CENGAGE

Statistics
for Management and Economics

12th Edition

Gerald Keller

Identify Compute Interpret

Statistics
for Management and Economics

12e

GERALD KELLER

CENGAGE

Australia • Brazil • Canada • Mexico • Singapore • United Kingdom • United States

CENGAGE

Statistics for Management and Economics, 12e
Gerald Keller

SVP, Higher Education Product Management: Erin Joyner

VP, Product Management, Learning Experiences: Thais Alencar

Product Director: Joe Sabatino

Senior Product Manager: Aaron Arnsparger

Product Assistant: Flannery Cowan

Senior Learning Designer: Brandon Foltz

Senior Content Manager: Conor Allen

Associate Subject-Matter Expert: Nancy Marchant

Digital Delivery Quality Partner: Steven McMillian

Director, Product Marketing: Danae April

IP Analyst: Ashley Maynard

IP Project Manager: Kumaresan Chandrakumar, Integra

Production Service: MPS Limited

Designer: Chris Doughman

Cover Image Source: IR Stone/ShutterStock.com

Interior Image Source: cmillerdesign

Library of Congress Control Number: 2021913565

ISBN: 978-0-357-71427-0

Cengage
200 Pier 4 Boulevard
Boston, MA 02210
USA

Cengage is a leading provider of customized learning solutions with employees residing in nearly 40 different countries and sales in more than 125 countries around the world. Find your local representative at **www.cengage.com.**

To learn more about Cengage platforms and services, register or access your online learning solution, or purchase materials for your course, visit **www.cengage.com.**

Printed in the United States of America
Print Number: 01 Print Year: 2022

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 statistical 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 genuine 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 interpret 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 now 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 10 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 four ways: manually (with the aid of a calculator), using Excel, XLSTAT, or Stata. 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 that 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 1,283 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 interpret 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 by providing 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 businesses 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-five application boxes are scattered throughout the book.

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 five 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

The use of statistical software has been reorganized. First, Excel can be used for all statistical applications. Second, XLSTAT output and instructions, which were introduced in the 11th edition, have been placed in the appendixes to Chapters 2 to 4, and 10 to 19. Third, Stata has been included for the first time with output and instructions in appendixes similar to the treatment of XLSTAT.

The data from the last 10 General Social Surveys and the last five Surveys of Consumer Finances have been included, which produced hundreds of new exercises. Students will have the opportunity to convert real data into information. Instructors can use these data sets to create hundreds of additional examples and exercises.

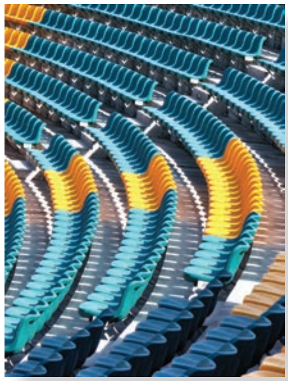
Many of the examples, exercises, and cases using real data in the 11th edition have been updated. These include the data on wins, payrolls, and attendance in baseball, basketball, football, and hockey; returns on stocks listed on the New York Stock Exchange, NASDAQ, and Toronto Stock Exchange; and global warming.

I've created many new examples and exercises. Here are the numbers for the 12th edition: 137 solved examples, 2,573 exercises, 32 cases, and 1,283 data sets.

New! MindTap Courseware....

Assign this textbook through MindTap to provide online homework and assessment, study tools, and seamless access to the eBook—inside or outside of your campus Learning Management System. MindTap includes chapter quizzes, Exploring Statistics applets with teaching videos and activities, assignable exercises from the textbook with algorithmic versions and solutions, auto-graded Excel problems, an algorithmic test bank, and more! Contact your Cengage representative for more information about accessing MindTap.

Data Driven: The Big Picture



Izf/Shutterstock.com

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.

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.

EXAMPLE 13.1*

DATA Xm13-01

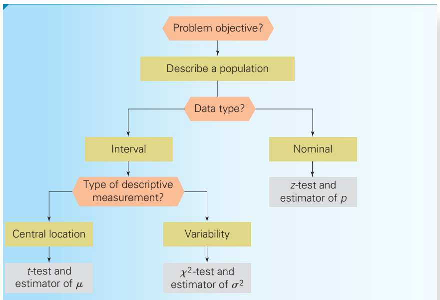
Direct and Broker-Purchased Mutual Funds

Millions of investors buy mutual funds (see page 175 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?

**Source:* D. Bergstresser, J. Chalmers, and P. Tufano, "Assessing the Costs and Benefits of Brokers in the Mutual Fund Industry."



Appendixes 13, 14, 15, 16, 17, and **19** 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.C 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.C, we provide a table of the techniques, a flowchart to help you identify the correct technique, and 34 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 μ
Estimator of μ (including estimator of $N\mu$)
χ^2 test of σ^2
Estimator of σ^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 μ_D
Estimator of μ_D
F-test of σ_1^2/σ_2^2
Estimator of σ_1^2/σ_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 μ_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, found in Appendix C 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. Here is part of the guide.

A Guide to Statistical Techniques

Problem Objectives

DATA TYPES		Describe a Population	Compare Two Populations	Compare Two or More Populations
	Interval	Histogram **Section 3-1** Line chart **Section 3-2** Mean, median, and mode **Section 4-1** Range, variance, and standard deviation **Section 4-2** Percentiles and quartiles **Section 4-3** *t*-test and estimator of a mean **Section 12-1** 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** Unequal-variances *t*-test and estimator of the difference between two means: independent samples **Section 13-1** *t*-test and estimator of mean difference **Section 13-3** *F*-test and estimator of ratio of two variances **Section 13-4** Wilcoxon rank sum test **Section 19-1** Wilcoxon signed rank sum test **Section 19-2**	One-way analysis of variance **Section 14-1** LSD multiple comparison method **Section 14-2** Tukey's multiple comparison method **Section 14-2** Two-way analysis of variance **Section 14-4** Two-factor analysis of varianc **Section 14-5** Kruskal-Wallis test **Section 19-3** Friedman test **Section 19-3**
	Nominal	Frequency distribution **Section 2-2** Bar chart **Section 2-2** Pie chart **Section 2-2** *Z*-test and estimator of a proportion **Section 12-3** Chi-squared goodness-of-fit test **Section 15-1**	*Z*-test and estimator of the difference between two proportions **Section 13-5** Chi-squared test of a contingency table **Section 15-2**	Chi-squared test of a contingency table **Section 15-2**
	Ordinal	Median **Section 4-1** Percentiles and quartiles **Section 4-3**	Wilcoxon rank sum test **Section 19-1** Sign test **Section 19-2**	Kruskal-Wallis test **Section 19-3** Friedman test **Section 19-3**

More Data Sets

A total of 1,283 data sets available to be downloaded provide ample practice. These data sets contain real data, including stock market returns, climate change temperature anomalies and atmospheric carbon dioxide, baseball, basketball, football and hockey team payrolls, wins, and attendance.

homes. Each was asked how many minutes they spent reading their newspapers. Can we infer that the amount of time reading differs between the two groups?

13.228 Xr13-228 In recent years, a number of state governments have passed mandatory seat-belt laws. Although the use of seat belts is known to save lives and reduce serious injuries, compliance with seat-belt laws is not universal. In an effort to increase the use of seat belts, a government agency sponsored a 2-year study. Among its objectives was to determine whether there was enough evidence to infer that seat-belt usage increased between last year and this year. To test this belief, random samples of drivers last year and this year were asked whether they always use their seat belts (2 = Wear seat belt, 1 = Do not wear seat belt). Can we infer that seat-belt usage has increased over the last year?

13.229 Xr13-229 An important component of the cost of living is the amount of money spent on housing. Housing costs include rent (for tenants), mortgage payments and property tax (for home owners), heating, electricity, and water. An economist undertook a 5-year study to determine how housing costs have changed. Random samples of 200 households this year and 5 years ago were drawn and the percentage of total income spent on housing was recorded.

a. Conduct a test to determine whether the economist can infer that housing cost as a percentage of total income has increased over the last 5 years.
b. Use whatever statistical method you deem appropriate to check the required condition(s) of the test used in part (a).

13.230 Xr13-230 In designing advertising campaigns to sell magazines, it is important to know how much time each of a number of demographic group spends reading magazines. In a preliminary study 40 people were randomly selected. All were asked how much time per week they spend reading magazines; additionally, each was categorized by gender (1 = Male, 2 = Female) and by income level (1 = Low, 2 = High).

a. Is there sufficient evidence to conclude that

consisted of a female and a male student who were matched according to their GPAs, courses taken, ages, and previous work expe… salary offered (in thousands of d… uate was recorded.

a. Is there enough evidence to … a factor in salary offers?
b. Discuss why the experiment … the way it was.
c. Is the required condition fo… satisfied?

13.232 Xr13-232 Refer to Exercise 1… poll also asked whether respon… view (1) or not (0) of the fede… the responses for 2016 and 201… evidence that Americans in 201… government less positively in 2…

13.233 Xr13-233 Before deciding which of two types of stamping machines should be purchased, the plant manager of an automotive parts manufacturer wants to determine the number of units that each produces. The two machines differ in cost, reliability, and productivity. The firm's accountant has calculated that machine A must produce 25 more non-defective units per hour than machine B to warrant buying machine A. To help decide, both machines were operated for 24 hours. The total number of units and the number of defective units produced by each machine per hour were recorded. These data are stored in the following way: column A = Total number of units produced by machine A; column B = Number of defectives produced by machine A; column C = Total number of units produced by machine B; column D = Number

EXAMPLE 13.9 — **Test Marketing of Package Designs, Part 1**

DATA Xm13-09

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 design is more

CASE 3.2 — **What Is Happening to the Polar Ice Caps**

DATA C03-02a C03-02b C03-02c C03-02d

The most serious consequence of possible rising temperatures is that the ice in one or both poles will melt, leading to a rise in the world's oceans with trillions of dollars in damages to coastal cities. This raises the question: Are the oceans rising? Since 1993, satellite measurements of the sea level have been recorded. These measurements allow for estimation the proportion of the year that has elapsed. The sea level is measured in millimeters and has been reset so that the value of the last days of 1992 is set to 0.

We have recorded the monthly extent of Northern hemisphere (C03-02b) and Southern hemisphere (C03-02c) sea ice (measured in millions of kilometers squared) starting in January 1979. The anomalies are

level. Describe what you have discovered.

b. For each hemisphere, use a suitable graph to display the trends in ice coverage.
c. For each hemisphere, use a graphical technique to determine whether the polar ice coverage is related to temperature anomalies (C03-02d).
d. What do the graphs you

APPENDIX A

Data File Sample Statistics

Chapter 10

10.34 $\bar{x} = 252.38$
10.35 $\bar{x} = 1{,}810.16$
10.36 $\bar{x} = 12.10$
10.37 $\bar{x} = 10.21$
10.38 $\bar{x} = .510$
10.39 $\bar{x} = 26.81$

12.52 $\bar{x} = 1{,}158$, $s = 396.5$, $n = 325$
12.53 $\bar{x} = 530.7$, $s = 97.17$, $n = 485$
12.77 $s^2 = 270.58$, $n = 25$
12.78 $s^2 = 22.56$, $n = 245$
12.79 $s^2 = 4.725$, $n = 90$
12.80 $s^2 = 174.47$, $n = 100$
12.81 $s^2 = 19.68$, $n = 25$

13.20 Oat bran: $\bar{x}_1 = 10.01$, $s_1 = 4.43$, $n_1 = 120$; Other: $\bar{x}_2 = 9.12$, $s_2 = 4.45$, $n_2 = 120$
13.21 Regular income: $\bar{x}_1 = 208.5$, $s_1 = 30.86$, $n_1 = 177$; Stimulus: $\bar{x}_2 = 217.5$, $s_2 = 37.17$,

Appendix A provides summary statistics for many of the exercises with large data sets. This feature offers unparalleled flexibility allowing students to solve most exercises by hand or by computer!

Real Data Sets

The data from the last 10 General Social Surveys and the last five Surveys of Consumer Finances are included. These feature thousands of observations and dozens of selected variables. Solving more than 500 exercises associated with these surveys encourages students to uncover interesting aspects of the society. For example, students can determine the incomes, education, and working hours of people who are self-employed and compare them to people who work for someone else. They can see the effect of education on income, assets, investments, and net worth. Instructors can use the data to create their own examples and exercises.

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.

COMPUTE

MANUALLY:

We need four values to construct the confidence interval estimate of μ. They are

$$\bar{x}, z_{\alpha/2}, \sigma, n$$

Using a calculator, we determine the summation $\sum x_i = 9{,}254$. From this, we find

$$\bar{x} = \frac{\sum x_i}{n} = \frac{9{,}254}{25} = 370.16$$

The confidence level is set at 95%;
thus, $1 - \alpha = .95$, $\alpha = 1 - .95 = .05$, and $\alpha/2 = .025$.
From Table 3 in Appendix B or from Table 10.1, we find

$$z_{\alpha/2} = z_{.025} = 1.96$$

The population standard deviation is $\sigma = 75$, and the sample size is 25. Substituting $\bar{x}$, $z_{\alpha/2}$, σ, and n into the confidence interval estimator, we find

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 370.16 \pm z_{.025}\frac{75}{\sqrt{25}} = 370.16 \pm 1.96\frac{75}{\sqrt{25}} = 370.16 \pm 29.40$$

The lower and upper confidence limits are LCL = 340.76 and UCL = 399.56, respectively.

Excel Workbook

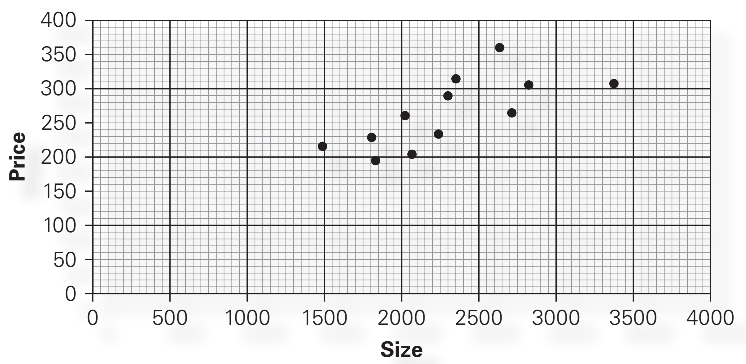
	A	B	C	D	E
1	z-Estimate of a Mean				
2					
3	Sample mean	370.16	Confidence Interval Estimate		
4	Population standard deviation	75	370.16	±	29.40
5	Sample size	25	Lower confidence limit		340.76
6	Confidence level	0.95	Upper confidence limit		399.56

INSTRUCTIONS

1. Type or import the data into one column. (Open Xm10-01.) In any empty cell, calculate the sample mean (=AVERAGE(A1:A26)).
2. Open the **Estimators Workbook** and click the **z-Estimate_Mean** tab. In cell B3, type or copy the value of the sample mean. If you use **Copy** also use **Paste Special** and **Values**. In cells B4–B6, type the value of σ (75), the value of n (25), and the confidence level (.95), respectively.

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

Step-by-step instructions in the use of **Excel** 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. Additionally, instructions and printouts for XLSTAT and Stata are provided in the appendixes to most chapters.



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 graphic outputs are given for students to compare to their own results.

Interpret the Results

INTERPRET

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.

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 42 **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. Another optional section, "Applications in Marketing: Market Segmentation" demonstrates how to estimate the size of a market segment.

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

APPLICATIONS in **OPERATIONS MANAGEMENT**

Michaeljung/Shutterstock.com

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.

The Number of Unemployed

DATA GSS2018

One of the most important economic statistics is the unemployment rate. Unfortunately, it is a very poor measure because it is misleading. The United States Bureau of Labor Statistics (BLS) defines the unemployment rate as the percentage of unemployed persons who are currently in the labor force. In order to be in the labor force, a person either must have a job or have looked for work in the last 4 weeks. This leaves out a lot of people. Some are left out because they have not done anything to find work in more than 4 weeks and as a result became discouraged, and some are left out because they are not available for work at the moment. Yet to leave this group out significantly underestimates the unemployment rate.

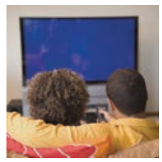
Jupiterimages/Getty images

See page 423 for our answer.

(Continued)

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.

The Number of Unemployed: Solution



Jupiterimages/Getty images

IDENTIFY

The problem objective is to describe the population of work status of American adults. The data are nominal. The combination of problem objective and data type make the parameter to be estimated the proportion of the entire population that is unemployed. The confidence interval estimator of the population is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

COMPUTE

MANUALLY:

To solve manually we count the number of 3s and 4s in the WRKSTAT column. They are 53 and 84, respectively. The sample size is 2,346. (There are two blanks representing missing data.) Thus,

$$\hat{p} = \frac{53 + 84}{2{,}346} = .0584$$

The confidence level is $1 - \alpha = .95$. It follows that $\alpha = .05$, $\alpha/2 = .025$, $z_{\alpha/2} = z_{.025} = 1.96$. The 95% confidence interval estimate of p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = .0584 \pm 1.96\sqrt{\frac{.0584(1-.0584)}{2{,}346}} = .0584 \pm .0095$$

$$LCL = .0489 \qquad UCL = .0679$$

EXCEL Workbook

	A	B	C	D	E
1	z-Estimate of a Proportion				
2					
3	Sample proportion	0.0584	Confidence Interval Estimate		
4	Sample size	2346	0.0584	±	0.0095
5	Confidence level	0.95	Lower confidence limit		0.0489
6			Upper confidence limit		0.0679

INSTRUCTIONS

1. Type or import the data into one column. (Open GSS2018.) (We copied column X into another spreadsheet.) In any empty cell, calculate the number of "successes" (=COUNTIF (A1:A2349,3) and =COUNTIF (A1:A2349,4)). Divide this number (53 + 84) by the sample size (2,346) to obtain the sample proportion.
2. Open the **Estimators Workbook** and click the **z-Estimate_Proportion** tab. Type or copy the sample proportion. Type the value of the sample size and the value of α.

INTERPRET

We estimate that the proportion of unemployed American adults lies between 4.89% and 6.79%. To determine the number of unemployed people, multiply the lower and upper limits by the population size 255,200,373. Thus,

$$LCL = 255{,}200{,}373(.0489) = 12{,}479{,}298$$
$$UCL = 255{,}200{,}373(.0679) = 17{,}328{,}105$$

CASE 12.5 Bias in Roulette Betting

The game of roulette consists of a wheel with 38 colored and numbered slots. The numbers are 1 to 36, 0 and 00. Half of the slots numbered 1 to 36 are red and the other half are black. The two "zeros" are green. The wheel is spun and an iron ball is rolled, which eventually comes to rest in one of the slots. Gamblers can make several different kinds of bets. Most players bet on one or more numbers or on

DATA C12-05

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.

Chapter summaries briefly review the material and list important terms, symbols, and formulas.

CHAPTER SUMMARY

The inferential methods presented in this chapter address the problem of describing a single population. When the data are interval, the parameters of interest are the population mean μ and the population variance σ^2. The Student t-distribution is used to test and estimate the mean when the population standard deviation is unknown. The chi-squared distribution is used to make inferences about a population variance. When the data are nominal, the parameter to be tested and estimated is the population proportion p. The sample proportion follows an approximate normal distribution, which produces the test statistic and the interval estimator. We also discussed how to determine the sample size required to estimate a population proportion. We introduced market segmentation and described how statistical techniques presented in this chapter can be used to estimate the size of a segment.

IMPORTANT TERMS:

t-statistic 397
Student t-distribution 397
Robust 402
Chi-squared statistic 412

SYMBOLS:

Symbol	Pronounced	Represents
ν	nu	Degrees of freedom
χ^2	chi squared	Chi-squared statistic
$\hat{p}$	p hat	Sample proportion
$\tilde{p}$	p tilde	Wilson estimator

FORMULAS:

Test statistic for μ

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Confidence interval estimator of μ

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Test statistic for σ^2

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Confidence interval estimator of σ^2

$$\text{LCL} = \frac{(n-1)s^2}{\chi^2_{\alpha/2}}$$

$$\text{UCL} = \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

Test statistic for p

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

Confidence interval estimator of p

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

Sample size to estimate p

$$n = \left(\frac{z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{B}\right)^2$$

Wilson estimator

$$\tilde{p} = \frac{x+2}{n+4}$$

Confidence interval estimator of p using the Wilson estimator

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/(n+4)}$$

Confidence interval estimator of the total of a large finite population

$$N\left[\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

Confidence interval estimator of the total number of successes in a large finite population

$$N\left[\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Instructor and Student Resources

Additional instructor and student resources for this product are available online. Instructor assets include an Instructor's Manual, Solutions and Answer Guide, Educator's Guide, PowerPoint® slides, and a test bank powered by Cognero®. New to this edition for instructors, Excel solutions files are now available for certain exercises in the book that require computer generated solutions. Student assets include Excel datasets, Excel workbooks, and more. Sign up or sign in at www.cengage.com to search for and access this product and its online resources.

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 Product Manager Aaron Arnsparger, Senior Content Manager Conor Allen, Senior Learning Designer Brandon Foltz, and Associate Subject-Matter Expert Nancy Marchant. (For all remaining errors, place the blame where it belongs—on me.) Their advice and suggestions made my task considerably easier.

Paolo Catasti, Virginia Commonwealth University, produced the Instructor PowerPoint slides and the Instructor's Manual.

The author extends thanks also to the survey participants and reviewers of the previous editions: Roger Bailey, Vanderbilt University; Paul Baum, California State University–Northridge; Nagraj Balakrishnan, Clemson University; Chen-Huei Chou, College of Charleston; Howard Clayton, Auburn University; Philip Cross, Georgetown University; Barry Cuffe, Wingate University; Ernest Demba, Washington University–St. Louis; Michael Douglas, Millersville University; Neal Duffy, State University of New York–Plattsburgh; John Dutton, North Carolina State University; Ehsan Elahi, University of Massachusetts–Boston; Erick Elder, University of Arkansas; Mohammed El-Saidi, Ferris State University; Grace Esimai, University of Texas–Arlington; Leila Farivar, The Ohio State University; Homi Fatemi, Santa Clara University; Abe Feinberg, California State University–Northridge; Samuel Graves, Boston College; Robert Gould, UCLA; Darren Grant, Sam Houston State University; Shane Griffith, Lee University; Paul Hagstrom, Hamilton College; John Hebert, Virginia Tech; James Hightower, California State University, Fullerton; Bo Honore, Princeton University; Ira Horowitz, University of Florida; Onisforos Iordanou, Hunter College; Torsten Jochem, University of Pittsburgh; Gordon Johnson, California State University–Northridge; Hilke Kayser, Hamilton College; Kenneth Klassen, California State University–Northridge; Roger Kleckner, Bowling Green State University–Firelands; Eylem Koca, Fairleigh Dickinson University; Harry Kypraios, Rollins College; John Lawrence, California State University–Fullerton; Tae H. Lee, University of California–Riverside; Dennis Lin, Pennsylvania State University; Jialu Liu, Allegheny College; Chung-Ping Loh, University of North Florida; Neal Long, Stetson University; Jayashree Mahajan, University of Florida; 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; Robert G. Morris, University of Texas–Dallas; Patricia Mullins, University of Wisconsin; Adam Munson, University of Florida; David Murphy, Boston College; 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; Bruce Pietrykowski, University of Michigan–Dearborn; 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; Yaron Raviv, Claremont McKenna College; Jason Reed, Wayne State University; Phil Roth, Clemson University; Deb Rumsey, The Ohio State University; Farhad Saboori, Albright College; Don St. Jean, George Brown College; Hedayeh Samavati, Indiana–Purdue University; Sandy Shroeder, Ohio Northern University; Chris Silvia, University of Kansas; 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; Gordon M. Stringer, University of Colorado–Colorado Springs; Arnold Stromberg, University of Kentucky; Pandu Tadikamalla, University of Pittsburgh; Patrick Thompson, University of Florida; Steve Thorpe, University of Northern Iowa; Sheldon Vernon, Houston Baptist University; John J. Wiorkowski, University of Texas–Dallas; and W. F. Younkin, University of Miami.

1

iStockphoto.com/leluconcepts

WHAT IS STATISTICS?

CHAPTER OUTLINE

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)

Students enrolled in a business program are attending their first class of the required statistics course. The students are somewhat apprehensive because they believe the myth that the course is difficult. To alleviate their anxiety, the professor provides a list of the final marks, which are composed of term work plus the final exam. What information can students obtain from the list?

This is a typical statistics problem. The students have the data (marks) and need to apply statistical techniques to get the information they require. 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 that students were told that the average mark last year was 67. Is this enough information to reduce their anxiety? Students would likely respond "No" because they would like to know whether most of the marks were close to 67 or were scattered far below and above the average. They need 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 students must determine more about the marks. In particular, they need 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 individuals who work with the mathematics of statistics. Their work involves research that develops techniques and concepts, which 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, the student 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.

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 Cengage's website plus hundreds of data sets 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 U.S. 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. Since then, this experience has likely been 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.

1-1a 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.

1-1b 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.

1-1c 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 provide readers with some background of these applications by describing their functions in two ways.

1-2a Application Sections and Subsections

We feature five sections that describe statistical applications in the functional areas of business. In Section 4-5, we discuss an application in finance, the market model, which introduces an important concept in investing. Section 7-3 describes another application in finance that describes a financial analyst's use of probability and statistics to construct portfolios that decrease risk. Section 12-4 is an application in marketing, market segmentation. In Section 14-6, we present an application in operations management, finding and reducing variation. In Section 18-3, we provide an application in human resources, pay equity. A subsection in Section 6-4 presents an application in medical testing (useful in the medical insurance industry).

1-2b 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 Section 4-1, we discuss the geometric mean and why it is used instead of the arithmetic mean to measure variables that are rates of change.

1-3 Large Real Data Sets

The author believes that you learn statistics by doing statistics. For their lives after college and university, we expect graduates to have access to large amounts of real data that must be summarized to acquire the information needed to make decisions. We include the data from two sources: the General Social Survey (GSS) and the Survey of Consumer Finances (SCF). We have scattered examples, exercises, and cases for these surveys throughout the book.

1-3a General Social Survey

Since 1972, the GSS has been tracking American attitudes on a wide variety of topics. With the exception of the U.S. Census, the GSS is the most frequently used source of information about American society. The surveys are conducted every second year and feature hundreds of variables and thousands of observations. The data for the 10 most recent surveys are stored in files GSS2000, GSS2002, GSS2004, GSS2006, GSS2008, GSS2010, GSS2012, GSS2014, GSS2016, and GSS2018. The sample sizes are 2,817, 2,765, 2,812, 4,510, 2,023, 2,044, 1,974, 2,538, 2,868, and 2,348, respectively. We downloaded the variables that we think would be of interest to students of business and economics only. We removed the missing data codes representing "No answer," and "Don't know," for most variables and replaced them with blanks.

A list of all the variables and their definitions is available as an online appendix.

1-3b Survey of Consumer Finances

The SCF is conducted every 3 years to provide detailed information on the finances of U.S. households. The study is sponsored by the Federal Reserve Board in cooperation with the Department of the Treasury. Since 1992, data have been collected by the National Opinion Research Center (NORC) at the University of Chicago. The data for the five most recent surveys are stored in folders SCF2007, SCF2010, SCF2013, SCF2016, and SCF2019. The sample sizes are 4,417, 6,482, 6,015, 6,248, and 5,777, respectively. As we did with the General Social Surveys, we downloaded only a fraction of the variables in the original surveys. Because the samples are so large and the range of some of the variables so wide, there are problems summarizing and describing the data. To solve the problem, we have created subsamples based on percentiles of the net worth of the households being sampled. Here is a list of the subsamples.

L20: Lowest 20%

Lower Middle Class (LMC): 20%–40%

Middle Class (MC): 40%–60%

Upper Middle Class (UMC): 60%–80%

Upper Class (UC): 80%–90%

Wealthy (W): 90%–95%

Super Rich (SR): 95%–99%

T1: Top 1%

A complete list of the variables and their definitions is available as an online appendix.

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 in the belief that virtually all university graduates use it now and will in the future.

Additionally, we have included chapter appendixes (for Chapters 2, 3, 4, 10, 11, 12, 13, 14, 15, 16, 17, and 19) displaying output and step-by-step instructions for two popular statistical software packages, XLSTAT and Stata. This will allow instructors to use any one of Excel, XLSTAT, and Stata without requiring students to acquire instruction manuals.

1-4a Excel

Excel can perform statistical procedures in several ways.

1. **Statistical** (which includes probability) and other functions ***fx*****:** We use some of these functions to draw graphs and charts in Chapter 2, calculate statistics in Chapters 4 and 15, and to compute probabilities in Chapters 7 and 8.
2. **Analysis ToolPak:** This group of procedures comes with every version of Excel. The techniques are accessed by clicking Data and Data Analysis. One of its drawbacks is that it does not offer a complete set of the statistical techniques we introduce in this book. The methods not included with Data Analysis will be performed by Excel spreadsheets and Do It Yourself Excel.
3. **Spreadsheets:** We use statistical functions to create spreadsheets that calculate statistical inference methods in Chapters 10–16 and 19. These can be downloaded from Cengage's website. Additionally, the spreadsheets can be used to conduct what–if analyses. The rationale for their use is described in subsection 1-4d.
4. **Do It Yourself:** We provide step-by-step instructions on how to use Excel to perform the remaining inference methods.

1-4b File Names and Notation

A large proportion of the examples, exercises, and cases feature large data sets. These are denoted with the file name next to the exercise number. The data sets associated with examples are denoted as Xm. To illustrate, the data for Example 2.2 are stored in file Xm02-02 in the Chapter 2 folder. The data for exercises and cases are stored in files prefixed by Xr and C, respectively. The prefixes GSS and SCF designate data from the General Social Surveys and Surveys of Consumer Finances, respectively.

In many real applications of statistics, additional data are collected. For instance, in Example 12.5, the pollster often records the gender and asks for other information including race, religion, education, and income. 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 a plus sign on the file name.

1-4c Our Approach

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.

1-4d Excel 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 readers 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.)

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 the politician.
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 Xr01-08 The owner of a large fleet of taxis is trying to estimate costs for next year's operations. One major cost is fuel purchase. To estimate fuel purchase, the owner needs to know the total distance the 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 the fleet's taxis to operate on propane. The propane mileage (in miles per gallon) for 50 taxis has been measured and recorded.

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 / MATERIAL TO DOWNLOAD

Additional Instructor and Student Resources

Additional instructor and student resources for this product are available online. Instructor assets include an Instructor's Manual, Educator's Guide, PowerPoint® slides, and a test bank powered by Cognero®. Student assets include data sets and Excel workbooks. Sign up or sign in at www.cengage.com to search for and access this product and its online resources.

2

© Steve Cole/Digital Vision/Getty Images

GRAPHICAL DESCRIPTIVE TECHNIQUES I

CHAPTER OUTLINE

Do Male and Female Americans Differ in Their Political Party Affiliation?

DATA
GSS2018

iStockPhoto/hermosawave

On page 40, we will provide our answer.

In Chapter 1, we introduced the General Social Survey (GSS), which is conducted every 2 years with the objective to track the experiences, behaviors, and attitudes of Americans. One question that has been asked in all General Social Surveys is "Generally speaking, do you think of yourself as Republican, Democrat, Independent, or what?" The responses are as follows:

0. Strong Democrat
1. Not Strong Democrat
2. Independent, Near Democrat

3. Independent
4. Independent, Near Republican
5. Not Strong Republican
6. Strong Republican
7. Other Party

Respondents are also identified by sex: 1 = Male, and 2 = Female. The data are stored in the file GSS2018. Note that the file contains other variables that are not needed in this example. The variable SEX is stored in column B and PARTYID is stored in Column AG. Some of the data are listed here.

ID	SEX	PARTYID
62467	1	5
62468	2	2
62469	1	4
⋮	⋮	⋮
64812	2	3
64813	1	4
64814	2	3

Determine whether American males and females differ in their political affiliations. See page 40 for our solution.

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 customers when they 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	Tesla
5	Oracle

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 Tesla? 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.

2-1a 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 14 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*.

2-1b 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 89 on an accounting course exam gives far more information about the performance of that student than does a letter grade of B, 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.

2-1c 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.

2-1d 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.

2.11 A survey of taxpayers who complete their own tax returns were asked the following questions. Determine the type of data each question produces.
a. Did you use software?
b. How long did it take you to complete this year's return?
c. Rate the ease with which you completed this year's return (very easy, quite easy, neither easy or difficult, quite difficult, very difficult)

2.12 A random sample of car owners was asked these questions. Identify the type of data.
a. Make of car
b. Age of your car in months
c. Amount of annual insurance
d. Number of miles on odometer

In many surveys, respondents are asked to report variables such as age, income, and education. It is often the case that respondents are reluctant to provide the actual amount (e.g., income). Surveyors have found that providing a list of intervals makes it more likely that the respondent will provide that information. Exercises 2.13 to 2.19 show two ways to record the variable. Briefly describe why the second variable listed provides less information than the first one listed.

2.13 a. Age of adult respondent
b. Age categories

1. $21 < x \leq 40$
2. $41 < x \leq 60$
3. More than 60

2.14 a. Income
b. Income categories

1. $x \leq \$25,000$
2. $\$25,000 < x \leq \$50,000$
3. $\$50,000 < x \leq \$75,000$
4. More than $75,000

2.15 a. How stressful do you find your job? Select any number between 0 (Not at all stressful) and 100 (Extremely stressful)
b. Categories

1. Not at all stressful
2. Somewhat stressful
3. Very stressful
4. Extremely stressful

2.16 a. Years of education
b. Education category
1. No high school diploma
2. High school diploma
3. Some college
4. College degree

2.17 a. Mark on a Statistics course
b. Grades: A, B, C, D, F

2.18 a. Assets
b. Asset categories
1. $x \leq \$100,000$
2. $\$100,000 < x \leq \$500,000$
3. $\$500,000 < x \leq \$1,000,000$
4. More than $1,000,000

2.19 How likely is it that you will vote in the upcoming election?
a. Select any number between 0 (Will not vote) and 100 (Will definitely vote)
b. Categories
1. Not at all likely
2. Somewhat likely
3. Very likely
4. Extremely likely

2.20 Xr02-20 At WLU (a famous Canadian university), suppose that instructors mark exams out of 100 and then must convert that mark to a letter grade using the following conversions.

A: 80–100
B: 70–79.9
C: 60–69.9
D: 50–59.9
F: Less than 50

Student X's marks on each course are 81, 83, 81, 71, 72, 74, 63, 61, 51, 35.

Student Y's marks are 98, 96, 99, 79, 78, 79, 68, 69, 57, 49.

a. Calculate each student's average mark.
b. Convert the individual marks into letter grades and calculate the grade point average where A = 4, B = 3, C = 2, D = 1, and F = 0.
c. Convert the average of all 10 marks into a letter grade for each student. What have you observed?

2.21 Xr02-21 Refer to Exercise 2.20. Another pair of students achieved the following marks.

Student P: 94, 91, 88, 77, 75, 74, 66, 65, 54, 52
Student Q: 93, 91, 88, 75, 75, 73, 67, 66, 55, 51

a. Calculate the average mark and the grade point average for each student.
b. Compare the mark average and the grade point average of this pair of students with the ones in Exercise 2.20.
c. Discuss the differences between the results in Exercises 2.20 and 2.21.

2.22 Xr02-22 Refer to Exercise 2.20. At another university, the marks are converted into letter grades using the following conversions.

A+: 90–100
A: 85–89.9
A−: 80–84.9
B+: 77–79.9
B: 73–76.9
B−: 70–72.9
C+: 67–69.9
C: 63–66.9
C−: 60–62.9
D+: 57–59.9
D: 53–56.9
D−: 50–52.9
F: Less than 50

The marks of two students are listed below. For each, calculate the average mark and the grade point average using a 13-point scale where A+ = 12, A = 11, and so on until F = 0.

Student M: 90, 82, 82, 73, 71, 70, 67, 61, 53, 51

Student N: 98, 84, 84, 76, 72, 72, 69, 62, 56, 52

Describe what you have determined.

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 General Social Survey

DATA
GSS2018

A major problem with the official unemployment rate is that it excludes people who have given up trying to find a job even though they would like to find employment. In an effort to track the numbers of Americans in the various categories of work status the GSS asked "Last week were you working full time, part time, going to school, keeping house, or what?" The responses are as follows:

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 is stored in Column X. Construct a frequency and relative frequency distribution for these data and graphically summarize the data by producing a bar chart and a pie chart.

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

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 2,348 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 the relative frequency distributions are combined in Table 2.1.

There were two individuals who refused to answer resulting in a total of 2,346 observations.

As promised in Chapter 1 (and in the preface) here are the Excel outputs and instructions on how to produce the frequency and the relative frequency distributions and specifically how to get the results shown 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	1,134	48.34
Working part time	2	259	11.04
Temporarily not working	3	53	2.26
Unemployed, laid off	4	84	3.58
Retired	5	445	18.97
School	6	81	3.45
Keeping house	7	242	10.32
Other	8	48	2.05
Total		2,346	100

EXCEL

INSTRUCTIONS

(Specific commands for this example are highlighted.)

1. Type or import the data into one or more columns. (Open GSS2018)
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 X1:X2349. The criteria are the codes you want to count: (1) (2) (3) (4) (5) (6) (7) (8). For example, to count the number of 1s (Working full time), type

=COUNTIF (X1:X2349, 1)

and the frequency will appear in the active cell. Change the criteria to produce the frequency of the remaining seven categories. To produce the results below, we copied the WRKSTAT variable into Column A in a new spreadsheet. We typed the codes 1 to 8 in rows 1 to 8 in Column B and then employed COUNTIF to get the frequencies. Column C calculated the relative frequencies.

INTERPRET

Only 48.34% of respondents were working full time, 18.97% were retired, 11.04% were working part time, 10.32% were keeping house, and the remaining 11.34% were in one of 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**

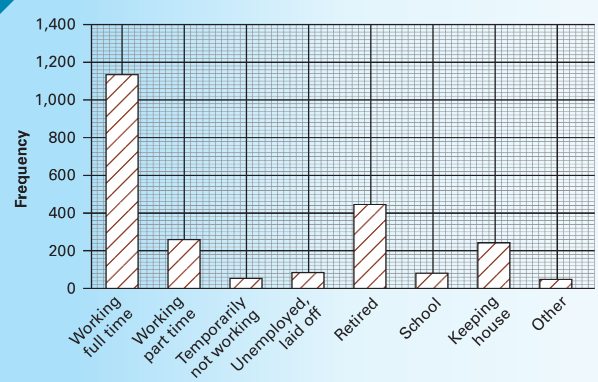
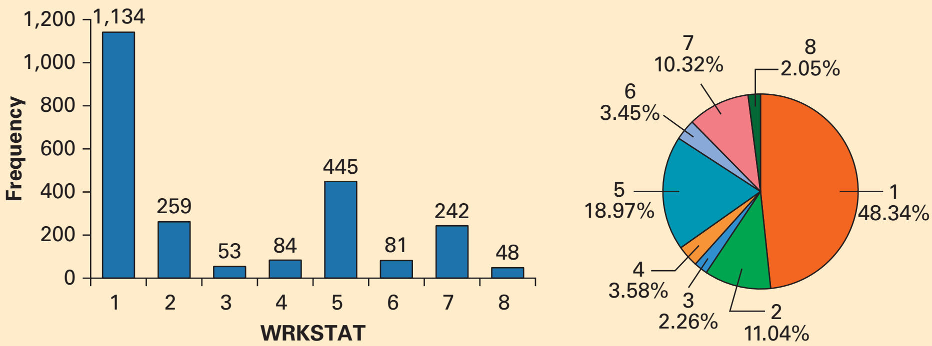


TABLE **2.2** **Proportions in Each Category in Example 2.1**

WORK STATUS	RELATIVE FREQUENCY (%)	SLICE OF THE PIE (DEGREES)
Working full time	48.34	174.0
Working part time	11.04	39.7
Temporarily not working	2.26	8.1
Unemployed, laid off	3.58	12.9
Retired	18.97	68.3
School	3.45	12.4
Keeping house	10.32	37.1
Other	2.05	7.4
Total	100.0	360

EXCEL

Here is the Excel version of the bar chart and pie chart.



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**. You can make changes to the chart. We removed the **Gridlines**, **Legend**, and clicked the **Data Labels** to create the titles.
3. For a pie chart, click **Pie** and **Chart Tools** to edit the graph.

Working with the General Social Survey and the Survey of Consumer Finances

There are hundreds of exercises scattered throughout this book that use the data in the General Social Survey and the Survey of Consumer Finances. Here are some hints on how to work with these data.

1. After you have downloaded all the data sets we recommend that you store them all onto your computer. Make no changes to these files.
2. To work with one or more columns we suggest that you copy the column or columns into a new spreadsheet. For example, in this example we need Column X (WRKSTAT) in the GSS2018 file. Copy the entire column into Column A of another worksheet. Then use the new spreadsheet to conduct any graphical or numerical technique.
3. To take a subset of any column of data use the DATA and SORT commands. Suppose that we are interested in the work status of respondents

who have completed a graduate degree (DEGREE: 4 = graduate). Proceed as follows.

Copy Column T (DEGREE) into Column A of a new worksheet.

Copy Column X (WRKSTAT) into Column B.

Highlight both columns.

Click **DATA** and **SORT**.

Specify **Sort by** Column A (or use the name of the variable).

Scroll down Column A until you reach the rows containing 4s.

Use the data in Column B to conduct your statistical analysis.

2-2a 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.com

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 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 2019

Table 2.3 lists the total energy consumption of the United States from all sources in 2019 (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, 2019

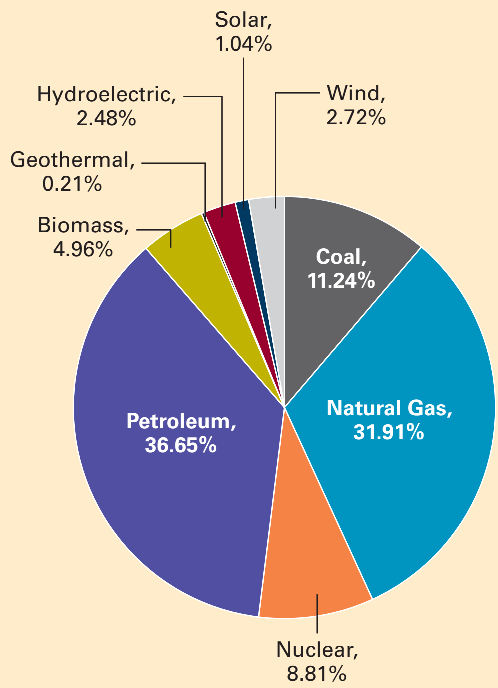
ENERGY SOURCES	QUADRILLIONS OF BTUs
Nonrenewable Energy Sources	
Coal	11.31
Natural Gas	32.10
Nuclear	8.862
Petroleum	36.87
Renewable Energy Sources	
Biomass	4.985
Geothermal	.2093
Hydroelectric	2.492
Solar	1.043
Wind	2.732
Total	100.6

Source: U.S. Energy Information Administration

SOLUTION:

We are interested in describing the proportion of total energy consumption from each source. Thus, the appropriate method is the pie chart. The next step is to determine the proportion and sizes of the pie slices from which the Excel pie chart is drawn.

EXCEL Chart



INTERPRET

The United States depends heavily on the three fossil fuels. Coal, natural gas, and petroleum constitute 80% of U.S. consumption. The renewables amount to about 11%, of which about half is biomass (wood, garbage, and crop waste) and a quarter is hydroelectric (mostly from dams). Wind and solar produce a little less than 4% of U.S. consumption. To see if and how the mix has changed since 1960, 1980, and 2000 go to Exercise 2.34.

EXAMPLE 2.3

DATA Xm02-03

Beer Consumption (Top 20 Countries)

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

TABLE **2.4** **Beer Consumption, Top 20 Countries**

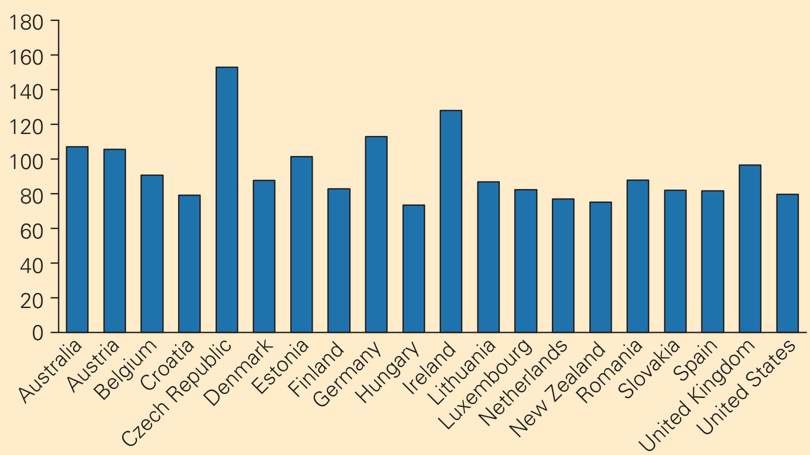
COUNTRY	PER CAPITA BEER CONSUMPTION (LITERS/YEAR)
Australia	109.9
Austria	108.3
Belgium	93.0
Croatia	81.2
Czech Republic	156.9
Denmark	89.9
Estonia	104.0
Finland	85.0
Germany	115.8
Hungary	75.3
Ireland	131.3
Lithuania	89.0
Luxembourg	84.5
Netherlands	79.0
New Zealand	77.0
Romania	90.0
Slovakia	84.1
Spain	83.8
United Kingdom	99.0
United States	81.6

Source: www.beerinfo.com.

SOLUTION:

In this example, we are primarily interested in the numbers. There is no use in presenting proportions here. The following is Excel version of the bar chart.

EXCEL Chart



INTERPRET

The Czech Republic, Ireland, and Germany head the list. Both the United States and the United Kingdom rank far lower, and Canada did not make it into the top 20 list.

2-2b 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, and Bar and Pie Charts

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

Exercises

2.23 Xr02-23 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 reserves of the top 15 countries.

a. Use a bar chart to describe these figures. Describe your findings.
b. Why is a bar chart a more appropriate graphical technique than a pie chart?
c. What other figures are needed to make a pie chart a better choice?

Country	Oil Reserves (Millions of Barrels)
Brazil	12,999
Canada	167,896
China	25,620
Iran	155,600
Iraq	145,019
Kazakhstan	30,000
Kuwait	101,500
Libya	48,363
Nigeria	36,972
Qatar	25,244
Russia	80,000
Saudi Arabia	267,026
United Arab Emirates	97,800
United States	47,053
Venezuela	302,809

2.24 Xr02-24 The table on the next page lists the number of medals won by the top 15 countries in the 2016 Summer Olympic Games.

a. Use a bar chart to summarize these figures.
b. Would a pie chart be a better graphical technique?
c. What other figure or figures would be necessary to have in order for a pie chart to be a better method?