Hugging the center line, 281

Hugging the control limits, 282–284

Hurdle rate, 369

Hypergeometric distribution, 111

Hypothesis testing

  alternative hypothesis, 163, 165, 167

  common types, 168

  critical value, 167, 169

  decision rules, 166–169

  definition, 163

  error types in, 166

  Excel support for, 168

  formulation of hypothesis, 164–165

  null hypothesis, 163, 171

  one-sample hypothesis test, 164–165

  rejection regions in, 167

  significance level, 165–166

  spreadsheet support for, 169

  steps involved, 164

  two-sample hypothesis test, 164–165

  in U.S. legal system, 163–164

## I

In-control processes, 274–275, 280, 282, 286, 288, 290

Independent events, 94

Independent sample, 149

Index of Leading Indicators, 240

Indexes

  for forecasting, 239–240

  indicators in, 240

  of leading indicators, 240

Indicators

  for forecasting, 239–240

  in index, 240

  lagging, 240

  leading, 239–240

Infeasibility, 446

Influence diagram, for medical services cost, 427–428

Innis Investments, 456

Integer (linear) optimization model

  with binary variables, 487–495

  computer configuration and, 491–494

  cutting stock problem and, 483–484

  optimal distribution center (DC)

  locations, 494

  project selection, 487–488

  site location, 488–491

  solving of, 484–486

  supply chain facility location, 494–495

Interaction, defined, 221

INTERCEPT function, 201

Internal rate of return (IRR), 369

Interquartile range (IQR), 65

Interval data, 34–35

Interval estimate, 137

Intuition, decision models and, 298–299

Inventory management decision model, 321

Investment return payoff, 390

Investment risk

  simple regression and, 202–203

  specific, 202

  systematic, 202

  types of, 202

IQR. *See* Interquartile range (IQR)

IRR. *See* Internal rate of return (IRR)

## J

J&G Bank, 362

J&M Manufacturing linear optimization model, 464–469

Jockey, 405

Joint probability distribution, 113–114

Judgment sampling, 125

Judgmental forecasting methods, 238–240

  Delphi method, 239

  historical analogy, 239

  indicators and indexes, 239–240

Juran, Joseph M., 273

## K

Kolmogorov–Smirnov test, 344

K<sup>th</sup> percentile, 62

Kurtosis, 68, 84

## L

Lagging indicators, 240

Lagrange multipliers, 502, 504

Laplace strategy. *See* Average payoff strategy

Last come, first served (LCFS), 405

Latin Hypercube sampling, 334

Leadership and governance outcomes, 30

Leading indicators, 239–240

Least-squares fit, 203

Least-squares regression, 200–202

Level of confidence, 137

Level of significance, of the test, 166

Limitations, 452

Limits Report, 450

Line charts, 47

Linear functions, 305

Linear optimization

  applications of, 446–472

  blending models of, 454–456

  bounded variables model of, 464–469

  building models of, 436–440

  characteristics of, 439–440

  constraints, identifying, 438

  constraints in models of, 453

  decision variables, 438

  Deercrest model, 437–439

  financial investment planning, 456–457

  generic examples, 452

  interpretation of reduced costs, 461

  Jordanelle model, 437–439

  mathematical expressions, 438–439

  multiperiod optimization models, 463–464

  multiperiod production planning, 461–463

  objective function, identifying, 438

  portfolio investment models of, 456–457

  process selection models, 453–454

  production/marketing allocation model of, 469–472

  spreadsheet models for optimization problems, 440–442

  SSC model of, 442–446

  transportation model of, 457–460

  using Solver, 440–442

Linear program, 439

Linearity, 208

Little, John D.C., 408

Little's Law, 408–409

L.L. Bean, 263

Local optimum solution, 503

Location parameter, 104

Logarithmic functions, 305

Logistic distribution, 111  
Lognormal distribution, 111  
Lot size, 424  
Lower specification limit (LSL), 275, 290

## M

MAD. *See* Mean absolute deviation (MAD)

Make-or-buy decisions, 453

Malcolm Baldrige Award Criteria for Performance Excellence, 30, 81–82

Manual process simulation, of single-server queue, 410

Manufacturing inspection model with decision points, 421–424

Manufacturing processes, 417

Marginal probability distribution, 113–114

Market researchers, 124

Market value, as a function, 197

Markowitz portfolio model, 503–506

risk vs return profile for, 505

sensitivity report, 504–505

variance of a portfolio, 503

Mathematical functions, used in models  
exponential, 305–306  
linear, 305–306  
logarithmic, 305–306  
polynomial, 305–306  
power, 305

Mathematical model, 297

Maximax strategy, 374

Maximin strategy, 374

Mean. *See* Arithmetic mean

Mean absolute deviation (MAD), 244

Mean absolute percentage error (MAPE), 244

Mean interarrival time, 404

Mean queuing statistics, as a function of simulation run time, 416

Mean square error (MSE), 244–245

Measurement, 31

Measures, 31

Measures of association, 69–71

Measures of central tendency, 56

Measures of dispersion, 64–67

Measures of location, 63–64

Measures of shape, 67–68

Median, 64

Metaheuristics, 506

Metric, 31

Midrange, 64

Minimax regret strategy, 374

Mixed integer linear optimization model  
definition, 483

with fixed costs, 497–499

plant location model, 495–497

Mixture, 282

Mode, 64

Model, defined, 296

Monte Carlo sampling, 326

Monte Carlo simulation, 257, 377

analyzing results of, 334–338

cash budgeting application of, 349–352

forecasting and, 332

new product development application of, 352–353

newsvendor model application of, 343–348

for outsourcing decision model, 326

overbooking model application of, 348–349

project management and, 353–358

running of, 332–334

saving runs of, 334  
uncertainty and, 328–332

Moore Pharmaceuticals model, 310, 327, 369

Mortgage instruments, 370

Moving average methods, 241

double, 265

Excel support for, 242–244

simple, 242

weighted, 244

MSE. *See* Mean square error (MSE)

Multicollinearity, 212–214

Multiple correlation coefficient, 212

Multiple linear regression

for the *College and Universities* data, 211–212

correlation and, 212–214

correlation matrix for variables in  
*Colleges and Universities* data, 213

form of, 210

interpreting results from, 212

model for La Quinta Motor Inns

proposed sites, 210

multicollinearity, 212–214

Multiple regression, 198

Multiplication law of probability, 93

Multiplicative seasonality, 266

Multivariate variables, 33

Mutually exclusive event, 92

## N

National Institute of Standards and  
Technology (NIST), 81

Negative binomial distribution, 111

Net present value (NPV), 323, 369

New product development

decision model for, 309–311

Monte Carlo simulation and, 352–353

Newsvendor model

*Crystal Ball* suite implementation of,  
344–348

flaw of averages in, 380

mutually exclusive alternatives,  
370–371

using Monte Carlo simulation, 343–348

NIST. *See* National Institute of Standards  
and Technology (NIST)

Nonlinear optimization model, 483

hotel pricing, 499–501

Markowitz portfolio model, 503–506

*Solver* and, 504–505

solving of, 501–503

Non-mutually exclusive alternatives,  
369–370

Nonnegativity constraint, 436

Nonparametric test, 184

Nonrejection region, 167

Nonsampling error, 127

Nonsmooth optimization, 506–512

constraints needed, 506

job sequencing model, 509–511

rectilinear location model using,  
508–509

Normal distribution, 105–108, 119

standard, 105

Normality of errors, 208–209

*NORMINV* function, 326

NPV. *See* Net present value (NPV)

Null hypothesis, 163

Numerical data, 58–62

## O

Objective coefficient, 448

Objective function, 436

Observed significance level, 171

Ogive, 61

One-sample hypothesis test, 164–165

for means, 169–170

for proportions, 172–174

*p-value*, 171–172

for variance, 174–175

One-simulation method, of *Bootstrap* tool,  
366

One-tailed tests of hypothesis, 167

One-way data table, 299

Operating characteristics, of queuing  
systems, 406

Opportunity loss strategy, 374

Opportunity loss table, 385

Optimal solution, 436

Optimization model, 298, 302–304

airline pricing model, 303–304

algorithms, 302

*Changing Variable Cells*, 303

constraints in, 303–304

heuristics in, 304

risk analysis and, 512–514

*Solver* add-in, 302–304

Optimization models, 298

*OptQuest*

*Add Requirement* button, 523

adding a requirement, 521–523

basic process for using, 516–524

*Constraints* screen, 519

creating new file, 517

*Decision Variables* screen, 517–519

*Define Decision* option, 517

interpretation of results, 520–521

*Performance Chart*, 520

portfolio allocation model, 515–516

*Run* button, 520

run options, 519–520

solving optimization problem, 520

Order quantity, 424

Ordering costs, 321

Ordinal data, 34

Outcomes of an experiment, 90

Outliers, 64, 74

Out-of-control conditions, 274, 280–282

Outsourcing, decision model for, 296–297

Overbooking model, 348–349

Overlay chart, 339

## P

Paired samples, 149

Parallel servers, 405

Pareto distribution, 112

Parsimony, 220

Partial regression coefficients, 211

Payback period, 369

Payoff table, 372

*p-charts*

calculations worksheet, 285–286

construction of, 292

*PHStat* and, 285–286, 292

Percentiles, 62

Perfect information, 385

Periodic sampling. *See* Systematic  
(periodic) sampling

PERT. *See* Program Evaluation and  
Review Technique (PERT)

*PHStat*. *See* Prentice-Hall Statistics (*PHStat*)

Pie charts, 47

*PivotTables* tool, 74–78

Plant location mixed linear optimization  
model, 495–497

Point estimates, 134–136

Poisson distribution, 100–101, 119

Poisson process, 404

- Polynomial functions, 305  
Population, 36  
  calling, 404  
  frame, 124  
  sampling error and, 127  
Portfolio  
  allocation model using *Crystal Ball* and *OptQuest*, 515–523  
  definition of, 378  
  Markowitz, 503–506  
Power curve, 177  
Power functions, 305  
Power of a test, 166, 177  
Prediction intervals, 148–149, 233  
Predictive statistics, 37  
*Premium Solver*, 302, 442  
*Prentice-Hall Statistics (PHStat)*, 43–44, 73, 77, 97  
best-subsets regression, 201, 218, 236  
*Calculations Worksheet for R & x Bar Charts*, 279–280  
for chi-Square test of variance, 174  
computation of Durbin–Watson statistic, 210  
confidence and prediction intervals, 206–207  
confidence intervals, 160–161  
*Correlation* tool, 235  
creating a frequency distribution and histogram, 85  
creating box plots, 87  
creating PivotTables, 87  
decision making, 379  
determining sample size, 161  
expected monetary value (EMV) tool, 378  
*Expected Monetary Value* tool, 379, 385, 399–400  
generating probabilities in, 121–122  
hypothesis tests, 168  
linear regression, 201  
*Multiple Regression* tool, 201, 214  
normal probability tools, 120–121  
*One-Way and Two-Way Tables & Charts*, 87–88  
*One-Way ANOVA* tool, 185  
*p*-chart calculations worksheet, 285–286  
portfolio risk analysis, 379  
probability distributions in, 110  
*Random Sample Generator* tool, 125  
random sampling tools, 158–161  
*Regression* tool, 235  
sampling from probability distributions, 159–160  
*Stepwise Regression* tool, 201, 217, 235–236  
*t*-test statistic, 171  
two-tailed *t*-test, 172  
using correlation tool, 86–87  
using descriptive statistics tool, 85–86  
*VLOOKUP* function, 159  
*x*- and *R*-charts, 291  
Prescriptive decision models, 298  
Principle of insufficient reason, 370  
Probabilistic activity time, 313, 353, 404–405  
Probabilistic sampling, 125  
Probability  
  basic concepts, 90–94  
  classical definition of, 90  
  complement of an event, 92  
  conditional, 92–94, 387–389  
  experiment, defined, 90  
  independent events, 94  
  multiplication law of, 93  
  outcomes of an experiment, 90  
  for quality measurements, 118  
  relative frequency definition of, 91  
  rules and formulas, 91–92  
  subjective definition of, 91  
  of a type II error, 166  
Probability density function, 102  
Probability distributions, 94–95  
  Bernoulli, 99, 159  
  beta, 111  
  binomial, 99–100  
  continuous, 102–112  
  cumulative distribution function, 103  
  discrete, 97–101  
  empirical, 95–96  
  expected value, of a random variable, 98  
  exponential, 109–110  
  extreme value, 112  
  fitting to data, 344, 363–365  
  gamma, 111  
  geometric, 111  
  hypergeometric, 111  
  joint, 113–114  
  logistic, 111  
  lognormal, 111  
  marginal, 113–114  
  negative binomial, 111  
  normal, 105–108  
  Pareto, 112  
  *PHStat* support in, 110  
  Poisson, 100–102  
  triangular, 108–109  
  uniform, 104–105  
  Weibull, 111  
Probability interval, 137  
Probability mass function, 97  
Process average, sudden shift in, 281  
Process capability analysis  
  elements of, 288–289  
  illustration of, 289–290  
  role of, 288  
Process capability index, 290  
Process selection models, 453–454  
Process simulation, 403  
  sequence of activities, 409–410  
  with *SimQuick*, 410–427  
  single-server queue, 409–410  
Product and process outcomes, 30  
Product development decision model  
  decision model for, 309–311  
  Monte Carlo simulation and, 352–353  
Production/marketing allocation model, 469–472  
Program Evaluation and Review Technique (PERT), 354  
Project management  
  analytical critical path calculations, 356  
  *Bootstrap* tool used in, 353–358  
  *Crystal Ball* model, 355  
  model development, 313–315  
  project completion estimated in, 356–358  
  uncertain activity time data, 355  
Proportion (s)  
  confidence intervals for, 142–143, 150, 157–158  
  differences between, 150, 157–158  
  one-sample hypothesis test, 172–174  
  sample, 71  
  sampling distribution of, 154  
  two-sample hypothesis test for, 179–180  
Proportional relationships, 452  
Pull system supply chain, 424–426  
*p*-value, 205, 213, 216, 220, 259

## Q

- Qualitative and judgmental forecasting techniques, 238–240  
Delphi method, 239  
historical analogy, 239  
indicators and indexes, 239–240

## Qualitative events, 371

- Quality control  
  common causes of variation and, 273  
  role of statistics and data analysis in, 273  
  special causes of variation and, 273  
  statistical process control in, 274–280

## Quantitative events, 371

### Quartiles, 63

### Queue discipline, 405

### Queuing systems

- analytical models, 406–409  
  basic concepts, 403–404  
  customer characteristics, 404–405  
  operating characteristics, 406–407  
  performance measures, 406  
  process simulation model, 409–410  
  queue characteristics, 405  
  service characteristics, 405  
  single-server model, 407–408  
  system configurations, 405

## R

### Radar chart, 49

### RAND function, 127–128, 131, 327, 349

### Random number, 127

- in Monte Carlo simulation, 127  
  seed, 160

### Random sampling

- from common probability distributions, 129–130  
  from discrete probability distributions, 128–129  
  Excel-based, 158–159  
  *PHStat* and, 158–159  
  simple, 125

### Random variables, 94

- continuous, 94  
  discrete, 94  
  spreadsheet models with Monte Carlo simulation, 326–327  
  statistically independent, 113

### Random variate, 129, 325, 332, 339, 342, 351

### Range, 64

### Rank and Percentile tool, 62–63

### Ratio data, 35

### R-chart, 275–280

### Reduced cost, 448

### Regression analysis. *See also* Multiple linear regression; Simple linear regression

- adjusted  $R^2$ , 233  
  as analysis of variance, 231–233  
  with categorical independent variables, 220–225  
  with categorical variables with two levels, 223–225  
  confidence intervals, 233  
  definition, 198  
  models, development of, 214–220  
  with non-linear terms, 225–227  
  prediction intervals, 233  
  residual analysis and, 206–210  
  standard error of the estimate, 233

- Regression-based forecasting methods, 248–249  
with causal variables, 255–257
- Rejection region, 167
- Relative frequency distribution, 58, 61
- Renege, 405
- Reorder point (ROP) inventory system, 424
- Requirements, 452
- Residual analysis, in regression analysis, 206–210
- Residuals, 200
- Resources, in *SimQuick*, 418–421
- Return on investment (ROI), 370
- Return to risk, 376
- Revenue management decision tree, 384
- Risk  
beta, 203  
decision making and, 371–377  
decision tree and, 382–384  
of exceeding capacity, 513  
in financial investment analysis, 375  
investment, 202–203  
in portfolio risk analysis, 378–379  
premium, 391  
profile, 383  
return to, 376  
specific, 202  
systematic, 202  
variability and, 375–377
- Risk analysis, concept of, 325  
and optimization, 512–514
- Risk Solver Platform* add-in, 302
- Risk-averse decision makers, 391–392
- RMSE. *See* Root mean square error (RMSE)
- Root mean square error (RMSE), 244
- ROUND function, 326
- ROUNDUP function, 298
- R-square, 204, 215
- Run chart, 274
- ## S
- Sales, predictive model of, 29
- Sample, 36  
correlation coefficient, 85  
information, 384  
proportion, 71  
size, determination of, 155  
space, 90
- Sampling  
cluster, 126  
from common probability distributions, 129–130  
from a continuous process, 126  
convenience, 125  
from discrete probability distributions, 128–129  
Excel-based, 158–159  
judgment, 125  
Latin Hypercube, 332–334  
Monte Carlo, 332–334  
periodic, 126  
*PHStat* and, 158–159  
plan, 124  
probabilistic, 125  
random. *See* Random sampling  
simple random, 125  
statistical error, 127, 131–133  
stratified, 126  
systematic, 126
- Sampling distribution  
created by *Bootstrap* tool, 366  
of mean, 133–134
- nonrejection and rejection regions in, 167  
in *PHStat*, 159–160  
of proportion, 154
- Scale parameter, 104
- Scatter diagrams, 48
- Scenario Analysis* tool, 343
- Scoring model, 370
- Seasonal additive model, 252
- Seasonal multiplicative model, 252
- Seasonality  
*CB Predictor* support for, 259–260  
forecasting models with, 252, 255  
in regression models, 253–255
- Second-order autoregressive model, 250
- Second-order polynomial, 305  
fit, 307
- Sensitivity analysis, 299, 339  
on an optimization model, 368  
in decision tree, 384
- Sensitivity chart, 339–340, 363
- Service characteristics, in queuing systems, 405
- Shadow price, 449
- Shape parameter, 104
- Sharpe ratio, 376
- Shewhart, Walter, 274
- Shewhart charts, 274
- Significance F value, 206
- Significance level, 165–166
- Significance of regression, 205
- Simple bounds, 451–452
- Simple exponential smoothing, 246–247
- Simple linear regression, 198–203  
application to investment risk, 202–203  
confidence and prediction intervals for X-values, 206  
confidence intervals for regression coefficients, 206  
*Home Market Value*, 198–199, 202  
least-squares regression, 200–202  
regression statistics, 204–205  
scatter chart, 198, 202  
statistical hypothesis tests, 208–210  
testing hypotheses for regression coefficients, 205–206  
value of regression, explanation, 199
- Simple moving average method, 242, 244, 248, 265
- Simple random sampling, 125
- Simple regression, 198
- Simplex method algorithm, 473
- SimQuick*  
*Buffers* button, 412–413  
car wash simulation results, 414–415  
*Custom Schedules*, 426–427  
*Discrete Distributions*, 426–427  
*Entrances* button, 412  
*Exit* element, 412, 424  
five-element structures in, 412  
flow process maps of, 417  
getting started, 411  
grocery store checkout model with resources of, 418–421  
manufacturing inspection model with decision points of, 421–424  
*Other Features* button, 412, 426–427  
pull system supply chain with exit schedules of, 424–426  
queues in series with blocking and, 417–418  
resources in, 418–421  
standard deviation in, 411, 415, 422, 424  
*View Model* button, 414  
*Work Stations* worksheet, 414
- SimQuick-v26.xls*, 43
- Simulation modeling, continuous, 427–430
- Simulation statistics, 416
- Single alternative, decision making with, 369
- Single-factor analysis of variance, 195
- Site location model, 488–491
- Six Sigma, 36, 66
- Skewness, 67, 84  
characteristics of, 67–68  
coefficient of, 67, 84
- Sklenga Skis (SSC) model, 437–441
- SLOPE function, 201–202
- Smoothing constant, 246
- Solver* add-in, 302–304, 460  
*Add Constraint* dialog, 488  
decision variables, 440  
difficulties with, 451  
interpretation of reports, 446–450  
lower and upper bounds in, 465–467  
mathematical algorithm of, 473  
model for plant location, 497  
nonlinear optimization model, 504–505  
for nonsmooth optimization, 506–512  
outcomes and solution messages, 446  
reports, 451  
risk analysis and optimization, 512–514  
solution for 467-room capacity, 514  
*Standard Evolutionary* algorithm, 507–508  
working of, 473
- S&P 524 index, 203
- Special causes of variation and quality control, 273–274
- Specific risk, 202
- Spider charts, 353–354
- Spreadsheet, 296  
engineering, 308–309  
models, for optimization problems, 440–442
- Squares of the errors, 200
- Stack Data, 44
- Stacked column chart, 45
- Standard deviation, 65, 83  
in binomial distribution, 285  
control charts and, 274–276  
in *Crystal Ball*, 329, 334, 337, 362, 428, 521  
in financial investment analysis, 375–376, 378, 456  
Monte Carlo simulation and, 326–329  
in *OptQuest*, 521–523  
in *PHStat*, 287, 379–380  
in portfolio risk analysis, 378–379  
process capability and, 290  
risk and, 375–376  
in *SimQuick*, 411, 415, 422, 424  
SPC and, 280  
using *Bootstrap* tool, 343
- Standard error  
of estimate, 205  
of the mean, 133
- Standard Evolutionary* algorithm, 507–508
- Standard normal distribution, 105
- Standard residuals, 208
- Standardized normal values, 119
- Standardized z-values, 119
- State variables, 427
- Stationary arrival rate, 404
- Stationary time-series data, 146
- Statistical distributions  
in *SimQuick*, 411
- Statistical inference, 37

- S**  
Statistical measures  
for grouped data, 84  
visual display of, 73–74  
Statistical process control (SPC), 274  
Statistical quality control. *See* Quality control  
Statistical sampling  
errors in sampling, 127  
experiment in finance, 130  
population frame, 124  
sample design, 124–125  
sampling methods, 125–126  
sampling plan, 124  
Statistical thinking, 35–37  
Statistical time-series models, forecasting with, 242–249  
error metrics, 244–246  
exponential smoothing models, 246–247  
forecast accuracy, 244–246  
moving average, 242–244  
weighted moving averages, 244  
Statistics, 29  
role in quality control, 273  
Steady state, 408  
Stem-and-leaf displays, 73  
Stepwise regression, 217  
Straight line, equation of, 298  
Stratified sampling, 126  
Strengthening Global Effectiveness (SGE), 494  
Studentized range  $Q$ , 543–544  
Subjective sampling, 125  
Subscripted variables, 453  
Sum of squares of observed errors, 201  
SUMPRODUCT function, 441, 459, 490  
Supply chain facility location model, 494–495  
*Survey of Current Business*, 240  
System dynamics, 427, 429  
Systematic (periodic) sampling, 126  
Systematic risk, 202
- T**  
 $T$  Stat, 205  
 $T$  values, 536–538  
 $T$ -distributions, 169  
Test for equality of variances, 192, 194–195  
Test statistic, 205  
Textile linear programming model, 453–454  
Theil's U statistic, 257, 259  
Third-order polynomial, 305–306  
Third-order polynomial fit, 306–307
- Time-series model, 238  
Time-series regression, 198  
*Tornado Chart* tool, 342, 365–366  
Transient period, 408  
*TreePlan* add-in, 43, 381  
TREND function, 201–202  
*Trendline* tool, 306  
Triangular probability distribution, 108–109  
 $T$ -statistic, 216  
 $T$ -test statistic, 171  
Two-asset portfolio, 378  
*Two-dimensional Simulation* tool, 343  
Two-sample hypothesis test, 164–165  
confidence intervals and, 180–181  
for equality of variance, 181–182  
for mean, 177–178  
for means with paired samples, 179  
for proportions, 179–180  
using Excel, 179  
Two-tailed tests of hypothesis, 167, 171  
Two-tailed  $t$ -test, 172  
Two-way data tables, 300  
Type I error, 166, 175  
Type II error, 166, 175–177
- U**  
Unbiased estimators, 136  
Unboundedness, 446  
Uncertainty  
analysis using *Crystal Ball*, 339, 354–358  
Becker Consulting project management model, estimation in, 354–358  
in decision models, 328–332  
*Two-dimensional Simulation* tool and, 343  
Uncontrollable variables, 296  
Uniform distribution, 104–105, 119  
Unimodal distributions, 67  
Unique optimal solution, 446  
Univariate variables, 33  
Upper specification limit (USL), 275, 290  
Utility theory, 389–394
- V**  
Validity, 315  
Value of information, 384–389  
Variability, risk and, 375–377  
Variables, 33  
decision, 296  
uncontrollable, 296  
Variables data, 274  
Variance, 65, 83, 250, 288, 356, 376, 378. *See also* Analysis of variance (ANOVA)
- of Bernoulli distribution, 99  
of binomial distribution, 100  
confidence intervals for, 143–145, 158, 161  
for a continuous random variable, 104  
of a discrete random variable, 98  
in Excel, 57  
of exponential distribution, 109  
of gamma distribution, 111  
of normal distribution, 105  
one sample test for, 174–175  
population, 135–136, 142, 191–194  
of a portfolio, 503  
of a random variable, 118–119  
sample, 135–136, 141–142  
*Sensitivity Chart*, 339–340  
test for equality of, 181–182, 194–195  
of a triangular random variable, 108  
two sample test for, 193–194  
of a uniform random variable, 104  
using *PHStat* tools, 138  
Variance inflation factor (VIF), 214  
Variation  
common causes of, 273  
decision making under, 375–377  
in design specifications, 288–289  
distribution of output, 274  
lower specification limit, 275  
nominal specification, 275  
permissible dimension, 275–276  
processes in and out of control and, 35–36, 274  
special causes of, 273  
statistical measure of coefficient of, 376  
upper specification limit, 275  
Verification, 308
- W**  
Weibull distributions, 111  
Weighted moving averages, 244  
What-if analysis, 409  
data tables, 299–301  
*Goal Seek* tool, 302  
*Scenario Manager* tool, 301  
Work stations, in *SimQuick*, 411  
Work-focused outcomes, 30
- X**  
 $x$  -chart, 275–280
- Z**  
 $Z$ -values, 119–120, 177

Using the <i>PHStat Stack Data</i> and <i>Unstack Data</i> Tools	p. 52
One- and Two-Way Tables and Charts	p. 87
Normal Probability Tools	p. 120
Generating Probabilities in <i>PHStat</i>	p. 121
Confidence Intervals for the Mean	p. 160
Confidence Intervals for Proportions	p. 160
Confidence Intervals for the Population Variance	p. 161
Determining Sample Size	p. 161
One-Sample Test for the Mean, Sigma Unknown	p. 193
One-Sample Test for Proportions	p. 193
Using Two-Sample <i>t</i> -Test Tools	p. 193
Testing for Equality of Variances	p. 194
Chi-Square Test for Independence	p. 195
Using Regression Tools	p. 233
Stepwise Regression	p. 235
Best-Subsets Regression	p. 236
Creating $\bar{x}$ - and <i>R</i> -Charts	p. 291
Creating <i>p</i> -Charts	p. 292
Using the <i>Expected Monetary Value</i> Tool	p. 399

*Excel Notes*

---

Creating Charts in Excel 2010	p. 53
Creating a Frequency Distribution and Histogram	p. 85
Using the Descriptive Statistics Tool	p. 85
Using the Correlation Tool	p. 86
Creating Box Plots	p. 87
Creating PivotTables	p. 87
Excel-Based Random Sampling Tools	p. 158
Using the <i>VLOOKUP</i> Function	p. 159
Sampling from Probability Distributions	p. 159
Single-Factor Analysis of Variance	p. 195
Using the Trendline Option	p. 233
Using Regression Tools	p. 233
Using the Correlation Tool	p. 235
Forecasting with Moving Averages	p. 267
Forecasting with Exponential Smoothing	p. 267
Using <i>CB Predictor</i>	p. 268
Creating Data Tables	p. 322
<i>Data Table</i> Dialog	p. 322
Using the <i>Scenario Manager</i>	p. 322
Using <i>Goal Seek</i>	p. 323
Net Present Value and the NPV Function	p. 323
Using the IRR Function	p. 399

*Crystal Ball Notes*

---

Customizing <i>Define Assumption</i>	p. 362
Sensitivity Charts	p. 363
Distribution Fitting with Crystal Ball	p. 363
<i>Correlation Matrix</i> Tool	p. 365
Tornado Charts	p. 365
<i>Bootstrap</i> Tool	p. 366

*TreePlan Note*

---

Constructing Decision Trees in Excel	p. 400
--------------------------------------	--------

# **STUDENTS**

With the purchase of a new copy of this textbook, you immediately have access to the subscription content on the **Statistics, Data Analysis, and Decision Modeling, Fifth Edition** Companion Website.

**Subscription content provides you with:**

- Risk Solver Platform for Education
- Oracle Crystal Ball 140-day Trial
- SimQuick

## **Access Code Here**

**Use a coin to scratch off the coating and reveal your student access code.**

**Do not use a knife or other sharp object as it may damage the code.**

To access the **Statistics, Data Analysis, and Decision Modeling, Fifth Edition** subscription content for the first time, you will need to register online. This process takes just a couple of minutes and only needs to be completed once.

1. Go to <http://www.pearsoninternationaleditions.com/evans>
2. Click on **Companion Website**.
3. Click on **Subscription Content**.
4. On the registration page, enter your student access code. Do not type the dashes.
5. Follow the on-screen instructions. If you need help at any time during the online registration process, simply click the **Help?** icon.
6. Once your personal Login Name and Password are confirmed, you will be given access to the directions for downloading the subscription content.

To log in after you have registered:

You only need to register for this Companion Website once. After that, you can log in any time at <http://www.pearsoninternationaleditions.com/evans> by providing your Login Name and Password when prompted.

### **IMPORTANT:**

The access code can only be used once. If this access code has already been revealed, it may no longer be valid.