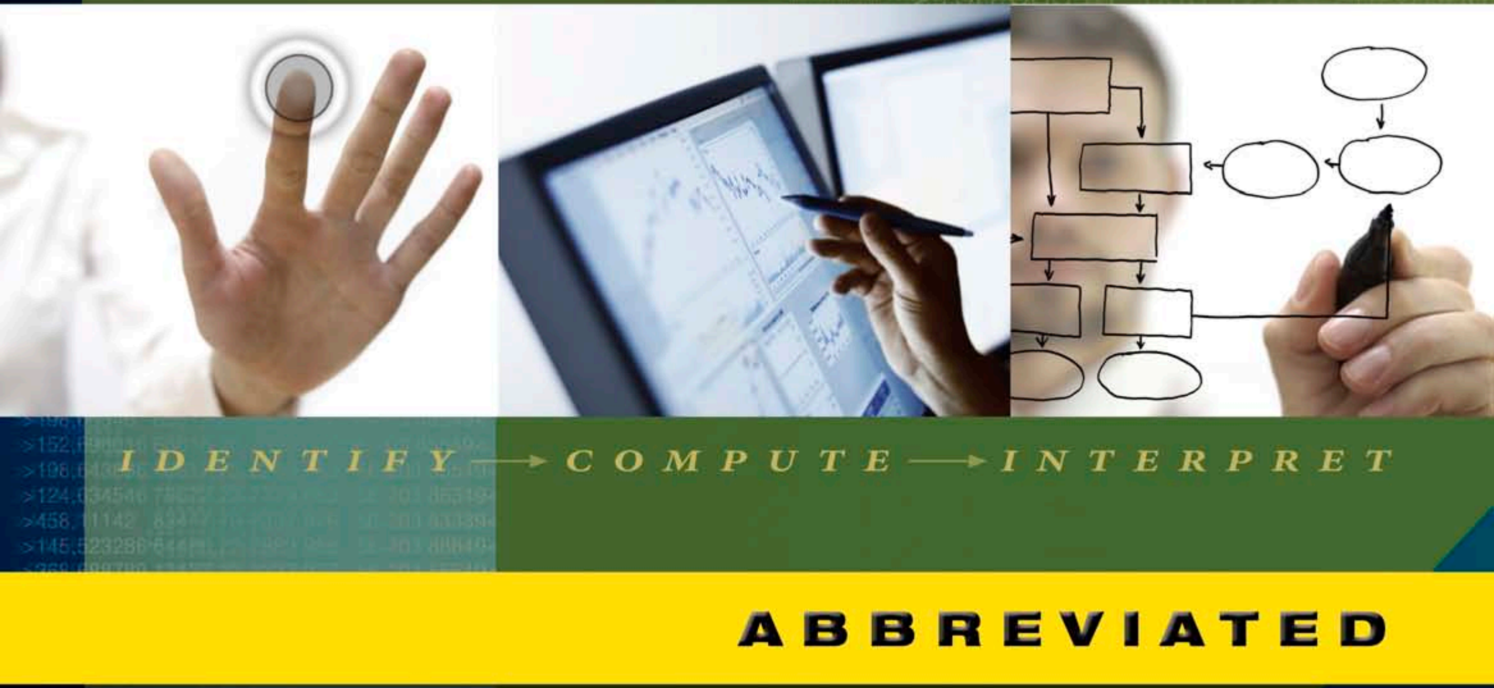
**GERALD KELLER**

# Statistics

## FOR MANAGEMENT AND ECONOMICS



*IDENTIFY → COMPUTE → INTERPRET*

**ABBREVIATED**

9e

# A GUIDE TO STATISTICAL TECHNIQUES

*Problem Objectives*

| | | Describe a Population | Compare Two Populations | Compare Two or More Populations | Analyze Relationship between Two Variables | Analyze Relationship among Two or More Variables |
|---|---|---|---|---|---|---|
| DATA TYPES | Interval | Histogram **Section 3.1**<br>Ogive **Section 3.1**<br>Stem-and-leaf **Section 3.1**<br>Line chart **Section 3.2**<br>Mean, median, and mode **Section 4.1**<br>Range, variance, and standard deviation **Section 4.2**<br>Percentiles and quartiles **Section 4.3**<br>Box plot **Section 4.3**<br>*t*-test and estimator of a mean **Section 12.1**<br>Chi-squared test and estimator of a variance **Section 12.2** | Equal-variances *t*-test and estimator of the difference between two means: independent samples **Section 13.1**<br>Unequal-variances *t*-test and estimator of the difference between two means: independent samples **Section 13.1**<br>*t*-test and estimator of mean difference **Section 13.3**<br>*F*-test and estimator of ratio of two variances **Section 13.4** | One-way analysis of variance **Section 14.1**<br>LSD multiple comparison method **Section 14.2**<br>Tukey's multiple comparison method **Section 14.2**<br>Two-way analysis of variance **Section 14.4**<br>Two-factor analysis of variance **Section 14.5** | Scatter diagram **Section 3.3**<br>Covariance **Section 4.4**<br>Coefficient of correlation **Section 4.4**<br>Coefficient of determination **Section 4.4**<br>Least squares line **Section 4.4**<br>Simple linear regression and correlation **Chapter 16** | Multiple regression **Chapter 17** |
| | Nominal | Frequency distribution **Section 2.2**<br>Bar chart **Section 2.2**<br>Pie chart **Section 2.2**<br>*z*-test and estimator of a proportion **Section 12.3**<br>Chi-squared goodness-of-fit test **Section 15.1** | *z*-test and estimator of the difference between two proportions **Section 13.5**<br>Chi-squared test of a contingency table **Section 15.2** | Chi-squared test of a contingency table **Section 15.2** | Chi-squared test of a contingency table **Section 15.2** | Not covered |
| | Ordinal | Median **Section 4.1**<br>Percentiles and quartiles **Section 4.3**<br>Box plot **Section 4.3** | | | | Not covered |

# AMERICAN NATIONAL ELECTION SURVEY AND GENERAL SOCIAL SURVEY EXERCISES

| Chapter | ANES | Page | GSS | Page |
|---|---|---|---|---|
| 2 | | | 2.34–2.37 | 31 |
| 3 | 3.62–3.67 | 82 | 3.25–3.28 | 64 |
| | 3.68–3.71 | 82 | | |
| 4 | 4.37–4.38 | 117 | 4.39 | 117 |
| | 4.58–4.60 | 126 | 4.61–4.62 | 126 |
| | 4.86 | 144 | 4.84–4.85 | 144 |
| 12 | 12.51–12.53 | 413 | 12.46–12.50 | 413 |
| | 12.116–12.123 | 435 | 12.103–12.115 | 434 |
| 13 | 13.42–13.44 | 472 | 13.38–13.41 | 472 |
| | 13.73 | 488 | 13.70–13.72 | 488 |
| | 13.123–13.125 | 512 | 13.113–13.122 | 511 |
| | A13.27–A13.30 | 524 | A13.18–A13.126 | 523 |
| 14 | 14.27–14.32 | 543 | 14.21–14.26 | 542 |
| | 14.47–14.50 | 553 | 14.43–14.46 | 552 |
| | 14.66–14.67 | 563 | | |
| | A14.23–A14.25 | 594 | A14.19–A14.22 | 594 |
| 15 | 15.17 | 604 | 15.18–15.21 | 604 |
| | 15.43–15.46 | 615 | 15.39–15.42 | 614 |
| | A15.25–A15.28 | 631 | A15.17–A15.24 | 630 |
| 16 | 16.45–16.49 | 665 | 16.50–16.53 | 666 |
| | 16.73–16.76 | 671 | 16.77–16.80 | 671 |
| | A16.27–A16.28 | 689 | A16.17–A16.26 | 689 |
| 17 | 17.21–17.22 | 713 | 17.16–17.20 | 712 |
| | A17.27–A17.28 | 734 | A17.17–A17.26 | 733 |

# APPLICATION SECTIONS

**Section 4.5** (Optional) Application in Professional Sports Management: Determinants of the Number of Wins in a Baseball Season (illustrating an application of the least squares method and correlation) 144

**Section 4.6** (Optional) Application in Finance: Market Model (illustrating using a least squares lines and coefficient of determination to estimate a stock's market-related risk and its firm-specific risk) 147

**Section 7.3** (Optional) Application in Finance: Portfolio Diversification and Asset Allocation (illustrating the laws of expected value and variance and covariance) 237

**Section 12.4** (Optional) Application in Marketing: Market Segmentation (using inference about a proportion to estimate the size of a market segment) 435

**Section 14.6** (Optional) Application in Operations Management: Finding and Reducing Variation (using analysis of variance to actively experiment to find sources of variation) 578

# APPLICATION SUBSECTION

**Section 6.4** (Optional) Application in Medicine and Medical Insurance: Medical Screening (using Bayes's Law to calculate probabilities after a screening test) 199

*This page intentionally left blank*

# Statistics

## FOR MANAGEMENT AND ECONOMICS ABBREVIATED

9e

# Statistics

## FOR MANAGEMENT AND ECONOMICS ABBREVIATED

9e

GERALD KELLER

*Wilfred Laurier University*

SOUTH-WESTERN
CENGAGE Learning™

Australia • Brazil • Japan • Korea • Mexico • Singapore • Spain • United Kingdom • United States

SOUTH-WESTERN
CENGAGE Learning™

**Statistics for Management and Economics Abbreviated, Ninth Edition**

**Gerald Keller**

VP/Editorial Director:
Jack W. Calhoun

Publisher:
Joe Sabatino

Senior Acquisitions Editor:
Charles McCormick, Jr.

Developmental Editor:
Elizabeth Lowry

Editorial Assistant:
Nora Heink

Senior Marketing Communications Manager:
Libby Shipp

Marketing Manager:
Adam Marsh

Content Project Manager:
Jacquelyn K Featherly

Media Editor:
Chris Valentine

Manufacturing Buyer:
Miranda Klapper

Production House/Compositor:
MPS Limited, a Macmillan Company

Senior Rights Specialist:
John Hill

Senior Art Director:
Stacy Jenkins Shirley

Internal Designer:
KeDesign/cmiller design

Cover Designer:
Cmiller design

Cover Images:
© iStock Photo

For product information and technology assistance, contact us at
**Cengage Learning Customer & Sales Support, 1-800-354-9706**

For permission to use material from this text or product,
submit all requests online at **www.cengage.com/permissions**
Further permissions questions can be emailed to
**permissionrequest@cengage.com**
ExamView® and ExamView Pro® are registered trademarks
of FSCreations, Inc. Windows is a registered trademark of the
Microsoft Corporation used herein under license. Macintosh and
Power Macintosh are registered trademarks of Apple Computer, Inc.
used herein under license.

Library of Congress Control Number: 2010928228

Abbreviated Student Edition ISBN 13: 978-1-111-52708-2

Abbreviated Student Edition ISBN 10: 1-111-52708-3

Package Abbreviated Student Edition ISBN 13: 978-1-111-52732-7

Package Abbreviated Student Edition ISBN 10: 1-111-52732-6

**South-Western Cengage Learning**
5191 Natorp Boulevard
Mason, OH 45040
USA

Cengage Learning products are represented in Canada by
Nelson Education, Ltd.

For your course and learning solutions, visit **www.cengage.com**

Purchase any of our products at your local college store or at our
preferred online store **www.cengagebrain.com**

Printed in the United States of America
1 2 3 4 5 6 7 14 13 12 11 10

# BRIEF CONTENTS

# CONTENTS

# PREFACE

Businesses are increasingly using statistical techniques to convert data into information. For students preparing for the business world, it is not enough merely to focus on mastering a diverse set of statistical techniques and calculations. A course and its attendant textbook must provide a complete picture of statistical concepts and their applications to the real world. *Statistics for Management and Economics* is designed to demonstrate that statistics methods are vital tools for today's managers and economists.

Fulfilling this objective requires the several features that I have built into this book. First, I have included data-driven examples, exercises, and cases that demonstrate statistical applications that are and can be used by marketing managers, financial analysts, accountants, economists, operations managers, and others. Many are accompanied by large and either genuine or realistic data sets. Second, I reinforce the applied nature of the discipline by teaching students how to choose the correct statistical technique. Third, I teach students the concepts that are essential to interpreting the statistical results.

## Why I Wrote This Book

Business is complex and requires effective management to succeed. Managing complexity requires many skills. There are more competitors, more places to sell products, and more places to locate workers. As a consequence, effective decision making is more crucial than ever before. On the other hand, managers have more access to larger and more detailed data that are potential sources of information. However, to achieve this potential requires that managers know how to convert data into information. This knowledge extends well beyond the arithmetic of calculating statistics. Unfortunately, this is what most textbooks offer—a series of unconnected techniques illustrated mostly with manual calculations. This continues a pattern that goes back many years. What is required is a complete approach to applying statistical techniques.

When I started teaching statistics in 1971, books demonstrated how to calculate statistics and, in some cases, how various formulas were derived. One reason for doing so was the belief that by doing calculations by hand, students would be able to understand the techniques and concepts. When the first edition of this book was published in 1988, an important goal was to teach students to identify the correct technique. Through the next eight editions, I refined my approach to emphasize interpretation and decision making equally. I now divide the solution of statistical problems into three stages and include them in every appropriate example: (1) *identify* the technique, (2) *compute* the statistics, and (3) *interpret* the results. The compute stage can be completed in any or all of three ways: manually (with the aid of a calculator), using Excel 2010, and using Minitab. For those courses that wish to use the computer extensively, manual calculations can be played down or omitted completely. Conversely, those that wish to emphasize manual calculations may easily do so, and the computer solutions can be selectively introduced or skipped entirely. This approach is designed to provide maximum flexibility, and it leaves to the instructor the decision of if and when to introduce the computer.

I believe that my approach offers several advantages.

- An emphasis on identification and interpretation provides students with practical skills they can apply to real problems they will face regardless of whether a course uses manual or computer calculations.
- Students learn that statistics is a method of converting data into information. With 878 data files and corresponding problems that ask students to interpret statistical results, students are given ample opportunities to practice data analysis and decision making.
- The optional use of the computer allows for larger and more realistic exercises and examples.

Placing calculations in the context of a larger problem allows instructors to focus on more important aspects of the decision problem. For example, more attention needs to be devoted to interpreting statistical results. Proper interpretation of statistical results requires an understanding of the probability and statistical concepts that underlie the techniques and an understanding of the context of the problems. An essential aspect of my approach is teaching students the concepts. I do so in two ways.

1. Nineteen Java applets allow students to see for themselves how statistical techniques are derived without going through the sometimes complicated mathematical derivations.
2. Instructions are provided about how to create Excel worksheets that allow students to perform "what-if" analyses. Students can easily see the effect of changing the components of a statistical technique, such as the effect of increasing the sample size.

Efforts to teach statistics as a valuable and necessary tool in business and economics are made more difficult by the positioning of the statistics course in most curricula. The required statistics course in most undergraduate programs appears in the first or second year. In many graduate programs, the statistics course is offered in the first semester of a three-semester program and the first year of a two-year program. Accounting, economics, finance, human resource management, marketing, and operations management are usually taught after the statistics course. Consequently, most students will not be able to understand the general context of the statistical application. This deficiency is addressed in this book by "Applications in . . ." sections, subsections, and boxes. Illustrations of statistical applications in business that students are unfamiliar with are preceded by an explanation of the background material.

- For example, to illustrate graphical techniques, we use an example that compares the histograms of the returns on two different investments. To explain what financial analysts look for in the histograms requires an understanding that risk is measured by the amount of variation in the returns. The example is preceded by an "Applications in Finance" box that discusses how return on investment is computed and used.
- Later when I present the normal distribution, I feature another "Applications in Finance" box to show why the standard deviation of the returns measures the risk of that investment.
- Thirty-six application boxes are scattered throughout the book.
- I've added Do-It-Yourself Excel exercises will teach students to compute spreadsheets on their own.

Some applications are so large that I devote an entire section or subsection to the topic. For example, in the chapter that introduces the confidence interval estimator of a proportion, I also present market segmentation. In that section, I show how the confidence interval estimate of a population proportion can yield estimates of the sizes of market segments. In other chapters, I illustrate various statistical techniques by showing how marketing managers can apply these techniques to determine the differences that exist between market segments. There are six such sections and one subsection in this book. The "Applications in . . ." segments provide great motivation to the student who asks, "How will I ever use this technique?"

## New in This Edition

Six large real data sets are the sources of 150 new exercises. Students will have the opportunity to convert real data into information. Instructors can use the data sets for hundreds of additional examples and exercises.

Many of the examples, exercises, and cases using real data in the eighth edition have been updated. These include the data on wins, payrolls, and attendance in baseball, basketball, football, and hockey; returns on stocks listed on the New York Stock Exchange, NASDAQ, and Toronto Stock Exchange; and global warming.

Chapter 2 in the eighth edition, which presented graphical techniques, has been split into two chapters–2 and 3. Chapter 2 describes graphical techniques for nominal data, and Chapter 3 presents graphical techniques for interval data. Some of the material in the eighth edition Chapter 3 has been incorporated into the new Chapter 3.

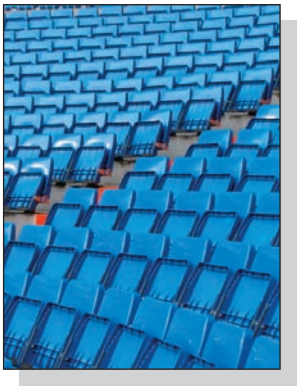
To make room for the new additional exercises we have removed Section 12.5, Applications in Accounting: Auditing.

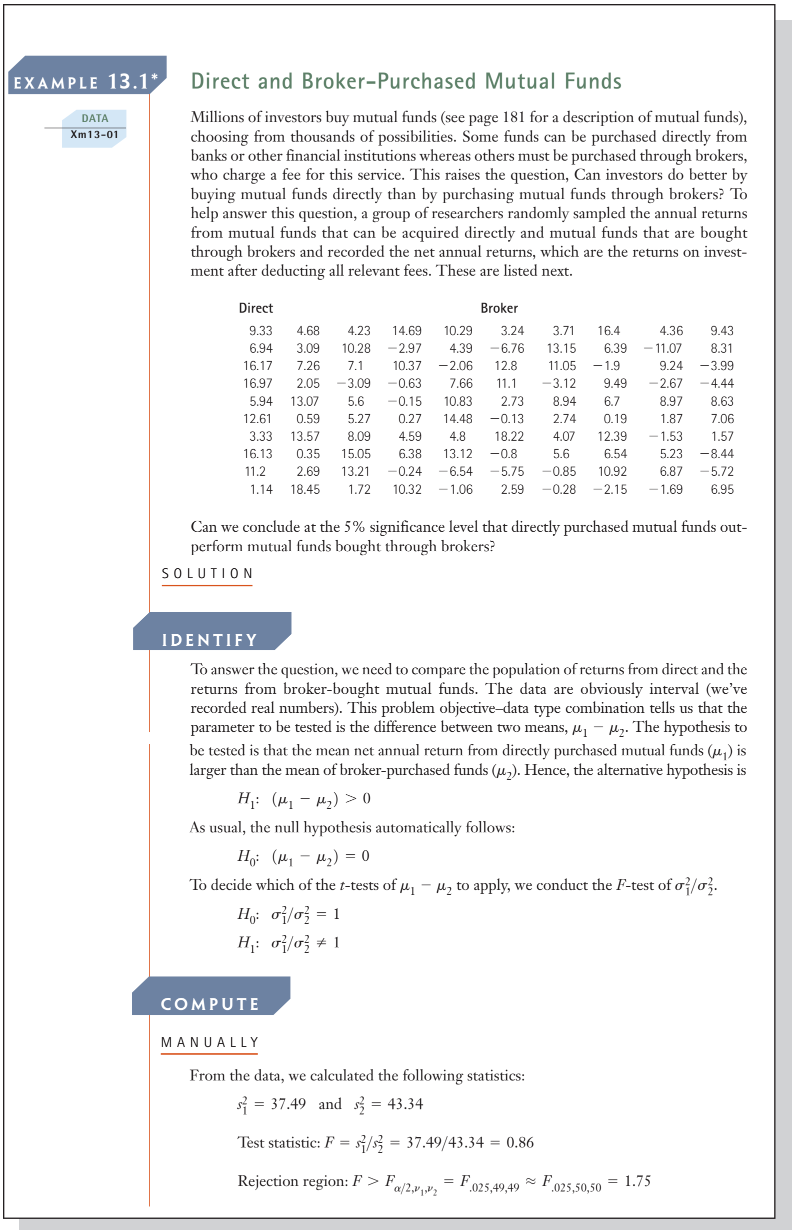
I've created many new examples and exercises. Here are the numbers for the Abbreviated ninth edition: 116 solved examples, 1727 exercises, 26 cases, 690 data sets, 35 appendixes containing 37 solved examples, 98 exercises, and 25 data sets for a grand total of 153 solved examples, 1825 exercises, 26 cases, and 715 data sets.

# Guided Book Tour

## Data Driven: The Big Picture

Solving statistical problems begins with a problem and data. The ability to select the right method by problem objective and data type is **a valuable tool for business**. Because business decisions are driven by data, students will leave this course equipped with the tools they need to make effective, informed decisions in all areas of the business world.





EXAMPLE 13.1*

DATA
Xm13-01

### Direct and Broker-Purchased Mutual Funds

Millions of investors buy mutual funds (see page 181 for a description of mutual funds), choosing from thousands of possibilities. Some funds can be purchased directly from banks or other financial institutions whereas others must be purchased through brokers, who charge a fee for this service. This raises the question, Can investors do better by buying mutual funds directly than by purchasing mutual funds through brokers? To help answer this question, a group of researchers randomly sampled the annual returns from mutual funds that can be acquired directly and mutual funds that are bought through brokers and recorded the net annual returns, which are the returns on investment after deducting all relevant fees. These are listed next.

| Direct | | | | | Broker | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 9.33 | 4.68 | 4.23 | 14.69 | 10.29 | 3.24 | 3.71 | 16.4 | 4.36 | 9.43 |
| 6.94 | 3.09 | 10.28 | −2.97 | 4.39 | −6.76 | 13.15 | 6.39 | −11.07 | 8.31 |
| 16.17 | 7.26 | 7.1 | 10.37 | −2.06 | 12.8 | 11.05 | −1.9 | 9.24 | −3.99 |
| 16.97 | 2.05 | −3.09 | −0.63 | 7.66 | 11.1 | −3.12 | 9.49 | −2.67 | −4.44 |
| 5.94 | 13.07 | 5.6 | −0.15 | 10.83 | 2.73 | 8.94 | 6.7 | 8.97 | 8.63 |
| 12.61 | 0.59 | 5.27 | 0.27 | 14.48 | −0.13 | 2.74 | 0.19 | 1.87 | 7.06 |
| 3.33 | 13.57 | 8.09 | 4.59 | 4.8 | 18.22 | 4.07 | 12.39 | −1.53 | 1.57 |
| 16.13 | 0.35 | 15.05 | 6.38 | 13.12 | −0.8 | 5.6 | 6.54 | 5.23 | −8.44 |
| 11.2 | 2.69 | 13.21 | −0.24 | −6.54 | −5.75 | −0.85 | 10.92 | 6.87 | −5.72 |
| 1.14 | 18.45 | 1.72 | 10.32 | −1.06 | 2.59 | −0.28 | −2.15 | −1.69 | 6.95 |

Can we conclude at the 5% significance level that directly purchased mutual funds outperform mutual funds bought through brokers?

SOLUTION

IDENTIFY

To answer the question, we need to compare the population of returns from direct and the returns from broker-bought mutual funds. The data are obviously interval (we've recorded real numbers). This problem objective–data type combination tells us that the parameter to be tested is the difference between two means, $\mu_1 - \mu_2$. The hypothesis to be tested is that the mean net annual return from directly purchased mutual funds ($\mu_1$) is larger than the mean of broker-purchased funds ($\mu_2$). Hence, the alternative hypothesis is

$$H_1: \ (\mu_1 - \mu_2) > 0$$

As usual, the null hypothesis automatically follows:

$$H_0: \ (\mu_1 - \mu_2) = 0$$

To decide which of the $t$-tests of $\mu_1 - \mu_2$ to apply, we conduct the $F$-test of $\sigma_1^2/\sigma_2^2$.

$$H_0: \ \sigma_1^2/\sigma_2^2 = 1$$
$$H_1: \ \sigma_1^2/\sigma_2^2 \neq 1$$

COMPUTE

MANUALLY

From the data, we calculated the following statistics:

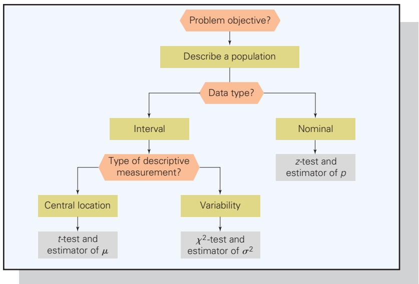
$$s_1^2 = 37.49 \quad \text{and} \quad s_2^2 = 43.34$$

$$\text{Test statistic: } F = s_1^2/s_2^2 = 37.49/43.34 = 0.86$$

$$\text{Rejection region: } F > F_{\alpha/2,\nu_1,\nu_2} = F_{.025,49,49} \approx F_{.025,50,50} = 1.75$$

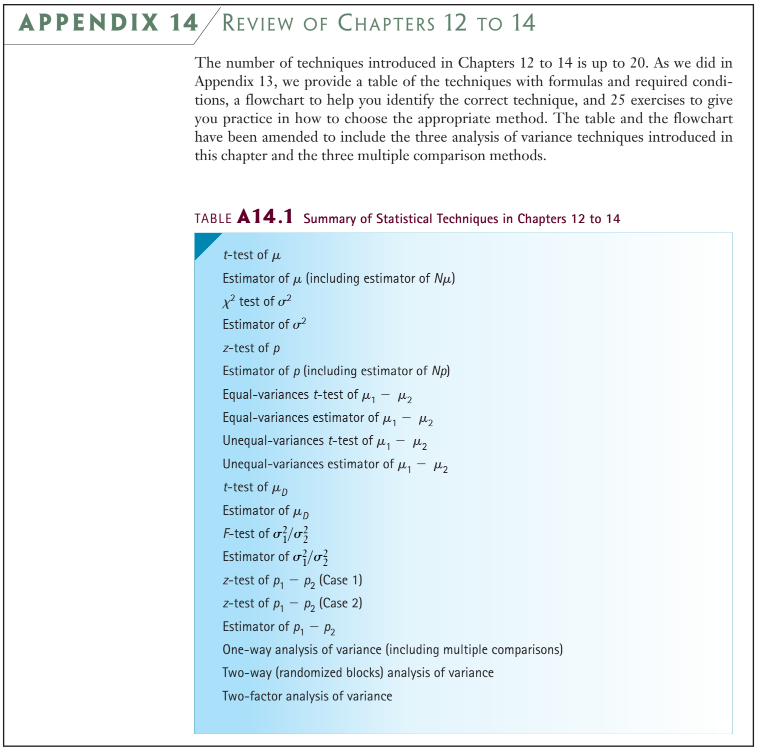
## *Identify* the Correct Technique

**Examples** introduce the first crucial step in this three-step (*identify–compute–interpret*) approach. Every example's solution begins by examining the data type and problem objective and then identifying the right technique to solve the problem.


Problem objective?
Describe a population
Data type?
Interval
Nominal
Type of descriptive measurement?
z-test and estimator of p
Central location
Variability
t-test and estimator of μ
χ²-test and estimator of σ²

**Appendixes 13, 14, 15, 16,** and **17** reinforce this problem-solving approach and allow students to hone their skills.

**Flowcharts**, found within the appendixes, help students develop the logical process for choosing the correct technique, reinforce the learning process, and provide easy review material for students.

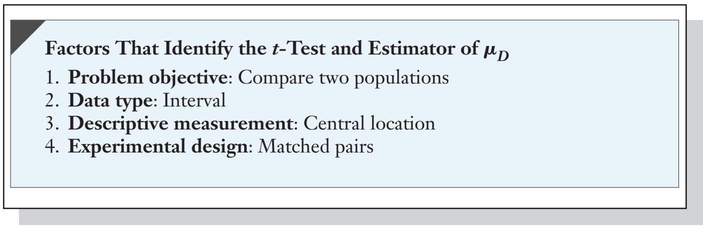


## APPENDIX 14 / Review of Chapters 12 to 14

The number of techniques introduced in Chapters 12 to 14 is up to 20. As we did in Appendix 13, we provide a table of the techniques with formulas and required conditions, a flowchart to help you identify the correct technique, and 25 exercises to give you practice in how to choose the appropriate method. The table and the flowchart have been amended to include the three analysis of variance techniques introduced in this chapter and the three multiple comparison methods.

TABLE **A14.1** Summary of Statistical Techniques in Chapters 12 to 14

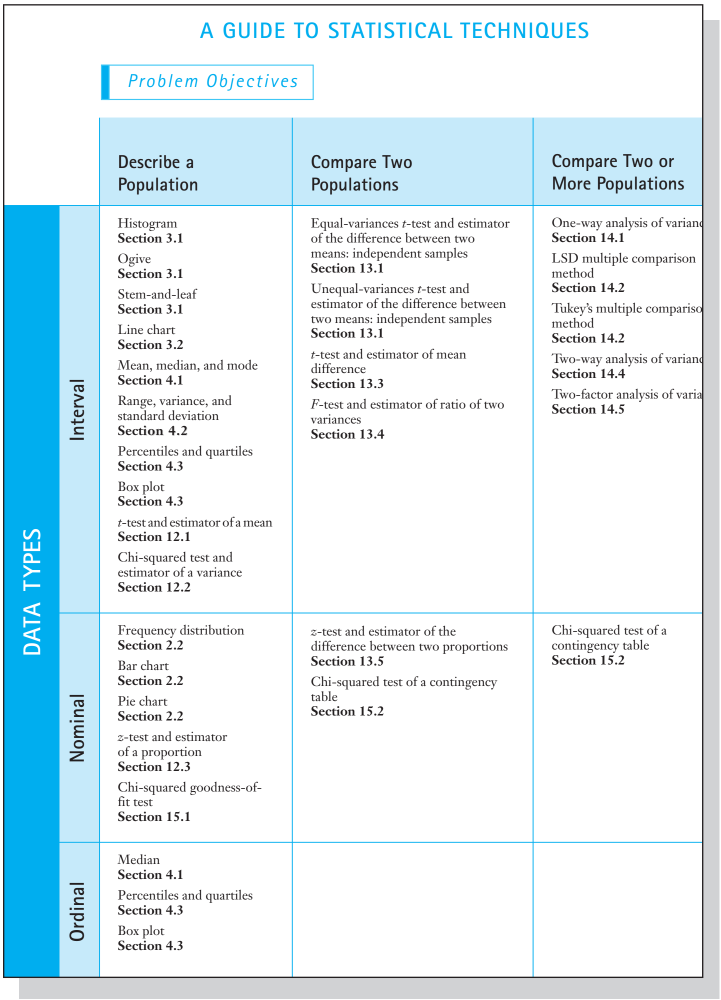
$t$-test of $\mu$
Estimator of $\mu$ (including estimator of $N\mu$)
$\chi^2$ test of $\sigma^2$
Estimator of $\sigma^2$
$z$-test of $p$
Estimator of $p$ (including estimator of $Np$)
Equal-variances $t$-test of $\mu_1 - \mu_2$
Equal-variances estimator of $\mu_1 - \mu_2$
Unequal-variances $t$-test of $\mu_1 - \mu_2$
Unequal-variances estimator of $\mu_1 - \mu_2$
$t$-test of $\mu_D$
Estimator of $\mu_D$
$F$-test of $\sigma_1^2/\sigma_2^2$
Estimator of $\sigma_1^2/\sigma_2^2$
$z$-test of $p_1 - p_2$ (Case 1)
$z$-test of $p_1 - p_2$ (Case 2)
Estimator of $p_1 - p_2$
One-way analysis of variance (including multiple comparisons)
Two-way (randomized blocks) analysis of variance
Two-factor analysis of variance



**Factors That Identify the *t*-Test and Estimator of $\mu_D$**

1. **Problem objective**: Compare two populations
2. **Data type**: Interval
3. **Descriptive measurement**: Central location
4. **Experimental design**: Matched pairs

**Factors That Identify . . .** boxes are found in each chapter after a technique or concept has been introduced. These boxes allow students to see a technique's essential requirements and give them a way to easily review their understanding. These essential requirements are revisited in the review chapters, where they are coupled with other concepts illustrated in flowcharts.



A GUIDE TO STATISTICAL TECHNIQUES

*Problem Objectives*

| DATA TYPES | | Describe a Population | Compare Two Populations | Compare Two or More Populations |
|---|---|---|---|---|
| | Interval | Histogram<br>**Section 3.1**<br>Ogive<br>**Section 3.1**<br>Stem-and-leaf<br>**Section 3.1**<br>Line chart<br>**Section 3.2**<br>Mean, median, and mode<br>**Section 4.1**<br>Range, variance, and standard deviation<br>**Section 4.2**<br>Percentiles and quartiles<br>**Section 4.3**<br>Box plot<br>**Section 4.3**<br>*t*-test and estimator of a mean<br>**Section 12.1**<br>Chi-squared test and estimator of a variance<br>**Section 12.2** | Equal-variances *t*-test and estimator of the difference between two means: independent samples<br>**Section 13.1**<br>Unequal-variances *t*-test and estimator of the difference between two means: independent samples<br>**Section 13.1**<br>*t*-test and estimator of mean difference<br>**Section 13.3**<br>*F*-test and estimator of ratio of two variances<br>**Section 13.4** | One-way analysis of varianc<br>**Section 14.1**<br>LSD multiple comparison method<br>**Section 14.2**<br>Tukey's multiple compariso method<br>**Section 14.2**<br>Two-way analysis of varianc<br>**Section 14.4**<br>Two-factor analysis of varia<br>**Section 14.5** |
| | Nominal | Frequency distribution<br>**Section 2.2**<br>Bar chart<br>**Section 2.2**<br>Pie chart<br>**Section 2.2**<br>*z*-test and estimator of a proportion<br>**Section 12.3**<br>Chi-squared goodness-of-fit test<br>**Section 15.1** | *z*-test and estimator of the difference between two proportions<br>**Section 13.5**<br>Chi-squared test of a contingency table<br>**Section 15.2** | Chi-squared test of a contingency table<br>**Section 15.2** |
| | Ordinal | Median<br>**Section 4.1**<br>Percentiles and quartiles<br>**Section 4.3**<br>Box plot<br>**Section 4.3** | | |

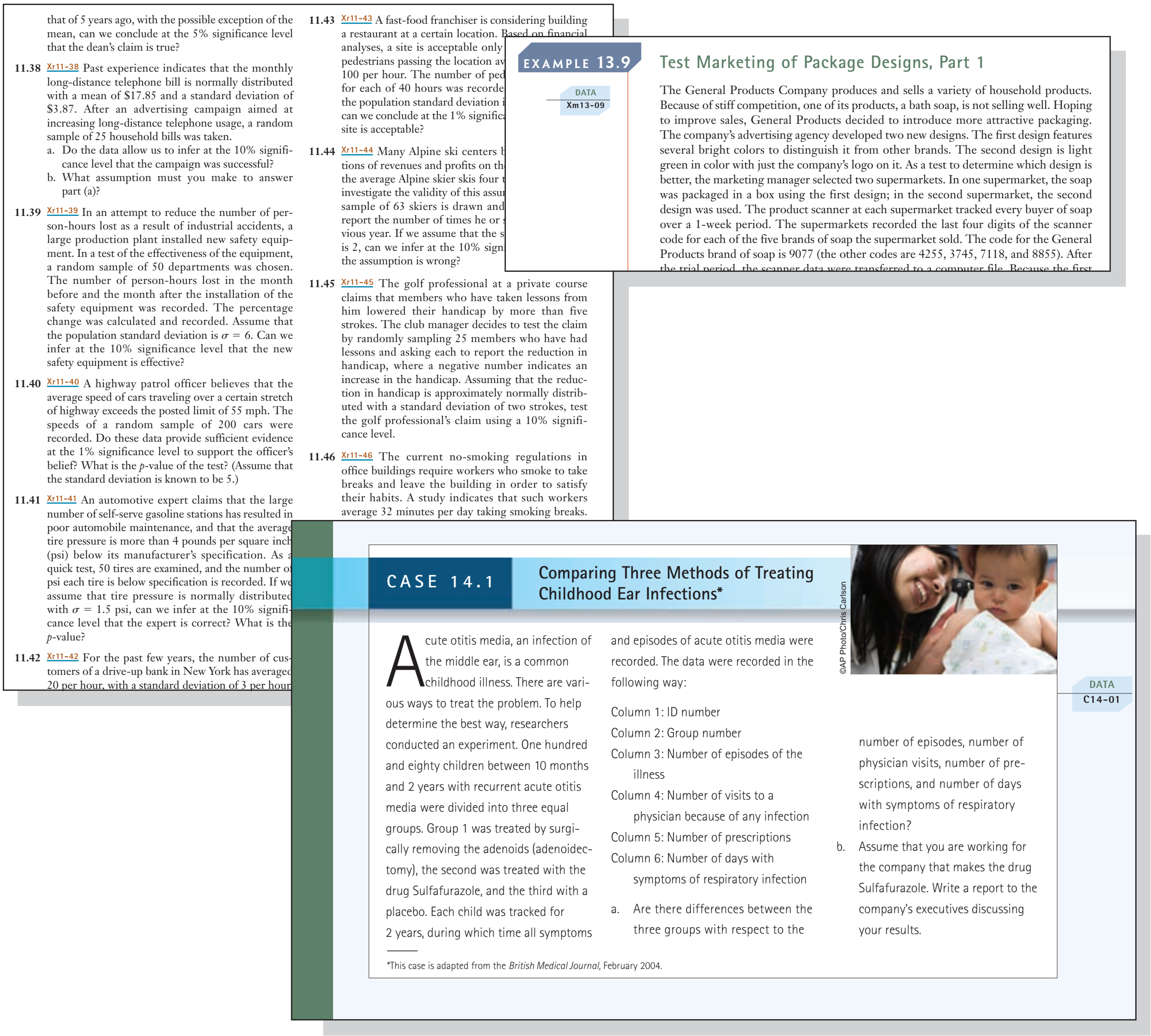
**A Guide to Statistical Techniques**, found on the inside front cover of the text, pulls everything together into one useful table that helps students identify which technique to perform based on the problem objective and data type.

## More Data Sets

**A total of 715 data sets** available to be downloaded provide ample practice. These data sets often contain real data, are typically large, and are formatted for Excel, Minitab, SPSS, SAS, JMP IN, and ASCII.

**Prevalent use of data in examples, exercises, and cases** is highlighted by the accompanying data icon, which alerts students to go to Keller's website.

DATA
Xm13-02



that of 5 years ago, with the possible exception of the mean, can we conclude at the 5% significance level that the dean's claim is true?

**11.38** Xr11-38 Past experience indicates that the monthly long-distance telephone bill is normally distributed with a mean of $17.85 and a standard deviation of $3.87. After an advertising campaign aimed at increasing long-distance telephone usage, a random sample of 25 household bills was taken.

a. Do the data allow us to infer at the 10% significance level that the campaign was successful?

b. What assumption must you make to answer part (a)?

**11.39** Xr11-39 In an attempt to reduce the number of person-hours lost as a result of industrial accidents, a large production plant installed new safety equipment. In a test of the effectiveness of the equipment, a random sample of 50 departments was chosen. The number of person-hours lost in the month before and the month after the installation of the safety equipment was recorded. The percentage change was calculated and recorded. Assume that the population standard deviation is $\sigma = 6$. Can we infer at the 10% significance level that the new safety equipment is effective?

**11.40** Xr11-40 A highway patrol officer believes that the average speed of cars traveling over a certain stretch of highway exceeds the posted limit of 55 mph. The speeds of a random sample of 200 cars were recorded. Do these data provide sufficient evidence at the 1% significance level to support the officer's belief? What is the *p*-value of the test? (Assume that the standard deviation is known to be 5.)

**11.41** Xr11-41 An automotive expert claims that the large number of self-serve gasoline stations has resulted in poor automobile maintenance, and that the average tire pressure is more than 4 pounds per square inch (psi) below its manufacturer's specification. As a quick test, 50 tires are examined, and the number of psi each tire is below specification is recorded. If we assume that tire pressure is normally distributed with $\sigma = 1.5$ psi, can we infer at the 10% significance level that the expert is correct? What is the *p*-value?

**11.42** Xr11-42 For the past few years, the number of customers of a drive-up bank in New York has averaged 20 per hour, with a standard deviation of 3 per hour

**11.43** Xr11-43 A fast-food franchiser is considering building a restaurant at a certain location. Based on financial analyses, a site is acceptable only pedestrians passing the location av 100 per hour. The number of ped for each of 40 hours was recorde the population standard deviation i can we conclude at the 1% significa site is acceptable?

**11.44** Xr11-44 Many Alpine ski centers b tions of revenues and profits on th the average Alpine skier skis four t investigate the validity of this assur sample of 63 skiers is drawn and report the number of times he or s vious year. If we assume that the s is 2, can we infer at the 10% signi the assumption is wrong?

**11.45** Xr11-45 The golf professional at a private course claims that members who have taken lessons from him lowered their handicap by more than five strokes. The club manager decides to test the claim by randomly sampling 25 members who have had lessons and asking each to report the reduction in handicap, where a negative number indicates an increase in the handicap. Assuming that the reduction in handicap is approximately normally distributed with a standard deviation of two strokes, test the golf professional's claim using a 10% significance level.

**11.46** Xr11-46 The current no-smoking regulations in office buildings require workers who smoke to take breaks and leave the building in order to satisfy their habits. A study indicates that such workers average 32 minutes per day taking smoking breaks.

EXAMPLE 13.9

DATA Xm13-09

Test Marketing of Package Designs, Part 1

The General Products Company produces and sells a variety of household products. Because of stiff competition, one of its products, a bath soap, is not selling well. Hoping to improve sales, General Products decided to introduce more attractive packaging. The company's advertising agency developed two new designs. The first design features several bright colors to distinguish it from other brands. The second design is light green in color with just the company's logo on it. As a test to determine which design is better, the marketing manager selected two supermarkets. In one supermarket, the soap was packaged in a box using the first design; in the second supermarket, the second design was used. The product scanner at each supermarket tracked every buyer of soap over a 1-week period. The supermarkets recorded the last four digits of the scanner code for each of the five brands of soap the supermarket sold. The code for the General Products brand of soap is 9077 (the other codes are 4255, 3745, 7118, and 8855). After the trial period, the scanner data were transferred to a computer file. Because the first

CASE 14.1

Comparing Three Methods of Treating Childhood Ear Infections*

©AP Photo/Chris Carlson

DATA C14-01

Acute otitis media, an infection of the middle ear, is a common childhood illness. There are various ways to treat the problem. To help determine the best way, researchers conducted an experiment. One hundred and eighty children between 10 months and 2 years with recurrent acute otitis media were divided into three equal groups. Group 1 was treated by surgically removing the adenoids (adenoidectomy), the second was treated with the drug Sulfafurazole, and the third with a placebo. Each child was tracked for 2 years, during which time all symptoms and episodes of acute otitis media were recorded. The data were recorded in the following way:

Column 1: ID number

Column 2: Group number

Column 3: Number of episodes of the illness

Column 4: Number of visits to a physician because of any infection

Column 5: Number of prescriptions

Column 6: Number of days with symptoms of respiratory infection

a. Are there differences between the three groups with respect to the number of episodes, number of physician visits, number of prescriptions, and number of days with symptoms of respiratory infection?

b. Assume that you are working for the company that makes the drug Sulfafurazole. Write a report to the company's executives discussing your results.

*This case is adapted from the *British Medical Journal*, February 2004.

## Flexible to Use

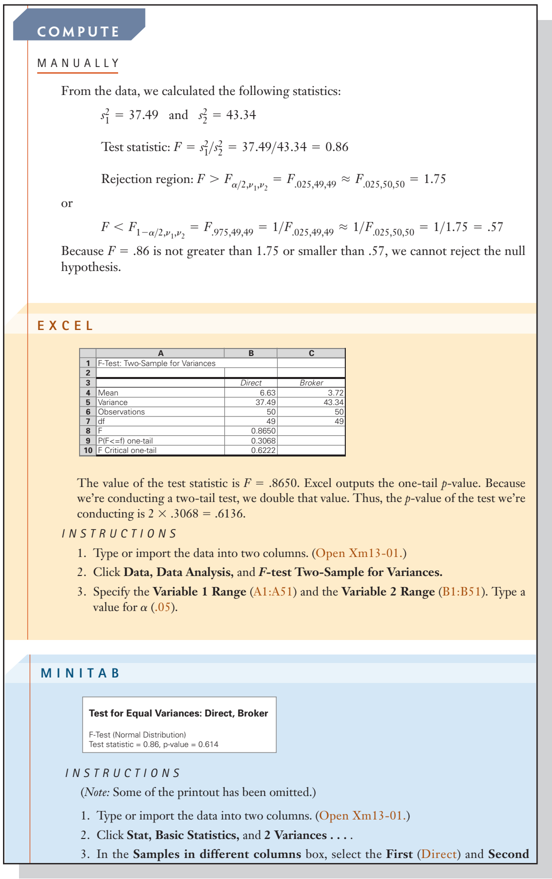
Although many texts today incorporate the use of the computer, ***Statistics for Management and Economics*** is designed for maximum flexibility and ease of use for both instructors and students. To this end, parallel illustration of both manual and computer printouts is provided throughout the text. This approach **allows you to choose** which, if any, computer program to use. Regardless of the method or software you choose, the output and instructions that you need are provided!

## *Compute* the Statistics

Once the correct technique has been identified, examples take students to the next level within the solution by asking them to compute the statistics.

**Manual calculation** of the problem is presented first in each "Compute" section of the examples.

**Step-by-step instructions** in the use of **Excel** and **Minitab** immediately follow the manual presentation. Instruction appears in the book with the printouts—there's no need to incur the extra expense of separate software manuals. SPSS and JMP IN are also available at no cost on the Keller companion website.



COMPUTE

MANUALLY

From the data, we calculated the following statistics:

$$s_1^2 = 37.49 \quad \text{and} \quad s_2^2 = 43.34$$

$$\text{Test statistic: } F = s_1^2/s_2^2 = 37.49/43.34 = 0.86$$

$$\text{Rejection region: } F > F_{\alpha/2,\nu_1,\nu_2} = F_{.025,49,49} \approx F_{.025,50,50} = 1.75$$

or

$$F < F_{1-\alpha/2,\nu_1,\nu_2} = F_{.975,49,49} = 1/F_{.025,49,49} \approx 1/F_{.025,50,50} = 1/1.75 = .57$$

Because $F = .86$ is not greater than 1.75 or smaller than .57, we cannot reject the null hypothesis.

EXCEL

| | A | B | C |
|---|---|---|---|
| 1 | F-Test: Two-Sample for Variances | | |
| 2 | | | |
| 3 | | *Direct* | *Broker* |
| 4 | Mean | 6.63 | 3.72 |
| 5 | Variance | 37.49 | 43.34 |
| 6 | Observations | 50 | 50 |
| 7 | df | 49 | 49 |
| 8 | F | 0.8650 | |
| 9 | P(F<=f) one-tail | 0.3068 | |
| 10 | F Critical one-tail | 0.6222 | |

The value of the test statistic is $F = .8650$. Excel outputs the one-tail $p$-value. Because we're conducting a two-tail test, we double that value. Thus, the $p$-value of the test we're conducting is $2 \times .3068 = .6136$.

*INSTRUCTIONS*

1. Type or import the data into two columns. (Open Xm13-01.)
2. Click **Data, Data Analysis,** and ***F*-test Two-Sample for Variances.**
3. Specify the **Variable 1 Range** (A1:A51) and the **Variable 2 Range** (B1:B51). Type a value for $\alpha$ (.05).

MINITAB

**Test for Equal Variances: Direct, Broker**

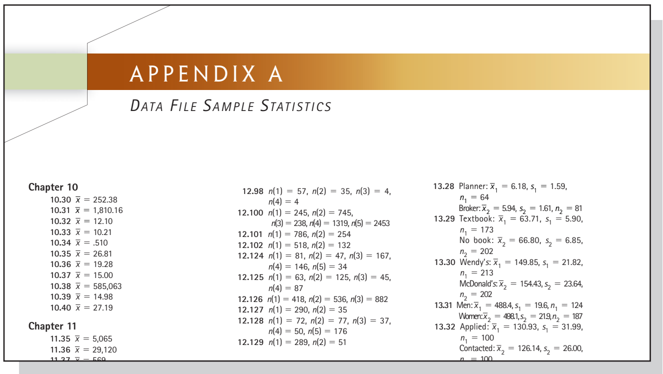
F-Test (Normal Distribution)
Test statistic = 0.86, p-value = 0.614

*INSTRUCTIONS*

(*Note:* Some of the printout has been omitted.)

1. Type or import the data into two columns. (Open Xm13-01.)
2. Click **Stat, Basic Statistics,** and **2 Variances . . . .**
3. In the **Samples in different columns** box, select the **First** (Direct) and **Second**

**Appendix A** provides summary statistics that allow students to solve applied exercises with data files by hand. Offering unparalleled flexibility, this feature allows virtually *all* exercises to be solved by hand!



APPENDIX A

*DATA FILE SAMPLE STATISTICS*

**Chapter 10**

**10.30** $\bar{x} = 252.38$
**10.31** $\bar{x} = 1{,}810.16$
**10.32** $\bar{x} = 12.10$
**10.33** $\bar{x} = 10.21$
**10.34** $\bar{x} = .510$
**10.35** $\bar{x} = 26.81$
**10.36** $\bar{x} = 19.28$
**10.37** $\bar{x} = 15.00$
**10.38** $\bar{x} = 585{,}063$
**10.39** $\bar{x} = 14.98$
**10.40** $\bar{x} = 27.19$

**Chapter 11**

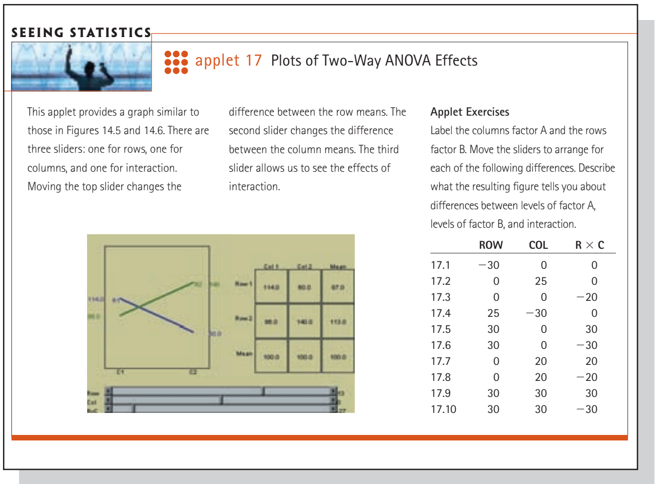
**11.35** $\bar{x} = 5{,}065$
**11.36** $\bar{x} = 29{,}120$
**11.37** $\bar{x} = 569$

**12.98** $n(1) = 57, n(2) = 35, n(3) = 4, n(4) = 4$
**12.100** $n(1) = 245, n(2) = 745, n(3) = 238, n(4) = 1319, n(5) = 2453$
**12.101** $n(1) = 786, n(2) = 254$
**12.102** $n(1) = 518, n(2) = 132$
**12.124** $n(1) = 81, n(2) = 47, n(3) = 167, n(4) = 146, n(5) = 34$
**12.125** $n(1) = 63, n(2) = 125, n(3) = 45, n(4) = 87$
**12.126** $n(1) = 418, n(2) = 536, n(3) = 882$
**12.127** $n(1) = 290, n(2) = 35$
**12.128** $n(1) = 72, n(2) = 77, n(3) = 37, n(4) = 50, n(5) = 176$
**12.129** $n(1) = 289, n(2) = 51$

**13.28** Planner: $\bar{x}_1 = 6.18, s_1 = 1.59, n_1 = 64$
Broker: $\bar{x}_2 = 5.94, s_2 = 1.61, n_2 = 81$
**13.29** Textbook: $\bar{x}_1 = 63.71, s_1 = 5.90, n_1 = 173$
No book: $\bar{x}_2 = 66.80, s_2 = 6.85, n_2 = 202$
**13.30** Wendy's: $\bar{x}_1 = 149.85, s_1 = 21.82, n_1 = 213$
McDonald's: $\bar{x}_2 = 154.43, s_2 = 23.64, n_2 = 202$
**13.31** Men: $\bar{x}_1 = 488.4, s_1 = 19.6, n_1 = 124$
Women: $\bar{x}_2 = 498.1, s_2 = 21.9, n_2 = 187$
**13.32** Applied: $\bar{x}_1 = 130.93, s_1 = 31.99, n_1 = 100$
Contacted: $\bar{x}_2 = 126.14, s_2 = 26.00, n_2 = 100$

## Flexible Learning

For visual learners, the **Seeing Statistics** feature refers to online Java applets developed by Gary McClelland of the University of Colorado, which use the interactive nature of the web to illustrate key statistical concepts. With 19 applets and 82 follow-up exercises, students can explore and interpret statistical concepts, leading them to greater intuitive understanding. All Seeing Statistics applets can be found on CourseMate.



**SEEING STATISTICS**

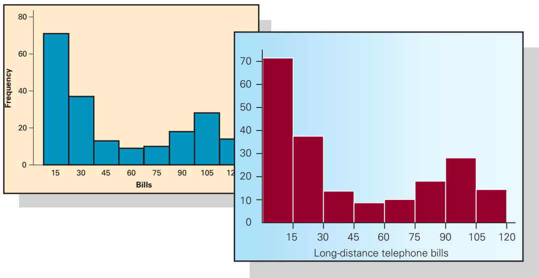
applet 17 Plots of Two-Way ANOVA Effects

This applet provides a graph similar to those in Figures 14.5 and 14.6. There are three sliders: one for rows, one for columns, and one for interaction. Moving the top slider changes the difference between the row means. The second slider changes the difference between the column means. The third slider allows us to see the effects of interaction.

**Applet Exercises**

Label the columns factor A and the rows factor B. Move the sliders to arrange for each of the following differences. Describe what the resulting figure tells you about differences between levels of factor A, levels of factor B, and interaction.

| | ROW | COL | R × C |
|---|---|---|---|
| 17.1 | −30 | 0 | 0 |
| 17.2 | 0 | 25 | 0 |
| 17.3 | 0 | 0 | −20 |
| 17.4 | 25 | −30 | 0 |
| 17.5 | 30 | 0 | 30 |
| 17.6 | 30 | 0 | −30 |
| 17.7 | 0 | 20 | 20 |
| 17.8 | 0 | 20 | −20 |
| 17.9 | 30 | 30 | 30 |
| 17.10 | 30 | 30 | −30 |



**Ample use of graphics** provides students many opportunities to see statistics in all its forms. In addition to manually presented figures throughout the text, Excel and Minitab graphic outputs are given for students to compare to their own results.

## Applied: Bridging the Gap

In the real world, it is not enough to know *how* to generate the statistics. To be truly effective, a business person must also know how to **interpret and articulate** the results. Furthermore, students need a framework to understand and apply statistics **within a realistic setting** by using realistic data in exercises, examples, and case studies.

### *Interpret* the Results

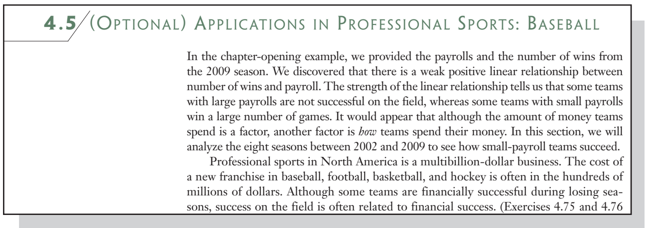
INTERPRET

Examples round out the final component of the identify–compute–interpret approach by asking students to interpret the results in the context of a business-related decision. This final step motivates and shows how statistics is used in everyday business situations.

## An Applied Approach

With **Applications in . . .** sections and boxes, ***Statistics for Management and Economics*** now includes 45 **applications** (in finance, marketing, operations management, human resources, economics, and accounting) highlighting how statistics is used in those professions. For example, "Applications in Finance: Portfolio Diversification and Asset Allocation" shows how probability is used to help select stocks to minimize risk. A new optional section, "Applications in Professional Sports: Baseball" contains a subsection on the success of the Oakland Athletics.

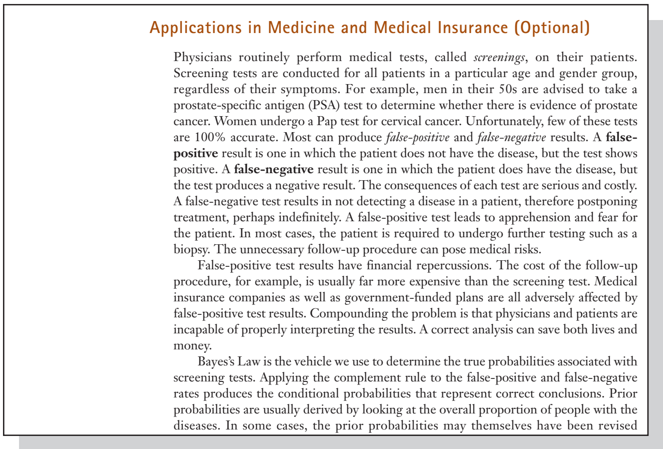
In addition to sections and boxes, **Applications in . . . exercises** can be found within the exercise sections to further reinforce the big picture.



4.5 (OPTIONAL) APPLICATIONS IN PROFESSIONAL SPORTS: BASEBALL

In the chapter-opening example, we provided the payrolls and the number of wins from the 2009 season. We discovered that there is a weak positive linear relationship between number of wins and payroll. The strength of the linear relationship tells us that some teams with large payrolls are not successful on the field, whereas some teams with small payrolls win a large number of games. It would appear that although the amount of money teams spend is a factor, another factor is *how* teams spend their money. In this section, we will analyze the eight seasons between 2002 and 2009 to see how small-payroll teams succeed.

Professional sports in North America is a multibillion-dollar business. The cost of a new franchise in baseball, football, basketball, and hockey is often in the hundreds of millions of dollars. Although some teams are financially successful during losing seasons, success on the field is often related to financial success. (Exercises 4.75 and 4.76

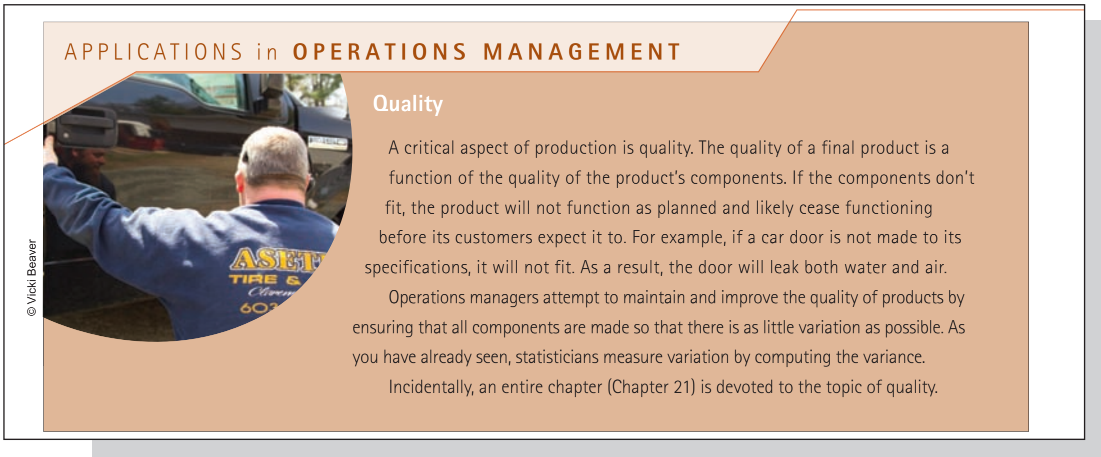


Applications in Medicine and Medical Insurance (Optional)

Physicians routinely perform medical tests, called *screenings*, on their patients. Screening tests are conducted for all patients in a particular age and gender group, regardless of their symptoms. For example, men in their 50s are advised to take a prostate-specific antigen (PSA) test to determine whether there is evidence of prostate cancer. Women undergo a Pap test for cervical cancer. Unfortunately, few of these tests are 100% accurate. Most can produce *false-positive* and *false-negative* results. A **false-positive** result is one in which the patient does not have the disease, but the test shows positive. A **false-negative** result is one in which the patient does have the disease, but the test produces a negative result. The consequences of each test are serious and costly. A false-negative test results in not detecting a disease in a patient, therefore postponing treatment, perhaps indefinitely. A false-positive test leads to apprehension and fear for the patient. In most cases, the patient is required to undergo further testing such as a biopsy. The unnecessary follow-up procedure can pose medical risks.

False-positive test results have financial repercussions. The cost of the follow-up procedure, for example, is usually far more expensive than the screening test. Medical insurance companies as well as government-funded plans are all adversely affected by false-positive test results. Compounding the problem is that physicians and patients are incapable of properly interpreting the results. A correct analysis can save both lives and money.

Bayes's Law is the vehicle we use to determine the true probabilities associated with screening tests. Applying the complement rule to the false-positive and false-negative rates produces the conditional probabilities that represent correct conclusions. Prior probabilities are usually derived by looking at the overall proportion of people with the diseases. In some cases, the prior probabilities may themselves have been revised



APPLICATIONS in OPERATIONS MANAGEMENT

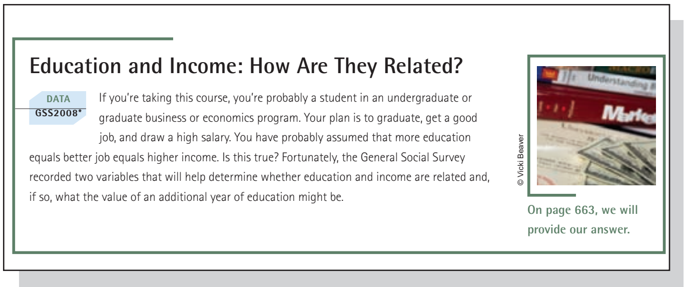
© Vicki Beaver

Quality

A critical aspect of production is quality. The quality of a final product is a function of the quality of the product's components. If the components don't fit, the product will not function as planned and likely cease functioning before its customers expect it to. For example, if a car door is not made to its specifications, it will not fit. As a result, the door will leak both water and air.

Operations managers attempt to maintain and improve the quality of products by ensuring that all components are made so that there is as little variation as possible. As you have already seen, statisticians measure variation by computing the variance.

Incidentally, an entire chapter (Chapter 21) is devoted to the topic of quality.



## Education and Income: How Are They Related?

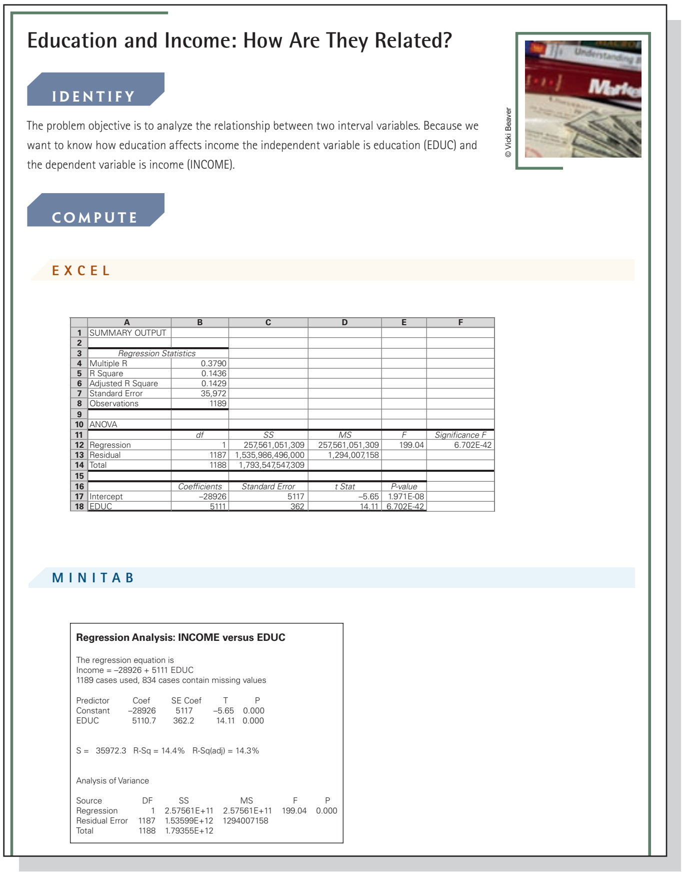
DATA
GSS2008*

If you're taking this course, you're probably a student in an undergraduate or graduate business or economics program. Your plan is to graduate, get a good job, and draw a high salary. You have probably assumed that more education equals better job equals higher income. Is this true? Fortunately, the General Social Survey recorded two variables that will help determine whether education and income are related and, if so, what the value of an additional year of education might be.

© Vicki Beaver

On page 663, we will provide our answer.

**Chapter-opening examples and solutions** present compelling discussions of how the techniques and concepts introduced in that chapter are applied to real-world problems. These examples are then revisited with a solution as each chapter unfolds, applying the methodologies introduced in the chapter.



## Education and Income: How Are They Related?

© Vicki Beaver

IDENTIFY

The problem objective is to analyze the relationship between two interval variables. Because we want to know how education affects income the independent variable is education (EDUC) and the dependent variable is income (INCOME).

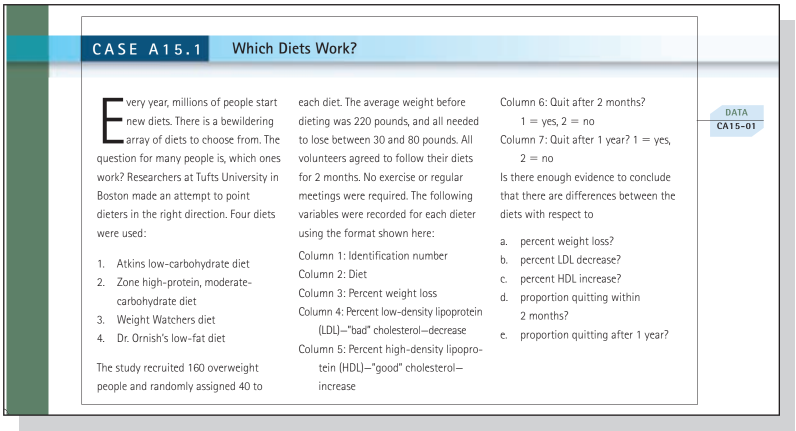
COMPUTE

EXCEL

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | SUMMARY OUTPUT | | | | | |
| 2 | | | | | | |
| 3 | *Regression Statistics* | | | | | |
| 4 | Multiple R | 0.3790 | | | | |
| 5 | R Square | 0.1436 | | | | |
| 6 | Adjusted R Square | 0.1429 | | | | |
| 7 | Standard Error | 35,972 | | | | |
| 8 | Observations | 1189 | | | | |
| 9 | | | | | | |
| 10 | ANOVA | | | | | |
| 11 | | *df* | *SS* | *MS* | *F* | *Significance F* |
| 12 | Regression | 1 | 257,561,051,309 | 257,561,051,309 | 199.04 | 6.702E-42 |
| 13 | Residual | 1187 | 1,535,986,496,000 | 1,294,007,158 | | |
| 14 | Total | 1188 | 1,793,547,547,309 | | | |
| 15 | | | | | | |
| 16 | | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | |
| 17 | Intercept | –28926 | 5117 | –5.65 | 1.971E-08 | |
| 18 | EDUC | 5111 | 362 | 14.11 | 6.702E-42 | |

MINITAB

```
Regression Analysis: INCOME versus EDUC

The regression equation is
Income = –28926 + 5111 EDUC
1189 cases used, 834 cases contain missing values

Predictor      Coef    SE Coef       T      P
Constant     –28926     5117     –5.65  0.000
EDUC         5110.7     362.2    14.11  0.000


S =  35972.3  R-Sq = 14.4%  R-Sq(adj) = 14.3%


Analysis of Variance

Source           DF         SS           MS         F       P
Regression        1  2.57561E+11  2.57561E+11  199.04  0.000
Residual Error 1187  1.53599E+12  1294007158
Total          1188  1.79355E+12
```



CASE A15.1 Which Diets Work?

DATA CA15-01

Every year, millions of people start new diets. There is a bewildering array of diets to choose from. The question for many people is, which ones work? Researchers at Tufts University in Boston made an attempt to point dieters in the right direction. Four diets were used:

1. Atkins low-carbohydrate diet
2. Zone high-protein, moderate-carbohydrate diet
3. Weight Watchers diet
4. Dr. Ornish's low-fat diet

The study recruited 160 overweight people and randomly assigned 40 to each diet. The average weight before dieting was 220 pounds, and all needed to lose between 30 and 80 pounds. All volunteers agreed to follow their diets for 2 months. No exercise or regular meetings were required. The following variables were recorded for each dieter using the format shown here:

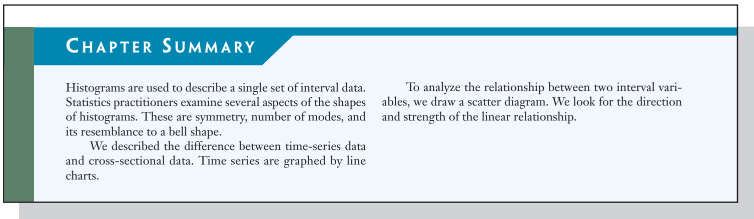
Column 1: Identification number  
Column 2: Diet  
Column 3: Percent weight loss  
Column 4: Percent low-density lipoprotein (LDL)—"bad" cholesterol—decrease  
Column 5: Percent high-density lipoprotein (HDL)—"good" cholesterol—increase  
Column 6: Quit after 2 months? 1 = yes, 2 = no  
Column 7: Quit after 1 year? 1 = yes, 2 = no

Is there enough evidence to conclude that there are differences between the diets with respect to

a. percent weight loss?
b. percent LDL decrease?
c. percent HDL increase?
d. proportion quitting within 2 months?
e. proportion quitting after 1 year?

Many of the **examples, exercises, and cases are based on actual studies** performed by statisticians and published in journals, newspapers, and magazines, or presented at conferences. Many data files were recreated to produce the original results.

**A total of 1825 exercises**, many of them new or updated, offer ample practice for students to use statistics in an applied context.



CHAPTER SUMMARY

Histograms are used to describe a single set of interval data. Statistics practitioners examine several aspects of the shapes of histograms. These are symmetry, number of modes, and its resemblance to a bell shape.

We described the difference between time-series data and cross-sectional data. Time series are graphed by line charts.

To analyze the relationship between two interval variables, we draw a scatter diagram. We look for the direction and strength of the linear relationship.

# Resources

## Learning Resources

**The Essential Textbook Resources website:** At the Keller website, you'll find materials previously on the student CD, including: Interactive concept simulation exercises from ***Seeing Statistics***, the **Data Analysis Plus** add-in, 715 data sets, optional topics, and 35 appendixes (for more information, please visit www.cengage.com/bstatistics/keller).

**Student Solutions Manual** (ISBN: 1111531889): Students can check their understanding with this manual, which includes worked solutions of even-numbered exercises from the text.

***Seeing Statistics* by Gary McClelland**: This online product is flexible and addresses different learning styles. It presents the visual nature of statistical concepts using more than 150 Java applets to create an intuitive learning environment. Also included are relevant links to examples, exercises, definitions, and search and navigation capabilities.

## Teaching Resources

To access both student and faculty resources, please visit **www.cengage.com/login.**

# Acknowledgments

Although there is only one name on the cover of this book, the number of people who made contributions is large. I would like to acknowledge the work of all of them, with particular emphasis on the following: Paul Baum, California State University, Northridge, and John Lawrence, California State University, Fullerton, reviewed the page proofs. Their job was to find errors in presentation, arithmetic, and composition.The following individuals played important roles in the production of this book Senior Acquisitions Editor: Charles McCormick, Jr., Developmental Editor: Elizabeth Lowry, Content Project Manager: Jacquelyn K Featherly and Project Manager Lynn Lustberg (For all remaining errors, place the blame where it belongs–on me.) Their advice and suggestions made my task considerably easier.

Fernando Rodriguez produced the test bank stored on the Instructor's Suite CD-ROM.

Trent Tucker, Wilfrid Laurier University, and Zvi Goldstein, California State University, Fullerton, each produced a set of PowerPoint slides.

The author extends thanks also to the survey participants and reviewers of the previous editions: Paul Baum, California State University, Northridge; Nagraj Balakrishnan, Clemson University; Howard Clayton, Auburn University; Philip Cross, Georgetown University; Barry Cuffe, Wingate University; Ernest Demba, Washington University–St. Louis; Neal Duffy, State University of New York, Plattsburgh; John Dutton, North Carolina State University; Erick Elder, University of Arkansas; Mohammed El-Saidi, Ferris State University; Grace Esimai, University of Texas at Arlington; Abe Feinberg, California State University, Northridge; Samuel Graves, Boston College; Robert Gould, UCLA; John Hebert, Virginia Tech; James Hightower, California State University, Fullerton; Bo Honore, Princeton University; Onisforos Iordanou, Hunter College; Gordon Johnson, California State University,

Northridge; Hilke Kayser, Hamilton College; Kenneth Klassen, California State University, Northridge; Roger Kleckner, Bowling Green State University–Firelands; Harry Kypraios, Rollins College; John Lawrence, California State University, Fullerton; Dennis Lin, Pennsylvania State University; Neal Long, Stetson University; George Marcoulides, California State University, Fullerton; Paul Mason, University of North Florida; Walter Mayer, University of Mississippi; John McDonald, Flinders University; Richard McGowan, Boston College; Richard McGrath, Bowling Green State University; Amy Miko, St. Francis College; Janis Miller, Clemson University; Glenn Milligan, Ohio State University; James Moran, Oregon State University; Patricia Mullins, University of Wisconsin; Kevin Murphy, Oakland University; Pin Ng, University of Illinois; Des Nicholls, Australian National University; Andrew Paizis, Queens College; David Pentico, Duquesne University; Ira Perelle, Mercy College; Nelson Perera, University of Wollongong; Amy Puelz, Southern Methodist University; Lawrence Ries, University of Missouri; Colleen Quinn, Seneca College; Tony Quon, University of Ottawa; Madhu Rao, Bowling Green State University; Phil Roth, Clemson University; Farhad Saboori, Albright College; Don St. Jean, George Brown College; Hedayeh Samavati, Indiana–Purdue University; Sandy Shroeder, Ohio Northern University; Jineshwar Singh, George Brown College; Natalia Smirnova, Queens College; Eric Sowey, University of New South Wales; Cyrus Stanier, Virginia Tech; Stan Stephenson, Southwest Texas State University; Arnold Stromberg, University of Kentucky; Steve Thorpe, University of Northern Iowa; Sheldon Vernon, Houston Baptist University; and W. F. Younkin, University of Miami.

# Statistics

## FOR MANAGEMENT AND ECONOMICS ABBREVIATED

9e

*This page intentionally left blank*

1

© sellingpix/Shutterstock

# WHAT IS STATISTICS?

## Introduction

Statistics is a way to get information from data. That's it! Most of this textbook is devoted to describing how, when, and why managers and statistics practitioners* conduct statistical procedures. You may ask, "If that's all there is to statistics, why is this book (and most other statistics books) so large?" The answer is that students of applied statistics will be exposed to different kinds of information and data. We demonstrate some of these with a case and two examples that are featured later in this book.

The first may be of particular interest to you.

*The term *statistician* is used to describe so many different kinds of occupations that it has ceased to have any meaning. It is used, for example, to describe a person who calculates baseball statistics as well as an individual educated in statistical principles. We will describe the former as a *statistics practitioner* and the

(*continued*)

## EXAMPLE 3.3 Business Statistics Marks (See Chapter 3)

A student enrolled in a business program is attending his first class of the required statistics course. The student is somewhat apprehensive because he believes the myth that the course is difficult. To alleviate his anxiety, the student asks the professor about last year's marks. Because this professor is friendly and helpful, like all other statistics professors, he obliges the student and provides a list of the final marks, which are composed of term work plus the final exam. What information can the student obtain from the list?

This is a typical statistics problem. The student has the data (marks) and needs to apply statistical techniques to get the information he requires. This is a function of **descriptive statistics**.

## Descriptive Statistics

Descriptive statistics deals with methods of organizing, summarizing, and presenting data in a convenient and informative way. One form of descriptive statistics uses graphical techniques that allow statistics practitioners to present data in ways that make it easy for the reader to extract useful information. In Chapters 2 and 3 we will present a variety of graphical methods.

Another form of descriptive statistics uses numerical techniques to summarize data. One such method that you have already used frequently calculates the average or mean. In the same way that you calculate the average age of the employees of a company, we can compute the mean mark of last year's statistics course. Chapter 4 introduces several numerical statistical measures that describe different features of the data.

The actual technique we use depends on what specific information we would like to extract. In this example, we can see at least three important pieces of information. The first is the "typical" mark. We call this a *measure of central location*. The average is one such measure. In Chapter 4, we will introduce another useful measure of central location, the median. Suppose the student was told that the average mark last year was 67. Is this enough information to reduce his anxiety? The student would likely respond "No" because he would like to know whether most of the marks were close to 67 or were scattered far below and above the average. He needs a *measure of variability*. The simplest such measure is the *range*, which is calculated by subtracting the smallest number from the largest. Suppose the largest mark is 96 and the smallest is 24. Unfortunately, this provides little information since it is based on only two marks. We need other measures—these will be introduced in Chapter 4. Moreover, the student must determine more about the marks. In particular, he needs to know how the marks are distributed between 24 and 96. The best way to do this is to use a graphical technique, the histogram, which will be introduced in Chapter 3.

latter as a *statistician*. A statistics practitioner is a person who uses statistical techniques properly. Examples of statistics practitioners include the following:

1. a financial analyst who develops stock portfolios based on historical rates of return;
2. an economist who uses statistical models to help explain and predict variables such as inflation rate, unemployment rate, and changes in the gross domestic product; and
3. a market researcher who surveys consumers and converts the responses into useful information.

Our goal in this book is to convert you into one such capable individual.

The term *statistician* refers to an individual who works with the mathematics of statistics. His or her work involves research that develops techniques and concepts that in the future may help the statistics practitioner. Statisticians are also statistics practitioners, frequently conducting empirical research and consulting. If you're taking a statistics course, your instructor is probably a statistician.

### Case 12.1 Pepsi's Exclusivity Agreement with a University (see Chapter 12)

In the last few years, colleges and universities have signed exclusivity agreements with a variety of private companies. These agreements bind the university to sell these companies' products exclusively on the campus. Many of the agreements involve food and beverage firms.

A large university with a total enrollment of about 50,000 students has offered Pepsi-Cola an exclusivity agreement that would give Pepsi exclusive rights to sell its products at all university facilities for the next year with an option for future years. In return, the university would receive 35% of the on-campus revenues and an additional lump sum of $200,000 per year. Pepsi has been given 2 weeks to respond.

The management at Pepsi quickly reviews what it knows. The market for soft drinks is measured in terms of 12-ounce cans. Pepsi currently sells an average of 22,000 cans per week over the 40 weeks of the year that the university operates. The cans sell for an average of one dollar each. The costs, including labor, total 30 cents per can. Pepsi is unsure of its market share but suspects it is considerably less than 50%. A quick analysis reveals that if its current market share were 25%, then, with an exclusivity agreement, Pepsi would sell 88,000 (22,000 is 25% of 88,000) cans per week or 3,520,000 cans per year. The gross revenue would be computed as follows[†]:

Gross revenue = 3,520,000 × $1.00/can = $3,520,000

This figure must be multiplied by 65% because the university would rake in 35% of the gross. Thus,

Gross revenue after deducting 35% university take
= 65% × $3,520,000 = $2,288,000

The total cost of 30 cents per can (or $1,056,000) and the annual payment to the university of $200,000 are subtracted to obtain the net profit:

Net profit = $2,288,000 − $1,056,000 − $200,000 = $1,032,000

Pepsi's current annual profit is

40 weeks × 22,000 cans/week × $.70 = $616,000

If the current market share is 25%, the potential gain from the agreement is

$1,032,000 – $616,000 = $416,000

The only problem with this analysis is that Pepsi does not know how many soft drinks are sold weekly at the university. Coke is not likely to supply Pepsi with information about its sales, which together with Pepsi's line of products constitute virtually the entire market.

Pepsi assigned a recent university graduate to survey the university's students to supply the missing information. Accordingly, she organizes a survey that asks 500 students to keep track of the number of soft drinks they purchase in the next 7 days. The responses are stored in a file C12-01 available to be downloaded. See Appendix 1 for instructions.

## Inferential Statistics

The information we would like to acquire in Case 12.1 is an estimate of annual profits from the exclusivity agreement. The data are the numbers of cans of soft drinks consumed in 7 days by the 500 students in the sample. We can use descriptive techniques to

[†] We have created an Excel spreadsheet that does the calculations for this case. See Appendix 1 for instructions on how to download this spreadsheet from Keller's website plus hundreds of datasets and much more.

learn more about the data. In this case, however, we are not so much interested in what the 500 students are reporting as in knowing the mean number of soft drinks consumed by all 50,000 students on campus. To accomplish this goal we need another branch of statistics: **inferential statistics**.

Inferential statistics is a body of methods used to draw conclusions or inferences about characteristics of populations based on sample data. The population in question in this case is the university's 50,000 students. The characteristic of interest is the soft drink consumption of this population. The cost of interviewing each student in the population would be prohibitive and extremely time consuming. Statistical techniques make such endeavors unnecessary. Instead, we can sample a much smaller number of students (the sample size is 500) and infer from the data the number of soft drinks consumed by all 50,000 students. We can then estimate annual profits for Pepsi.

## EXAMPLE 12.5 Exit Polls (see Chapter 12)

When an election for political office takes place, the television networks cancel regular programming to provide election coverage. After the ballots are counted, the results are reported. However, for important offices such as president or senator in large states, the networks actively compete to see which one will be the first to predict a winner. This is done through **exit polls** in which a random sample of voters who exit the polling booth are asked for whom they voted. From the data, the sample proportion of voters supporting the candidates is computed. A statistical technique is applied to determine whether there is enough evidence to infer that the leading candidate will garner enough votes to win. Suppose that the exit poll results from the state of Florida during the year 2000 elections were recorded. Although several candidates were running for president, the exit pollsters recorded only the votes of the two candidates who had any chance of winning: Republican George W. Bush and Democrat Albert Gore. The results (765 people who voted for either Bush or Gore) were stored in file Xm12-05. The network analysts would like to know whether they can conclude that George W. Bush will win the state of Florida.

Example 12.5 describes a common application of statistical inference. The population the television networks wanted to make inferences about is the approximately 5 million Floridians who voted for Bush or Gore for president. The sample consisted of the 765 people randomly selected by the polling company who voted for either of the two main candidates. The characteristic of the population that we would like to know is the proportion of the Florida total electorate that voted for Bush. Specifically, we would like to know whether more than 50% of the electorate voted for Bush (counting only those who voted for either the Republican or Democratic candidate). It must be made clear that we cannot predict the outcome with 100% certainty because we will not ask all 5 million actual voters for whom they voted. This is a fact that statistics practitioners and even students of statistics must understand. A sample that is only a small fraction of the size of the population can lead to correct inferences only a certain percentage of the time. You will find that statistics practitioners can control that fraction and usually set it between 90% and 99%.

Incidentally, on the night of the United States election in November 2000, the networks goofed badly. Using exit polls as well as the results of previous elections, all four networks concluded at about 8 P.M. that Al Gore would win Florida. Shortly after 10 P.M., with a large percentage of the actual vote having been counted, the networks reversed course and declared that George W. Bush would win the state. By 2 A.M., another verdict was declared: The result was too close to call. In the future, this experience will likely be used by statistics instructors when teaching how *not* to use statistics.

Notice that, contrary to what you probably believed, data are not necessarily numbers. The marks in Example 3.3 and the number of soft drinks consumed in a week in Case 12.1, of course, are numbers; however, the votes in Example 12.5 are not. In Chapter 2, we will discuss the different types of data you will encounter in statistical applications and how to deal with them.

## 1.1 Key Statistical Concepts

Statistical inference problems involve three key concepts: the population, the sample, and the statistical inference. We now discuss each of these concepts in more detail.

### Population

A **population** is the group of all items of interest to a statistics practitioner. It is frequently very large and may, in fact, be infinitely large. In the language of statistics, *population* does not necessarily refer to a group of people. It may, for example, refer to the population of ball bearings produced at a large plant. In Case 12.1, the population of interest consists of the 50,000 students on campus. In Example 12.5, the population consists of the Floridians who voted for Bush or Gore.

A descriptive measure of a population is called a **parameter**. The parameter of interest in Case 12.1 is the mean number of soft drinks consumed by all the students at the university. The parameter in Example 12.5 is the proportion of the 5 million Florida voters who voted for Bush. In most applications of inferential statistics the parameter represents the information we need.

### Sample

A **sample** is a set of data drawn from the studied population. A descriptive measure of a sample is called a **statistic**. We use statistics to make inferences about parameters. In Case 12.1, the statistic we would compute is the mean number of soft drinks consumed in the last week by the 500 students in the sample. We would then use the sample mean to infer the value of the population mean, which is the parameter of interest in this problem. In Example 12.5, we compute the proportion of the sample of 765 Floridians who voted for Bush. The sample statistic is then used to make inferences about the population of all 5 million votes—that is, we predict the election results even before the actual count.

### Statistical Inference

**Statistical inference** is the process of making an estimate, prediction, or decision about a population based on sample data. Because populations are almost always very large, investigating each member of the population would be impractical and expensive. It is far easier and cheaper to take a sample from the population of interest and draw conclusions or make estimates about the population on the basis of information provided by the sample. However, such conclusions and estimates are not always going to be correct. For this reason, we build into the statistical inference a measure of reliability. There are two such measures: the **confidence level** and the **significance level**. The *confidence level* is the proportion of times that an estimating procedure will be correct. For example, in Case 12.1, we will produce an estimate of the average number of soft drinks to be consumed by all 50,000 students that has a confidence level of 95%. In other words, estimates based on this form of statistical inference will be correct 95% of the time. When the purpose of the statistical inference is to draw a conclusion about a population, the *significance level* measures how frequently the conclusion will be wrong. For example, suppose that, as a result of the analysis in Example 12.5, we conclude that more than 50% of the electorate will vote for George W. Bush, and thus he will win the state of Florida. A 5% significance level means that samples that lead us to conclude that Bush wins the election, will be wrong 5% of the time.

## 1.2 Statistical Applications in Business

An important function of statistics courses in business and economics programs is to demonstrate that statistical analysis plays an important role in virtually all aspects of business and economics. We intend to do so through examples, exercises, and cases. However, we assume that most students taking their first statistics course have not taken courses in most of the other subjects in management programs. To understand fully how statistics is used in these and other subjects, it is necessary to know something about them. To provide sufficient background to understand the statistical application we introduce applications in accounting, economics, finance, human resources management, marketing, and operations management. We will provide readers with some background to these applications by describing their functions in two ways.

### Application Sections and Subsections

We feature five sections that describe statistical applications in the functional areas of business. For example, in Section 7.3 we show an application in finance that describes a financial analyst's use of probability and statistics to construct portfolios that decrease risk.

One section and one subsection demonstrate the uses of probability and statistics in specific industries. Section 4.5 introduces an interesting application of statistics in professional baseball. A subsection in Section 6.4 presents an application in medical testing (useful in the medical insurance industry).

### Application Boxes

For other topics that require less detailed description, we provide application boxes with a relatively brief description of the background followed by examples or exercises. These boxes are scattered throughout the book. For example, in Chapter 3 we discuss a job a marketing manager may need to undertake to determine the appropriate price for a product. To understand the context, we need to provide a description of marketing management. The statistical application will follow.

## 1.3 Large Real Data Sets

The authors believe that you learn statistics by doing statistics. For their lives after college and university, we expect our students to have access to large amounts of real data that must be summarized to acquire the information needed to make decisions. To provide practice in this vital skill we have created six large real datasets, available to be downloaded from Keller's website. Their sources are the General Social Survey (GSS) and the American National Election Survey (ANES).

## General Social Survey

Since 1972, the General Social Survey has been tracking American attitudes on a wide variety of topics. Except for the United States census, the GSS is the most frequently used sources of information about American society. The surveys now conducted every second year measure hundreds of variables and thousands of observations. We have included the results of the last four surveys (years 2002, 2004, 2006, and 2008) stored as GSS2002, GSS2004, GSS2006, and GSS2008, respectively. The survey sizes are 2,765, 2,812, 4,510, and 2,023, respectively. We have reduced the number of variables to about 60 and have deleted the responses that are known as *missing data* ("Don't know," "Refused," etc.).

We have included some demographic variables such as, age, gender, race, income, and education. Others measure political views, support for various government activities, and work. The full lists of variables for each year are stored in Appendixes GSS2002, GSS2004, GSS2006, and GSSS2008 that can be downloaded from Keller's website.

We have scattered throughout this book examples and exercises drawn from these data sets.

## American National Election Survey

The goal of the American National Election Survey is to provide data about why Americans vote as they do. The surveys are conducted in presidential election years. We have included data from the 2004 and 2008 surveys. Like the General Social Survey, the ANES includes demographic variables. It also deals with interest in the presidential election as well as variables describing political beliefs and affiliations. Online Appendixes ANES2004 and ANES2008 contain the names and definitions of the variables.

The 2008 surveys overly sampled African American and Hispanic voters. We have "adjusted" these data by randomly deleting responses from these two racial groups.

As is the case with the General Social Surveys, we have removed missing data.

# 1.4 Statistics and the Computer

In virtually all applications of statistics, the statistics practitioner must deal with large amounts of data. For example, Case 12.1 (Pepsi-Cola) involves 500 observations. To estimate annual profits, the statistics practitioner would have to perform computations on the data. Although the calculations do not require any great mathematical skill, the sheer amount of arithmetic makes this aspect of the statistical method time-consuming and tedious.

Fortunately, numerous commercially prepared computer programs are available to perform the arithmetic. We have chosen to use Microsoft Excel, which is a spreadsheet program, and Minitab, which is a statistical software package. (We use the latest versions of both software: Office 2010 and Minitab 16.) We chose Excel because we believe that it is and will continue to be the most popular spreadsheet package. One of its drawbacks is that it does not offer a complete set of the statistical techniques we introduce in this book. Consequently, we created add-ins that can be loaded onto your computer to enable you to use Excel for all statistical procedures introduced in this book. The add-ins can be downloaded and, when installed, will appear as *Data Analysis Plus*© on Excel's Add-Ins menu. Also available are introductions to Excel and Minitab, and detailed instructions for both software packages.

Appendix 1 describes the material that can be downloaded and provides instructions on how to acquire the various components.

A large proportion of the examples, exercises, and cases feature large data sets. These are denoted with the file name on an orange background. We demonstrate the solution to the statistical examples in three ways: manually, by employing Excel, and by using Minitab. Moreover, we will provide detailed instructions for all techniques.

The files contain the data needed to produce the solution. However, in many real applications of statistics, additional data are collected. For instance, in Example 12.5, the pollster often records the voter's gender and asks for other information including race, religion, education, and income. Many other data sets are similarly constructed. In later chapters, we will return to these files and require other statistical techniques to extract the needed information. (Files that contain additional data are denoted by an asterisk on the file name.)

The approach we prefer to take is to minimize the time spent on manual computations and to focus instead on selecting the appropriate method for dealing with a problem and on interpreting the output after the computer has performed the necessary computations. In this way, we hope to demonstrate that statistics can be as interesting and as practical as any other subject in your curriculum.

## Applets and Spreadsheets

Books written for statistics courses taken by mathematics or statistics majors are considerably different from this one. It is not surprising that such courses feature mathematical proofs of theorems and derivations of most procedures. When the material is covered in this way, the underlying concepts that support statistical inference are exposed and relatively easy to see. However, this book was created for an applied course in business and economics statistics. Consequently, we do not address directly the mathematical principles of statistics. However, as we pointed out previously, one of the most important functions of statistics practitioners is to properly interpret statistical results, whether produced manually or by computer. And, to correctly interpret statistics, students require an understanding of the principles of statistics.

To help students understand the basic foundation, we offer two approaches. First, we will teach readers how to create Excel spreadsheets that allow for *what-if* analyses. By changing some of the input value, students can see for themselves how statistics works. (The term is derived from *what* happens to the statistics *if* I change this value?) These spreadsheets can also be used to calculate many of the same statistics that we introduce later in this book. Second, we offer *applets*, which are computer programs that perform similar what-if analyses or simulations. The applets and the spreadsheet applications appear in several chapters and explained in greater detail.

# CHAPTER SUMMARY

### IMPORTANT TERMS

Descriptive statistics 2
Inferential statistics 4
Exit polls 4
Population 5
Parameter 5

Sample 5
Statistic 5
Statistical inference 5
Confidence level 5
Significance level 5

## Chapter Exercises

**1.1** In your own words, define and give an example of each of the following statistical terms.

a. population
b. sample
c. parameter
d. statistic
e. statistical inference

**1.2** Briefly describe the difference between descriptive statistics and inferential statistics.

**1.3** A politician who is running for the office of mayor of a city with 25,000 registered voters commissions a survey. In the survey, 48% of the 200 registered voters interviewed say they plan to vote for her.

a. What is the population of interest?
b. What is the sample?
c. Is the value 48% a parameter or a statistic? Explain.

**1.4** A manufacturer of computer chips claims that less than 10% of its products are defective. When 1,000 chips were drawn from a large production, 7.5% were found to be defective.

a. What is the population of interest?
b. What is the sample?
c. What is the parameter?
d. What is the statistic?
e. Does the value 10% refer to the parameter or to the statistic?
f. Is the value 7.5% a parameter or a statistic?
g. Explain briefly how the statistic can be used to make inferences about the parameter to test the claim.

**1.5** Suppose you believe that, in general, graduates who have majored in *your* subject are offered higher salaries upon graduating than are graduates of other programs. Describe a statistical experiment that could help test your belief.

**1.6** You are shown a coin that its owner says is fair in the sense that it will produce the same number of heads and tails when flipped a very large number of times.

a. Describe an experiment to test this claim.
b. What is the population in your experiment?
c. What is the sample?
d. What is the parameter?
e. What is the statistic?
f. Describe briefly how statistical inference can be used to test the claim.

**1.7** Suppose that in Exercise 1.6 you decide to flip the coin 100 times.

a. What conclusion would you be likely to draw if you observed 95 heads?
b. What conclusion would you be likely to draw if you observed 55 heads?
c. Do you believe that, if you flip a perfectly fair coin 100 times, you will always observe exactly 50 heads? If you answered "no," then what numbers do you think are possible? If you answered "yes," how many heads would you observe if you flipped the coin twice? Try flipping a coin twice and repeating this experiment 10 times and report the results.

**1.8** Xm01-08 The owner of a large fleet of taxis is trying to estimate his costs for next year's operations. One major cost is fuel purchases. To estimate fuel purchases, the owner needs to know the total distance his taxis will travel next year, the cost of a gallon of fuel, and the fuel mileage of his taxis. The owner has been provided with the first two figures (distance estimate and cost of a gallon of fuel). However, because of the high cost of gasoline, the owner has recently converted his taxis to operate on propane. He has measured and recorded the propane mileage (in miles per gallon) for 50 taxis.

a. What is the population of interest?
b. What is the parameter the owner needs?
c. What is the sample?
d. What is the statistic?
e. Describe briefly how the statistic will produce the kind of information the owner wants.

# APPENDIX 1 / INSTRUCTIONS FOR KELLER'S WEBSITE

The Keller website that accompanies this book contains the following features:

Data Analysis Plus 9.0 in VBA, which works with new and earlier versions of Excel (Office 1997, 2000, XP, 2003, 2007, and 2010 Office for Mac 2004)

A help file for Data Analysis Plus 9.0 in VBA

Data files in the following formats: ASCII, Excel, JMP, Minitab, SAS, and SPSS

Excel workbooks

Seeing Statistics (Java applets that teach a number of important statistical concepts)

Appendices (40 additional topics that are not covered in the book)

Formula card listing every formula in the book

## Keller website Instructions

"Data Analysis Plus 9.0 in VBA" can be found on the Keller website. It will be installed into the XLSTART folder of the most recent version of Excel on your computer. If properly installed Data Analysis Plus will be a menu item in Excel. The help file for Data Analysis Plus will be stored directly in your computer's My Documents folder. It will appear when you click the Help button or when you make a mistake when using Data Analysis Plus.

The Data Sets will also be installed from a link within the Keller website.

The Excel workbooks, Seeing Statistics Applets, and Appendixes will be accessed from the Keller website. Alternatively, you can store the Excel workbooks and Appendixes to your hard drive.

The Keller website is available using the student access code accompanying all new books. For more information on how to access the Keller website, please visit www.cengage.com/bstatistics/keller.

For technical support, please visit www.cengage.com/support for contact options. Refer to Statistics for Management and Economics, Ninth edition, by Gerald Keller (ISBN 0-538-47749-0).

# 2

© Steve Cole/Digital Vision/Getty Images

# GRAPHICAL DESCRIPTIVE TECHNIQUES I

## Do Male and Female American Voters Differ in Their Party Affiliation?

DATA
ANES2008*

In Chapter 1, we introduced the American National Election Survey (ANES), which is conducted every 4 years with the objective of developing information about how Americans vote. One question in the 2008 survey was "Do you think of yourself as Democrat, Republican, Independent, or what?"

Responses were

1. Democrat
2. Republican
3. Independent

© AP Photo/David Smith

On page 37 we will provide our answer.

4. Other party
5. No preference

Respondents were also identified by gender: 1 = male, and 2 = female. The responses are stored in file ANES2008* on our Keller's website. The asterisk indicates that there are variables that are not needed for this example but which will be used later in this book. For Excel users, GENDER AND PARTY are in columns B and BD, respectively. For Minitab users, GENDER AND PARTY are in columns 2 and 56, respectively. Some of the data are listed here.

| ID | GENDER | PARTY |
|---|---|---|
| 1 | 1 | 3 |
| 2 | 2 | 1 |
| 3 | 2 | 2 |
| . | . | . |
| . | . | . |
| 1795 | 1 | 1 |
| 1796 | 1 | 2 |
| 1797 | 1 | 1 |

Determine whether American female and male voters differ in their political affiliations.

## Introduction

In Chapter 1, we pointed out that statistics is divided into two basic areas: descriptive statistics and inferential statistics. The purpose of this chapter, together with the next, is to present the principal methods that fall under the heading of descriptive statistics. In this chapter, we introduce graphical and tabular statistical methods that allow managers to summarize data visually to produce useful information that is often used in decision making. Another class of descriptive techniques, numerical methods, is introduced in Chapter 4.

Managers frequently have access to large masses of potentially useful data. But before the data can be used to support a decision, they must be organized and summarized. Consider, for example, the problems faced by managers who have access to the databases created by the use of debit cards. The database consists of the personal information supplied by the customer when he or she applied for the debit card. This information includes age, gender, residence, and the cardholder's income. In addition, each time the card is used the database grows to include a history of the timing, price, and brand of each product purchased. Using the appropriate statistical technique, managers can determine which segments of the market are buying their company's brands. Specialized marketing campaigns, including telemarketing, can be developed. Both descriptive and inferential statistics would likely be employed in the analysis.

**Descriptive statistics** involves arranging, summarizing, and presenting a set of data in such a way that useful information is produced. Its methods make use of graphical techniques and numerical descriptive measures (such as averages) to summarize and present the data, allowing managers to make decisions based on the information generated. Although descriptive statistical methods are quite straightforward, their importance should not be underestimated. Most management, business, and economics students will encounter numerous opportunities to make valuable use of graphical and numerical descriptive techniques when preparing reports and presentations in the workplace. According to a Wharton Business School study, top managers reach a consensus 25% more quickly when responding to a presentation in which graphics are used.

In Chapter 1, we introduced the distinction between a population and a sample. Recall that a **population** is the entire set of observations under study, whereas a **sample** is a subset of a population. The descriptive methods presented in this chapter and in Chapters 3 and 4 apply to both a set of data constituting a population and a set of data constituting a sample.

In both the preface and Chapter 1, we pointed out that a critical part of your education as statistics practitioners includes an understanding of not only *how* to draw graphs and calculate statistics (manually or by computer) but also *when* to use each technique that we cover. The two most important factors that determine the appropriate method to use are (1) the type of data and (2) the information that is needed. Both are discussed next.

## 2.1 Types of Data and Information

The objective of statistics is to extract information from data. There are different types of data and information. To help explain this important principle, we need to define some terms.

A **variable** is some characteristic of a population or sample. For example, the mark on a statistics exam is a characteristic of statistics exams that is certainly of interest to readers of this book. Not all students achieve the same mark. The marks will vary from student to student, thus the name *variable*. The price of a stock is another variable. The prices of most stocks vary daily. We usually represent the name of the variable using uppercase letters such as $X$, $Y$, and $Z$.

The **values** of the variable are the possible observations of the variable. The values of statistics exam marks are the integers between 0 and 100 (assuming the exam is marked out of 100). The values of a stock price are real numbers that are usually measured in dollars and cents (sometimes in fractions of a cent). The values range from 0 to hundreds of dollars.

**Data*** are the observed values of a variable. For example, suppose that we observe the following midterm test marks of 10 students:

67 74 71 83 93 55 48 82 68 62

These are the data from which we will extract the information we seek. Incidentally, *data* is plural for **datum**. The mark of one student is a datum.

When most people think of data, they think of sets of numbers. However, there are three types of data: interval, nominal, and ordinal.†

*Unfortunately, the term *data*, like the term *statistician*, has taken on several different meanings. For example, dictionaries define data as facts, information, or statistics. In the language of computers, data may refer to any piece of information such as this textbook or an essay you have written. Such definitions make it difficult for us to present *statistics* as a method of converting *data* into *information*. In this book, we carefully distinguish among the three terms.

†There are actually four types of data, the fourth being *ratio* data. However, for statistical purposes there is no difference between ratio and interval data. Consequently, we combine the two types.

**Interval** data are real numbers, such as heights, weights, incomes, and distances. We also refer to this type of data as **quantitative** or **numerical**.

The values of **nominal** data are categories. For example, responses to questions about marital status produce nominal data. The values of this variable are single, married, divorced, and widowed. Notice that the values are not numbers but instead are words that describe the categories. We often record nominal data by arbitrarily assigning a number to each category. For example, we could record marital status using the following codes:

single = 1, married = 2, divorced = 3, widowed = 4

However, any other numbering system is valid provided that each category has a different number assigned to it. Here is another coding system that is just as valid as the previous one.

Single = 7, married = 4, divorced = 13, widowed = 1

Nominal data are also called **qualitative** or **categorical**.

The third type of data is ordinal. **Ordinal** data appear to be nominal, but the difference is that the order of their values has meaning. For example, at the completion of most college and university courses, students are asked to evaluate the course. The variables are the ratings of various aspects of the course, including the professor. Suppose that in a particular college the values are

poor, fair, good, very good, and excellent

The difference between nominal and ordinal types of data is that the order of the values of the latter indicate a higher rating. Consequently, when assigning codes to the values, we should maintain the order of the values. For example, we can record the students' evaluations as

Poor = 1, Fair = 2, Good = 3, Very good = 4, Excellent = 5

Because the only constraint that we impose on our choice of codes is that the order must be maintained, we can use any set of codes that are in order. For example, we can also assign the following codes:

Poor = 6, Fair = 18, Good = 23, Very good = 45, Excellent = 88

As we discuss in Chapter 19, which introduces statistical inference techniques for ordinal data, the use of any code that preserves the order of the data will produce exactly the same result. Thus, it's not the magnitude of the values that is important, it's their order.

Students often have difficulty distinguishing between ordinal and interval data. The critical difference between them is that the intervals or differences between values of interval data are consistent and meaningful (which is why this type of data is called *interval*). For example, the difference between marks of 85 and 80 is the same five-mark difference that exists between 75 and 70—that is, we can calculate the difference and interpret the results.

Because the codes representing ordinal data are arbitrarily assigned except for the order, we cannot calculate and interpret differences. For example, using a 1-2-3-4-5 coding system to represent poor, fair, good, very good, and excellent, we note that the difference between excellent and very good is identical to the difference between good and fair. With a 6-18-23-45-88 coding, the difference between excellent and very good is 43, and the difference between good and fair is 5. Because both coding systems are valid, we cannot use either system to compute and interpret differences.

Here is another example. Suppose that you are given the following list of the most active stocks traded on the NASDAQ in descending order of magnitude:

| Order | Most Active Stocks |
|---|---|
| 1 | Microsoft |
| 2 | Cisco Systems |
| 3 | Dell Computer |
| 4 | Sun Microsystems |
| 5 | JDS Uniphase |

Does this information allow you to conclude that the difference between the number of stocks traded in Microsoft and Cisco Systems is the same as the difference in the number of stocks traded between Dell Computer and Sun Microsystems? The answer is "no" because we have information only about the order of the numbers of trades, which are ordinal, and not the numbers of trades themselves, which are interval. In other words, the difference between 1 and 2 is not necessarily the same as the difference between 3 and 4.

## Calculations for Types of Data

### Interval Data

All calculations are permitted on interval data. We often describe a set of interval data by calculating the average. For example, the average of the 10 marks listed on page 13 is 70.3. As you will discover, there are several other important statistics that we will introduce.

### Nominal Data

Because the codes of nominal data are completely arbitrary, we cannot perform any calculations on these codes. To understand why, consider a survey that asks people to report their marital status. Suppose that the first 10 people surveyed gave the following responses:

single, married, married, married, widowed, single, married, married, single, divorced

Using the codes

Single = 1, married = 2, divorced = 3, widowed = 4

we would record these responses as

1 2 2 2 4 1 2 2 1 3

The average of these numerical codes is 2.0. Does this mean that the average person is married? Now suppose four more persons were interviewed, of whom three are widowed and one is divorced. The data are given here:

1 2 2 2 4 1 2 2 1 3 4 4 4 3

The average of these 14 codes is 2.5. Does this mean that the average person is married—but halfway to getting divorced? The answer to both questions is an emphatic "no." This example illustrates a fundamental truth about nominal data: Calculations based on the codes used to store this type of data are meaningless. All that we are permitted to do with nominal data is count or compute the percentages of the occurrences of each category. Thus, we would describe the 14 observations by counting the number of each marital status category and reporting the frequency as shown in the following table.

| Category | Code | Frequency |
|---|---|---|
| Single | 1 | 3 |
| Married | 2 | 5 |
| Divorced | 3 | 2 |
| Widowed | 4 | 4 |

The remainder of this chapter deals with nominal data only. In Chapter 3, we introduce graphical techniques that are used to describe interval data.

### Ordinal Data

The most important aspect of ordinal data is the order of the values. As a result, the only permissible calculations are those involving a ranking process. For example, we can place all the data in order and select the code that lies in the middle. As we discuss in Chapter 4, this descriptive measurement is called the *median*.

## Hierarchy of Data

The data types can be placed in order of the permissible calculations. At the top of the list, we place the interval data type because virtually *all* computations are allowed. The nominal data type is at the bottom because *no* calculations other than determining frequencies are permitted. (We are permitted to perform calculations using the frequencies of codes, but this differs from performing calculations on the codes themselves.) In between interval and nominal data lies the ordinal data type. Permissible calculations are ones that rank the data.

Higher-level data types may be treated as lower-level ones. For example, in universities and colleges, we convert the marks in a course, which are interval, to letter grades, which are ordinal. Some graduate courses feature only a pass or fail designation. In this case, the interval data are converted to nominal. It is important to point out that when we convert higher-level data as lower-level we lose information. For example, a mark of 83 on an accounting course exam gives far more information about the performance of that student than does a letter grade of A, which might be the letter grade for marks between 80 and 90. As a result, we do not convert data unless it is necessary to do so. We will discuss this later.

It is also important to note that we cannot treat lower-level data types as higher-level types.

The definitions and hierarchy are summarized in the following box.

**Types of Data**

Interval

Values are real numbers.

All calculations are valid.

Data may be treated as ordinal or nominal.

Ordinal

Values must represent the ranked order of the data.

Calculations based on an ordering process are valid.

Data may be treated as nominal but not as interval.

Nominal

Values are the arbitrary numbers that represent categories.

Only calculations based on the frequencies or percentages of occurrence are valid.

Data may not be treated as ordinal or interval.

## Interval, Ordinal, and Nominal Variables

The variables whose observations constitute our data will be given the same name as the type of data. Thus, for example, interval data are the observations of an interval variable.

## Problem Objectives and Information

In presenting the different types of data, we introduced a critical factor in deciding which statistical procedure to use. A second factor is the type of information we need to produce from our data. We discuss the different types of information in greater detail in Section 11.4 when we introduce *problem objectives*. However, in this part of the book (Chapters 2–5), we will use statistical techniques to describe a set of data, compare two or more sets of data, and describe the relationship between two variables. In Section 2.2, we introduce graphical and tabular techniques employed to describe a set of nominal data. Section 2.3 shows how to describe the relationship between two nominal variables and compare two or more sets of nominal data.

## EXERCISES

**2.1** Provide two examples each of nominal, ordinal, and interval data.

**2.2** For each of the following examples of data, determine the type.

a. The number of miles joggers run per week
b. The starting salaries of graduates of MBA programs
c. The months in which a firm's employees choose to take their vacations
d. The final letter grades received by students in a statistics course

**2.3** For each of the following examples of data, determine the type.

a. The weekly closing price of the stock of Amazon.com
b. The month of highest vacancy rate at a La Quinta motel
c. The size of soft drink (small, medium, or large) ordered by a sample of McDonald's customers
d. The number of Toyotas imported monthly by the United States over the last 5 years
e. The marks achieved by the students in a statistics course final exam marked out of 100

**2.4** The placement office at a university regularly surveys the graduates 1 year after graduation and asks for the following information. For each, determine the type of data.

a. What is your occupation?
b. What is your income?
c. What degree did you obtain?
d. What is the amount of your student loan?
e. How would you rate the quality of instruction? (excellent, very good, good, fair, poor)

**2.5** Residents of condominiums were recently surveyed and asked a series of questions. Identify the type of data for each question.

a. What is your age?
b. On what floor is your condominium?

c. Do you own or rent?
d. How large is your condominium (in square feet)?
e. Does your condominium have a pool?

**2.6** A sample of shoppers at a mall was asked the following questions. Identify the type of data each question would produce.
a. What is your age?
b. How much did you spend?
c. What is your marital status?
d. Rate the availability of parking: excellent, good, fair, or poor
e. How many stores did you enter?

**2.7** Information about a magazine's readers is of interest to both the publisher and the magazine's advertisers. A survey of readers asked respondents to complete the following:
a. Age
b. Gender
c. Marital status
d. Number of magazine subscriptions
e. Annual income
f. Rate the quality of our magazine: excellent, good, fair, or poor

For each item identify the resulting data type.

**2.8** Baseball fans are regularly asked to offer their opinions about various aspects of the sport. A survey asked the following questions. Identify the type of data.
a. How many games do you attend annually?
b. How would you rate the quality of entertainment? (excellent, very good, good, fair, poor)
c. Do you have season tickets?
d. How would you rate the quality of the food? (edible, barely edible, horrible)

**2.9** A survey of golfers asked the following questions. Identify the type of data each question produces.
a. How many rounds of golf do you play annually?
b. Are you a member of a private club?
c. What brand of clubs do you own?

**2.10** At the end of the term, university and college students often complete questionnaires about their courses. Suppose that in one university, students were asked the following.
a. Rate the course (highly relevant, relevant, irrelevant)
b. Rate the professor (very effective, effective, not too effective, not at all effective)
c. What was your midterm grade (A, B, C, D, F)?

Determine the type of data each question produces.

## 2.2 Describing a Set of Nominal Data

As we discussed in Section 2.1, the only allowable calculation on nominal data is to count the frequency or compute the percentage that each value of the variable represents. We can summarize the data in a table, which presents the categories and their counts, called a **frequency distribution** A **relative frequency distribution** lists the categories and the proportion with which each occurs. We can use graphical techniques to present a picture of the data. There are two graphical methods we can use: the **bar chart** and the **pie chart**.

### Example 2.1 Work Status in the GSS 2008 Survey

DATA
GSS2008*

In Chapter 1, we briefly introduced the General Social Survey. In the 2008 survey respondents were asked the following.

"Last week were you working full time, part time, going to school, keeping house, or what"? The responses were

1. Working full-time
2. Working part-time
3. Temporarily not working
4. Unemployed, laid off
5. Retired

6. School
7. Keeping house
8. Other

The responses were recorded using the codes 1, 2, 3, 4, 5, 6, 7, and 8, respectively. The first 150 observations are listed here. The name of the variable is WRKSTAT, and the data are stored in the 16th column (column P in Excel, column 16 in Minitab).

Construct a frequency and relative frequency distribution for these data and graphically summarize the data by producing a bar chart and a pie chart.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 7 | 7 | 1 | 1 | 5 | 1 | 5 | 7 | 1 | 1 |
| 5 | 7 | 1 | 5 | 2 | 5 | 1 | 5 | 8 | 1 | 5 | 7 | 1 | 4 | 2 |
| 7 | 1 | 2 | 1 | 1 | 2 | 1 | 7 | 1 | 7 | 1 | 2 | 1 | 1 | 1 |
| 1 | 1 | 6 | 5 | 1 | 1 | 1 | 1 | 1 | 2 | 5 | 2 | 7 | 2 | 7 |
| 8 | 1 | 8 | 1 | 7 | 1 | 6 | 7 | 6 | 1 | 5 | 1 | 2 | 2 | 4 |
| 1 | 1 | 1 | 1 | 1 | 6 | 5 | 5 | 3 | 2 | 1 | 1 | 8 | 1 | 5 |
| 1 | 1 | 1 | 1 | 5 | 5 | 1 | 5 | 4 | 7 | 1 | 1 | 1 | 4 | 5 |
| 2 | 5 | 6 | 7 | 7 | 1 | 4 | 2 | 1 | 2 | 6 | 1 | 1 | 1 | 1 |
| 1 | 1 | 7 | 4 | 1 | 1 | 1 | 7 | 8 | 1 | 3 | 1 | 1 | 3 | 1 |
| 1 | 1 | 1 | 1 | 1 | 2 | 1 | 5 | 1 | 1 | 1 | 1 | 1 | 2 | 1 |

## SOLUTION

Scan the data. Have you learned anything about the responses of these 150 Americans? Unless you have special skills you have probably learned little about the numbers. If we had listed all 2,023 observations you would be even less likely to discover anything useful about the data. To extract useful information requires the application of a statistical or graphical technique. To choose the appropriate technique we must first identify the type of data. In this example the data are nominal because the numbers represent categories. The only calculation permitted on nominal data is to count the number of occurrences of each category. Hence, we count the number of 1s, 2s, 3s, 4s, 5s, 6s, 7s, and 8s. The list of the categories and their counts constitute the frequency distribution. The relative frequency distribution is produced by converting the frequencies into proportions. The frequency and relative frequency distributions are combined in Table 2.1.

TABLE **2.1** Frequency and Relative Frequency Distributions for Example 2.1

| WORK STATUS | CODE | FREQUENCY | RELATIVE FREQUENCY (%) |
|---|---|---|---|
| Working full-time | 1 | 1003 | 49.6 |
| Working part-time | 2 | 211 | 10.4 |
| Temporarily not working | 3 | 53 | 2.6 |
| Unemployed, laid off | 4 | 74 | 3.7 |
| Retired | 5 | 336 | 16.6 |
| School | 6 | 57 | 2.8 |
| Keeping house | 7 | 227 | 11.2 |
| Other | 8 | 60 | 3.0 |
| Total | | 2021 | 100 |

There were two individuals who refused to answer hence the number of observations is the sample size 2,023 minus 2, which equals 2,021.

As we promised in Chapter 1 (and the preface), we demonstrate the solution of all examples in this book using three approaches (where feasible): manually, using Excel, and using Minitab. For Excel and Minitab, we provide not only the printout but also instructions to produce them.

## EXCEL

### *INSTRUCTIONS*

(Specific commands for this example are highlighted.)

1. Type or import the data into one or more columns. (Open GSS2008.)
2. Activate any empty cell and type

**=COUNTIF** ([Input range], [Criteria])

Input range are the cells containing the data. In this example, the range is P1:P2024. The criteria are the codes you want to count: (1) (2) (3) (4) (5) (6) (7) (8). To count the number of 1s (“Working full-time”), type

**=COUNTIF** (P1:P2024, 1)

and the frequency will appear in the dialog box. Change the criteria to produce the frequency of the other categories.

## MINITAB

```
WRKSTAT      Count     Percent
------------------------------
1             1003       49.63
2              211       10.44
3               53        2.62
4               74        3.66
5              336       16.63
6               57        2.82
7              227       11.23
8               60        2.97
------------------------------
            N= 2021
             *= 2
```

### *INSTRUCTIONS*

(Specific commands for this example are highlighted.)

1. Type or import the data into one column. (Open GSS2008.)
2. Click **Stat**, **Tables**, and **Tally Individual Variables**.
3. Type or use the **Select** button to specify the name of the variable or the column where the data are stored in the **Variables** box (WRKSTAT). Under **Display**, click **Counts** and **Percents.**

## INTERPRET

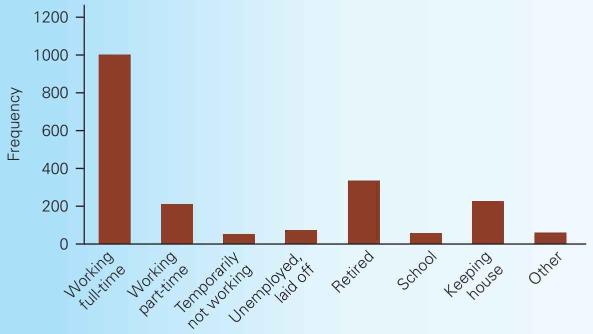
Almost 50% of respondents are working full-time, 16.6% are retired, 11.2% are keeping house, 10.4% are working part-time, and the remaining 12.1% are divided almost equally among the other four categories.

## Bar and Pie Charts

The information contained in the data is summarized well in the table. However, graphical techniques generally catch a reader's eye more quickly than does a table of numbers. Two graphical techniques can be used to display the results shown in the table. A **bar chart** is often used to display frequencies; a **pie chart** graphically shows relative frequencies.

The bar chart is created by drawing a rectangle representing each category. The height of the rectangle represents the frequency. The base is arbitrary. Figure 2.1 depicts the manually drawn bar chart for Example 2.1.

FIGURE **2.1** Bar Chart for Example 2.1



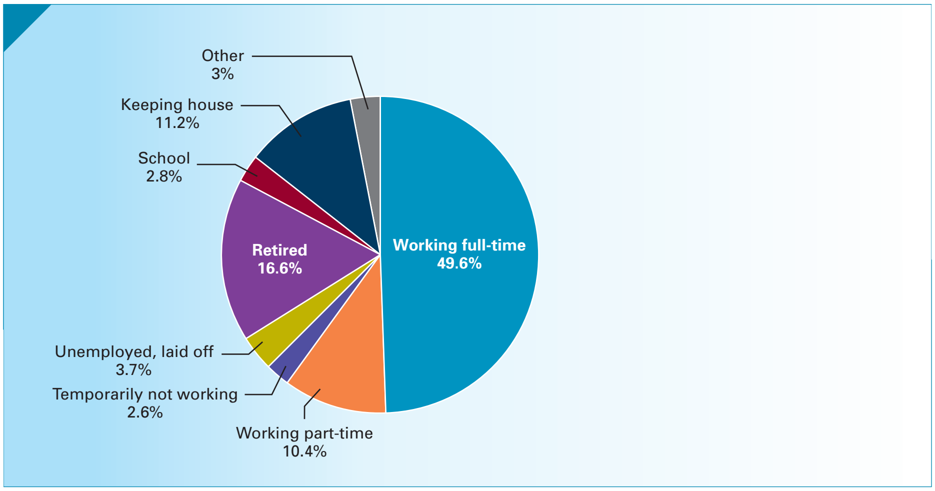
If we wish to emphasize the relative frequencies instead of drawing the bar chart, we draw a pie chart. A pie chart is simply a circle subdivided into slices that represent the categories. It is drawn so that the size of each slice is proportional to the percentage corresponding to that category. For example, because the entire circle is composed of 360 degrees, a category that contains 25% of the observations is represented by a slice of the pie that contains 25% of 360 degrees, which is equal to 90 degrees. The number of degrees for each category in Example 2.1 is shown in Table 2.2.

TABLE **2.2** Proportion in Each Category in Example 2.1

| WORK STATUS | RELATIVE FREQUENCY (%) | SLICE OF THE PIE (°) |
|---|---|---|
| Working full-time | 49.6 | 178.7 |
| Working part-time | 10.4 | 37.6 |
| Temporarily not working | 2.6 | 9.4 |
| Unemployed, laid off | 3.7 | 13.2 |
| Retired | 16.6 | 59.9 |
| School | 2.8 | 10.2 |
| Keeping house | 11.2 | 40.4 |
| Other | 3.0 | 10.7 |
| Total | 100.0 | 360 |

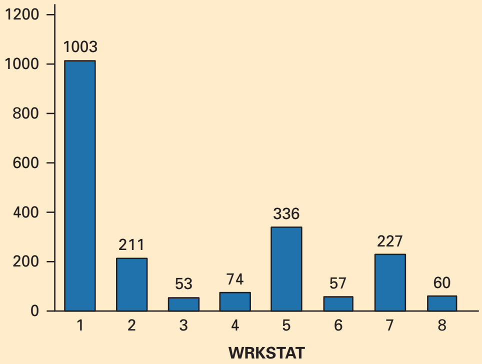
Figure 2.2 was drawn from these results.

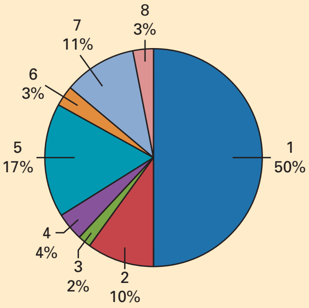
FIGURE **2.2** Pie Chart for Example 2.1



## EXCEL

Here are Excel's bar and pie charts.

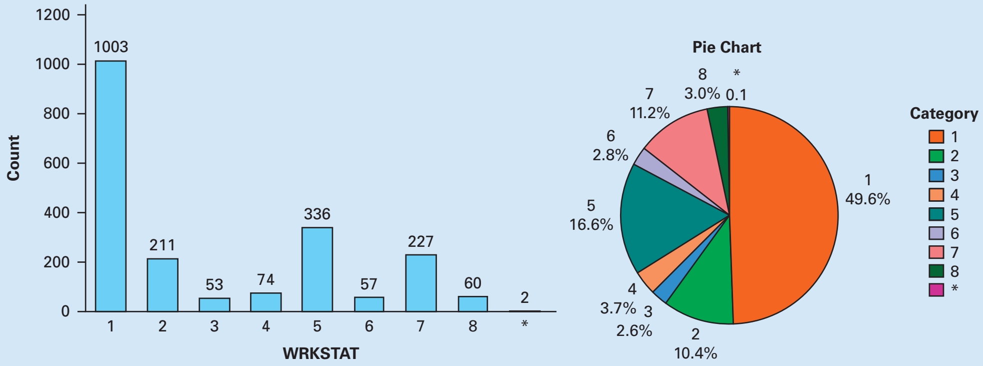




### INSTRUCTIONS

1. After creating the frequency distribution, highlight the column of frequencies.
2. For a bar chart, click **Insert, Column,** and the first **2-D Column.**
3. Click **Chart Tools** (if it does not appear, click inside the box containing the bar chart) and **Layout.** This will allow you to make changes to the chart. We removed the **Gridlines**, the **Legend**, and clicked the **Data Labels** to create the titles.
4. For a pie chart, click **Pie** and **Chart Tools** to edit the graph.

## MINITAB



### INSTRUCTIONS

1. Type or import the data into one column. (Open GSS2008.)

For a bar chart:

2. Click **Graph** and **Bar Chart**.
3. In the **Bars represent** box, click **Counts of unique values** and select **Simple**.
4. Type or use the **Select** button to specify the variable in the **Variables** box (WRKSTAT).

We clicked **Labels** and added the title and clicked **Data Labels** and **Use y-value labels** to display the frequencies at the top of the columns.

For a pie chart:

2. Click **Graph** and **Pie Chart**.
3. Click **Chart, Counts of unique values,** and in the **Categorical variables** box type or use the **Select** button to specify the variable (WRKSTAT).

We clicked **Labels** and added the title. We clicked **Slice Labels** and clicked **Category name** and **Percent.**

## INTERPRET

The bar chart focuses on the frequencies and the pie chart focuses on the proportions.

## Other Applications of Pie Charts and Bar Charts

Pie and bar charts are used widely in newspapers, magazines, and business and government reports. One reason for this appeal is that they are eye-catching and can attract the reader's interest whereas a table of numbers might not. Perhaps no one understands this better than the newspaper *USA Today*, which typically has a colored graph on the front page and others inside. Pie and bar charts are frequently used to simply present numbers associated with categories. The only reason to use a bar or pie chart in such a situation would be to enhance the reader's ability to grasp the substance of the data. It might, for example, allow the reader to more quickly recognize the relative sizes of the categories, as in the breakdown of a budget. Similarly, treasurers might use pie charts to show the breakdown of a firm's revenues by department, or university students might use pie charts to show the amount of time devoted to daily activities (e.g., eat 10%, sleep 30%, and study statistics 60%).

## APPLICATIONS in ECONOMICS

### Macroeconomics

Macroeconomics is a major branch of economics that deals with the behavior of the economy as a whole. Macroeconomists develop mathematical models that predict variables such as gross domestic product, unemployment rates, and inflation. These are used by governments and corporations to help develop strategies. For example, central banks attempt to control inflation by lowering or raising interest rates. To do this requires that economists determine the effect of a variety of variables, including the supply and demand for energy.

## APPLICATIONS in ECONOMICS

© Aaron Kohr/Shutterstock

### Energy Economics

One variable that has had a large influence on the economies of virtually every country is energy. The 1973 oil crisis in which the price of oil quadrupled over a short period of time is generally considered to be one of the largest financial shocks to our economy. In fact, economists often refer to two different economies: before the 1973 oil crisis and after.

Unfortunately, the world will be facing more shocks to our economy because of energy for two primary reasons. The first is the depletion of nonrenewable sources of energy and the resulting price increases. The second is the possibility that burning fossil fuels and the creation of carbon dioxide may be the cause of global warming. One economist predicted that the cost of global warming will be calculated in the trillions of dollars. Statistics can play an important role by determining whether Earth's temperature has been increasing and, if so, whether carbon dioxide is the cause. (See Case 3.1.)

In this chapter, you will encounter other examples and exercises that involve the issue of energy.

EXAMPLE 2.2

DATA
Xm02-02

## Energy Consumption in the United States in 2007

Table 2.3 lists the total energy consumption of the United States from all sources in 2007 (latest data available at publication). To make it easier to see the details, the table measures the energy in quadrillions of British thermal units (BTUs). Use an appropriate graphical technique to depict these figures.

TABLE **2.3** Energy Consumption in the United States by Source, 2007

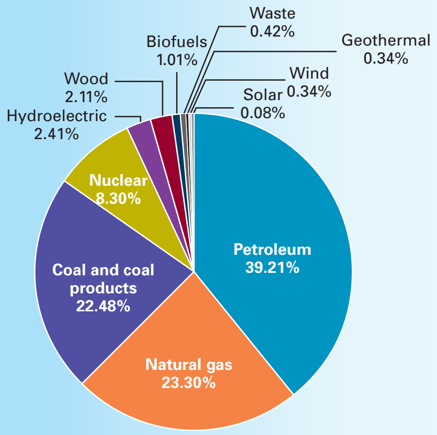
| ENERGY SOURCES | QUADRILLIONS OF BTUS |
|---|---|
| Nonrenewable | |
| Petroleum | 39.773 |
| Natural Gas | 23.637 |
| Coal and coal products | 22.801 |
| Nuclear | 8.415 |
| Renewable Energy Sources | |
| Hydroelectric | 2.446 |
| Wood derived fuels | 2.142 |
| Biofuels | 1.024 |
| Waste | 0.430 |
| Geothermal | 0.349 |
| Wind | 0.341 |
| Solar/photovoltaic | 0.081 |
| Total | 101.439 |

*Sources:* Non-renewable energy: Energy Information Administration (EIA), Monthly Energy Review (MER) December 2008, DOE/EIA-0035 (2008/12) (Washington, DC: December 2008) Tables 1.3, 1.4a, and 1.4b; Renewable Energy: Table 1.2 of this report.

## SOLUTION

We're interested in describing the proportion of total energy consumption for each source. Thus, the appropriate technique is the pie chart. The next step is to determine the proportions and sizes of the pie slices from which the pie chart is drawn. The following pie chart was created by Excel. Minitab's would be similar.

FIGURE **2.3** Pie Chart for Example 2.2



## INTERPRET

The United States depends heavily on petroleum, coal, and, natural gas. About 85% of national energy use is based on these sources. The renewable energy sources amount to less than 7%, of which about a third is hydroelectric and probably cannot be expanded much further. Wind and solar barely appear in the chart.

See Exercises 2.11 to 2.15 for more information on the subject.

## EXAMPLE 2.3 Per Capita Beer Consumption (10 Selected Countries)

DATA
Xm02-03

Table 2.4 lists the per capita beer consumption for each of 20 countries around the world. Graphically present these numbers.

TABLE **2.4** Per Capita Beer Consumption 2008

| COUNTRY | BEER CONSUMPTION(L/YR) |
|---|---|
| Australia | 119.2 |
| Austria | 106.3 |
| Belgium | 93.0 |
| Canada | 68.3 |
| Croatia | 81.2 |
| Czech Republic | 138.1 |
| Denmark | 89.9 |
| Finland | 85.0 |
| Germany | 147.8 |
| Hungary | 75.3 |
| Ireland | 138.3 |
| Luxembourg | 84.4 |
| Netherlands | 79.0 |
| New Zealand | 77.0 |
| Poland | 69.1 |
| Portugal | 59.6 |
| Slovakia | 84.1 |
| Spain | 83.8 |
| United Kingdom | 96.8 |
| United States | 81.6 |

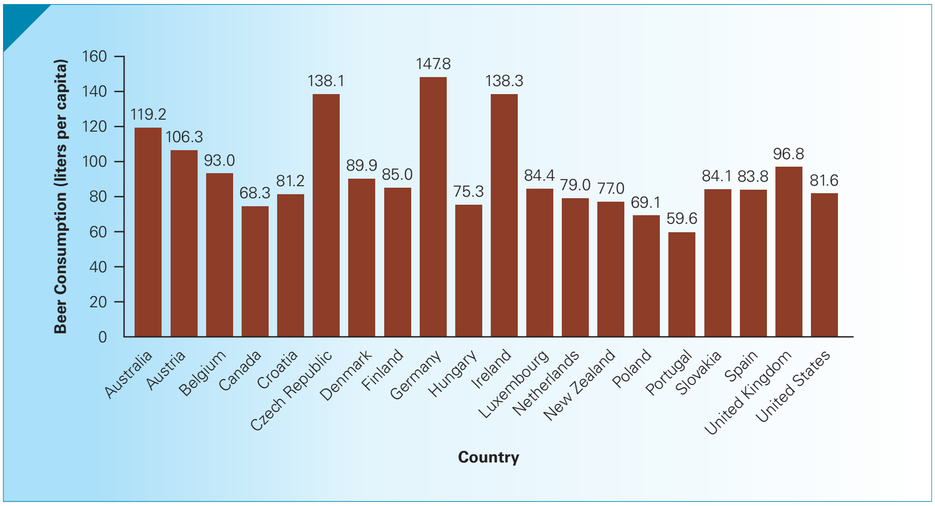
*Source:* www.beerinfo.com

## SOLUTION

In this example, we're primarily interested in the numbers. There is no use in presenting proportions here.

The following is Excel's bar chart.

FIGURE 2.4 EXCEL Bar Chart for Example 2.3



INTERPRET

Germany, the Czech Republic, Ireland, Australia, and Austria head the list. Both the United States and Canada rank far lower. Surprised?

## Describing Ordinal Data

There are no specific graphical techniques for ordinal data. Consequently, when we wish to describe a set of ordinal data, we will treat the data as if they were nominal and use the techniques described in this section. The only criterion is that the bars in bar charts should be arranged in ascending (or descending) ordinal values; in pie charts, the wedges are typically arranged clockwise in ascending or descending order.

We complete this section by describing when bar and pie charts are used to summarize and present data.

**Factors That Identify When to Use Frequency and Relative Frequency Tables, Bar and Pie Charts**

1. **Objective**: Describe a single set of data.
2. **Data type**: Nominal or ordinal

## EXERCISES

**2.11** Xr02-11 When will the world run out of oil? One way to judge is to determine the oil reserves of the countries around the world. The next table displays the known oil reserves of the top 15 countries. Graphically describe the figures.

| Country | Reserves |
|---|---|
| Brazil | 12,620,000,000 |
| Canada | 178,100,000,000 |
| China | 16,000,000,000 |

(Continued)

| Country | Reserves |
|---|---|
| Iran | 136,200,000,000 |
| Iraq | 115,000,000,000 |
| Kazakhstan | 30,000,000,000 |
| Kuwait | 104,000,000,000 |
| Libya | 43,660,000,000 |
| Nigeria | 36,220,000,000 |
| Qatar | 15,210,000,000 |
| Russia | 60,000,000,000 |
| Saudi Arabia | 266,700,000,000 |
| United Arab Emirates | 97,800,000,000 |
| United States | 21,320,000,000 |
| Venezuela | 99,380,000,000 |

*Source:* CIA World Factbook.

**2.12** Refer to Exercise 2.11. The total reserves in the world are 1,348,528,420,000 barrels. The total reserves of the top 15 countries are 1,232,210,000,000 barrels. Use a graphical technique that emphasizes the percentage breakdown of the top 15 countries plus others. Briefly describe your findings.

**2.13** Xr02-13 The following table lists the average oil consumption per day for the top 15 oil-consuming countries. Use a graphical technique to present these figures.

| Country | Consumption (barrels per day) |
|---|---|
| Brazil | 2,520,000 |
| Canada | 2,260,000 |
| China | 7,850,000 |
| France | 1,986,000 |
| Germany | 2,569,000 |
| India | 2,940,000 |
| Iran | 1,755,000 |
| Italy | 1,639,000 |
| Japan | 4,785,000 |
| Mexico | 2,128,000 |
| Russia | 2,900,000 |
| Saudi Arabia | 2,380,000 |
| South Korea | 2,175,000 |
| United Kingdom | 1,710,000 |
| United States | 19,500,000 |

*Source:* CIA World Factbook.

**2.14** Xr02-14 There are 42 gallons in a barrel of oil. The number of products produced and the proportion of the total are listed in the following table. Draw a graph to depict these numbers. What can you conclude from your graph?

| Product | Percent of Total (%) |
|---|---|
| Gasoline | 51.4 |
| Distillate fuel oil | 15.3 |
| Jet fuel | 12.6 |
| Still gas | 5.4 |
| Marketable coke | 5.0 |
| Residual fuel oil | 3.3 |
| Liquefied refinery gas | 2.8 |
| Asphalt and road oil | 1.9 |
| Lubricants | .9 |
| Other | 1.5 |

*Source*: California Energy Commission based on 2004 data.

**2.15** Xr02-15* The following table displays the energy consumption pattern of Australia. The figures measure the heat content in metric tons (1,000 kilograms) of oil equivalent. Draw a graph that depicts these numbers and explain what you have learned.

| Energy Sources | Heat Content |
|---|---|
| Nonrenewable | |
| Coal and coal products | 55,385 |
| Oil | 33,185 |
| Natural Gas | 20,350 |
| Nuclear | 0 |
| Renewable Energy Sources | |
| Hydroelectric | 1,388 |
| Solid Biomass | 4,741 |
| Other (Liquid biomass, geothermal, solar, wind, and tide, wave, and ocean) | 347 |
| Total | 115,396 |

*Source*: International Energy Association.

**2.16** Xr02-16 The planet may be threatened by global warming, which may be caused by the burning of fossil fuels (petroleum, natural gas, and coal) that produces carbon dioxide ($CO_2$). The following table lists the top 15 producers of $CO_2$ and the annual amounts (million of metric tons) from fossil fuels. Graphically depict these figures. Explain what you have learned.

| Country | $CO_2$ | Country | $CO_2$ |
|---|---|---|---|
| Australia | 406.6 | Japan | 1230.4 |
| Canada | 631.3 | Korea, South | 499.6 |
| China | 5322.7 | Russia | 1696.0 |
| France | 415.3 | Saudi Arabia | 412.4 |
| Germany | 844.2 | South Africa | 423.8 |
| India | 1165.7 | United Kingdom | 577.2 |
| Iran | 450.7 | United States | 5957.0 |
| Italy | 466.6 | | |

*Source: Statistical Abstract of the United States*, 2009, Table 1304.

**2.17** Xr02-17 The production of steel has often been used as a measure of the economic strength of a country. The following table lists the steel production in the 20 largest steel-producing nations in 2008. The units are millions of metric tons. Use a graphical technique to display these figures.

| Country | Steel production | Country | Steel production |
|---|---|---|---|
| Belgium | 10.7 | Mexico | 17.2 |
| Brazil | 33.7 | Poland | 9.7 |
| Canada | 14.8 | Russia | 68.5 |
| China | 500.5 | South Korea | 53.6 |
| France | 17.9 | Spain | 18.6 |
| Germany | 45.8 | Taiwan | 19.9 |
| India | 55.2 | Turkey | 26.8 |
| Iran | 10 | Ukraine | 37.1 |
| Italy | 30.6 | United Kingdom | 13.5 |
| Japan | 118.7 | United States | 91.4 |

*Source:* World Steel Association.

**2.18** Xr02-18 In 2003 (latest figures available) the United States generated 251.3 million tons of garbage. The following table lists the amounts by source. Use one or more graphical techniques to present these figures.

| Source | Amount (millions of tons) |
|---|---|
| Paper and paperboard | 85.2 |
| Glass | 13.3 |
| Metals | 19.1 |
| Plastics | 29.4 |
| Rubber and leather | 6.5 |
| Textiles | 11.8 |
| Wood | 13.8 |
| Food scraps | 31.2 |
| Yard trimmings | 32.4 |
| Other | 8.6 |

*Source: Statistical Abstract of the United States*, 2009, Table 361.

**2.19** Xr02-19 In the last five years, the city of Toronto has intensified its efforts to reduce the amount of garbage that is taken to landfill sites. [Currently, the Greater Toronto Area (GTA) disposes of its garbage in a dump site in Michigan.] A current analysis of GTA reveals that 36% of waste collected is taken from residences and 64% from businesses and public institutions (hospitals, schools, universities, etc.). A further breakdown is listed below. (*Source:* Toronto City Summit Alliance.)

a. Draw a pie chart for residential waste including both recycled and disposed waste.

b. Repeat part (a) for nonresidential waste.

**Residential**

| Recycled | Pct | Disposed | Pct |
|---|---|---|---|
| Recycled Plastic | 1% | Plastic | 7% |
| Recycled Glass | 3% | Paper | 12% |
| Recycled Paper | 14% | Metal | 2% |
| Recycled Metal | 1% | Organic | 23% |
| Recycled Organic/Food | 7% | Other | 17% |
| Recycled Organic/Yard | 10% | | |
| Recycled Other | 4% | | |

**Non-Residential**

| Recycled | Pct | Disposed | Pct |
|---|---|---|---|
| Recycled Glass | 1% | Plastic | 10% |
| Recycled Paper | 11% | Glass | 3% |
| Recycled Metal | 3% | Paper | 31% |
| Recycled Organic | 1% | Metal | 8% |
| Recycled Construction/Demolition | 1% | Organic | 18% |
| Recycled Other | 1% | Construction/Demolition | 7% |
| | | Other | 6% |

**2.20** Xr02-20 The following table lists the top 10 countries and amounts of oil (millions of barrels annually) they exported to the United States in 2007.

| Country | Oil Imports |
|---|---|
| Algeria | 162 |
| Angola | 181 |
| Canada | 681 |
| Iraq | 177 |
| Kuwait | 64 |
| Mexico | 514 |
| Nigeria | 395 |
| Saudi Arabia | 530 |
| United Kingdom | 37 |
| Venezuela | 420 |

*Source: Statistical Abstract of the United States*, 2009, Table 895.

a. Draw a bar chart.

b. Draw a pie chart.

c. What information is conveyed by each chart?

**2.21** Xr02-21 The following table lists the percentage of males and females in five age groups that did not have health insurance in the United States in September 2008. Use a graphical technique to present these figures.

| Age Group | Male | Female |
|---|---|---|
| Under 18 | 8.5 | 8.5 |
| 18–24 | 32.3 | 24.9 |
| 25–34 | 30.4 | 21.4 |
| 35–44 | 21.3 | 17.1 |
| 45–64 | 13.5 | 13.0 |

*Source:* National Health Interview Survey.

**2.22** Xr02-22 The following table lists the average costs for a family of four to attend a game at a National Football League (NFL) stadium compared to a Canadian Football League (CFL) stadium. Use a graphical technique that allows the reader to compare each component of the total cost.

| | NFL | CFL |
|---|---|---|
| Four tickets | 274.12 | 171.16 |
| Parking | 19.75 | 10.85 |
| Two ball caps | 31.12 | 44.26 |
| Two beers | 11.90 | 11.24 |
| Two drinks | 7.04 | 7.28 |
| Four hot dogs | 15.00 | 16.12 |

*Source:* Team Market Report, Matthew Coutts.

**2.23** Xr02-23 Productivity growth is critical to the economic well-being of companies and countries. In the table below we list the average annual growth rate (in percent) in productivity for the Organization for Economic Co-Operation and Development (OECD) countries. Use graphical technique to present these figures.

| Country | Productivity Growth | Country | Productivity Growth |
|---|---|---|---|
| Australia | 1.6 | Japan | 2.775 |
| Austria | 1.5 | Korea | 5.55 |
| Belgium | 1.975 | Luxembourg | 2.6 |
| Canada | 1.25 | Mexico | 1.2 |
| Czech Republic | 3.3 | Netherlands | 2 |
| | | New Zealand | 1.5 |
| Denmark | 2.175 | Norway | 2.575 |
| Finland | 2.775 | Portugal | 2.8 |
| France | 2.35 | Slovak Republic | 4.8 |
| Germany | 2.3 | Spain | 2 |
| Greece | 1.6 | Sweden | 1.775 |
| Hungary | 3.6 | Switzerland | 1.033 |
| Iceland | 0.775 | United Kingdom | 2.2 |
| Ireland | 3.775 | | |
| Italy | 1.525 | United States | 1.525 |

*Source:* OECD Labor Productivity Database July 2007.

*The following exercises require a computer and software.*

**2.24** Xr02-24 In an attempt to stimulate the economy in 2008, the U.S. government issued rebate checks totaling $107 billion. A survey conducted by the National Retail Federation (NRF) asked recipients what they intended to do with their rebates. The choices are:

1. Buy something
2. Pay down debt
3. Invest
4. Pay medical bills
5. Save
6. Other

Use a graphical technique to summarize and present these data. Briefly describe your findings.

**2.25** Xr02-25 Refer to Exercise 2.24. Those who responded that they planned to buy something were asked what they intended to buy. Here is a list of their responses.

1. Home improvement project
2. Purchase appliances
3. Purchase automobles
4. Purchase clothing
5. Purchase electronics
6. Purchase furniture
7. Purchase gas
8. Spa or salon time
9. Purchase vacation
10. Purchase groceries
11. Impulse purchase
12. Down payment on house

Graphically summarize these data. What can you conclude from the chart?

**2.26** Xr02-26 What are the most important characteristics of colleges and universities? This question was asked of a sample of college-bound high school seniors. The responses are:

1. Location
2. Majors
3. Academic reputation
4. Career focus
5. Community
6. Number of students

The results are stored using the codes. Use a graphical technique to summarize and present the data.

**2.27** Xr02-27 Where do consumers get information about cars? A sample of recent car buyers was asked to identify the most useful source of information about the cars they purchased. The responses are:

1. Consumer guide
2. Dealership
3. Word of mouth
4. Internet

The responses were stored using the codes. Graphically depict the responses. *Source: Automotive Retailing Today*, The Gallup Organization.

**2.28** Xr02-28 A survey asked 392 homeowners which area of their homes they would most like to renovate. The responses and frequencies are shown next. Use a graphical technique to present these results. Briefly summarize your findings.

| Area | Code |
|---|---|
| Basement | 1 |
| Bathroom | 2 |
| Bedroom | 3 |
| Kitchen | 4 |
| Living/dining room | 5 |

*Source: Toronto Star*, November 23, 2004.

**2.29** Xr02-29 Subway train riders frequently pass the time by reading a newspaper. New York City has a subway and four newspapers. A sample of 360 subway riders who regularly read a newspaper was asked to identify that newspaper. The responses are:

1. *New York Daily News*
2. *New York Post*
3. *New York Times*
4. *Wall Street Journal*

The responses were recorded using the numerical codes shown.

a. Produce a frequency distribution and a relative frequency distribution.
b. Draw an appropriate graph to summarize the data. What does the graph tell you?

**2.30** Xr02-30 Who applies to MBA programs? To help determine the background of the applicants, a sample of 230 applicants to a university's business school was asked to report their undergraduate degrees. The degrees were recorded using these codes.

1. BA
2. BBA
3 BEng
4. BSc
5. Other

a. Determine the frequency distribution.
b. Draw a bar chart.
c. Draw a pie chart.
d. What do the charts tell you about the sample of MBA applicants?

**2.31** Xr02-31 Many business and economics courses require the use of computer, so students often must buy their own computers. A survey asks students to identify which computer brands they have purchased. The responses are:

1. IBM
2. Compaq
3. Dell
4. Other

a. Use a graphical technique that depicts the frequencies.
b. Graphically depict the proportions.
c. What do the charts tell you about the brands of computers used by the students?

**2.32** Xr02-32 An increasing number of statistics courses use a computer and software rather than manual calculations. A survey of statistics instructors asked them to report the software their courses use. The responses are:

1. Excel
2. Minitab
3. SAS
4. SPSS
5. Other

a. Produce a frequency distribution.
b. Graphically summarize the data so that the proportions are depicted.
c. What do the charts tell you about the software choices?

**2.33** Xr02-33* The total light beer sales in the United States are approximately 3 million gallons annually. With this large of a market, breweries often need to know more about who is buying their product. The marketing manager of a major brewery wanted to analyze the light beer sales among college and university students who drink light beer. A random sample of 285 graduating students was asked to report which of the following is their favorite light beer:

1. Bud Light
2. Busch Light
3. Coors Light
4. Michelob Light
5. Miller Lite
6. Natural Light
7. Other brands

The responses were recorded using the codes 1, 2, 3, 4, 5, 6, and 7, respectively. Use a graphical to summarize these data. What can you conclude from the chart?

## General Social Survey Exercises

*The following exercises are based on the GSS described above.*

**2.34** GSS2008* In the 2008 General Social Survey, respondents were asked to identify their race (RACE) using the following categories:

1. White
2. Black
3. Other

Summarize the results using an appropriate graphical technique and interpret your findings.

**2.35** GSS2008* Several questions deal with education. One question in the 2008 survey asked respondents to indicate their highest degree (DEGREE). The responses are:

0. Left high school
1. Completed high school
2. Completed junior college
3. Completed bachelor's degree
4. Completed graduate degree

Use a graphical technique to summarize the data. Describe what the graph tells you.

**2.36** GSS2006* Refer to the GSS in 2006. Responses to the question about marital status (MARITAL) were:

1. Married
2. Widowed
3. Divorced
4. Separated
5. Never Married

a. Create a frequency distribution
b. Use a graphical method to present these data and briefly explain what the graph reveals.

**2.37** GSS2004* Refer to the 2004 GSS, which asked about individual's class (CLASS). The responses were:

1. Lower class
2. Working class
3. Middle class
4. Upper class

Summarize the data using a graphical method and describe your findings.

## 2.3 Describing the Relationship between Two Nominal Variables and Comparing Two or More Nominal Data Sets

In Section 2.2, we presented graphical and tabular techniques used to summarize a set of nominal data. Techniques applied to single sets of data are called **univariate**. There are many situations where we wish to depict the relationship between variables; in such cases, **bivariate** methods are required. A **cross-classification table** (also called a **cross-tabulation table**) is used to describe the relationship between two nominal variables. A variation of the bar chart introduced in Section 2.2 is employed to graphically describe the relationship. The same technique is used to compare two or more sets of nominal data.

### Tabular Method of Describing the Relationship between Two Nominal Variables

To describe the relationship between two nominal variables, we must remember that we are permitted only to determine the frequency of the values. As a first step, we need to produce a cross-classification table that lists the frequency of each combination of the values of the two variables.

### EXAMPLE 2.4 Newspaper Readership Survey

DATA
Xm02-04

A major North American city has four competing newspapers: the *Globe and Mail (G&M)*, *Post*, *Star and Sun*. To help design advertising campaigns, the advertising managers of the newspapers need to know which segments of the newspaper market are reading their papers. A survey was conducted to analyze the relationship between newspapers read and occupation. A sample of newspaper readers was asked to report which newspaper they read—*Globe and Mail* (1), *Post* (2), *Star* (3), *Sun* (4)—and indicate whether they were blue-collar workers (1), white-collar workers (2), or professionals (3). Some of the data are listed here.

| Reader | Occupation | Newspaper |
|---|---|---|
| 1 | 2 | 2 |
| 2 | 1 | 4 |
| 3 | 2 | 1 |
| . | . | . |
| . | . | . |
| 352 | 3 | 2 |
| 353 | 1 | 3 |
| 354 | 2 | 3 |

Determine whether the two nominal variables are related.

## SOLUTION

By counting the number of times each of the 12 combinations occurs, we produced the Table 2.5.

TABLE **2.5** Cross-Classification Table of Frequencies for Example 2.4

| | NEWSPAPER | | | | |
|---|---|---|---|---|---|
| OCCUPATION | *G&M* | *POST* | *STAR* | *SUN* | TOTAL |
| Blue collar | 27 | 18 | 38 | 37 | 120 |
| White collar | 29 | 43 | 21 | 15 | 108 |
| Professional | 33 | 51 | 22 | 20 | 126 |
| Total | 89 | 112 | 81 | 72 | 354 |

If occupation and newspaper are related, there will be differences in the newspapers read among the occupations. An easy way to see this is to convert the frequencies in each row (or column) to relative frequencies in each row (or column). That is, compute the row (or column) totals and divide each frequency by its row (or column) total, as shown in Table 2.6. Totals may not equal 1 because of rounding.

TABLE **2.6** Row Relative Frequencies for Example 2.4

| | NEWSPAPER | | | | |
|---|---|---|---|---|---|
| OCCUPATION | *G&M* | *POST* | *STAR* | *SUN* | TOTAL |
| Blue collar | .23 | .15 | .32 | .31 | 1.00 |
| White collar | .27 | .40 | .19 | .14 | 1.00 |
| Professional | .26 | .40 | .17 | .16 | 1.00 |
| Total | .25 | .32 | .23 | .20 | 1.00 |

## EXCEL

Excel can produce the cross-classification table using several methods. We will use and describe the PivotTable in two ways: (1) to create the cross-classification table featuring the counts and (2) to produce a table showing the row relative frequencies.

| **Count of Reader** | **Newspaper** | | | | |
|---|---|---|---|---|---|
| Occupation | G&M | Post | Star | Sun | Grand Total |
| Blue collar | 27 | 18 | 38 | 37 | 120 |
| White collar | 29 | 43 | 21 | 15 | 108 |
| Professional | 33 | 51 | 22 | 20 | 126 |
| Grand Total | 89 | 112 | 81 | 72 | 354 |

| **Count of Reader** | **Newspaper** | | | | |
|---|---|---|---|---|---|
| Occupation | G&M | Post | Star | Sun | Grand Total |
| Blue collar | 0.23 | 0.15 | 0.32 | 0.31 | 1.00 |
| White collar | 0.27 | 0.40 | 0.19 | 0.14 | 1.00 |
| Professional | 0.26 | 0.40 | 0.17 | 0.16 | 1.00 |
| Grand Total | 0.25 | 0.32 | 0.23 | 0.20 | 1.00 |

### INSTRUCTIONS

The data must be stored in (at least) three columns as we have done in Xm02-04. Put the cursor somewhere in the data range.

1. Click **Insert** and **PivotTable**.

2. Make sure that the Table/Range is correct.

3. Drag the Occupation button to the **ROW** section of the box. Drag the Newspaper button to the **COLUMN** section. Drag the Reader button to the **DATA** field. Right-click any number in the table, click **Summarize Data By**, and check **Count.** To convert to row percentages, right-click any number, click **Summarize Data By, More options . . .** , and **Show values as**. Scroll down and click **% of rows**. **(**We then formatted the data into decimals.) To improve both tables, we substituted the names of the occupations and newspapers.

## MINITAB

Tabulated statistics: Occupation, Newspaper

```
Rows: Occupation                    Columns: Newspaper
            1          2          3          4        All
1           27         18         38         37       120
            22.50      15.00      31.67      30.83    100.00
2           29         43         21         15       108
            26.85      39.81      19.44      13.89    100.00
3           33         51         22         20       126
            26.19      40.48      17.46      15.87    100.00
All         89         112        81         72       354
            25.14      31.64      22.88      20.34    100.00
Cell Contents:   Count
                 % of Row
```

*INSTRUCTIONS*

1. Type or import the data into two columns. (Open xM02-04)
2. Click **Stat**, **Tables**, and **Cross Tabulation and Chi-square**.
3. Type or use the **Select** button to specify the **Categorical variables**: **For rows** (Occupation) and **For columns** (Newspaper)
4. Under **Display**, click **Counts** and **Row percents** (or any you wish)

## INTERPRET

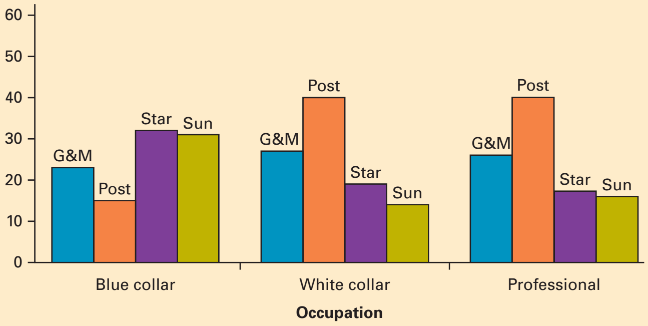
Notice that the relative frequencies in the second and third rows are similar and that there are large differences between row 1 and rows 2 and 3. This tells us that blue-collar workers tend to read different newspapers from both white-collar workers and professionals and that white-collar workers and professionals are quite similar in their newspaper choices.

## Graphing The Relationship between Two Nominal Variables

We have chosen to draw three bar charts, one for each occupation depicting the four newspapers. We'll use Excel and Minitab for this purpose. The manually drawn charts are identical.

## EXCEL

There are several ways to graphically display the relationship between two nominal variables. We have chosen two dimensional bar charts for each of the three occupations. The charts can be created from the output of the PivotTable (either counts as we have done) or row proportions.

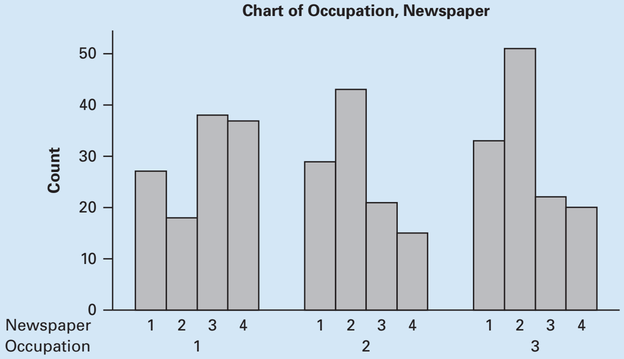


*INSTRUCTIONS*

From the cross-classification table, click **Insert** and **Column.** You can do the same from any completed cross-classification table.

## MINITAB

Minitab can draw bar charts from the raw data.



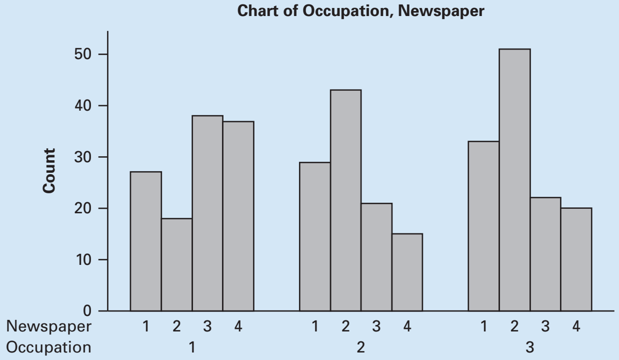
*INSTRUCTIONS*

1. Click **Graph** and **Bar Chart**.
2. In the **Bars represent** box, specify **Counts of unique values**. Select **Cluster**.
3. In the **Categorical variables** box, type or select the two variables (Newspaper Occupation).

If you or someone else has created the cross-classification table, Minitab can draw bar charts directly from the table.

*INSTRUCTIONS*

1. Start with a completed cross-classification table such as Table 2.9.
2. Click **Graph** and Bar **Chart**
3. In the **Bars represent** box click **Values from a table.** Choose **Two-way table Cluster**.
4. In the **Graph variables** box, **Select** the columns of numbers in the table. In the **Row labels** box, **Select** the column with the categories.



## INTERPRET

If the two variables are unrelated, then the patterns exhibited in the bar charts should be approximately the same. If some relationship exists, then some bar charts will differ from others.

The graphs tell us the same story as did the table. The shapes of the bar charts for occupations 2 and 3 (white-collar and professional) are very similar. Both differ considerably from the bar chart for occupation 1 (blue-collar).

## Comparing Two or More Sets of Nominal Data

We can interpret the results of the cross-classification table of the bar charts in a different way. In Example 2.4, we can consider the three occupations as defining three different populations. If differences exist between the columns of the frequency distributions (or between the bar charts), then we can conclude that differences exist among the three populations. Alternatively, we can consider the readership of the four newspapers as four different populations. If differences exist among the frequencies or the bar charts, then we conclude that there are differences between the four populations.

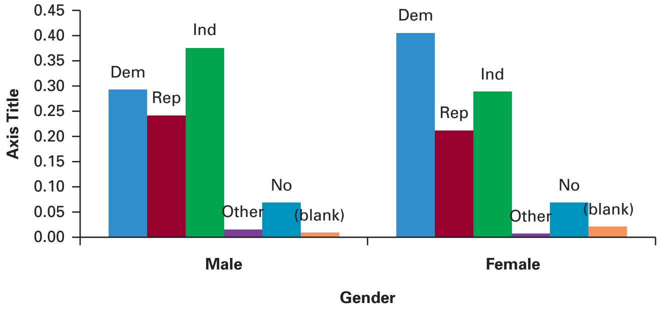
## Do Male and Female American Voters Differ in Their Party Affiliation?

DATA
ANES2008*

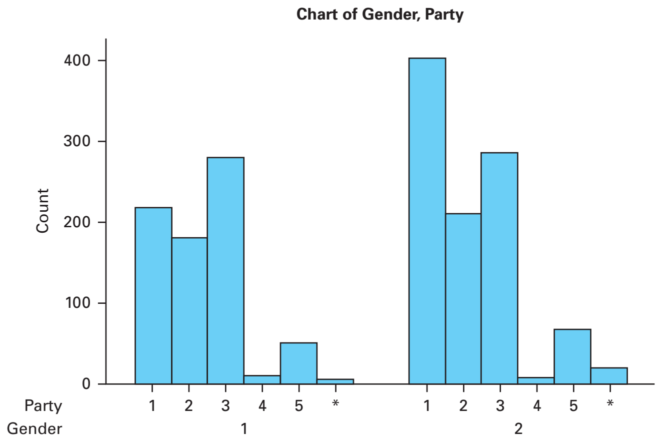
Using the technique introduced above, we produced the bar charts below.

©AP Photo/David Smith

### EXCEL



## MINITAB



## INTERPRET

As you can see, there are substantial differences between the bar charts for men and women. We can conclude that gender and party affiliation are related. However, we can also conclude that differences in party affiliation exist between American male and female voters: Specifically, men tend to identify themselves as independents, whereas women support the Democratic party.

Historically, women tend to be Democrats, and men lean toward the Republican party. However, in this survey, both genders support the Democrats over the Republicans, which explains the results of the 2008 election.

## Data Formats

There are several ways to store the data to be used in this section to produce a table or a bar or pie chart.

1. The data are in two columns. The first column represents the categories of the first nominal variable, and the second column stores the categories for the second variable. Each row represents one observation of the two variables. The number of observations in each column must be the same. Excel and Minitab can produce a cross-classification table from these data. (To use Excel's PivotTable, there also must be a third variable representing the observation number.) This is the way the data for Example 2.4 were stored.

2. The data are stored in two or more columns, with each column representing the same variable in a different sample or population. For example, the variable may be the type of undergraduate degree of applicants to an MBA program, and there may be five universities we wish to compare. To produce a cross-classification table, we would have to count the number of observations of each category (undergraduate degree) in each column.

3. The table representing counts in a cross-classification table may have already been created.

We complete this section with the factors that identify the use of the techniques introduced here.

**Factors That Identify When to Use a Cross-Classification Table**

1. **Objective**: Describe the relationship between two variables and compare two or more sets of data.
2. **Data type**: Nominal

## EXERCISES

2.38 Xr02-38 Has the educational level of adults changed over the past 15 years? To help answer this question, the Bureau of Labor Statistics compiled the following table; it lists the number (1,000) of adults 25 years of age and older who are employed. Use a graphical technique to present these figures. Briefly describe what the chart tells you.

| Educational level | 1995 | 1999 | 2003 | 2007 |
|---|---|---|---|---|
| Less than high school | 12,021 | 12,110 | 12,646 | 12,408 |
| High school | 36,746 | 35,335 | 33,792 | 32,634 |
| Some college | 30,908 | 30,401 | 30,338 | 30,389 |
| College graduate | 31,176 | 33,651 | 35,454 | 37,321 |

*Source: Statistical Abstract of the United States*, 2009, Table 572.

2.39 Xr02-39 How do governments spend the tax dollars they collect, and has this changed over the past 15 years? The following table displays the amounts spent by the federal, state, and local governments on consumption expenditures and gross investments. Consumption expenditures are services (such as education). Gross investments ($billions) consist of expenditures on fixed assets (such as roads, bridges, and highways). Use a graphical technique to present these figures. Have the ways governments spend money changed over the previous 15 years?

| Level of Government and Type | 1990 | 1995 | 2000 | 2004 |
|---|---|---|---|---|
| Federal national defense | | | | |
| Consumption | 308.1 | 297.3 | 321.5 | 477.5 |
| Gross | 65.9 | 51.4 | 48.8 | 70.4 |
| Federal nondefense | | | | |
| Consumption | 111.7 | 143.2 | 177.8 | 227.0 |
| Gross | 22.6 | 27.3 | 30.7 | 35.0 |
| State and local | | | | |
| Consumption | 544.6 | 696.1 | 917.8 | 1,099.7 |
| Gross | 127.2 | 154.0 | 225.0 | 274.3 |

*Source: Statistical Abstract of the United States*, 2006, Table 419.

2.40 Xr02-15* The table below displays the energy consumption patterns of Australia and New Zealand. The figures measure the heat content in metric tons (1,000 kilograms) of oil equivalent. Use a graphical technique to display the differences between the sources of energy for the two countries.

| Energy Sources | Australia | New Zealand |
|---|---|---|
| Nonrenewable | | |
| Coal & coal products | 55,385 | 1,281 |
| Oil | 33,185 | 6,275 |
| Natural Gas | 20,350 | 5,324 |
| Nuclear | 0 | 0 |
| Renewable | | |
| Hydroelectric | 1,388 | 1,848 |
| Solid Biomass | 4,741 | 805 |
| Other (Liquid biomass, geothermal, solar, wind, and tide, wave, & ocean) | 347 | 2,761 |
| Total | 115,396 | 18,294 |

*Source*: International Energy Association.

*The following exercises require a computer and software.*

2.41 Xr02-41 The average loss from a robbery in the United States in 2004 was $1,308 (*Source:* U.S. Federal Bureau of Investigation). Suppose that a government agency wanted to know whether the type of robbery differed between 1990, 1995, 2000, and 2006. A random sample of robbery reports was taken from each of these years, and the types were recorded using the codes below. Determine whether there are differences in the types of robbery over the 16-year span. (Adapted from *Statistical Abstract of the United States*, 2009, Table 308.)

1. Street or highway
2. Commercial house
3. Gas station
4. Convenience store
5. Residence
6. Bank
7. Other

2.42 Xr02-42 The associate dean of a business school was looking for ways to improve the quality of the applicants to its MBA program. In particular, she wanted to know whether the undergraduate degree of applicants differed among her school and the three nearby universities with MBA programs. She sampled 100 applicants of her program and an equal number from each of the other universities. She recorded their undergraduate degrees (1 = BA, 2 = BEng, 3 = BBA, 4 = other) as well the university (codes 1, 2, 3, and 4). Use a tabular technique to determine whether the undergraduate degree and the university each person applied to appear to be related.

2.43 Xr02-43 Is there brand loyalty among car owners in their purchases of gasoline? To help answer the question, a random sample of car owners was asked to record the brand of gasoline in their last two purchases (1 = Exxon, 2 = Amoco, 3 = Texaco, 4 = Other). Use a tabular technique to formulate your answer.

2.44 Xr02-44 The costs of smoking for individuals, companies for whom they work, and society in general is in the many billions of dollars. In an effort to reduce smoking, various government and non-government organizations have undertaken information campaigns about the dangers of smoking. Most of these have been directed at young people. This raises the question: Are you more likely to smoke if your parents smoke? To shed light on the issue, a sample of 20- to 40-year-old people were asked whether they smoked and whether their parents smoked. The results are stored the following way:

Column 1: 1 = do not smoke, 2 = smoke
Column 2: 1 = neither parent smoked,
2 = father smoked, 3 = mother smoked,
4 = both parents smoked

Use a tabular technique to produce the information you need.

2.45 Xr02-45 In 2007, 3,882,000 men and 3,196,000 women were unemployed at some time during the year (*Source*: U.S. Bureau of Labor Statistics). A statistics practitioner wanted to investigate the reasons for unemployment and whether the reasons differed by gender. A random sample of people 16 years of age and older was drawn. The reasons given for their status are:

1. Lost job
2. Left job
3. Reentrants
4. New entrants

Determine whether there are differences between unemployed men and women in terms of the reasons for unemployment. (*Source:* Adapted from *Statistical Abstract of the United States*, 2009 Table 604.)

2.46 Xr02-46 In 2004, the total number of prescriptions sold in the United States was 3,274,000,000 (*Source*: National Association of Drug Store Chains). The sales manager of a chain of drugstores wanted to determine whether changes were made in the way the prescriptions were filled. A survey of prescriptions was undertaken in 1995, 2000, and 2007. The year and type of each prescription were recorded using the codes below. Determine whether there are differences between the years. (*Source:* Adapted from the *Statistical Abstract of the United States*, 2009, Table 151.)

1. Traditional chain store
2. Independent drugstore
3. Mass merchant
4. Supermarket
5. Mail order

2.47 Xr02-33* Refer to Exercise 2.33. Also recorded was the gender of the respondents. Use a graphical technique to determine whether the choice of light beers differs between genders.

## CHAPTER SUMMARY

Descriptive statistical methods are used to summarize data sets so that we can extract the relevant information. In this chapter, we presented graphical techniques for nominal data.

Bar charts, pie charts, and frequency distributions are employed to summarize single sets of nominal data. Because of the restrictions applied to this type of data, all that we can show is the frequency and proportion of each category.

To describe the relationship between two nominal variables, we produce cross classification tables and bar charts.

## IMPORTANT TERMS

## COMPUTER OUTPUT AND INSTRUCTIONS

| Graphical Technique | Excel | Minitab |
|---|---|---|
| Bar chart | 22 | 23 |
| Pie chart | 22 | 23 |

# Chapter Exercises

*The following exercises require a computer and software.*

2.48 Xr02-48 A sample of 200 people who had purchased food at the concession stand at Yankee Stadium was asked to rate the quality of the food. The responses are:

a. Poor
b. Fair
c. Good
d. Very good
e. Excellent

Draw a graph that describes the data. What does the graph tell you?

2.49 Xr02-49 There are several ways to teach applied statistics. The most popular approaches are:

a. Emphasize manual calculations
b. Use a computer combined with manual calculations
c. Use a computer exclusively with no manual calculations

A survey of 100 statistics instructors asked each one to report his or her approach. Use a graphical method to extract the most useful information about the teaching approaches.

2.50 Xr02-50 Which Internet search engines are the most popular? A survey undertaken by the *Financial Post* (May 14, 2004) asked random samples of Americans and Canadians that question. The responses were:

1. Google
2. Microsoft (MSN)
3. Yahoo
4. Other

Use a graphical technique that compares the proportions of Americans' and Canadians' use of search engines

2.51 Xr02-51 The Wilfrid Laurier University bookstore conducts annual surveys of its customers. One question asks respondents to rate the prices of textbooks. The wording is, "The bookstore's prices of textbooks are reasonable." The responses are:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

The responses for a group of 115 students were recorded. Graphically summarize these data and report your findings.

2.52 Xr02-52 The Red Lobster restaurant chain conducts regular surveys of its customers to monitor the performance of individual restaurants. One question asks customers to rate the overall quality of their last visit. The listed responses are poor (1), fair (2), good (3), very good (4), and excellent (5). The survey also asks respondents whether their children accompanied them to the restaurant (1 = yes, 2 = no). Graphically depict these data and describe your findings.

2.53 Xr02-53 Many countries are lowering taxes on corporations in an effort to make their countries more attractive for investment. In the next table, we list the marginal effective corporate tax rates among Organization for Economic Co-Operation and Development (OECD) countries. Develop a graph that depicts these figures. Briefly describe your results.

| Country | Manufacturers | Services | Aggregate |
|---|---|---|---|
| Australia | 27.7 | 26.6 | 26.7 |
| Austria | 21.6 | 19.5 | 19.9 |
| Belgium | −6.0 | −4.1 | −0.5 |
| Canada | 20.0 | 29.2 | 25.2 |
| Czech Republic | 1.0 | 7.8 | 8.4 |
| Denmark | 16.5 | 12.7 | 13.4 |
| Finland | 22.4 | 22.9 | 22.8 |
| France | 33.0 | 31.7 | 31.9 |
| Germany | 30.8 | 29.4 | 29.7 |
| Greece | 18.0 | 13.2 | 13.8 |
| Hungary | 12.9 | 12.0 | 12.2 |
| Iceland | 19.5 | 17.6 | 17.9 |
| Ireland | 12.7 | 11.7 | 12.0 |
| Italy | 24.6 | 28.6 | 27.8 |
| Japan | 35.2 | 30.4 | 31.3 |
| Korea | 32.8 | 31.0 | 31.5 |
| Luxembourg | 24.1 | 20.3 | 20.6 |
| Mexico | 17.1 | 12.1 | 13.1 |
| Netherlands | 18.3 | 15.0 | 15.5 |
| New Zealand | 27.1 | 25.4 | 25.7 |
| Norway | 25.8 | 23.2 | 23.5 |
| Poland | 14.4 | 15.0 | 14.9 |
| Portugal | 14.8 | 16.1 | 15.9 |
| Slovak | 13.3 | 11.7 | 12.0 |
| Spain | 27.2 | 25.2 | 25.5 |
| Sweden | 19.3 | 17.5 | 17.8 |
| Switzerland | 14.8 | 15.0 | 14.9 |
| Turkey | 22.7 | 20.2 | 20.8 |
| United Kingdom | 22.7 | 27.8 | 26.9 |
| United States | 32.7 | 39.9 | 36.9 |

2.54 Xr02-54* A survey of the business school graduates undertaken by a university placement office asked, among other questions, the area in which each person was employed. The areas of employment are:

a. Accounting
b. Finance
c. General management
d. Marketing/Sales
e. Other

Additional questions were asked and the responses were recorded in the following way.

| Column | Variable |
|---|---|
| 1 | Identification number |
| 2 | Area |
| 3 | Gender (1 = female, 2 = male) |
| 4 | Job satisfaction (4 = very, 3 = quite, 2 = little, 1 = none) |

The placement office wants to know the following:

a. Do female and male graduates differ in their areas of employment? If so, how?
b. Are area of employment and job satisfaction related?

3

© Tonis Valling/Shutterstock

# GRAPHICAL DESCRIPTIVE TECHNIQUES II

## Were Oil Companies Gouging Customers 2000–2009?

DATA
Xm03-00

The price of oil has been increasing for several reasons. First, oil is a finite resource; the world will eventually run out. In January 2009, the world was consuming more than 100 million barrels per day—more than 36 billion barrels per year. The total proven world reserves of oil are 1,348.5 billion barrels. At today's consumption levels, the proven reserves will be exhausted in 37 years. (It should be noted, however, that in 2005 the proven reserves of oil amounted to 1,349.4 billion barrels, indicating that new oil discoveries are offsetting increasing usage.) Second, China's and India's industries are rapidly increasing and require ever-increasing amounts of oil. Third, over the last 10 years, hurricanes have threatened the oil rigs in the Gulf of Mexico.

The result of the price increases in oil is reflected in the price of gasoline. In January 2000, the average retail price of gasoline in the United States was $1.301 per U.S. gallon (one U.S.

© Comstock Images/Jupiterimages

On page 78 you will find our answer

gallon equals 3.79 liters) and the price of oil (West Texas intermediate crude) was $27.18 per barrel (one barrel equals 42 U.S. gallons). (*Sources:* U.S Department of Energy.) Over the next 10 years, the price of both oil and gasoline substantially increased. Many drivers complained that the oil companies were guilty of price gouging; that is, they believed that when the price of oil increased, the price of gas also increased, but when the price of oil decreased, the decrease in the price of gasoline seemed to lag behind. To determine whether this perception is accurate, we determined the monthly figures for both commodities. Were oil and gas prices related?

## Introduction

Chapter 2 introduced graphical techniques used to summarize and present nominal data. In this chapter, we do the same for interval data. Section 3.1 presents techniques to describe a set of interval data, Section 3.2 introduces time series and the method used to present time series data, and Section 3.3 describes the technique we use to describe the relationship between two interval variables. We complete this chapter with a discussion of how to properly use graphical methods in Section 3.4.

## 3.1 Graphical Techniques to Describe a Set of Interval Data

In this section, we introduce several graphical methods that are used when the data are interval. The most important of these graphical methods is the histogram. As you will see, the histogram not only is a powerful graphical technique used to summarize interval data but also is used to help explain an important aspect of probability (see Chapter 8).

### APPLICATIONS in MARKETING

© AP Photo/Paul Sakuma

#### Pricing

Traditionally, marketing has been defined in terms of the four P's: product, price, promotion, and place. *Marketing management* is the functional area of business that focuses on the development of a product, together with its pricing, promotion, and distribution. Decisions are made in these four areas with a view to satisfying the wants and needs of consumers while also satisfying the firm's objective.

The pricing decision must be addressed both for a new product, and, from time to time, for an existing product. Anyone buying a product such as a personal computer has been confronted with a wide variety of prices, accompanied by a correspondingly wide variety of features. From a vendor's standpoint, establishing the appropriate price and corresponding set of attributes for a product is complicated and must be done in the context of the overall marketing plan for the product.

EXAMPLE 3.1

## Analysis of Long-Distance Telephone Bills

DATA
Xm03-01

Following deregulation of telephone service, several new companies were created to compete in the business of providing long-distance telephone service. In almost all cases, these companies competed on price because the service each offered is similar. Pricing a service or product in the face of stiff competition is very difficult. Factors to be considered include supply, demand, price elasticity, and the actions of competitors. Long-distance packages may employ per minute charges, a flat monthly rate, or some combination of the two. Determining the appropriate rate structure is facilitated by acquiring information about the behaviors of customers, especially the size of monthly long-distance bills.

As part of a larger study, a long-distance company wanted to acquire information about the monthly bills of new subscribers in the first month after signing with the company. The company's marketing manager conducted a survey of 200 new residential subscribers and recorded the first month's bills. These data are listed here. The general manager planned to present his findings to senior executives. What information can be extracted from these data?

**Long-Distance Telephone Bills**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 42.19 | 39.21 | 75.71 | 8.37 | 1.62 | 28.77 | 35.32 | 13.9 | 114.67 | 15.3 |
| 38.45 | 48.54 | 88.62 | 7.18 | 91.1 | 9.12 | 117.69 | 9.22 | 27.57 | 75.49 |
| 29.23 | 93.31 | 99.5 | 11.07 | 10.88 | 118.75 | 106.84 | 109.94 | 64.78 | 68.69 |
| 89.35 | 104.88 | 85 | 1.47 | 30.62 | 0 | 8.4 | 10.7 | 45.81 | 35 |
| 118.04 | 30.61 | 0 | 26.4 | 100.05 | 13.95 | 90.04 | 0 | 56.04 | 9.12 |
| 110.46 | 22.57 | 8.41 | 13.26 | 26.97 | 14.34 | 3.85 | 11.27 | 20.39 | 18.49 |
| 0 | 63.7 | 70.48 | 21.13 | 15.43 | 79.52 | 91.56 | 72.02 | 31.77 | 84.12 |
| 72.88 | 104.84 | 92.88 | 95.03 | 29.25 | 2.72 | 10.13 | 7.74 | 94.67 | 13.68 |
| 83.05 | 6.45 | 3.2 | 29.04 | 1.88 | 9.63 | 5.72 | 5.04 | 44.32 | 20.84 |
| 95.73 | 16.47 | 115.5 | 5.42 | 16.44 | 21.34 | 33.69 | 33.4 | 3.69 | 100.04 |
| 103.15 | 89.5 | 2.42 | 77.21 | 109.08 | 104.4 | 115.78 | 6.95 | 19.34 | 112.94 |
| 94.52 | 13.36 | 1.08 | 72.47 | 2.45 | 2.88 | 0.98 | 6.48 | 13.54 | 20.12 |
| 26.84 | 44.16 | 76.69 | 0 | 21.97 | 65.9 | 19.45 | 11.64 | 18.89 | 53.21 |
| 93.93 | 92.97 | 13.62 | 5.64 | 17.12 | 20.55 | 0 | 83.26 | 1.57 | 15.3 |
| 90.26 | 99.56 | 88.51 | 6.48 | 19.7 | 3.43 | 27.21 | 15.42 | 0 | 49.24 |
| 72.78 | 92.62 | 55.99 | 6.95 | 6.93 | 10.44 | 89.27 | 24.49 | 5.2 | 9.44 |
| 101.36 | 78.89 | 12.24 | 19.6 | 10.05 | 21.36 | 14.49 | 89.13 | 2.8 | 2.67 |
| 104.8 | 87.71 | 119.63 | 8.11 | 99.03 | 24.42 | 92.17 | 111.14 | 5.1 | 4.69 |
| 74.01 | 93.57 | 23.31 | 9.01 | 29.24 | 95.52 | 21 | 92.64 | 3.03 | 41.38 |
| 56.01 | 0 | 11.05 | 84.77 | 15.21 | 6.72 | 106.59 | 53.9 | 9.16 | 45.77 |

### SOLUTION

Little information can be developed just by casually reading through the 200 observations. The manager can probably see that most of the bills are under $100, but that is likely to be the extent of the information garnered from browsing through the data. If he examines the data more carefully, he may discover that the smallest bill is $0 and the largest is $119.63. He has now developed some information. However, his presentation to senior executives will be most unimpressive if no other information is produced. For example, someone is likely to ask how the numbers are distributed between 0 and 119.63. Are there many small bills and few large bills? What is the "typical" bill? Are the bills somewhat similar or do they vary considerably?

To help answer these questions and others like them, the marketing manager can construct a frequency distribution from which a histogram can be drawn. In the previous section a frequency distribution was created by counting the number of times each category of the nominal variable occurred. We create a frequency distribution for interval data by counting the number of observations that fall into each of a series of intervals, called **classes**, that cover the complete range of observations. We discuss how to decide the number of classes and the upper and lower limits of the intervals later. We have chosen eight classes defined in such a way that each observation falls into one and only one class. These classes are defined as follows:

**Classes**

Amounts that are less than or equal to 15

Amounts that are more than 15 but less than or equal to 30

Amounts that are more than 30 but less than or equal to 45

Amounts that are more than 45 but less than or equal to 60

Amounts that are more than 60 but less than or equal to 75

Amounts that are more than 75 but less than or equal to 90

Amounts that are more than 90 but less than or equal to 105

Amounts that are more than 105 but less than or equal to 120

Notice that the intervals do not overlap, so there is no uncertainty about which interval to assign to any observation. Moreover, because the smallest number is 0 and the largest is 119.63, every observation will be assigned to a class. Finally, the intervals are equally wide. Although this is not essential, it makes the task of reading and interpreting the graph easier.

To create the frequency distribution manually, we count the number of observations that fall into each interval. Table 3.1 presents the frequency distribution.

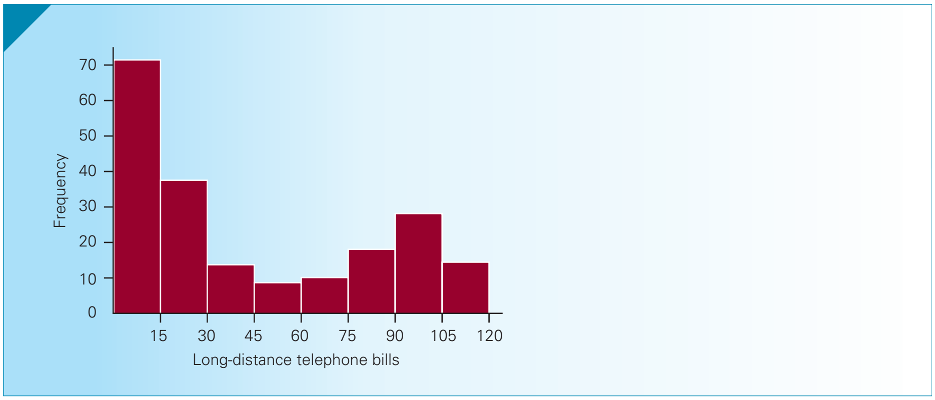
TABLE 3.1 Frequency Distribution of the Long-Distance Bills in Example 3.1

| CLASS LIMITS | FREQUENCY |
|---|---|
| 0 to 15* | 71 |
| 15 to 30 | 37 |
| 30 to 45 | 13 |
| 45 to 60 | 9 |
| 60 to 75 | 10 |
| 75 to 90 | 18 |
| 90 to 105 | 28 |
| 105 to 120 | 14 |
| Total | 200 |

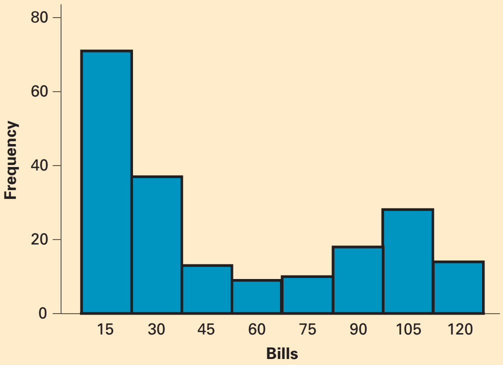
*Classes contain observations greater than their lower limits (except for the first class) and less than or equal to their upper limits.

Although the frequency distribution provides information about how the numbers are distributed, the information is more easily understood and imparted by drawing a picture or graph. The graph is called a **histogram**. A histogram is created by drawing rectangles whose bases are the intervals and whose heights are the frequencies. Figure 3.1 exhibits the histogram that was drawn by hand.

FIGURE **3.1** Histogram for Example 3.1



## EXCEL



### INSTRUCTIONS

1. Type or import the data into one column. (Open Xm03-01.) In another column, type the upper limits of the class intervals. Excel calls them *bins*. (You can put any name in the first row; we typed "Bills.")
2. Click **Data**, **Data Analysis**, and **Histogram.** If Data Analysis does not appear in the menu box, see our Keller's website, Appendix A1.
3. Specify the **Input Range** (A1:A201) and the **Bin Range** (B1:B9). Click **Chart Output**. Click **Labels** if the first row contains names.
4. To remove the gaps, place the cursor over one of the rectangles and click the right button of the mouse. Click (with the left button) **Format Data Series . . . .** move the pointer to **Gap Width** and use the slider to change the number from 150 to 0.

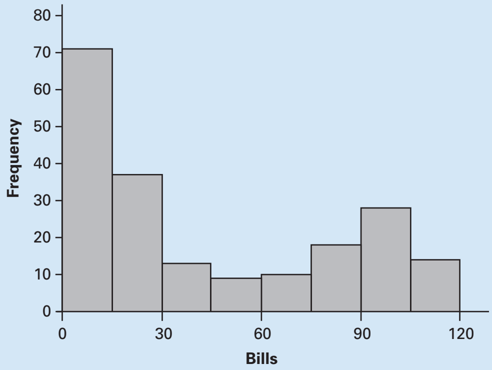
Except for the first class, Excel counts the number of observations in each class that are greater than the lower limit and less than or equal to the upper limit.

Note that the numbers along the horizontal axis represent the upper limits of each class although they appear to be placed in the centers. If you wish, you can replace these numbers with the actual midpoints by making changes to the frequency distribution in cells A1:B14 (change 15 to 7.5, 30 to 22.5, . . . , and 120 to 112.5).

You can also convert the histogram to list relative frequencies instead of frequencies. To do so, change the frequencies to relative frequencies by dividing each frequency by 200; that is, replace 71 by .355, 37 by .185, . . . , and 14 by .07.

If you have difficulty with this technique, turn to the website Appendix A2 or A3, which provides step-by-step instructions for Excel and provides troubleshooting tips.

## MINITAB



Note that Minitab counts the number of observations in each class that are strictly less than their upper limits.

### INSTRUCTIONS

1. Type or import the data into one column. (Open Xm03-01.)
2. Click **Graph**, **Histogram . . .** , and **Simple**.
3. Type or use the **Select** button to specify the name of the variable in the **Graph Variables** box (Bills). Click **Data View**.
4. Click **Data Display** and **Bars**. Minitab will create a histogram using its own choices of class intervals.
5. To choose your own classes, double-click the horizontal axis. Click **Binning**.
6. Under **Interval Type**, choose **Cutpoint**. Under **Interval Definition**, choose **Midpoint/Cutpoint positions** and type in your choices (0 15 30 45 60 75 90 105 120) to produce the histogram shown here.

## INTERPRET

The histogram gives us a clear view of the way the bills are distributed. About half the monthly bills are small ($0 to $30), a few bills are in the middle range ($30 to $75), and a relatively large number of long-distance bills are at the high end of the range. It would appear from this sample of first-month long-distance bills that the company's customers are split unevenly between light and heavy users of long-distance telephone service. If the company assumes that this pattern will continue, it must address a number of pricing issues. For example, customers who incurred large monthly bills may be targets of competitors who offer flat rates for 15-minute or 30-minute calls. The company needs to know more about these customers. With the additional information, the marketing manager may suggest altering the company's pricing.

## Determining the Number of Class Intervals

The number of class intervals we select depends entirely on the number of observations in the data set. The more observations we have, the larger the number of class intervals we need to use to draw a useful histogram. Table 3.2 provides guidelines on choosing the number of classes. In Example 3.1, we had 200 observations. The table tells us to use 7, 8, 9, or 10 classes.

TABLE **3.2** Approximate Number of Classes in Histograms

| NUMBER OF OBSERVATIONS | NUMBER OF CLASSES |
|---|---|
| Less than 50 | 5–7 |
| 50–200 | 7–9 |
| 200–500 | 9–10 |
| 500–1,000 | 10–11 |
| 1,000–5,000 | 11–13 |
| 5,000–50,000 | 13–17 |
| More than 50,000 | 17–20 |

An alternative to the guidelines listed in Table 3.2 is to use Sturges's formula, which recommends that the number of class intervals be determined by the following:

$$\text{Number of class intervals} = 1 + 3.3 \log(n)$$

For example, if n = 50 Sturges's formula becomes

$$\text{Number of class intervals} = 1 + 3.3 \log(50) = 1 + 3.3(1.7) = 6.6$$

which we round to 7.

**Class Interval Widths** We determine the approximate width of the classes by subtracting the smallest observation from the largest and dividing the difference by the number of classes. Thus,

$$\text{Class width} = \frac{\textit{Largest Observation} - \textit{Smallest Observation}}{\textit{Number of Classes}}$$

In Example 3.1, we calculated

$$\text{Class width} = \frac{119.63 - 0}{8} = 14.95$$

We often round the result to some convenient value. We then define our class limits by selecting a lower limit for the first class from which all other limits are determined. The only condition we apply is that the first class interval must contain the smallest observation. In Example 3.1, we rounded the class width to 15 and set the lower limit of the first class to 0. Thus, the first class is defined as "Amounts that are greater than or equal to 0 but less than or equal to 15." (Minitab users should remember that the classes are defined as the number of observations that are *strictly less* than their upper limits.)

Table 3.2 and Sturges's formula are guidelines only. It is more important to choose classes that are easy to interpret. For example, suppose that we have recorded the marks on an exam of the 100 students registered in the course where the highest mark is 94 and the lowest is 48. Table 3.2 suggests that we use 7, 8, or 9 classes, and Sturges's formula computes the approximate number of classes as

$$\text{Number of class intervals} = 1 + 3.3 \log(100) = 1 + 3.3(2) = 7.6$$

which we round to 8. Thus,

$$\text{Class width} = \frac{94 - 48}{8} = 5.75$$

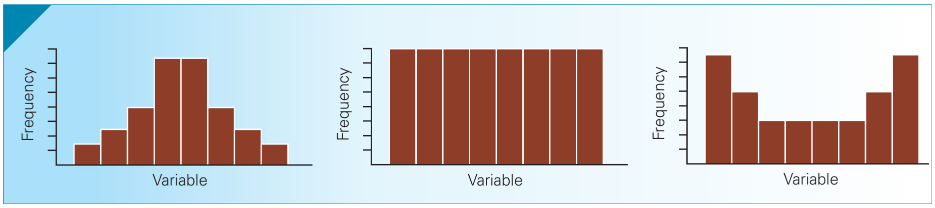
which we would round to 6. We could then produce a histogram whose upper limits of the class intervals are 50, 56, 62, . . . , 98. Because of the rounding and the way in which we defined the class limits, the number of classes is 9. However, a histogram that is easier to interpret would be produced using classes whose widths are 5; that is, the upper limits would be 50, 55, 60, . . . , 95. The number of classes in this case would be 10.

## Shapes of Histograms

The purpose of drawing histograms, like that of all other statistical techniques, is to acquire information. Once we have the information, we frequently need to describe what we've learned to others. We describe the shape of histograms on the basis of the following characteristics.

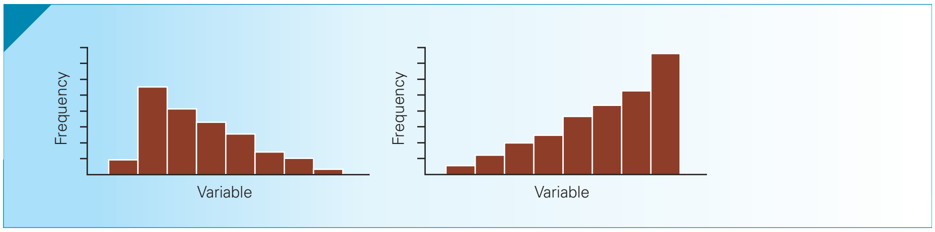
**Symmetry** A histogram is said to be **symmetric** if, when we draw a vertical line down the center of the histogram, the two sides are identical in shape and size. Figure 3.2 depicts three symmetric histograms.

FIGURE **3.2** Three Symmetric Histograms



**Skewness** A skewed histogram is one with a long tail extending to either the right or the left. The former is called **positively skewed**, and the latter is called **negatively skewed**. Figure 3.3 shows examples of both. Incomes of employees in large firms tend to be positively skewed because there is a large number of relatively low-paid workers and a small number of well-paid executives. The time taken by students to write exams is frequently negatively skewed because few students hand in their exams early; most prefer to reread their papers and hand them in near the end of the scheduled test period.

FIGURE **3.3** Positively and Negatively Skewed Histograms



**Number of Modal Classes** As we discuss in Chapter 4, a *mode* is the observation that occurs with the greatest frequency. A **modal class** is the class with the largest number of observations. A **unimodal histogram** is one with a single peak. The histogram in

Figure 3.4 is unimodal. A **bimodal histogram** is one with two peaks, not necessarily equal in height. Bimodal histograms often indicate that two different distributions are present. (See Example 3.4.) Figure 3.5 depicts bimodal histograms.

FIGURE **3.4** A Unimodal Histogram

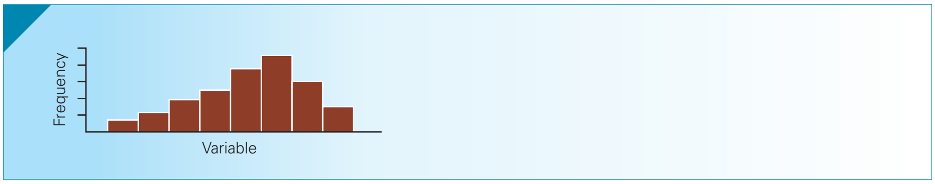
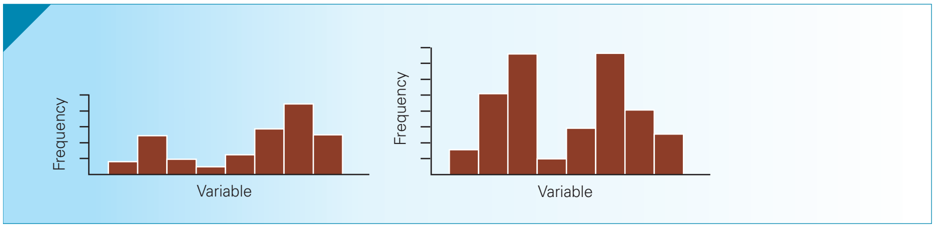
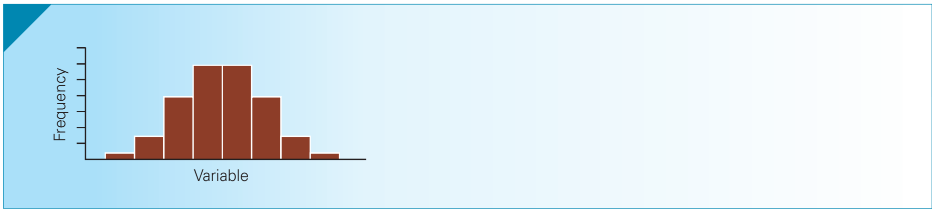


FIGURE **3.5** Bimodal Histograms



**Bell Shape** A special type of symmetric unimodal histogram is one that is bell shaped. In Chapter 8 we will explain why this type of histogram is important. Figure 3.6 exhibits a bell-shaped histogram.

FIGURE **3.6** Bell-Shaped Histogram



Now that we know what to look for, let's examine some examples of histograms and see what we can discover.

## APPLICATIONS in FINANCE

© BanaStock/Jupiter Images

### Stock and Bond Valuation

A basic understanding of how financial assets, such as stocks and bonds, are valued is critical to good financial management. Understanding the basics of valuation is necessary for capital budgeting and capital structure decisions. Moreover, understanding the basics of valuing investments such as stocks and bonds is at the heart of the huge and growing discipline known as *investment management*.

A financial manager must be familiar with the main characteristics of the capital markets where long-term financial assets such as stocks and bonds trade. A well-functioning capital market provides managers with useful information concerning the appropriate prices and rates of return that are required for a variety of financial securities with differing levels of risk. Statistical methods can be used to analyze capital markets and summarize their characteristics, such as the shape of the distribution of stock or bond returns.

## APPLICATIONS in FINANCE

© Vicki Beaver

### Return on Investment

The return on an investment is calculated by dividing the gain (or loss) by the value of the investment. For example, a $100 investment that is worth $106 after 1 year has a 6% rate of return. A $100 investment that loses $20 has a −20% rate of return. For many investments, including individual stocks and stock portfolios (combinations of various stocks), the rate of return is a variable. In other words, the investor does not know in advance what the rate of return will be. It could be a positive number, in which case the investor makes money—or negative, and the investor loses money.

Investors are torn between two goals. The first is to maximize the rate of return on investment. The second goal is to reduce risk. If we draw a histogram of the returns for a certain investment, the location of the center of the histogram gives us some information about the return one might expect from that investment. The spread or variation of the histogram provides us with guidance about the risk. If there is little variation, an investor can be quite confident in predicting what his or her rate of return will be. If there is a great deal of variation, the return becomes much less predictable and thus riskier. Minimizing the risk becomes an important goal for investors and financial analysts.

## EXAMPLE 3.2 Comparing Returns on Two Investments

DATA
Xm03-02

Suppose that you are facing a decision about where to invest that small fortune that remains after you have deducted the anticipated expenses for the next year from the earnings from your summer job. A friend has suggested two types of investment, and to help make the decision you acquire some rates of return from each type. You would like to know what you can expect by way of the return on your investment, as well as other types of information, such as whether the rates are spread out over a wide range (making the investment risky) or are grouped tightly together (indicating relatively low risk). Do the data indicate that it is possible that you can do extremely well with little likelihood of a large loss? Is it likely that you could lose money (negative rate of return)?

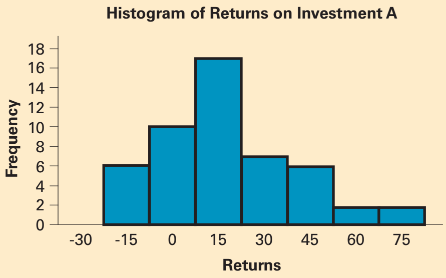
The returns for the two types of investments are listed here. Draw histograms for each set of returns and report on your findings. Which investment would you choose and why?

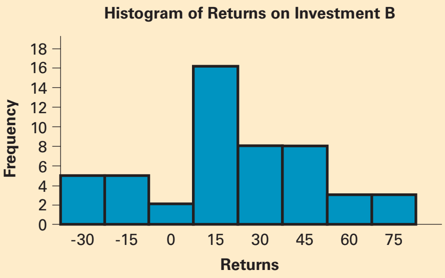
| Returns on Investment A | | | | Returns on Investment B | | | |
|---|---|---|---|---|---|---|---|
| 30.00 | 6.93 | 13.77 | −8.55 | 30.33 | −34.75 | 30.31 | 24.3 |
| −2.13 | −13.24 | 22.42 | −5.29 | −30.37 | 54.19 | 6.06 | −10.01 |
| 4.30 | −18.95 | 34.40 | −7.04 | −5.61 | 44.00 | 14.73 | 35.24 |
| 25.00 | 9.43 | 49.87 | −12.11 | 29.00 | −20.23 | 36.13 | 40.7 |
| 12.89 | 1.21 | 22.92 | 12.89 | −26.01 | 4.16 | 1.53 | 22.18 |
| −20.24 | 31.76 | 20.95 | 63.00 | 0.46 | 10.03 | 17.61 | 3.24 |
| 1.20 | 11.07 | 43.71 | −19.27 | 2.07 | 10.51 | 1.2 | 25.1 |
| −2.59 | 8.47 | −12.83 | −9.22 | 29.44 | 39.04 | 9.94 | −24.24 |
| 33.00 | 36.08 | 0.52 | −17.00 | 11 | 24.76 | −33.39 | −38.47 |
| 14.26 | −21.95 | 61.00 | 17.30 | −25.93 | 15.28 | 58.67 | 13.44 |
| −15.83 | 10.33 | −11.96 | 52.00 | 8.29 | 34.21 | 0.25 | 68.00 |
| 0.63 | 12.68 | 1.94 | | 61.00 | 52.00 | 5.23 | |
| 38.00 | 13.09 | 28.45 | | −20.44 | −32.17 | 66 | |

## SOLUTION

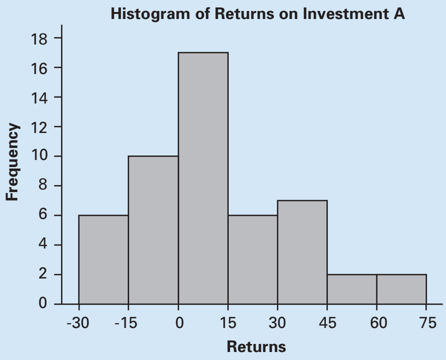
We draw the histograms of the returns on the two investments. We'll use Excel and Minitab to do the work.

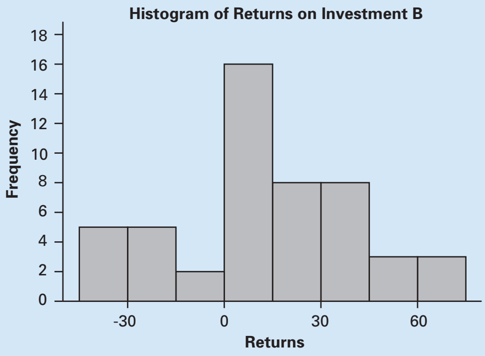
## EXCEL


Histogram of Returns on Investment A


Histogram of Returns on Investment B

## MINITAB


Histogram of Returns on Investment A


Histogram of Returns on Investment B

## INTERPRET

Comparing the two histograms, we can extract the following information:

1. The center of the histogram of the returns of investment A is slightly lower than that for investment B.
2. The spread of returns for investment A is considerably less than that for investment B.
3. Both histograms are slightly positively skewed.

These findings suggest that investment A is superior. Although the returns for A are slightly less than those for B, the wider spread for B makes it unappealing to most investors. Both investments allow for the possibility of a relatively large return.

The interpretation of the histograms is somewhat subjective. Other viewers may not concur with our conclusion. In such cases, numerical techniques provide the detail and precision lacking in most graphs. We will redo this example in Chapter 4 to illustrate how numerical techniques compare to graphical ones.

## EXAMPLE 3.3 Business Statistics Marks

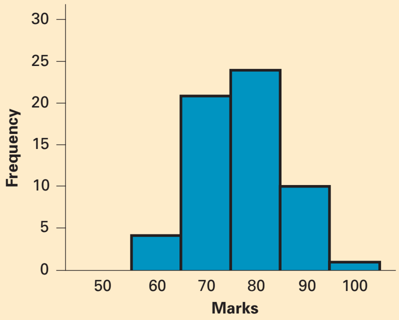
DATA
Xm03-03*

A student enrolled in a business program is attending the first class of the required statistics course. The student is somewhat apprehensive because he believes the myth that the course is difficult. To alleviate his anxiety, the student asks the professor about last year's marks. The professor obliges and provides a list of the final marks, which is composed of term work plus the final exam. Draw a histogram and describe the result, based on the following marks:

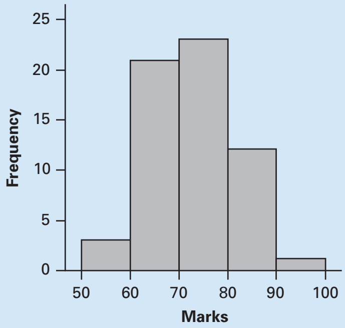
| | | | |
|---|---|---|---|
| 65 | 81 | 72 | 59 |
| 71 | 53 | 85 | 66 |
| 66 | 70 | 72 | 71 |
| 79 | 76 | 77 | 68 |
| 65 | 73 | 64 | 72 |
| 82 | 73 | 77 | 75 |
| 80 | 85 | 89 | 74 |
| 86 | 83 | 87 | 77 |
| 67 | 80 | 78 | 69 |
| 64 | 67 | 79 | 60 |
| 62 | 78 | 59 | 92 |
| 74 | 68 | 63 | 69 |
| 67 | 67 | 84 | 69 |
| 72 | 62 | 74 | 73 |
| 68 | 83 | 74 | 65 |

## SOLUTION

### EXCEL



### MINITAB



### INTERPRET

The histogram is unimodal and approximately symmetric. There are no marks below 50, with the great majority of marks between 60 and 90. The modal class is 70 to 80, and the center of the distribution is approximately 75.

## EXAMPLE 3.4 Mathematical Statistics Marks

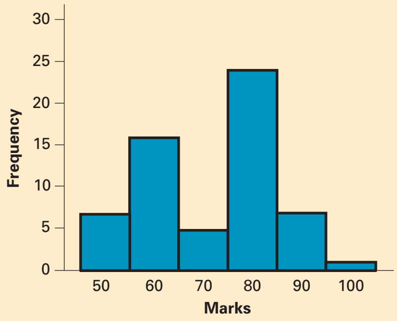
DATA
Xm03-04*

Suppose the student in Example 3.3 obtained a list of last year's marks in a mathematical statistics course. This course emphasizes derivations and proofs of theorems. Use the accompanying data to draw a histogram and compare it to the one produced in Example 3.3. What does this histogram tell you?

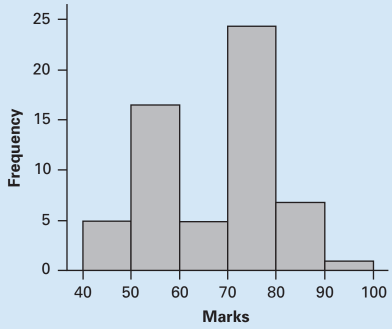
| | | | |
|---|---|---|---|
| 77 | 67 | 53 | 54 |
| 74 | 82 | 75 | 44 |
| 75 | 55 | 76 | 54 |
| 75 | 73 | 59 | 60 |
| 67 | 92 | 82 | 50 |
| 72 | 75 | 82 | 52 |
| 81 | 75 | 70 | 47 |
| 76 | 52 | 71 | 46 |
| 79 | 72 | 75 | 50 |
| 73 | 78 | 74 | 51 |
| 59 | 83 | 53 | 44 |
| 83 | 81 | 49 | 52 |
| 77 | 73 | 56 | 53 |
| 74 | 72 | 61 | 56 |
| 78 | 71 | 61 | 53 |

## SOLUTION

### EXCEL



### MINITAB



INTERPRET

The histogram is bimodal. The larger modal class is composed of the marks in the 70s. The smaller modal class includes the marks that are in the 50s. There appear to be few marks in the 60s. This histogram suggests that there are two groups of students. Because of the emphasis on mathematics in the course, one may conclude that those who performed poorly in the course are weaker mathematically than those who performed well. The histograms in this example and in Example 3.3 suggest that the courses are quite different from one another and have a completely different distribution of marks.

## Stem-and-Leaf Display

One of the drawbacks of the histogram is that we lose potentially useful information by classifying the observations. In Example 3.1, we learned that there are 71 observations that fall between 0 and 15. By classifying the observations we did acquire useful information. However, the histogram focuses our attention on the frequency of each class and by doing so sacrifices whatever information was contained in the actual observations. A statistician named John Tukey introduced the **stem-and-leaf display**, which is a method that to some extent overcomes this loss.

The first step in developing a stem-and-leaf display is to split each observation into two parts, a stem and a leaf. There are several different ways of doing this. For example, the number 12.3 can be split so that the stem is 12 and the leaf is 3. In this definition the stem consists of the digits to the left of the decimal and the leaf is the digit to the right of the decimal. Another method can define the stem as 1 and the leaf as 2 (ignoring the 3). In this definition the stem is the number of tens and the leaf is the number of ones. We'll use this definition to create a stem-and-leaf display for Example 3.1.

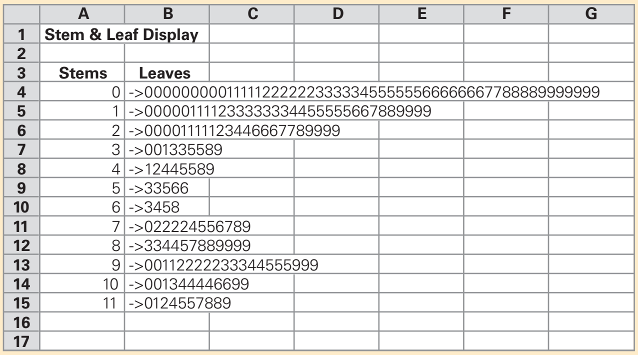
The first observation is 42.19. Thus, the stem is 4 and the leaf is 2. The second observation is 38.45, which has a stem of 3 and a leaf of 8. We continue converting each number in this way. The stem-and-leaf display consists of listing the stems 0, 1, 2, . . . , 11. After each stem, we list that stem's leaves, usually in ascending order. Figure 3.7 depicts the manually created stem-and-leaf display.

FIGURE 3.7 Stem-and-Leaf Display for Example 3.1

| Stem | Leaf |
|---|---|
| 0 | 000000000111112222223333345555556666666778888999999 |
| 1 | 0000011112333333344555556678899999 |
| 2 | 00001111123446667789999 |
| 3 | 001335589 |
| 4 | 124445589 |
| 5 | 33566 |
| 6 | 3458 |
| 7 | 022224556789 |
| 8 | 334457889999 |
| 9 | 00112222233344555999 |
| 10 | 001344446699 |
| 11 | 0124557889 |

As you can see the stem-and-leaf display is similar to a histogram turned on its side. The length of each line represents the frequency in the class interval defined by the stems. The advantage of the stem-and-leaf display over the histogram is that we can see the actual observations.

## EXCEL

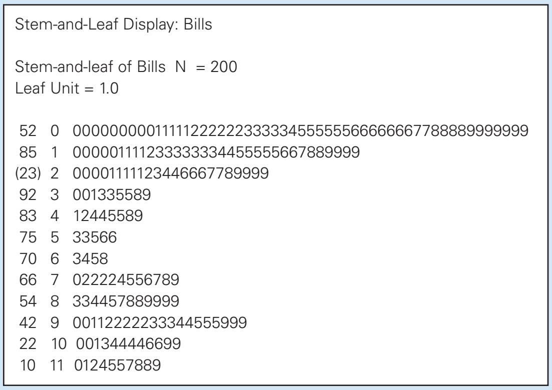


|  | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Stem & Leaf Display |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |
| 3 | Stems | Leaves |  |  |  |  |  |
| 4 | 0 | ->0000000001111122222233333455555566666667788889999999 |  |  |  |  |  |
| 5 | 1 | ->000001111233333334455555667889999 |  |  |  |  |  |
| 6 | 2 | ->00001111123446667789999 |  |  |  |  |  |
| 7 | 3 | ->001335589 |  |  |  |  |  |
| 8 | 4 | ->12445589 |  |  |  |  |  |
| 9 | 5 | ->33566 |  |  |  |  |  |
| 10 | 6 | ->3458 |  |  |  |  |  |
| 11 | 7 | ->022224556789 |  |  |  |  |  |
| 12 | 8 | ->334457889999 |  |  |  |  |  |
| 13 | 9 | ->00112222233344555999 |  |  |  |  |  |
| 14 | 10 | ->001344446699 |  |  |  |  |  |
| 15 | 11 | ->0124557889 |  |  |  |  |  |
| 16 |  |  |  |  |  |  |  |
| 17 |  |  |  |  |  |  |  |

### INSTRUCTIONS

1. Type or import the data into one column. (Open Xm03-01.)
2. Click **Add-ins, Data Analysis Plus**, and **Stem-and-Leaf Display**.
3. Specify the **Input Range** (A1:A201). Click one of the values of **Increment** (the increment is the difference between stems) (10).

## MINITAB



```
Stem-and-Leaf Display: Bills

Stem-and-leaf of Bills  N  = 200
Leaf Unit = 1.0

 52   0   0000000001111122222233333455555566666667788889999999
 85   1   000001111233333334455555667889999
(23)  2   00001111123446667789999
 92   3   001335589
 83   4   12445589
 75   5   33566
 70   6   3458
 66   7   022224556789
 54   8   334457889999
 42   9   00112222233344555999
 22   10  001344446699
 10   11  0124557889
```

The numbers in the left column are called **depths**. Each depth counts the number of observations that are on its line or beyond. For example, the second depth is 85, which means that there are 85 observations that are less than 20. The third depth is displayed in parentheses, which indicates that the third interval contains the observation that falls in the middle of all the observations, a statistic we call the *median* (to be presented in Chapter 4). For this interval, the depth tells us the frequency of the interval; that is, 23 observations are greater than or equal to 20 but less than 30. The fourth depth is 92, which tells us that 92 observations are greater than or equal to 30. Notice that for classes below the median, the depth reports the number of observations that are less than the upper limit of that class. For classes that are above the median, the depth reports the number of observations that are greater than or equal to the lower limit of that class.

*INSTRUCTIONS*

1. Type or import the data into one column. (Open Xm03-01.)
2. Click **Graph** and **Stem-and-Leaf. . . .**
3. Type or use the **Select** button to specify the variable in the **Variables** box (Bills). Type the increment in the **Increment** box (10).

## Ogive

The frequency distribution lists the number of observations that fall into each class interval. We can also create a **relative frequency distribution** by dividing the frequencies by the number of observations. Table 3.3 displays the relative frequency distribution for Example 3.1.

TABLE **3.3** Relative Frequency Distribution for Example 3.1

| CLASS LIMITS | RELATIVE FREQUENCY |
|---|---|
| 0 to 15 | 71/200 = .355 |
| 15 to 30 | 37/200 = .185 |
| 30 to 45 | 13/200 = .065 |
| 45 to 60 | 9/200 = .045 |
| 60 to 75 | 10/200 = .050 |
| 75 to 90 | 18/200 = .090 |
| 90 to 105 | 28/200 = .140 |
| 105 to 120 | 14/200 = .070 |
| Total | 200/200 = 1.0 |

As you can see, the relative frequency distribution highlights the proportion of the observations that fall into each class. In some situations, we may wish to highlight the proportion of observations that lie below each of the class limits. In such cases, we create the **cumulative relative frequency distribution**. Table 3.4 displays this type of distribution for Example 3.1.

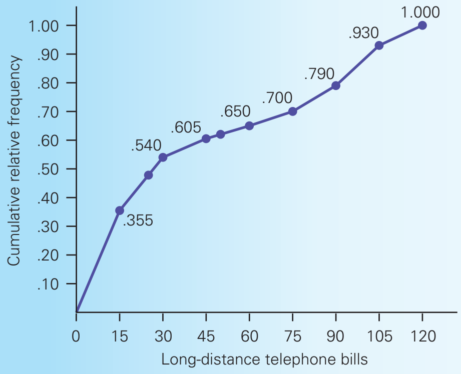
From Table 3.4, you can see that, for example, 54% of the bills were less than or equal to $30 and that 79% of the bills were less than or equal to $90.

Another way of presenting this information is the **ogive**, which is a graphical representation of the cumulative relative frequencies. Figure 3.8 is the manually drawn ogive for Example 3.1.

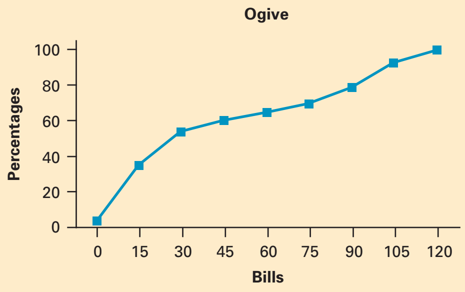
TABLE **3.4** Cumulative Relative Frequency Distribution for Example 3.1

| CLASS LIMITS | RELATIVE FREQUENCY | CUMULATIVE RELATIVE FREQUENCY |
|---|---|---|
| 0 to 15 | 71/200 = .355 | 71/200 = .355 |
| 15 to 30 | 37/200 = .185 | 108/200 = .540 |
| 30 to 45 | 13/200 = .065 | 121/200 = .605 |
| 45 to 60 | 9/200 = .045 | 130/200 = .650 |
| 60 to 75 | 10/200 = .05 | 140/200 = .700 |
| 75 to 90 | 18/200 = .09 | 158/200 = .790 |
| 90 to 105 | 28/200 = .14 | 186/200 = .930 |
| 105 to 120 | 14/200 = .07 | 200/200 = 1.00 |

FIGURE **3.8** Ogive for Example 3.1



## EXCEL



*INSTRUCTIONS*

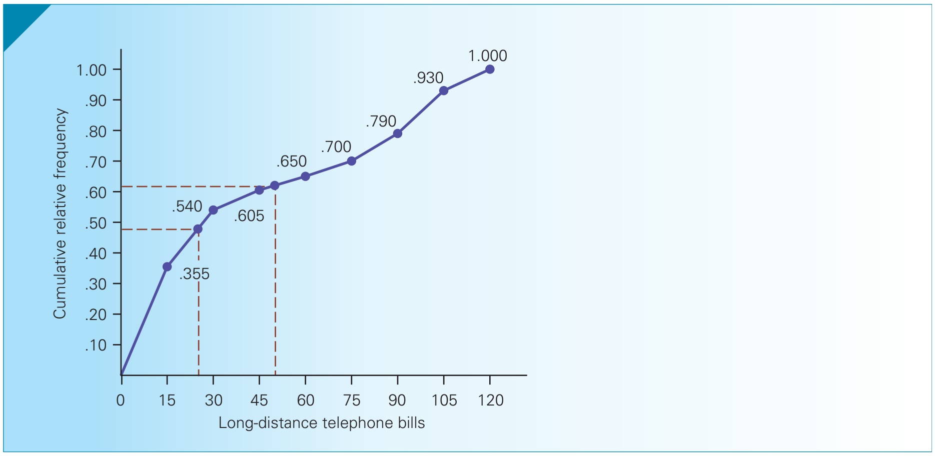
Follow instructions to create a histogram. Make the first bin's upper limit a number that is slightly smaller than the smallest number in the data set. Move the cursor to **Chart Output** and click. Do the same for **Cumulative Percentage**. Remove the "More" category. Click on any of the rectangles and click Delete. Change the **Scale**, if necessary. (Right-click the vertical or horizontal axis, click **Format Axis . . .** , and change the **Maximum** value of **Y** equal to 1.0.)

**MINITAB**

Minitab does not draw ogives.

We can use the ogive to estimate the cumulative relative frequencies of other values. For example, we estimate that about 62% of the bills lie below $50 and that about 48% lie below $25. (See Figure 3.9.)

FIGURE **3.9** Ogive with Estimated Relative Frequencies for Example 3.1



Here is a summary of this section's techniques.

**Factors That Identify When to Use a Histogram, Ogive, or Stem-and-Leaf Display**

1. Objective: Describe a single set of data
2. Data type: Interval

## EXERCISES

**3.1** How many classes should a histogram contain if the number of observations is 250?

**3.2** Determine the number of classes of a histogram for 700 observations.

**3.3** A data set consists of 125 observations that range between 37 and 188.

a. What is an appropriate number of classes to have in the histogram?

b. What class intervals would you suggest?

**3.4** A statistics practitioner would like to draw a histogram of 62 observations that range from 5.2 to 6.1.

a. What is an appropriate number of class intervals?

b. Define the upper limits of the classes you would use.

**3.5** Xr03-05 The number of items rejected daily by a manufacturer because of defects was recorded for the past 30 days. The results are as follows.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 9 | 13 | 7 | 5 | 8 | 12 | 15 | 5 | 7 | 3 |
| 8 | 15 | 17 | 19 | 6 | 4 | 10 | 8 | 22 | 16 | 9 |
| 5 | 3 | 9 | 19 | 14 | 13 | 18 | 7 | | | |

a. Construct a histogram.

b. Construct an ogive.

c. Describe the shape of the histogram.

**3.6** Xr03-06 The final exam in a third-year organizational behavior course requires students to write several essay-style answers. The numbers of pages for a sample of 25 exams were recorded. These data are shown here.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 8 | 9 | 3 | 12 | 8 | 5 | 7 | 3 | 8 | 9 | 5 | 2 |
| 7 | 12 | 9 | 6 | 3 | 8 | 7 | 10 | 9 | 12 | 7 | 3 | |

a. Draw a histogram.
b. Draw an ogive.
c. Describe what you've learned from the answers to parts (a) and (b).

**3.7** Xr03-07 A large investment firm on Wall Street wants to review the distribution of ages of its stockbrokers. The firm believes that this information can be useful in developing plans to recruit new brokers. The ages of a sample of 40 brokers are shown here.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 46 | 28 | 51 | 34 | 29 | 40 | 38 | 33 | 41 | 52 |
| 53 | 40 | 50 | 33 | 36 | 41 | 25 | 38 | 37 | 41 |
| 36 | 50 | 46 | 33 | 61 | 48 | 32 | 28 | 30 | 49 |
| 41 | 37 | 26 | 39 | 35 | 39 | 46 | 26 | 31 | 35 |

a. Draw a stem-and-leaf display.
b. Draw a histogram.
c. Draw an ogive.
d. Describe what you have learned.

**3.8** Xr03-08 The numbers of weekly sales calls by a sample of 30 telemarketers are listed here. Draw a histogram of these data and describe it.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 14 | 8 | 6 | 12 | 21 | 4 | 9 | 3 | 25 | 17 |
| 9 | 5 | 8 | 18 | 16 | 3 | 17 | 19 | 10 | 15 |
| 5 | 20 | 17 | 14 | 19 | 7 | 10 | 15 | 10 | 8 |

**3.9** Xr03-09 The amount of time (in seconds) needed to complete a critical task on an assembly line was measured for a sample of 50 assemblies. These data are as follows:

| | | | | |
|---|---|---|---|---|
| 30.3 | 34.5 | 31.1 | 30.9 | 33.7 |
| 31.9 | 33.1 | 31.1 | 30.0 | 32.7 |
| 34.4 | 30.1 | 34.6 | 31.6 | 32.4 |
| 32.8 | 31.0 | 30.2 | 30.2 | 32.8 |
| 31.1 | 30.7 | 33.1 | 34.4 | 31.0 |
| 32.2 | 30.9 | 32.1 | 34.2 | 30.7 |
| 30.7 | 30.7 | 30.6 | 30.2 | 33.4 |
| 36.8 | 30.2 | 31.5 | 30.1 | 35.7 |
| 30.5 | 30.6 | 30.2 | 31.4 | 30.7 |
| 30.6 | 37.9 | 30.3 | 34.1 | 30.4 |

a. Draw a stem-and-leaf display.
b. Draw a histogram.
c. Describe the histogram.

**3.10** Xr03-10 A survey of individuals in a mall asked 60 people how many stores they will enter during this visit to the mall. The responses are listed here.

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

a. Draw a histogram.
b. Draw an ogive.
c. Describe your findings.

**3.11** Xr03-11 A survey asked 50 baseball fans to report the number of games they attended last year. The results are listed here. Use an appropriate graphical technique to present these data and describe what you have learned.

| | | | | |
|---|---|---|---|---|
| 5 | 15 | 14 | 7 | 8 |
| 16 | 26 | 6 | 15 | 23 |
| 11 | 15 | 6 | 4 | 7 |
| 8 | 19 | 16 | 9 | 9 |
| 8 | 7 | 10 | 5 | 8 |
| 8 | 6 | 6 | 21 | 10 |
| 5 | 24 | 5 | 28 | 9 |
| 11 | 20 | 24 | 5 | 13 |
| 14 | 9 | 25 | 10 | 24 |
| 10 | 18 | 22 | 12 | 17 |

**3.12** Xr03-12 To help determine the need for more golf courses, a survey was undertaken. A sample of 75 self-declared golfers was asked how many rounds of golf they played last year. These data are as follows:

| | | | | |
|---|---|---|---|---|
| 18 | 26 | 16 | 35 | 30 |
| 15 | 18 | 15 | 18 | 29 |
| 25 | 30 | 35 | 14 | 20 |
| 18 | 24 | 21 | 25 | 18 |
| 29 | 23 | 15 | 19 | 27 |
| 28 | 9 | 17 | 28 | 25 |
| 23 | 20 | 24 | 28 | 36 |
| 20 | 30 | 26 | 12 | 31 |
| 13 | 26 | 22 | 30 | 29 |
| 26 | 17 | 32 | 36 | 24 |
| 29 | 18 | 38 | 31 | 36 |
| 24 | 30 | 20 | 13 | 23 |
| 3 | 28 | 5 | 14 | 24 |
| 13 | 18 | 10 | 14 | 16 |
| 28 | 19 | 10 | 42 | 22 |

a. Draw a histogram.
b. Draw a stem-and-leaf display.
c. Draw an ogive.
d. Describe what you have learned.

*The following exercises require a computer and statistical software.*

**3.13** Xr03-13 The annual incomes for a sample of 200 first-year accountants were recorded. Summarize these data using a graphical method. Describe your results.

**3.14** Xr03-14 The real estate board in a suburb of Los Angeles wanted to investigate the distribution of the prices (in $ thousands) of homes sold during the past year.

a Draw a histogram.
b. Draw an ogive.
c. Draw a stem-and-leaf display (if your software allows it).
d. Describe what you have learned.

**3.15** Xr03-15 The number of customers entering a bank in the first hour of operation for each of the last 200 days was recorded. Use a graphical technique to extract information. Describe your findings.

**3.16** Xr03-16 The lengths of time (in minutes) to serve 420 customers at a local restaurant were recorded.

a. How many bins should a histogram of these data contain?
b. Draw a histogram using the number of bins specified in part (a).
c. Is the histogram symmetric or skewed?
d. Is the histogram bell shaped?

**3.17** Xr03-17 The marks of 320 students on an economics midterm test were recorded. Use a graphical technique to summarize these data. What does the graph tell you?

**3.18** Xr03-18 The lengths (in inches) of 150 newborn babies were recorded. Use whichever graphical technique you judge suitable to describe these data. What have you learned from the graph?

**3.19** Xr03-19 The number of copies made by an office copier was recorded for each of the past 75 days. Graph the data using a suitable technique. Describe what the graph tells you.

**3.20** Xr03-20 Each of a sample of 240 tomatoes grown with a new type of fertilizer was weighed (in ounces) and recorded. Draw a histogram and describe your findings.

**3.21** Xr03-21 The volume of water used by each of a sample of 350 households was measured (in gallons) and recorded. Use a suitable graphical statistical method to summarize the data. What does the graph tell you?

**3.22** Xr03-22 The number of books shipped out daily by Amazon.com was recorded for 100 days. Draw a histogram and describe your findings.

## APPLICATIONS in BANKING

© Comstock Images/Jupiterimages

### Credit Scorecards

**Credit scorecards** are used by banks and financial institutions to determine whether applicants will receive loans. The scorecard is the product of a statistical technique that converts questions about income, residence, and other variables into a score. The higher the score, the higher the probability that the applicant will repay. The scorecard is a formula produced by a statistical technique called *logistic regression*, which is available as an appendix on the Keller's website. For example, a scorecard may score age categories in the following way:

| | |
|---|---|
| Less than 25 | 20 points |
| 25 to 39 | 24 |
| 40 to 55 | 30 |
| Over 55 | 38 |

Other variables would be scored similarly. The sum for all variables would be the applicant's score. A cutoff score would be used to predict those who will repay and those who will default. Because no scorecard is perfect, it is possible to make two types of error: granting credit to those who will default and not lending money to those who would have repaid.

*(Continued)*

## EXERCISES

**3.23** Xr03-23 A small bank that had not yet used a scorecard wanted to determine whether a scorecard would be advantageous. The bank manager took a random sample of 300 loans that were granted and scored each on a scorecard borrowed from a similar bank. This scorecard is based on the responses supplied by the applicants to questions such as age, marital status, and household income. The cutoff is 650, which means that those scoring below are predicted to default and those scoring above are predicted to repay. Two hundred twenty of the loans were repaid, the rest were not. The scores of those who repaid and the scores of those who defaulted were recorded.

a. Use a graphical technique to present the scores of those who repaid.
b. Use a graphical technique to present the scores of those who defaulted.
c. What have you learned about the scorecard?

**3.24** Xr03-24 Refer to Exercise 3.23. The bank decided to try another scorecard, this one based not on the responses of the applicants but on credit bureau reports, which list problems such as late payments and previous defaults. The scores using the new scorecard of those who repaid and the scores of those who did not repay were recorded. The cutoff score is 650.

a. Use a graphical technique to present the scores of those who repaid.
b. Use a graphical technique to present the scores of those who defaulted.
c. What have you learned about the scorecard?
d. Compare the results of this exercise with those of Exercise 3.23. Which scorecard appears to be better?

## GENERAL SOCIAL SURVEY EXERCISES

**3.25** The GSS asked respondents to specify their highest year of school completed (EDUC).

a. Is this type of data interval, ordinal, or nominal?
b. GSS2008* Use a graphical technique to present these data for the 2008 survey.
c. Briefly describe your results.

**3.26** GSS2008* Graphically display the results of the GSS 2008 question, On average days how many hours do you spend watching television (TVHOURS)? Briefly describe what you have discovered.

**3.27** GSS2008* Employ a graphical technique to present the ages (AGE) of the respondents in the 2008 survey. Describe your results.

**3.28** GSS2008* The survey in 2008 asked "If working, full- or part-time, how many hours did you work last week at all jobs (HRS)?" Summarize these data with a graphical technique.

## 3.2 DESCRIBING TIME-SERIES DATA

Besides classifying data by type, we can also classify them according to whether the observations are measured at the same time or whether they represent measurements at successive points in time. The former are called **cross-sectional data**, and the latter **time-series data**.

The techniques described in Section 3.1 are applied to cross-sectional data. All the data for Example 3.1 were probably determined within the same day. We can probably say the same thing for Examples 3.2 to 3.4.

To give another example, consider a real estate consultant who feels that the selling price of a house is a function of its size, age, and lot size. To estimate the specific form of the function, she samples, say, 100 homes recently sold and records the price, size, age, and lot size for each home. These data are cross-sectional: They all are observations at the same point in time. The real estate consultant is also working on a separate project to forecast the monthly housing starts in the northeastern United States over the next year. To do so, she collects the monthly housing starts in this region for each of the past 5 years. These 60 values (housing starts) represent time-series data because they are observations taken over time.

Note that the original data may be interval or nominal. All the illustrations above deal with interval data. A time series can also list the frequencies and relative frequencies of a nominal variable over a number of time periods. For example, a brand-preference survey asks consumers to identify their favorite brand. These data are nominal. If we repeat the survey once a month for several years, the proportion of consumers who prefer a certain company's product each month would constitute a time series.

## Line Chart

Time-series data are often graphically depicted on a **line chart**, which is a plot of the variable over time. It is created by plotting the value of the variable on the vertical axis and the time periods on the horizontal axis.

The chapter-opening example addresses the issue of the relationship between the price of gasoline and the price of oil. We will introduce the technique we need to answer the question in Section 3.3. Another question arises: Is the recent price of gasoline high compared to the past prices?

## EXAMPLE 3.5 Price of Gasoline

DATA
Xm03-05

We recorded the monthly average retail price of gasoline (in cents per gallon) since January 1976. Some of these data are displayed below. Draw a line chart to describe these data and briefly describe the results.

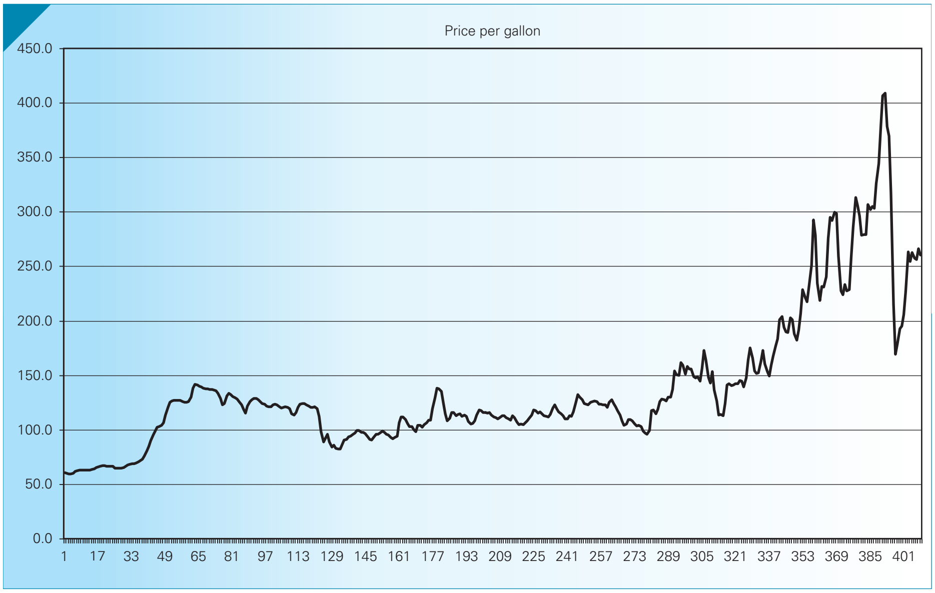
| Year | Month | Price per gallon |
|---|---|---|
| 1976 | 1 | 60.5 |
| 1976 | 2 | 60.0 |
| 1976 | 3 | 59.4 |
| 1976 | 4 | 59.2 |
| 1976 | 5 | 60.0 |
| 1976 | 6 | 61.6 |
| 1976 | 7 | 62.3 |
| 1976 | 8 | 62.8 |
| 1976 | 9 | 63.0 |
| 1976 | 10 | 62.9 |
| 1976 | 11 | 62.9 |
| 1976 | 12 | 62.6 |
| 2009 | 1 | 178.7 |
| 2009 | 2 | 192.8 |
| 2009 | 3 | 194.9 |
| 2009 | 4 | 205.6 |

| | | |
|---|---|---|
| 2009 | 5 | 226.5 |
| 2009 | 6 | 263.1 |
| 2009 | 7 | 254.3 |
| 2009 | 8 | 262.7 |
| 2009 | 9 | 257.4 |
| 2009 | 10 | 256.1 |
| 2009 | 11 | 266.0 |
| 2009 | 12 | 260.0 |

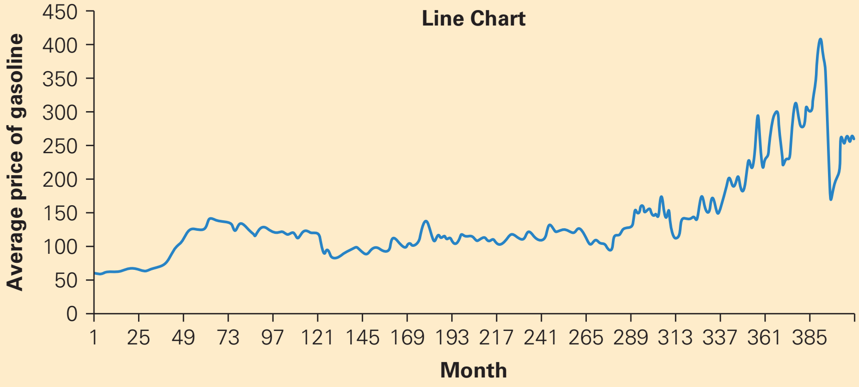
## SOLUTION

Here are the line charts produced manually, and by Excel and Minitab.

FIGURE **3.10** Line Chart for Example 3.5



## EXCEL

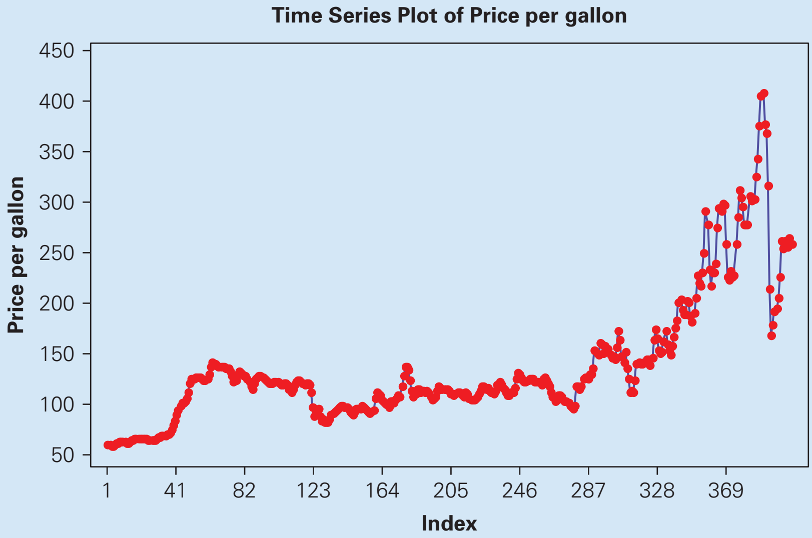


*INSTRUCTIONS*

1. Type or import the data into one column. (Open Xm03-05.)
2. Highlight the column of data. Click **Insert, Line,** and the first **2-D Line**. Click **Chart Tools** and **Layout** to make whatever changes you wish.

You can draw two or more line charts (for two or more variables) by highlighting all columns of data you wish to graph.

## MINITAB



*INSTRUCTIONS*

1. Type or import the data into one column. (Open Xm03-05.)
2. Click **Graph** and **Time Series Plot . . . .** Click **Simple.**
3. In the **series** box type or use the **Select** button to specify the variable (Price). Click **Time/Scale**.
4. Click the **Time** tab, and under **Time Scale** click **Index**.

## INTERPRET

The price of gasoline rose from about $.60 to more than a dollar in the late 1970s (months 1 to 49), fluctuated between $.90 and $1.50 until 2000 (months 49 to 289), then generally rose with large fluctuations (months 289 to 380), then declined sharply before rallying in the last 10 months.

## APPLICATIONS in ECONOMICS

### Measuring Inflation: Consumer Price Index*

Inflation is the increase in the prices for goods and services. In most countries, inflation is measured using the Consumer Price Index (CPI). The Consumer Price Index works with a basket of some 300 goods and services in the United States (and a similar number in other countries), including such diverse items as food, housing, clothing, transportation, health, and recreation. The basket is defined for the "typical" or "average" middle-income family, and the set of items and their weights are revised periodically (every 10 years in the United States and every 7 years in Canada).

Prices for each item in this basket are computed on a monthly basis and the CPI is computed from these prices. Here is how it works. We start by setting a period of time as the base. In the United States the base is the years 1982–1984. Suppose that the basket of goods and services cost $1,000 during this period. Thus, the base is $1,000, and the CPI is set at 100. Suppose that in the next month (January 1985) the price increases to $1,010. The CPI for January 1985 is calculated in the following way:

$$\text{CPI(January 1985)} = \frac{1{,}010}{1{,}000} \times 100 = 101$$

If the price increases to $1,050 in the next month, the CPI is

$$\text{CPI(February 1985)} = \frac{1{,}050}{1{,}000} \times 100 = 105$$

The CPI, despite never really being intended to serve as the official measure of inflation, has come to be interpreted in this way by the general public. Pension-plan payments, old-age Social Security, and some labor contracts are automatically linked to the CPI and automatically indexed (so it is claimed) to the level of inflation. Despite its flaws, the CPI is used in numerous applications. One application involves adjusting prices by removing the effect of inflation, making it possible to track the "real" changes in a time series of prices.

In Example 3.5, the figures shown are the actual prices measured in what are called *current* dollars. To remove the effect of inflation, we divide the monthly prices by the CPI for that month and multiply by 100. These prices are then measured in *constant* 1982–1984 dollars. This makes it easier to see what has happened to the prices of the goods and services of interest.

We created two data sets to help you calculate prices in constant 1982–1984 dollars. File Ch03:\\CPI-Annual and Ch03:\\CPI-Monthly list the values of the CPI where 1982–1984 is set at 100 for annual values and monthly values, respectively.

*Keller's website Appendix Index Numbers, located at www.cengage.com/bstatistics/keller, describes index numbers and how they are calculated.

## EXAMPLE 3.6 Price of Gasoline in 1982–1984 Constant Dollars

DATA
Xm03-06

Remove the effect of inflation in Example 3.5 to determine whether gasoline prices are higher than they have been in the past.

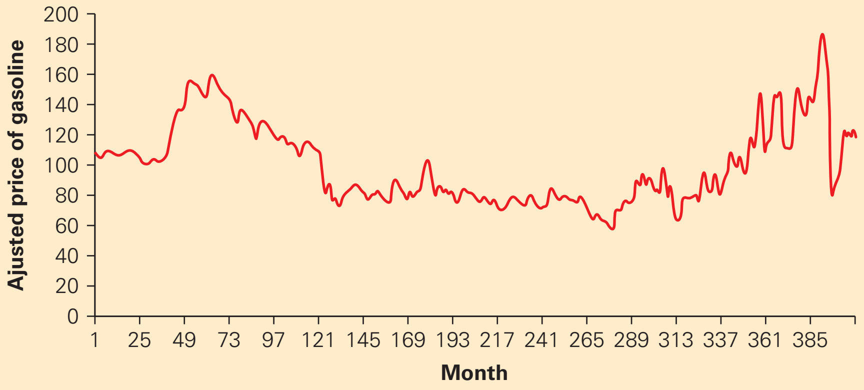
### SOLUTION

Here are the 1976 and 2009 average monthly prices of gasoline, the CPI, and the adjusted prices.

The adjusted figures for all months were used in the line chart produced by Excel. Minitab's chart is similar.

| Year | Month | Price per gallon | CPI | Adjusted price |
|---|---|---|---|---|
| 1976 | 1 | 60.5 | 55.8 | 108.4 |
| 1976 | 2 | 60.0 | 55.9 | 107.3 |
| 1976 | 3 | 59.4 | 56.0 | 106.1 |
| 1976 | 4 | 59.2 | 56.1 | 105.5 |
| 1976 | 5 | 60.0 | 56.4 | 106.4 |
| 1976 | 6 | 61.6 | 56.7 | 108.6 |
| 1976 | 7 | 62.3 | 57.0 | 109.3 |
| 1976 | 8 | 62.8 | 57.3 | 109.6 |
| 1976 | 9 | 63.0 | 57.6 | 109.4 |
| 1976 | 10 | 62.9 | 57.9 | 108.6 |
| 1976 | 11 | 62.9 | 58.1 | 108.3 |
| 1976 | 12 | 62.6 | 58.4 | 107.2 |
| 2009 | 1 | 178.7 | 212.17 | 84.2 |
| 2009 | 2 | 192.8 | 213.01 | 90.5 |
| 2009 | 3 | 194.9 | 212.71 | 91.6 |
| 2009 | 4 | 205.6 | 212.67 | 96.7 |
| 2009 | 5 | 226.5 | 212.88 | 106.4 |
| 2009 | 6 | 263.1 | 214.46 | 122.7 |
| 2009 | 7 | 254.3 | 214.47 | 118.6 |
| 2009 | 8 | 262.7 | 215.43 | 121.9 |
| 2009 | 9 | 257.4 | 215.79 | 119.3 |
| 2009 | 10 | 256.1 | 216.39 | 118.4 |
| 2009 | 11 | 266.0 | 217.25 | 122.4 |
| 2009 | 12 | 260.0 | 217.54 | 119.5 |

### EXCEL



### INTERPRET

Using constant 1982–1984 dollars, we can see that the average price of a gallon of gasoline hit its peak in the middle of 2008 (month 390). From there it dropped rapidly and in late 2009 was about equal to the adjusted price in 1976.

There are two more factors to consider in judging whether the price of gasoline is high. The first is distance traveled and the second is fuel consumption. Exercise 3.41 deals with this issue.

## EXERCISES

**3.29** Xr03-29 The fees television broadcasters pay to cover the summer Olympic Games has become the largest source of revenue for the host country. Below we list the year, city, and revenue in millions of U.S. dollars paid by television broadcasters around the world. Draw a chart to describe these prices paid by the networks.

| Year | City | Broadcast Revenue |
|---|---|---|
| 1960 | Rome | 1.2 |
| 1964 | Tokyo | 1.6 |
| 1968 | Mexico City | 9.8 |
| 1972 | Munich | 17.8 |
| 1976 | Montreal | 34.9 |
| 1980 | Moscow | 88.0 |
| 1984 | Los Angeles | 266.9 |
| 1988 | Seoul | 402.6 |
| 1992 | Barcelona | 636.1 |
| 1996 | Atlanta | 898.3 |
| 2000 | Sydney | 1331.6 |
| 2004 | Athens | 1494.0 |
| 2008 | Beijing | 1737.0 |

*Source:* Bloomberg News.

**3.30** Xr03-30 The number of females enlisted in the United States Army from 1971 to 2007 are listed here. Draw a line chart, and describe what the chart tells you.

| Year | Females enlisted | Year | Females enlisted |
|---|---|---|---|
| 1971 | 11.8 | 1990 | 71.2 |
| 1972 | 12.3 | 1991 | 67.8 |
| 1973 | 16.5 | 1992 | 61.7 |
| 1974 | 26.3 | 1993 | 60.2 |
| 1975 | 37.7 | 1994 | 59.0 |
| 1976 | 43.8 | 1995 | 57.3 |
| 1977 | 46.1 | 1996 | 59.0 |
| 1978 | 50.5 | 1997 | 62.4 |
| 1979 | 55.2 | 1998 | 61.4 |
| 1980 | 61.7 | 1999 | 61.5 |
| 1981 | 65.3 | 2000 | 62.9 |
| 1982 | 64.1 | 2001 | 63.4 |
| 1983 | 66.5 | 2002 | 63.2 |
| 1984 | 67.1 | 2003 | 63.5 |
| 1985 | 68.4 | 2004 | 61.0 |
| 1986 | 69.7 | 2005 | 57.9 |
| 1987 | 71.6 | 2006 | 58.5 |
| 1988 | 72.0 | 2007 | 58.8 |
| 1989 | 74.3 | | |

*(Continued)*

*Source: Statistical Abstract of the United States, 2009*, Table 494.

**3.31** Xr03-31 The United States spends more money on health care than any other country. To gauge how fast costs are increasing, the following table was produced, listing the total health-care expenditures in the United States annually for 1981 to 2006 (costs are in $billions).

a. Graphically present these data.
b. Use the data in CPI-Annual to remove the effect of inflation. Graph the results and describe your findings.

| Year | Health expenditures | Year | Health expenditures |
|---|---|---|---|
| 1981 | 294 | 1994 | 962 |
| 1982 | 331 | 1995 | 1017 |
| 1983 | 365 | 1996 | 1069 |
| 1984 | 402 | 1997 | 1125 |
| 1985 | 440 | 1998 | 1191 |
| 1986 | 472 | 1999 | 1265 |
| 1987 | 513 | 2000 | 1353 |
| 1988 | 574 | 2001 | 1470 |
| 1989 | 639 | 2002 | 1603 |
| 1990 | 714 | 2003 | 1732 |
| 1991 | 782 | 2004 | 1852 |
| 1992 | 849 | 2005 | 1973 |
| 1993 | 913 | 2006 | 2106 |

*Source: Statistical Abstract of the United States, 2009*, Table 124.

**3.32** Xr03-32 The number of earned degrees (thousands) for males and females is listed below for the years 1987 to 2006. Graph both sets of data. What do the graphs tell you?

| Year | Female | Male |
|---|---|---|
| 1987 | 941 | 882 |
| 1988 | 954 | 881 |
| 1989 | 986 | 887 |
| 1990 | 1035 | 905 |
| 1991 | 1097 | 928 |
| 1992 | 1147 | 961 |
| 1993 | 1182 | 985 |
| 1994 | 1211 | 995 |
| 1995 | 1223 | 995 |
| 1996 | 1255 | 993 |
| 1997 | 1290 | 998 |
| 1998 | 1304 | 994 |
| 1999 | 1330 | 993 |
| 2000 | 1369 | 1016 |
| 2001 | 1391 | 1025 |
| 2002 | 1441 | 1053 |
| 2003 | 1517 | 1104 |
| 2004 | 1603 | 1152 |
| 2005 | 1666 | 1185 |
| 2006 | 1725 | 1211 |

*Source*: U.S. National Center for Education Statistics, *Statistical Abstract of the United States, 2009*, Table 288.

**3.33** Xr03-33 The number of property crimes (burglary, larceny, theft, car theft) (in thousands) for the years 1992 to 2006 are listed next. Draw a line chart and interpret the results.

| Year | Crimes | Year | Crimes |
|---|---|---|---|
| 1992 | 12506 | 2000 | 10183 |
| 1993 | 12219 | 2001 | 10437 |
| 1994 | 12132 | 2002 | 10455 |
| 1995 | 12064 | 2003 | 10443 |
| 1996 | 11805 | 2004 | 10319 |
| 1997 | 11558 | 2005 | 10175 |
| 1998 | 10952 | 2006 | 9984 |
| 1999 | 10208 | | |

*Source:* U.S. Federal Bureau of Investigation *Statistical Abstract of the United States, 2009*, Table 295.

**3.34** Xr03-34 Refer to Exercise 3.33. Another way of measuring the number of property crimes is to calculate the number of crimes per 100,000 of population. This allows us to remove the effect of the increasing population. Graph these data and interpret your findings.

| Year | Crimes | Year | Crimes |
|---|---|---|---|
| 1992 | 4868 | 2000 | 3606 |
| 1993 | 4695 | 2001 | 3658 |
| 1994 | 4605 | 2002 | 3628 |
| 1995 | 4526 | 2003 | 3589 |
| 1996 | 4378 | 2004 | 3515 |
| 1997 | 4235 | 2005 | 3434 |
| 1998 | 3966 | 2006 | 3337 |
| 1999 | 3655 | | |

**3.35** Xr03-35 The gross national product (GNP) is the sum total of the economic output of a the citizens (nationals) of a country. It is an important measure of the wealth of a country. The following table lists the year and the GNP in billions of current dollars for the United States.

a. Graph the GNP. What have you learned?
b. Use the data in CPI-Annual to compute the per capita GNP in constant 1982–1984 dollars. Graph the results and describe your findings.

| Year | GNP | Year | GNP |
|---|---|---|---|
| 1980 | 2822 | 1995 | 7444 |
| 1981 | 3160 | 1996 | 7870 |
| 1982 | 3290 | 1997 | 8356 |
| 1983 | 3572 | 1998 | 8811 |
| 1984 | 3967 | 1999 | 9381 |
| 1985 | 4244 | 2000 | 9989 |
| 1986 | 4478 | 2001 | 10338 |
| 1987 | 4754 | 2002 | 10691 |
| 1988 | 5124 | 2003 | 11211 |
| 1989 | 5508 | 2004 | 11959 |
| 1990 | 5835 | 2005 | 12736 |
| 1991 | 6022 | 2006 | 13471 |
| 1992 | 6371 | 2007 | 14193 |
| 1993 | 6699 | 2008 | 14583 |
| 1994 | 7109 | | |

*Source*: U.S. Bureau of Economic Activity.

**3.36** Xr03-36 The average daily U.S. oil consumption and production (thousands of barrels) is shown for the years 1973 to 2007. Use a graphical technique to describe these figures. What does the graph tell you?

| Year | 1973 | 1974 | 1975 | 1976 | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 |
|---|---|---|---|---|---|---|---|---|---|---|
| Consumption | 17,318 | 16,655 | 16,323 | 17,460 | 18,443 | 18,857 | 18,527 | 17,060 | 16,061 | 15,301 |
| Production | 9,209 | 8,776 | 8,376 | 8,132 | 8,245 | 8,706 | 8,551 | 8,597 | 8,572 | 8,649 |
| **Year** | **1983** | **1984** | **1985** | **1986** | **1987** | **1988** | **1989** | **1990** | **1991** | **1992** |
| Consumption | 15,228 | 15,722 | 15,726 | 16,277 | 16,666 | 17,284 | 17,327 | 16,988 | 16,710 | 17,031 |
| Production | 8,689 | 8,879 | 8,972 | 8,683 | 8,349 | 8,140 | 7,615 | 7,356 | 7,418 | 7,172 |
| **Year** | **1993** | **1994** | **1995** | **1996** | **1997** | **1998** | **1999** | **2000** | **2001** | **2002** |
| Consumption | 17,328 | 17,721 | 17,730 | 18,308 | 18,618 | 18,913 | 19,515 | 19,699 | 19,647 | 19,758 |
| Production | 6,847 | 6,662 | 6,561 | 6,465 | 6,452 | 6,253 | 5,882 | 5,822 | 5,801 | 5,746 |
| **Year** | **2003** | **2004** | **2005** | **2006** | **2007** | | | | | |
| Consumption | 20,034 | 20,731 | 20,799 | 20,800 | 20,680 | | | | | |
| Production | 5,682 | 5,419 | 5,179 | 5,102 | 5,065 | | | | | |

*Source*: U.S. Department of Energy: Monthly Energy Review.

**3.37** Xr03-37 Has housing been a hedge against inflation in the last 20 years? To answer this question, we produced the following table, which lists the average selling price of one-family homes in all of the United States, the Northeast, Midwest, South, and West for the years 1988 to 2007, as well as the annual CPI. For the entire country and for each area, use a graphical technique to determine whether housing prices stayed ahead of inflation.

| Year | All | Northeast | Midwest | South | West | CPI |
|---|---|---|---|---|---|---|
| 1988 | 89,300 | 143,000 | 68,400 | 82,200 | 124,900 | 118.3 |
| 1989 | 94,600 | 147,700 | 73,100 | 85,600 | 138,400 | 124.0 |
| 1990 | 97,300 | 146,200 | 76,700 | 86,300 | 141,200 | 130.7 |
| 1991 | 102,700 | 149,300 | 81,000 | 89,800 | 147,400 | 136.2 |
| 1992 | 105,500 | 149,000 | 84,600 | 92,900 | 143,300 | 140.3 |
| 1993 | 109,100 | 149,300 | 87,600 | 95,800 | 144,400 | 144.5 |
| 1994 | 113,500 | 149,300 | 90,900 | 97,200 | 151,900 | 148.2 |
| 1995 | 117,000 | 146,500 | 96,500 | 99,200 | 153,600 | 152.4 |
| 1996 | 122,600 | 147,800 | 102,800 | 105,000 | 160,200 | 156.9 |
| 1997 | 129,000 | 152,400 | 108,900 | 111,300 | 169,000 | 160.5 |
| 1998 | 136,000 | 157,100 | 116,300 | 118,000 | 179,500 | 163.0 |
| 1999 | 141,200 | 160,700 | 121,600 | 122,100 | 189,400 | 166.6 |
| 2000 | 147,300 | 161,200 | 125,600 | 130,300 | 199,200 | 172.2 |
| 2001 | 156,600 | 169,400 | 132,300 | 139,600 | 211,700 | 177.1 |
| 2002 | 167,600 | 190,100 | 138,300 | 149,700 | 234,300 | 179.9 |
| 2003 | 180,200 | 220,300 | 143,700 | 159,700 | 254,700 | 184.0 |
| 2004 | 195,200 | 254,400 | 151,500 | 171,800 | 289,100 | 188.9 |
| 2005 | 219,000 | 281,600 | 168,300 | 181,100 | 340,300 | 195.3 |
| 2006 | 221,900 | 280,300 | 164,800 | 183,700 | 350,500 | 201.6 |
| 2007 | 217,900 | 288,100 | 161,400 | 178,800 | 342,500 | 207.3 |

*Source*: *Statistical Abstract of the United States, 2009*, Table 935.

**3.38** Xr03-38 How has the size of government changed? To help answer this question, we recorded the U.S. federal budget receipts and outlays (billions of current dollars) for the years 1980 to 2007.

a. Use a graphical technique to describe the receipts and outlays of the annual U.S. federal government budgets since 1980.

b. Calculate the difference between receipts and outlays. If the difference is positive the result is a surplus; if the difference is negative the result is a deficit. Graph the surplus/deficit variable and describe the results.

| Year | 1980 | 1981 | 1982 | 1983 | 1984 | 1985 | 1986 | 1987 | 1988 | 1989 |
|---|---|---|---|---|---|---|---|---|---|---|
| Receipts | 517.1 | 599.3 | 617.8 | 600.6 | 666.5 | 734.1 | 769.2 | 854.4 | 909.3 | 991.2 |
| Outlays | 590.9 | 678.2 | 745.7 | 808.4 | 851.9 | 946.4 | 990.4 | 1,004.1 | 1,064.5 | 1,143.6 |
| **Year** | **1990** | **1991** | **1992** | **1993** | **1994** | **1995** | **1996** | **1997** | **1998** | **1999** |
| Receipts | 1,032.0 | 1,055.0 | 1,091.3 | 1,154.4 | 1,258.6 | 1,351.8 | 1,453.1 | 1,579.3 | 1,721.8 | 1,827.5 |
| Outlays | 1,253.2 | 1,324.4 | 1,381.7 | 1,409.5 | 1,461.9 | 1,515.8 | 1,560.5 | 1,601.3 | 1,652.6 | 1,701.9 |
| **Year** | **2000** | **2001** | **2002** | **2003** | **2004** | **2005** | **2006** | **2007** | | |
| Receipts | 2,025.2 | 1,991.2 | 1,853.2 | 1,782.3 | 1,880.1 | 2,153.9 | 2,407.3 | 2,568.2 | | |
| Outlays | 1,789.1 | 1,863.9 | 2,011.0 | 2,159.9 | 2,293.0 | 2,472.2 | 2,655.4 | 2,730.2 | | |

*Source*: *Statistical Abstract of the United States, 2009*, Table 451.

**3.39** Refer to Exercise 3.38. Another way of judging the size of budget surplus/deficits is to calculate the deficit as a percentage of GNP. Use the data in Exercises 3.35 and 3.38 to calculate this variable and use a graphical technique to display the results.

**3.40** Repeat Exercise 3.39 using the CPI-Annual file to convert all amounts to constant 1982–1984 dollars. Draw a line chart to show these data.

**3.41** Xr03-41 Refer to Example 3.5. The following table lists the average gasoline consumption in miles per gallon (MPG) and the average distance (thousands of miles) driven by cars in each of the years 1980 to 2006. (The file contains the average price for each year, the annual CPI, fuel consumption and distance (thousands).) For each year calculate the inflation-adjusted cost per year of driving. Use a graphical technique to present the results.

| Year | 1980 | 1981 | 1982 | 1983 | 1984 | 1985 | 1986 | 1987 | 1988 | 1989 |
|---|---|---|---|---|---|---|---|---|---|---|
| MPG | 13.3 | 13.6 | 14.1 | 14.2 | 14.5 | 14.6 | 14.7 | 15.1 | 15.6 | 15.9 |
| Distance | 9.5 | 9.5 | 9.6 | 9.8 | 10.0 | 10.0 | 10.1 | 10.5 | 10.7 | 10.9 |
| **Year** | **1990** | **1991** | **1992** | **1993** | **1994** | **1995** | **1996** | **1997** | **1998** | **1999** |
| MPG | 16.4 | 16.9 | 16.9 | 16.7 | 16.7 | 16.8 | 16.9 | 17.0 | 16.9 | 17.7 |
| Distance | 11.1 | 11.3 | 11.6 | 11.6 | 11.7 | 11.8 | 11.8 | 12.1 | 12.2 | 12.2 |
| **Year** | **2000** | **2001** | **2002** | **2003** | **2004** | **2005** | **2006** | | | |
| MPG | 16.9 | 17.1 | 16.9 | 17.0 | 17.1 | 17.2 | 17.1 | | | |
| Distance | 12.2 | 11.9 | 12.2 | 12.2 | 12.2 | 12.1 | 12.4 | | | |

*Source: Statistical Abstract of the United States, 2009*, Tables 1061 and 1062.

*The following exercises require a computer and software.*

**3.42** Xr03-42 The monthly value of U.S. exports to Canada (in $millions) and imports from Canada from 1985 to 2009 were recorded. (*Source*: Federal Reserve Economic Data.)

a. Draw a line chart of U.S. exports to Canada.
b. Draw a line chart of U.S. imports from Canada.
c. Calculate the trade balance and draw a line chart.
d. What do all the charts reveal?

**3.43** Xr03-43 The monthly value of U.S. exports to Japan (in $ millions) and imports from Japan from 1985 to 2009 were recorded. (*Source*: Federal Reserve Economic Data.)

a. Draw a line chart of U.S. exports to Japan.
b. Draw a line chart of U.S. imports from Japan.
c. Calculate the trade balance and draw a line chart.
d. What do all the charts reveal?

**3.44** Xr03-44 The value of the Canadian dollar in U.S. dollars was recorded monthly for the period 1971 to 2009. Draw a graph of these figures and interpret your findings.

**3.45** Xr03-45 The value of the Japanese yen in U.S. dollars was recorded monthly for the period 1971 to 2009. Draw a graph of these figures and interpret your findings.

**3.46** Xr03-46 The Dow Jones Industrial Average was recorded monthly for the years 1950 to 2009. Use a graph to describe these numbers. (*Source: Wall Street Journal.*)

**3.47** Refer to Exercise 3.46. Use the CPI-monthly file to measure the Dow Jones Industrial Average in 1982–1984 constant dollars. What have you learned?

## 3.3 Describing the Relationship between Two Interval Variables

Statistics practitioners frequently need to know how two interval variables are related. For example, financial analysts need to understand how the returns of individual stocks are related to the returns of the entire market. Marketing managers need to understand the relationship between sales and advertising. Economists develop statistical techniques to describe the relationship between such variables as unemployment rates and inflation. The technique is called a **scatter diagram**.

To draw a scatter diagram, we need data for two variables. In applications where one variable depends to some degree on the other variable, we label the dependent variable *Y* and the other, called the *independent variable*, *X*. For example, an individual's income depends somewhat on the number of years of education. Accordingly, we identify income as the dependent variable and label it *Y*, and we identify years of education as the independent variable and label it *X*. In other cases where no dependency is evident, we label the variables arbitrarily.

### Example 3.7

DATA Xm03-07

### Analyzing the Relationship between Price and Size of Houses

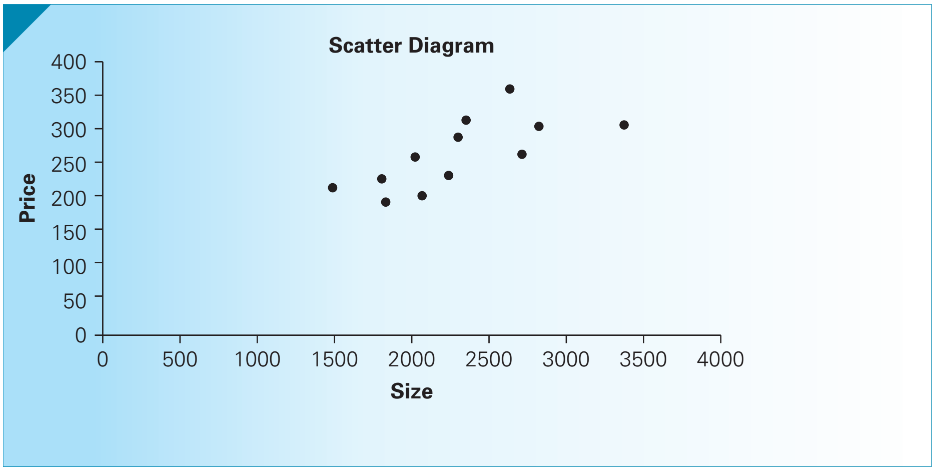
A real estate agent wanted to know to what extent the selling price of a home is related to its size. To acquire this information, he took a sample of 12 homes that had recently sold, recording the price in thousands of dollars and the size in square feet. These data are listed in the accompanying table. Use a graphical technique to describe the relationship between size and price.

| Size (ft$^2$) | Price ($1,000) |
|---|---|
| 2,354 | 315 |
| 1,807 | 229 |
| 2,637 | 355 |
| 2,024 | 261 |
| 2,241 | 234 |
| 1,489 | 216 |
| 3,377 | 308 |
| 2,825 | 306 |
| 2,302 | 289 |
| 2,068 | 204 |
| 2,715 | 265 |
| 1,833 | 195 |

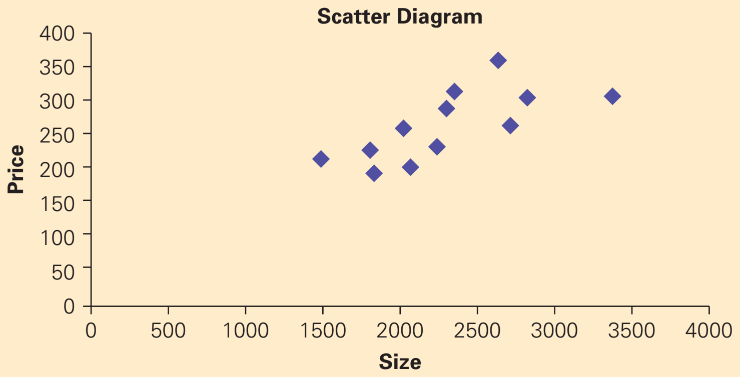
## SOLUTION

Using the guideline just stated, we label the price of the house $Y$ (dependent variable) and the size $X$ (independent variable). Figure 3.11 depicts the scatter diagram.

FIGURE **3.11** Scatter Diagram for Example 3.7



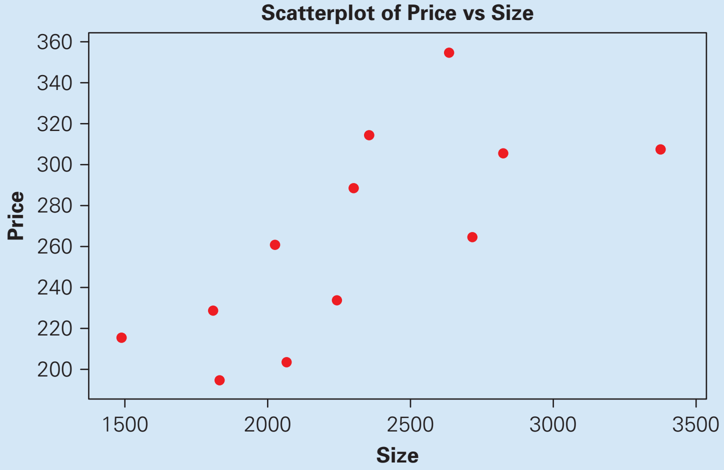
## EXCEL



### INSTRUCTIONS

1. Type or import the data into two adjacent columns. Store variable $X$ in the first column and variable $Y$ in the next column. (Open Xm03-07.)
2. Click **Insert** and **Scatter**.
3. To make cosmetic changes, click **Chart Tools** and **Layout**. (We chose to add titles and remove the gridlines.) If you wish to change the scale, click **Axes, Primary Horizontal Axis** or **Primary Vertical Axis, More Primary Horizontal** or **Vertical Axis Options . . .** , and make the changes you want.

## MINITAB



*INSTRUCTIONS*

1. Type or import the data into two columns. (Open Xm03-07.)
2. Click **Graph** and **Scatterplot. . . .**
3. Click Simple.
4. Type or use the **Select** button to specify the variable to appear on the *Y*-axis (Price) and the *X*-axis (Size).

## INTERPRET

The scatter diagram reveals that, in general, the greater the size of the house, the greater the price. However, there are other variables that determine price. Further analysis may reveal what these other variables are.

## Patterns of Scatter Diagrams

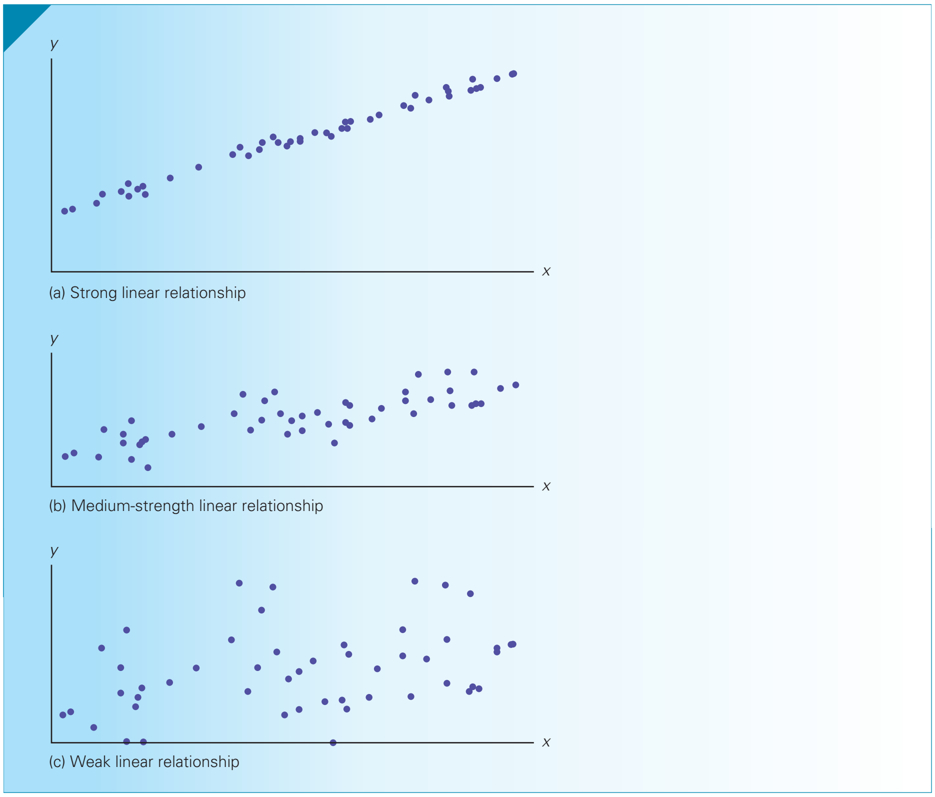
As was the case with histograms, we frequently need to describe verbally how two variables are related. The two most important characteristics are the strength and direction of the linear relationship.

## Linearity

To determine the strength of the linear relationship, we draw a straight line through the points in such a way that the line represents the relationship. If most of the points fall close to the line, we say that there is a **linear relationship**. If most of the points appear to be scattered randomly with only a semblance of a straight line, there is no, or at best, a weak linear relationship. Figure 3.12 depicts several scatter diagrams that exhibit various levels of linearity.

In drawing the line freehand, we would attempt to draw it so that it passes through the middle of the data. Unfortunately, different people drawing a straight line through the same set of data will produce somewhat different lines. Fortunately, statisticians have produced an objective way to draw the straight line. The method is called the *least squares method*, and it will be presented in Chapter 4 and employed in Chapters 16, 17, and 18.

FIGURE **3.12** Scatter Diagrams Depicting Linearity



(a) Strong linear relationship

(b) Medium-strength linear relationship

(c) Weak linear relationship

Note that there may well be some other type of relationship, such as a quadratic or exponential one.

## Direction

In general, if one variable increases when the other does, we say that there is a **positive linear relationship**. When the two variables tend to move in opposite directions, we describe the nature of their association as a **negative linear relationship**. (The terms *positive* and *negative* will be explained in Chapter 4.) See Figure 3.13 for examples of scatter diagrams depicting a positive linear relationship, a negative linear relationship, no relationship, and a nonlinear relationship.

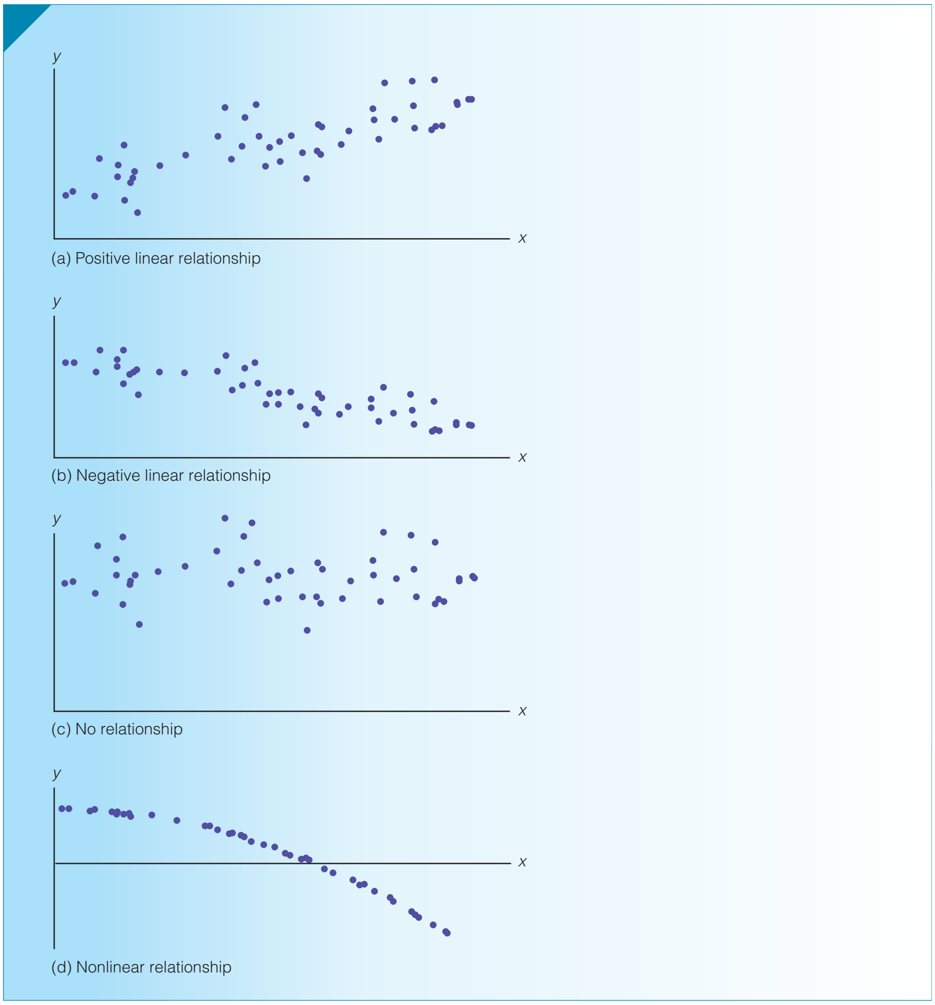
## Interpreting a Strong Linear Relationship

In interpreting the results of a scatter diagram it is important to understand that if two variables are linearly related it does not mean that one is causing the other. In fact, we can never conclude that one variable causes another variable. We can express this more eloquently as

Correlation is not causation.

Now that we know what to look for, we can answer the chapter-opening example.

FIGURE 3.13 Scatter Diagrams Describing Direction



(a) Positive linear relationship

(b) Negative linear relationship

(c) No relationship

(d) Nonlinear relationship

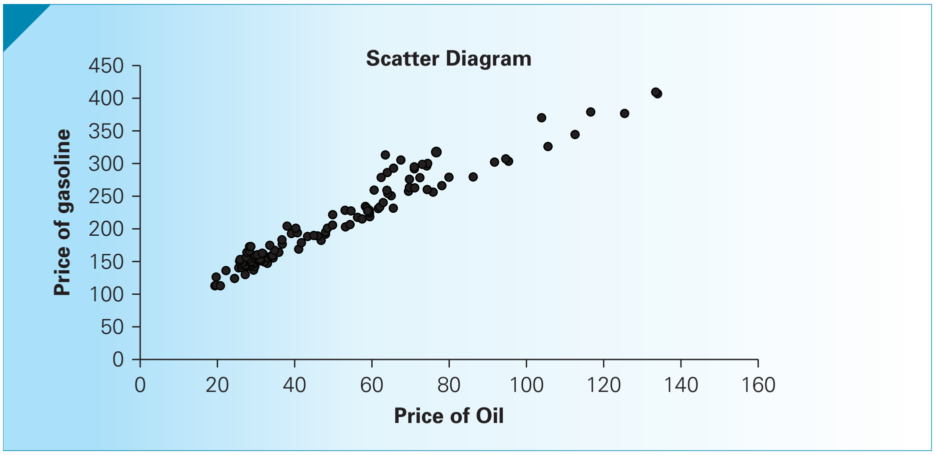
## Were Oil Companies Gouging Customers 2000–2009: Solution

© Comstock Images/Jupiterimages

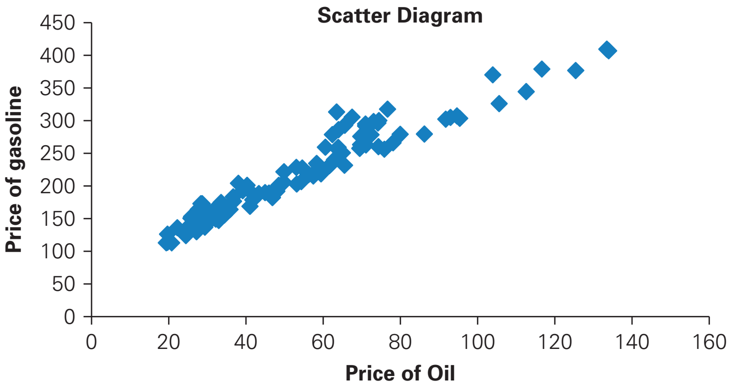
To determine whether drivers' perceptions that oil companies were gouging consumers, we need to determine whether and to what extent the two variables are related. The appropriate statistical technique is the scatter diagram.

We label the price of gasoline $Y$ and the price of oil $X$. Figure 3.14 displays the scatter diagram.

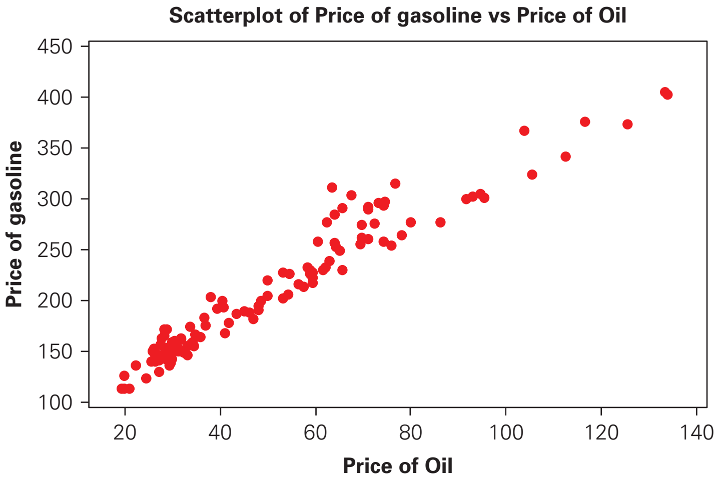
FIGURE 3.14 Scatter Diagram for Chapter-Opening Example



EXCEL



MINITAB



*(Continued)*

### INTERPRET

The scatter diagram reveals that the two prices are strongly related linearly. When the price of oil was below $40, the relationship between the two was stronger than when the price of oil exceeded $40.

We close this section by reviewing the factors that identify the use of the scatter diagram.

**Factors That Identify When to Use a Scatter Diagram**

1. **Objective**: Describe the relationship between two variables
2. **Data type**: Interval

## Exercises

**3.48** Xr03-48 Between 2002 and 2005, there was a decrease in movie attendance. There are several reasons for this decline. One reason may be the increase in DVD sales. The percentage of U.S. homes with DVD players and the movie attendance (billions) in the United States for the years 2000 to 2005 are shown next. Use a graphical technique to describe the relationship between these variables.

| Year | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 |
|---|---|---|---|---|---|---|
| DVD percentage | 12 | 23 | 37 | 42 | 59 | 74 |
| Movie attendance | 1.41 | 1.49 | 1.63 | 1.58 | 1.53 | 1.40 |

*Sources*: Northern Technology & Telecom Research and Motion Picture Association.

**3.49** Xr03-49 Because inflation reduces the purchasing power of the dollar, investors seek investments that will provide higher returns when inflation is higher. It is frequently stated that common stocks provide just such a hedge against inflation. The annual percentage rates of return on common stock and annual inflation rates for a recent 10-year period are listed here.

| Year | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Returns | 25 | 8 | 6 | 11 | 21 | −15 | 12 | −1 | 33 | 0 |
| Inflation | 4.4 | 4.2 | 4.1 | 4.0 | 5.2 | 5.0 | 3.8 | 2.1 | 1.7 | 0.2 |

a. Use a graphical technique to depict the relationship between the two variables.

b. Does it appear that the returns on common stocks and inflation are linearly related?

**3.50** Xr03-50 In a university where calculus is a prerequisite for the statistics course, a sample of 15 students was drawn. The marks for calculus and statistics were recorded for each student. The data are as follows:

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Calculus | 65 | 58 | 93 | 68 | 74 | 81 | 58 | 85 | 88 | 75 | 63 | 79 | 80 | 54 | 72 |
| Statistics | 74 | 72 | 84 | 71 | 68 | 85 | 63 | 73 | 79 | 65 | 62 | 71 | 74 | 68 | 73 |

a. Draw a scatter diagram of the data.

b. What does the graph tell you about the relationship between the marks in calculus and statistics?

**3.51** Xr03-51 The cost of repairing cars involved in accidents is one reason that insurance premiums are so high. In an experiment, 10 cars were driven into a wall. The speeds were varied between 2 and 20 mph. The costs of repair were estimated and are listed here. Draw an appropriate graph to analyze the relationship between the two variables. What does the graph tell you?

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Speed | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 |
| Cost of Repair ($) | 88 | 124 | 358 | 519 | 699 | 816 | 905 | 1,521 | 1,888 | 2,201 |

**3.52** Xr03-52 The growing interest in and use of the Internet have forced many companies into considering ways to sell their products on the Web. Therefore, it is of interest to these companies to determine who is using the Web. A statistics practitioner undertook a study to determine how education and Internet use are connected. She took a random sample of 15 adults (20 years of age and older) and asked each to report the years of education they had completed and the number of hours of Internet use in the previous week. These data follow.

a. Employ a suitable graph to depict the data.
b. Does it appear that there is a linear relationship between the two variables? If so, describe it.

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Education | 11 | 11 | 8 | 13 | 17 | 11 | 11 | 11 | 19 | 13 | 15 | 9 | 15 | 15 | 11 |
| Internet use | 10 | 5 | 0 | 14 | 24 | 0 | 15 | 12 | 20 | 10 | 5 | 8 | 12 | 15 | 0 |

**3.53** Xr03-53 A statistics professor formed the opinion that students who handed in quiz and exams early outperformed students who handed in their papers later. To develop data to decide whether her opinion is valid, she recorded the amount of time (in minutes) taken by students to submit their midterm tests (time limit 90 minutes) and the subsequent mark for a sample of 12 students.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Time | 90 | 73 | 86 | 85 | 80 | 87 | 90 | 78 | 84 | 71 | 72 | 88 |
| Mark | 68 | 65 | 58 | 94 | 76 | 91 | 62 | 81 | 75 | 83 | 85 | 74 |

*The following exercises require a computer and software.*

**3.54** Xr03-54 In an attempt to determine the factors that affect the amount of energy used, 200 households were analyzed. The number of occupants and the amount of electricity used were measured for each household.

a. Draw a graph of the data.
b. What have you learned from the graph?

**3.55** Xr03-55 Many downhill skiers eagerly look forward to the winter months and fresh snowfalls. However, winter also entails cold days. How does the temperature affect skiers' desire? To answer this question, a local ski resort recorded the temperature for 50 randomly selected days and the number of lift tickets they sold. Use a graphical technique to describe the data and interpret your results.

**3.56** Xr03-56 One general belief held by observers of the business world is that taller men earn more money than shorter men. In a University of Pittsburgh study, 250 MBA graduates, all about 30 years old, were polled and asked to report their height (in inches) and their annual income (to the nearest $1,000).

a. Draw a scatter diagram of the data.
b. What have you learned from the scatter diagram?

**3.57** Xr03-57 Do chief executive officers (CEOs) of publicly traded companies earn their compensation? Every year the *National Post's Business* magazine attempts to answer the question by reporting the CEO's annual compensation ($1,000), the profit (or loss) ($1,000), and the three-year share return (%) for the top 50 Canadian companies. Use a graphical technique to answer the question.

**3.58** Xr03-58 Are younger workers less likely to stay with their jobs? To help answer this question, a random sample of workers was selected. All were asked to report their ages and how many months they had been employed with their current employers. Use a graphical technique to summarize these data. (Adapted from *Statistical Abstract of the United States, 2006*, Table 599.)

**3.59** Xr03-59 A very large contribution to profits for a movie theater is the sales of popcorn, soft drinks, and candy. A movie theater manager speculated that the longer the time between showings of a movie, the greater the sales of concession items. To acquire more information, the manager conducted an experiment. For a month he varied the amount of time between movie showings and calculated the sales. Use a graphical technique to help the manager determine whether longer time gaps produces higher concession stand sales.

**3.60** Xr03-60 An analyst employed at a commodities trading firm wanted to explore the relationship between prices of grains and livestock. Theoretically, the prices should move in the same direction because, as the price of livestock increases, more livestock are bred, resulting in a greater demand for grains to feed them. The analyst recorded the monthly grains and livestock subindexes for 1971 to 2008. (Subindexes are based on the prices of several similar commodities. For example, the livestock subindex represents the prices of cattle and hogs.) Using a graphical technique, describe the relationship between the two subindexes and report your findings. (*Source:* Bridge Commodity Research Bureau.)

**3.61** Xr03-61 It is generally believed that higher interest rates result in less employment because companies are more reluctant to borrow to expand their business. To determine whether there is a relationship between bank prime rate and unemployment, an economist collected the monthly prime bank rate and the monthly unemployment rate for the years 1950 to 2009. Use a graphical technique to supply your answer. (*Source:* Bridge Commodity Research Bureau.)

## American National Election Survey Exercises

3.62 ANES2004* Do younger people have more education (EDUC) than older people (AGE)? Use the American National Election Survey from 2004 and a graphical technique to help answer the question.

In the 2008 survey American adults were asked to report the amount of time (in minutes) that each person spent in an average day watching, reading, or listening about news in four different media. They are

Internet (TIME1)
Television (TIME2)
Printed newspaper (TIME3)
Radio (TIME4)

**3.63** ANES2008* Use a graphical technique to determine whether people who spend more time reading news on the Internet also devote more time to watching news on television.

**3.64** ANES2008* Analyze the relationship between the amount time reading news on the Internet and reading news in a printed newspaper. Does it appear that they are linearly related?

**3.65** ANES2008* Refer to Exercise 3.64. Study the scatter diagram. Does it appear that something is wrong with the data? If so, how do you correct the problem and determine whether a linear relationship exists?

**3.66** ANES2008* Graphically describe the relationship between the amount of time watching news on television and listening to news on the radio. Are the two linearly related?

**3.67** ANES2008* Do younger people spend more time reading news on the Internet than older people? Use a graphical technique to help answer the question.

## Genral Social Survey Exercises

**3.68** GSS2008* Do more educated people tend to marry people with more education? Draw a scatter diagram of EDUC and SPEDUC to answer the question.

**3.69** GSS2008* Do the children of more educated men (PAEDUC) have more education (EDUC)? Produce a graph that helps answer the question.

**3.70** GSS2008* Is there a positive linear relationship between the amount of education of mothers (MAEDUC) and their children (EDUC)? Draw a scatter diagram to answer the question.

**3.71** GSS2008* If one member of a married couple works more hours (HRS) does his or her spouse work less hours (SPHRS)? Draw a graph to produce the information you need.

## 3.4 Art and Science of Graphical Presentations

In this chapter and in Chapter 2, we introduced a number of graphical techniques. The emphasis was on how to construct each one manually and how to command the computer to draw them. In this section, we discuss how to use graphical techniques effectively. We introduce the concept of **graphical excellence**, which is a term we apply to techniques that are informative and concise and that impart information clearly to their viewers. Additionally, we discuss an equally important concept: graphical integrity and its enemy graphical deception.

### Graphical Excellence

Graphical excellence is achieved when the following characteristics apply.

1. **The graph presents large data sets concisely and coherently.** Graphical techniques were created to summarize and describe large data sets. Small data sets are easily summarized with a table. One or two numbers can best be presented in a sentence.

2. **The ideas and concepts the statistics practitioner wants to deliver are clearly understood by the viewer.** The chart is designed to describe what would otherwise be described in words. An excellent chart is one that can replace a thousand words and still be clearly comprehended by its readers.

3. **The graph encourages the viewer to compare two or more variables.** Graphs displaying only one variable provide very little information. Graphs are often best used to depict relationships between two or more variables or to explain how and why the observed results occurred.

4. **The display induces the viewer to address the substance of the data and not the form of the graph.** The form of the graph is supposed to help present the substance. If the form replaces the substance, the chart is not performing its function.

5. **There is no distortion of what the data reveal.** You cannot make statistical techniques say whatever you like. A knowledgeable reader will easily see through distortions and deception. We will endeavor to make you a knowledgeable reader by describing graphical deception later in this section.

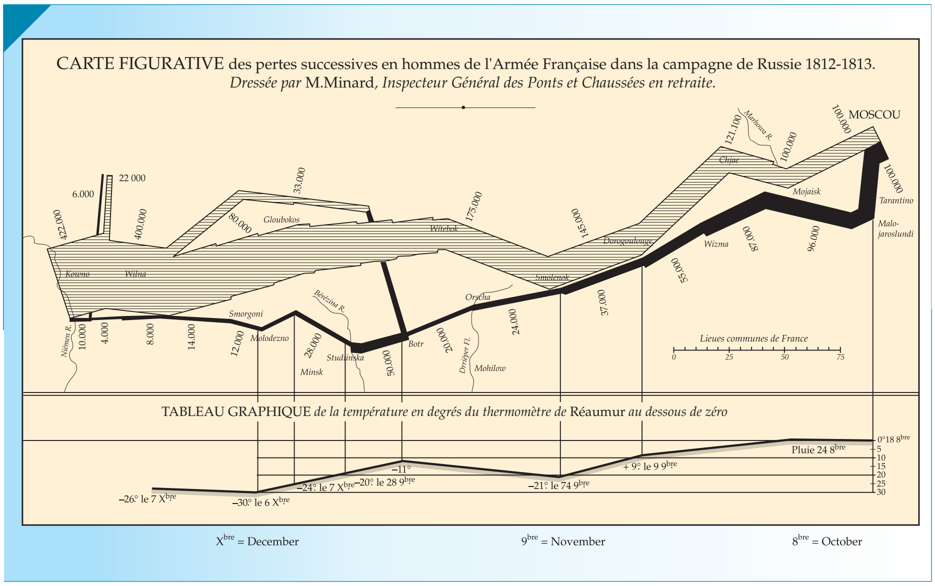
Edward Tufte, professor of statistics at Yale University, summarized graphical excellence this way:

1. Graphical excellence is the well-designed presentation of interesting data—a matter of substance, of statistics, and of design.
2. Graphical excellence is that which gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
3. Graphical excellence is nearly always multivariate.
4. And graphical excellence requires telling the truth about the data.

Now let's examine the chart that has been acclaimed the best chart ever drawn.

Figure 3.15 depicts Minard's graph. The striped band is a time series depicting the size of the army at various places on the map, which is also part of the chart. When

FIGURE **3.15** Chart Depicting Napoleon's Invasion and Retreat from Russia in 1812



*Source* : Edward Tufte, *The Visual Display of Quantitative Information* (Cheshire, CT: Graphics Press, 1983), p. 41.

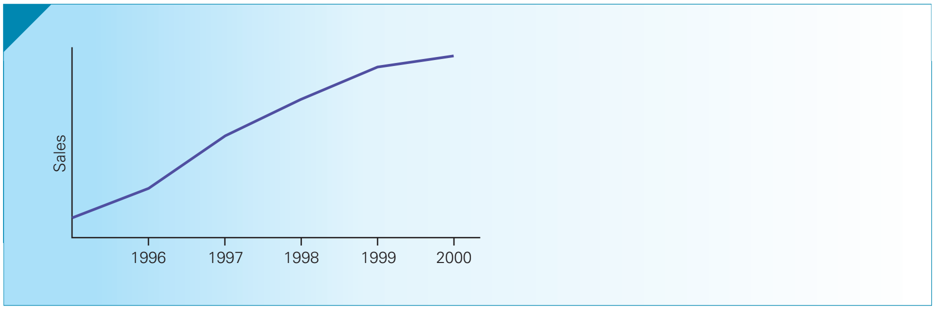
Napoleon invaded Russia by crossing the Niemen River on June 21, 1812, there were 422,000 soldiers. By the time the army reached Moscow, the number had dwindled to 100,000. At that point, the army started its retreat. The black band represents the army in retreat. At the bottom of the chart, we see the dates starting with October 1813. Just above the dates, Minard drew another time series, this one showing the temperature. It was bitterly cold during the fall, and many soldiers died of exposure. As you can see, the temperature dipped to –30 on December 6. The chart is effective because it depicts five variables clearly and succinctly.

## Graphical Deception

The use of graphs and charts is pervasive in newspapers, magazines, business and economic reports, and seminars, in large part because of the increasing availability of computers and software that allow the storage, retrieval, manipulation, and summary of large masses of raw data. It is therefore more important than ever to be able to evaluate critically the information presented by means of graphical techniques. In the final analysis, graphical techniques merely create a visual impression, which is easy to distort. In fact, distortion is so easy and commonplace that in 1992 the Canadian Institute of Chartered Accountants found it necessary to begin setting guidelines for financial graphics, after a study of hundreds of the annual reports of major corporations found that 8% contained at least one misleading graph that covered up bad results. Although the heading for this section mentions deception, it is quite possible for an inexperienced person inadvertently to create distorted impressions with graphs. In any event, you should be aware of possible methods of graphical deception. This section illustrates a few of them.

The first thing to watch for is a graph without a scale on one axis. The line chart of a firm's sales in Figure 3.16 might represent a growth rate of 100% or 1% over the 5 years depicted, depending on the vertical scale. It is best simply to ignore such graphs.

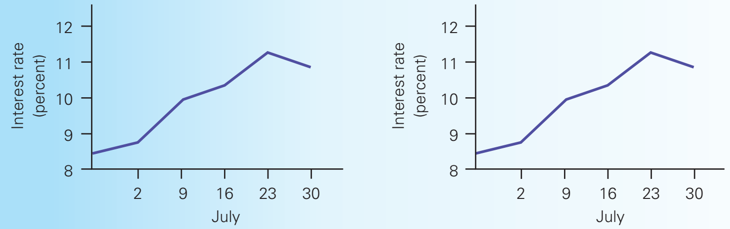
FIGURE **3.16** Graph without Scale



A second trap to avoid is being influenced by a graph's caption. Your impression of the trend in interest rates might be different, depending on whether you read a newspaper carrying caption (a) or caption (b) in Figure 3.17.

Perspective is often distorted if only absolute changes in value, rather than percentage changes, are reported. A $1 drop in the price of your $2 stock is relatively more distressing than a $1 drop in the price of your $100 stock. On January 9, 1986, newspapers throughout North America displayed graphs similar to the one shown in Figure 3.18 and reported that the stock market, as measured by the Dow Jones Industrial Average (DJIA), had suffered its worst 1-day loss ever on the previous day.

FIGURE 3.17 Graphs with Different Captions

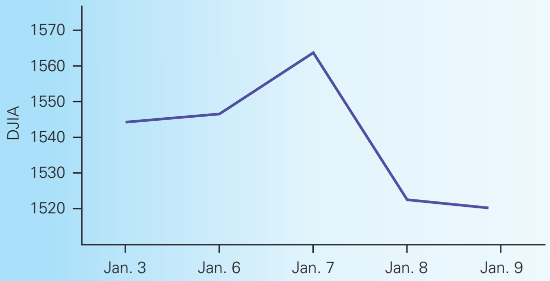


**(a)** Interest rates have finally begun to turn downward.

**(b)** Last week provided temporary relief from the upward trend in interest rates.

The loss was 39 points, exceeding even the loss of Black Tuesday: October 28, 1929. While the loss was indeed a large one, many news reports failed to mention that the 1986 level of the DJIA was much higher than the 1929 level. A better perspective on the situation could be gained by noticing that the loss on January 8, 1986, represented a 2.5% decline, whereas the decline in 1929 was 12.8%. As a point of interest, we note that the stock market was 12% higher within 2 months of this historic drop and 40% higher 1 year later. The largest one-day percentage drop in the DJIA is 24.4% (December 12, 1914).

FIGURE 3.18 Graph Showing Drop in the DJIA



We now turn to some rather subtle methods of creating distorted impressions with graphs. Consider the graph in Figure 3.19, which depicts the growth in a firm's quarterly sales during the past year, from $100 million to $110 million. This 10% growth in quarterly sales can be made to appear more dramatic by stretching the vertical axis—a technique that involves changing the scale on the vertical axis so that a given dollar amount is represented by a greater height than before. As a result, the rise in sales appears to be greater because the slope of the graph is visually (but not numerically) steeper. The expanded scale is usually accommodated by employing a break in the vertical axis, as in Figure 3.20(a), or by truncating the vertical axis, as in Figure 3.20(b), so that the vertical scale begins at a point greater than zero. The effect of making slopes appear steeper can also be created by shrinking the horizontal axis, in which case points on the horizontal axis are moved closer together.

FIGURE 3.19 Graph Showing Growth in Quarterly Sales 1

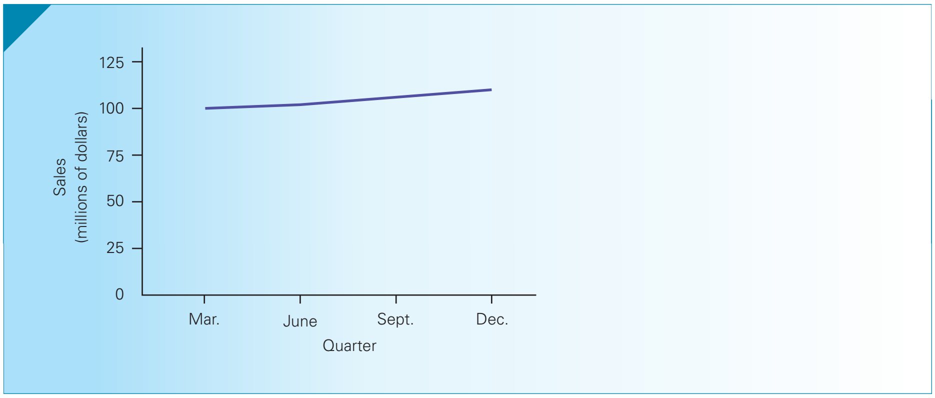
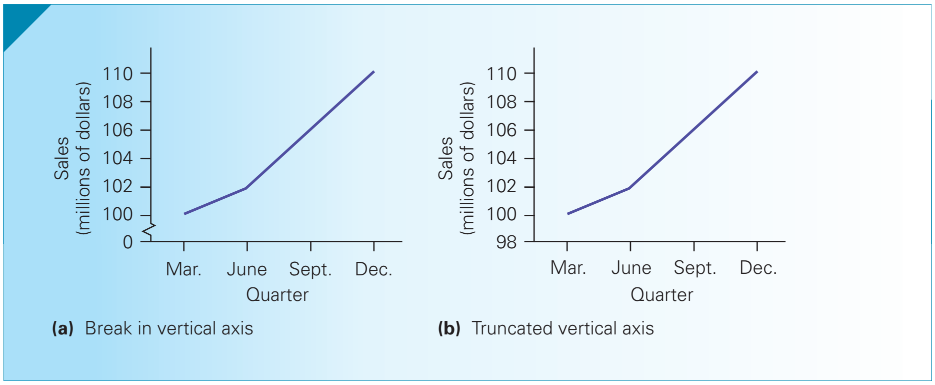


FIGURE 3.20 Graph Showing Growth in Quarterly Sales 2

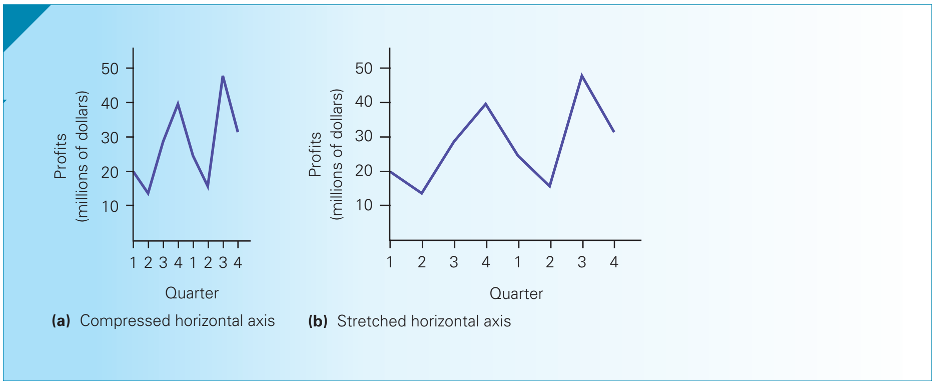


**(a)** Break in vertical axis

**(b)** Truncated vertical axis

Just the opposite effect is obtained by stretching the horizontal axis; that is, spreading out the points on the horizontal axis to increase the distance between them so that slopes and trends will appear to be less steep. The graph of a firm's profits presented in Figure 3.21(a) shows considerable swings, both upward and downward in the profits from one quarter to the next. However, the firm could convey the impression of reasonable stability in profits from quarter to quarter by stretching the horizontal axis, as shown in Figure 3.21(b).

FIGURE 3.21 Graph Showing Considerable Swings or Relative Stability

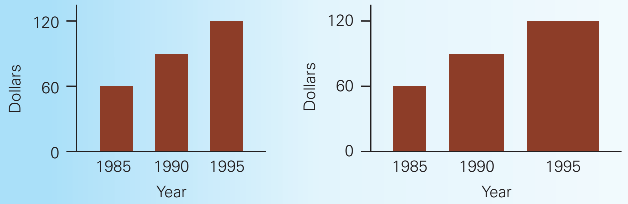


**(a)** Compressed horizontal axis

**(b)** Stretched horizontal axis

Similar illusions can be created with bar charts by stretching or shrinking the vertical or horizontal axis. Another popular method of creating distorted impressions with bar charts is to construct the bars so that their widths are proportional to their heights. The bar chart in Figure 3.22(a) correctly depicts the average weekly amount spent on food by Canadian families during three particular years. This chart correctly uses bars of equal width so that both the height and the area of each bar are proportional to the expenditures they represent. The growth in food expenditures is exaggerated in Figure 3.22(b), in which the widths of the bars increase with their heights. A quick glance at this bar chart might leave the viewer with the mistaken impression that food expenditures increased fourfold over the decade, because the 1995 bar is four times the size of the 1985 bar.

FIGURE **3.22** Correct and Distorted Bar Charts

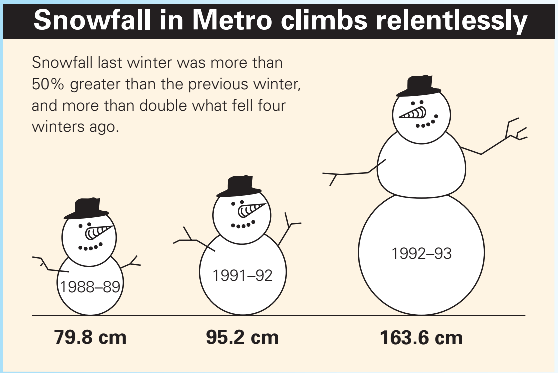


**(a)** Correct bar chart

**(b)** Increasing bar widths to create distortion

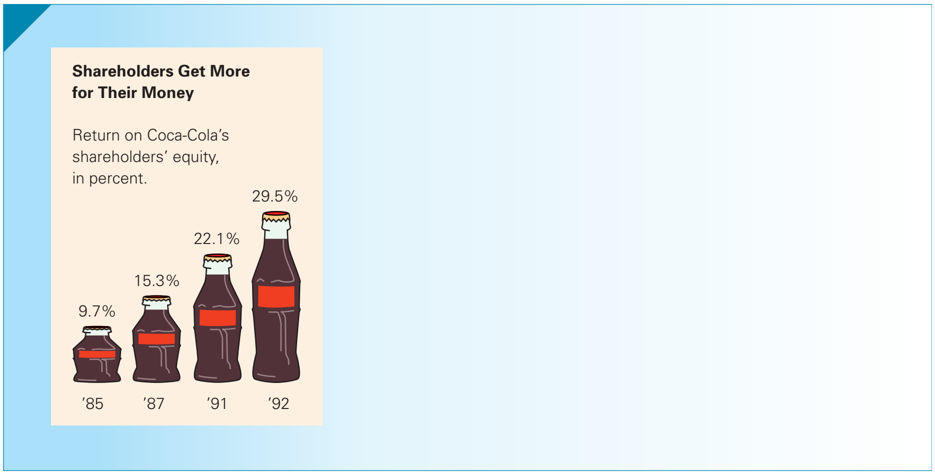
You should be on the lookout for size distortions, particularly in pictograms, which replace the bars with pictures of objects (such as bags of money, people, or animals) to enhance the visual appeal. Figure 3.23 displays the misuse of a pictogram—the snowman grows in width as well as height. The proper use of a pictogram is shown in Figure 3.24, which effectively uses pictures of Coca-Cola bottles.

FIGURE **3.23** Misuse of Pictogram


Snowfall in Metro climbs relentlessly
Snowfall last winter was more than 50% greater than the previous winter, and more than double what fell four winters ago.
1988–89
1991–92
1992–93
79.8 cm
95.2 cm
163.6 cm

The preceding examples of creating a distorted impression using graphs are not exhaustive, but they include some of the more popular methods. They should also serve to make the point that graphical techniques are used to create a visual impression, and

FIGURE 3.24 Correct Pictogram



the impression you obtain may be a distorted one unless you examine the graph with care. You are less likely to be misled if you focus your attention on the numerical values that the graph represents. Begin by carefully noting the scales on both axes; graphs with unmarked axes should be ignored completely.

# Exercises

**3.72** Xr03-72 A computer company has diversified its operations into financial services, construction, manufacturing, and hotels. In a recent annual report, the following tables were provided. Create charts to present these data so that the differences between last year and the previous year are clear. (*Note:* It may be necessary to draw the charts manually.)

| | Sales (Millions of Dollars) by Region | |
|---|---|---|
| **Region** | **Last Year** | **Previous Year** |
| United States | 67.3 | 40.4 |
| Canada | 20.9 | 18.9 |
| Europe | 37.9 | 35.5 |
| Australasia | 26.2 | 10.3 |
| Total | 152.2 | 105.1 |

| | Sales (Millions of Dollars) by Division | |
|---|---|---|
| **Division** | **Last Year** | **Previous Year** |
| Customer service | 54.6 | 43.8 |
| Library systems | 49.3 | 30.5 |
| Construction/property management | 17.5 | 7.7 |
| Manufacturing and distribution | 15.4 | 8.9 |
| Financial systems | 9.4 | 10.9 |
| Hotels and clubs | 5.9 | 3.4 |

**3.73** Xr03-73 The following table lists the number (in thousands) of violent crimes and property crimes committed annually in 1985 to 2006 (the last year data were available).

a. Draw a chart that displays both sets of data.
b. Does it appear that crime rates are decreasing? Explain.
c. Is there another variable that should be included to show the trends in crime rates?

| Year | Violent crimes | Property crimes |
|---|---|---|
| 1985 | 1,328 | 11,103 |
| 1986 | 1,489 | 11,723 |
| 1987 | 1,484 | 12,025 |
| 1988 | 1,566 | 12,357 |

| | | |
|---|---|---|
| 1989 | 1,646 | 12,605 |
| 1990 | 1,820 | 12,655 |
| 1991 | 1,912 | 12,961 |
| 1992 | 1,932 | 12,506 |
| 1993 | 1,926 | 12,219 |
| 1994 | 1,858 | 12,132 |
| 1995 | 1,799 | 12,064 |
| 1996 | 1,689 | 11,805 |
| 1997 | 1,636 | 11,558 |
| 1998 | 1,534 | 10,952 |
| 1999 | 1,426 | 10,208 |
| 2000 | 1,425 | 10,183 |
| 2001 | 1,439 | 10,437 |
| 2002 | 1,424 | 10,455 |
| 2003 | 1,384 | 10,443 |
| 2004 | 1,360 | 10,319 |
| 2005 | 1,391 | 10,175 |
| 2006 | 1,418 | 9,984 |

*Source: Statistical Abstract of the United States, 2006,* Table 293; and *2009,* Table 293.

**3.74** Xr03-74 Refer to Exercise 3.73. We've added the United States population.

a. Incorporate this variable into your charts to show crime rate trends.
b. Summarize your findings.
c. Can you think of another demographic variable that may explain crime rate trends?

**3.75** Xr03-75 Refer to Exercises 3.73 and 3.74. We've included the number of Americans aged 15 to 24.

a. What is the significance of adding the populations aged 15 to 24?
b. Include these data in your analysis. What have you discovered?

**3.76** Xr03-76 To determine premiums for automobile insurance, companies must have an understanding of the variables that affect whether a driver will have an accident. The age of the driver may top the list of variables. The following table lists the number of drivers in the United States, the number of fatal accidents, and the number of total accidents in each age group in 2002.

a. Calculate the accident rate (per driver) and the fatal accident rate (per 1,000 drivers) for each age group.
b. Graphically depict the relationship between the ages of drivers, their accident rates, and their fatal accident rates (per 1,000 drivers).
c. Briefly describe what you have learned.

| Age Group | Number of Drivers (1,000s) | Number of Accidents (1,000s) | Number of Fatal Accidents |
|---|---|---|---|
| Under 20 | 9,508 | 3,543 | 6,118 |
| 20–24 | 16,768 | 2,901 | 5,907 |
| 25–34 | 33,734 | 7,061 | 10,288 |
| 35–44 | 41,040 | 6,665 | 10,309 |
| 45–54 | 38,711 | 5,136 | 8,274 |
| 55–64 | 25,609 | 2,775 | 5,322 |
| 65–74 | 15,812 | 1,498 | 2,793 |
| Over 74 | 12,118 | 1,121 | 3,689 |
| Total | 193,300 | 30,700 | 52,700 |

*Source:* National Safety Council.

**3.77** Xr03-77 During 2002 in the state of Florida, a total of 365,474 drivers were involved in car accidents. The accompanying table breaks down this number by the age group of the driver and whether the driver was injured or killed. (There were actually 371,877 accidents, but the driver's age was not recorded in 6,413 of these.)

a. Calculate the injury rate (per 100 accidents) and the death rate (per accident) for each age group.
b. Graphically depict the relationship between the ages of drivers, their injury rate (per 100 accidents), and their death rate.
c. Briefly describe what you have learned from these graphs.
d. What is the difference between the information extracted from Exercise 3.9 and this one?

| Age Group | Number of Accidents | Drivers Injured | Drivers Killed |
|---|---|---|---|
| 20 or less | 52,313 | 21,762 | 217 |
| 21–24 | 38,449 | 16,016 | 185 |
| 25–34 | 78,703 | 31,503 | 324 |
| 35–44 | 76,152 | 30,542 | 389 |
| 45–54 | 54,699 | 22,638 | 260 |
| 55–64 | 31,985 | 13,210 | 167 |
| 65–74 | 18,896 | 7,892 | 133 |
| 75–84 | 11,526 | 5,106 | 138 |
| 85 or more | 2,751 | 1,223 | 65 |
| Total | 365,474 | 149,892 | 1,878 |

*Source*: Florida Department of Highway Safety and Motor Vehicles.

**3.78** Xr03-78 The accompanying table lists the average test scores in the Scholastic Assessment Test (SAT) for the years 1967, 1970, 1975, 1980, 1985, 1990, 1995, and 1997 to 2007.

| Year | Verbal All | Verbal Male | Verbal Female | Math All | Math Male | Math Female |
|---|---|---|---|---|---|---|
| 1967 | 543 | 540 | 545 | 516 | 535 | 595 |
| 1970 | 537 | 536 | 538 | 512 | 531 | 493 |
| 1975 | 512 | 515 | 509 | 498 | 518 | 479 |
| 1980 | 502 | 506 | 498 | 492 | 515 | 473 |
| 1985 | 509 | 514 | 503 | 500 | 522 | 480 |
| 1990 | 500 | 505 | 496 | 501 | 521 | 483 |
| 1995 | 504 | 505 | 502 | 506 | 525 | 490 |
| 1997 | 505 | 507 | 503 | 511 | 530 | 494 |
| 1998 | 505 | 509 | 502 | 512 | 531 | 496 |
| 1999 | 505 | 509 | 502 | 511 | 531 | 495 |
| 2000 | 505 | 507 | 504 | 514 | 533 | 498 |
| 2001 | 506 | 509 | 502 | 514 | 533 | 498 |
| 2002 | 504 | 507 | 502 | 516 | 534 | 500 |
| 2003 | 507 | 512 | 503 | 519 | 537 | 503 |
| 2004 | 508 | 512 | 504 | 518 | 537 | 501 |
| 2005 | 508 | 513 | 505 | 520 | 538 | 504 |
| 2006 | 503 | 505 | 502 | 518 | 536 | 502 |
| 2007 | 502 | 504 | 502 | 515 | 533 | 499 |

*Source: Statistical Abstract of the United States, 2003*, Table 264; *2006*, Table 252; *2009*, Table 258.

Draw a chart for each of the following.

a. You wish to show that both verbal and mathematics test scores for all students have not changed much over the years.
b. The exact opposite of part (a).
c. You want to claim that there are no differences between genders.
d. You want to "prove" that differences between genders exist.

**3.79** Xr03-79 The monthly unemployment rate in one state for the past 12 months is listed here.

a. Draw a bar chart of these data with 6.0% as the lowest point on the vertical axis.
b. Draw a bar chart of these data with 0.0% as the lowest point on the vertical axis.
c. Discuss the impression given by the two charts.
d. Which chart would you use? Explain.

| Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rate | 7.5 | 7.6 | 7.5 | 7.3 | 7.2 | 7.1 | 7.0 | 6.7 | 6.4 | 6.5 | 6.3 | 6.0 |

**3.80** Xr03-80 The accompanying table lists the federal minimum wage from 1955 to 2007. The actual and adjusted minimum wages (in constant 1996 dollars) are listed.

a. Suppose you wish to show that the federal minimum wage has grown rapidly over the years. Draw an appropriate chart.
b. Draw a chart to display the actual changes in the federal minimum wage.

| Year | Current Dollars | Constant 1996 Dollars | Year | Current Dollars | Constant 1996 Dollars |
|---|---|---|---|---|---|
| 1955 | 0.75 | 4.39 | 1982 | 3.35 | 5.78 |
| 1956 | 1.00 | 5.77 | 1983 | 3.35 | 5.28 |
| 1957 | 1.00 | 5.58 | 1984 | 3.35 | 5.06 |
| 1958 | 1.00 | 5.43 | 1985 | 3.35 | 4.88 |
| 1959 | 1.00 | 5.39 | 1986 | 3.35 | 4.80 |
| 1960 | 1.00 | 5.30 | 1987 | 3.35 | 4.63 |
| 1961 | 1.15 | 6.03 | 1988 | 3.35 | 4.44 |
| 1962 | 1.15 | 5.97 | 1989 | 3.35 | 4.24 |
| 1963 | 1.25 | 6.41 | 1990 | 3.80 | 4.56 |
| 1964 | 1.25 | 6.33 | 1991 | 4.25 | 4.90 |
| 1965 | 1.25 | 6.23 | 1992 | 4.25 | 4.75 |
| 1966 | 1.25 | 6.05 | 1993 | 4.25 | 4.61 |
| 1967 | 1.40 | 6.58 | 1994 | 4.25 | 4.50 |

| | | | | | |
|---|---|---|---|---|---|
| 1968 | 1.60 | 7.21 | 1995 | 4.25 | 4.38 |
| 1969 | 1.60 | 6.84 | 1996 | 4.75 | 4.75 |
| 1970 | 1.60 | 6.47 | 1997 | 5.15 | 5.03 |
| 1971 | 1.60 | 6.20 | 1998 | 5.15 | 4.96 |
| 1972 | 1.60 | 6.01 | 1999 | 5.15 | 4.85 |
| 1973 | 1.60 | 5.65 | 2000 | 5.15 | 4.69 |
| 1974 | 2.00 | 6.37 | 2001 | 5.15 | 4.56 |
| 1975 | 2.10 | 6.12 | 2002 | 5.15 | 4.49 |
| 1976 | 2.30 | 6.34 | 2003 | 5.15 | 4.39 |
| 1977 | 2.30 | 5.95 | 2004 | 5.15 | 4.28 |
| 1978 | 2.65 | 6.38 | 2005 | 5.15 | 4.14 |
| 1979 | 2.90 | 6.27 | 2006 | 5.15 | 4.04 |
| 1980 | 3.10 | 5.90 | 2007 | 5.85 | 4.41 |
| 1981 | 3.35 | 5.78 | | | |

*Source*: U.S. Department of Labor.

**3.81** Xr03-81 The following table shows school enrollment (in thousands) for public and private schools for the years 1965 to 2005.

a. Draw charts that allow you to claim that enrollment in private schools is "skyrocketing."

b. Draw charts that "prove" public school enrollment is stagnant.

| Year | Public_K_8 | Private_K_8 | Public_9_12 | Private_9_12 | College_Public | College_Private |
|---|---|---|---|---|---|---|
| 1965 | 30,563 | 4,900 | 11,610 | 1,400 | 3,970 | 1,951 |
| 1966 | 31,145 | 4,800 | 11,894 | 1,400 | 4,349 | 2,041 |
| 1967 | 31,641 | 4,600 | 12,250 | 1,400 | 4,816 | 2,096 |
| 1968 | 32,226 | 4,400 | 12,718 | 1,400 | 5,431 | 2,082 |
| 1969 | 32,513 | 4,200 | 13,037 | 1,300 | 5,897 | 2,108 |
| 1970 | 32,558 | 4,052 | 13,336 | 1,311 | 6,428 | 2,153 |
| 1971 | 32,318 | 3,900 | 13,753 | 1,300 | 6,804 | 2,144 |
| 1972 | 31,879 | 3,700 | 13,848 | 1,300 | 7,071 | 2,144 |
| 1973 | 31,401 | 3,700 | 14,044 | 1,300 | 7,420 | 2,183 |
| 1974 | 30,971 | 3,700 | 14,103 | 1,300 | 7,989 | 2,235 |
| 1975 | 30,515 | 3,700 | 14,304 | 1,300 | 8,835 | 2,350 |
| 1976 | 29,997 | 3,825 | 14,314 | 1,342 | 8,653 | 2,359 |
| 1977 | 29,375 | 3,797 | 14,203 | 1,343 | 8,847 | 2,439 |
| 1978 | 28,463 | 3,732 | 14,088 | 1,353 | 8,786 | 2,474 |
| 1979 | 28,034 | 3,700 | 13,616 | 1,300 | 9,037 | 2,533 |
| 1980 | 27,647 | 3,992 | 13,231 | 1,339 | 9,457 | 2,640 |
| 1981 | 27,280 | 4,100 | 12,764 | 1,400 | 9,647 | 2,725 |
| 1982 | 27,161 | 4,200 | 12,405 | 1,400 | 9,696 | 2,730 |
| 1983 | 26,981 | 4,315 | 12,271 | 1,400 | 9,683 | 2,782 |
| 1984 | 26,905 | 4,300 | 12,304 | 1,400 | 9,477 | 2,765 |
| 1985 | 27,034 | 4,195 | 12,388 | 1,362 | 9,479 | 2,768 |
| 1986 | 27,420 | 4,116 | 12,333 | 1,336 | 9,714 | 2,790 |
| 1987 | 27,933 | 4,232 | 12,076 | 1,247 | 9,973 | 2,793 |
| 1988 | 28,501 | 4,036 | 11,687 | 1,206 | 10,161 | 2,894 |
| 1989 | 29,152 | 4,035 | 11,390 | 1,163 | 10,578 | 2,961 |
| 1990 | 29,878 | 4,084 | 11,338 | 1,150 | 10,845 | 2,974 |
| 1991 | 30,506 | 4,518 | 11,541 | 1,163 | 11,310 | 3,049 |
| 1992 | 31,088 | 4,528 | 11,735 | 1,148 | 11,385 | 3,102 |
| 1993 | 31,504 | 4,536 | 11,961 | 1,132 | 11,189 | 3,116 |
| 1994 | 31,898 | 4,624 | 12,213 | 1,162 | 11,134 | 3,145 |
| 1995 | 32,341 | 4,721 | 12,500 | 1,197 | 11,092 | 3,169 |
| 1996 | 32,764 | 4,720 | 12,847 | 1,213 | 11,121 | 3,247 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 1997 | 33,073 | 4,726 | 13,054 | 1,218 | 11,196 | 3,306 |
| 1998 | 33,346 | 4,748 | 13,193 | 1,240 | 11,138 | 3,369 |
| 1999 | 33,488 | 4,765 | 13,369 | 1,254 | 11,309 | 3,482 |
| 2000 | 33,688 | 4,878 | 13,515 | 1,292 | 11,753 | 3,560 |
| 2001 | 33,938 | 4,993 | 13,734 | 1,326 | 12,233 | 3,695 |
| 2002 | 34,116 | 4,886 | 14,067 | 1,334 | 12,752 | 3,860 |
| 2003 | 34,202 | 4,761 | 14,338 | 1,338 | 12,857 | 4,043 |
| 2004 | 34,178 | 4,731 | 14,617 | 1,356 | 12,980 | 4,292 |
| 2005 | 34,205 | 4,699 | 14,909 | 1,374 | 13,022 | 4,466 |

*Source: Statistical Abstract of the United States, 2009*, Table 211.

**3.82** Xr03-82 The following table lists the percentage of single and married women in the United States who had jobs outside the home during the period 1970 to 2007.

a. Construct a chart that shows that the percentage of married women who are working outside the home has not changed much in the past 47 years.

b. Use a chart to show that the percentage of single women in the workforce has increased "dramatically."

| Year | Single | Married |
|---|---|---|
| 1970 | 56.8 | 40.5 |
| 1971 | 56.4 | 40.6 |
| 1972 | 57.5 | 41.2 |
| 1973 | 58.6 | 42.3 |
| 1974 | 59.5 | 43.3 |
| 1975 | 59.8 | 44.3 |
| 1976 | 61.0 | 45.3 |
| 1977 | 62.1 | 46.4 |
| 1978 | 63.7 | 47.8 |
| 1979 | 64.6 | 49.0 |
| 1980 | 64.4 | 49.8 |
| 1981 | 64.5 | 50.5 |
| 1982 | 65.1 | 51.1 |
| 1983 | 65.0 | 51.8 |
| 1984 | 65.6 | 52.8 |
| 1985 | 66.6 | 53.8 |
| 1986 | 67.2 | 54.9 |
| 1987 | 67.4 | 55.9 |
| 1988 | 67.7 | 56.7 |
| 1989 | 68.0 | 57.8 |
| 1990 | 66.7 | 58.4 |
| 1991 | 66.2 | 58.5 |
| 1992 | 66.2 | 59.3 |
| 1993 | 66.2 | 59.4 |
| 1994 | 66.7 | 60.7 |
| 1995 | 66.8 | 61.0 |
| 1996 | 67.1 | 61.2 |
| 1997 | 67.9 | 61.6 |
| 1998 | 68.5 | 61.2 |
| 1999 | 68.7 | 61.2 |
| 2000 | 68.9 | 61.1 |
| 2001 | 68.1 | 61.2 |
| 2002 | 67.4 | 61.0 |
| 2003 | 66.2 | 61.0 |
| 2004 | 65.9 | 60.5 |
| 2005 | 66.0 | 60.7 |
| 2006 | 65.7 | 61.0 |
| 2007 | 65.3 | 61.0 |

*Source: Statistical Abstract of the United States, 2009*, Table 286.

## CHAPTER SUMMARY

Histograms are used to describe a single set of interval data. Statistics practitioners examine several aspects of the shapes of histograms. These are symmetry, number of modes, and its resemblance to a bell shape.

We described the difference between time-series data and cross-sectional data. Time series are graphed by line charts.

To analyze the relationship between two interval variables, we draw a scatter diagram. We look for the direction and strength of the linear relationship.

## IMPORTANT TERMS

| | |
|---|---|
| Classes | 46 |
| Histogram | 46 |
| Symmetry | 50 |
| Positively skewed | 50 |
| Negatively skewed | 50 |
| Modal class | 50 |
| Unimodal | 50 |
| Bimodal | 51 |
| Stem-and-leaf display | 57 |
| Depths | 58 |
| Relative frequency distribution | 59 |
| Cumulative relative frequency distribution | 59 |
| Ogive | 60 |
| Credit scorecard | 63 |
| Cross-sectional data | 64 |
| Time-series data | 64 |
| Line chart | 65 |
| Scatter diagram | 74 |
| Linear relationship | 76 |
| Positive linear relationship | 77 |
| Negative linear relationship | 77 |
| Graphical excellence | 82 |
| Graphical deception | 82 |

## COMPUTER OUTPUT AND INSTRUCTIONS

| Graphical Technique | Excel | Minitab |
|---|---|---|
| Histogram | 47 | 48 |
| Stem-and-leaf display | 58 | 58 |
| Ogive | 60 | 61 |
| Line chart | 66 | 67 |
| Scatter diagram | 75 | 76 |

# CHAPTER EXERCISES

*The following exercises require a computer and software.*

**3.83** Xr03-83 Gold and other precious metals have traditionally been considered a hedge against inflation. If this is true, we would expect that a fund made up of precious metals (gold, silver, platinum, and others) would have a strong positive relationship with the inflation rate. To see whether this is true, a statistics practitioner collected the monthly CPI and the monthly precious metals subindex, which is based on the prices of gold, silver, platinum, etc., for the years 1975 to 2008. These figures were used to calculate the monthly inflation rate and the monthly return on the precious metals subindex. Use a graphical technique to determine the nature of the relationship between the inflation rate and the return on the subindex. What does the graph tell you? (*Source:* U.S. Treasury and Bridge Commodity Research Bureau.)

**3.84** Xr03-84 The monthly values of one Australian dollar measured in American dollars since 1971 were recorded. Draw a graph that shows how the exchange rate has varied over the 38-year period. (*Source*: Federal Reserve Economic Data.)

**3.85** Xr03-85 Studies of twins may reveal more about the "nature or nurture" debate. The issue being debated is whether nature or the environment has more of an effect on individual traits such as intelligence. Suppose that a sample of identical twins was selected and their IQs measured. Use a suitable graphical technique to depict the data, and describe what it tells you about the relationship between the IQs of identical twins.

**3.86** Xr03-86 An economist wanted to determine whether a relationship existed between interest rates and currencies (measured in U.S. dollars). He recorded the monthly interest rate and the currency indexes for the years 1982 to 2008. Graph the data and describe the results. (*Source:* Bridge Commodity Research Bureau.)

**3.87** Xr03-87 One hundred students who had reported that they use their computers for at least 20 hours per week were asked to keep track of the number of crashes their computers incurred during a 12-week period. Using an appropriate statistical method, summarize the data. Describe your findings.

**3.88** Xr03-88 In Chapters 16, 17, and 18, we introduce regression analysis, which addresses the relationships among variables. One of the first applications of regression analysis was to analyze the relationship between the heights of fathers and sons. Suppose that a sample of 80 fathers and sons was drawn. The heights of the fathers and of the adult sons were measured.

a. Draw a scatter diagram of the data. Draw a straight line that describes the relationship.
b. What is the direction of the line?
c. Does it appear that there is a linear relationship between the two variables?

**3.89** Xr03-89 When the Dow Jones Industrial Averages index increases, it usually means that the economy is growing, which in turn usually means that the unemployment rate is low. A statistics professor pointed out that in numerous periods (including when this edition was being written), the stock market had been booming while the rest of the economy was performing poorly. To learn more about the issue, the monthly closing DJIA and the monthly unemployment rates were recorded for the years 1950 to 2009. Draw a graph of the data and report your results. (*Source:* Federal Reserve Economic Data and the *Wall Street Journal*.)

**3.90** Xr03-90 The monthly values of one British pound measured in American dollars since 1987 were recorded. Produce a graph that shows how the exchange rate has varied over the past 23 years. (*Source*: Federal Reserve Economic Data.)

**3.91** Xr03-91 Do better golfers play faster than poorer ones? To determine whether a relationship exists, a sample of 125 foursomes was selected. Their total scores and the amount of time taken to complete the 18 holes were recorded. Graphically depict the data, and describe what they tell you about the relationship between score and time.

**3.92** Xr03-92 The value of monthly U.S. exports to Mexico and imports from Mexico (in $ millions) since 1985 were recorded. (*Source*: Federal Reserve Economic Data.)

a. Draw a chart that depicts exports.
b. Draw a chart that exhibits imports.
c. Compute the trade balance and graph these data.
d. What do these charts tell you?

**3.93** Xr03-93 An increasing number of consumers prefer to use debit cards in place of cash or credit cards. To analyze the relationship between the amounts of purchases made with debit and credit cards, 240 people were interviewed and asked to report the amount of money spent on purchases using debit cards and the amount spent using credit cards during the last month. Draw a graph of the data and summarize your findings.

**3.94** Xr03-94 Most publicly traded companies have boards of directors. The rate of pay varies considerably. A survey was undertaken by the *Globe and Mail* (February 19, 2001) wherein 100 companies were surveyed and asked to report how much their directors were paid annually. Use a graphical technique to present these data.

**3.95** Xr03-95 Refer to Exercise 3.94. In addition to reporting the annual payment per director, the survey recorded the number of meetings last year. Use a graphical technique to summarize and present these data.

**3.96** Xr03-96 Is airline travel becoming safer? To help answer this question, a student recorded the number of fatal accidents and the number of deaths that occurred in the years 1986 to 2007 for scheduled airlines. Use a graphical method to answer the question. (*Source: Statistical Abstract of the United States, 2009*, Table 1036.)

**3.97** Xr03-97 Most car-rental companies keep their cars for about a year and then sell them to used car dealerships. Suppose one company decided to sell the used cars themselves. Because most used car buyers make their decision on what to buy and how much to spend based on the car's odometer reading, this would be an important issue for the car-rental company. To develop information about the mileage shown on the company's rental cars, the general manager took a random sample of 658 customers and recorded the average number of miles driven per day. Use a graphical technique to display these data.

**3.98** Xr03-98 Several years ago, the Barnes Exhibit toured major cities all over the world, with millions of people flocking to see it. Dr. Albert Barnes was a wealthy art collector who accumulated a large number of impressionist masterpieces; the total exceeds 800 paintings. When Dr. Barnes died in 1951, he stated in his will that his collection was not to be allowed to tour. However, because of the deterioration of the exhibit's home near Philadelphia, a judge ruled that the collection could go on tour to raise enough money to renovate the building. Because of the size and value of the collection, it was predicted (correctly) that in each city a large number of people would come to view the paintings. Because space was limited, most galleries had to sell tickets that were valid at one time (much like a play). In this way, they were able to control the number of visitors at any one time. To judge how many people to let in at any time, it was necessary to know the length of time people would spend at the exhibit; longer times would dictate smaller audiences; shorter times would allow for the sale of more tickets. The manager of a gallery that will host the exhibit realized her facility can comfortably and safely hold about

250 people at any one time. Although the demand will vary throughout the day and from weekday to weekend, she believes that the demand will not drop below 500 at any time. To help make a decision about how many tickets to sell, she acquired the amount of time a sample of 400 people spent at the exhibit from another city. What ticket procedure should the museum management institute?

*The following exercises are based on data sets that include additional data referenced in previously presented examples and exercises.*

**3.99** Xm03-03* Xm03-04* Examples 3.3 and 3.4 listed the final marks in the business statistics course and the mathematical statistics course. The professor also provided the final marks in the first-year required calculus course. Graphically describe the relationship between calculus and statistics marks. What information were you able to extract?

**3.100** Xm03-03* Xm03-04* In addition to the previously discussed data in Examples 3.3 and 3.4, the professor listed the midterm mark. Conduct an analysis of the relationship between final exam mark and midterm mark in each course. What does this analysis tell you?

**3.101** Xr02-54* Two other questions were asked in Exercise 2.54:

Number of weeks job searching?
Salary ($ thousands)?

*The placement office wants the following:*

a. Graphically describe salary.
b. Is salary related to the number of weeks needed to land the job?

## CASE 3.1 The Question of Global Warming

DATA
C03-01a
C03-01b

In the last part of the 20th century, scientists developed the theory that the planet was warming and that the primary cause was the increasing amounts of atmospheric carbon dioxide ($CO_2$), which are the product of burning oil, natural gas, and coal (fossil fuels). Although many climatologists believe in the so-called greenhouse effect, many others do not subscribe to this theory. There are three critical questions that need to be answered in order to resolve the issue.

1. Is Earth actually warming? To answer this question, we need accurate temperature measurements over a large number of years. But how do we measure the temperature before the invention of accurate thermometers? Moreover, how do we go about measuring Earth's temperature even with accurate thermometers?
2. If the planet is warming, is there a human cause or is it natural fluctuation? Earth's temperature has increased and decreased many times in its long history. We've had higher temperatures, and we've had lower temperatures, including various ice ages. In fact, a period called the "Little Ice Age" ended around the middle to the end of the 19th century. Then the temperature rose until about 1940, at which point it decreased until 1975. In fact, an April 28, 1975, *Newsweek* article discussed the possibility of global cooling, which seemed to be the consensus among scientists.
3. If the planet is warming, is $CO_2$ the cause? There are greenhouse gases in the atmosphere, without which Earth would be considerably colder. These gases include methane, water vapor, and carbon dioxide. All occur naturally in nature. Carbon dioxide is vital to our life on Earth because it is necessary for growing plants. The amount of $CO_2$ produced by fossil fuels is a relatively small proportion of all the $CO_2$ in the atmosphere.

The generally accepted procedure is to record monthly temperature anomalies. To do so, we calculate the average for each month over many years. We then calculate any deviations between the latest month's temperature reading and its average. A positive anomaly would represent a month's temperature that is above the average. A negative anomaly indicates a month where the temperature is less than the average. One key question is how we measure the temperature.

Although there are many different sources of data, we have chosen to provide you with one, the National Climatic Data Center (NCDC), which is affiliated with the National Oceanic and

(Case 3.1 continued)

Atmospheric Administration (NOAA). (Other sources tend to agree with the NCDC's data.) C03-01a stores the monthly temperature anomalies from 1880 to 2009.

The best measures of $CO_2$ levels in the atmosphere come from the Mauna Loa Observatory in Hawaii, which has measured this variable since December 1958. However, attempts to estimate $CO_2$ levels prior to 1958 are as controversial as the methods used to estimate temperatures. These techniques include taking ice-core samples from the arctic and measuring the amount of carbon dioxide trapped in the ice from which estimates of atmospheric $CO_2$ are produced. To avoid this controversy, we will use the Mauna Loa Observatory numbers only. These data are stored in file C03-01b. (Note that some of the original data are missing and were replaced by interpolated values.)

1. Use whichever techniques you wish to determine whether there is global warming.
2. Use a graphical technique to determine whether there is a relationship between temperature anomalies and $CO_2$ levels.

## CASE 3.2 Economic Freedom and Prosperity

DATA
C03-02a
C03-02b

Adam Smith published *The Wealth of Nations* in 1776. In that book he argued that when institutions protect the liberty of individuals, greater prosperity results for all. Since 1995, the *Wall Street Journal* and the Heritage Foundation, a think tank in Washington, D.C., have produced the Index of Economic Freedom for all countries in the world. The index is based on a subjective score for 10 freedoms: business freedom, trade freedom, fiscal freedom, government size, monetary freedom, investment freedom, financial freedom, property rights, freedom from corruption, and labor freedom. We downloaded the scores for the years 1995 to 2009 and stored them in C03-02a. From the *CIA Factbook*, we determined the gross domestic product (GDP), measured in terms purchasing power parity (PPP), which makes it possible to compare the GDP for all countries. The GDP PPP figures for 2008 (the latest year available) are stored in C03-02b. Use the 2009 Freedom Index scores, the GDP PPP figures, and a graphical technique to see how freedom and prosperity are related.

© R.L./Shutterstock

# 4

# NUMERICAL DESCRIPTIVE TECHNIQUES

## The Cost of One More Win in Major League Baseball

© AP Photo/Charles Krupa

DATA
Xm04-00

In the era of free agency, professional sports teams must compete for the services of the best players. It is generally believed that only teams whose salaries place them in the top quarter have a chance of winning the championship. Efforts have been made to provide balance by establishing salary caps or some form of equalization. To examine the problem, we gathered data from the 2009 baseball season. For each team in major league baseball, we recorded the number of wins and the team payroll.

To make informed decisions, we need to know how the number of wins and the team payroll are related. After the statistical technique is presented, we return to this problem and solve it.

## Introduction

In Chapters 2 and 3, we presented several graphical techniques that describe data. In this chapter we introduce numerical descriptive techniques that allow the statistics practitioner to be more precise in describing various characteristics of a sample or population. These techniques are critical to the development of statistical inference.

As we pointed out in Chapter 2, arithmetic calculations can be applied to interval data only. Consequently, most of the techniques introduced here may be used only to numerically describe interval data. However, some of the techniques can be used for ordinal data, and one of the techniques can be employed for nominal data.

When we introduced the histogram, we commented that there are several bits of information that we look for. The first is the location of the center of the data. In Section 4.1, we will present **measures of central location**. Another important characteristic that we seek from a histogram is the spread of the data. The spread will be measured more precisely by measures of variability, which we present in Section 4.2. Section 4.3 introduces measures of relative standing and another graphical technique, the box plot.

In Section 3.3, we introduced the scatter diagram, which is a graphical method that we use to analyze the relationship between two interval variables. The numerical counterparts to the scatter diagram are called *measures of linear relationship*, and they are presented in Section 4.4.

Sections 4.5 and 4.6 feature statistical applications in baseball and finance, respectively. In Section 4.7, we compare the information provided by graphical and numerical techniques. Finally, we complete this chapter by providing guidelines on how to explore data and retrieve information.

## Sample Statistic or Population Parameter

Recall the terms introduced in Chapter 1: population, sample, parameter, and statistic. A parameter is a descriptive measurement about a population, and a statistic is a descriptive measurement about a sample. In this chapter, we introduce a dozen descriptive measurements. For each one, we describe how to calculate both the population parameter and the sample statistic. However, in most realistic applications, populations are very large—in fact, virtually infinite. The formulas describing the calculation of parameters are not practical and are seldom used. They are provided here primarily to teach the concept and the notation. In Chapter 7, we introduce probability distributions, which describe populations. At that time we show how parameters are calculated from probability distributions. In general, small data sets of the type we feature in this book are samples.

## 4.1 Measures of Central Location

### Arithmetic Mean

There are three different measures that we use to describe the center of a set of data. The first is the best known, the *arithmetic mean*, which we'll refer to simply as the **mean**. Students may be more familiar with its other name, the *average*. The mean is computed by summing the observations and dividing by the number of observations.

We label the observations in a sample $x_1, x_2, \ldots, x_n$, where $x_1$ is the first observation, $x_2$ is the second, and so on until $x_n$, where $n$ is the sample size. As a result, the sample mean is denoted $\bar{x}$. In a population, the number of observations is labeled $N$ and the population mean is denoted by $\mu$ (Greek letter *mu*).

**Mean**

$$\textbf{Population mean: } \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\textbf{Sample mean: } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

## EXAMPLE 4.1 Mean Time Spent on the Internet

A sample of 10 adults was asked to report the number of hours they spent on the Internet the previous month. The results are listed here. Manually calculate the sample mean.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 7 | 12 | 5 | 33 | 14 | 8 | 0 | 9 | 22 |

SOLUTION

Using our notation, we have $x_1 = 0, x_2 = 7, \ldots, x_{10} = 22$, and $n = 10$. The sample mean is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0 + 7 + 12 + 5 + 33 + 14 + 8 + 0 + 9 + 22}{10} = \frac{110}{10} = 11.0$$

## EXAMPLE 4.2 Mean Long-Distance Telephone Bill

DATA
Xm03-01

Refer to Example 3.1. Find the mean long-distance telephone bill.

SOLUTION

To calculate the mean, we add the observations and divide the sum by the size of the sample. Thus,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{42.19 + 38.45 + \cdots + 45.77}{200} = \frac{8717.52}{200} = 43.59$$

## Using the Computer

There are several ways to command Excel and Minitab to compute the mean. If we simply want to compute the mean and no other statistics, we can proceed as follows.

### EXCEL

*INSTRUCTIONS*

Type or import the data into one or more columns. (Open Xm03-01.) Type into any empty cell

=**AVERAGE**([Input range])

For Example 4.2, we would type into any cell

=**AVERAGE**(A1:A201)

The active cell would store the mean as 43.5876.

### MINITAB

*INSTRUCTIONS*

1. Type or import the data into one column. (Open Xm03-01.)
2. Click **Calc** and **Column Statistics . . . .** Specify **Mean** in the **Statistic** box. Type or use the **Select** button to specify the **Input variable** and click **OK**. The sample mean is outputted in the session window as 43.5876.

## Median

The second most popular measure of central location is the *median*.

**Median**

The **median** is calculated by placing all the observations in order (ascending or descending). The observation that falls in the middle is the median. The sample and population medians are computed in the same way.

When there is an even number of observations, the median is determined by averaging the two observations in the middle.

EXAMPLE 4.3

### Median Time Spent on Internet

Find the median for the data in Example 4.1.

SOLUTION

When placed in ascending order, the data appear as follows:

0 0 5 7 8 9 12 14 22 33

The median is the average of the fifth and sixth observations (the middle two), which are 8 and 9, respectively. Thus, the median is 8.5.

## EXAMPLE 4.4 Median Long-Distance Telephone Bill

DATA
Xm03-01

Find the median of the 200 observations in Example 3.1.

### SOLUTION

All the observations were placed in order. We observed that the 100th and 101st observations are 26.84 and 26.97, respectively. Thus, the median is the average of these two numbers:

$$\text{Median} = \frac{26.84 + 26.97}{2} = 26.905$$

### EXCEL

*INSTRUCTIONS*

To calculate the median, substitute **MEDIAN** in place of **AVERAGE** in the instructions for the mean (page 100). The median is reported as 26.905.

### MINITAB

*INSTRUCTIONS*

Follow the instructions for the mean to compute the mean except click **Median** instead of **Mean**. The median is outputted as 26.905 in the session window.

### INTERPRET

Half the observations are below 26.905, and half the observations are above 26.905.

## Mode

The third and last measure of central location that we present here is the *mode*.

**Mode**

The **mode** is defined as the observation (or observations) that occurs with the greatest frequency. Both the statistic and parameter are computed in the same way.

For populations and large samples, it is preferable to report the **modal class**, which we defined in Chapter 2.

There are several problems with using the mode as a measure of central location. First, in a small sample it may not be a very good measure. Second, it may not be unique.

## EXAMPLE 4.5 Mode Time Spent on Internet

Find the mode for the data in Example 4.1.

### SOLUTION

All observations except 0 occur once. There are two 0s. Thus, the mode is 0. As you can see, this is a poor measure of central location. It is nowhere near the center of the data. Compare this with the mean 11.0 and median 8.5 and you can appreciate that in this example the mean and median are superior measures.

## EXAMPLE 4.6 Mode of Long-Distance Bill

DATA
Xm03-01

Determine the mode for Example 3.1.

### SOLUTION

An examination of the 200 observations reveals that, except for 0, it appears that each number is unique. However, there are 8 zeroes, which indicates that the mode is 0.

### EXCEL

*INSTRUCTIONS*

To compute the mode, substitute **MODE** in place of **AVERAGE** in the previous instructions. Note that if there is more than one mode, Excel prints only the smallest one, without indicating whether there are other modes. In this example, Excel reports that the mode is 0.

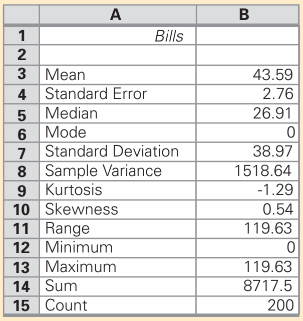
### MINITAB

Follow the instructions to compute the mean except click **Mode** instead of **Mean**. The mode is outputted as 0 in the session window. (See page 20.)

**Excel and Minitab: Printing All the Measures of Central Location plus Other Statistics** Both Excel and Minitab can produce the measures of central location plus a variety of others that we will introduce in later sections.

## EXCEL

### Excel Output for Examples 4.2, 4.4, and 4.6



|  | A | B |
|---|---|---|
| 1 | Bills |  |
| 2 |  |  |
| 3 | Mean | 43.59 |
| 4 | Standard Error | 2.76 |
| 5 | Median | 26.91 |
| 6 | Mode | 0 |
| 7 | Standard Deviation | 38.97 |
| 8 | Sample Variance | 1518.64 |
| 9 | Kurtosis | -1.29 |
| 10 | Skewness | 0.54 |
| 11 | Range | 119.63 |
| 12 | Minimum | 0 |
| 13 | Maximum | 119.63 |
| 14 | Sum | 8717.5 |
| 15 | Count | 200 |

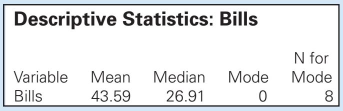
Excel reports the mean, median, and mode as the same values we obtained previously. Most of the other statistics will be discussed later.

*INSTRUCTIONS*

1. Type or import the data into one column. (Open Xm03-01.)
2. Click **Data, Data Analysis**, and **Descriptive Statistics.**
3. Specify the **Input Range** (A1:A201) and click **Summary statistics**.

## MINITAB

### Minitab Output for Examples 4.2, 4.4, and 4.6



```
Descriptive Statistics: Bills

                                      N for
Variable   Mean   Median   Mode       Mode
Bills      43.59   26.91      0          8
```

*INSTRUCTIONS*

1. Type or import the data into one column. (Open Xm03-01.)
2. Click **Stat, Basic Statistics,** and **Display Descriptive Statistics . . . .**
3. Type or use **Select** to identify the name of the variable or column (Bills). Click **Statistics . . .** to add or delete particular statistics.

## Mean, Median, Mode: Which Is Best?

With three measures from which to choose, which one should we use? There are several factors to consider when making our choice of measure of central location. The mean is generally our first selection. However, there are several circumstances when the median is better. The mode is seldom the best measure of central location. One advantage the median holds is that it is not as sensitive to extreme values as is the mean.

To illustrate, consider the data in Example 4.1. The mean was 11.0, and the median was 8.5. Now suppose that the respondent who reported 33 hours actually reported 133 hours (obviously an Internet addict). The mean becomes

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0 + 7 + 12 + 5 + 133 + 14 + 8 + 0 + 22}{10} = \frac{210}{10} = 21.0$$

This value is exceeded by only 2 of the 10 observations in the sample, making this statistic a poor measure of *central* location. The median stays the same. When there is a relatively small number of extreme observations (either very small or very large, but not both), the median usually produces a better measure of the center of the data.

To see another advantage of the median over the mean, suppose you and your classmates have written a statistics test and the instructor is returning the graded tests. What piece of information is most important to you? The answer, of course, is *your* mark. What is the next important bit of information? The answer is how well you performed relative to the class. Most students ask their instructor for the class mean. This is the wrong statistic to request. You want the *median* because it divides the class into two halves. This information allows you to identify which half of the class your mark falls into. The median provides this information; the mean does not. Nevertheless, the mean can also be useful in this scenario. If there are several sections of the course, the section means can be compared to determine whose class performed best (or worst).

## Measures of Central Location for Ordinal and Nominal Data

When the data are interval, we can use any of the three measures of central location. However, for ordinal and nominal data, the calculation of the mean is not valid. Because the calculation of the median begins by placing the data in order, this statistic is appropriate for ordinal data. The mode, which is determined by counting the frequency of each observation, is appropriate for nominal data. However, nominal data do not have a "center," so we cannot interpret the mode of nominal data in that way. It is generally pointless to compute the mode of nominal data.

## APPLICATIONS in FINANCE

© Image State Royalty-free

### Geometric Mean

The arithmetic mean is the single most popular and useful measure of central location. We noted certain situations where the median is a better measure of central location. However, there is another circumstance where neither the mean nor the median is the best measure. When the variable is a growth rate or rate of change, such as the value of an investment over periods of time, we need another measure. This will become apparent from the following illustration.

Suppose you make a 2-year investment of $1,000, and it grows by 100% to $2,000 during the first year. During the second year, however, the investment suffers a 50% loss, from $2,000 back to $1,000. The rates of return for years 1 and 2 are $R_1 = 100\%$ and $R_2 = -50\%$, respectively. The arithmetic mean (and the median) is computed as

$$\overline{R} = \frac{R_1 + R_2}{2} = \frac{100 + (-50)}{2} = 25\%$$

But this figure is misleading. Because there was no change in the value of the investment from the beginning to the end of the 2-year period, the "average" compounded rate of return is 0%. As you will see, this is the value of the *geometric mean.*

Let $R_i$ denote the rate of return (in decimal form) in period $i$ ($i = 1, 2, \ldots, n$). The **geometric mean** $R_g$ of the returns $R_1, R_2, \ldots, R_n$ is defined such that

$$(1 + R_g)^n = (1 + R_1)(1 + R_2) \cdots (1 + R_n)$$

Solving for $R_g$, we produce the following formula:

$$R_g = \sqrt[n]{(1 + R_1)(1 + R_2) \cdots (1 + R_n)} - 1$$

The geometric mean of our investment illustration is

$$R_g = \sqrt[n]{(1 + R_1)(1 + R_2) \cdots (1 + R_n)} - 1 = \sqrt[2]{(1 + 1)(1 + [-.50])} - 1 = 1 - 1 = 0$$

The geometric mean is therefore 0%. This is the single "average" return that allows us to compute the value of the investment at the end of the investment period from the beginning value. Thus, using the formula for compound interest with the rate = 0%, we find

$$\text{Value at the end of the investment period} = 1{,}000(1 + R_g)^2 = 1{,}000(1 + 0)^2 = 1{,}000$$

The geometric mean is used whenever we wish to find the "average" growth rate, or rate of change, in a variable *over time*. However, the arithmetic mean of $n$ returns (or growth rates) is the appropriate mean to calculate if you wish to estimate the mean rate of return (or growth rate) for any *single* period in the future; that is, in the illustration above if we wanted to estimate the rate of return in year 3, we would use the arithmetic mean of the two annual rates of return, which we found to be 25%.

## EXCEL

### *INSTRUCTIONS*

1. Type or import the values of $1 + R_i$ into a column.
2. Follow the instructions to produce the mean (page 100) except substitute **GEOMEAN** in place of **AVERAGE**.
3. To determine the geometric mean, subtract 1 from the number produced.

## MINITAB

Minitab does not compute the geometric mean.

Here is a summary of the numerical techniques introduced in this section and when to use them.

### Factors That Identify When to Compute the Mean

1. **Objective**: Describe a single set of data
2. **Type of data**: Interval
3. **Descriptive measurement**: Central location

### Factors That Identify When to Compute the Median

1. **Objective**: Describe a single set of data
2. **Type of data**: Ordinal or interval (with extreme observations)
3. **Descriptive measurement**: Central location

### Factors That Identify When to Compute the Mode

1. **Objective**: Describe a single set of data
2. **Type of data**: Nominal, ordinal, interval

### Factors That Identify When to Compute the Geometric Mean

1. **Objective**: Describe a single set of data
2. **Type of data**: Interval; growth rates

## Exercises

**4.1** A sample of 12 people was asked how much change they had in their pockets and wallets. The responses (in cents) are

52 25 15 0 104 44
60 30 33 81 40 5

Determine the mean, median, and mode for these data.

**4.2** The number of sick days due to colds and flu last year was recorded by a sample of 15 adults. The data are

5 7 0 3 15 6 5 9
3 8 10 5 2 0 12

Compute the mean, median, and mode.

**4.3** A random sample of 12 joggers was asked to keep track and report the number of miles they ran last week. The responses are

5.5 7.2 1.6 22.0 8.7 2.8
5.3 3.4 12.5 18.6 8.3 6.6

a. Compute the three statistics that measure central location.
b. Briefly describe what each statistic tells you.

**4.4** The midterm test for a statistics course has a time limit of 1 hour. However, like most statistics exams this one was quite easy. To assess how easy, the professor recorded the amount of time taken by a sample of nine students to hand in their test papers. The times (rounded to the nearest minute) are

33 29 45 60 42 19 52 38 36

a. Compute the mean, median, and mode.
b. What have you learned from the three statistics calculated in part (a)?

**4.5** The professors at Wilfrid Laurier University are required to submit their final exams to the registrar's office 10 days before the end of the semester. The exam coordinator sampled 20 professors and recorded the number of days before the final exam that each submitted his or her exam. The results are

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 14 | 8 | 3 | 2 | 6 | 4 | 9 | 13 | 10 | 12 |
| 7 | 4 | 9 | 13 | 15 | 8 | 11 | 12 | 4 | 0 |

a. Compute the mean, median, and mode.
b. Briefly describe what each statistic tells you.

**4.6** Compute the geometric mean of the following rates of return.

.25 −.10 .50

**4.7** What is the geometric mean of the following rates of return?

.50 .30 −.50 −.25

**4.8** The following returns were realized on an investment over a 5-year period.

| Year | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Rate of Return | .10 | .22 | .06 | −.05 | .20 |

a. Compute the mean and median of the returns.
b. Compute the geometric mean.
c. Which one of the three statistics computed in parts (a) and (b) best describes the return over the 5-year period? Explain.

**4.9** An investment you made 5 years ago has realized the following rates of return.

| Year | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Rate of Return | −.15 | −.20 | .15 | −.08 | .50 |

a. Compute the mean and median of the rates of return.
b. Compute the geometric mean.
c. Which one of the three statistics computed in parts (a) and (b) best describes the return over the 5-year period? Explain.

**4.10** An investment of $1,000 you made 4 years ago was worth $1,200 after the first year, $1,200 after the second year, $1,500 after the third year, and $2,000 today.
a. Compute the annual rates of return.
b. Compute the mean and median of the rates of return.
c. Compute the geometric mean.
d. Discuss whether the mean, median, or geometric mean is the best measure of the performance of the investment.

**4.11** Suppose that you bought a stock 6 years ago at $12.The stock's price at the end of each year is shown here.

| Year | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Price | 10 | 14 | 15 | 22 | 30 | 25 |

a. Compute the rate of return for each year.
b. Compute the mean and median of the rates of return.
c. Compute the geometric mean of the rates of return.
d. Explain why the best statistic to use to describe what happened to the price of the stock over the 6-year period is the geometric mean.

*The following exercises require the use of a computer and software.*

**4.12** Xr04-12 The starting salaries of a sample of 125 recent MBA graduates are recorded.
a. Determine the mean and median of these data.
b. What do these two statistics tell you about the starting salaries of MBA graduates?

**4.13** Xr04-13 To determine whether changing the color of its invoices would improve the speed of payment, a company selected 200 customers at random and sent their invoices on blue paper. The number of days until the bills were paid was recorded. Calculate the mean and median of these data. Report what you have discovered.

**4.14** Xr04-14 A survey undertaken by the U.S. Bureau of Labor Statistics, Annual Consumer Expenditure, asks American adults to report the amount of money spent on reading material in 2006. (*Source:* Adapted from *Statistical Abstract of the United States*, *2009*, Table 664.)
a. Compute the mean and median of the sample.
b. What do the statistics computed in part (a) tell you about the reading materials expenditures?

**4.15** Xr04-15 A survey of 225 workers in Los Angeles and 190 workers in New York asked each to report the average amount of time spent commuting to work. (*Source:* Adapted from *Statistical Abstract of the United States*, *2009*, Table 1060.)
a. Compute the mean and median of the commuting times for workers in Los Angeles.
b. Repeat part (a) for New York workers.
c. Summarize your findings.

**4.16** Xr04-16 In the United States, banks and financial institutions often require buyers of houses to pay fees in order to arrange mortgages. In a survey conducted by the U.S. Federal Housing Finance Board, 350 buyers of new houses who received a mortgage from a bank were asked to report the amount of fees (fees include commissions, discounts, and points) they paid as a percentage of the whole mortgage. (*Source:* Adapted from *Statistical Abstract of the United States*, *2009*, Table 1153.)
a. Compute the mean and median.
b. Interpret the statistics you computed.

**4.17** Xr04-17 In an effort to slow drivers, traffic engineers painted a solid line 3 feet from the curb over the entire length of a road and filled the space with diagonal lines. The lines made the road look narrower. A sample of car speeds was taken after the lines were drawn.

a. Compute the mean, median, and mode of these data.
b. Briefly describe the information you acquired from each statistic calculated in part (a).

**4.18** Xr04-18 How much do Americans spend on various food groups? Two hundred American families were surveyed and asked to report the amount of money spent annually on fruits and vegetables. Compute the mean and median of these data and interpret the results. (*Source:* Adapted from *Statistical Abstract of the United States, 2009*, Table 662.)

# 4.2 Measures of Variability

The statistics introduced in Section 4.1 serve to provide information about the central location of the data. However, as we have already discussed in Chapter 2, there are other characteristics of data that are of interest to practitioners of statistics. One such characteristic is the spread or variability of the data. In this section, we introduce four **measures of variability**. We begin with the simplest.

## Range

**Range**

Range = Largest observation − Smallest observation

The advantage of the **range** is its simplicity. The disadvantage is also its simplicity. Because the range is calculated from only two observations, it tells us nothing about the other observations. Consider the following two sets of data.

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Set 1: | 4 | 4 | 4 | 4 | 4 | 50 |
| Set 2: | 4 | 8 | 15 | 24 | 39 | 50 |

The range of both sets is 46. The two sets of data are completely different, yet their ranges are the same. To measure variability, we need other statistics that incorporate all the data and not just two observations.

## Variance

The **variance** and its related measure, the **standard deviation**, are arguably the most important statistics. They are used to measure variability, but, as you will discover, they play a vital role in almost all statistical inference procedures.

**Variance**

**Population variance:** $$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

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

The population variance is represented by $\sigma^2$(Greek letter *sigma* squared).

Examine the formula for the sample variance $s^2$. It may appear to be illogical that in calculating $s^2$ we divide by $n - 1$ rather than by $n$.* However, we do so for the following reason. Population parameters in practical settings are seldom known. One objective of statistical inference is to estimate the parameter from the statistic. For example, we estimate the population mean $\mu$ from the sample mean $\bar{x}$. Although it is not obviously logical, the statistic created by dividing $\sum(x_i - \bar{x})^2$ by $n - 1$ is a better estimator than the one created by dividing by $n$. We will discuss this issue in greater detail in Section 10.1.

To compute the sample variance $s^2$, we begin by calculating the sample mean $\bar{x}$. Next we compute the difference (also call the **deviation**) between each observation and the mean. We square the deviations and sum. Finally, we divide the sum of squared deviations by $n - 1$.

We'll illustrate with a simple example. Suppose that we have the following observations of the numbers of hours five students spent studying statistics last week:

8 4 9 11 3

The mean is

$$\bar{x} = \frac{8 + 4 + 9 + 11 + 3}{5} = \frac{35}{5} = 7$$

For each observation, we determine its deviation from the mean. The deviation is squared, and the sum of squares is determined as shown in Table 4.1.

TABLE **4.1** Calculation of Sample Variance

| $x_i$ | $(x_i - \bar{x})$ | $(x_i - \bar{x})^2$ |
|---|---|---|
| 8 | $(8 - 7) = 1$ | $(1)^2 = 1$ |
| 4 | $(4 - 7) = -3$ | $(-3)^2 = 9$ |
| 9 | $(9 - 7) = 2$ | $(2)^2 = 4$ |
| 11 | $(11 - 7) = 4$ | $(4)^2 = 16$ |
| 3 | $(3 - 7) = -4$ | $(-4)^2 = 16$ |
|  | $\sum_{l=1}^{5}(x_i - \bar{x}) = 0$ | $\sum_{l=1}^{5}(x_i - \bar{x})^2 = 46$ |

The sample variance is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1} = \frac{46}{5 - 1} = 11.5$$

The calculation of this statistic raises several questions. Why do we square the deviations before averaging? If you examine the deviations, you will see that some of the

*Technically, the variance of the sample is calculated by dividing the sum of squared deviations by $n$. The statistic computed by dividing the sum of squared deviations by $n - 1$ is called the *sample variance corrected for the mean*. Because this statistic is used extensively, we will shorten its name to *sample variance*.

deviations are positive and some are negative. When you add them together, the sum is 0. This will always be the case because the sum of the positive deviations will always equal the sum of the negative deviations. Consequently, we square the deviations to avoid the "canceling effect."

Is it possible to avoid the canceling effect without squaring? We could average the *absolute* value of the deviations. In fact, such a statistic has already been invented. It is called the **mean absolute deviation** or MAD. However, this statistic has limited utility and is seldom calculated.

What is the unit of measurement of the variance? Because we squared the deviations, we also squared the units. In this illustration the units were hours (of study). Thus, the sample variance is 11.5 hours$^2$.

## EXAMPLE 4.7 Summer Jobs

The following are the number of summer jobs a sample of six students applied for. Find the mean and variance of these data.

17 15 23 7 9 13

### SOLUTION

The mean of the six observations is

$$\bar{x} = \frac{17 + 15 + 23 + 7 + 9 + 13}{6} = \frac{84}{6} = 14 \text{ jobs}$$

The sample variance is

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

$$= \frac{(17 - 14)^2 + (15 - 14)^2 + (23 - 14)^2 + (7 - 14)^2 + (9 - 14)^2 + (13 - 14)^2}{6 - 1}$$

$$= \frac{9 + 1 + 81 + 49 + 25 + 1}{5} = \frac{166}{5} = 33.2 \text{ jobs}^2$$

**(Optional) Shortcut Method for Variance** The calculations for larger data sets are quite time-consuming. The following shortcut for the sample variance may help lighten the load.

**Shortcut for Sample Variance**

$$s^2 = \frac{1}{n - 1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

To illustrate, we'll do Example 4.7 again.

$$\sum_{I=1}^{n} x_i^2 = 17^2 + 15^2 + 23^2 + 7^2 + 9^2 + 13^2 = 1{,}342$$

$$\sum_{i=1}^{n} x_i = 17 + 15 + 23 + 7 + 9 + 13 = 84$$

$$\left(\sum_{i=1}^{n} x_i\right)^2 = 84^2 = 7{,}056$$

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{6-1}\left[1342 - \frac{7056}{6}\right] = 33.2 \text{ jobs}^2$$

Notice that we produced the same exact answer.

## EXCEL

*INSTRUCTIONS*

Follow the instructions to compute the mean (page 100) except type VAR instead of AVERAGE.

## MINITAB

1. Type or import data into one column.
2. Click **Stat, Basic Statistics, Display Descriptive Statistics . . .** , and select the variable.
3. Click **Statistics** and **Variance**.

## Interpreting the Variance

We calculated the variance in Example 4.7 to be 33.2 jobs$^2$. What does this statistic tell us? Unfortunately, the variance provides us with only a rough idea about the amount of variation in the data. However, this statistic is useful when comparing two or more sets of data of the same type of variable. If the variance of one data set is larger than that of a second data set, we interpret that to mean that the observations in the first set display more variation than the observations in the second set.

The problem of interpretation is caused by the way the variance is computed. Because we squared the deviations from the mean, the unit attached to the variance is the square of the unit attached to the original observations. In other words, in Example 4.7 the unit of the data is jobs; the unit of the variance is jobs squared. This contributes to the problem of interpretation. We resolve this difficulty by calculating another related measure of variability.

## Standard Deviation

**Standard Deviation**

Population standard deviation: $\sigma = \sqrt{\sigma^2}$

Sample standard deviation: $s = \sqrt{s^2}$

The standard deviation is simply the positive square root of the variance. Thus, in Example 4.7, the sample standard deviation is

$$s = \sqrt{s^2} = \sqrt{33.2} = 5.76 \text{ jobs}$$

Notice that the unit associated with the standard deviation is the unit of the original data set.

EXAMPLE 4.8

DATA Xm04-08

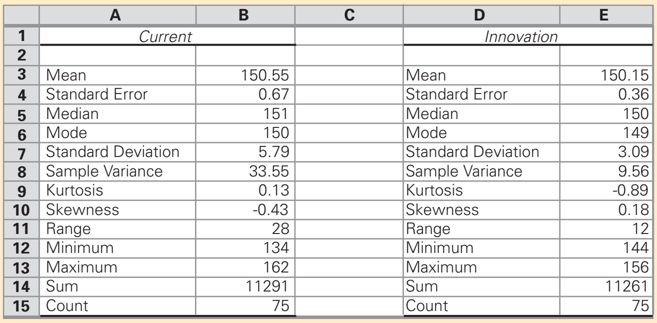
## Comparing the Consistency of Two Types of Golf Clubs

Consistency is the hallmark of a good golfer. Golf equipment manufacturers are constantly seeking ways to improve their products. Suppose that a recent innovation is designed to improve the consistency of its users. As a test, a golfer was asked to hit 150 shots using a 7 iron, 75 of which were hit with his current club and 75 with the new innovative 7 iron. The distances were measured and recorded. Which 7 iron is more consistent?

### SOLUTION

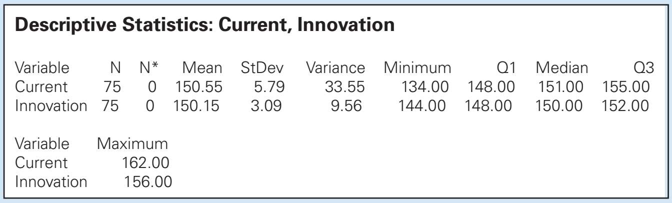
To gauge the consistency, we must determine the standard deviations. (We could also compute the variances, but as we just pointed out, the standard deviation is easier to interpret.) We can get Excel and Minitab to print the sample standard deviations. Alternatively, we can calculate all the descriptive statistics, a course of action we recommend because we often need several statistics. The printouts for both 7 irons are shown here.

### EXCEL



| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | *Current* | | | *Innovation* | |
| 2 | | | | | |
| 3 | Mean | 150.55 | | Mean | 150.15 |
| 4 | Standard Error | 0.67 | | Standard Error | 0.36 |
| 5 | Median | 151 | | Median | 150 |
| 6 | Mode | 150 | | Mode | 149 |
| 7 | Standard Deviation | 5.79 | | Standard Deviation | 3.09 |
| 8 | Sample Variance | 33.55 | | Sample Variance | 9.56 |
| 9 | Kurtosis | 0.13 | | Kurtosis | -0.89 |
| 10 | Skewness | -0.43 | | Skewness | 0.18 |
| 11 | Range | 28 | | Range | 12 |
| 12 | Minimum | 134 | | Minimum | 144 |
| 13 | Maximum | 162 | | Maximum | 156 |
| 14 | Sum | 11291 | | Sum | 11261 |
| 15 | Count | 75 | | Count | 75 |

## MINITAB



```
Descriptive Statistics: Current, Innovation

Variable     N   N*    Mean   StDev   Variance  Minimum      Q1   Median      Q3
Current     75    0  150.55    5.79      33.55   134.00  148.00   151.00  155.00
Innovation  75    0  150.15    3.09       9.56   144.00  148.00   150.00  152.00

Variable    Maximum
Current      162.00
Innovation   156.00
```

## INTERPRET

The standard deviation of the distances of the current 7 iron is 5.79 yards whereas that of the innovative 7 iron is 3.09 yards. Based on this sample, the innovative club is more consistent. Because the mean distances are similar it would appear that the new club is indeed superior.

**Interpreting the Standard Deviation** Knowing the mean and standard deviation allows the statistics practitioner to extract useful bits of information. The information depends on the shape of the histogram. If the histogram is bell shaped, we can use the **Empirical Rule**.

### Empirical Rule

1. Approximately 68% of all observations fall within one standard deviation of the mean.
2. Approximately 95% of all observations fall within two standard deviations of the mean.
3. Approximately 99.7% of all observations fall within three standard deviations of the mean.

EXAMPLE 4.9

## Using the Empirical Rule to Interpret Standard Deviation

After an analysis of the returns on an investment, a statistics practitioner discovered that the histogram is bell shaped and that the mean and standard deviation are 10% and 8%, respectively. What can you say about the way the returns are distributed?

### SOLUTION

Because the histogram is bell shaped, we can apply the Empirical Rule:

1. Approximately 68% of the returns lie between 2% (the mean minus one standard deviation = 10 − 8) and 18% (the mean plus one standard deviation = 10 + 8).

2. Approximately 95% of the returns lie between −6% [the mean minus two standard deviations = 10 − 2(8)] and 26% [the mean plus two standard deviations = 10 + 2(8)].
3. Approximately 99.7% of the returns lie between −14% [the mean minus three standard deviations = 10 − 3(8)] and 34% [the mean plus three standard deviations = 10 − 3(8)].

A more general interpretation of the standard deviation is derived from *Chebysheff's Theorem*, which applies to all shapes of histograms.

**Chebysheff's Theorem**

The proportion of observations in any sample or population that lie within $k$ standard deviations of the mean is at least

$$1 - \frac{1}{k^2} \quad \text{for} \quad k > 1$$

When $k = 2$, **Chebysheff's Theorem** states that at least three-quarters (75%) of all observations lie within two standard deviations of the mean. With $k = 3$, Chebysheff's Theorem states that at least eight-ninths (88.9%) of all observations lie within three standard deviations of the mean.

Note that the Empirical Rule provides approximate proportions, whereas Chebysheff's Theorem provides lower bounds on the proportions contained in the intervals.

## EXAMPLE 4.10 Using Chebysheff's Theorem to Interpret Standard Deviation

The annual salaries of the employees of a chain of computer stores produced a positively **skewed** histogram. The mean and standard deviation are \$28,000 and \$3,000, respectively. What can you say about the salaries at this chain?

### SOLUTION

Because the histogram is not bell shaped, we cannot use the Empirical Rule. We must employ Chebysheff's Theorem instead.

The intervals created by adding and subtracting two and three standard deviations to and from the mean are as follows:

1. At least 75% of the salaries lie between \$22,000 [the mean minus two standard deviations = 28,000 − 2(3,000)] and \$34,000 [the mean plus two standard deviations = 28,000 + 2(3,000)].
2. At least 88.9% of the salaries lie between \$19,000 [the mean minus three standard deviations = 28,000 − 3(3,000)] and \$37,000 [the mean plus three standard deviations = 28,000 + 3(3,000)].

## Coefficient of Variation

Is a standard deviation of 10 a large number indicating great variability or a small number indicating little variability? The answer depends somewhat on the magnitude of the observations in the data set. If the observations are in the millions, then a standard deviation of 10 will probably be considered a small number. On the other hand, if the observations are less than 50, then the standard deviation of 10 would be seen as a large number. This logic lies behind yet another measure of variability, the *coefficient of variation*.

### Coefficient of Variation

The **coefficient of variation** of a set of observations is the standard deviation of the observations divided by their mean:

$$\text{Population coefficient of variation: CV} = \frac{\sigma}{\mu}$$

$$\text{Sample coefficient of variation: cv} = \frac{s}{\bar{x}}$$

## Measures of Variability for Ordinal and Nominal Data

The measures of variability introduced in this section can be used only for interval data. The next section will feature a measure that can be used to describe the variability of ordinal data. There are no measures of variability for nominal data.

## Approximating the Mean and Variance from Grouped Data

The statistical methods presented in this chapter are used to compute descriptive statistics from data. However, in some circumstances, the statistics practitioner does not have the raw data but instead has a frequency distribution. This is often the case when data are supplied by government organizations. In Appendix Approximating Means and Variances for Grouped Data on Keller's website we provide the formulas used to approximate the sample mean and variance.

We complete this section by reviewing the factors that identify the use of measures of variability.

### Factors That Identify When to Compute the Range, Variance, Standard Deviation, and Coefficient of Variation

1. **Objective**: Describe a single set of data
2. **Type of Data**: Interval
3. **Descriptive measurement**: Variability

## Exercises

**4.19** Calculate the variance of the following data.

9 3 7 4 1 7 5 4

**4.20** Calculate the variance of the following data.

4 5 3 6 5 6 5 6

**4.21** Determine the variance and standard deviation of the following sample.

12 6 22 31 23 13 15 17 21

**4.22** Find the variance and standard deviation of the following sample.

0 −5 −3 6 4 −4 1 −5 0 3

**4.23** Examine the three samples listed here. Without performing any calculations, indicate which sample has the largest amount of variation and which sample has the smallest amount of variation. Explain how you produced your answer.

| | | | | | |
|---|---|---|---|---|---|
| a. | 17 | 29 | 12 | 16 | 11 |
| b. | 22 | 18 | 23 | 20 | 17 |
| c. | 24 | 37 | 6 | 39 | 29 |

**4.24** Refer to Exercise 4.23. Calculate the variance for each part. Was your answer in Exercise 4.23 correct?

**4.25** A friend calculates a variance and reports that it is –25.0. How do you know that he has made a serious calculation error?

**4.26** Create a sample of five numbers whose mean is 6 and whose standard deviation is 0.

**4.27** A set of data whose histogram is bell shaped yields a mean and standard deviation of 50 and 4, respectively. Approximately what proportion of observations

a. are between 46 and 54?
b. are between 42 and 58?
c. are between 38 and 62?

**4.28** Refer to Exercise 4.27. Approximately what proportion of observations

a. are less than 46?
b. are less than 58?
c. are greater than 54?

**4.29** A set of data whose histogram is extremely skewed yields a mean and standard deviation of 70 and 12, respectively. What is the minimum proportion of observations that

a. are between 46 and 94?
b. are between 34 and 106?

**4.30** A statistics practitioner determined that the mean and standard deviation of a data set were 120 and 30, respectively. What can you say about the proportions of observations that lie between each of the following intervals?

a. 90 and 150
b. 60 and 180
c. 30 and 210

*The following exercises require a computer and software.*

**4.31** Xr04-31 There has been much media coverage of the high cost of medicinal drugs in the United States. One concern is the large variation from pharmacy to pharmacy. To investigate, a consumer advocacy group took a random sample of 100 pharmacies around the country and recorded the price (in dollars per 100 pills) of Prozac. Compute the range, variance, and standard deviation of the prices. Discuss what these statistics tell you.

**4.32** Xr04-32 Many traffic experts argue that the most important factor in accidents is not the average speed of cars but the amount of variation. Suppose that the speeds of a sample of 200 cars were taken over a stretch of highway that has seen numerous accidents. Compute the variance and standard deviation of the speeds, and interpret the results.

**4.33** Xr04-33 Three men were trying to make the football team as punters. The coach had each of them punt the ball 50 times, and the distances were recorded.

a. Compute the variance and standard deviation for each punter.
b. What do these statistics tell you about the punters?

**4.34** Xr04-34 Variance is often used to measure quality in production-line products. Suppose that a sample of steel rods that are supposed to be exactly 100 cm long is taken. The length of each is determined, and the results are recorded. Calculate the variance and the standard deviation. Briefly describe what these statistics tell you.

**4.35** Xr04-35 To learn more about the size of withdrawals at a banking machine, the proprietor took a sample of 75 withdrawals and recorded the amounts. Determine the mean and standard deviation of these data, and describe what these two statistics tell you about the withdrawal amounts.

**4.36** Xr04-36 Everyone is familiar with waiting lines or queues. For example, people wait in line at a supermarket to go through the checkout counter. There are two factors that determine how long the queue becomes. One is the speed of service. The other is the number of arrivals at the checkout counter. The mean number of arrivals is an important number, but so is the standard deviation. Suppose that a consultant for the supermarket counts the number of arrivals per hour during a sample of 150 hours.

a. Compute the standard deviation of the number of arrivals.
b. Assuming that the histogram is bell shaped, interpret the standard deviation.

## American National Election Survey Exercises

**4.37** ANES2008* The ANES in 2008 asked respondents to state their ages stored as AGE.

a. Calculate the mean, variance, and standard deviation.
b. Draw a histogram.
c. Use the Empirical Rule, if applicable, or Chebysheff's Theorem to interpret the mean and standard deviation.

**4.38** ANES2008* Respondents were asked to report the number of minutes spent watching news on television during a typical day (TIME2).

a. Calculate the mean and standard deviation.
b. Draw a histogram.
c. Use the Empirical Rule, if applicable, or Chebysheff's Theorem to interpret the mean and standard deviation.

## General Social Survey Exercise

**4.39** GSS2008* One of the questions in the 2008 General Social Survey was, If you were born outside the United States, at what age did you permanently move to the United States (AGECMEUS)?

a. Calculate the mean, variance, and standard deviation.
b. Draw a histogram.
c. Use the Empirical Rule, if applicable, or Chebysheff's Theorem to interpret the mean and standard deviation.

# 4.3 Measures of Relative Standing and Box Plots

Measures of relative standing are designed to provide information about the position of particular values relative to the entire data set. We've already presented one measure of relative standing, the median, which is also a measure of central location. Recall that the median divides the data set into halves, allowing the statistics practitioner to determine which half of the data set each observation lies in. The statistics we're about to introduce will give you much more detailed information.

**Percentile**

The *P*th **percentile** is the value for which *P* percent are less than that value and (100–*P*)% are greater than that value.

The scores and the percentiles of the Scholastic Achievement Test (SAT) and the Graduate Management Admission Test (GMAT), as well as various other admissions tests, are reported to students taking them. Suppose for example, that your SAT score is reported to be at the 60th percentile. This means that 60% of all the other marks are below yours and 40% are above it. You now know exactly where you stand relative to the population of SAT scores.

We have special names for the 25th, 50th, and 75th percentiles. Because these three statistics divide the set of data into quarters, these measures of relative standing are also called **quartiles**. The *first* or *lower quartile* is labeled $Q_1$. It is equal to the 25th percentile. The *second quartile*, $Q_2$, is equal to the 50th percentile, which is also the median. The *third* or *upper quartile*, $Q_3$, is equal to the 75th percentile. Incidentally, many people confuse the terms *quartile* and *quarter*. A common error is to state that someone is in the lower *quartile* of a group when they actually mean that someone is in the lower *quarter* of a group.

Besides quartiles, we can also convert percentiles into quintiles and deciles. *Quintiles* divide the data into fifths, and *deciles* divide the data into tenths.

## Locating Percentiles

The following formula allows us to approximate the location of any percentile.

**Location of a Percentile**

$$L_P = (n + 1)\frac{P}{100}$$

where $L_P$ is the location of the $P$th percentile.

## EXAMPLE 4.11 Percentiles of Time Spent on Internet

Calculate the 25th, 50th, and 75th percentiles (first, second, and third quartiles) of the data in Example 4.1.

### SOLUTION

Placing the 10 observations in ascending order we get

0 0 5 7 8 9 12 14 22 33

The location of the 25th percentile is

$$L_{25} = (10 + 1)\frac{25}{100} = (11)(.25) = 2.75$$

The 25th percentile is three-quarters of the distance between the second (which is 0) and the third (which is 5) observations. Three-quarters of the distance is

$$(.75)(5 - 0) = 3.75$$

Because the second observation is 0, the 25th percentile is 0 + 3.75 = 3.75.

To locate the 50th percentile, we substitute P = 50 into the formula and produce

$$L_{50} = (10 + 1)\frac{50}{100} = (11)(.5) = 5.5$$

which means that the 50th percentile is halfway between the fifth and sixth observations. The fifth and sixth observations are 8 and 9, respectively. The 50th percentile is 8.5. This is the median calculated in Example 4.3.

The 75th percentile's location is

$$L_{75} = (10 + 1)\frac{75}{100} = (11)(.75) = 8.25$$

Thus, it is located one-quarter of the distance between the eighth and ninth observations, which are 14 and 22, respectively. One-quarter of the distance is

$$(.25)(22 - 14) = 2$$

which means that the 75th percentile is

$$14 + 2 = 16$$

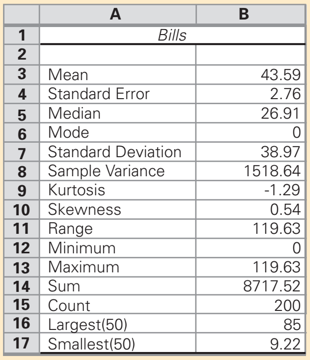
## EXAMPLE 4.12 Quartiles of Long-Distance Telephone Bills

DATA Xm03-01

Determine the quartiles for Example 3.1.

SOLUTION

EXCEL



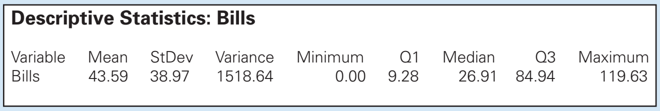
| | A | B |
|---|---|---|
| 1 | *Bills* | |
| 2 | | |
| 3 | Mean | 43.59 |
| 4 | Standard Error | 2.76 |
| 5 | Median | 26.91 |
| 6 | Mode | 0 |
| 7 | Standard Deviation | 38.97 |
| 8 | Sample Variance | 1518.64 |
| 9 | Kurtosis | -1.29 |
| 10 | Skewness | 0.54 |
| 11 | Range | 119.63 |
| 12 | Minimum | 0 |
| 13 | Maximum | 119.63 |
| 14 | Sum | 8717.52 |
| 15 | Count | 200 |
| 16 | Largest(50) | 85 |
| 17 | Smallest(50) | 9.22 |

*INSTRUCTIONS*

Follow the instructions for **Descriptive Statistics** (page 103). In the dialog box, click **Kth Largest** and type in the integer closest to $n/4$. Repeat for **Kth Smallest**, typing in the integer closest to $n/4$.

Excel approximates the third and first percentiles in the following way. The **Largest(50)** is 85, which is the number such that 150 numbers are below it and 49 numbers are above it. The **Smallest(50)** is 9.22, which is the number such that 49 numbers are below it and 150 numbers are above it. The median is 26.91, a statistic we discussed in Example 4.4.

## MINITAB



```
Descriptive Statistics: Bills

Variable  Mean   StDev  Variance  Minimum    Q1   Median     Q3   Maximum
Bills     43.59  38.97  1518.64      0.00  9.28    26.91  84.94    119.63
```

Minitab outputs the first and third quartiles as Q1 (9.28) and Q3 (84.94), respectively. (See page 103.)

We can often get an idea of the shape of the histogram from the quartiles. For example, if the first and second quartiles are closer to each other than are the second and third quartiles, then the histogram is positively skewed. If the first and second quartiles are farther apart than the second and third quartiles, then the histogram is negatively skewed. If the difference between the first and second quartiles is approximately equal to the difference between the second and third quartiles, then the histogram is approximately symmetric. The box plot described subsequently is particularly useful in this regard.

## Interquartile Range

The quartiles can be used to create another measure of variability, the **interquartile range**, which is defined as follows.

**Interquartile Range**

$$\text{Interquartile range} = Q_3 - Q_1$$

The interquartile range measures the spread of the middle 50% of the observations. Large values of this statistic mean that the first and third quartiles are far apart, indicating a high level of variability.

## EXAMPLE 4.13 Interquartile Range of Long-Distance Telephone Bills

DATA
Xm03-01

Determine the interquartile range for Example 3.1.

### SOLUTION

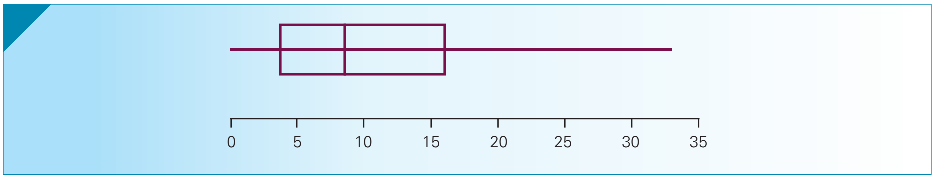
Using Excel's approximations of the first and third quartiles, we find

$$\text{Interquartile range} = Q_3 - Q_1 = 85 - 9.22 = 75.78$$

## Box Plots

Now that we have introduced quartiles we can present one more graphical technique, the **box plot**. This technique graphs five statistics: the minimum and maximum observations, and the first, second, and third quartiles. It also depicts other features of a set of data. Figure 4.1 exhibits the box plot of the data in Example 4.1.

FIGURE **4.1** Box Plot for Example 4.1



The three vertical lines of the box are the first, second, and third quartiles. The lines extending to the left and right are called *whiskers*. Any points that lie outside the whiskers are called *outliers*. The whiskers extend outward to the smaller of 1.5 times the interquartile range or to the most extreme point that is not an outlier.

**Outliers** **Outliers** are unusually large or small observations. Because an outlier is considerably removed from the main body of the data set, its validity is suspect. Consequently, outliers should be checked to determine that they are not the result of an error in recording their values. Outliers can also represent unusual observations that should be investigated. For example, if a salesperson's performance is an outlier on the high end of the distribution, the company could profit by determining what sets that salesperson apart from the others.

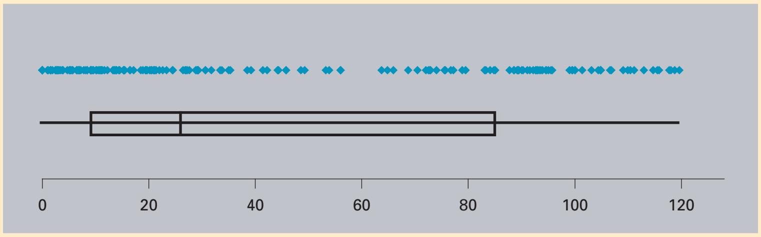
## EXAMPLE 4.14 Box Plot of Long-Distance Telephone Bills

DATA
Xm03-01

Draw the box plot for Example 3.1.

SOLUTION

EXCEL



INSTRUCTIONS

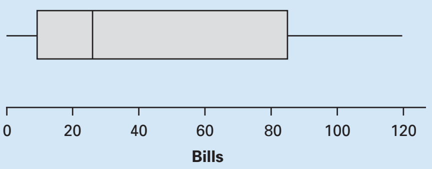
1. Type or import the data into one column or two or more adjacent columns. (Open Xm03-01.)
2. Click **Add-Ins, Data Analysis Plus**, and **Box Plot**.
3. Specify the **Input Range** (A1:A201).

A box plot will be created for each column of data that you have specified or highlighted.

Notice that the quartiles produced in the **Box Plot** are not exactly the same as those produced by **Descriptive Statistics**. The **Box Plot** command uses a slightly different method than the **Descriptive Methods** command.

## MINITAB

Box Plot of Bills



### INSTRUCTIONS

1. Type or import the data into one column or more columns. (Open Xm03-01.)
2. Click **Graph** and **Box Plot . . . .**
3. Click **Simple** if there is only one column of data or **Multiple Y's** if there are two or more columns.
4. Type or **Select** the variable or variables in the **Graph variables** box (Bills).
5. The box plot will be drawn so that the values will appear on the vertical axis. To turn the box plot on its side click **Scale . . .** , **Axes and Ticks**, and **Transpose value and category scales**.

## INTERPRET

The smallest value is 0, and the largest is 119.63. The first, second, and third quartiles are 9.275, 26.905, and 84.9425, respectively. The interquartile range is 75.6675. One and one-half times the interquartile range is 1.5 × 75.6675 = 113.5013. Outliers are defined as any observations that are less than 9.275 − 113.5013 = −104.226 and any observations that are larger than 84.9425 + 113.5013 = 198.4438. The whisker to the left extends only to 0, which is the smallest observation that is not an outlier. The whisker to the right extends to 119.63, which is the largest observation that is not an outlier. There are no outliers.

The box plot is particularly useful when comparing two or more data sets.

## EXAMPLE 4.15 Comparing Service Times of Fast-Food Restaurants' Drive-Throughs

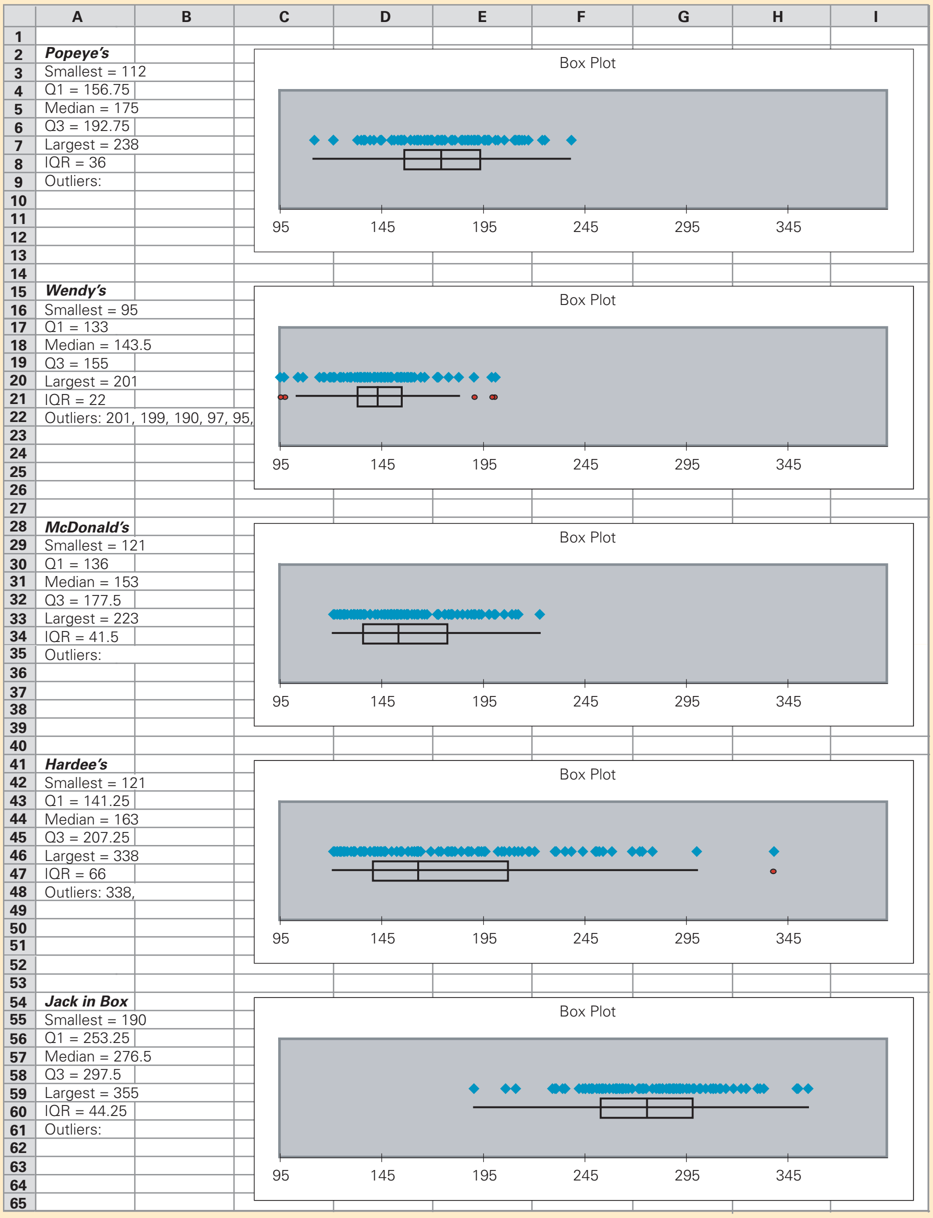
DATA Xm04-15

A large number of fast-food restaurants with drive-through windows offer drivers and their passengers the advantages of quick service. To measure how good the service is, an organization called QSR planned a study in which the amount of time taken by a sample of drive-through customers at each of five restaurants was recorded. Compare the five sets of data using a box plot and interpret the results.

### SOLUTION

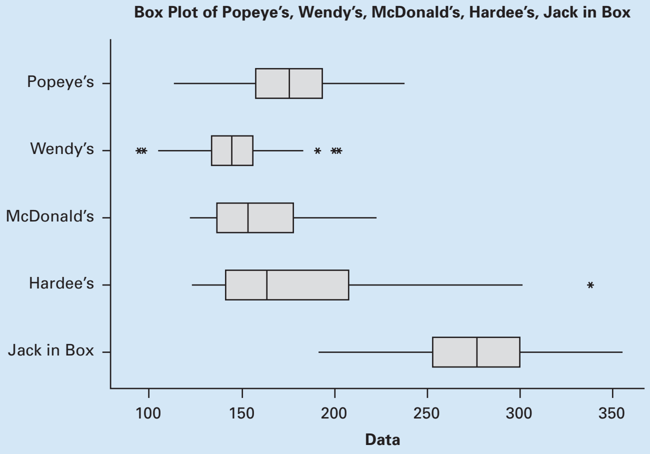
We use the computer and our software to produce the box plots.

## EXCEL



| | A |
|---|---|
| 1 | |
| 2 | ***Popeye's*** |
| 3 | Smallest = 112 |
| 4 | Q1 = 156.75 |
| 5 | Median = 175 |
| 6 | Q3 = 192.75 |
| 7 | Largest = 238 |
| 8 | IQR = 36 |
| 9 | Outliers: |
| 10 | |
| 11 | |
| 12 | |
| 13 | |
| 14 | |
| 15 | ***Wendy's*** |
| 16 | Smallest = 95 |
| 17 | Q1 = 133 |
| 18 | Median = 143.5 |
| 19 | Q3 = 155 |
| 20 | Largest = 201 |
| 21 | IQR = 22 |
| 22 | Outliers: 201, 199, 190, 97, 95, |
| 23 | |
| 24 | |
| 25 | |
| 26 | |
| 27 | |
| 28 | ***McDonald's*** |
| 29 | Smallest = 121 |
| 30 | Q1 = 136 |
| 31 | Median = 153 |
| 32 | Q3 = 177.5 |
| 33 | Largest = 223 |
| 34 | IQR = 41.5 |
| 35 | Outliers: |
| 36 | |
| 37 | |
| 38 | |
| 39 | |
| 40 | |
| 41 | ***Hardee's*** |
| 42 | Smallest = 121 |
| 43 | Q1 = 141.25 |
| 44 | Median = 163 |
| 45 | Q3 = 207.25 |
| 46 | Largest = 338 |
| 47 | IQR = 66 |
| 48 | Outliers: 338, |
| 49 | |
| 50 | |
| 51 | |
| 52 | |
| 53 | |
| 54 | ***Jack in Box*** |
| 55 | Smallest = 190 |
| 56 | Q1 = 253.25 |
| 57 | Median = 276.5 |
| 58 | Q3 = 297.5 |
| 59 | Largest = 355 |
| 60 | IQR = 44.25 |
| 61 | Outliers: |
| 62 | |
| 63 | |
| 64 | |
| 65 | |

MINITAB



INTERPRET

Wendy's times appear to be the lowest and most consistent. The service times for Hardee's display considerably more variability. The slowest service times are provided by Jack in the Box. The service times for Popeye's, Wendy's, and Jack in the Box seem to be symmetric. However, the times for McDonald's and Hardee's are positively skewed.

## Measures of Relative Standing and Variability for Ordinal Data

Because the measures of relative standing are computed by ordering the data, these statistics are appropriate for ordinal as well as for interval data. Furthermore, because the interquartile range is calculated by taking the difference between the upper and lower quartiles, it too can be employed to measure the variability of ordinal data.

Here are the factors that tell us when to use the techniques presented in this section.

### Factors That Identify When to Compute Percentiles and Quartiles

1. **Objective**: Describe a single set of data
2. **Type of data**: Interval or ordinal
3. **Descriptive measurement**: Relative standing

**Factors That Identify When to Compute the Interquartile Range**

1. **Objective**: Describe a single set of data
2. **Type of data**: Interval or ordinal
3. **Descriptive measurement**: Variability

## EXERCISES

**4.40** Calculate the first, second, and third quartiles of the following sample.

5 8 2 9 5 3 7 4 2 7 4 10 4 3 5

**4.41** Find the third and eighth deciles (30th and 80th percentiles) of the following data set.

| | | | | |
|---|---|---|---|---|
| 26 | 23 | 29 | 31 | 24 |
| 22 | 15 | 31 | 30 | 20 |

**4.42** Find the first and second quintiles (20th and 40th percentiles) of the data shown here.

| | | | | |
|---|---|---|---|---|
| 52 | 61 | 88 | 43 | 64 |
| 71 | 39 | 73 | 51 | 60 |

**4.43** Determine the first, second, and third quartiles of the following data.

| | | | | | | |
|---|---|---|---|---|---|---|
| 10.5 | 14.7 | 15.3 | 17.7 | 15.9 | 12.2 | 10.0 |
| 14.1 | 13.9 | 18.5 | 13.9 | 15.1 | 14.7 | |

**4.44** Calculate the 3rd and 6th deciles of the accompanying data.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 7 | 18 | 12 | 17 | 29 | 18 | 4 | 27 | 30 | 2 |
| 4 | 10 | 21 | 5 | 8 | | | | | |

**4.45** Refer to Exercise 4.43. Determine the interquartile range.

**4.46** Refer to Exercise 4.40. Determine the interquartile range.

**4.47** Compute the interquartile range from the following data.

5 8 14 6 21 11 9 10 18 2

**4.48** Draw the box plot of the following set of data.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 9 | 28 | 15 | 21 | 12 | 22 | 29 | |
| 20 | 23 | 31 | 11 | 19 | 24 | 16 | 13 |

*The following exercises require a computer and software.*

**4.49** Xr04-49 Many automotive experts believe that speed limits on highways are too low. One particular expert has stated that he thinks that most drivers drive at speeds that they consider safe. He suggested that the "correct" speed limit should be set at the 85th percentile. Suppose that a random sample of 400 speeds on a highway where the limit is 60 mph was recorded. Find the "correct" speed limit.

**4.50** Xr04-50 Accountemps, a company that supplies temporary workers, sponsored a survey of 100 executives. Each was asked to report the number of minutes they spend screening each job resume they receive.

a. Compute the quartiles.

b. What information did you derive from the quartiles? What does this suggest about writing your resume?

**4.51** Xr04-51 How much do pets cost? A random sample of dog and cat owners was asked to compute the amounts of money spent on their pets (exclusive of pet food). Draw a box plot for each data set and describe your findings.

**4.52** Xr04-52 The Travel Industry Association of America sponsored a poll that asked a random sample of people how much they spent in preparation for pleasure travel. Determine the quartiles and describe what they tell you.

**4.53** Xr04-53 The career-counseling center at a university wanted to learn more about the starting salaries of the university's graduates. They asked each graduate to report the highest salary offer received. The survey also asked each graduate to report the degree and starting salary (column 1 = BA, column 2 = BSc, column 3 = BBA, column 4 = other). Draw box plots to compare the four groups of starting salaries. Report your findings.

**4.54** Xr04-54 A random sample of Boston Marathon runners was drawn and the times to complete the race were recorded.

a. Draw the box plot.

b. What are the quartiles?

c. Identify outliers.
d. What information does the box plot deliver?

**4.55** Xr04-55 Do golfers who are members of private courses play faster than players on a public course? The amount of time taken for a sample of private-course and public-course golfers was recorded.
a. Draw box plots for each sample.
b. What do the box plots tell you?

**4.56** Xr04-56 For many restaurants, the amount of time customers linger over coffee and dessert negatively affect profits. To learn more about this variable, a sample of 200 restaurant groups was observed, and the amount of time customers spent in the restaurant was recorded.
a. Calculate the quartiles of these data.
b. What do these statistics tell you about the amount of time spent in this restaurant?

**4.57** Xr04-57 In the United States, taxpayers are allowed to deduct mortgage interest from their incomes before calculating the amount of income tax they are required to pay. In 2005, the Internal Revenue Service sampled 500 tax returns that had a mortgage-interest deduction. Calculate the quartiles and describe what they tell you. (Adapted from *Statistical Abstract of the United States*, *2009*, Table 471.)

## American National Election Survey Exercises

**4.58 ANES2008* In the 2008 survey, people were asked to indicate the amount of time they spent in a typical day receiving news about the election on the Internet (TIME1) and on television (TIME2). Compare the two amounts of time by drawing box plots (using the same scale) and describe what the graphs tell you. (*Excel users:* You must have adjacent columns. We recommend that you copy the two columns into adjacent columns in a separate spreadsheet.)**

**4.59 ANES2008* Draw a box plot of the ages (AGE) of respondents in the 2008 survey.**

**4.60 ANES2008* Draw a box plot of the education level of both married spouses (EDUC and SPEDUC). Describe your findings.**

## General Social Survey Exercises

**4.61 GSS2008* Draw a box plot of the ages (AGE) of respondents from the 2008 survey. Briefly describe the graph.**

**4.62 GSS2008* Produce a box plot of the amount of television watched (TVHOURS). State what the graph tells you.**

# 4.4 Measures of Linear Relationship

In Chapter 3, we introduced the scatter diagram, a graphical technique that describes the relationship between two interval variables. At that time, we pointed out that we were particularly interested in the direction and strength of the linear relationship. We now present three numerical measures of linear relationship that provide this information: *covariance*, *coefficient of correlation*, and *coefficient of determination*. Later in this section we discuss another related numerical technique, the *least squares line*.

## Covariance

As we did in Chapter 3, we label one variable $X$ and the other $Y$.

### Covariance

$$\text{Population covariance: } \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

$$\text{Sample covariance: } s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

The denominator in the calculation of the sample **covariance** is $n - 1$, not the more logical $n$ for the same reason we divide by $n - 1$ to calculate the sample variance (see page 109). If you plan to compute the sample covariance manually, here is a shortcut calculation.

### Shortcut for Sample Covariance

$$s_{xy} = \frac{1}{n - 1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right]$$

To illustrate how covariance measures the linear relationship, examine the following three sets of data.

**Set 1**

| $x_i$ | $y_i$ | $(x_i - \bar{x})$ | $(y_i - \bar{y})$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|---|---|---|---|---|
| 2 | 13 | –3 | –7 | 21 |
| 6 | 20 | 1 | 0 | 0 |
| 7 | 27 | 2 | 7 | 14 |
| | $\bar{x} = 5$ | $\bar{y} = 20$ | | $s_{xy} = 35/2 = 17.5$ |

**Set 2**

| $x_i$ | $y_i$ | $(x_i - \bar{x})$ | $(y_i - \bar{y})$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|---|---|---|---|---|
| 2 | 27 | –3 | 7 | –21 |
| 6 | 20 | 1 | 0 | 0 |
| 7 | 13 | 2 | –7 | –14 |
| | $\bar{x} = 5$ | $\bar{y} = 20$ | | $s_{xy} = -35/2 = -17.5$ |

**Set 3**

| $x_i$ | $y_i$ | $(x_i - \bar{x})$ | $(y_i - \bar{y})$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|---|---|---|---|---|
| 2 | 20 | −3 | 0 | 0 |
| 6 | 27 | 1 | 7 | 7 |
| 7 | 13 | 2 | −7 | −14 |
| | $\bar{x} = 5$ | $\bar{y} = 20$ | | $s_{xy} = -7/2 = -3.5$ |

Notice that the values of $x$ are the same in all three sets and that the values of $y$ are also the same. The only difference is the *order* of the values of $y$.

In set 1, as $x$ increases so does $y$. When $x$ is larger than its mean, $y$ is at least as large as its mean. Thus $(x_i - \bar{x})$ and $(y_i - \bar{y})$ have the same sign or 0. Their product is also positive or 0. Consequently, the covariance is a positive number. Generally, when two variables move in the same direction (both increase or both decrease), the covariance will be a large positive number.

If you examine set 2, you will discover that as $x$ increases, y decreases. When $x$ is larger than its mean, $y$ is less than or equal to its mean. As a result when $(x_i - \bar{x})$ is positive, $(y_i - \bar{y})$ is negative or 0. Their products are either negative or 0. It follows that the covariance is a negative number. In general, when two variables move in opposite directions, the covariance is a large negative number.

In set 3, as $x$ increases, $y$ does not exhibit any particular direction. One of the products $(x_i - \bar{x})(y_i - \bar{y})$ is 0, one is positive, and one is negative. The resulting covariance is a small number. In general, when there is no particular pattern, the covariance is a small number.

We would like to extract two pieces of information. The first is the sign of the covariance, which tells us the nature of the relationship. The second is the magnitude, which describes the strength of the association. Unfortunately, the magnitude may be difficult to judge. For example, if you're told that the covariance between two variables is 500, does this mean that there is a strong linear relationship? The answer is that it is impossible to judge without additional statistics. Fortunately, we can improve on the information provided by this statistic by creating another one.

## Coefficient of Correlation

The **coefficient of correlation** is defined as the covariance divided by the standard deviations of the variables.

### Coefficient of Correlation

$$\text{Population coefficient of correlation: } \rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$\text{Sample coefficient of correlation: } r = \frac{s_{xy}}{s_x s_y}$$

The population parameter is denoted by the Greek letter *rho*.

The advantage that the coefficient of correlation has over the covariance is that the former has a set lower and upper limit. The limits are −1 and +1, respectively—that is,

$$-1 \le r \le +1 \qquad \text{and} \qquad -1 \le \rho \le +1$$

When the coefficient of correlation equals –1, there is a negative linear relationship and the scatter diagram exhibits a straight line. When the coefficient of correlation equals +1, there is a perfect positive relationship. When the coefficient of correlation equals 0, there is no linear relationship. All other values of correlation are judged in relation to these three values. The drawback to the coefficient of correlation is that—except for the three values −1, 0, and +1—we cannot interpret the correlation. For example, suppose that we calculated the coefficient of correlation to be −.4. What does this tell us? It tells us two things. The minus sign tells us the relationship is negative and because .4 is closer to 0 than to 1, we judge that the linear relationship is weak. In many applications, we need a better interpretation than the "linear relationship is weak." Fortunately, there is yet another measure of the strength of a linear relationship, which gives us more information. It is the *coefficient of determination*, which we introduce later in this section.

## EXAMPLE 4.16 Calculating the Coefficient of Correlation

Calculate the coefficient of correlation for the three sets of data on pages 126–127.

### SOLUTION

Because we've already calculated the covariances we need to compute only the standard deviations of $X$ and $Y$.

$$\bar{x} = \frac{2 + 6 + 7}{3} = 5.0$$

$$\bar{y} = \frac{13 + 20 + 27}{3} = 20.0$$

$$s_x^2 = \frac{(2 - 5)^2 + (6 - 5)^2 + (7 - 5)^2}{3 - 1} = \frac{9 + 1 + 4}{2} = 7.0$$

$$s_y^2 = \frac{(13 - 20)^2 + (20 - 20)^2 + (27 - 20)^2}{3 - 1} = \frac{49 + 0 + 49}{2} = 49.0$$

The standard deviations are

$$s_x = \sqrt{7.0} = 2.65$$

$$s_y = \sqrt{49.0} = 7.00$$

The coefficients of correlation are:

$$\textbf{Set 1:} \quad r = \frac{s_{xy}}{s_x s_y} = \frac{17.5}{(2.65)(7.0)} = .943$$

$$\textbf{Set 2:} \quad r = \frac{s_{xy}}{s_x s_y} = \frac{-17.5}{(2.65)(7.0)} = -.943$$

$$\textbf{Set 3:} \quad r = \frac{s_{xy}}{s_x s_y} = \frac{-3.5}{(2.65)(7.0)} = -.189$$

It is now easier to see the strength of the linear relationship between $X$ and $Y$.

## Comparing the Scatter Diagram, Covariance, and Coefficient of Correlation

The scatter diagram depicts relationships graphically; the covariance and the coefficient of correlation describe the linear relationship numerically. Figures 4.2, 4.3, and 4.4 depict three scatter diagrams. To show how the graphical and numerical techniques compare, we calculated the covariance and the coefficient of correlation for each. (The data are stored in files Fig04-02, Fig04-03, and Fig04-04.) As you can see, Figure 4.2 depicts a strong positive relationship between the two variables. The covariance is 36.87, and the coefficient of correlation is .9641. The variables in Figure 4.3 produced a relatively strong negative linear relationship; the covariance and coefficient of correlation are −34.18 and −.8791, respectively. The covariance and coefficient of correlation for the data in Figure 4.4 are 2.07 and .1206, respectively. There is no apparent linear relationship in this figure.

FIGURE **4.2** Strong Positive Linear Relationship

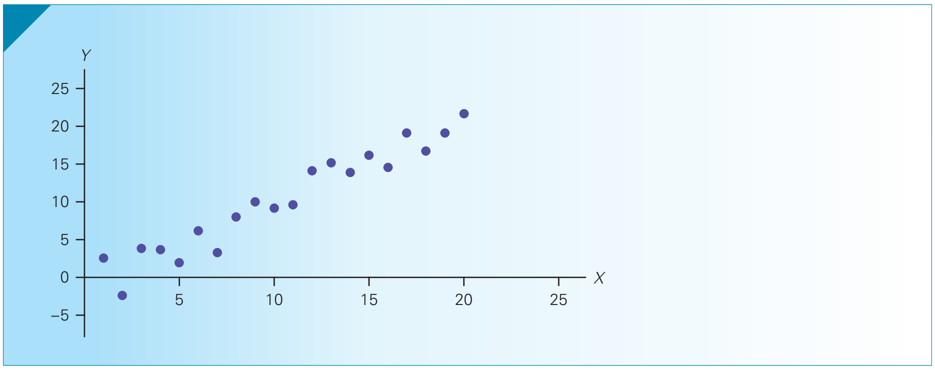


FIGURE **4.3** Strong Negative Linear Relationship

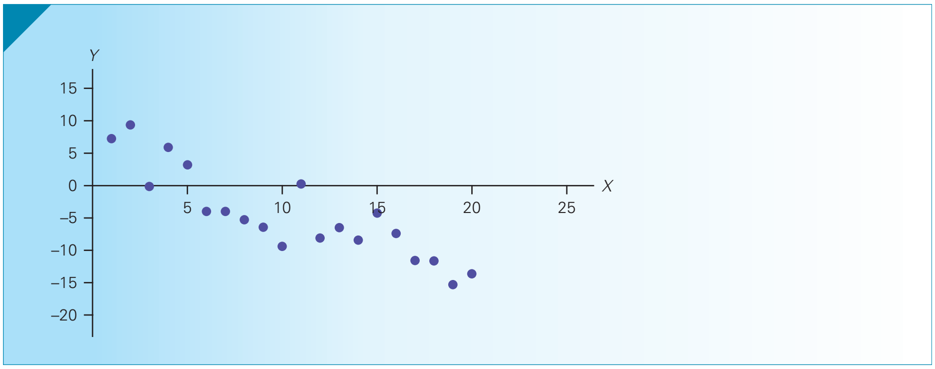
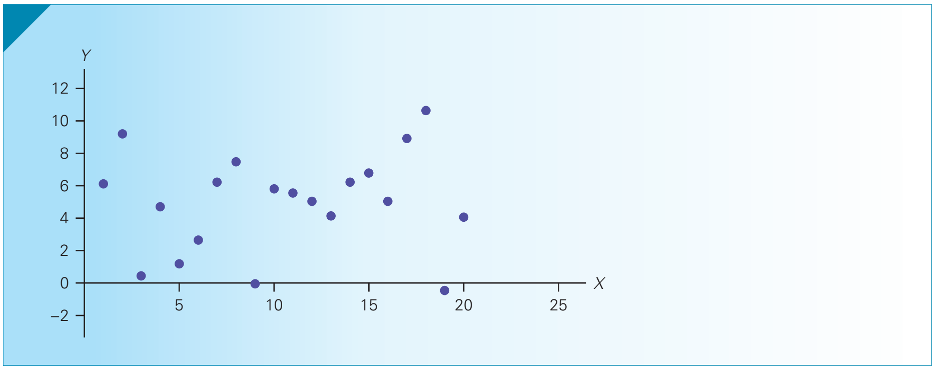


FIGURE **4.4** No Linear Relationship



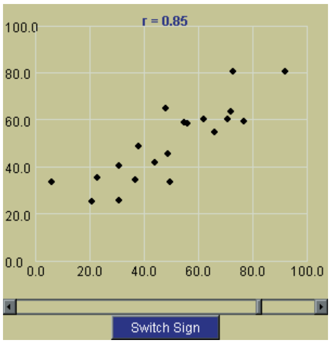
SEEING STATISTICS

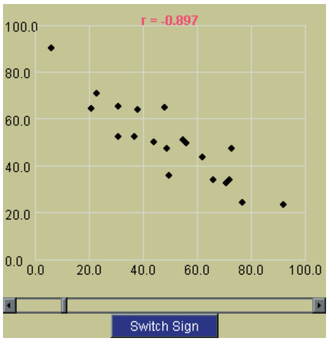
## applet 1 Scatter Diagrams and Correlation

In Section 1.3, we introduced applets as a method to show students of applied statistics how statistical techniques work and gain insights into the underlying principles. The applets are stored on Keller's website that accompanies this book. See the README file for instructions on how to use them.

### Instructions for Applet 1

Use your mouse to move the slider in the graph. As you move the slider, observe how the coefficient of correlation changes as the points become more "organized" in the scatter diagram. If you click **Switch sign**, you can see the difference between positive and negative coefficients. The following figures displays the applet for two values of *r*.





### Applet Exercises

1.1 Drag the slider to the right until the correlation coefficient *r* is 1.0. Describe the pattern of the data points.

1.2 Drag the slider to the left until the correlation coefficient *r* is −1.0. Describe the pattern of the data points. In what way does it differ from the case where *r* = 1.0?

1.3 Drag the slider toward the center until the correlation coefficient *r* is 0 (approximately). Describe the pattern of the data points. Is there a pattern? Or do the points appear to be scattered randomly?

1.4 Drag the slider until the correlation coefficient *r* is .5 (approximately). Can you detect a pattern? Now click on the **Switch Sign** button to change the correlation coefficient *r* to −.5. How does the pattern change when the sign switches? Switch back and forth several times so you can see the changes.

SEEING STATISTICS

## applet 2 Scatter Patterns and Correlation

This applet allows you to place points on a graph and see the resulting value of the coefficient of correlation.

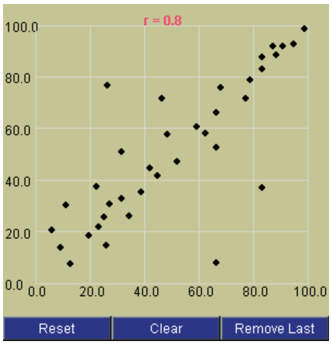
### Instructions

Click on the graph to place a point. As you add points, the correlation coefficient is recalculated. Click to add points in various patterns to see how the correlation does (or does not) reflect those patterns. Click on the **Reset** button to clear all points. The figure shown here depicts a scatter diagram and its coefficient of correlation.

### Applet Exercises

2.1 Create a scatter diagram where *r* is approximately 0. Describe how you did it.

2.2 Create a scatter diagram where *r* is approximately 1. Describe how this was done.



2.3 Plot points such that $r$ is approximately .5. How would you describe the resulting scatter diagram?

2.4 Plot the points on a scatter diagram where $r$ is approximately 1. Now add one more point, decreasing $r$ by as much as possible. What does this tell you about extreme points?

2.5 Repeat Applet Exercise 2.4, adding two points. How close to $r = 0$ did you get?

## Least Squares Method

When we presented the scatter diagram in Section 3.3, we pointed out that we were interested in measuring the strength and direction of the linear relationship. Both can be more easily judged by drawing a straight line through the data. However, if different people draw a line through the same data set, it is likely that each person's line will differ from all the others. Moreover, we often need to know the equation of the line. Consequently, we need an objective method of producing a straight line. Such a method has been developed; it is called the **least squares method**.

The least squares method produces a straight line drawn through the points so that the sum of squared deviations between the points and the line is minimized. The line is represented by the equation:

$$\hat{y} = b_0 + b_1 x$$

where $b_0$ is the $y$-intercept (where the line intercepts the $y$-axis), and $b_1$ is the slope (defined as rise/run), and $\hat{y}$ ($y$ *hat*) is the value of $y$ determined by the line. The coefficients $b_0$ and $b_1$ are derived using calculus so that we minimize the sum of squared deviations:

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

**Least Squares Line Coefficients**

$$b_1 = \frac{s_{xy}}{s_x^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

APPLICATIONS in ACCOUNTING

### Breakeven Analysis

**Breakeven analysis** is an extremely important business tool, one that you will likely encounter repeatedly in your course of studies. It can used to determine how much sales volume your business needs to start making a profit.

In Section 3.1 (page 44) we briefly introduced the four P's of marketing and illustrated the problem of pricing with Example 3.1. Breakeven analysis is especially useful when managers are attempting to determine the appropriate price for the company's products and services.

A company's profit can be calculated simply as

Profit = (Price per unit – variable cost per unit) × (Number of units sold) – Fixed costs

The breakeven point is the number of units sold such that the profit is 0. Thus, the breakeven point is calculated as

Number of units sold = Fixed cost / (Price – Variable cost)

Managers can use the formula to help determine the price that will produce a profit. However, to do so requires knowledge of the fixed and variable costs. For example, suppose that a bakery sells only loaves of bread. The bread sells for $1.20, the variable cost is $0.40, and the fixed annual costs are $10,000. The breakeven point is

Number of units sold = 10,000/ (1.20 – 0.40) = 12,500

The bakery must sell more than 12,500 loaves per year to make a profit.

In the next application box, we discuss fixed and variable costs.

## APPLICATIONS in ACCOUNTING

© Steve Allen/Brand X Pictures/ Jupiter images

### Fixed and Variable Costs

Fixed costs are costs that must be paid whether or not any units are produced. These costs are "fixed" over a specified period of time or range of production. Variable costs are costs that vary directly with the number of products produced. For the previous bakery example, the fixed costs would include rent and maintenance of the shop, wages paid to employees, advertising costs, telephone, and any other costs that are not related to the number of loaves baked. The variable cost is primarily the cost of ingredients, which rises in relation to the number of loaves baked.

Some expenses are mixed. For the bakery example, one such cost is the cost of electricity. Electricity is needed for lights, which is considered a fixed cost, but also for the ovens and other equipment, which are variable costs.

There are several ways to break the mixed costs into fixed and variable components. One such method is the least squares line; that is, we express the total costs of some component as

$$y = b_0 + b_1x$$

where $y$ = total mixed cost, $b_0$ = fixed cost, $b_1$ = variable cost, and $x$ is the number of units.

## EXAMPLE 4.17 Estimating Fixed and Variable Costs

DATA
Xm04-17

A tool and die maker operates out of a small shop making specialized tools. He is considering increasing the size of his business and needs to know more about his costs. One such cost is electricity, which he needs to operate his machines and lights. (Some jobs require that he turn on extra bright lights to illuminate his work.) He keeps track of his daily electricity costs and the number of tools that he made that day. These data are listed next. Determine the fixed and variable electricity costs.

| Day | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Number of tools | 7 | 3 | 2 | 5 | 8 | 11 | 5 | 15 | 3 | 6 |
| Electricity cost | 23.80 | 11.89 | 15.98 | 26.11 | 31.79 | 39.93 | 12.27 | 40.06 | 21.38 | 18.65 |

### SOLUTION

The dependent variable is the daily cost of electricity, and the independent variable is the number of tools. To calculate the coefficients of the least squares line and other statistics (calculated below), we need the sum of $X$, $Y$, $XY$, $X^2$, and $Y^2$.

| Day | $X$ | $Y$ | $XY$ | $X^2$ | $Y^2$ |
|---|---|---|---|---|---|
| 1 | 7 | 23.80 | 166.60 | 49 | 566.44 |
| 2 | 3 | 11.89 | 35.67 | 9 | 141.37 |
| 3 | 2 | 15.98 | 31.96 | 4 | 255.36 |
| 4 | 5 | 26.11 | 130.55 | 25 | 681.73 |
| 5 | 8 | 31.79 | 254.32 | 64 | 1010.60 |
| 6 | 11 | 39.93 | 439.23 | 121 | 1594.40 |
| 7 | 5 | 12.27 | 61.35 | 25 | 150.55 |
| 8 | 15 | 40.06 | 600.90 | 225 | 1604.80 |
| 9 | 3 | 21.38 | 64.14 | 9 | 457.10 |
| 10 | 6 | 18.65 | 111.90 | 36 | 347.82 |
| Total | 65 | 241.86 | 1896.62 | 567 | 6810.20 |

Covariance:

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right] = \frac{1}{10-1}\left[1896.62 - \frac{(65)(241.86)}{10}\right] = 36.06$$

Variance of $X$:

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{10-1}\left[567 - \frac{(65)^2}{10}\right] = 16.06$$

Sample means

$$\bar{x} = \frac{\sum x_I}{n} = \frac{65}{10} = 6.5$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{241.86}{10} = 24.19$$

The coefficients of the least squares line are

Slope

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{36.06}{16.06} = 2.25$$

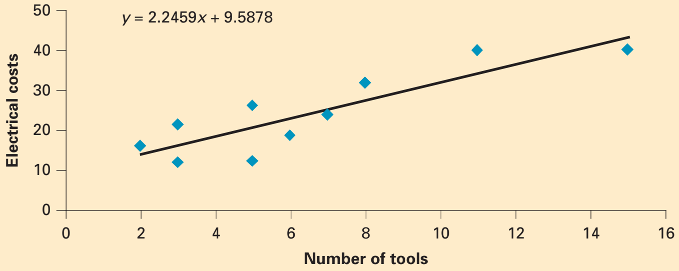
$y$-intercept:

$$b_0 = \bar{y} - b_1\bar{x} = 24.19 - (2.25)(6.5) = 9.57$$

The least squares line is

$$\hat{y} = 9.57 + 2.25x$$

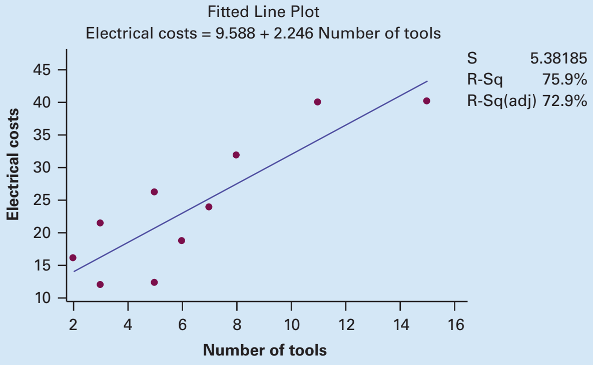
## EXCEL



*INSTRUCTIONS*

1. Type or import the data into two columns where the first column stores the values of $X$ and the second stores $Y$. (Open Xm04-17.) Highlight the columns containing the variables. Follow the instructions to draw a scatter diagram (page 75).
2. In the **Chart Tools** and **Layout** menu, click **Trendline** and **Linear Trendline.**
3. Click **Trendline** and **More Trendline Options . . . .** Click **Display Equation on Chart.**

## MINITAB



### INSTRUCTIONS

1. Type or import the data into two columns. (Open Xm04-17.)
2. Click **Stat, Regression**, and **Fitted Line Plot**.
3. Specify the **Response [Y]** (Electrical cost) and the **Predictor [X]** (Number of tools) variables. Specify **Linear.**

## INTERPRET

The slope is defined as rise/run, which means that it is the change in $y$ (rise) for a one-unit increase in $x$ (run). Put less mathematically, the slope measures the *marginal* rate of change in the dependent variable. The marginal rate of change refers to the effect of increasing the independent variable by one additional unit. In this example, the slope is 2.25, which means that in this sample, for each one-unit increase in the number of tools, the marginal increase in the electricity cost is \$2.25. Thus, the estimated variable cost is \$2.25 per tool.

The $y$-intercept is 9.57; that is, the line strikes the $y$-axis at 9.57. This is simply the value of $\hat{y}$ when $x = 0$. However, when $x = 0$, we are producing no tools and hence the estimated fixed cost of electricity is \$9.57 per day.

Because the costs are estimates based on a straight line, we often need to know how well the line fits the data.

## EXAMPLE 4.18 Measuring the Strength of the Linear Relationship

DATA
Xm04-17

Calculate the coefficient of correlation for Example 4.17.

### SOLUTION

To calculate the coefficient of correlation, we need the covariance and the standard deviations of both variables. The covariance and the variance of $X$ were calculated in Example 4.17. The covariance is

$$s_{xy} = 36.06$$

and the variance of $X$ is

$$s_x^2 = 16.06$$

Standard deviation of $X$ is

$$s_x = \sqrt{s_x^2} = \sqrt{16.06} = 4.01$$

All we need is the standard deviation of $Y$.

$$s_y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] = \frac{1}{10-1}\left[6810.20 - \frac{(241.86)^2}{10}\right] = 106.73$$

$$s_y = \sqrt{s_y^2} = \sqrt{106.73} = 10.33$$

The coefficient of correlation is

$$r = \frac{s_{XY}}{s_x s_y} = \frac{36.06}{(4.01)(10.33)} = .8705$$

## EXCEL

As with the other statistics introduced in this chapter, there is more than one way to calculate the coefficient of correlation and the covariance. Here are the instructions for both.

### INSTRUCTIONS

1. Type or import the data into two columns. (Open Xm04-17.) Type the following into any empty cell.

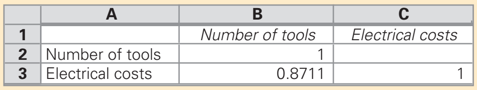
= **CORREL**([Input range of one variable], [Input range of second variable])

In this example, we would enter

= **CORREL**(B1:B11, C1:C11)

To calculate the covariance, replace **CORREL** with **COVAR.**

Another method, which is also useful if you have more than two variables and would like to compute the coefficient of correlation or the covariance for each pair of variables, is to produce the correlation matrix and the variance–covariance matrix. We do the correlation matrix first.



| | A | B | C |
|---|---|---|---|
| 1 | | *Number of tools* | *Electrical costs* |
| 2 | Number of tools | 1 | |
| 3 | Electrical costs | 0.8711 | 1 |

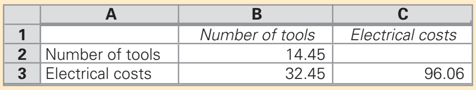
### INSTRUCTIONS

1. Type or import the data into adjacent columns. (Open Xm04-17.)
2. Click **Data, Data Analysis**, and **Correlation**.
3. Specify the **Input Range (B1:C11).**

The coefficient of correlation between number of tools and electrical costs is .8711 (slightly different from the manually calculated value). (The two 1s on the diagonal of the matrix are the coefficients of number of tools and number of tools, and electrical costs and electrical costs, telling you the obvious.)

Incidentally, the formula for the population parameter $\rho$ (Greek letter *rho*) and for the sample statistic $r$ produce exactly the same value.

The variance–covariance matrix is shown next.

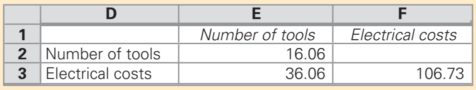


| | A | B | C |
|---|---|---|---|
| 1 | | *Number of tools* | *Electrical costs* |
| 2 | Number of tools | 14.45 | |
| 3 | Electrical costs | 32.45 | 96.06 |

### INSTRUCTIONS

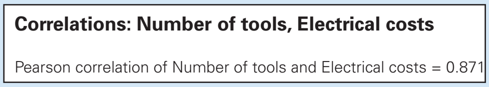
1. Type or import the data into adjacent columns. (Open Xm04-17.)
2. Click **Data, Data Analysis**, and **Covariance**.
3. Specify the **Input Range (B1:C11).**

Unfortunately, Excel computes the population parameters. In other words, the variance of the number of tools is $\sigma_x^2 = 14.45$, the variance of the electrical costs is $\sigma_y^2 = 96.06$, and the covariance is $\sigma_{xy} = 32.45$. You can convert these parameters to statistics by multiplying each by $n/(n-1)$.



| | D | E | F |
|---|---|---|---|
| 1 | | *Number of tools* | *Electrical costs* |
| 2 | Number of tools | 16.06 | |
| 3 | Electrical costs | 36.06 | 106.73 |

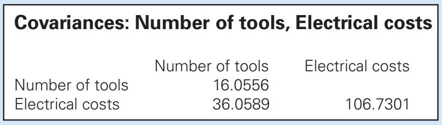
## MINITAB



```
Correlations: Number of tools, Electrical costs

Pearson correlation of Number of tools and Electrical costs = 0.871
```

### INSTRUCTIONS

1. Type or import the data into two columns. (Open Xm04-17.)
2. Click **Calc, Basic Statistics** and **Correlation . . . .**
3. In the **Variables** box, type **Select** the variables (**Number of Tools, Electrical Costs**).



```
Covariances: Number of tools, Electrical costs

                   Number of tools   Electrical costs
Number of tools            16.0556
Electrical costs           36.0589           106.7301
```

### INSTRUCTIONS

Click **Covariance . . .** instead of **Correlation . . .** in step 2 above.

## INTERPRET

The coefficient of correlation is .8711, which tells us that there is a positive linear relationship between the number of tools and the electricity cost. The coefficient of correlation tells us that the linear relationship is quite strong and thus the estimates of the fixed and variable costs should be good.

## Coefficient of Determination

When we introduced the coefficient of correlation (page 128), we pointed out that except for $-1$, 0, and $+1$ we cannot precisely interpret its meaning. We can judge the coefficient of correlation in relation to its proximity to only $-1$, 0, and $+1$. Fortunately, we have another measure that can be precisely interpreted. It is the coefficient of determination, which is calculated by squaring the coefficient of correlation. For this reason, we denote it $R^2$ .

The coefficient of determination measures the amount of variation in the dependent variable that is explained by the variation in the independent variable. For example, if the coefficient of correlation is $-1$ or $+1$, a scatter diagram would display all the points lining up in a straight line. The coefficient of determination is 1, which we interpret to mean that 100% of the variation in the dependent variable $Y$ is explained by the variation in the independent variable $X$. If the coefficient of correlation is 0, then there is no linear relationship between the two variables, $R^2 = 0$, and none of the variation in $Y$ is explained by the variation in $X$. In Example 4.18, the coefficient of correlation was calculated to be $r = .8711$. Thus, the coefficient of determination is

$$r^2 = (.8711)^2 = .7588$$

This tells us that 75.88% of the variation in electrical costs is explained by the number of tools. The remaining 24.12% is unexplained.

## Using the Computer

### EXCEL

You can use Excel to calculate the coefficient of correlation and then square the result. Alternatively, use Excel to draw the least squares line. After doing so, click **Trendline, Trendline Options**, and **Display R-squared value on chart.**

### MINITAB

Minitab automatically prints the coefficient of determination.

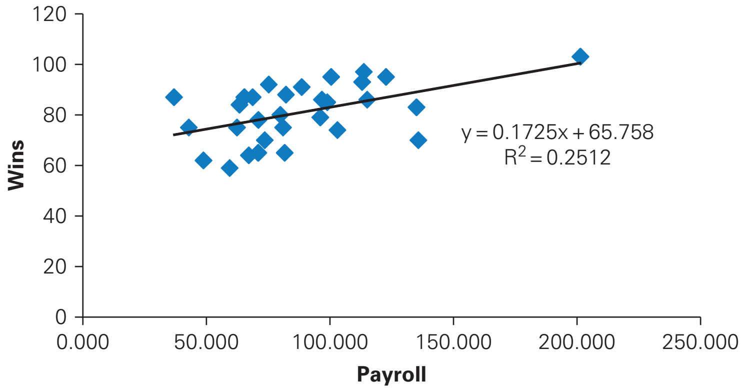
The concept of explained variation is an extremely important one in statistics. We return to this idea repeatedly in Chapters 13, 14, 16, 17, and 18. In Chapter 16, we explain why we interpret the coefficient of determination in the way that we do.

## Cost of One More Win: Solution

To determine the cost of an additional win, we must describe the relationship between two variables. To do so, we use the least squares method to produce a straight line through the data. Because we believe that the number of games a baseball team wins depends to some extent on its team payroll, we label Wins as the dependent variable and Payroll as the independent variable. Because of rounding problems, we expressed the payroll in the number of millions of dollars.

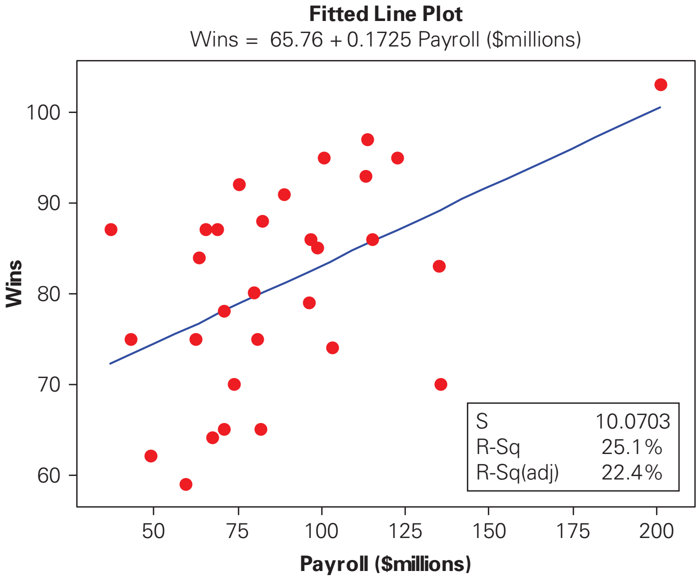
© AP Photo/Charles Krupa

### EXCEL



As you can see, Excel outputs the least squares line and the coefficient of determination.

### MINITAB



INTERPRET

The least squares line is

$$\hat{y} = 65.758 + .1725\, x$$

The slope is equal to .1725, which is the marginal rate of change in games won for each one-unit increase in payroll. Because payroll is measured in millions of dollars, we estimate that for each $1 million increase in the payroll, the number of games won increases on average by .1725. Thus, to win one more game requires on average an additional expenditure of an incredible $5,797,101 (calculated as 1 million/.1725).

Besides analyzing the least squares line, we should determine the strength of the linear relationship. The coefficient of determination is .2512, which means that the variation in the team's payroll explains 25.12% of the variation in the team's number of games won. This suggests that some teams win a small number of games with large payrolls, whereas others win a large number of games with small payrolls. In the next section, we will return to this issue and examine why some teams perform better than predicted by the least squares line.

## Interpreting Correlation

Because of its importance, we remind you about the correct interpretation of the analysis of the relationship between two interval variables that we discussed in Chapter 3. In other words, if two variables are linearly related, it does not mean that *X* causes *Y*. It may mean that another variable causes both *X* and *Y* or that *Y* causes *X*. Remember

Correlation is not Causation

We complete this section with a review of when to use the techniques introduced in this section.

**Factors That Identify When to Compute Covariance, Coefficient of Correlation, Coefficient of Determination, and Least Squares Line**

1. **Objective**: Describe the relationship between two variables
2. **Type of data**: Interval

## EXERCISES

**4.63** The covariance of two variables has been calculated to be –150. What does the statistic tell you about the relationship between the two variables?

**4.64** Refer to Exercise 4.63. You've now learned that the two sample standard deviations are 16 and 12.

a. Calculate the coefficient of correlation. What does this statistic tell you about the relationship between the two variables?

b. Calculate the coefficient of determination and describe what this says about the relationship between the two variables.

**4.65** Xr04-65 A retailer wanted to estimate the monthly fixed and variable selling expenses. As a first step, she collected data from the past 8 months. The total selling expenses ($1,000) and the total sales ($1,000) were recorded and listed below.

| Total Sales | Selling Expenses |
|---|---|
| 20 | 14 |
| 40 | 16 |
| 60 | 18 |
| 50 | 17 |
| 50 | 18 |
| 55 | 18 |
| 60 | 18 |
| 70 | 20 |

a. Compute the covariance, the coefficient of correlation, and the coefficient of determination and describe what these statistics tell you.

b. Determine the least squares line and use it to produce the estimates the retailer wants.

**4.66** Xr04-66 Are the marks one receives in a course related to the amount of time spent studying the subject? To analyze this mysterious possibility, a student took a random sample of 10 students who had enrolled in an accounting class last semester. He asked each to report his or her mark in the course and the total number of hours spent studying accounting. These data are listed here.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Marks** | 77 | 63 | 79 | 86 | 51 | 78 | 83 | 90 | 65 | 47 |
| **Time spent studying** | 40 | 42 | 37 | 47 | 25 | 44 | 41 | 48 | 35 | 28 |

a. Calculate the covariance.

b. Calculate the coefficient of correlation.

c. Calculate the coefficient of determination.

d. Determine the least squares line.

e. What do the statistics calculated above tell you about the relationship between marks and study time?

**4.67** Xr04-67 Students who apply to MBA programs must take the Graduate Management Admission Test (GMAT). University admissions committees use the GMAT score as one of the critical indicators of how well a student is likely to perform in the MBA program. However, the GMAT may not be a very strong indicator for all MBA programs. Suppose that an MBA program designed for middle managers who wish to upgrade their skills was launched 3 years ago. To judge how well the GMAT score predicts MBA performance, a sample of 12 graduates was taken. Their grade point averages in the MBA program (values from 0 to 12) and their GMAT score (values range from 200 to 800) are listed here. Compute the covariance, the coefficient of correlation, and the coefficient of determination. Interpret your findings.

GMAT and GPA Scores for 12 MBA Students

| | | | | | | |
|---|---|---|---|---|---|---|
| GMAT | 599 | 689 | 584 | 631 | 594 | 643 |
| MBA GPA | 9.6 | 8.8 | 7.4 | 10.0 | 7.8 | 9.2 |
| GMAT | 656 | 594 | 710 | 611 | 593 | 683 |
| MBA GPA | 9.6 | 8.4 | 11.2 | 7.6 | 8.8 | 8.0 |

*The following exercises require a computer and software.*

**4.68** Xr04-68 The unemployment rate is an important measure of a country's economic health. The unemployment rate measures the percentage of people who are looking for work and who are without jobs. Another way of measuring this economic variable is to calculate the employment rate, which is the percentage of adults who are employed. Here are the unemployment rates and employment rates of 19 countries. Calculate the coefficient of determination and describe what you have learned.

| Country | Unemployment Rate | Employment Rate |
|---|---|---|
| Australia | 6.7 | 70.7 |
| Austria | 3.6 | 74.8 |
| Belgium | 6.6 | 59.9 |
| Canada | 7.2 | 72.0 |
| Denmark | 4.3 | 77.0 |
| Finland | 9.1 | 68.1 |
| France | 8.6 | 63.2 |
| Germany | 7.9 | 69.0 |
| Hungary | 5.8 | 55.4 |
| Ireland | 3.8 | 67.3 |
| Japan | 5.0 | 74.3 |
| Netherlands | 2.4 | 65.4 |
| New Zealand | 5.3 | 62.3 |
| Poland | 18.2 | 53.5 |
| Portugal | 4.1 | 72.2 |
| Spain | 13.0 | 57.5 |
| Sweden | 5.1 | 73.0 |
| United Kingdom | 5.0 | 72.2 |
| United States | 4.8 | 73.1 |

(*Source:* National Post Business.)

**4.69** Xr04-69 All Canadians have government-funded health insurance, which pays for any medical care they require. However, when traveling out of the country, Canadians usually acquire supplementary health insurance to cover the difference between the costs incurred for emergency treatment and what the government program pays. In the United States, this cost differential can be prohibitive. Until recently, private insurance companies (such as BlueCross BlueShield) charged everyone the same weekly rate, regardless of age. However, because of rising costs and the realization that older people frequently incur greater medical emergency expenses, insurers had to change their premium plans. They decided to offer rates that depend on the age of the customer. To help determine the new rates, one insurance company gathered data concerning the age and mean daily medical expenses of a random sample of 1,348 Canadians during the previous 12-month period.

a. Calculate the coefficient of determination.
b. What does the statistic calculated in part (a) tell you?
c. Determine the least squares line.
d. Interpret the coefficients.
e. What rate plan would you suggest?

**4.70** Xr04-70 A real estate developer of single-family dwellings across the country is in the process of developing plans for the next several years. An analyst for the company believes that interest rates are likely to increase but remain at low levels. To help make decisions about the number of homes to build, the developer acquired the monthly bank prime rate and the number of new single-family homes sold monthly (thousands) from 1963 to 2009. (*Source:* Federal Reserve Statistics and U.S. Census Bureau.)

Calculate the coefficient of determination. Explain what this statistic tells you about the relationship between the prime bank rate and the number of single-family homes sold.

**4.71** Xr04-71 When the price of crude oil increases, do oil companies drill more oil wells? To determine the strength and nature of the relationship, an economist recorded the price of a barrel of domestic crude oil (West Texas crude) and the number of exploratory oil wells drilled for each month from 1973 to 2009. Analyze the data and explain what you have discovered. (*Source:* U.S. Department of Energy.)

**4.72** Xr04-72 One way of measuring the extent of unemployment is through the help wanted index, which measures the number of want ads in the nation's newspapers. The higher the index, the greater the demand for workers. Another measure is the unemployment rate among insured workers. An economist wanted to know whether these two variables are related and, if so, how. He acquired the help wanted index and unemployment rates for each month between 1951 and 2006 (last year available). Determine the strength and direction of the relationship. (*Source*: U.S. Department of Labor Statistics.)

**4.73** Xr04-73 A manufacturing firm produces its products in batches using sophisticated machines and equipment. The general manager wanted to investigate the relationship between direct labor costs and the number of units produced per batch. He recorded the data from the last 30 batches. Determine the fixed and variable labor costs.

**4.74** Xr04-74 A manufacturer has recorded its cost of electricity and the total number of hours of machine time for each of 52 weeks. Estimate the fixed and variable electricity costs.

**4.75** Xr04-75 The chapter-opening example showed that there is a linear relationship between a baseball team's payroll and the number of wins. This raises the question, are success on the field and attendance related? If the answer is no, then profit-driven owners may not be inclined to spend money to improve their teams. The statistics practitioner recorded the number of wins and the average home attendance for the 2009 baseball season.
a. Calculate whichever parameters you wish to help guide baseball owners.
b. Estimate the marginal number of tickets sold for each additional game won.

**4.76** Xr04-76 Refer to Exercise 4.75. The practitioner also recorded the average away attendance for each team. Visiting teams take a share of the gate, so every owner should be interested in this analysis.
a. Are visiting team attendance figures related to number of wins?
b. Estimate the marginal number of tickets sold for each additional game won.

**4.77** Xr04-77 The number of wins and payrolls for the each team in the National Basketball Association (NBA) in the 2008–2009 season were recorded.
a. Determine the marginal cost of one more win.
b. Calculate the coefficient of determination and describe what this number tells you.

**4.78** Xr04-78 The number of wins and payrolls for each team in the National Football League (NFL) in the 2009–2010 season were recorded.
a. Determine the marginal cost of one more win.
b. Calculate the coefficient of determination and describe what this number tells you.

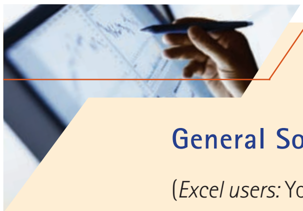
**4.79** Xr04-79 The number of wins and payrolls for each team in the National Hockey League (NHL) in the 2008–2009 season were recorded.
a. Determine the marginal cost of one more win.
b. Calculate the coefficient of determination and describe what this number tells you.

**4.80** Xr04-80 We recorded the home and away attendance for the NBA for the 2008–2009 season.
a. Analyze the relationship between the number of wins and home attendance.
b. Perform a similar analysis for away attendance.

**4.81** Xr04-81 Refer to Exercise 4.77. The relatively weak relationship between the number of wins and home attendance may be explained by the size of the arena each team plays in. The ratio of home attendance to the arena's capacity was calculated. Is percent of capacity more strongly related to the number of wins than average home attendance? Explain.

**4.82** Xr04-82 Analyze the relationship between the number of wins and home and away attendance in the National Football League in the 2009–2010 season.

**4.83** Xr04-83 Repeat Exercise 4.81 for the NFL.



## General Social Survey Exercises

(*Excel users:* You must have adjacent columns. We recommend that you copy the two columns into adjacent columns in a separate spreadsheet.)

**4.84 GSS2008* Do more educated people watch less television?**

a. To answer this question use the least squares method to determine how education (EDUC) affects the amount of time spent watching television (TVHOURS).

b. Measure the strength of the linear relationship using an appropriate statistic and explain what the statistic tells you.

**4.85 GSS2006* Using the 2006 survey, determine whether the number of years of education (EDUC) of the respondent is linearly related to the number of years of education of his or her father (PAEDUC).**

## American National Election Survey Exercise

**4.86 ANES2008* Determine whether the age (AGE) of the respondent and the amount of time he or she watches television news in a typical week (TIME2) are linearly related.**

# 4.5 (Optional) Applications in Professional Sports: Baseball

In the chapter-opening example, we provided the payrolls and the number of wins from the 2009 season. We discovered that there is a weak positive linear relationship between number of wins and payroll. The strength of the linear relationship tells us that some teams with large payrolls are not successful on the field, whereas some teams with small payrolls win a large number of games. It would appear that although the amount of money teams spend is a factor, another factor is *how* teams spend their money. In this section, we will analyze the eight seasons between 2002 and 2009 to see how small-payroll teams succeed.

Professional sports in North America is a multibillion-dollar business. The cost of a new franchise in baseball, football, basketball, and hockey is often in the hundreds of millions of dollars. Although some teams are financially successful during losing seasons, success on the field is often related to financial success. (Exercises 4.75 and 4.76 reveal that there is a financial incentive to win more games.)

It is obvious that winning teams have better players. But how does a team get better players? Teams acquire new players in three ways:

1. They can draft players from high school and college.
2. They can sign free agents on their team or on other teams.
3. They can trade with other teams.

## Drafting Players

Every year, high school and university players are drafted by major league baseball teams. The order of the draft is in reverse order of the winning percentage the previous season. Teams that rank low in the standings rank high in the draft. A team that drafts and signs a player owns the rights to that player for his first 7 years in the minor leagues and his first 6 years in the major leagues. The decision on whom to draft and in what order is made by the general manager and a group of scouts who travel the country watching high school and college games. Young players are often invited to a camp where variables such as running speed, home run power, and, for pitchers, velocity are measured. They are often judged by whether a young man "looks" like a player. That is, taller, more athletic men are judged to be better than shorter, heavier ones.

## Free Agency

For the first 3 years in the major leagues, the team can pay a player the minimum, which in 2009 was $400,000 per year. After 3 years, the player is eligible for arbitration. A successful player can usually increase his salary from $2 million to $3 million through arbitration. After 6 years, the player can become a free agent and can sign with any major league team. The top free agents can make well in excess of $10 million per year in a multiyear contract.

## Trading

Teams will often trade with each other hoping that the players they acquire will help them more than the players they traded away. Many trades produce little improvement in both teams. However, in the history of baseball, there have been some very one-sided trades that resulted in great improvement in one team and the weakening of the other.

As you can see from the solved chapter-opening example "Cost of One More Win," there is a great variation in team payrolls. In 2009, the New York Yankees spent $201 million, while the Florida Marlins spent $37 million (the amounts listed are payrolls at the beginning of the season). To a very large extent, the ability to finance expensive teams is a function of the city in which the team plays. For example, teams in New York, Los Angeles, Atlanta, and Arlington, Texas, are in large markets. Tampa Bay, Oakland, and Minnesota are small-market teams. Large-market teams can afford higher salaries because of the higher gate receipts and higher fees for local television. This means that small-market teams cannot compete for the services of the top free agency players, and thus are more likely to be less successful than large-market teams.

The question arises, can small-market teams be successful on the field and, if so, how? The answer lies in how players are assessed for the draft and for trades. The decisions about whom to draft, whom to trade for, and whom to give in return are made by the team's general manager with the assistance of his assistants and the team's scouts. Because scouts are usually former major league and minor league players who were trained by other former minor league and major league players, they tend to generally agree on the value of the players in the draft. Similarly, teams making trades often trade players of "equal" value. As a result, most teams evaluate players in the same way, yielding little differences in a player's worth. This raises the question, how can a team get the edge on other teams? The answer lies in statistics.

You won't be surprised to learn that the two most important variables in determining the number of wins are the number of runs the team scores and the number of runs the team allows. The number of runs allowed is a function of the quality of the team's pitchers and, to a lesser extent, the defense. Most major league teams evaluate pitchers on the velocity of their fastball. Velocities in the 90 to 100 mile per hour range get the scouts' attention. High school and college pitchers with fastball speeds in the 80s are seldom drafted in the early rounds even when they appear to allow fewer runs by opposing teams.

Scouts also seek out high school and college players with high batting averages and who hit home runs in high school and college.

The only way that small-budget teams can succeed is for them to evaluate players more accurately. In practice, this means that they need to judge players differently from the other teams. In the following analysis, we concentrate on the number of runs a team scores and the statistics that are related to this variable.

If the scouts are correct in their method of evaluating young players, the variables that would be most strongly related to the number of runs a team scores are batting average (BA) and the number of home runs (HR). (A player's batting average is computed by calculating the number of times the player hits divided by the number of at bats less bases on balls.) The coefficients of correlation for seasons 2002 to 2009 are listed here.

| | Year | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Coefficients of Correlation | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
| Number of runs & batting average | .828 | .889 | .803 | .780 | .672 | .762 | .680 | .748 |
| Number of runs & home runs | .682 | .747 | .765 | .713 | .559 | .536 | .617 | .744 |

Are there better statistics? In other words, are there other team statistics that correlate more highly with the number of runs a team scores? There are two candidates. The first is the teams' on-base average (OBA); the second is the slugging percentage (SLG). The OBA is the number of hits plus bases on balls plus being hit by the pitcher divided by the number of at bats. The SLG is calculated by dividing the total number of bases (single = 1, double = 2, triple = 3, and home run = 4) by the number of at bats minus bases on balls and hit by pitcher. The coefficients of correlation are listed here.

| | Year | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Coefficients of Correlation | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
| Number of runs and on-base average | .835 | .916 | .875 | .783 | .800 | .875 | .834 | .851 |
| Number of runs and slugging percentage | .913 | .951 | .929 | .790 | .854 | .885 | .903 | .911 |

As you can see, for all eight seasons the OBA had a higher correlation with runs than did the BA.

Comparing the coefficients of correlation of runs with HR and SLG, we can see that in all eight seasons SLG was more strongly correlated than was HR.

As we've pointed out previously, we cannot definitively conclude a causal relationship from these statistics. However, because most decisions are based on the BA and HR, these statistics suggest that general managers should place much greater weight on batters' ability to get on base instead of simply reading the batting averages.

## The Oakland Athletics (and a Statistics) Success Story*

From 2002 to 2006, no team was as successful as the Oakland Athletics in converting a small payroll into a large number of wins. In 2002, Oakland's payroll was $40 million and the team won 103 games. In the same season, the New York Yankees spent $126 million and won the same number of games. In 2003, Oakland won 96 games, second to

*The Oakland success story is described in the book Moneyball : *The Art of Winning an Unfair Game* by Michael Lewis, New York London: W.W. Norton

New York's 101 games. Oakland's payroll was $50 million, whereas New York's was $153 million. In 2004, Oakland won 91 games with a payroll of $59 million, and the Yankees won 101 games with a payroll of $184 million. Oakland won 88 games in 2005, and the Yankees won 95 games. Payrolls were Oakland $55 million, Yankees $208 million. In 2006, the team payrolls were Oakland $62 million, Yankees $199 million. Team wins were Oakland 93, Yankees 97.

The Athletics owe their success to general managers who were willing to rethink how teams win. The first of these general managers was Sandy Alderson, who was hired by Oakland in 1993. He was a complete outsider with no baseball experience. This was an unusual hire in an organization in which managers and general managers are either former players or individuals who worked their way up the organization after years of service in a variety of jobs. At the time, Oakland was a high-payroll team that was successful on the field. It was in the World Series in 1988, 1989, and 1990, and had the highest payroll in 1991. The team was owned by Walter A. Haas, Jr., who was willing to spend whatever was necessary to win. When he died in 1995, the new owners decided that the payroll was too large and limited it. This forced Alderson to rethink strategy.

Sandy Alderson was a lawyer and a former marine. Because he was an outsider, he approached the job in a completely different way. He examined each aspect of the game and, among other things, concluded that before three outs everything was possible, but after three outs nothing was possible. This led him to the conclusion that the way to score runs is to minimize each player's probability of making an out. Rather than judge a player by his batting average, which is the way every other general manger assessed players, it would make more sense to judge the player on his on-base average. The on-base average (explained previously) is the probability of *not* making an out. Thus was born something quite rare in baseball—a new idea.

Alderson's replacement is Billy Beane, who continued and extended Alderson's thinking, including hiring a Harvard graduate to help manage the statistics.

In the previous edition of this book, we asked the question, why don't other teams do the same? The answer in 2010 is that other teams have. In the last three years, Oakland has not been anywhere nearly as successful as it was in the previous five (winning about 75 games each year). Apparently, Oakland's approach to evaluating players has influenced other teams. However, there is still a weak linear relationship between team success and team payroll. This means that other variables affect how many games a team wins besides what the team pays its players. Perhaps some clever general manager will find these variables. If he or she does, it may be best not to publicize the discovery in another book.

## 4.6 (Optional) Applications in Finance: Market Model

In the Applications in Finance box in Chapter 3 (page 52), we introduced the terms *return on investment* and *risk*. We described two goals of investing. The first is to maximize the expected or mean return and the second is to minimize the risk. Financial analysts use a variety of statistical techniques to achieve these goals. Most investors are risk averse, which means that for them minimizing risk is of paramount importance. In Section 4.2, we pointed out that variance and standard deviation are used to measure the risk associated with investments.

## APPLICATIONS in FINANCE

### Stock Market Indexes

Stock markets such as the New York Stock Exchange (NYSE), NASDAQ, Toronto Stock Exchange (TSE), and many others around the world calculate indexes to provide information about the prices of stocks on their exchanges. A stock market index is composed of a number of stocks that more or less represent the entire market. For example, the Dow Jones Industrial Average (DJIA) is the average price of a group of 30 NYSE stocks of large publicly traded companies. The Standard and Poor's 500 Index (S&P) is the average price of 500 NYSE stocks. These indexes represent their stock exchanges and give readers a quick view of how well the exchange is doing as well as the economy of the country as a whole. The NASDAQ 100 is the average price of the 100 largest nonfinancial companies on the NASDAQ exchange. The S&P/TSX Composite Index is composed of the largest companies on the TSE.

In this section, we describe one of the most important applications of the use of a least squares line. It is the well-known and often applied *market model.* This model assumes that the rate of return on a stock is linearly related to the rate of return on the stock market index. The return on the index is calculated in the same way the return on a single stock is computed. For example, if the index at the end of last year was 10,000 and the value at the end of this year is 11,000, then the market index annual return is 10%. The return on the stock is the dependent variable $Y$, and the return on the index is the independent variable $X$.

We use the least squares line to represent the linear relationship between $X$ and $Y$. The coefficient $b_1$ is called the stock's *beta coefficient*, which measures how sensitive the stock's rate of return is to changes in the level of the overall market. For example, if $b_1$ is greater than 1, then the stock's rate of return is more sensitive to changes in the level of the overall market than the average stock. To illustrate, suppose that $b_1 = 2$. Then a 1% increase in the index results in an average increase of 2% in the stock's return. A 1% decrease in the index produces an average 2% decrease in the stock's return. Thus, a stock with a beta coefficient greater than 1 will tend to be more volatile than the market.

## EXAMPLE 4.19 Market Model for Research in Motion

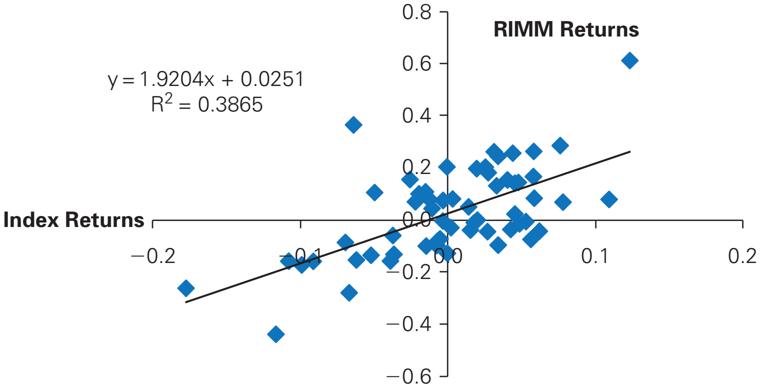
DATA
Xm04-19

The monthly rates of return for Research in Motion, maker of the BlackBerry (symbol RIMM), and the NASDAQ index (a measure of the overall NASDAQ stock market) were recorded for each month between January 2005 and December 2009. Some of these data are shown below. Estimate the market model and analyze the results.

| Month-Year | Index | RIMM |
|---|---|---|
| Jan-05 | −0.05196 | −0.13506 |
| Feb-05 | −0.00518 | −0.07239 |
| Mar-05 | −0.02558 | 0.15563 |
| Apr-05 | −0.03880 | −0.15705 |
| May-05 | 0.07627 | 0.28598 |
| Jun-05 | −0.00544 | −0.10902 |
| Jul-09 | 0.07818 | 0.06893 |
| Aug-09 | 0.01545 | −0.03856 |
| Sep-09 | 0.05642 | −0.07432 |
| Oct-09 | −0.03643 | −0.13160 |
| Nov-09 | 0.04865 | −0.01430 |
| Dec-09 | 0.05808 | 0.16670 |

SOLUTION

Excel's scatter diagram and least squares line are shown below. (Minitab produces a similar result.) We added the equation and the coefficient of determination to the scatter diagram.



We note that the slope coefficient for RIMM is 1.9204. We interpret this to mean that for each 1% increase in the NASDAQ index return in this sample, the average increase in RIMM's return is 1.9204%. Because $b_1$ is greater than 1, we conclude that the return on investing in Research in Motion is more volatile and therefore riskier than the entire NASDAQ market.

## Systematic and Firm-Specific Risk

The slope coefficient $b_1$ is a measure of the stock's *market-related* (or *systematic*) *risk* because it measures the volatility of the stock price that is related to the overall market volatility. The slope coefficient only informs us about the nature of the relationship between the two sets of returns. It tells us nothing about the *strength* of the linear relationship.

The coefficient of determination measures the proportion of the total risk that is market related. In this case, we see that 38.65% of RIMM's total risk is market related. That is, 38.65% of the variation in RIMM's returns is explained by the variation in the NASDAQ index's returns. The remaining 61.35% is the proportion of the risk that is associated with events specific to RIMM rather than the market. Financial analysts (and most everyone else) call this the *firm-specific* (or *nonsystematic) risk*. The firm-specific risk is attributable to variables and events not included in the market model, such as the effectiveness of RIMM's sales force and managers. This is the part of the risk that can be "diversified away" by creating a portfolio of stocks (as will be discussed in Section 7.3). We cannot, however, diversify away the part of the risk that is market related.

When a portfolio has been created, we can estimate its beta by averaging the betas of the stocks that compose the portfolio. If an investor believes that the market is likely to rise, then a portfolio with a beta coefficient greater than 1 is desirable. Risk-averse investors or ones who believe that the market will fall will seek out portfolios with betas less than 1.

## Exercises

*The following exercises require the use of a computer and software.*

**4.87** Xr04-87 We have recorded the monthly returns for the S&P 500 index and the following six stocks listed on the New York Stock Exchange for the period January 2005 to December 2009.

AT&T
Aetna
Cigna
Coca-Cola
Disney
Ford
McDonald's

Calculate the beta coefficient for each stock and briefly describe what it means. (*Excel users:* To use the scatter diagram to compute the beta coefficient, the data must be stored in two adjacent columns. The first must contain the returns on the index, and the second stores the returns for whichever stock whose coefficient you wish to calculate.)

**4.88** Xm04-88 Monthly returns for the Toronto Stock Exchange index and the following stocks on the Toronto Stock Exchange were recorded for the years 2005 to 2009.

Barrick Gold
Bell Canada Enterprises (BCE)
Bank of Montreal (BMO)
Enbridge
Fortis
Methanex
Research in Motion (RIM)
Telus
Trans Canada Pipeline

Calculate the beta coefficient for each stock and discuss what you have learned about each stock.

**4.89** X04-89 We calculated the returns on the NASDAQ index and the following stocks on the NASDAQ exchange for the period January 2005 to December 2009.

Amazon
Amgen
Apple
Cisco Systems
Google
Intel
Microsoft
Oracle
Research in Motion

Calculate the beta coefficient for each stock and briefly describe what it means.

## 4.7 Comparing Graphical and Numerical Techniques

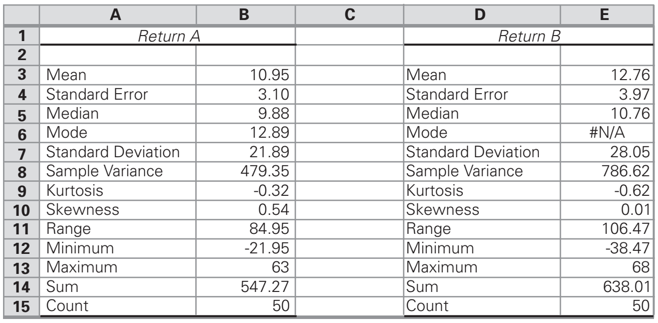
As we mentioned before, graphical techniques are useful in producing a quick picture of the data. For example, you learn something about the location, spread, and shape of a set of interval data when you examine its histogram. Numerical techniques provide the same approximate information. We have measures of central location, measures of variability, and measures of relative standing that do what the histogram does. The scatter diagram graphically describes the relationship between two interval variables, but so do the numerical measures covariance, coefficient of correlation, coefficient of determination, and least squares line. Why then do we need to learn both categories of techniques? The answer is that they differ in the information each provides. We illustrate the difference between graphical and numerical methods by redoing four examples we used to illustrate graphical techniques in Chapter 3.

### EXAMPLE 3.2 Comparing Returns on Two Investments

In Example 3.2, we wanted to judge which investment appeared to be better. As we discussed in the Applications in Finance: Return on Investment (page 52), we judge investments in terms of the return we can expect and its risk. We drew histograms and attempted to interpret them. The centers of the histograms provided us with information about the expected return and their spreads gauged the risk. However, the histograms were not clear. Fortunately, we can use numerical measures. The mean and median provide us with information about the return we can expect, and the variance or standard deviation tell us about the risk associated with each investment.

Here are the descriptive statistics produced by Excel. Minitab's are similar. (We combined the output into one worksheet.)

### Microsoft Excel Output for Example 3.2



| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | *Return A* | | | *Return B* | |
| 2 | | | | | |
| 3 | Mean | 10.95 | | Mean | 12.76 |
| 4 | Standard Error | 3.10 | | Standard Error | 3.97 |
| 5 | Median | 9.88 | | Median | 10.76 |
| 6 | Mode | 12.89 | | Mode | #N/A |
| 7 | Standard Deviation | 21.89 | | Standard Deviation | 28.05 |
| 8 | Sample Variance | 479.35 | | Sample Variance | 786.62 |
| 9 | Kurtosis | -0.32 | | Kurtosis | -0.62 |
| 10 | Skewness | 0.54 | | Skewness | 0.01 |
| 11 | Range | 84.95 | | Range | 106.47 |
| 12 | Minimum | -21.95 | | Minimum | -38.47 |
| 13 | Maximum | 63 | | Maximum | 68 |
| 14 | Sum | 547.27 | | Sum | 638.01 |
| 15 | Count | 50 | | Count | 50 |

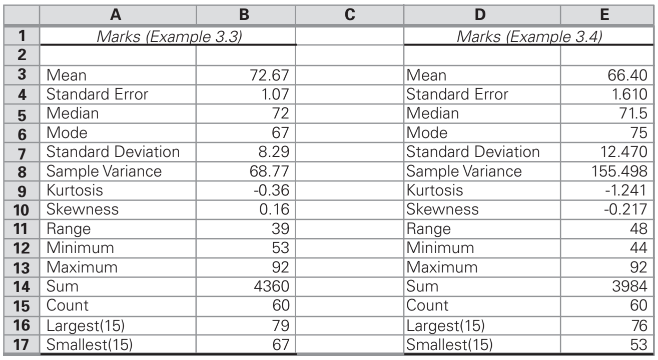
We can now see that investment B has a larger mean and median but that investment A has a smaller variance and standard deviation. If an investor were interested in low-risk investments, then he or she would choose investment A. If you reexamine the histograms from Example 3.2 (page 53), you will see that the precision provided by the numerical techniques (mean, median, and standard deviation) provides more useful information than did the histograms.

EXAMPLES 3.3 AND 3.4

## Business Statistics Marks; Mathematical Statistical Marks

In these examples we wanted to see what differences existed between the marks in the two statistics classes. Here are the descriptive statistics. (We combined the two printouts in one worksheet.)

### Microsoft Excel Output for Examples 3.3 and 3.4



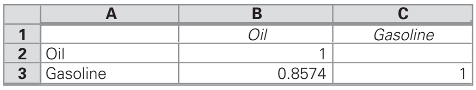
| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | *Marks (Example 3.3)* | | | *Marks (Example 3.4)* | |
| 2 | | | | | |
| 3 | Mean | 72.67 | | Mean | 66.40 |
| 4 | Standard Error | 1.07 | | Standard Error | 1.610 |
| 5 | Median | 72 | | Median | 71.5 |
| 6 | Mode | 67 | | Mode | 75 |
| 7 | Standard Deviation | 8.29 | | Standard Deviation | 12.470 |
| 8 | Sample Variance | 68.77 | | Sample Variance | 155.498 |
| 9 | Kurtosis | -0.36 | | Kurtosis | -1.241 |
| 10 | Skewness | 0.16 | | Skewness | -0.217 |
| 11 | Range | 39 | | Range | 48 |
| 12 | Minimum | 53 | | Minimum | 44 |
| 13 | Maximum | 92 | | Maximum | 92 |
| 14 | Sum | 4360 | | Sum | 3984 |
| 15 | Count | 60 | | Count | 60 |
| 16 | Largest(15) | 79 | | Largest(15) | 76 |
| 17 | Smallest(15) | 67 | | Smallest(15) | 53 |

The statistics tell us that the mean and median of the marks in the business statistics course (Example 3.3) are higher than in the mathematical statistics course (Example 3.4). We found that the histogram of the mathematical statistics marks was bimodal, which we interpreted to mean that this type of approach created differences between students. The unimodal histogram of the business statistics marks informed us that this approach eliminated those differences.

## Chapter 3 Opening Example

In this example, we wanted to know whether the prices of gasoline and oil were related. The scatter diagram did reveal a strong positive linear relationship. We can improve on the quality of this information by computing the coefficient of correlation and drawing the least squares line.

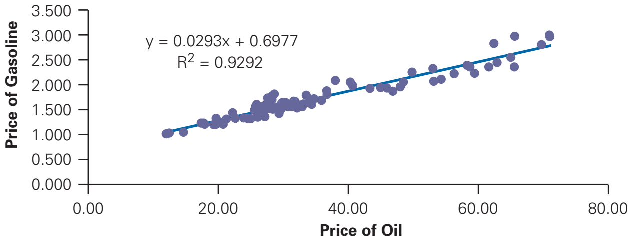
### Excel Output for Chapter 3 Opening Example: Coefficient of Correlation



| | A | B | C |
|---|---|---|---|
| 1 | | *Oil* | *Gasoline* |
| 2 | Oil | 1 | |
| 3 | Gasoline | 0.8574 | 1 |

The coefficient of correlation seems to confirm what we learned from the scatter diagram: There is a moderately strong positive linear relationship between the two variables.

### Excel Output for Chapter 3 Opening Example: Least Squares Line



The slope coefficient tells us that for each dollar increase in the price of a barrel of oil, the price of a (U.S.) gallon of gasoline increases an average of 2.9 cents. However, because there are 42 gallons per barrel, we would expect a dollar increase in a barrel of oil to yield a 2.4† cents per gallon increase (calculated as $1.00/42). It does appear that the oil companies are taking some small advantage by adding an extra half-cent per gallon. The coefficient of determination is .929, which indicates that 92.9% of the variation in gasoline prices is explained by the variation in oil prices.

†This is a simplification. In fact a barrel of oil yields a variety of other profitable products. See Exercise 2.14.

## EXERCISES

*The following exercises require a computer and statistical software.*

**4.90** Xr03-23 Refer to Exercise 3.23

a. Calculate the mean, median, and standard deviation of the scores of those who repaid and of those who defaulted.

b. Do these statistics produce more useful information than the histograms?

**4.91** Xr03-24 Refer to Exercise 3.24.

a. Draw box plots of the scores of those who repaid and of those who defaulted.

b. Compare the information gleaned from the histograms to that contained in the box plots. Which are better?

**4.92** Xr03-50 Calculate the coefficient of determination for Exercise 3.50. Is this more informative than the scatter diagram?

**4.93** Xr03-51 Refer to Exercise 3.51. Compute the coefficients of the least squares line and compare your results with the scatter diagram.

**4.94** Xr03-56 Compute the coefficient of determination and the least squares line for Exercise 3.56. Compare this information with that developed by the scatter diagram alone.

**4.95** Xr03-59 Refer to Exercise 3.59. Calculate the coefficient of determination and the least squares line. Is this more informative than the scatter diagram?

**4.96** Xm03-07 a. Calculate the coefficients of the least squares line for the data in Example 3.7.

b. Interpret the coefficients.

c. Is this information more useful than the information extracted from the scatter diagram?

**4.97** Xr04-53 In Exercise 4.53, you drew box plots. Draw histograms instead and compare the results.

**4.98** Xr04-55 Refer to Exercise 4.55. Draw histograms of the data. What have you learned?

## 4.8 GENERAL GUIDELINES FOR EXPLORING DATA

The purpose of applying graphical and numerical techniques is to describe and summarize data. Statisticians usually apply graphical techniques as a first step because we need to know the shape of the distribution. The shape of the distribution helps answer the following questions:

1. Where is the approximate center of the distribution?
2. Are the observations close to one another, or are they widely dispersed?
3. Is the distribution unimodal, bimodal, or multimodal? If there is more than one mode, where are the peaks, and where are the valleys?
4. Is the distribution symmetric? If not, is it skewed? If symmetric, is it bell shaped?

Histograms and box plots provide most of the answers. We can frequently make several inferences about the nature of the data from the shape. For example, we can assess the relative risk of investments by noting their spreads. We can attempt to improve the teaching of a course by examining whether the distribution of final grades is bimodal or skewed.

The shape can also provide some guidance on which numerical techniques to use. As we noted in this chapter, the central location of highly skewed data may be more appropriately measured by the median. We may also choose to use the interquartile range instead of the standard deviation to describe the spread of skewed data.

When we have an understanding of the structure of the data, we may do additional analysis. For example, we often want to determine how one variable, or several variables, affects another. Scatter diagrams, covariance, and the coefficient of correlation are useful techniques for detecting relationships between variables. A number of techniques to be introduced later in this book will help uncover the nature of these associations.

## CHAPTER SUMMARY

This chapter extended our discussion of descriptive statistics, which deals with methods of summarizing and presenting the essential information contained in a set of data. After constructing a frequency distribution to obtain a general idea about the distribution of a data set, we can use numerical measures to describe the central location and variability of interval data. Three popular measures of central location, or averages, are the mean, the median, and the mode. Taken by themselves, these measures provide an inadequate description of the data because they say nothing about the extent to which the data vary. Information regarding the variability of interval data is conveyed by such numerical measures as the range, variance, and standard deviation.

For the special case in which a sample of measurements has a mound-shaped distribution, the Empirical Rule provides a good approximation of the percentages of measurements that fall within one, two, and three standard deviations of the mean. Chebysheff's Theorem applies to all sets of data no matter the shape of the histogram.

Measures of relative standing that were presented in this chapter are percentiles and quartiles. The box plot graphically depicts these measures as well as several others. The linear relationship between two interval variables is measured by the covariance, the coefficient of correlation, the coefficient of determination, and the least squares line.

### IMPORTANT TERMS

| | |
|---|---|
| Measures of central location | 98 |
| Mean | 98 |
| Median | 100 |
| Mode | 101 |
| Modal class | 102 |
| Geometric mean | 105 |
| Measures of variability | 108 |
| Range | 108 |
| Variance | 108 |
| Standard deviation | 108 |
| Deviation | 109 |
| Mean absolute deviation | 110 |
| Empirical Rule | 113 |
| Chebysheff's Theorem | 114 |
| Skewed | 114 |
| Coefficient of variation | 115 |
| Percentiles | 117 |
| Quartiles | 118 |
| Interquartile range | 120 |
| Box plots | 120 |
| Outlier | 121 |
| Covariance | 127 |
| Coefficient of correlation | 128 |
| Least squares method | 132 |

### SYMBOLS

| Symbol | Pronounced | Represents |
|---|---|---|
| $\mu$ | mu | Population mean |
| $\sigma^2$ | sigma squared | Population variance |
| $\sigma$ | sigma | Population standard deviation |
| $\rho$ | rho | Population coefficient of correlation |
| $\sum$ | Sum of | Summation |

| Symbol | Pronounced | Represents |
|---|---|---|
| $\sum_{i=1}^{n} x_i$ | Sum of $x_i$ from 1 to $n$ | Summation of $n$ numbers |
| $\hat{y}$ | $y$ hat | Fitted or calculated value of $y$ |
| $b_0$ | $b$ zero | $y$-Intercept |
| $b_1$ | $b$ one | Slope coefficient |

## FORMULAS

Population mean

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

Sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Range

Largest observation – Smallest observation

Population variance

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

Sample variance

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

Population standard deviation

$$\sigma = \sqrt{\sigma^2}$$

Sample standard deviation

$$s = \sqrt{s^2}$$

Population covariance

$$\sigma_{xy} = \frac{\sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)}{N}$$

Sample covariance

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Population coefficient of correlation

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Sample coefficient of correlation

$$r = \frac{s_{xy}}{s_x s_y}$$

Coefficient of determination

$$R^2 = r^2$$

Slope coefficient

$$b_1 = \frac{s_{xy}}{s_x^2}$$

$y$-intercept

$$b_0 = \bar{y} - b_1 \bar{x}$$

## COMPUTER OUTPUT AND INSTRUCTIONS

| Technique | Excel | Minitab |
|---|---|---|
| Mean | 100 | 100 |
| Median | 101 | 101 |
| Mode | 102 | 102 |
| Variance | 111 | 111 |
| Standard deviation | 112 | 113 |
| Descriptive statistics | 119 | 120 |
| Box plot | 121 | 122 |
| Least squares line | 135 | 136 |

*(Continued)*

| Technique | Excel | Minitab |
|---|---|---|
| Covariance | 137 | 138 |
| Correlation | 137 | 138 |
| Coefficient of determination | 139 | 139 |

## Chapter Exercises

**4.99** Xr04-99* Osteoporosis is a condition in which bone density decreases, often resulting in broken bones. Bone density usually peaks at age 30 and decreases thereafter. To understand more about the condition, researchers recruited a random sample of women aged 50 and older. Each woman's bone density loss was recorded.

a. Compute the mean and median of these data.
b. Compute the standard deviation of the bone density losses.
c. Describe what you have learned from the statistics.

**4.100** Xr04-100* The temperature in December in Buffalo, New York, is often below 40 degrees Fahrenheit (4 degrees Celsius). Not surprisingly, when the National Football League Buffalo Bills play at home in December, hot coffee is a popular item at the concession stand. The concession manager would like to acquire more information so that he can manage inventories more efficiently. The number of cups of coffee sold during 50 games played in December in Buffalo were recorded.

a. Determine the mean and median.
b. Determine the variance and standard deviation.
c. Draw a box plot.
d. Briefly describe what you have learned from your statistical analysis.

**4.101** Refer to Exercise 4.99. In addition to the bone density losses, the ages of the women were also recorded. Compute the coefficient of determination and describe what this statistic tells you.

**4.102** Refer to Exercise 4.100. Suppose that in addition to recording the coffee sales, the manager also recorded the average temperature (measured in degrees Fahrenheit) during the game. These data together with the number of cups of coffee sold were recorded.

a. Compute the coefficient of determination.
b. Determine the coefficients of the least squares line.
c. What have you learned from the statistics calculated in parts (a) and (b) about the relationship between the number of cups of coffee sold and the temperature?
d. Discuss the information obtained here and in Exercise 4.100. Which is more useful to the manager?

**4.103** Xr04-103* Chris Golfnut loves the game of golf. Chris also loves statistics. Combining both passions, Chris records a sample of 100 scores.

a. What statistics should Chris compute to describe the scores?
b. Calculate the mean and standard deviation of the scores.
c. Briefly describe what the statistics computed in part (b) divulge.

**4.104** Xr04-104* The Internet is growing rapidly with an increasing number of regular users. However, among people older than 50, Internet use is still relatively low. To learn more about this issue, a sample of 250 men and women older than 50 who had used the Internet at least once were selected. The number of hours on the Internet during the past month was recorded.

a. Calculate the mean and median.
b. Calculate the variance and standard deviation.
c. Draw a box plot.
d. Briefly describe what you have learned from the statistics you calculated.

**4.105** Refer to Exercise 4.103. For each score, Chris also recorded the number of putts as well as his scores. Conduct an analysis of both sets of data. What conclusions can be achieved from the statistics?

**4.106** Refer to Exercise 4.104. In addition to Internet use, the numbers of years of education were recorded.

a. Compute the coefficient of determination.
b. Determine the coefficients of the least squares line.
c. Describe what these statistics tell you about the relationship between Internet use and education.
d. Discuss the information obtained here and in Exercise 4.104.

**4.107** Xr04-107* A sample was drawn of one-acre plots of land planted with corn. The crop yields were recorded. Calculate the descriptive statistics you judge to be useful. Interpret these statistics.

**4.108** Refer to Exercise 4.107. For each plot, the amounts of rainfall were also recorded.

a. Compute the coefficient of determination.

b. Determine the coefficients of the least squares line.

c. Describe what these statistics tell you about the relationship between crop yield and rainfall.

d. Discuss the information obtained here and in Exercise 4.107.

**4.109** Refer to Exercise 4.107. For each plot, the amounts of fertilizer were recorded.

a. Compute the coefficient of determination.

b. Determine the coefficients of the least squares line.

c. Describe what these statistics tell you about the relationship between crop yield and the amount of fertilizer.

d. Discuss the information obtained here and in Exercise 4.107.

**4.110** Xr04-110 Increasing tuition has resulted in some students being saddled with large debts at graduation. To examine this issue, a random sample of recent graduates was asked to report whether they had student loans, and, if so, how much was the debt at graduation.

a. Compute all three measures of central location.

b. What do these statistics reveal about student loan debt at graduation?

## CASE 4.1 Return to the Global Warming Question

DATA
C04-01a
C04-01b

Now that we have presented techniques that allow us to conduct more precise analyses we'll return to Case 3.1. Recall that there are two issues in this discussion. First, is there global warming; second, if so, is carbon dioxide the cause? The only tools available at the end of Chapter 3 were graphical techniques including line charts and scatter diagrams. You are now invited to apply the more precise techniques in this chapter to answer the same questions.

Here are the data sets you can work with.

C04-01a: Column 1: Months numbered 1 to 1559
Column 2: Temperature anomalies produced by the National Climatic Data Center

C04-01b: Column 1: Monthly carbon dioxide levels measured by the Mauna Loa Observatory
Column 2: Temperature anomalies produced by the National Climatic Data Center

a. Use the least squares method to estimate average monthly changes in temperature anomalies.

b. Calculate the least squares line and the coefficient of correlation between $CO_2$ levels and temperature anomalies and describe your findings.

## CASE 4.2 Another Return to the Global Warming Question

DATA
C04-02a
C04-02b
C04-02c
C04-02d

Did you conclude in Case 4.1 that Earth has warmed since 1880 and that there is some linear relationship between $CO_2$ and temperature anomalies? If so, here is another look at the same data. C04-02a lists the temperature anomalies from 1880 to 1940, C04-02b lists the data from 1941 to 1975, C04-02c stores temperature anomalies from 1976 to 1997, and C04-02d contains the data from 1998 to 2009. For each set of data, calculate the least squares line and the coefficient of determination. Report your findings.

## CASE 4.3 The Effect of the Players' Strike in the 2004-05 Hockey Season

DATA
C04-03a
C04-03b

The 2004–2005 hockey season was canceled because of a players' strike. The key issue in the labor dispute was a "salary cap." The team owners wanted a salary cap to cut their costs. The owners of small-market teams wanted the cap to help their teams become competitive. Of course, caps on salaries would lower the salaries of most players; as a result, the players association fought against it. The team owners prevailed, and the collective bargaining agreement specified a salary cap of $39 million and a floor of $21.5 million for the 2005–2006 season.

Conduct an analysis of the 2003–2004 season (C04-02a) and the 2005–2006 season (C04-02b). For each season:

a. Estimate how much on average a team needs to spend to win one more game.
b. Measure the strength of the linear relationship.
c. Discuss the differences between the two seasons.

## CASE 4.4 Quebec Referendum Vote: Was There Electoral Fraud?*

© Bettman/Corbis

DATA
C04-04

Since the 1960s, Quebecois (citizens of the province of Quebec) have been debating whether to separate from Canada and form an independent nation. A referendum was held on October 30, 1995, in which the people of Quebec voted not to separate. The vote was extremely close with the "no" side winning by only 52,448 votes. A large number of no votes was cast by the non-Francophone (non-French speaking) people of Quebec, who make up about 20% of the population and who very much want to remain Canadians. The remaining 80% are Francophones, a majority of whom voted "yes."

After the votes were counted, it became clear that the tallied vote was much closer than it should have been. Supporters of the no side charged that poll scrutineers, all of whom were appointed by the proseparatist provincial government, rejected a disproportionate number of ballots in ridings (electoral districts) where the percentage of yes votes was low and where there are large numbers of Allophone (people whose first language is neither English nor French) and Anglophone (English-speaking) residents. (Electoral laws require the rejection of ballots that do not appear to be properly marked.) They were outraged that in a strong democracy like Canada, votes would be rigged much as they are in many nondemocratic countries around the world.

If, in ridings where there was a low percentage of "yes" votes, there was a high percentage of rejected ballots, this would be evidence of electoral fraud. Moreover, if, in ridings where there were large percentages of Allophone or Anglophone voters (or both), there were high percentages of rejected ballots, this too would constitute evidence of fraud on the part of the scrutineers and possibly the government.

To determine the veracity of the charges, the following variables were recorded for each riding.

Percentage of rejected ballots in referendum

Percentage of "yes" votes

Percentage of Allophones

Percentage of Anglophones

Conduct a statistical analysis of these data to determine whether there are indications that electoral fraud took place.

*This case is based on "Voting Irregularities in the 1995 Referendum on Quebec Sovereignty" by Jason Cawley and Paul Sommers, *Chance*, Vol. 9, No. 4, Fall, 1996. We are grateful to Dr. Paul Sommers, Middlebury College, for his assistance in writing this case.

# APPENDIX 4 Review of Descriptive Techniques

Here is a list of the statistical techniques introduced in Chapters 2, 3, and 4. This is followed by a flowchart designed to help you select the most appropriate method to use to address any problem requiring a descriptive method.

To provide practice in identifying the correct descriptive method to use we have created a number of review exercises. These are in Keller's website Appendix Descriptive Techniques Review Exercises.

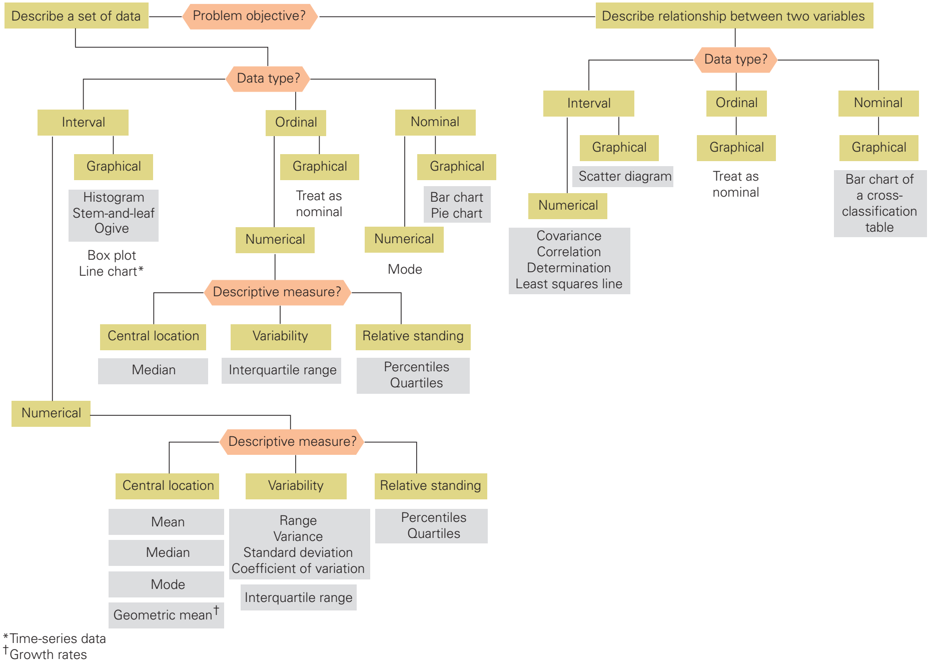
## Graphical Techniques

Histogram
Stem-and-leaf display
Ogive
Bar chart
Pie chart
Scatter diagram
Line chart (time series)
Box plot

## Numerical Techniques

Measures of Central Location
Mean
Median
Mode
Geometric mean (growth rates)
Measures of Variability
Range
Variance
Standard deviation
Coefficient of variation
Interquartile range
Measures of Relative Standing
Percentiles
Quartiles
Measures of Linear Relationship
Covariance
Coefficient of correlation
Coefficient of determination
Least squares line

## Flowchart: Graphical and Numerical Techniques



*Time-series data
†Growth rates

© Robert Hardholt/Shutterstock

# 5

# DATA COLLECTION AND SAMPLING

## Sampling and the Census

Courtesy, US Census Bureau

The census, which is conducted every 10 years in the United States, serves an important function. It is the basis for deciding how many congressional representatives and how many votes in the electoral college each state will have. Businesses often use the information derived from the census to help make decisions about products, advertising, and plant locations.

One of the problems with the census is the issue of undercounting, which occurs when some people are not included. For example, the 1990 census reported that 12.05% of adults were African American; the true value was 12.41%. To address undercounting, the Census Bureau adjusts the numbers it gets from the census. The adjustment is based on another survey. The mechanism is called the Accuracy and Coverage Evaluation. Using sampling methods described in this chapter, the Census Bureau is able to adjust the numbers in American subgroups. For example, the Bureau may discover that the number of Hispanics has been undercounted or that the number of people living in California has not been accurately counted.

Later in this chapter we'll discuss how the sampling is conducted and how the adjustments are made.

## Introduction

In Chapter 1, we briefly introduced the concept of statistical inference—the process of inferring information about a population from a sample. Because information about populations can usually be described by parameters, the statistical technique used generally deals with drawing inferences about population parameters from sample statistics. (Recall that a parameter is a measurement about a population, and a statistic is a measurement about a sample.)

Working within the covers of a statistics textbook, we can assume that population parameters are known. In real life, however, calculating parameters is virtually impossible because populations tend to be very large. As a result, most population parameters are not only unknown but also unknowable. The problem that motivates the subject of statistical inference is that we often need information about the value of parameters in order to make decisions. For example, to make decisions about whether to expand a line of clothing, we may need to know the mean annual expenditure on clothing by North American adults. Because the size of this population is approximately 200 million, determining the mean is prohibitive. However, if we are willing to accept less than 100% accuracy, we can use statistical inference to obtain an estimate. Rather than investigating the entire population, we select a sample of people, determine the annual expenditures on clothing in this group, and calculate the sample mean. Although the probability that the sample mean will equal the population mean is very small, we would expect them to be close. For many decisions, we need to know how close. We postpone that discussion until Chapters 10 and 11. In this chapter, we will discuss the basic concepts and techniques of sampling itself. But first we take a look at various sources for collecting data.

## 5.1 Methods of Collecting Data

Most of this book addresses the problem of converting data into information. The question arises, where do data come from? The answer is that a large number of methods produce data. Before we proceed however, we'll remind you of the definition of data introduced in Section 2.1. Data are the observed values of a variable; that is, we define a variable or variables that are of interest to us and then proceed to collect observations of those variables.

### Direct Observation

The simplest method of obtaining data is by direct observation. When data are gathered in this way, they are said to be **observational**. For example, suppose that a researcher for a pharmaceutical company wants to determine whether aspirin actually reduces the incidence of heart attacks. Observational data may be gathered by selecting a sample of men and women and asking each whether he or she has taken aspirin regularly over the past 2 years. Each person would be asked whether he or she had suffered a heart attack over the same period. The proportions reporting heart attacks would be compared and a statistical technique that is introduced in Chapter 13 would be used to determine whether aspirin is effective in reducing the likelihood of heart attacks. There are many drawbacks to this method. One of the most critical is that it is difficult to produce useful information in this way. For example, if the statistics practitioner concludes that people who take aspirin suffer fewer heart attacks, can we conclude that aspirin is effective? It may be that people who take aspirin tend to be more health conscious, and health-conscious people tend to have fewer heart attacks. The one advantage to direct observation is that it is relatively inexpensive.

## Experiments

A more expensive but better way to produce data is through experiments. Data produced in this manner are called **experimental**. In the aspirin illustration, a statistics practitioner can randomly select men and women. The sample would be divided into two groups. One group would take aspirin regularly, and the other would not. After 2 years, the statistics practitioner would determine the proportion of people in each group who had suffered heart attacks, and statistical methods again would be used to determine whether aspirin works. If we find that the aspirin group suffered fewer heart attacks, then we may more confidently conclude that taking aspirin regularly is a healthy decision.

## Surveys

One of the most familiar methods of collecting data is the **survey**, which solicits information from people concerning such things as their income, family size, and opinions on various issues. We're all familiar, for example, with opinion polls that accompany each political election. The Gallup Poll and the Harris Survey are two well-known surveys of public opinion whose results are often reported by the media. But the majority of surveys are conducted for private use. Private surveys are used extensively by market researchers to determine the preferences and attitudes of consumers and voters. The results can be used for a variety of purposes, from helping to determine the target market for an advertising campaign to modifying a candidate's platform in an election campaign. As an illustration, consider a television network that has hired a market research firm to provide the network with a profile of owners of luxury automobiles, including what they watch on television and at what times. The network could then use this information to develop a package of recommended time slots for Cadillac commercials, including costs, which it would present to General Motors. It is quite likely that many students reading this book will one day be marketing executives who will "live and die" by such market research data.

An important aspect of surveys is the **response rate**. The response rate is the proportion of all people who were selected who complete the survey. As we discuss in the next section, a low response rate can destroy the validity of any conclusion resulting from the statistical analysis. Statistics practitioners need to ensure that data are reliable.

**Personal Interview** Many researchers feel that the best way to survey people is by means of a personal interview, which involves an interviewer soliciting information from a respondent by asking prepared questions. A personal interview has the advantage of having a higher expected response rate than other methods of data collection. In addition, there will probably be fewer incorrect responses resulting from respondents misunderstanding some questions because the interviewer can clarify misunderstandings when asked to. But the interviewer must also be careful not to say too much for fear of biasing the response. To avoid introducing such biases, as well as to reap the potential benefits of a personal interview, the interviewer must be well trained in proper interviewing techniques and well informed on the purpose of the study. The main disadvantage of personal interviews is that they are expensive, especially when travel is involved.

**Telephone Interview** A telephone interview is usually less expensive, but it is also less personal and has a lower expected response rate. Unless the issue is of interest, many people will refuse to respond to telephone surveys. This problem is exacerbated by telemarketers trying to sell something.

**Self-Administered Survey** A third popular method of data collection is the self-administered questionnaire, which is usually mailed to a sample of people. This is an inexpensive method of conducting a survey and is therefore attractive when the number of people to be surveyed is large. But self-administered questionnaires usually have a low response rate and may have a relatively high number of incorrect responses due to respondents misunderstanding some questions.

**Questionnaire Design** Whether a questionnaire is self-administered or completed by an interviewer, it must be well designed. Proper questionnaire design takes knowledge, experience, time, and money. Some basic points to consider regarding questionnaire design follow.

1. First and foremost, the questionnaire should be kept as short as possible to encourage respondents to complete it. Most people are unwilling to spend much time filling out a questionnaire.
2. The questions themselves should also be short, as well as simply and clearly worded, to enable respondents to answer quickly, correctly, and without ambiguity. Even familiar terms such as "*unemployed*" and "*family*" must be defined carefully because several interpretations are possible.
3. Questionnaires often begin with simple demographic questions to help respondents get started and become comfortable quickly.
4. Dichotomous questions (questions with only two possible responses such as "yes" and "no" and multiple-choice questions) are useful and popular because of their simplicity, but they also have possible shortcomings. For example, a respondent's choice of yes or no to a question may depend on certain assumptions not stated in the question. In the case of a multiple-choice question, a respondent may feel that none of the choices offered is suitable.
5. Open-ended questions provide an opportunity for respondents to express opinions more fully, but they are time consuming and more difficult to tabulate and analyze.
6. Avoid using leading questions, such as "Wouldn't you agree that the statistics exam was too difficult?" These types of questions tend to lead the respondent to a particular answer.

7. Time permitting, it is useful to pretest a questionnaire on a small number of people in order to uncover potential problems such as ambiguous wording.

8. Finally, when preparing the questions, think about how you intend to tabulate and analyze the responses. First, determine whether you are soliciting values (i.e., responses) for an interval variable or a nominal variable. Then consider which type of statistical techniques—descriptive or inferential—you intend to apply to the data to be collected, and note the requirements of the specific techniques to be used. Thinking about these questions will help ensure that the questionnaire is designed to collect the data you need.

Whatever method is used to collect primary data, we need to know something about sampling, the subject of the next section.

## EXERCISES

**5.1** Briefly describe the difference between observational and experimental data.

**5.2** A soft drink manufacturer has been supplying its cola drink in bottles to grocery stores and in cans to small convenience stores. The company is analyzing sales of this cola drink to determine which type of packaging is preferred by consumers.
a. Is this study observational or experimental? Explain your answer.
b. Outline a better method for determining whether a store will be supplied with cola in bottles or in cans so that future sales data will be more helpful in assessing the preferred type of packaging.

**5.3** a. Briefly describe how you might design a study to investigate the relationship between smoking and lung cancer.
b. Is your study in part (a) observational or experimental? Explain why.

**5.4** a. List three methods of conducting a survey of people.
b. Give an important advantage and disadvantage of each of the methods listed in part (a).

**5.5** List five important points to consider when designing a questionnaire.

## 5.2 SAMPLING

The chief motive for examining a sample rather than a population is cost. Statistical inference permits us to draw conclusions about a population parameter based on a sample that is quite small in comparison to the size of the population. For example, television executives want to know the proportion of television viewers who watch a network's programs. Because 100 million people may be watching television in the United States on a given evening, determining the actual proportion of the population that is watching certain programs is impractical and prohibitively expensive. The Nielsen ratings provide approximations of the desired information by observing what is watched by a sample of 5,000 television viewers. The proportion of households watching a particular program can be calculated for the households in the Nielsen sample. This sample proportion is then used as an **estimate** of the proportion of all households (the population proportion) that watched the program.

Another illustration of sampling can be taken from the field of quality management. To ensure that a production process is operating properly, the operations manager needs to know what proportion of items being produced is defective. If the quality technician must destroy the item to determine whether it is defective, then there is no alternative to sampling: A complete inspection of the product population would destroy the entire output of the production process.

We know that the sample proportion of television viewers or of defective items is probably not exactly equal to the population proportion we want to estimate. Nonetheless, the sample statistic can come quite close to the parameter it is designed to estimate if the **target population** (the population about which we want to draw inferences) and the **sampled population** (the actual population from which the sample has been taken) are the same. In practice, these may not be the same. One of statistics' most famous failures illustrates this phenomenon.

The *Literary Digest* was a popular magazine of the 1920s and 1930s that had correctly predicted the outcomes of several presidential elections. In 1936, the *Digest* predicted that the Republican candidate, Alfred Landon, would defeat the Democratic incumbent, Franklin D. Roosevelt, by a 3 to 2 margin. But in that election, Roosevelt defeated Landon in a landslide victory, garnering the support of 62% of the electorate. The source of this blunder was the sampling procedure, and there were two distinct mistakes.* First, the *Digest* sent out 10 million sample ballots to prospective voters. However, most of the names of these people were taken from the *Digest*'s subscription list and from telephone directories. Subscribers to the magazine and people who owned telephones tended to be wealthier than average and such people then, as today, tended to vote Republican. In addition, only 2.3 million ballots were returned resulting in a self-selected sample.

**Self-selected samples** are almost always biased because the individuals who participate in them are more keenly interested in the issue than are the other members of the population. You often find similar surveys conducted today when radio and television stations ask people to call and give their opinion on an issue of interest. Again, only listeners who are concerned about the topic and have enough patience to get through to the station will be included in the sample. Hence, the sampled population is composed entirely of people who are interested in the issue, whereas the target population is made up of all the people within the listening radius of the radio station. As a result, the conclusions drawn from such surveys are frequently wrong.

An excellent example of this phenomenon occurred on ABC's *Nightline* in 1984. Viewers were given a 900 telephone number (cost: 50 cents) and asked to phone in their responses to the question of whether the United Nations should continue to be located in the United States. More than 186,000 people called, with 67% responding "no." At the same time, a (more scientific) market research poll of 500 people revealed that 72% wanted the United Nations to remain in the United States. In general, because the true value of the parameter being estimated is never known, these surveys give the impression of providing useful information. In fact, the results of such surveys are likely to be no more accurate than the results of the 1936 *Literary Digest* poll or *Nightline*'s phone-in show. Statisticians have coined two terms to describe these polls: SLOP (self-selected opinion poll) and *Oy vey* (from the Yiddish lament), both of which convey the contempt that statisticians have for such data-gathering processes.

* Many statisticians ascribe the *Literary Digest*'s statistical debacle to the wrong causes. For an understanding of what really happened, read Maurice C. Bryson, "The Literary Digest Poll: Making of a Statistical Myth" *American Statistician* 30(4) (November 1976): 184–185.

## EXERCISES

**5.6** For each of the following sampling plans, indicate why the target population and the sampled population are not the same.

a. To determine the opinions and attitudes of customers who regularly shop at a particular mall, a surveyor stands outside a large department store in the mall and randomly selects people to participate in the survey.

b. A library wants to estimate the proportion of its books that have been damaged. The librarians decide to select one book per shelf as a sample by measuring 12 inches from the left edge of each shelf and selecting the book in that location.

c. Political surveyors visit 200 residences during one afternoon to ask eligible voters present in the house at the time whom they intend to vote for.

**5.7** a. Describe why the *Literary Digest* poll of 1936 has become infamous.

b. What caused this poll to be so wrong?

**5.8** a. What is meant by *self-selected sample*?

b. Give an example of a recent poll that involved a self-selected sample.

c. Why are self-selected samples not desirable?

**5.9** A regular feature in a newspaper asks readers to respond via e-mail to a survey that requires a yes or no response. In the following day's newspaper, the percentage of yes and no responses are reported. Discuss why we should ignore these statistics.

**5.10** Suppose your statistics professor distributes a questionnaire about the course. One of the questions asks, "Would you recommend this course to a friend?" Can the professor use the results to infer something about all statistics courses? Explain.

## 5.3 Sampling Plans

Our objective in this section is to introduce three different sampling plans: simple random sampling, stratified random sampling, and cluster sampling. We begin our presentation with the most basic design.

### Simple Random Sampling

**Simple Random Sample**

A **simple random sample** is a sample selected in such a way that every possible sample with the same number of observations is equally likely to be chosen.

One way to conduct a simple random sample is to assign a number to each element in the population, write these numbers on individual slips of paper, toss them into a hat, and draw the required number of slips (the sample size, *n*) from the hat. This is the kind of procedure that occurs in raffles, when all the ticket stubs go into a large rotating drum from which the winners are selected.

Sometimes the elements of the population are already numbered. For example, virtually all adults have Social Security numbers (in the United States) or Social Insurance numbers (in Canada); all employees of large corporations have employee numbers; many people have driver's license numbers, medical plan numbers, student numbers, and so on. In such cases, choosing which sampling procedure to use is simply a matter of deciding how to select from among these numbers.

In other cases, the existing form of numbering has built-in flaws that make it inappropriate as a source of samples. Not everyone has a phone number, for example, so the telephone book does not list all the people in a given area. Many households have two (or more) adults but only one phone listing. Couples often list the phone number under the man's name, so telephone listings are likely to be disproportionately male. Some people do not have phones, some have unlisted phone numbers, and some have more than one phone; these differences mean that each element of the population does not have an equal probability of being selected.

After each element of the chosen population has been assigned a unique number, sample numbers can be selected at random. A random number table can be used to select these sample numbers. (See, for example, *CRC Standard Management Tables*, W. H. Beyer, ed., Boca Raton FL: CRC Press.) Alternatively, we can use Excel to perform this function.

## EXAMPLE 5.1 Random Sample of Income Tax Returns

A government income tax auditor has been given responsibility for 1,000 tax returns. A computer is used to check the arithmetic of each return. However, to determine whether the returns have been completed honestly, the auditor must check each entry and confirm its veracity. Because it takes, on average, 1 hour to completely audit a return and she has only 1 week to complete the task, the auditor has decided to randomly select 40 returns. The returns are numbered from 1 to 1,000. Use a computer random-number generator to select the sample for the auditor.

### SOLUTION

We generated 50 numbers between 1 and 1,000 even though we needed only 40 numbers. We did so because it is likely that there will be some duplicates. We will use the first 40 unique random numbers to select our sample. The following numbers were generated by Excel. The instructions for both Excel and Minitab are provided here. [Notice that the 24th and 36th (counting down the columns) numbers generated were the same—467.]

Computer-Generated Random Numbers

| | | | | |
|---|---|---|---|---|
| 383 | 246 | 372 | 952 | 75 |
| 101 | 46 | 356 | 54 | 199 |
| 597 | 33 | 911 | 706 | 65 |
| 900 | 165 | 467 | 817 | 359 |
| 885 | 220 | 427 | 973 | 488 |
| 959 | 18 | 304 | 467 | 512 |
| 15 | 286 | 976 | 301 | 374 |
| 408 | 344 | 807 | 751 | 986 |
| 864 | 554 | 992 | 352 | 41 |
| 139 | 358 | 257 | 776 | 231 |

### EXCEL

*INSTRUCTIONS*

1. Click **Data, Data Analysis**, and **Random Number Generation**.
2. Specify the **Number of Variables** (1) and the **Number of Random Numbers** (50).
3. Select **Uniform Distribution.**
4. Specify the range of the uniform distribution (**Parameters**) (0 and 1).
5. Click **OK**. Column A will fill with 50 numbers that range between 0 and 1.

6. Multiply column A by 1,000 and store the products in column B.
7. Make cell C1 active, and click $f_x$, **Math & Trig**, **ROUNDUP**, and **OK**.
8. Specify the first number to be rounded (**B1**).
9. Type the **number of digits** (decimal places) (**0**). Click **OK**.
10. Complete column C.

The first five steps command Excel to generate 50 uniformly distributed random numbers between 0 and 1 to be stored in column A. Steps 6 through 10 convert these random numbers to integers between 1 and 1,000. Each tax return has the same probability (1/1,000 = .001) of being selected. Thus, each member of the population is equally likely to be included in the sample.

## MINITAB

### *INSTRUCTIONS*

1. Click **Calc**, **Random Data**, and **Integer . . .**.
2. Type the number of random numbers you wish (**50**).
3. Specify where the numbers are to be stored (**C1**).
4. Specify the **Minimum value** (**1**).
5. Specify the **Maximum value** (**1000**). Click **OK**.

## INTERPRET

The auditor would examine the tax returns selected by the computer. She would pick returns numbered 383, 101, 597, . . . , 352, 776, and 75 (the first 40 unique numbers). Each of these returns would be audited to determine whether it is fraudulent. If the objective is to audit these 40 returns, no statistical procedure would be employed. However, if the objective is to estimate the proportion of all 1,000 returns that are dishonest, then she would use one of the inferential techniques presented later in this book.

## Stratified Random Sampling

In making inferences about a population, we attempt to extract as much information as possible from a sample. The basic sampling plan, simple random sampling, often accomplishes this goal at low cost. Other methods, however, can be used to increase the amount of information about the population. One such procedure is *stratified random sampling*.

### Stratified Random Sample

A **stratified random sample** is obtained by separating the population into mutually exclusive sets, or strata, and then drawing simple random samples from each stratum.

Examples of criteria for separating a population into strata (and of the strata themselves) follow.

1. Gender
    - male
    - female
2. Age
    - under 20
    - 20–30
    - 31–40
    - 41–50
    - 51–60
    - over 60
3. Occupation
    - professional
    - clerical
    - blue-collar
    - other
4. Household income
    - under $25,000
    - $25,000–$39,999
    - $40,000–$60,000
    - over $60,000

To illustrate, suppose a public opinion survey is to be conducted to determine how many people favor a tax increase. A stratified random sample could be obtained by selecting a random sample of people from each of the four income groups we just described. We usually stratify in a way that enables us to obtain particular kinds of information. In this example, we would like to know whether people in the different income categories differ in their opinions about the proposed tax increase, because the tax increase will affect the strata differently. We avoid stratifying when there is no connection between the survey and the strata. For example, little purpose is served in trying to determine whether people within religious strata have divergent opinions about the tax increase.

One advantage of stratification is that, besides acquiring information about the entire population, we can also make inferences within each stratum or compare strata. For instance, we can estimate what proportion of the lowest income group favors the tax increase, or we can compare the highest and lowest income groups to determine whether they differ in their support of the tax increase.

Any stratification must be done in such a way that the strata are mutually exclusive: Each member of the population must be assigned to exactly one stratum. After the population has been stratified in this way, we can use simple random sampling to generate the complete sample. There are several ways to do this. For example, we can draw random samples from each of the four income groups according to their proportions in the population. Thus, if in the population the relative frequencies of the four groups are as listed here, our sample will be stratified in the same proportions. If a total sample of 1,000 is to be drawn, then we will randomly select 250 from stratum 1, 400 from stratum 2, 300 from stratum 3, and 50 from stratum 4.

| Stratum | Income Categories ($) | Population Proportions (%) |
|---|---|---|
| 1 | Less than 25,000 | 25 |
| 2 | 25,000–39,999 | 40 |
| 3 | 40,000–60,000 | 30 |
| 4 | More than 60,000 | 5 |

The problem with this approach, however, is that if we want to make inferences about the last stratum, a sample of 50 may be too small to produce useful information. In such cases, we usually increase the sample size of the smallest stratum to ensure that the sample data provide enough information for our purposes. An adjustment must then be made before we attempt to draw inferences about the entire population. The required procedure is beyond the level of this book. We recommend that anyone planning such a survey consult an expert statistician or a reference book on the subject. Better still, become an expert statistician yourself by taking additional statistics courses.

## Cluster Sampling

**Cluster Sample**

A **cluster sample** is a simple random sample of groups or clusters of elements.

Cluster sampling is particularly useful when it is difficult or costly to develop a complete list of the population members (making it difficult and costly to generate a simple random sample). It is also useful whenever the population elements are widely dispersed geographically. For example, suppose we wanted to estimate the average annual household income in a large city. To use simple random sampling, we would need a complete list of households in the city from which to sample. To use stratified random sampling, we would need the list of households, and we would also need to have each household categorized by some other variable (such as age of household head) in order to develop the strata. A less-expensive alternative would be to let each block within the city represent a cluster. A sample of clusters could then be randomly selected, and every household within these clusters could be questioned to determine income. By reducing the distances the surveyor must cover to gather data, cluster sampling reduces the cost.

But cluster sampling also increases sampling error (see Section 5.4) because households belonging to the same cluster are likely to be similar in many respects, including household income. This can be partially offset by using some of the cost savings to choose a larger sample than would be used for a simple random sample.

## Sample Size

Whichever type of sampling plan you select, you still have to decide what size sample to use. Determining the appropriate sample size will be addressed in detail in Chapters 10 and 12. Until then, we can rely on our intuition, which tells us that the larger the sample size is, the more accurate we can expect the sample estimates to be.

## Sampling and the Census

Courtesy, US Census Bureau

To adjust for undercounting, the Census Bureau conducts cluster sampling. The clusters are geographic blocks. For the year 2000 census, the bureau randomly sampled 11,800 blocks, which contained 314,000 housing units. Each unit was intensively revisited to ensure that all residents were counted. From the results of this survey, the Census Bureau estimated the number of people missed by the first census in various subgroups, defined by several variables including gender, race, and age. Because of the importance of determining state populations, adjustments were made to state totals. For example, by comparing the results of the census and of the sampling, the Bureau determined that the undercount in the state of Texas was 1.7087%. The official census produced a state population of 20,851,820. Taking 1.7087% of this total produced an adjustment of 356,295. Using this method changed the population of the state of Texas to 21,208,115.

It should be noted that this process is contentious. The controversy concerns the way in which subgroups are defined. Changing the definition alters the undercounts, making this statistical technique subject to politicking.

## EXERCISES

**5.11** A statistics practitioner would like to conduct a survey to ask people their views on a proposed new shopping mall in their community. According to the latest census, there are 500 households in the community. The statistician has numbered each household (from 1 to 500), and she would like to randomly select 25 of these households to participate in the study. Use Excel or Minitab to generate the sample.

**5.12** A safety expert wants to determine the proportion of cars in his state with worn tire treads. The state license plate contains six digits. Use Excel or Minitab to generate a sample of 20 cars to be examined.

**5.13** A large university campus has 60,000 students. The president of the students' association wants to conduct a survey of the students to determine their views on an increase in the student activity fee. She would like to acquire information about all the students but would also like to compare the school of business, the faculty of arts and sciences, and the graduate school. Describe a sampling plan that accomplishes these goals.

**5.14** A telemarketing firm has recorded the households that have purchased one or more of the company's products. These number in the millions. The firm would like to conduct a survey of purchasers to acquire information about their attitude concerning the timing of the telephone calls. The president of the company would like to know the views of all purchasers but would also like to compare the attitudes of people in the West, South, North, and East. Describe a suitable sampling plan.

**5.15** The operations manager of a large plant with four departments wants to estimate the person-hours lost per month from accidents. Describe a sampling plan that would be suitable for estimating the plantwide loss and for comparing departments.

**5.16** A statistics practitioner wants to estimate the mean age of children in his city. Unfortunately, he does not have a complete list of households. Describe a sampling plan that would be suitable for his purposes.

## 5.4 SAMPLING AND NONSAMPLING ERRORS

Two major types of error can arise when a sample of observations is taken from a population: *sampling error* and *nonsampling error*. Anyone reviewing the results of sample surveys and studies, as well as statistics practitioners conducting surveys and applying statistical techniques, should understand the sources of these errors.

### Sampling Error

**Sampling error** refers to differences between the sample and the population that exists only because of the observations that happened to be selected for the sample. Sampling error is an error that we expect to occur when we make a statement about a population that is based only on the observations contained in a sample taken from the population.

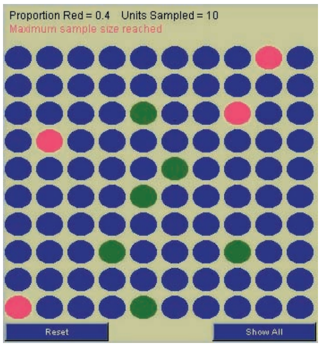
To illustrate, suppose that we wish to determine the mean annual income of North American blue-collar workers. To determine this parameter we would have to ask each North American blue-collar worker what his or her income is and then calculate the mean of all the responses. Because the size of this population is several million, the task is both expensive and impractical. We can use statistical inference to estimate the mean income $\mu$ of the population if we are willing to accept less than 100% accuracy. We record the incomes of a sample of the workers and find the mean $\bar{x}$ of this sample of incomes. This sample mean is an estimate, of the desired, population mean. But the value of the sample mean will deviate from the population mean simply by chance because the value of the sample mean depends on which incomes just happened to be selected for the sample. The difference between the true (unknown) value of the population mean and its estimate, the sample mean, is the sampling error. The size of this deviation may be large simply because of bad luck—bad luck that a particularly unrepresentative sample happened to be selected. The only way we can reduce the expected size of this error is to take a larger sample.

Given a fixed sample size, the best we can do is to state the probability that the sampling error is less than a certain amount (as we will discuss in Chapter 10). It is common today for such a statement to accompany the results of an opinion poll. If an opinion poll states that, based on sample results, the incumbent candidate for mayor has the support of 54% of eligible voters in an upcoming election, the statement may be accompanied by the following explanatory note: “This percentage is correct to within three percentage points, 19 times out of 20.” This statement means that we estimate that the actual level of support for the candidate is between 51% and 57%, and that in the long run this type of procedure is correct 95% of the time.

SEEING STATISTICS

## applet 3 Sampling

When you select this applet, you will see 100 circles. Imagine that each of the circles represents a household. You want to estimate the proportion of households having high-speed Internet access (DSL, cable modem, etc.). You may collect data from a sample of 10 households by clicking on a household's circle. If the circle turns red, the household has high-speed Internet access. If the circle turns green, the household does not have high-speed access. After collecting your sample and obtaining your estimate, click on the **Show All** button to see information for all the households. How well did your sample estimate the true proportion? Click the **Reset** button to try again. (*Note*: This page uses a randomly determined base proportion each time it is loaded or reloaded.)



Proportion Red = 0.4 Units Sampled = 10
Maximum sample size reached

**Applet Exercises**

3.1 Run the applet 25 times. How many times did the sample proportion equal the population proportion?

3.2 Run the applet 20 times. For each simulation, record the sample proportion of homes with high-speed Internet access as well as the population proportion. Compute the average sampling error.

## Nonsampling Error

Nonsampling error is more serious than sampling error because taking a larger sample won't diminish the size, or the possibility of occurrence, of this error. Even a census can (and probably will) contain nonsampling errors. **Nonsampling errors** result from mistakes made in the acquisition of data or from the sample observations being selected improperly.

1. *Errors in data acquisition*. This type of error arises from the recording of incorrect responses. Incorrect responses may be the result of incorrect measurements being

taken because of faulty equipment, mistakes made during transcription from primary sources, inaccurate recording of data because terms were misinterpreted, or inaccurate responses were given to questions concerning sensitive issues such as sexual activity or possible tax evasion.

2. *Nonresponse error*: **Nonresponse error** refers to error (or **bias**) introduced when responses are not obtained from some members of the sample. When this happens, the sample observations that are collected may not be representative of the target population, resulting in biased results (as was discussed in Section 5.2). Nonresponse can occur for a number of reasons. An interviewer may be unable to contact a person listed in the sample, or the sampled person may refuse to respond for some reason. In either case, responses are not obtained from a sampled person, and bias is introduced. The problem of nonresponse is even greater when self-administered questionnaires are used rather than an interviewer, who can attempt to reduce the nonresponse rate by means of callbacks. As noted previously, the *Literary Digest* fiasco was largely the result of a high nonresponse rate, resulting in a biased, self-selected sample.

3. *Selection bias*. **Selection bias** occurs when the sampling plan is such that some members of the target population cannot possibly be selected for inclusion in the sample. Together with nonresponse error, selection bias played a role in the *Literary Digest* poll being so wrong, as voters without telephones or without a subscription to *Literary Digest* were excluded from possible inclusion in the sample taken.

## EXERCISES

**5.17** a. Explain the difference between sampling error and nonsampling error.
b. Which type of error in part (a) is more serious? Why?

**5.18** Briefly describe three types of nonsampling error.

**5.19** Is it possible for a sample to yield better results than a census? Explain.

## CHAPTER SUMMARY

Because most populations are very large, it is extremely costly and impractical to investigate each member of the population to determine the values of the parameters. As a practical alternative, we take a sample from the population and use the sample statistics to draw inferences about the parameters. Care must be taken to ensure that the **sampled population** is the same as the **target population**.

We can choose from among several different sampling plans, including **simple random sampling**, **stratified random sampling**, and **cluster sampling**. Whatever sampling plan is used, it is important to realize that both **sampling error** and **nonsampling error** will occur and to understand what the sources of these errors are.

### IMPORTANT TERMS

Observational 162
Experimental 163
Survey 163
Response rate 163
Estimate 165
Target population 166
Sampled population 166
Self-selected sample 166
Simple random sample 167
Stratified random sample 169
Cluster sample 171
Sampling error 172
Nonsampling error 173
Nonresponse error (bias) 174
Selection bias 174

6

© Gary Buss/Taxi/Getty Images

# PROBABILITY

## Auditing Tax Returns

Government auditors routinely check tax returns to determine whether calculation errors were made. They also attempt to detect fraudulent returns. There are several methods that dishonest taxpayers use to evade income tax. One method is not to declare various sources of income. Auditors have several detection methods, including spending patterns. Another form of tax fraud is to invent deductions that are not real. After analyzing the returns of thousands of self-employed taxpayers, an auditor has determined that 45% of fraudulent returns contain two suspicious deductions, 28% contain one suspicious deduction, and the rest no suspicious deductions. Among honest returns the rates are 11% for two deductions, 18% for one deduction, and 71% for no deductions. The auditor believes that 5% of the returns of self-employed individuals contain significant fraud. The auditor has just received a tax return for a self-employed individual that contains one suspicious expense deduction. What is the probability that this tax return contains significant fraud?

© Gary Buss/Taxi/Getty Images

See page 202 for the answer.

## Introduction

In Chapters 2, 3, and 4, we introduced graphical and numerical descriptive methods. Although the methods are useful on their own, we are particularly interested in developing statistical inference. As we pointed out in Chapter 1, statistical inference is the process by which we acquire information about populations from samples. A critical component of inference is *probability* because it provides the link between the population and the sample.

Our primary objective in this and the following two chapters is to develop the probability-based tools that are at the basis of statistical inference. However, probability can also play a critical role in decision making, a subject we explore in Chapter 22.

## 6.1 Assigning Probability to Events

To introduce probability, we must first define a *random experiment*.

**Random Experiment**

A **random experiment** is an action or process that leads to one of several possible outcomes.

Here are six illustrations of random experiments and their outcomes.

**Illustration 1.** Experiment: Flip a coin.
Outcomes: Heads and tails

**Illustration 2.** Experiment: Record marks on a statistics test (out of 100).
Outcomes: Numbers between 0 and 100

**Illustration 3.** Experiment: Record grade on a statistics test.
Outcomes: A, B, C, D, and F

**Illustration 4.** Experiment: Record student evaluations of a course.
Outcomes: Poor, fair, good, very good, and excellent

**Illustration 5.** Experiment: Measure the time to assemble a computer.
Outcomes: Number whose smallest possible value is 0 seconds with no predefined upper limit

**Illustration 6.** Experiment: Record the party that a voter will vote for in an upcoming election.
Outcomes: Party A, Party B, . . .

The first step in assigning probabilities is to produce a list of the outcomes. The listed outcomes must be **exhaustive**, which means that all possible outcomes must be included. In addition, the outcomes must be **mutually exclusive**, which means that no two outcomes can occur at the same time.

To illustrate the concept of exhaustive outcomes consider this list of the outcomes of the toss of a die:

1 2 3 4 5

This list is not exhaustive, because we have omitted 6.

The concept of mutual exclusiveness can be seen by listing the following outcomes in illustration 2:

0–50  50–60  60–70  70–80  80–100

If these intervals include both the lower and upper limits, then these outcomes are not mutually exclusive because two outcomes can occur for any student. For example, if a student receives a mark of 70, both the third and fourth outcomes occur.

Note that we could produce more than one list of exhaustive and mutually exclusive outcomes. For example, here is another list of outcomes for illustration 3:

Pass and fail

A list of exhaustive and mutually exclusive outcomes is called a *sample space* and is denoted by $S$. The outcomes are denoted by $O_1, O_2, \ldots, O_k$.

**Sample Space**

A **sample space** of a random experiment is a list of all possible outcomes of the experiment. The outcomes must be exhaustive and mutually exclusive.

Using set notation, we represent the sample space and its outcomes as

$$S = \{O_1, O_2, \ldots, O_k\}$$

Once a sample space has been prepared we begin the task of assigning probabilities to the outcomes. There are three ways to assign probability to outcomes. However it is done, there are two rules governing probabilities as stated in the next box.

**Requirements of Probabilities**

Given a sample space $S = \{O_1, O_2, \ldots, O_k\}$, the probabilities assigned to the outcomes must satisfy two requirements.

1. The probability of any outcome must lie between 0 and 1; that is,

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

[Note: $\mathrm{P}(O_i)$ is the notation we use to represent the probability of outcome *i*.]

2. The sum of the probabilities of all the outcomes in a sample space must be 1. That is,

$$\sum_{i=1}^{k} P(O_i) = 1$$

## Three Approaches to Assigning Probabilities

The **classical approach** is used by mathematicians to help determine probability associated with games of chance. For example, the classical approach specifies that the probabilities of heads and tails in the flip of a balanced coin are equal to each other.

Because the sum of the probabilities must be 1, the probability of heads and the probability of tails are both 50%. Similarly, the six possible outcomes of the toss of a balanced die have the same probability; each is assigned a probability of 1/6. In some experiments, it is necessary to develop mathematical ways to count the number of outcomes. For example, to determine the probability of winning a lottery, we need to determine the number of possible combinations. For details on how to count events, see Keller's website Appendix Counting Formulas.

The **relative frequency approach** defines probability as the long-run relative frequency with which an outcome occurs. For example, suppose that we know that of the last 1,000 students who took the statistics course you're now taking, 200 received a grade of *A*. The relative frequency of *A*'s is then 200/1000 or 20%. This figure represents an estimate of the probability of obtaining a grade of *A* in the course. It is only an estimate because the relative frequency approach defines probability as the "long-run" relative frequency. One thousand students do not constitute the long run. The larger the number of students whose grades we have observed, the better the estimate becomes. In theory, we would have to observe an infinite number of grades to determine the exact probability.

When it is not reasonable to use the classical approach and there is no history of the outcomes, we have no alternative but to employ the **subjective approach**. In the subjective approach, we define probability as the degree of belief that we hold in the occurrence of an event. An excellent example is derived from the field of investment. An investor would like to know the probability that a particular stock will increase in value. Using the subjective approach, the investor would analyze a number of factors associated with the stock and the stock market in general and, using his or her judgment, assign a probability to the outcomes of interest.

## Defining Events

An individual outcome of a sample space is called a *simple event*. All other events are composed of the simple events in a sample space.

**Event**

An **event** is a collection or set of one or more simple events in a sample space.

In illustration 2, we can define the event, achieve a grade of *A*, as the set of numbers that lie between 80 and 100, inclusive. Using set notation, we have

$$A = \{80, 81, 82, \ldots, 99, \ 100\}$$

Similarly,

$$F = \{0, 1, 2, \ldots, 48, 49\}$$

## Probability of Events

We can now define the probability of any event.

**Probability of an Event**

The probability of an event is the sum of the probabilities of the simple events that constitute the event.

For example, suppose that in illustration 3, we employed the relative frequency approach to assign probabilities to the simple events as follows:

$$P(A) = .20$$
$$P(B) = .30$$
$$P(C) = .25$$
$$P(D) = .15$$
$$P(F) = .10$$

The probability of the event, pass the course, is

$$P(\text{Pass the course}) = P(A) + P(B) + P(C) + P(D) = .20 + .30 + .25 + .15 = .90$$

## Interpreting Probability

No matter what method was used to assign probability, we interpret it using the relative frequency approach for an infinite number of experiments. For example, an investor may have used the subjective approach to determine that there is a 65% probability that a particular stock's price will increase over the next month. However, we interpret the 65% figure to mean that if we had an infinite number of stocks with exactly the same economic and market characteristics as the one the investor will buy, 65% of them will increase in price over the next month. Similarly, we can determine that the probability of throwing a 5 with a balanced die is 1/6. We may have used the classical approach to determine this probability. However, we interpret the number as the proportion of times that a 5 is observed on a balanced die thrown an infinite number of times.

This relative frequency approach is useful to interpret probability statements such as those heard from weather forecasters or scientists. You will also discover that this is the way we link the population and the sample in statistical inference.

## EXERCISES

**6.1** The weather forecaster reports that the probability of rain tomorrow is 10%.
a. Which approach was used to arrive at this number?
b. How do you interpret the probability?

**6.2** A sportscaster states that he believes that the probability that the New York Yankees will win the World Series this year is 25%.
a. Which method was used to assign that probability?
b. How would you interpret the probability?

**6.3** A quiz contains a multiple-choice question with five possible answers, only one of which is correct. A student plans to guess the answer because he knows absolutely nothing about the subject.
a. Produce the sample space for each question.
b. Assign probabilities to the simple events in the sample space you produced.
c. Which approach did you use to answer part (b)?
d. Interpret the probabilities you assigned in part (b).

**6.4** An investor tells you that in her estimation there is a 60% probability that the Dow Jones Industrial Averages index will increase tomorrow.

a. Which approach was used to produce this figure?
b. Interpret the 60% probability.

**6.5** The sample space of the toss of a fair die is

$$S = \{1, 2, 3, 4, 5, 6\}$$

If the die is balanced each simple event has the same probability. Find the probability of the following events.

a. An even number
b. A number less than or equal to 4
c. A number greater than or equal to 5

**6.6** Four candidates are running for mayor. The four candidates are Adams, Brown, Collins, and Dalton. Determine the sample space of the results of the election.

**6.7** Refer to Exercise 6.6. Employing the subjective approach a political scientist has assigned the following probabilities:

$P(\text{Adams wins}) = .42$

$P(\text{Brown wins}) = .09$

$P(\text{Collins wins}) = .27$

$P(\text{Dalton wins}) = .22$

Determine the probabilities of the following events.

a. Adams loses.
b. Either Brown or Dalton wins.
c. Adams, Brown, or Collins wins.

**6.8** The manager of a computer store has kept track of the number of computers sold per day. On the basis of this information, the manager produced the following list of the number of daily sales.

| Number of Computers Sold | Probability |
|---|---|
| 0 | .08 |
| 1 | .17 |
| 2 | .26 |
| 3 | .21 |
| 4 | .18 |
| 5 | .10 |

a. If we define the experiment as observing the number of computers sold tomorrow, determine the sample space.
b. Use set notation to define the event, sell more than three computers.
c. What is the probability of selling five computers?
d. What is the probability of selling two, three, or four computers?
e. What is the probability of selling six computers?

**6.9** Three contractors (call them contractors 1, 2, and 3) bid on a project to build a new bridge. What is the sample space?

**6.10** Refer to Exercise 6.9. Suppose that you believe that contractor 1 is twice as likely to win as contractor 3 and that contractor 2 is three times as likely to win as contactor 3. What are the probabilities of winning for each contractor?

**6.11** Shoppers can pay for their purchases with cash, a credit card, or a debit card. Suppose that the proprietor of a shop determines that 60% of her customers use a credit card, 30% pay with cash, and the rest use a debit card.

a. Determine the sample space for this experiment.
b. Assign probabilities to the simple events.
c. Which method did you use in part (b)?

**6.12** Refer to Exercise 6.11.

a. What is the probability that a customer does not use a credit card?
b. What is the probability that a customer pays in cash or with a credit card?

**6.13** A survey asks adults to report their marital status. The sample space is $S = \{\text{single, married, divorced, widowed}\}$. Use set notation to represent the event the adult is not married.

**6.14** Refer to Exercise 6.13. Suppose that in the city in which the survey is conducted, 50% of adults are married, 15% are single, 25% are divorced, and 10% are widowed.

a. Assign probabilities to each simple event in the sample space.
b. Which approach did you use in part (a)?

**6.15** Refer to Exercises 6.13 and 6.14. Find the probability of each of the following events.

a. The adult is single.
b. The adult is not divorced
c. The adult is either widowed or divorced.

## 6.2 Joint, Marginal, and Conditional Probability

In the previous section, we described how to produce a sample space and assign probabilities to the simple events in the sample space. Although this method of determining probability is useful, we need to develop more sophisticated methods. In this section, we discuss how to calculate the probability of more complicated events from the probability of related events. Here is an illustration of the process.

The sample space for the toss of a die is

$$S = \{1, 2, 3, 4, 5, 6\}$$

If the die is balanced, the probability of each simple event is 1/6. In most parlor games and casinos, players toss two dice. To determine playing and wagering strategies, players need to compute the probabilities of various totals of the two dice. For example, the probability of tossing a total of 3 with two dice is 2/36. This probability was derived by creating combinations of the simple events. There are several different types of combinations. One of the most important types is the *intersection* of two events.

## Intersection

**Intersection of Events *A* and *B***

The **intersection** of events *A* and *B* is the event that occurs when both *A* and *B* occur. It is denoted as

*A* and *B*

The probability of the intersection is called the **joint probability**.

For example, one way to toss a 3 with two dice is to toss a 1 on the first die *and* a 2 on the second die, which is the intersection of two simple events. Incidentally, to compute the probability of a total of 3, we need to combine this intersection with another intersection, namely, a 2 on the first die and a 1 on the second die. This type of combination is called a *union* of two events, and it will be described later in this section. Here is another illustration.

APPLICATIONS in FINANCE

© AP Photo/Charles Bennett

### Mutual funds

A mutual fund is a pool of investments made on behalf of people who share similar objectives. In most cases, a professional manager who has been educated in finance and statistics manages the fund. He or she makes decisions to buy and sell individual stocks and bonds in accordance with a specified investment philosophy. For example, there are funds that concentrate on other publicly traded mutual fund companies. Other mutual funds specialize in Internet stocks (so-called dot-coms), whereas others buy stocks of biotechnology firms. Surprisingly, most mutual funds do not outperform the market; that is, the increase in the net asset value (NAV) of the mutual fund is often less than the increase in the value of stock indexes that represent their stock markets. One reason for this is the management expense ratio (MER) which is a measure of the costs charged to the fund by the manager to cover expenses, including the salary and bonus of the managers. The MERs for most funds range from .5% to more than 4%. The ultimate success of the fund depends on the skill and knowledge of the fund manager. This raises the question, which managers do best?

## EXAMPLE 6.1 Determinants of Success among Mutual Fund Managers—Part 1*

Why are some mutual fund managers more successful than others? One possible factor is the university where the manager earned his or her master of business administration (MBA). Suppose that a potential investor examined the relationship between how well the mutual fund performs and where the fund manager earned his or her MBA. After the analysis, Table 6.1, a table of joint probabilities, was developed. Analyze these probabilities and interpret the results.

TABLE **6.1** Determinants of Success among Mutual Fund Managers, Part 1*

| | MUTUAL FUND OUTPERFORMS MARKET | MUTUAL FUND DOES NOT OUTPERFORM MARKET |
|---|---|---|
| Top-20 MBA program | .11 | .29 |
| Not top-20 MBA program | .06 | .54 |

Table 6.1 tells us that the joint probability that a mutual fund outperforms the market *and* that its manager graduated from a top-20 MBA program is .11; that is, 11% of all mutual funds outperform the market and their managers graduated from a top-20 MBA program. The other three joint probabilities are defined similarly:

The probability that a mutual fund outperforms the market and its manager did not graduate from a top-20 MBA program is .06.

The probability that a mutual fund does not outperform the market and its manager graduated from a top-20 MBA program is .29.

The probability that a mutual fund does not outperform the market and its manager did not graduate from a top-20 MBA program is .54.

To help make our task easier, we'll use notation to represent the events. Let

$A_1$ = Fund manager graduated from a top-20 MBA program

$A_2$ = Fund manager did not graduate from a top-20 MBA program

$B_1$ = Fund outperforms the market

$B_2$ = Fund does not outperform the market

Thus,

$$P(A_1 \text{ and } B_1) = .11$$
$$P(A_2 \text{ and } B_1) = .06$$
$$P(A_1 \text{ and } B_2) = .29$$
$$P(A_2 \text{ and } B_2) = .54$$

*This example is adapted from "Are Some Mutual Fund Managers Better than Others? Cross-Sectional Patterns in Behavior and Performance" by Judith Chevalier and Glenn Ellison, Working paper 5852, National Bureau of Economic Research.

## Marginal Probability

The joint probabilities in Table 6.1 allow us to compute various probabilities. **Marginal probabilities**, computed by adding across rows or down columns, are so named because they are calculated in the margins of the table.

Adding across the first row produces

$$P(A_1 \text{ and } B_1) + P(A_1\text{and } B_2) = .11 + .29 = .40$$

Notice that both intersections state that the manager graduated from a top-20 MBA program (represented by $A_1$). Thus, when randomly selecting mutual funds, the probability that its manager graduated from a top-20 MBA program is .40. Expressed as relative frequency, 40% of all mutual fund managers graduated from a top-20 MBA program.

Adding across the second row:

$$P(A_2 \text{ and } B_1) + P(A_2 \text{ and } B_2) = .06 + .54 = .60$$

This probability tells us that 60% of all mutual fund managers did not graduate from a top-20 MBA program (represented by $A_2$). Notice that the probability that a mutual fund manager graduated from a top-20 MBA program and the probability that the manager did not graduate from a top-20 MBA program add to 1.

Adding down the columns produces the following marginal probabilities.

Column 1: $P(A_1 \text{ and } B_1) + P(A_2 \text{ and } B_1) = .11 + .06 = .17$

Column 2: $P(A_1 \text{ and } B_2) + P(A_2 \text{ and } B_2) = .29 + .54 = .83$

These marginal probabilities tell us that 17% of all mutual funds outperform the market and that 83% of mutual funds do not outperform the market.

Table 6.2 lists all the joint and marginal probabilities.

TABLE **6.2** Joint and Marginal Probabilities

| | MUTUAL FUND OUTPERFORMS MARKET | MUTUAL FUND DOES NOT OUTPERFORM MARKET | TOTALS |
|---|---|---|---|
| Top-20 MBA program | $P(A_1 \text{ and } B_1) = .11$ | $P(A_1 \text{ and } B_2) = .29$ | $P(A_1) = .40$ |
| Not top-20 MBA program | $P(A_2 \text{ and } B_1) = .06$ | $P(A_2 \text{ and } B_2) = .54$ | $P(A_2) = .60$ |
| Totals | $P(B_1) = .17$ | $P(B_2) = .83$ | 1.00 |

## Conditional Probability

We frequently need to know how two events are related. In particular, we would like to know the probability of one event given the occurrence of another related event. For example, we would certainly like to know the probability that a fund managed by a graduate of a top-20 MBA program will outperform the market. Such a probability will allow us to make an informed decision about where to invest our money. This probability is called a **conditional probability** because we want to know the probability that a fund will outperform the market *given* the condition that the manager graduated from a top-20 MBA program. The conditional probability that we seek is represented by

$$P(B_1|A_1)$$

where the "|" represents the word *given*. Here is how we compute this conditional probability.

The marginal probability that a manager graduated from a top-20 MBA program is .40, which is made up of two joint probabilities. They are (1) the probability that the mutual fund outperforms the market and the manager graduated from a top-20 MBA program [P($A_1$ and $B_1$)] and (2) the probability that the fund does not outperform the market and the manager graduated from a top-20 MBA program [P($A_1$ and $B_2$)]. Their joint probabilities are .11 and .29, respectively. We can interpret these numbers in the following way. On average, for every 100 mutual funds, 40 will be managed by a graduate of a top-20 MBA program. Of these 40 managers, on average 11 of them will manage a mutual fund that will outperform the market. Thus, the conditional probability is 11/40 = .275. Notice that this ratio is the same as the ratio of the joint probability to the marginal probability .11/.40. All conditional probabilities can be computed this way.

**Conditional Probability**

The probability of event *A* given event *B* is

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

The probability of event *B* given event *A* is

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

EXAMPLE 6.2

## Determinants of Success among Mutual Fund Managers—Part 2

Suppose that in Example 6.1 we select one mutual fund at random and discover that it did not outperform the market. What is the probability that a graduate of a top-20 MBA program manages it?

### SOLUTION

We wish to find a conditional probability. The condition is that the fund did not outperform the market (event $B_2$), and the event whose probability we seek is that the fund is managed by a graduate of a top-20 MBA program (event $A_1$). Thus, we want to compute the following probability:

$$P(A_1|B_2)$$

Using the conditional probability formula, we find

$$P(A_1|B_2) = \frac{P(A_1 \text{ and } B_2)}{P(B_2)} = \frac{.29}{.83} = .349$$

Thus, 34.9% of all mutual funds that do not outperform the market are managed by top-20 MBA program graduates.

The calculation of conditional probabilities raises the question of whether the two events, the fund outperformed the market and the manager graduated from a top-20 MBA program, are related, a subject we tackle next.

## Independence

One of the objectives of calculating conditional probability is to determine whether two events are related. In particular, we would like to know whether they are **independent events**.

**Independent Events**

Two events $A$ and $B$ are said to be independent if

$$P(A|B) = P(A)$$

or

$$P(B|A) = P(B)$$

Put another way, two events are independent if the probability of one event is not affected by the occurrence of the other event.

## EXAMPLE 6.3 Determinants of Success among Mutual Fund Managers—Part 3

Determine whether the event that the manager graduated from a top-20 MBA program and the event the fund outperforms the market are independent events.

### SOLUTION

We wish to determine whether $A_1$ and $B_1$ are independent. To do so, we must calculate the probability of $A_1$ given $B_1$; that is,

$$P(A_1|\, B_1) = \frac{P(A_1 \text{ and } B_1)}{P(B_1)} = \frac{.11}{.17} = .647$$

The marginal probability that a manager graduated from a top-20 MBA program is

$$P(A_1) = .40$$

Since the two probabilities are not equal, we conclude that the two events are dependent.

Incidentally, we could have made the decision by calculating $P(B_1|A_1) = .275$ and observing that it is not equal to $P(B_1) = .17$.

Note that there are three other combinations of events in this problem. They are ($A_1$ and $B_2$), ($A_2$ and $B_1$), ($A_2$ and $B_2$) [ignoring mutually exclusive combinations ($A_1$ and $A_2$) and ($B_1$ and $B_2$), which are dependent]. In each combination, the two events are dependent. In this type of problem, where there are only four combinations, if one combination is dependent, then all four will be dependent. Similarly, if one combination is independent, then all four will be independent. This rule does not apply to any other situation.

## Union

Another event that is the combination of other events is the *union*.

**Union of Events *A* and *B***

The **union** of events *A* and *B* is the event that occurs when either *A* or *B* or both occur. It is denoted as

$$A \text{ or } B$$

### EXAMPLE 6.4

## Determinants of Success among Mutual Fund Managers—Part 4

Determine the probability that a randomly selected fund outperforms the market or the manager graduated from a top-20 MBA program.

### SOLUTION

We want to compute the probability of the union of two events

$$P(A_1 \text{ or } B_1)$$

The union $A_1$ or $B_1$ consists of three events; That is, the union occurs whenever any of the following joint events occurs:

1. Fund outperforms the market and the manager graduated from a top-20 MBA program
2. Fund outperforms the market and the manager did not graduate from a top-20 MBA program
3. Fund does not outperform the market and the manager graduated from a top-20 MBA program

Their probabilities are

$$P(A_1 \text{ and } B_1) = .11$$
$$P(A_2 \text{ and } B_1) = .06$$
$$P(A_1 \text{ and } B_2) = .29$$

Thus, the probability of the union—the fund outperforms the market or the manager graduated from a top-20 MBA program—is the sum of the three probabilities; That is,

$$P(A_1 \text{ or } B_1) = P(A_1 \text{ and } B_1) + P(A_2 \text{ and } B_1) + P(A_1 \text{ and } B_2) = .11 + .06 + .29 = .46$$

Notice that there is another way to produce this probability. Of the four probabilities in Table 6.1, the only one representing an event that is not part of the union is the probability of the event the fund does not outperform the market and the manager did not graduate from a top-20 MBA program. That probability is

$$P(A_2 \text{ and } B_2) = .54$$

which is the probability that the union *does not* occur. Thus, the probability of the union is

$$P(A_1 \text{ or } B_1) = 1 - P(A_2 \text{ and } B_2) = 1 - .54 = .46.$$

Thus, we determined that 46% of mutual funds either outperform the market or are managed by a top-20 MBA program graduate or have both characteristics.

## EXERCISES

**6.16** Given the following table of joint probabilities, calculate the marginal probabilities.

| | $A_1$ | $A_2$ | $A_3$ |
|---|---|---|---|
| $B_1$ | .1 | .3 | .2 |
| $B_2$ | .2 | .1 | .1 |

**6.17** Calculate the marginal probabilities from the following table of joint probabilities.

| | $A_1$ | $A_2$ |
|---|---|---|
| $B_1$ | .4 | .3 |
| $B_2$ | .2 | .1 |

**6.18** Refer to Exercise 6.17.

a. Determine $P(A_1|B_1)$.

b. Determine $P(A_2|B_1)$.

c. Did your answers to parts (a) and (b) sum to 1? Is this a coincidence? Explain.

**6.19** Refer to Exercise 6.17. Calculate the following probabilities.

a. $P(A_1|B_2)$

b. $P(B_2|A_1)$

c. Did you expect the answers to parts (a) and (b) to be reciprocals? In other words, did you expect that $P(A_1|B_2) = 1/P(B_2|A_1)$? Why is this impossible (unless both probabilities are 1)?

**6.20** Are the events in Exercise 6.17 independent? Explain.

**6.21** Refer to Exercise 6.17. Compute the following.

a. P($A_1$ or $B_1$)

b. P($A_1$ or $B_2$)

c. P($A_1$ or $A_2$)

**6.22** Suppose that you have been given the following joint probabilities. Are the events independent? Explain.

| | $A_1$ | $A_2$ |
|---|---|---|
| $B_1$ | .20 | .60 |
| $B_2$ | .05 | .15 |

**6.23** Determine whether the events are independent from the following joint probabilities.

| | $A_1$ | $A_2$ |
|---|---|---|
| $B_1$ | .20 | .15 |
| $B_2$ | .60 | .05 |

**6.24** Suppose we have the following joint probabilities.

| | $A_1$ | $A_2$ | $A_3$ |
|---|---|---|---|
| $B_1$ | .15 | .20 | .10 |
| $B_2$ | .25 | .25 | .05 |

Compute the marginal probabilities.

**6.25** Refer to Exercise 6.24.

a. Compute $P(A_2|B_2)$.

b. Compute $P(B_2|A_2)$.

c. Compute $P(B_1|A_2)$.

**6.26** Refer to Exercise 6.24.

a. Compute P($A_1$ or $A_2$).

b. Compute P($A_2$ or $B_2$).

c. Compute P($A_3$ or $B_1$).

**6.27** Discrimination in the workplace is illegal, and companies that discriminate are often sued. The female instructors at a large university recently lodged a complaint about the most recent round of promotions from assistant professor to associate professor. An analysis of the relationship between gender and promotion produced the following joint probabilities.

| | Promoted | Not Promoted |
|---|---|---|
| Female | .03 | .12 |
| Male | .17 | .68 |

a. What is the rate of promotion among female assistant professors?
b. What is the rate of promotion among male assistant professors?
c. Is it reasonable to accuse the university of gender bias?

**6.28** A department store analyzed its most recent sales and determined the relationship between the way the customer paid for the item and the price category of the item. The joint probabilities in the following table were calculated.

| | Cash | Credit Card | Debit Card |
|---|---|---|---|
| Less than $20 | .09 | .03 | .04 |
| $20–$100 | .05 | .21 | .18 |
| More than $100 | .03 | .23 | .14 |

a. What proportion of purchases was paid by debit card?
b. Find the probability that a credit card purchase was more than $100.
c. Determine the proportion of purchases made by credit card or by debit card.

**6.29** The following table lists the probabilities of unemployed females and males and their educational attainment.

| | Female | Male |
|---|---|---|
| Less than high school | .077 | .110 |
| High school graduate | .154 | .201 |
| Some college or university—no degree | .141 | .129 |
| College or university graduate | .092 | .096 |

(*Source: Statistical Abstract of the United States, 2009*, Table 607.)

a. If one unemployed person is selected at random, what is the probability that he or she did not finish high school?
b. If an unemployed female is selected at random, what is the probability that she has a college or university degree?
c. If an unemployed high school graduate is selected at random, what is the probability that he is a male?

**6.30** The costs of medical care in North America are increasing faster than inflation, and with the baby boom generation soon to need health care, it becomes imperative that countries find ways to reduce both costs and demand. The following table lists the joint probabilities associated with smoking and lung disease among 60- to 65-year-old men.

| | He is a smoker | He is a nonsmoker |
|---|---|---|
| He has lung disease | .12 | .03 |
| He does not have lung disease | .19 | .66 |

One 60- to 65-year-old man is selected at random. What is the probability of the following events?
a. He is a smoker.
b. He does not have lung disease.
c. He has lung disease given that he is a smoker.
d. He has lung disease given that he does not smoke.

**6.31** Refer to Exercise 6.30. Are smoking and lung disease among 60- to 65-year-old men related?

**6.32** The method of instruction in college and university applied statistics courses is changing. Historically, most courses were taught with an emphasis on manual calculation. The alternative is to employ a computer and a software package to perform the calculations. An analysis of applied statistics courses investigated whether the instructor's educational background is primarily mathematics (or statistics) or some other field. The result of this analysis is the accompanying table of joint probabilities.

| | Statistics Course Emphasizes Manual Calculations | Statistics Course Employs Computer and Software |
|---|---|---|
| Mathematics or statistics education | .23 | .36 |
| Other education | .11 | .30 |

a. What is the probability that a randomly selected applied statistics course instructor whose education was in statistics emphasizes manual calculations?
b. What proportion of applied statistics courses employ a computer and software?
c. Are the educational background of the instructor and the way his or her course is taught independent?

**6.33** A restaurant chain routinely surveys its customers. Among other questions, the survey asks each customer whether he or she would return and to rate the quality of food. Summarizing hundreds of thousands of questionnaires produced this table of joint probabilities.

| Rating | Customer Will Return | Customer Will Not Return |
|---|---|---|
| Poor | .02 | .10 |
| Fair | .08 | .09 |
| Good | .35 | .14 |
| Excellent | .20 | .02 |

a. What proportion of customers say that they will return and rate the restaurant's food as good?
b. What proportion of customers who say that they will return rate the restaurant's food as good?
c. What proportion of customers who rate the restaurant's food as good say that they will return?
d. Discuss the differences in your answers to parts (a), (b), and (c).

**6.34** To determine whether drinking alcoholic beverages has an effect on the bacteria that cause ulcers, researchers developed the following table of joint probabilities.

| Number of Alcoholic Drinks per Day | Ulcer | No Ulcer |
|---|---|---|
| None | .01 | .22 |
| One | .03 | .19 |
| Two | .03 | .32 |
| More than two | .04 | .16 |

a. What proportion of people have ulcers?
b. What is the probability that a teetotaler (no alcoholic beverages) develops an ulcer?
c. What is the probability that someone who has an ulcer does not drink alcohol?
d. What is the probability that someone who has an ulcer drinks alcohol?

**6.35** An analysis of fired or laid-off workers, their age, and the reasons for their departure produced the following table of joint probabilities.

| | Age Category | | | |
|---|---|---|---|---|
| Reason for job loss | 20–24 | 25–54 | 55–64 | 65 and older |
| Plant or company closed or moved | .015 | .320 | .089 | .029 |
| Insufficient work | .014 | .180 | .034 | .011 |
| Position or shift abolished | .006 | .214 | .071 | .016 |

(*Source: Statistical Abstract of the United States, 2009*, Table 593.)

a. What is the probability that a 25- to 54-year-old employee was laid off or fired because of insufficient work?
b. What proportion of laid-off or fired workers is age 65 and older?
c. What is the probability that a laid-off or fired worker because the plant or company closed is 65 or older?

**6.36** Many critics of television claim that there is too much violence and that it has a negative effect on society. There may also be a negative effect on advertisers. To examine this issue, researchers developed two versions of a cops-and-robbers made-for-television movie. One version depicted several violent crimes, and the other removed these scenes. In the middle of the movie, one 60-second commercial was shown advertising a new product and brand name. At the end of the movie, viewers were asked to name the brand. After observing the results, the researchers produced the following table of joint probabilities.

| | Watch Violent Movie | Watch Nonviolent Movie |
|---|---|---|
| Remember the brand name | .15 | .18 |
| Do not remember the brand name | .35 | .32 |

a. What proportion of viewers remember the brand name?
b. What proportion of viewers who watch the violent movie remember the brand name?
c. Does watching a violent movie affect whether the viewer will remember the brand name? Explain.

**6.37** Is there a relationship between the male hormone testosterone and criminal behavior? To answer this question, medical researchers measured the testosterone level of penitentiary inmates and recorded whether they were convicted of murder. After analyzing the results, the researchers produced the following table of joint probabilities.

| Testosterone Level | Murderer | Other Felon |
|---|---|---|
| Above average | .27 | .24 |
| Below average | .21 | .28 |

a. What proportion of murderers have above-average testosterone levels?
b. Are levels of testosterone and the crime committed independent? Explain.

**6.38** The issue of health care coverage in the United States is becoming a critical issue in American politics. A large-scale study was undertaken to determine who is and is not covered. From this study, the following table of joint probabilities was produced.

| Age Category | Has Health Insurance | Does Not Have Health Insurance |
|---|---|---|
| 25–34 | .167 | .085 |
| 35–44 | .209 | .061 |
| 45–54 | .225 | .049 |
| 55–64 | .177 | .026 |

(*Source*: U.S. Department of Health and Human Services.)

If one person is selected at random, find the following probabilities.
a. *P*(Person has health insurance)
b. *P*(Person 55–64 has no health insurance)
c. *P*(Person without health insurance is between 25 and 34 years old)

**6.39** Violent crime in many American schools is an unfortunate fact of life. An analysis of schools and violent crime yielded the table of joint probabilities shown next.

| Level | Violent Crime Committed This Year | No Violent Crime Committed |
|---|---|---|
| Primary | .393 | .191 |
| Middle | .176 | .010 |
| High School | .134 | .007 |
| Combined | .074 | .015 |

(*Source: Statistical Abstract of the United States, 2009*, Table 237.)

If one school is randomly selected find the following probabilities.

a. Probability of at least one incident of violent crime during the year in a primary school
b. Probability of no violent crime during the year

**6.40** Refer to Exercise 6.39. A similar analysis produced these joint probabilities.

| Enrollment | Violent Crime Committed This Year | No Violent Crime Committed |
|---|---|---|
| Less than 300 | .159 | .091 |
| 300 to 499 | .221 | .065 |
| 500 to 999 | .289 | .063 |
| 1,000 or more | .108 | .004 |

(*Source: Statistical Abstract of the United States, 2009*, Table 237.)

a. What is the probability that a school with an enrollment of less than 300 had at least one violent crime during the year?
b. What is the probability that a school that has at least one violent crime had an enrollment of less than 300?

**6.41** A firm has classified its customers in two ways: (1) according to whether the account is overdue and (2) whether the account is new (less than 12 months) or old. An analysis of the firm's records provided the input for the following table of joint probabilities.

| | Overdue | Not Overdue |
|---|---|---|
| New | .06 | .13 |
| Old | .52 | .29 |

One account is randomly selected.

a. If the account is overdue, what is the probability that it is new?
b. If the account is new, what is the probability that it is overdue?
c. Is the age of the account related to whether it is overdue? Explain.

**6.42** How are the size of a firm (measured in terms of the number of employees) and the type of firm related? To help answer the question, an analyst referred to the U.S. Census and developed the following table of joint probabilities.

| | Industry | | |
|---|---|---|---|
| Number of Employees | Construction | Manufacturing | Retail |
| Fewer than 20 | .464 | .147 | .237 |
| 20 to 99 | .039 | .049 | .035 |
| 100 or more | .005 | .019 | .005 |

(*Source: Statistical Abstract of the United States, 2009*, Table 737.)

If one firm is selected at random, find the probability of the following events.

a. The firm employs fewer than 20 employees.
b. The firm is in the retail industry.
c. A firm in the construction industry employs between 20 and 99 workers.

**6.43** Credit scorecards are used by financial institutions to help decide to whom loans should be granted (see the Applications in Banking: Credit Scorecards summary on page 63). An analysis of the records of one bank produced the following probabilities.

| | Score | |
|---|---|---|
| Loan Performance | Under 400 | 400 or More |
| Fully repaid | .19 | .64 |
| Defaulted | .13 | .04 |

a. What proportion of loans are fully repaid?
b. What proportion of loans given to scorers of less than 400 fully repay?
c. What proportion of loans given to scorers of 400 or more fully repay?
d. Are score and whether the loan is fully repaid independent? Explain.

**6.44** A retail outlet wanted to know whether its weekly advertisement in the daily newspaper works. To acquire this critical information, the store manager surveyed the people who entered the store and determined whether each individual saw the ad and whether a purchase was made. From the information developed, the manager produced the following table of joint probabilities. Are the ads effective? Explain.

| | Purchase | No Purchase |
|---|---|---|
| See ad | .18 | .42 |
| Do not see ad | .12 | .28 |

**6.45** To gauge the relationship between education and unemployment, an economist turned to the U.S. Census from which the following table of joint probabilities was produced.

| Education | Employed | Unemployed |
|---|---|---|
| Not a high school graduate | .091 | .008 |
| High school graduate | .282 | .014 |

(Continued)

| | | |
|---|---|---|
| Some college, no degree | .166 | .007 |
| Associate's degree | .095 | .003 |
| Bachelor's degree | .213 | .004 |
| Advanced degree | .115 | .002 |

(*Source: Statistical Abstract of the United States, 2009*, Table 223.)

a. What is the probability that a high school graduate is unemployed?
b. Determine the probability that a randomly selected individual is employed.
c. Find the probability that an unemployed person possesses an advanced degree.
d. What is the probability that a randomly selected person did not finish high school?

**6.46** The decision about where to build a new plant is a major one for most companies. One factor that is often considered is the education level of the location's residents. Census information may be useful in this regard. After analyzing a recent census, a company produced the following joint probabilities.

| | Region | | | |
|---|---|---|---|---|
| **Education** | **Northeast** | **Midwest** | **South** | **West** |
| Not a high school graduate | .024 | .024 | .059 | .036 |
| High school graduate | .063 | .078 | .117 | .059 |
| Some college, no degree | .023 | .039 | .061 | .045 |
| Associate's degree | .015 | .021 | .030 | .020 |
| Bachelor's degree | .038 | .040 | .065 | .046 |
| Advanced degree | .024 | .020 | .032 | .023 |

(*Source: Statistical Abstract of the United States, 2009*, Table 223.)

a. Determine the probability that a person living in the West has a bachelor's degree.
b. Find the probability that a high school graduate lives in the Northeast.
c. What is the probability that a person selected at random lives in the South?
d. What is the probability that a person selected at random does not live in the South?

## 6.3 Probability Rules and Trees

In Section 6.2, we introduced intersection and union and described how to determine the probability of the intersection and the union of two events. In this section, we present other methods of determining these probabilities. We introduce three rules that enable us to calculate the probability of more complex events from the probability of simpler events.

### Complement Rule

The **complement** of event $A$ is the event that occurs when event $A$ does not occur. The complement of event $A$ is denoted by $A^C$. The **complement rule** defined here derives from the fact that the probability of an event and the probability of the event's complement must sum to 1.

**Complement Rule**

$$P(A^C) = 1 - P(A)$$

for any event $A$.

We will demonstrate the use of this rule after we introduce the next rule.

### Multiplication Rule

The **multiplication rule** is used to calculate the joint probability of two events. It is based on the formula for conditional probability supplied in the previous section; that is, from the following formula

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

we derive the multiplication rule simply by multiplying both sides by $P(B)$.

**Multiplication Rule**

The joint probability of any two events $A$ and $B$ is

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

or, altering the notation,

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

If $A$ and $B$ are independent events, $P(A|B) = P(A)$ and $P(B|A) = P(B)$. It follows that the joint probability of two independent events is simply the product of the probabilities of the two events. We can express this as a special form of the multiplication rule.

**Multiplication Rule for Independent Events**

The joint probability of any two independent events $A$ and $B$ is

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

## EXAMPLE 6.5* Selecting Two Students without Replacement

A graduate statistics course has seven male and three female students. The professor wants to select two students at random to help her conduct a research project. What is the probability that the two students chosen are female?

### SOLUTION

Let $A$ represent the event that the first student chosen is female and $B$ represent the event that the second student chosen is also female. We want the joint probability $P(A \text{ and } B)$. Consequently, we apply the multiplication rule:

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

Because there are 3 female students in a class of 10, the probability that the first student chosen is female is

$$P(A) = 3/10$$

*This example can be solved using the Hypergeometric distribution, which is described in the Keller's website Appendix Hypergeometric Distribution.

After the first student is chosen, there are only nine students left. Given that the first student chosen was female, there are only two female students left. It follows that

$$P(B|A) = 2/9$$

Thus, the joint probability is

$$P(A \text{ and } B) = P(A)P(B|A) = \left(\frac{3}{10}\right)\left(\frac{2}{9}\right) = \frac{6}{90} = .067$$

## EXAMPLE 6.6 Selecting Two Students with Replacement

Refer to Example 6.5. The professor who teaches the course is suffering from the flu and will be unavailable for two classes. The professor's replacement will teach the next two classes. His style is to select one student at random and pick on him or her to answer questions during that class. What is the probability that the two students chosen are female?

### SOLUTION

The form of the question is the same as in Example 6.5: We wish to compute the probability of choosing two female students. However, the experiment is slightly different. It is now possible to choose the *same* student in each of the two classes taught by the replacement. Thus, $A$ and $B$ are independent events, and we apply the multiplication rule for independent events:

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

The probability of choosing a female student in each of the two classes is the same; that is,

$$P(A) = 3/10 \text{ and } P(B) = 3/10$$

Hence,

$$P(A \text{ and } B) = P(A)P(B) = \left(\frac{3}{10}\right)\left(\frac{3}{10}\right) = \frac{9}{100} = .09$$

## Addition Rule

The **addition rule** enables us to calculate the probability of the union of two events.

**Addition Rule**

The probability that event $A$, or event $B$, or both occur is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

If you're like most students, you're wondering why we subtract the joint probability from the sum of the probabilities of $A$ and $B$. To understand why this is necessary, examine Table 6.2 (page 183 ), which we have reproduced here as Table 6.3.

TABLE **6.3** Joint and Marginal Probabilities

| | $B_1$ | $B_2$ | TOTALS |
|---|---|---|---|
| $A_1$ | $P(A_1 \text{ and } B_1) = .11$ | $P(A_1 \text{ and } B_2) = .29$ | $P(A_1) = .40$ |
| $A_2$ | $P(A_2 \text{ and } B_1) = .06$ | $P(A_2 \text{ and } B_2) = .54$ | $P(A_2) = .60$ |
| Totals | $P(B_1) = .17$ | $P(B_2) = .83$ | 1.00 |

This table summarizes how the marginal probabilities were computed. For example, the marginal probability of $A_1$ and the marginal probability of $B_1$ were calculated as

$$P(A_1) = P(A_1 \text{ and } B_1) + P(A_1 \text{ and } B_2) = .11 + .29 = .40$$

$$P(B_1) = P(A_1 \text{ and } B_1) + P(A_2 \text{ and } B_1) = .11 + .06 = .17$$

If we now attempt to calculate the probability of the union of $A_1$ and $B_1$ by summing their probabilities, we find

$$P(A_1) + P(B_1) = .11 + .29 + .11 + .06$$

Notice that we added the joint probability of $A_1$ and $B_1$ (which is .11) twice. To correct the double counting, we subtract the joint probability from the sum of the probabilities of $A_1$and $B_1$. Thus,

$$\begin{aligned} P(A_1 \text{ or } B_1) &= P(A_1) + P(B_1) - P(A_1 \text{ and } B_1) \\ &= [.11 + .29] + [.11 + .06] - .11 \\ &= .40 + .17 - .11 = .46 \end{aligned}$$

This is the probability of the union of $A_1$ and $B_1$, which we calculated in Example 6.4 (page 186).

As was the case with the multiplication rule, there is a special form of the addition rule. When two events are mutually exclusive (which means that the two events cannot occur together), their joint probability is 0.

**Addition Rule for Mutually Exclusive Events**

The probability of the union of two mutually exclusive events $A$ and $B$ is

$$P(A \text{ or } B) = P(A) + P(B)$$

EXAMPLE 6.7

## Applying the Addition Rule

In a large city, two newspapers are published, the *Sun* and the *Post*. The circulation departments report that 22% of the city's households have a subscription to the *Sun* and 35% subscribe to the *Post*. A survey reveals that 6% of all households subscribe to both newspapers. What proportion of the city's households subscribe to either newspaper?

SOLUTION

We can express this question as, what is the probability of selecting a household at random that subscribes to the *Sun*, the *Post*, or both? Another way of asking the question is, what is the probability that a randomly selected household subscribes to *at least one* of the newspapers? It is now clear that we seek the probability of the union, and we must apply the addition rule. Let $A$ = the household subscribes to the *Sun* and $B$ = the household subscribes to the *Pos*t. We perform the following calculation:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = .22 + .35 - .06 = .51$$

The probability that a randomly selected household subscribes to either newspaper is .51. Expressed as relative frequency, 51% of the city's households subscribe to either newspaper.

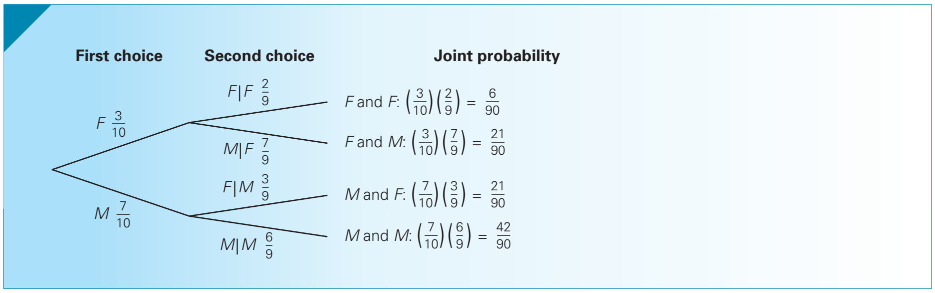
## Probability Trees

An effective and simpler method of applying the probability rules is the probability tree, wherein the events in an experiment are represented by lines. The resulting figure resembles a tree, hence the name. We will illustrate the probability tree with several examples, including two that we addressed using the probability rules alone.

In Example 6.5, we wanted to find the probability of choosing two female students, where the two choices had to be different. The tree diagram in Figure 6.1 describes this experiment. Notice that the first two branches represent the two possibilities, female and male students, on the first choice. The second set of branches represents the two possibilities on the second choice. The probabilities of female and male student chosen first are 3/10 and 7/10, respectively. The probabilities for the second set of branches are conditional probabilities based on the choice of the first student selected.

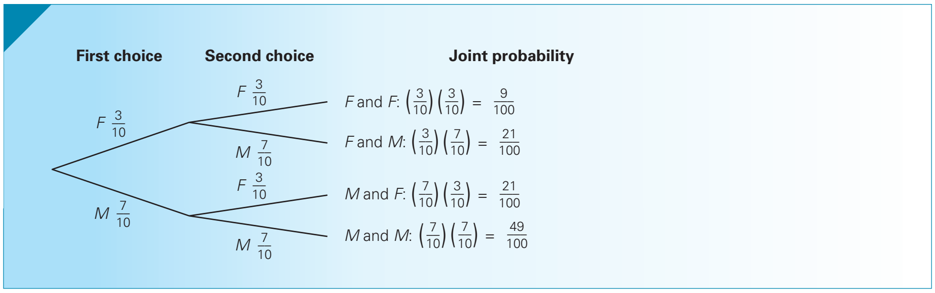
We calculate the joint probabilities by multiplying the probabilities on the linked branches. Thus, the probability of choosing two female students is $P(F \text{ and } F) = (3/10)(2/9) = 6/90$. The remaining joint probabilities are computed similarly.

FIGURE **6.1** Probability Tree for Example 6.5


First choice
Second choice
Joint probability
$F|F$ $\frac{2}{9}$
$F$ $\frac{3}{10}$
$M|F$ $\frac{7}{9}$
$F|M$ $\frac{3}{9}$
$M$ $\frac{7}{10}$
$M|M$ $\frac{6}{9}$
$F$ and $F$: $\left(\frac{3}{10}\right)\left(\frac{2}{9}\right) = \frac{6}{90}$
$F$ and $M$: $\left(\frac{3}{10}\right)\left(\frac{7}{9}\right) = \frac{21}{90}$
$M$ and $F$: $\left(\frac{7}{10}\right)\left(\frac{3}{9}\right) = \frac{21}{90}$
$M$ and $M$: $\left(\frac{7}{10}\right)\left(\frac{6}{9}\right) = \frac{42}{90}$

In Example 6.6, the experiment was similar to that of Example 6.5. However, the student selected on the first choice was returned to the pool of students and was eligible to be chosen again. Thus, the probabilities on the second set of branches remain the same as the probabilities on the first set, and the probability tree is drawn with these changes, as shown in Figure 6.2.

FIGURE 6.2 Probability Tree for Example 6.6


First choice
Second choice
Joint probability
F 3/10
M 7/10
F 3/10
M 7/10
F 3/10
M 7/10
F and F: (3/10)(3/10) = 9/100
F and M: (3/10)(7/10) = 21/100
M and F: (7/10)(3/10) = 21/100
M and M: (7/10)(7/10) = 49/100

The advantage of a probability tree on this type of problem is that it restrains its users from making the wrong calculation. Once the tree is drawn and the probabilities of the branches inserted, virtually the only allowable calculation is the multiplication of the probabilities of linked branches. An easy check on those calculations is available. The joint probabilities at the ends of the branches must sum to 1 because all possible events are listed. In both figures, notice that the joint probabilities do indeed sum to 1.

The special form of the addition rule for mutually exclusive events can be applied to the joint probabilities. In both probability trees, we can compute the probability that one student chosen is female and one is male simply by adding the joint probabilities. For the tree in Example 6.5, we have

$$P(F \text{ and } M) + P(M \text{ and } F) = 21/90 + 21/90 = 42/90$$

In the probability tree in Example 6.6, we find

$$P(F \text{ and } M) + P(M \text{ and } F) = 21/100 + 21/100 = 42/100$$

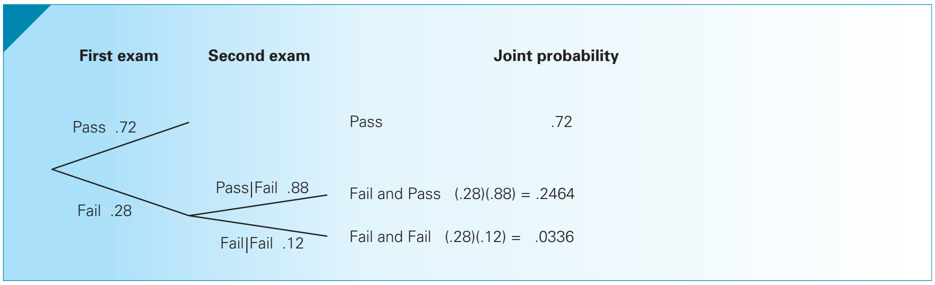
## EXAMPLE 6.8 Probability of Passing the Bar Exam

Students who graduate from law schools must still pass a bar exam before becoming lawyers. Suppose that in a particular jurisdiction the pass rate for first-time test takers is 72%. Candidates who fail the first exam may take it again several months later. Of those who fail their first test, 88% pass their second attempt. Find the probability that a randomly selected law school graduate becomes a lawyer. Assume that candidates cannot take the exam more than twice.

### SOLUTION

The probability tree in Figure 6.3 is employed to describe the experiment. Note that we use the complement rule to determine the probability of failing each exam.

FIGURE 6.3 Probability Tree for Example 6.8



We apply the multiplication rule to calculate P(Fail and Pass), which we find to be .2464. We then apply the addition rule for mutually exclusive events to find the probability of passing the first or second exam:

$$P(\text{Pass [on first exam]}) + P(\text{Fail [on first exam] and Pass [on second exam]})$$
$$= .72 + .2464 = .9664$$

Thus, 96.64% of applicants become lawyers by passing the first or second exam.

## EXERCISES

**6.47** Given the following probabilities, compute all joint probabilities.

$P(A) = .9$ $\quad P(A^C) = .1$

$P(B|A) = .4$ $\quad P(B|A^C) = .7$

**6.48** Determine all joint probabilities from the following.

$P(A) = .8$ $\quad P(A^C) = .2$

$P(B|A) = .4$ $\quad P(B|A^C) = .7$

**6.49** Draw a probability tree to compute the joint probabilities from the following probabilities.

$P(A) = .5$ $\quad P(A^C) = .2$

$P(B|A) = .4$ $\quad P(B|A^C) = .7$

**6.50** Given the following probabilities, draw a probability tree to compute the joint probabilities.

$P(A) = .8$ $\quad P(A^C) = .2$

$P(B|A) = .3$ $\quad P(B|A^C) = .3$

**6.51** Given the following probabilities, find the joint probability *P*(*A* and *B*).

$P(A) = .7$ $\quad P(B|A) = .3$

**6.52** Approximately 10% of people are left-handed. If two people are selected at random, what is the probability of the following events?

a. Both are right-handed.
b. Both are left-handed.
c. One is right-handed and the other is left-handed.
d. At least one is right-handed.

**6.53** Refer to Exercise 6.52. Suppose that three people are selected at random.

a. Draw a probability tree to depict the experiment.
b. If we use the notation RRR to describe the selection of three right-handed people, what are the descriptions of the remaining seven events? (Use L for left-hander.)
c. How many of the events yield no right-handers, one right-hander, two right-handers, three right-handers?
d. Find the probability of no right-handers, one right-hander, two right-handers, three right-handers.

**6.54** Suppose there are 100 students in your accounting class, 10 of whom are left-handed. Two students are selected at random.

a. Draw a probability tree and insert the probabilities for each branch.

What is the probability of the following events?

b. Both are right-handed.
c. Both are left-handed.
d. One is right-handed and the other is left-handed.
e. At least one is right-handed

**6.55** Refer to Exercise 6.54. Suppose that three people are selected at random.

a. Draw a probability tree and insert the probabilities of each branch.
b. What is the probability of no right-handers, one right-hander, two right-handers, three right-handers?

**6.56** An aerospace company has submitted bids on two separate federal government defense contracts. The company president believes that there is a 40% probability of winning the first contract. If they win the first contract, the probability of winning the second is 70%. However, if they lose the first contract, the president thinks that the probability of winning the second contract decreases to 50%.

a. What is the probability that they win both contracts?
b. What is the probability that they lose both contracts?
c. What is the probability that they win only one contract?

**6.57** A telemarketer calls people and tries to sell them a subscription to a daily newspaper. On 20% of her calls, there is no answer or the line is busy. She sells subscriptions to 5% of the remaining calls. For what proportion of calls does she make a sale?

**6.58** A foreman for an injection-molding firm admits that on 10% of his shifts, he forgets to shut off the injection machine on his line. This causes the machine to overheat, increasing the probability from 2% to 20% that a defective molding will be produced during the early morning run. What proportion of moldings from the early morning run is defective?

**6.59** A study undertaken by the Miami-Dade Supervisor of Elections in 2002 revealed that 44% of registered voters are Democrats, 37% are Republicans, and 19% are others. If two registered voters are selected at random, what is the probability that both of them have the same party affiliation? (*Source: Miami Herald*, April 11, 2002.)

**6.60** In early 2001, the U.S. Census Bureau started releasing the results of the 2000 census. Among many other pieces of information, the bureau recorded the race or ethnicity of the residents of every county in every state. From these results, the bureau calculated a "diversity index" that measures the probability that two people chosen at random are of different races or ethnicities. Suppose that the census determined that in a county in Wisconsin 80% of its residents are white, 15% are black, and 5% are Asian. Calculate the diversity index for this county.

**6.61** A survey of middle-aged men reveals that 28% of them are balding at the crown of their heads. Moreover, it is known that such men have an 18% probability of suffering a heart attack in the next 10 years. Men who are not balding in this way have an 11% probability of a heart attack. Find the probability that a middle-aged man will suffer a heart attack sometime in the next 10 years.

**6.62** The chartered financial analyst (CFA) is a designation earned after a candidate has taken three annual exams (CFA I, II, and III). The exams are taken in early June. Candidates who pass an exam are eligible to take the exam for the next level in the following year. The pass rates for levels I, II, and III are .57, .73, and .85, respectively. Suppose that 3,000 candidates take the level I exam, 2,500 take the level II exam, and 2,000 take the level III exam. Suppose that one student is selected at random. What is the probability that he or she has passed the exam? (*Source*: Institute of Financial Analysts.)

**6.63** The Nickels restaurant chain regularly conducts surveys of its customers. Respondents are asked to assess food quality, service, and price. The responses are

Excellent    Good    Fair

Surveyed customers are also asked whether they would come back. After analyzing the responses, an expert in probability determined that 87% of customers say that they will return. Of those who so indicate, 57% rate the restaurant as excellent, 36% rate it as good, and the remainder rate it as fair. Of those who say that they won't return, the probabilities are 14%, 32%, and 54%, respectively. What proportion of customers rate the restaurant as good?

**6.64** Researchers at the University of Pennsylvania School of Medicine have determined that children under 2 years old who sleep with the lights on have a 36% chance of becoming myopic before they are 16. Children who sleep in darkness have a 21% probability of becoming myopic. A survey indicates that 28% of children under 2 sleep with some light on. Find the probability that a child under 16 is myopic.

**6.65** All printed circuit boards (PCBs) that are manufactured at a certain plant are inspected. An analysis of the company's records indicates that 22% of all PCBs are flawed in some way. Of those that are flawed, 84% are reparable and the rest must be discarded. If a newly produced PCB is randomly selected, what is the probability that it does not have to be discarded?

**6.66** A financial analyst has determined that there is a 22% probability that a mutual fund will outperform the market over a 1-year period provided that it outperformed the market the previous year. If only 15% of mutual funds outperform the market during any year, what is the probability that a mutual fund will outperform the market 2 years in a row?

**6.67** An investor believes that on a day when the Dow Jones Industrial Average (DJIA) increases, the probability that the NASDAQ also increases is 77%. If the investor believes that there is a 60% probability that the DJIA will increase tomorrow, what is the probability that the NASDAQ will increase as well?

**6.68** The controls of an airplane have several backup systems or redundancies so that if one fails the plane will continue to operate. Suppose that the mechanism that controls the flaps has two backups. If the probability that the main control fails is .0001 and the probability that each backup will fail is .01, what is the probability that all three fail to operate?

**6.69** According to TNS Intersearch, 69% of wireless web users use it primarily for receiving and sending e-mail. Suppose that three wireless web users are selected at random. What is the probability that all of them use it primarily for e-mail?

**6.70** A financial analyst estimates that the probability that the economy will experience a recession in the next 12 months is 25%. She also believes that if the economy encounters a recession, the probability that her mutual fund will increase in value is 20%. If there is no recession, the probability that the mutual fund will increase in value is 75%. Find the probability that the mutual fund's value will increase.

## 6.4 Bayes's Law

Conditional probability is often used to gauge the relationship between two events. In many of the examples and exercises you've already encountered, conditional probability measures the probability that an event occurs given that a possible cause of the event has occurred. In Example 6.2, we calculated the probability that a mutual fund outperforms the market (the effect) given that the fund manager graduated from a top-20 MBA program (the possible cause). There are situations, however, where we witness a particular event and we need to compute the probability of one of its possible causes. **Bayes's Law** is the technique we use.

### EXAMPLE 6.9 Should an MBA Applicant Take a Preparatory Course?

The Graduate Management Admission Test (GMAT) is a requirement for all applicants of MBA programs. A variety of preparatory courses are designed to help applicants improve their GMAT scores, which range from 200 to 800. Suppose that a survey of MBA students reveals that among GMAT scorers above 650, 52% took a preparatory course; whereas among GMAT scorers of less than 650 only 23% took a preparatory course. An applicant to an MBA program has determined that he needs a score of more than 650 to get into a certain MBA program, but he feels that his probability of getting that high a score is quite low—10%. He is considering taking a preparatory course that costs $500. He is willing to do so only if his probability of achieving 650 or more doubles. What should he do?

#### SOLUTION

The easiest way to address this problem is to draw a tree diagram. The following notation will be used:

$A$ = GMAT score is 650 or more
$A^C$ = GMAT score less than 650
$B$ = Took preparatory course
$B^C$ = Did not take preparatory course

The probability of scoring 650 or more is

$$P(A) = .10$$

The complement rule gives us

$$P(A^C) = 1 - .10 = .90$$

Conditional probabilities are

$$P(B|A) = .52$$

and

$$P(B|A^C) = .23$$

Again using the complement rule, we find the following conditional probabilities:

$$P(B^C|A) = 1 - .52 = .48$$

and

$$P(B^C|A^C) = 1 - .23 = .77$$

We would like to determine the probability that he would achieve a GMAT score of 650 or more given that he took the preparatory course; that is, we need to compute

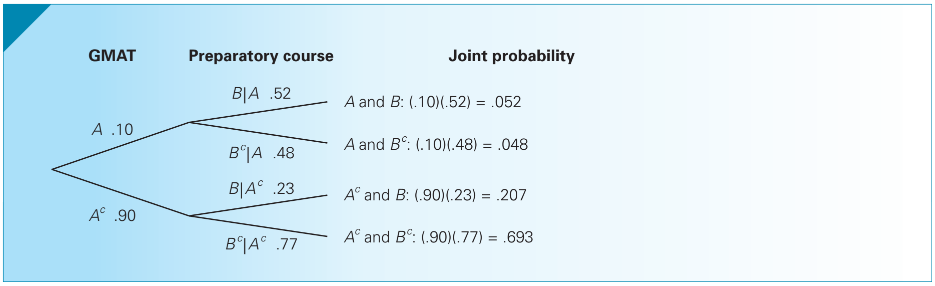
$$P(A|B)$$

Using the definition of conditional probability (page 184), we have

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

Neither the numerator nor the denominator is known. The probability tree (Figure 6.4) will provide us with the probabilities.

FIGURE **6.4** Probability Tree for Example 6.9


GMAT
Preparatory course
Joint probability
A .10
$A^c$ .90
B|A .52
$B^c$|A .48
B|$A^c$ .23
$B^c$|$A^c$ .77
A and B: (.10)(.52) = .052
A and $B^c$: (.10)(.48) = .048
$A^c$ and B: (.90)(.23) = .207
$A^c$ and $B^c$: (.90)(.77) = .693

As you can see,

$$P(A \text{ and } B) = (.10)(.52) = .052$$
$$P(A^C \text{ and } B) = (.90)(.23) = .207$$

and

$$P(B) = P(A \text{ and } B) + P(A^C \text{ and } B) = .052 + .207 = .259$$

Thus,

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{.052}{.259} = .201$$

The probability of scoring 650 or more on the GMAT doubles when the preparatory course is taken.

Thomas Bayes first employed the calculation of conditional probability as shown in Example 6.9 during the 18th century. Accordingly, it is called Bayes's Law.

The probabilities $P(A)$ and $P(A^C)$ are called **prior probabilities** because they are determined *prior* to the decision about taking the preparatory course. The conditional probabilities are called **likelihood probabilities** for reasons that are beyond the mathematics in this book. Finally, the conditional probability $P(A|B)$ and similar conditional probabilities $P(A^C|B)$, $P(A|B^C)$, and $P(A^C|B^C)$ are called **posterior probabilities** or **revised probabilities** because the prior probabilities are revised *after* the decision about taking the preparatory course.

You may be wondering why we did not get $P(A|B)$ directly. In other words, why not survey people who took the preparatory course and ask whether they received a score of 650 or more? The answer is that using the likelihood probabilities and using Bayes's Law allows individuals to set their own prior probabilities, which can then be revised. For example, another MBA applicant may assess her probability of scoring 650 or more as .40. Inputting the new prior probabilities produces the following probabilities:

$$P(A \text{ and } B) = (.40)(.52) = .208$$
$$P(A^C \text{ and } B) = (.60)(.23) = .138$$
$$P(B) = P(A \text{ and } B) + P(A^C\text{and } B) = .208 + .138 = .346$$
$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{.208}{.346} = .601$$

The probability of achieving a GMAT score of 650 or more increases by a more modest 50% (from .40 to .601).

## Bayes's Law Formula (Optional)

Bayes's Law can be expressed as a formula for those who prefer an algebraic approach rather than a probability tree. We use the following notation.

The event $B$ is the given event and the events

$$A_1, A_2, \ldots, A_k$$

are the events for which prior probabilities are known; that is,

$$P(A_1), P(A_2), \ldots, P(A_k)$$

are the prior probabilities.

The likelihood probabilities are

$$P(B|A_1), P(B|A_2), \ldots, P(B|A_k)$$

and

$$P(A_1|B), P(A_2|B), \ldots, P(A_k|B)$$

are the posterior probabilities, which represent the probabilities we seek.

**Bayes's Law Formula**

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_k)P(B|A_k)}$$

To illustrate the use of the formula, we'll redo Example 6.9. We begin by defining the events.

$A_1$ = GMAT score is 650 or more
$A_2$ = GMAT score less than 650
$B$ = Take preparatory course

The probabilities are

$$P(A_1) = .10$$

The complement rule gives us

$$P(A_2) = 1 - .10 = .90$$

Conditional probabilities are

$$P(B|A_1) = .52$$

and

$$P(B|A_2) = .23$$

Substituting the prior and likelihood probabilities into the Bayes's Law formula yields the following:

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)} = \frac{(.10)(.52)}{(.10)(.52) + (.90)(.23)}$$

$$= \frac{.052}{.052 + .207} = \frac{.052}{.259} = .201$$

As you can see, the calculation of the Bayes's Law formula produces the same results as the probability tree.

## Auditing Tax Returns: Solution

© Gary Buss/Taxi/Getty Images

We need to revise the prior probability that this return contains significant fraud. The tree shown in Figure 6.5 details the calculation.

$F$ = Tax return is fraudulent

$F^C$ = Tax return is honest

$E_0$ = Tax return contains no expense deductions

$E_1$ = Tax return contains one expense deduction

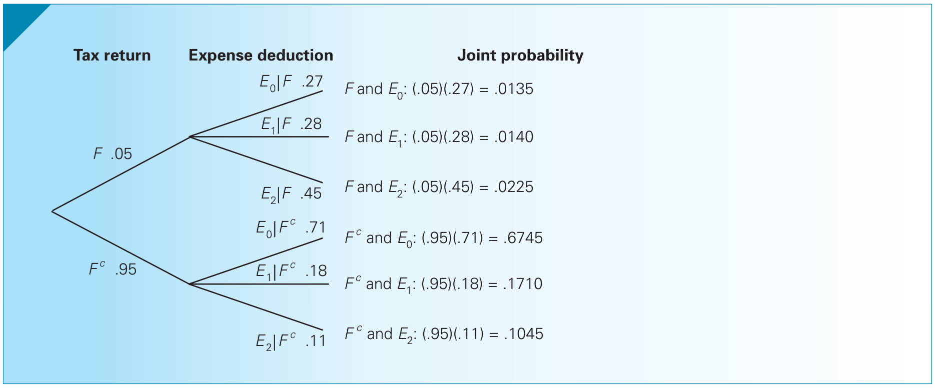
$E_2$ = tax return contains two expense deductions

$$P(E_1) = P(F \text{ and } E_1) + P(F^C \text{ and } E_1) = .0140 + .1710 = .1850$$

$$P(F|E_1) = P(F \text{ and } E_1)/P(E_1) = .0140/.1850 = .0757$$

The probability that this return is fraudulent is .0757.

FIGURE **6.5** Probability Tree for Auditing Tax Returns


Tax return
Expense deduction
Joint probability
F .05
F^c .95
E_0|F .27
E_1|F .28
E_2|F .45
E_0|F^c .71
E_1|F^c .18
E_2|F^c .11
F and E_0: (.05)(.27) = .0135
F and E_1: (.05)(.28) = .0140
F and E_2: (.05)(.45) = .0225
F^c and E_0: (.95)(.71) = .6745
F^c and E_1: (.95)(.18) = .1710
F^c and E_2: (.95)(.11) = .1045

## Applications in Medicine and Medical Insurance (Optional)

Physicians routinely perform medical tests, called *screenings*, on their patients. Screening tests are conducted for all patients in a particular age and gender group, regardless of their symptoms. For example, men in their 50s are advised to take a prostate-specific antigen (PSA) test to determine whether there is evidence of prostate cancer. Women undergo a Pap test for cervical cancer. Unfortunately, few of these tests are 100% accurate. Most can produce *false-positive* and *false-negative* results. A **false-positive** result is one in which the patient does not have the disease, but the test shows positive. A **false-negative** result is one in which the patient does have the disease, but the test produces a negative result. The consequences of each test are serious and costly. A false-negative test results in not detecting a disease in a patient, therefore postponing treatment, perhaps indefinitely. A false-positive test leads to apprehension and fear for the patient. In most cases, the patient is required to undergo further testing such as a biopsy. The unnecessary follow-up procedure can pose medical risks.

False-positive test results have financial repercussions. The cost of the follow-up procedure, for example, is usually far more expensive than the screening test. Medical insurance companies as well as government-funded plans are all adversely affected by false-positive test results. Compounding the problem is that physicians and patients are incapable of properly interpreting the results. A correct analysis can save both lives and money.

Bayes's Law is the vehicle we use to determine the true probabilities associated with screening tests. Applying the complement rule to the false-positive and false-negative rates produces the conditional probabilities that represent correct conclusions. Prior probabilities are usually derived by looking at the overall proportion of people with the diseases. In some cases, the prior probabilities may themselves have been revised because of heredity or demographic variables such as age or race. Bayes's Law allows us to revise the prior probability after the test result is positive or negative.

Example 6.10 is based on the actual false-positive and false-negative rates. Note however, that different sources provide somewhat different probabilities. The differences may be the result of the way positive and negative results are defined or the way technicians conduct the tests. Students who are affected by the diseases described in the example and exercises should seek clarification from their physicians.

EXAMPLE 6.10

### Probability of Prostate Cancer

Prostate cancer is the most common form of cancer found in men. The probability of developing prostate cancer over a lifetime is 16%. (This figure may be higher since many prostate cancers go undetected.) Many physicians routinely perform a PSA test, particularly for men over age 50. PSA is a protein produced only by the prostate gland and thus is fairly easy to detect. Normally, men have PSA levels between 0 and 4 mg/ml. Readings above 4 may be considered high and potentially indicative of cancer. However, PSA levels tend to rise with age even among men who are cancer free. Studies have shown that the test is not very accurate. In fact, the probability of having an elevated PSA level given that the man does not have cancer (false positive) is .135. If the man does have cancer, the probability of a normal PSA level (false negative) is almost .300. (This figure may vary by age and by the definition of *high* PSA level.) If a physician concludes that the PSA is high, a biopsy is performed. Besides the concerns and health needs of the men, there are also financial costs. The cost of the blood test is low (approximately $50). However, the cost of the biopsy is considerably higher (approximately $1,000). A false-positive PSA test will lead to an unnecessary biopsy. Because the PSA test is so inaccurate, some private and public medical plans do not pay for it. Suppose you are a manager in a medical insurance company and must decide on guidelines for whom should be routinely screened for prostate cancer. An analysis of prostate cancer incidence and age produces the following table of probabilities. (The probability of a man under 40 developing prostate cancer is less than .0001, or small enough to treat as 0.)

| Age | Probability of Developing Prostate Cancer |
|---|---|
| 40–49 | .010 |
| 50–59 | .022 |
| 60–69 | .046 |
| 70 and older | .079 |

Assume that a man in each of the age categories undergoes a PSA test with a positive result. Calculate the probability that each man actually has prostate cancer and the probability that he does not. Perform a cost–benefit analysis to determine the cost per cancer detected.

## SOLUTION

As we did in Example 6.9 and the chapter-opening example, we'll draw a probability tree (Figure 6.6). The notation is

$C$ = Has prostate cancer

$C^C$ = Does not have prostate cancer

$PT$ = Positive test result

$NT$ = Negative test result

Starting with a man between 40 and 50 years old, we have the following probabilities

**Prior**

$P(C) = .010$

$P(C^C) = 1 - .010 = .990$

**Likelihood probabilities**

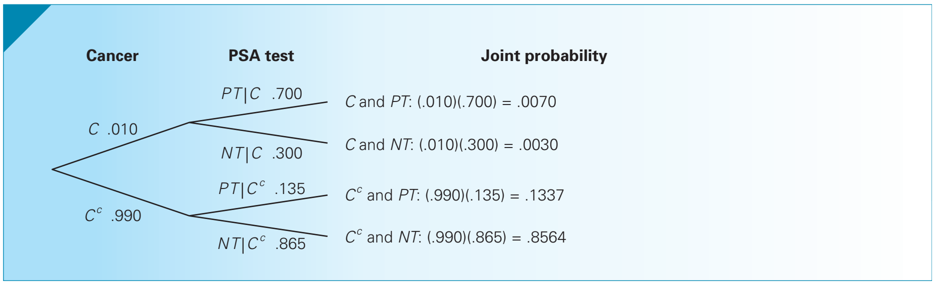
False negative: $P(NT|C) = .300$

True positive: $P(\mathrm{PT}|C) = 1 - .300 = .700$

False positive: $P(\mathrm{PT}|C^C) = .135$

True negative: $P(NT|C^C) = 1 - .135 = .865$

FIGURE 6.6 Probability Tree for Example 6.10


Cancer
PSA test
Joint probability
C .010
C^c .990
PT|C .700
NT|C .300
PT|C^c .135
NT|C^c .865
C and PT: (.010)(.700) = .0070
C and NT: (.010)(.300) = .0030
C^c and PT: (.990)(.135) = .1337
C^c and NT: (.990)(.865) = .8564

The tree allows you to determine the probability of obtaining a positive test result. It is

$$P(PT) = P(C \text{ and } PT) + P(C^C \text{ and } PT) = .0070 + .1337 = .1407$$

We can now compute the probability that the man has prostate cancer given a positive test result:

$$P(C|PT) = \frac{P(C \text{ and } PT)}{P(PT)} = \frac{.0070}{.1407} = .0498$$

The probability that he does not have prostate cancer is

$$P(C^C|PT) = 1 - P(C|PT) = 1 - .0498 = .9502$$

We can repeat the process for the other age categories. Here are the results.

| | Probabilities Given a Positive PSA Test | |
|---|---|---|
| Age | Has Prostate Cancer | Does Not Have Prostate Cancer |
| 40–49 | .0498 | .9502 |
| 50–59 | .1045 | .8955 |
| 60–69 | .2000 | .8000 |
| 70 and older | .3078 | .6922 |

The following table lists the proportion of each age category wherein the PSA test is positive [$P(PT)$]

| Age | Proportion of Tests That Are Positive | Number of Biopsies Performed per Million | Number of Cancers Detected | Number of Biopsies per Cancer Detected |
|---|---|---|---|---|
| 40–49 | .1407 | 140,700 | .0498(140,700) = 7,007 | 20.10 |
| 50–59 | .1474 | 147,400 | .1045(147,400) = 15,403 | 9.57 |
| 60–79 | .1610 | 161,000 | .2000(161,000) = 32,200 | 5.00 |
| 70 and older | .1796 | 179,600 | .3078(179,600) = 55,281 | 3.25 |

If we assume a cost of $1,000 per biopsy, the cost per cancer detected is $20,100 for 40 to 50, $9,570 for 50 to 60, $5,000 for 60 to 70, and $3,250 for over 70.

We have created an Excel spreadsheet to help you perform the calculations in Example 6.10. Open the **Excel Workbooks** folder and select **Medical screening**. There are three cells that you may alter. In cell B5, enter a new prior probability for prostate cancer. Its complement will be calculated in cell B15. In cells D6 and D15, type new values for the false-negative and false-positive rates, respectively. Excel will do the rest. We will use this spreadsheet to demonstrate some terminology standard in medical testing.

**Terminology** We will illustrate the terms using the probabilities calculated for the 40 to 50 age category.

The false-negative rate is .300. Its complement is the likelihood probability $P(PT|C)$, called the *sensitivity*. It is equal to $1 - .300 = .700$. Among men with prostate cancer, this is the proportion of men who will get a positive test result.

The complement of the false-positive rate (.135) is $P(NT|C^C)$, which is called the *specificity*. This likelihood probability is $1 - .135 = .865$

The posterior probability that someone has prostate cancer given a positive test result [$P(C|PT) = .0498$] is called the *positive predictive value*. Using Bayes's Law, we can compute the other three posterior probabilities.

The probability that the patient does not have prostate cancer given a positive test result is

$$P(C^C|PT) = .9502$$

The probability that the patient has prostate cancer given a negative test result is

$$P(C|NT) = .0035$$

The probability that the patient does not have prostate cancer given a negative test result:

$$P(C^C|NT) = .9965$$

This revised probability is called the *negative predictive value*.

## Developing an Understanding of Probability Concepts

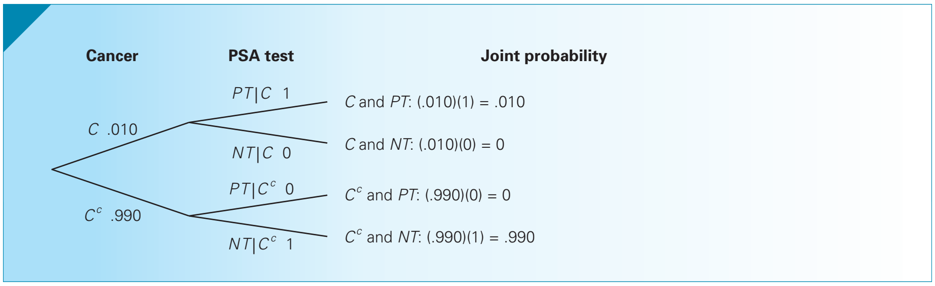
If you review the computations made previously, you'll realize that the prior probabilities are as important as the probabilities associated with the test results (the likelihood probabilities) in determining the posterior probabilities. The following table shows the prior probabilities and the revised probabilities.

| Age | Prior Probabilities for Prostate Cancer | Posterior Probabilities Given a Positive PSA Test |
|---|---|---|
| 40–49 | .010 | .0498 |
| 50–59 | .022 | .1045 |
| 60–69 | .046 | .2000 |
| 70 and older | .079 | .3078 |

As you can see, if the prior probability is low, then unless the screening test is quite accurate, the revised probability will still be quite low.

To see the effects of different likelihood probabilities, suppose the PSA test is a perfect predictor. In other words, the false-positive and false-negative rates are 0. Figure 6.7 displays the probability tree.

FIGURE 6.7 Probability Tree for Example 6.10 with a Perfect Predictor Test


Cancer
PSA test
Joint probability
$C$ .010
$C^c$ .990
$PT|C$ 1
$NT|C$ 0
$PT|C^c$ 0
$NT|C^c$ 1
$C$ and $PT$: (.010)(1) = .010
$C$ and $NT$: (.010)(0) = 0
$C^c$ and $PT$: (.990)(0) = 0
$C^c$ and $NT$: (.990)(1) = .990

We find

$$P(PT) = P(C \text{ and } PT) + P(C^C \text{ and } PT) = .01 + 0 = .01$$

$$P(C|PT) = \frac{P(C \text{ and } PT)}{P(PT)} = \frac{.01}{.01} = 1.00$$

Now we calculate the probability of prostate cancer when the test is negative.

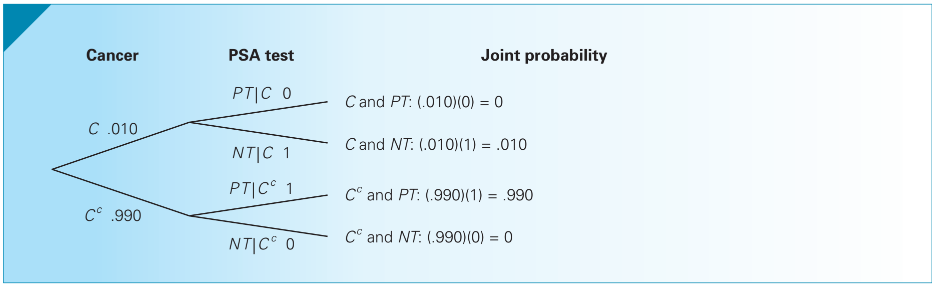
$$P(NT) \ = P(C \text{ and } NT) + P(C^C \text{ and } NT) = 0 + .99 = .99$$

$$P(C|NT) = \frac{P(C \text{ and } NT)}{P(NT)} = \frac{0}{.99} = 0$$

Thus, if the test is a perfect predictor and a man has a positive test, then as expected the probability that he has prostate cancer is 1.0. The probability that he does not have cancer when the test is negative is 0.

Now suppose that the test is always wrong; that is, the false-positive and false-negative rates are 100%. The probability tree is shown in Figure 6.8.

FIGURE **6.8** Probability Tree for Example 6.10 with a Test That Is Always Wrong


Cancer
PSA test
Joint probability
$PT|C$ 0
$C$ .010
$NT|C$ 1
$PT|C^c$ 1
$C^c$ .990
$NT|C^c$ 0
$C$ and $PT$: (.010)(0) = 0
$C$ and $NT$: (.010)(1) = .010
$C^c$ and $PT$: (.990)(1) = .990
$C^c$ and $NT$: (.990)(0) = 0

$$P(PT) = P(C \text{ and } PT) + P(C^C \text{ and } PT) = 0 + .99 = .99$$

$$P(C|PT) = \frac{P(C \text{ and } PT) \ \ 0}{P(PT) \quad .99} = 0$$

and

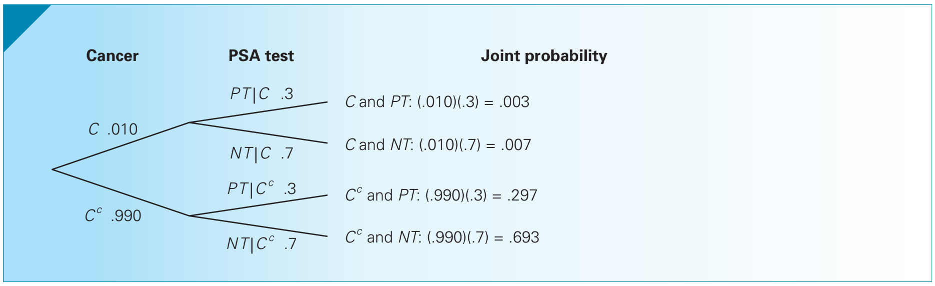
$$P(NT) \ = P(C \text{ and } NT) + P(C^C \text{ and } NT) = .01 + 0 = .01$$

$$P(C|NT) = \frac{P(C \text{ and } NT)}{P(NT)} = \frac{.01}{.01} = 1.00$$

Notice we have another perfect predictor except that it is reversed. The probability of prostate cancer given a positive test result is 0, but the probability becomes 1.00 when the test is negative.

Finally we consider the situation when the set of likelihood probabilities are the same. Figure 6.9 depicts the probability tree for a 40- to 50-year-old male and the probability of a positive test is (say) .3 and a the probability of a negative test is .7.

FIGURE **6.9** Probability Tree for Example 6.10 with Identical Likelihood Probabilities


Cancer
PSA test
Joint probability
$PT|C$ .3
$C$ .010
$NT|C$ .7
$PT|C^c$ .3
$C^c$ .990
$NT|C^c$ .7
$C$ and $PT$: (.010)(.3) = .003
$C$ and $NT$: (.010)(.7) = .007
$C^c$ and $PT$: (.990)(.3) = .297
$C^c$ and $NT$: (.990)(.7) = .693

$$P(PT) = P(C \text{ and } PT) + P(C^C \text{ and } PT) = .003 + .297 = .300$$

$$P(C|PT) = \frac{P(C \text{ and } PT)}{P(PT)} = \frac{.003}{.300} = .01$$

and

$$P(NT) = P(C \text{ and } NT) + P(C^C \text{ and } NT) = .007 + .693 = .700$$

$$P(C|NT) = \frac{P(C \text{ and } NT)}{P(NT)} = .007 + .700 = .01$$

As you can see, the posterior and prior probabilities are the same. That is, the PSA test does not change the prior probabilities. Obviously, the test is useless.

We could have used any probability for the false-positive and false-negative rates, including .5. If we had used .5, then one way of performing this PSA test is to flip a fair coin. One side would be interpreted as positive and the other side as negative. It is clear that such a test has no predictive power.

The exercises and Case 6.4 offer the probabilities for several other screening tests.

## EXERCISES

**6.71** Refer to Exercise 6.47. Determine $P(A|B)$.

**6.72** Refer to Exercise 6.48. Find the following.

a. $P(A|B)$
b. $P(A^C|B)$
c. $P(A|B^C)$
d. $P(A^C|B^C)$

**6.73** Refer to Example 6.9. An MBA applicant believes that the probability of scoring more than 650 on the GMAT without the preparatory course is .95. What is the probability of attaining that level after taking the preparatory course?

**6.74** Refer to Exercise 6.58. The plant manager randomly selects a molding from the early morning run and discovers it is defective. What is the probability that the foreman forgot to shut off the machine the previous night?

**6.75** The U.S. National Highway Traffic Safety Administration gathers data concerning the causes of highway crashes where at least one fatality has occurred. The following probabilities were determined from the 1998 annual study (BAC is blood-alcohol content). (*Source: Statistical Abstract of the United States, 2000*, Table 1042.)

$P(BAC = 0 \mid \text{Crash with fatality}) = .616$

$P(BAC \text{ is between } .01 \text{ and } .09 \mid \text{Crash with fatality}) = .300$

$P(BAC \text{ is greater than } .09 \mid \text{Crash with fatality}) = .084$

Over a certain stretch of highway during a 1-year period, suppose the probability of being involved in a crash that results in at least one fatality is .01. It has been estimated that 12% of the drivers on this highway drive while their BAC is greater than .09. Determine the probability of a crash with at least one fatality if a driver drives while legally intoxicated (BAC greater than .09).

**6.76** Refer to Exercise 6.62. A randomly selected candidate who took a CFA exam tells you that he has passed the exam. What is the probability that he took the CFA I exam?

**6.77** Bad gums may mean a bad heart. Researchers discovered that 85% of people who have suffered a heart attack had periodontal disease, an inflammation of the gums. Only 29% of healthy people have this disease. Suppose that in a certain community heart attacks are quite rare, occurring with only 10% probability. If someone has periodontal disease, what is the probability that he or she will have a heart attack?

**6.78** Refer to Exercise 6.77. If 40% of the people in a community will have a heart attack, what is the probability that a person with periodontal disease will have a heart attack?

**6.79** Data from the Office on Smoking and Health, Centers for Disease Control and Prevention, indicate that 40% of adults who did not finish high school, 34% of high school graduates, 24% of adults who completed some college, and 14% of college graduates smoke. Suppose that one individual is selected at random, and it is discovered that the individual smokes. What is the probability that the individual is a college graduate? Use the probabilities in Exercise 6.45 to calculate the probability that the individual is a college graduate.

**6.80** Three airlines serve a small town in Ohio. Airline A has 50% of all the scheduled flights, airline B has 30%, and airline C has the remaining 20%. Their on-time rates are 80%, 65%, and 40%, respectively. A plane has just left on time. What is the probability that it was airline A?

**6.81** Your favorite team is in the final playoffs. You have assigned a probability of 60% that it will win the championship. Past records indicate that when teams win the championship, they win the first game of the series 70% of the time. When they lose the series, they win the first game 25% of the time. The first game is over; your team has lost. What is the probability that it will win the series?

*The following exercises are based on the Applications in Medical Screening and Medical Insurance subsection.*

**6.82** Transplant operations have become routine. One common transplant operation is for kidneys. The most dangerous aspect of the procedure is the possibility that the body may reject the new organ. Several new drugs are available for such circumstances, and the earlier the drug is administered, the higher the probability of averting rejection. The *New England Journal of Medicine* recently reported the development of a new urine test to detect early warning signs that the body is rejecting a transplanted kidney. However, like most other tests, the new test is not perfect. When the test is conducted on someone whose kidney will be rejected, approximately one out of five tests will be negative (i.e., the test is wrong). When the test is conducted on a person whose kidney will not be rejected, 8% will show a positive test result (i.e., another incorrect result). Physicians know that in about 35% of kidney transplants the body tries to reject the organ. Suppose that the test was performed and the test is positive (indicating early warning of rejection). What is the probability that the body is attempting to reject the kidney?

**6.83** The Rapid Test is used to determine whether someone has HIV (the virus that causes AIDS). The false-positive and false-negative rates are .027 and .080, respectively. A physician has just received the Rapid Test report that his patient tested positive. Before receiving the result, the physician assigned his patient to the low-risk group (defined on the basis of several variables) with only a 0.5% probability of having HIV. What is the probability that the patient actually has HIV?

**6.84** What are the sensitivity, specificity, positive predictive value, and negative predictive value in the previous exercise?

**6.85** The Pap smear is the standard test for cervical cancer. The false-positive rate is .636; the false-negative rate is .180. Family history and age are factors that must be considered when assigning a probability of cervical cancer. Suppose that, after obtaining a medical history, a physician determines that 2% of women of his patient's age and with similar family histories have cervical cancer. Determine the effects a positive and a negative Pap smear test have on the probability that the patient has cervical cancer.

## 6.5 Identifying the Correct Method

As we've previously pointed out, the emphasis in this book will be on identifying the correct statistical technique to use. In Chapters 2 and 4, we showed how to summarize data by first identifying the appropriate method to use. Although it is difficult to offer strict rules on which probability method to use, we can still provide some general guidelines.

In the examples and exercises in this text (and most other introductory statistics books), the key issue is whether joint probabilities are provided or are required.

### Joint Probabilities Are Given

In Section 6.2, we addressed problems where the joint probabilities were given. In these problems, we can compute marginal probabilities by adding across rows and down columns. We can use the joint and marginal probabilities to compute conditional probabilities, for which a formula is available. This allows us to determine whether the events described by the table are independent or dependent.

We can also apply the addition rule to compute the probability that either of two events occur.

### Joint Probabilities Are Required

The previous section introduced three probability rules and probability trees. We need to apply some or all of these rules in circumstances where one or more joint probabilities are required. We apply the multiplication rule (either by formula or through a probability tree) to calculate the probability of intersections. In some problems, we're interested in adding these joint probabilities. We're actually applying the addition rule for mutually exclusive events here. We also frequently use the complement rule. In addition, we can also calculate new conditional probabilities using Bayes's Law.

## Chapter Summary

The first step in assigning probability is to create an **exhaustive** and **mutually exclusive** list of outcomes. The second step is to use the **classical**, **relative frequency**, or **subjective approach** and assign probability to the outcomes. A variety of methods are available to compute the probability of other events. These methods include **probability rules** and **trees**.

An important application of these rules is **Bayes's Law**, which allows us to compute conditional probabilities from other forms of probability.

### IMPORTANT TERMS

| | |
|---|---|
| Random experiment 176 | Independent events 185 |
| Exhaustive 176 | Union 186 |
| Mutually exclusive 176 | Complement 191 |
| Sample space 177 | Complement rule 191 |
| Classical approach 177 | Multiplication rule 191 |
| Relative frequency approach 178 | Addition rule 193 |
| Subjective approach 178 | Bayes's Law 199 |
| Event 178 | Prior probability 200 |
| Intersection 181 | Likelihood probability 200 |
| Joint probability 181 | Posterior probability 200 |
| Marginal probability 183 | False-positive 203 |
| Conditional probability 183 | False-negative 203 |

### FORMULAS

Conditional probability

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

Complement rule

$$P(A^C) = 1 - P(A)$$

Multiplication rule

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

Addition rule

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

## Chapter Exercises

**6.86** The following table lists the joint probabilities of achieving grades of A and not achieving A's in two MBA courses.

| | Achieve a Grade of A in Marketing | Does Not Achieve a Grade of A in Marketing |
|---|---|---|
| Achieve a grade of A in statistics | .053 | .130 |
| Does not achieve a grade of A in statistics | .237 | .580 |

a. What is the probability that a student achieves a grade of A in marketing?

b. What is the probability that a student achieves a grade of A in marketing, given that he or she does not achieve a grade of A in statistics?

c. Are achieving grades of A in marketing and statistics independent events? Explain.

**6.87** A construction company has bid on two contracts. The probability of winning contract A is .3. If the company wins contract A, then the probability of winning contract B is .4. If the company loses contract A, then the probability of winning contract B decreases to .2. Find the probability of the following events.

a. Winning both contracts

b. Winning exactly one contract

c. Winning at least one contract

**6.88** Laser surgery to fix shortsightedness is becoming more popular. However, for some people, a second procedure is necessary. The following table lists the joint probabilities of needing a second procedure and whether the patient has a corrective lens with a factor (diopter) of minus 8 or less.

| | Vision Corrective Factor of More Than Minus 8 | Vision Corrective Factor of Minus 8 or Less |
|---|---|---|
| First procedure is successful | .66 | .15 |
| Second procedure is required | .05 | .14 |

a. Find the probability that a second procedure is required.

b. Determine the probability that someone whose corrective lens factor is minus 8 or less does not require a second procedure.

c. Are the events independent? Explain your answer.

**6.89** The effect of an antidepressant drug varies from person to person. Suppose that the drug is effective on 80% of women and 65% of men. It is known that 66% of the people who take the drug are women. What is the probability that the drug is effective?

**6.90** Refer to Exercise 6.89. Suppose that you are told that the drug is effective. What is the probability that the drug taker is a man?

**6.91** In a four-cylinder engine there are four spark plugs. If any one of them malfunctions, the car will idle roughly and power will be lost. Suppose that for a certain brand of spark plugs the probability that a spark plug will function properly after 5,000 miles is .90. Assuming that the spark plugs operate independently, what is the probability that the car will idle roughly after 5,000 miles?

**6.92** A telemarketer sells magazine subscriptions over the telephone. The probability of a busy signal or no answer is 65%. If the telemarketer does make contact, the probability of 0, 1, 2, or 3 magazine subscriptions is .5, .25, .20, and .05, respectively. Find the probability that in one call she sells no magazines.

**6.93** A statistics professor believes that there is a relationship between the number of missed classes and the grade on his midterm test. After examining his records, he produced the following table of joint probabilities.

| | Student Fails the Test | Student Passes the Test |
|---|---|---|
| Student misses fewer than 5 classes | .02 | .86 |
| Student misses 5 or more classes | .09 | .03 |

a. What is the pass rate on the midterm test?

b. What proportion of students who miss five or more classes passes the midterm test?

c. What proportion of students who miss fewer than five classes passes the midterm test?

d. Are the events independent?

**6.94** In Canada, criminals are entitled to parole after serving only one-third of their sentence. Virtually all prisoners, with several exceptions including murderers, are released after serving two-thirds of their sentence. The government has proposed a new law that would create a special category of inmates based on whether they had committed crimes involving violence or drugs. Such criminals would be subject to additional detention if the Correction Services judges them highly likely to reoffend. Currently, 27% of prisoners who are released commit another crime within 2 years of release. Among those who have reoffended, 41% would have been detained under the new law, whereas 31% of those who have not reoffended would have been detained.

a. What is the probability that a prisoner who would have been detained under the new law does commit another crime within 2 years?

b. What is the probability that a prisoner who would not have been detained under the new law does commit another crime within 2 years?

**6.95** Casino Windsor conducts surveys to determine the opinions of its customers. Among other questions, respondents are asked to give their opinion about "Your overall impression of Casino Windsor." The responses are

Excellent Good Average Poor

In addition, the gender of the respondent is noted. After analyzing the results, the following table of joint probabilities was produced.

| Rating | Women | Men |
|---|---|---|
| Excellent | .27 | .22 |
| Good | .14 | .10 |
| Average | .06 | .12 |
| Poor | .03 | .06 |

a. What proportion of customers rate Casino Windsor as excellent?

b. Determine the probability that a male customer rates Casino Windsor as excellent.

c. Find the probability that a customer who rates Casino Windsor as excellent is a man.

d. Are gender and rating independent? Explain your answer.

**6.96** A customer-service supervisor regularly conducts a survey of customer satisfaction. The results of the latest survey indicate that 8% of customers were not satisfied with the service they received at their last visit to the store. Of those who are not satisfied, only 22% return to the store within a year. Of those who are satisfied, 64% return within a year. A customer has just entered the store. In response to your question, he informs you that it is less than 1 year since his last visit to the store. What is the probability that he was satisfied with the service he received?

**6.97** How does level of affluence affect health care? To address one dimension of the problem, a group of heart attack victims was selected. Each was categorized as a low-, medium-, or high-income earner. Each was also categorized as having survived or died. A demographer notes that in our society 21% fall into the low-income group, 49% are in the medium-income group, and 30% are in the high-income group. Furthermore, an analysis of heart attack victims reveals that 12% of low-income people, 9% of medium-income people, and 7% of high-income people die of heart attacks. Find the probability that a survivor of a heart attack is in the low-income group.

**6.98** A statistics professor and his wife are planning to take a 2-week vacation in Hawaii, but they can't decide whether to spend 1 week on each of the islands of Maui and Oahu, 2 weeks on Maui, or 2 weeks on Oahu. Placing their faith in random chance, they insert two Maui brochures in one envelope, two Oahu brochures in a second envelope, and one brochure from each island in a third envelope. The professor's wife will select one envelope at random, and their vacation schedule will be based on the brochures of the islands so selected. After his wife randomly selects an envelope, the professor removes one brochure from the envelope (without looking at the second brochure) and observes that it is a Maui brochure. What is the probability that the other brochure in the envelope is a Maui brochure? (Proceed with caution: The problem is more difficult than it appears.)

**6.99** The owner of an appliance store is interested in the relationship between the price at which an item is sold (regular or sale price) and the customer's decision on whether to purchase an extended warranty. After analyzing her records, she produced the following joint probabilities.

| | Purchased Extended Warranty | Did Not Purchase Extended Warranty |
|---|---|---|
| Regular price | .21 | .57 |
| Sale price | .14 | .08 |

a. What is the probability that a customer who bought an item at the regular price purchased the extended warranty?

b. What proportion of customers buy an extended warranty?

c. Are the events independent? Explain.

**6.100** Researchers have developed statistical models based on financial ratios that predict whether a company will go bankrupt over the next 12 months. In a test of one such model, the model correctly predicted the bankruptcy of 85% of firms that did in fact fail, and it correctly predicted nonbankruptcy for 74% of firms that did not fail. Suppose that we expect 8% of the firms in a particular city to fail over the next year. Suppose that the model predicts bankruptcy for a firm that you own. What is the probability that your firm will fail within the next 12 months?

**6.101** A union's executive conducted a survey of its members to determine what the membership felt were the important issues to be resolved during upcoming negotiations with management. The results indicate that 74% of members felt that job security was an important issue, whereas 65% identified pension benefits as an important issue. Of those who felt that pension benefits were important, 60% also felt that job security was an important issue. One member is selected at random.

a. What is the probability that he or she felt that both job security and pension benefits were important?

b. What is the probability that the member felt that at least one of these two issues was important?

**6.102** In a class on probability, a statistics professor flips two balanced coins. Both fall to the floor and roll under his desk. A student in the first row informs the professor that he can see both coins. He reports that at least one of them shows tails. What is the probability that the other coin is also tails? (Beware the obvious.)

**6.103** Refer to Exercise 6.102. Suppose the student informs the professor that he can see only one coin and it shows tails. What is the probability that the other coin is also tails?

## CASE 6.1 Let's Make a Deal

© Michael Newman/PhotoEdit

A number of years ago, there was a popular television game show called *Let's Make a Deal*. The host, Monty Hall, would randomly select contestants from the audience and, as the title suggests, he would make deals for prizes. Contestants would be given relatively modest prizes and would then be offered the opportunity to risk those prizes to win better ones.

Suppose that you are a contestant on this show. Monty has just given you a free trip touring toxic waste sites around the country. He now offers you a trade: Give up the trip in exchange for a gamble. On the stage are three curtains, A, B, and C. Behind one of them is a brand new car worth $20,000. Behind the other two curtains, the stage is empty. You decide to gamble and select curtain A. In an attempt to make things more interesting, Monty then exposes an empty stage by opening curtain C (he knows there is nothing behind curtain C). He then offers you the free trip again if you quit now or, if you like, he will propose another deal (i.e., you can keep your choice of curtain A or perhaps switch to curtain B). What do you do?

To help you answer that question, try first answering these questions.

1. Before Monty shows you what's behind curtain C, what is the probability that the car is behind curtain A? What is the probability that the car is behind curtain B?
2. After Monty shows you what's behind curtain C, what is the probability that the car is behind curtain A? What is the probability that the car is behind curtain B?

## CASE 6.2 To Bunt or Not to Bunt, That Is the Question

© AlBehrmanAP

No sport generates as many statistics as baseball. Reporters, managers, and fans argue and discuss strategies on the basis of these statistics. An article in *Chance* ("A Statistician Reads the Sports Page," Hal S. Stern, Vol. 1, Winter 1997) offers baseball lovers another opportunity to analyze numbers associated with the game. Table 1 lists the probabilities of scoring at least one run in situations that are defined by the number of outs and the bases occupied. For example, the probability of scoring at least one run when there are no outs and a man

(Case 6.4 continued)

| Mother's Age | False-Positive Rate | False-Negative Rate |
|---|---|---|
| Under 30 | .04 | .376 |
| 30–34 | .082 | .290 |
| 35–37 | .178 | .269 |
| Over 38 | .343 | .029 |

The probability that a baby has Down syndrome is primarily a function of the mother's age. The probabilities are listed here.

| Age | Probability of Down Syndrome |
|---|---|
| 25 | 1/1300 |
| 30 | 1/900 |
| 35 | 1/350 |
| 40 | 1/100 |
| 45 | 1/25 |
| 49 | 1/12 |

a. For each of the ages 25, 30, 35, 40, 45, and 49 determine the probability of Down syndrome if the maternity serum screening produces a positive result.

b. Repeat for a negative result.

## CASE 6.5 Probability That at Least Two People in the Same Room Have the Same Birthday

© Anna Jurkovska/Shutterstock

Suppose that there are two people in a room. The probability that they share the same birthday (date, not necessarily year) is 1/365, and the probability that they have different birthdays is 364/365. To illustrate, suppose that you're in a room with one other person and that your birthday is July 1. The probability that the other person does not have the same birthday is 364/365 because there are 364 days in the year that are not July 1. If a third person now enters the room, the probability that he or she has a different birthday from the first two people in the room is 363/365. Thus, the probability that three people in a room having different birthdays is (364/365)(363/365). You can continue this process for any number of people.

Find the number of people in a room so that there is about a 50% probability that at least two have the same birthday.

Hint1: Calculate the probability that they don't have the same birthday.

Hint 2: Excel users can employ the **product** function to calculate joint probabilities.