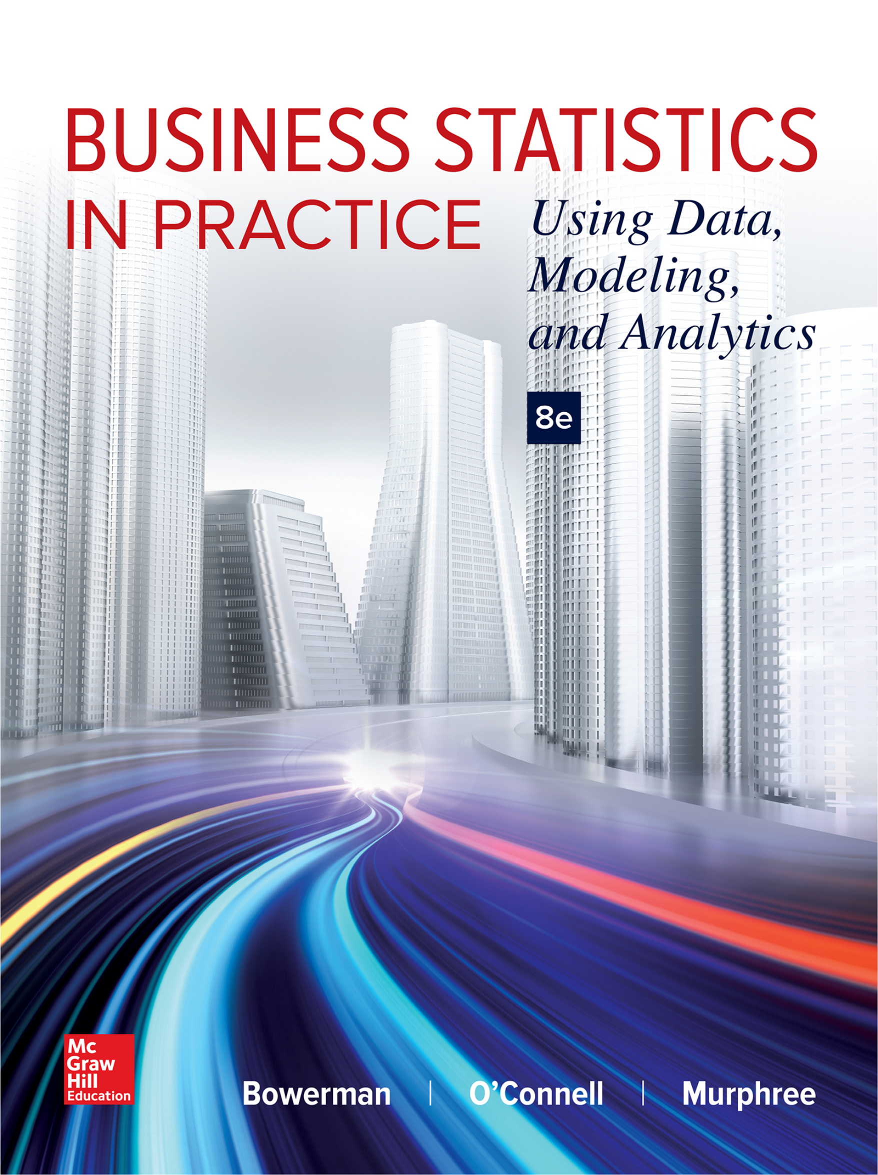


BUSINESS STATISTICS IN PRACTICE

Using Data, Modeling, and Analytics

8e

Bowerman | O'Connell | Murphree

Mc
Graw
Hill
Education

Bruce L. Bowerman
Miami University

Richard T. O'Connell
Miami University

Emily S. Murphree
Miami University

Business Statistics in Practice

Using Modeling, Data, and Analytics

EIGHTH EDITION

with major contributions by

Steven C. Huchendorf
University of Minnesota

Dawn C. Porter
University of Southern California

Patrick J. Schur
Miami University

Mc
Graw
Hill
Education

BUSINESS STATISTICS IN PRACTICE: USING DATA, MODELING, AND ANALYTICS, EIGHTH EDITION

Published by McGraw-Hill Education, 2 Penn Plaza, New York, NY 10121.

Some ancillaries, including electronic and print components, may not be available to customers outside the United States.

This book is printed on acid-free paper.

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

ISBN 978-1-259-54946-5
MHID 1-259-54946-1

Senior Vice President, Products & Markets: *Kurt L. Strand*
Vice President, General Manager, Products & Markets: *Marty Lange*
Vice President, Content Design & Delivery: *Kimberly Meriwether David*
Managing Director: *James Heine*
Senior Brand Manager: *Dolly Womack*
Director, Product Development: *Rose Koos*
Product Developer: *Camille Corum*
Marketing Manager: *Britney Hermsen*
Director of Digital Content: *Doug Ruby*
Digital Product Developer: *Tobi Philips*
Director, Content Design & Delivery: *Linda Avenarius*
Program Manager: *Mark Christianson*
Content Project Managers: *Harvey Yep (Core) / Bruce Gin (Digital)*
Buyer: *Laura M. Fuller*
Design: *Srdjan Savanovic*
Content Licensing Specialists: *Ann Marie Jannette (Image) / Beth Thole (Text)*
Cover Image: ©*Sergei Popov, Getty Images* and ©*teekid, Getty Images*
Compositor: *MPS Limited*
Printer: *R. R. Donnelley*

All credits appearing on page or at the end of the book are considered to be an extension of the copyright page.

Library of Congress Control Number: 2015956482

The Internet addresses listed in the text were accurate at the time of publication. The inclusion of a website does not indicate an endorsement by the authors or McGraw-Hill Education, and McGraw-Hill Education does not guarantee the accuracy of the information presented at these sites.

www.mhhe.com

ABOUT THE AUTHORS

Bruce L. Bowerman Bruce L. Bowerman is emeritus professor of information systems and analytics at Miami University in Oxford, Ohio. He received his Ph.D. degree in statistics from Iowa State University in 1974, and he has over 40 years of experience teaching basic statistics, regression analysis, time series forecasting, survey sampling, and design of experiments to both undergraduate and graduate students. In 1987 Professor Bowerman received an Outstanding Teaching award from the Miami University senior class, and in 1992 he received an Effective Educator award from the Richard T. Farmer School of Business Administration. Together with Richard T. O'Connell, Professor Bowerman has written 23 textbooks. These include *Forecasting, Time Series, and Regression: An Applied Approach* (also coauthored with Anne B. Koehler); *Linear Statistical Models: An Applied Approach*; *Regression Analysis: Unified Concepts, Practical Applications, and Computer Implementation* (also coauthored with Emily S. Murphree); and *Experimental Design: Unified Concepts, Practical Applications, and Computer Implementation* (also coauthored with Emily S. Murphree). The first edition of *Forecasting and Time Series* earned an Outstanding Academic Book award from *Choice* magazine. Professor Bowerman has also published a number of articles in applied stochastic process, time series forecasting, and statistical education. In his spare time, Professor Bowerman enjoys watching movies and sports, playing tennis, and designing houses.

Richard T. O'Connell Richard T. O'Connell is emeritus professor of information systems and analytics at Miami University in Oxford, Ohio. He has more than 35 years of experience teaching basic statistics, statistical quality control and process improvement, regression analysis, time series forecasting, and design of experiments to both undergraduate and graduate business students. He also has extensive consulting experience and has taught workshops dealing with statistical process control and process improvement for a variety of companies in the Midwest. In 2000 Professor O'Connell received an Effective Educator award from the Richard T. Farmer School of Business Administration. Together with Bruce L. Bowerman, he has written 23 textbooks. These include *Forecasting, Time Series, and Regression: An Applied Approach* (also coauthored with Anne B. Koehler); *Linear Statistical Models: An Applied Approach*; *Regression Analysis: Unified Concepts, Practical Applications, and Computer Implementation* (also coauthored with Emily S. Murphree); and *Experimental Design: Unified Concepts, Practical Applications, and Computer Implementation* (also coauthored with Emily S. Murphree). Professor O'Connell has published a number of articles in the area of innovative statistical education. He is one of the first college instructors in the United States to integrate statistical process control and process improvement methodology into his basic business statistics course. He (with Professor Bowerman) has written several articles advocating this approach. He has also given presentations on this subject at meetings such as the Joint Statistical Meetings of the American Statistical Association and the Workshop on Total Quality Management: Developing Curricula and Research Agendas (sponsored by the Production and Operations Management Society). Professor O'Connell received an M.S. degree in decision sciences from Northwestern University in 1973. In his spare time, Professor O'Connell enjoys fishing, collecting 1950s and 1960s rock music, and following the Green Bay Packers and Purdue University sports.

Emily S. Murphree Emily S. Murphree is emerita professor of statistics at Miami University in Oxford, Ohio. She received her Ph.D. degree in statistics from the University of North Carolina and does research in applied probability. Professor Murphree received Miami's College of Arts and Science Distinguished Educator Award in 1998. In 1996, she was named one of Oxford's Citizens of the Year for her work with Habitat for Humanity and for organizing annual Sonia Kovalevsky Mathematical Sciences Days for area high school girls. In 2012 she was recognized as "A Teacher Who Made a Difference" by the University of Kentucky.

AUTHORS' PREVIEW

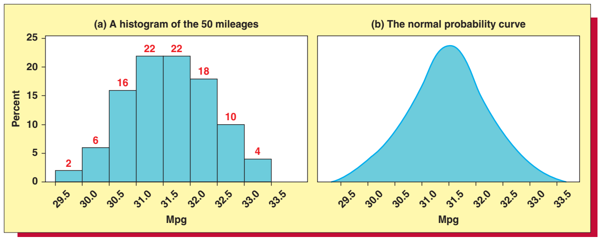
Business Statistics in Practice: Using Data, Modeling, and Analytics, Eighth Edition, provides a unique and flexible framework for teaching the introductory course in business statistics. This framework features:

- A new theme of statistical modeling introduced in Chapter 1 and used throughout the text.
- A substantial and innovative presentation of business analytics and data mining that provides instructors with a choice of different teaching options.
- Improved and easier to understand discussions of probability, probability modeling, traditional statistical inference, and regression and time series modeling.
- Continuing case studies that facilitate student learning by presenting new concepts in the context of familiar situations.
- Business improvement conclusions—highlighted in yellow and designated by icons BI in the page margins—that explicitly show how statistical analysis leads to practical business decisions.
- Many new exercises, with increased emphasis on students doing complete statistical analyses on their own.
- Use of Excel (including the Excel add-in MegaStat) and Minitab to carry out traditional statistical analysis and descriptive analytics. Use of JMP and the Excel add-in XLMiner to carry out predictive analytics.

We now discuss how these features are implemented in the book's 18 chapters.

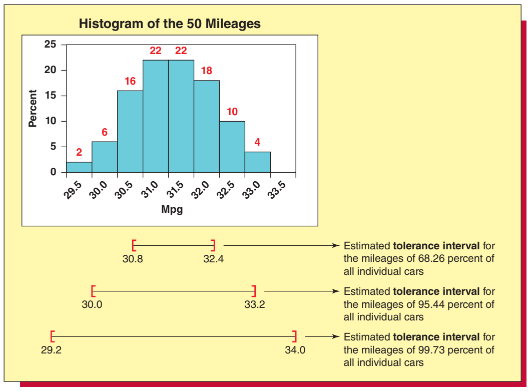
Chapters 1, 2, and 3: Introductory concepts and statistical modeling. Graphical and numerical descriptive methods. In Chapter 1 we discuss data, variables, populations, and how to select random and other types of samples (a topic formerly discussed in Chapter 7). A new section introduces statistical modeling by defining what a statistical model is and by using **The Car Mileage Case** to preview specifying a normal probability model describing the mileages obtained by a new midsize car model (see Figure 1.6):

FIGURE 1.6 A Histogram of the 50 Mileages and the Normal Probability Curve



In Chapters 2 and 3 we begin to formally discuss the statistical analysis used in statistical modeling and the statistical inferences that can be made using statistical models. For example, in Chapter 2 (graphical descriptive methods) we show how to construct the histogram of car mileages shown in Chapter 1, and in Chapter 3 (numerical descriptive methods) we use this histogram to help explain the Empirical Rule. As illustrated in Figure 3.15, this rule gives tolerance intervals providing estimates of the "lowest" and "highest" mileages that the new midsize car model should be expected to get in combined city and highway driving:

FIGURE 3.15 Estimated Tolerance Intervals in the Car Mileage Case



Chapters 1, 2, and 3: Six optional sections discussing business analytics and data mining. **The Disney Parks Case** is used in an optional section of Chapter 1 to introduce how business analytics and data mining are used to analyze big data. This case considers how Walt Disney World in Orlando, Florida, uses MagicBands worn by many of its visitors to collect massive amounts of real-time location, riding pattern, and purchase history data. These data help Disney improve visitor experiences and tailor its marketing messages to different types of visitors. At its Epcot park, Disney

FIGURE 2.36 Excel Output of a Bullet Graph of Disney's Predicted Waiting Times (in minutes) for the Seven Epcot Rides Posted at 3 P.M. on February 21, 2015 DisneyTimes

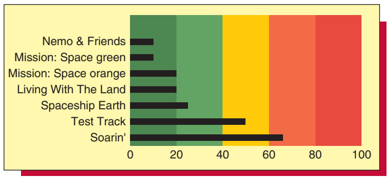


FIGURE 2.37 The Number of Ratings and the Mean Rating for Each of Seven Rides at Epcot (0 = Poor, 1 = Fair, 2 = Good, 3 = Very Good, 4 = Excellent, 5 = Superb) and an Excel Output of a Treemap of the Numbers of Ratings and the Mean Ratings

(b) Excel output of the treemap

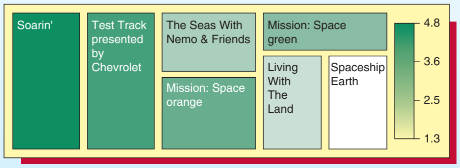
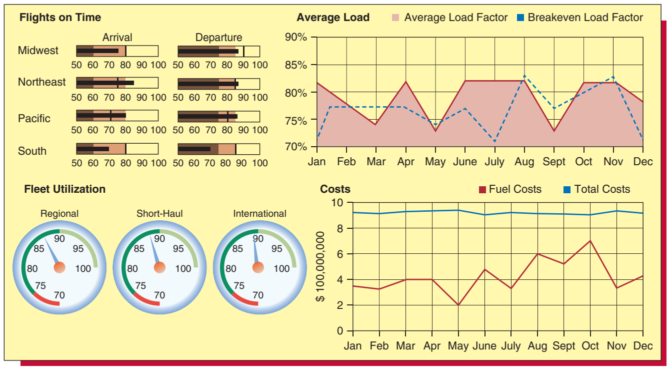


FIGURE 2.35 A Dashboard of the Key Performance Indicators for an Airline



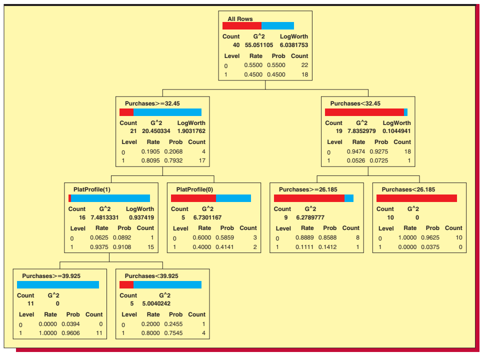
helps visitors choose their next ride by continuously summarizing predicted waiting times for seven popular rides on large screens in the park. Disney management also uses the riding pattern data it collects to make planning decisions, as is shown by the following business improvement conclusion from Chapter 1:

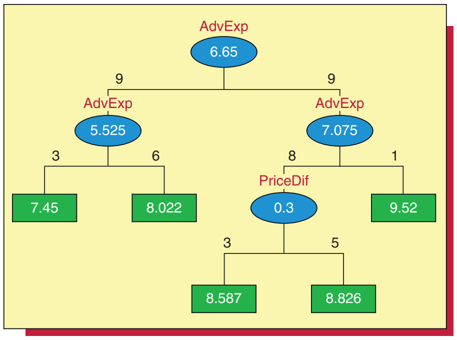
> . . . As a matter of fact, Channel 13 News in Orlando reported on March 6, 2015—during the writing of this case—that Disney had announced plans to add a third "theatre" for Soarin' (a virtual ride) in order to shorten long visitor waiting times.

The Disney Parks Case is also used in an optional section of Chapter 2 to help discuss descriptive analytics. Specifically, Figure 2.36 shows a bullet graph summarizing predicted waiting times for seven Epcot rides posted by Disney at 3 P.M. on February 21, 2015, and Figure 2.37 shows a treemap illustrating ficticious visitor ratings of the seven Epcot rides. Other graphics discussed in the optional section on descriptive analytics include gauges, sparklines, data drill-down graphics, and dashboards combining graphics illustrating a business's key performance indicators. For example, Figure 2.35 is a dashboard showing eight "flight on time" bullet graphs and three "flight utilization" gauges for an airline.

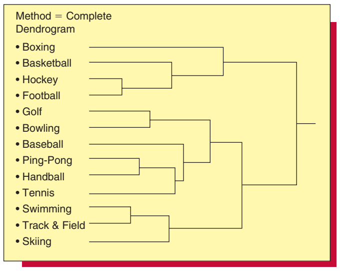
Chapter 3 contains four optional sections that discuss six methods of predictive analytics. The methods discussed are explained in an applied and practical way by using the numerical descriptive statistics previously discussed in Chapter 3. These methods are:

- Classification tree modeling and regression tree modeling (see Section 3.7 and the following figures):





- Hierarchical clustering and *k*-means clustering (see Section 3.8 and the following figures):



Sport	Cluster ID
Boxing	5
Basketball	4
Golf	2
Swimming	3
Skiing	5
Baseball	1
Ping-Pong	3
Hockey	4
Handball	3
Track & Field	3
Bowling	2
Tennis	3
Football	4

Cluster	Fast	Compl	Team	Easy	Ncon	Opp
Cluster-1	4.78	4.18	2.16	3.33	3.6	2.67
Cluster-2	5.6	4.825	5.99	3.475	1.71	3.92
Cluster-3	2.858	4.796	5.078	3.638	2.418	3.022
Cluster-4	1.99	3.253333	1.606667	4.62	5.773333	2.363333
Cluster-5	2.6	4.61	6.29	5	4.265	3.22

Cluster	#Obs	Avg. Dist
Cluster-1	1	0
Cluster-2	2	0.960547
Cluster-3	5	1.319782
Cluster-4	3	0.983933
Cluster-5	2	2.382945
Overall	13	1.249053

- Factor analysis and association rule mining (see Sections 3.9 and 3.10 and the following figures):

FIGURE 3.35 Minitab Output of a Factor Analysis of the Applicant Data (4 Factors Used)

Rotated Factor Loadings and Communalities
Varimax Rotation

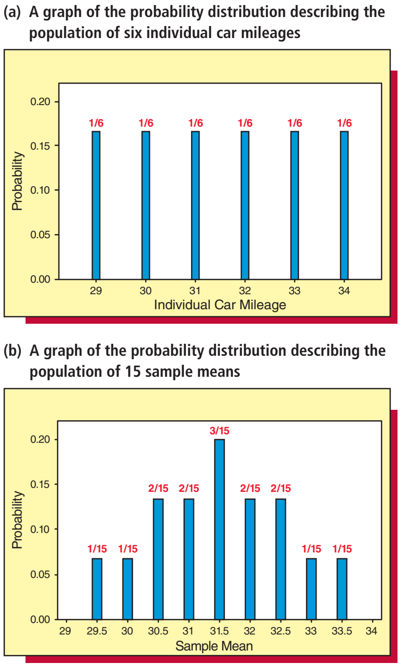
Variable	Factor1	Factor2	Factor3	Factor4	Communality
Var 1	0.114	−0.833✓	−0.111	−0.138	0.739
Var 2	0.440	−0.150	−0.394	0.226	0.422
Var 3	0.061	−0.127	−0.006	0.928✓	0.881
Var 4	0.216	−0.247	−0.874✓	−0.081	0.877
Var 5	0.919✓	0.104	−0.162	−0.062	0.885
Var 6	0.864✓	−0.102	−0.259	0.006	0.825
Var 7	0.217	0.246	−0.864✓	0.003	0.855
Var 8	0.918✓	−0.206	−0.088	−0.049	0.895
Var 9	0.085	−0.849✓	0.055	0.219	0.779
Var 10	0.796✓	−0.354	−0.160	−0.050	0.787
Var 11	0.916✓	−0.163	−0.105	−0.042	0.879
Var 12	0.804✓	−0.259	−0.340	0.152	0.852
Var 13	0.739✓	−0.329	−0.425	0.230	0.888
Var 14	0.436	−0.364	−0.541	−0.519	0.884
Var 15	0.379	−0.798✓	−0.078	0.082	0.794
Variance	5.7455	2.7351	2.4140	1.3478	12.2423
% Var	0.383	0.182	0.161	0.090	0.816

Row ID	Confidence%	Antecedent (*x*)	Consequent (*y*)	Lift Ratio
1	71.42857143	B	A	1.020408163
2	71.42857143	A	B	1.020408163
3	85.71428571	A	C	0.952380952
4	77.77777778	C	B	1.111111111
5	100	B	C	1.111111111
6	71.42857143	B & C	A	1.020408163
7	83.33333333	A & C	B	1.19047619
8	100	A & B	C	1.111111111
9	71.42857143	B	A & C	1.19047619
10	71.42857143	A	B & C	1.020408163
11	83.33333333	E	C	0.925925926
12	80	C & E	B	1.142857143
13	100	B & E	C	1.111111111

We believe that an early introduction to predictive analytics (in Chapter 3) will make statistics seem more useful and relevant from the beginning and thus motivate students to be more interested in the entire course. However, our presentation gives instructors various choices. This is because, after covering the optional introduction to business analytics in Chapter 1, the five optional sections on descriptive and predictive analytics in Chapters 2 and 3 can be covered in any order without loss of continuity. Therefore, the instructor can choose which of the six optional business analytics sections to cover early, as part of the main flow of Chapters 1–3, and which to discuss later. We recommend that sections chosen to be discussed later be covered after Chapter 14, which presents the further predictive analytics topics of multiple linear regression, logistic regression, and neural networks.

Chapters 4–8: Probability and probability modeling. Discrete and continuous probability distributions. Sampling distributions and confidence intervals. Chapter 4 discusses probability by featuring a new discussion of probability modeling and using motivating examples—The Crystal Cable Case and a real-world example of gender discrimination at a pharmaceutical company—to illustrate the probability rules. Chapters 5 and 6 give more concise discussions of discrete and continuous probability distributions (models) and feature practical examples illustrating the "rare event approach" to making a statistical inference. In Chapter 7, The Car Mileage Case is used to introduce sampling distributions and motivate the Central Limit Theorem (see Figures 7.1, 7.3, and 7.5). In Chapter 8, the automaker in The Car Mileage Case uses a confidence interval procedure specified by the Environmental Protection Agency (EPA) to find the EPA estimate of a new midsize model's true mean mileage and determine if the new midsize model deserves a federal tax credit (see Figure 8.2).

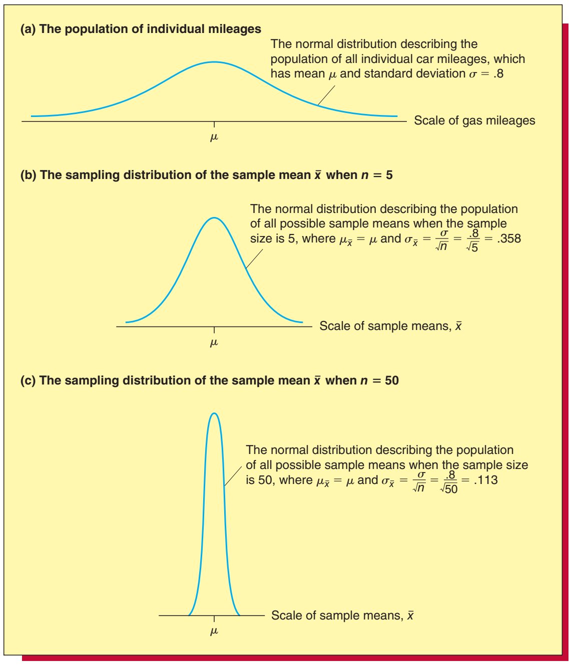
FIGURE 7.1 A Comparison of Individual Car Mileages and Sample Means



(a) A graph of the probability distribution describing the population of six individual car mileages

(b) A graph of the probability distribution describing the population of 15 sample means

FIGURE 7.3 A Comparison of (1) the Population of All Individual Car Mileages, (2) the Sampling Distribution of the Sample Mean $\bar{x}$ When $n = 5$, and (3) the Sampling Distribution of the Sample Mean $\bar{x}$ When $n = 50$



(a) The population of individual mileages

The normal distribution describing the population of all individual car mileages, which has mean μ and standard deviation $\sigma = .8$

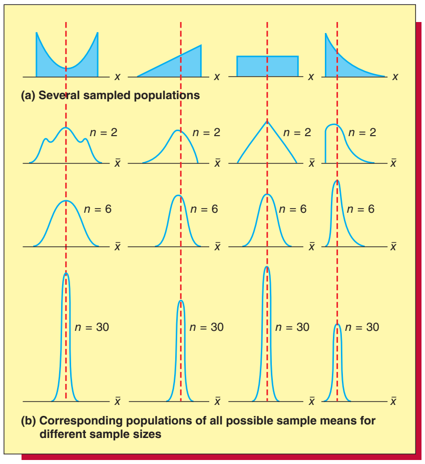
(b) The sampling distribution of the sample mean $\bar{x}$ when $n = 5$

The normal distribution describing the population of all possible sample means when the sample size is 5, where $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{.8}{\sqrt{5}} = .358$

(c) The sampling distribution of the sample mean $\bar{x}$ when $n = 50$

The normal distribution describing the population of all possible sample means when the sample size is 50, where $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{.8}{\sqrt{50}} = .113$

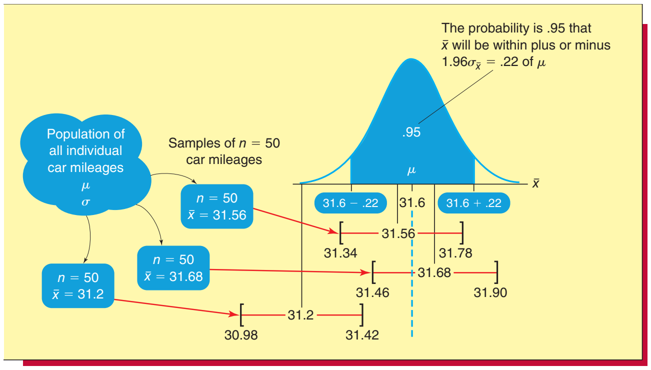
FIGURE 7.5 The Central Limit Theorem Says That the Larger the Sample Size Is, the More Nearly Normally Distributed Is the Population of All Possible Sample Means



(a) Several sampled populations

(b) Corresponding populations of all possible sample means for different sample sizes

FIGURE 8.2 Three 95 Percent Confidence Intervals for μ



The probability is .95 that $\bar{x}$ will be within plus or minus $1.96\sigma_{\bar{x}} = .22$ of μ

Chapters 9–12: Hypothesis testing. Two-sample procedures. Experimental design and analysis of variance. Chi-square tests. Chapter 9 discusses hypothesis testing and begins with a new section on formulating statistical hypotheses. Three cases—The Trash Bag Case, The e-billing Case, and The Valentine's Day Chocolate Case—are then used in a new section that explains the critical value and *p*-value approaches to testing a hypothesis about a population mean. A summary box visually illustrating these approaches is presented in the middle of this section (rather than at the end, as in previous editions) so that more of the section can be devoted to developing the summary box and showing how to use it. In addition, a five-step hypothesis testing procedure emphasizes that successfully using any of the book's hypothesis testing summary boxes requires simply identifying the alternative hypothesis being tested and then looking in the summary box for the corresponding critical value rule and/or *p*-value (see the next page).

The Five Steps of Hypothesis Testing

1. State the null hypothesis H_0 and the alternative hypothesis H_a.
2. Specify the level of significance α.
3. Plan the sampling procedure and select the test statistic.

Using a critical value rule:

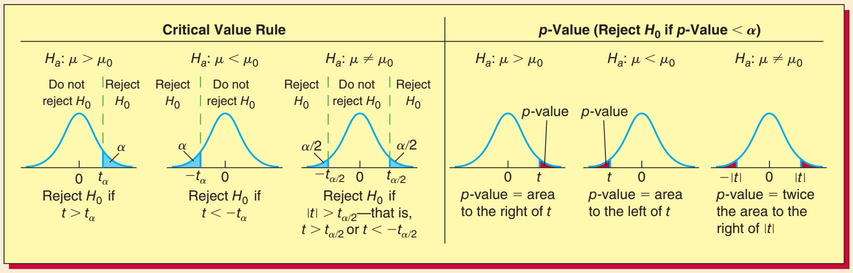
4. Use the summary box to find the critical value rule corresponding to the alternative hypothesis.
5. Collect the sample data, compute the value of the test statistic, and decide whether to reject H_0 by using the critical value rule. Interpret the statistical results.

Using a *p*-value rule:

4. Use the summary box to find the *p*-value corresponding to the alternative hypothesis. Collect the sample data, compute the value of the test statistic, and compute the *p*-value.
5. Reject H_0 at level of significance α if the *p*-value is less than α. Interpret the statistical results.

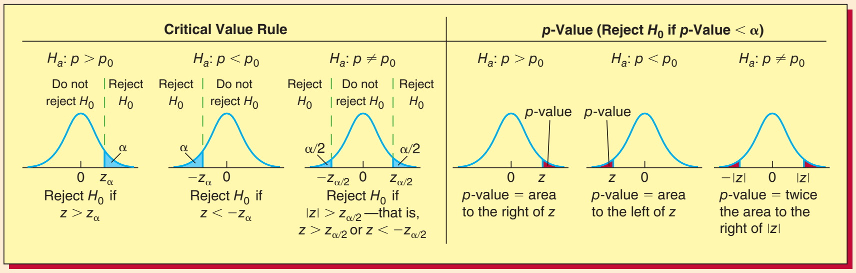
A *t* Test about a Population Mean: σ Unknown

Null Hypothesis $H_0\!: \mu = \mu_0$ **Test Statistic** $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ $df = n - 1$ **Assumptions** Normal population or Large sample size



A Large Sample Test about a Population Proportion

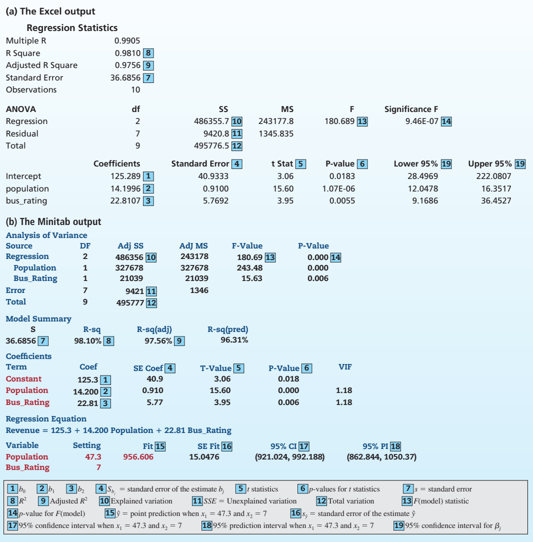
Null Hypothesis $H_0\!: p = p_0$ **Test Statistic** $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$ **Assumptions**[3] $np_0 \geq 5$ and $n(1 - p_0) \geq 5$



Hypothesis testing summary boxes are featured throughout Chapter 9, Chapter 10 (two-sample procedures), Chapter 11 (one-way, randomized block, and two-way analysis of variance), Chapter 12 (chi-square tests of goodness of fit and independence), and the remainder of the book. In addition, emphasis is placed throughout on estimating practical importance after testing for statistical significance.

Chapters 13–18: Simple and multiple regression analysis. Model building. Logistic regression and neural networks. Time series forecasting. Control charts. Nonparametric statistics. Decision theory. Chapters 13–15 present predictive analytics methods that are based on parametric regression and time series models. Specifically, Chapter 13 and the first seven sections of Chapter 14 discuss simple and basic multiple regression analysis by using a more streamlined organization and The Tasty Sub Shop (revenue prediction) Case (see Figure 14.4). The next five sections of Chapter 14 present five advanced modeling topics that can be covered in any order without loss of continuity: dummy variables (including a discussion of interaction); quadratic variables and quantitative interaction variables; model building and the effects of multicollinearity; residual analysis and diagnosing outlying and influential observations; and logistic regression (see Figure 14.36). The last section of Chapter 14 discusses neural networks and has logistic regression as a prerequisite. This section shows why neural network modeling is particularly useful when analyzing big data and how neural network models are used to make predictions (see Figures 14.37 and 14.38). Chapter 15 discusses time series forecasting, including Holt–Winters' exponential smoothing models, and refers readers to Appendix B (at the end of the book), which succinctly discusses the Box–Jenkins methodology. The book concludes with Chapter 16 (a clear discussion of control charts and process capability), Chapter 17 (nonparametric statistics), and Chapter 18 (decision theory, another useful predictive analytics topic).

FIGURE 14.4 Excel and Minitab Outputs of a Regression Analysis of the Tasty Sub Shop Revenue Data in Table 14.1 Using the Model $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \varepsilon$



(a) The Excel output

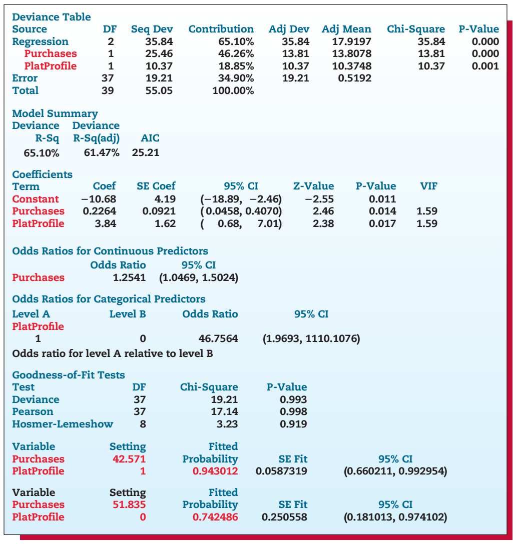
```
        Regression Statistics
Multiple R              0.9905
R Square                0.9810 [8]
Adjusted R Square       0.9756 [9]
Standard Error         36.6856 [7]
Observations                10

ANOVA                 df            SS            MS             F   Significance F
Regression             2   486355.7 [10]   243177.8   180.689 [13]    9.46E-07 [14]
Residual               7     9420.8 [11]   1345.835
Total                  9   495776.5 [12]

               Coefficients   Standard Error [4]  t Stat [5]  P-value [6]  Lower 95% [19]  Upper 95% [19]
Intercept       125.289 [1]        40.9333           3.06       0.0183        28.4969        222.0807
population      14.1996 [2]         0.9100          15.60     1.07E-06        12.0478         16.3517
bus_rating      22.8107 [3]         5.7692           3.95       0.0055         9.1686         36.4527
```

(b) The Minitab output

```
Analysis of Variance
Source        DF     Adj SS       AdJ MS    F-Value      P-Value
Regression     2   486356 [10]   243178   180.69 [13]   0.000 [14]
  Population   1   327678        327678   243.48        0.000
  Bus_Rating   1    21039         21039    15.63        0.006
Error          7     9421 [11]     1346
Total          9   495777 [12]

Model Summary
      S          R-sq      R-sq(adj)    R-sq(pred)
36.6856 [7]   98.10% [8]   97.56% [9]     96.31%

Coefficients
Term          Coef      SE Coef [4]   T-Value [5]   P-Value [6]   VIF
Constant     125.3 [1]     40.9          3.06          0.018
Population  14.200 [2]    0.910         15.60          0.000       1.18
Bus_Rating   22.81 [3]     5.77          3.95          0.006       1.18

Regression Equation
Revenue = 125.3 + 14.200 Population + 22.81 Bus_Rating

Variable     Setting     Fit [15]   SE Fit [16]      95% CI [17]             95% PI [18]
Population     47.3      956.606     15.0476     (921.024, 992.188)    (862.844, 1050.37)
Bus_Rating        7
```

[1] b_0 [2] b_1 [3] b_2 [4] S_{b_j} = standard error of the estimate b_j [5] t statistics [6] p-values for t statistics [7] s = standard error
[8] R^2 [9] Adjusted R^2 [10] Explained variation [11] SSE = Unexplained variation [12] Total variation [13] F(model) statistic
[14] p-value for F(model) [15] $\hat{y}$ = point prediction when $x_1 = 47.3$ and $x_2 = 7$ [16] $s_{\hat{y}}$ = standard error of the estimate $\hat{y}$
[17] 95% confidence interval when $x_1 = 47.3$ and $x_2 = 7$ [18] 95% prediction interval when $x_1 = 47.3$ and $x_2 = 7$ [19] 95% confidence interval for β_j

FIGURE 14.36 Minitab Output of a Logistic Regression of the Credit Card Upgrade Data



```
Deviance Table
Source        DF   Seq Dev   Contribution   Adj Dev   Adj Mean   Chi-Square   P-Value
Regression     2     35.84         65.10%     35.84    17.9197        35.84     0.000
  Purchases    1     25.46         46.26%     13.81    13.8078        13.81     0.000
  PlatProfile  1     10.37         18.85%     10.37    10.3748        10.37     0.001
Error         37     19.21         34.90%     19.21     0.5192
Total         39     55.05        100.00%

Model Summary
Deviance   Deviance
    R-Sq   R-Sq(adj)    AIC
  65.10%     61.47%   25.21

Coefficients
Term            Coef   SE Coef          95% CI         Z-Value   P-Value    VIF
Constant      -10.68      4.19   (-18.89,  -2.46)        -2.55     0.011
Purchases     0.2264    0.0921   ( 0.0458, 0.4070)        2.46     0.014   1.59
PlatProfile     3.84      1.62   (   0.68,    7.01)       2.38     0.017   1.59

Odds Ratios for Continuous Predictors
              Odds Ratio        95% CI
Purchases         1.2541   (1.0469, 1.5024)

Odds Ratios for Categorical Predictors
Level A         Level B      Odds Ratio          95% CI
PlatProfile
    1                 0         46.7564    (1.9693, 1110.1076)
Odds ratio for level A relative to level B

Goodness-of-Fit Tests
Test                DF    Chi-Square    P-Value
Deviance            37         19.21      0.993
Pearson             37         17.14      0.998
Hosmer-Lemeshow      8          3.23      0.919

Variable      Setting       Fitted
Purchases      42.571    Probability     SE Fit           95% CI
PlatProfile         1       0.943012  0.0587319    (0.660211, 0.992954)

Variable      Setting       Fitted
Purchases      51.835    Probability     SE Fit           95% CI
PlatProfile         0       0.742486   0.250558    (0.181013, 0.974102)
```

FIGURE 14.37 The Single Layer Perceptron

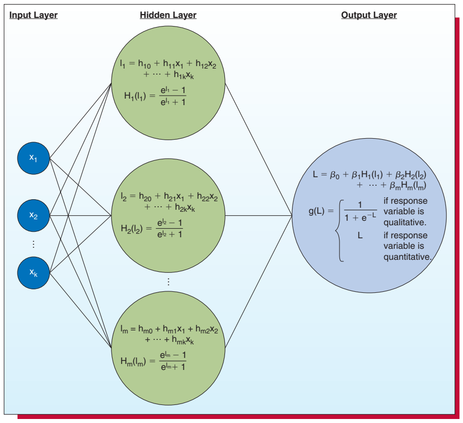
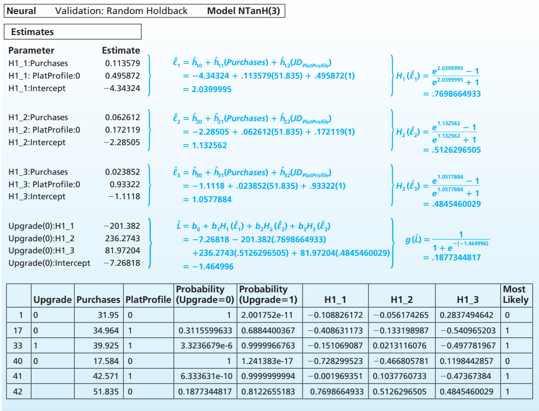

Input Layer
Hidden Layer
Output Layer
x_1
x_2
x_k
$l_1 = h_{10} + h_{11}x_1 + h_{12}x_2 + \cdots + h_{1k}x_k$
$H_1(l_1) = \frac{e^{l_1} - 1}{e^{l_1} + 1}$
$l_2 = h_{20} + h_{21}x_1 + h_{22}x_2 + \cdots + h_{2k}x_k$
$H_2(l_2) = \frac{e^{l_2} - 1}{e^{l_2} + 1}$
$l_m = h_{m0} + h_{m1}x_1 + h_{m2}x_2 + \cdots + h_{mk}x_k$
$H_m(l_m) = \frac{e^{l_m} - 1}{e^{l_m} + 1}$
$L = \beta_0 + \beta_1H_1(l_1) + \beta_2H_2(l_2) + \cdots + \beta_mH_m(l_m)$
$g(L) = \begin{cases} \frac{1}{1 + e^{-L}} & \text{if response variable is qualitative.} \\ L & \text{if response variable is quantitative.} \end{cases}$

FIGURE 14.38 JMP Output of Neural Network Estimation for the Credit Card Upgrade Data DS CardUpgrade



Neural Validation: Random Holdback Model NTanH(3)

Estimates

Parameter	Estimate
H1_1:Purchases	0.113579
H1_1: PlatProfile:0	0.495872
H1_1:Intercept	−4.34324
H1_2:Purchases	0.062612
H1_2: PlatProfile:0	0.172119
H1_2:Intercept	−2.28505
H1_3:Purchases	0.023852
H1_3: PlatProfile:0	0.93322
H1_3:Intercept	−1.1118
Upgrade(0):H1_1	−201.382
Upgrade(0):H1_2	236.2743
Upgrade(0):H1_3	81.97204
Upgrade(0):Intercept	−7.26818

$$\hat{\ell}_1 = \hat{h}_{10} + \hat{h}_{11}(Purchases) + \hat{h}_{12}(JD_{PlatProfile}) = -4.34324 + .113579(51.835) + .495872(1) = 2.0399995 \qquad H_1(\hat{\ell}_1) = \frac{e^{2.0399995} - 1}{e^{2.0399995} + 1} = .7698664933$$

$$\hat{\ell}_2 = \hat{h}_{20} + \hat{h}_{21}(Purchases) + \hat{h}_{22}(JD_{PlatProfile}) = -2.28505 + .062612(51.835) + .172119(1) = 1.132562 \qquad H_2(\hat{\ell}_2) = \frac{e^{1.132562} - 1}{e^{1.132562} + 1} = .5126296505$$

$$\hat{\ell}_3 = \hat{h}_{30} + \hat{h}_{31}(Purchases) + \hat{h}_{32}(JD_{PlatProfile}) = -1.1118 + .023852(51.835) + .93322(1) = 1.0577884 \qquad H_3(\hat{\ell}_3) = \frac{e^{1.0577884} - 1}{e^{1.0577884} + 1} = .4845460029$$

$$\hat{L} = b_0 + b_1H_1(\hat{\ell}_1) + b_2H_2(\hat{\ell}_2) + b_3H_3(\hat{\ell}_3) = -7.26818 - 201.382(.7698664933) + 236.2743(.5126296505) + 81.97204(.4845460029) = -1.464996 \qquad g(\hat{L}) = \frac{1}{1 + e^{-(-1.464996)}} = .1877344817$$

	Upgrade	Purchases	PlatProfile	Probability (Upgrade=0)	Probability (Upgrade=1)	H1_1	H1_2	H1_3	Most Likely
1	0	31.95	0	1	2.001752e-11	−0.108826172	−0.056174265	0.2837494642	0
17	0	34.964	1	0.3115599633	0.6884400367	−0.408631173	−0.133198987	−0.540965203	1
33	1	39.925	1	3.3236679e-6	0.9999966763	−0.151069087	0.0213116076	−0.497781967	1
40	0	17.584	0	1	1.241383e-17	−0.728299523	−0.466805781	0.1198442857	0
41		42.571	1	6.333631e-10	0.9999999994	−0.001969351	0.1037760733	−0.47367384	1
42		51.835	0	0.1877344817	0.8122655183	0.7698664933	0.5126296505	0.4845460029	1

WHAT SOFTWARE IS AVAILABLE

MEGASTAT® FOR MICROSOFT EXCEL® 2003, 2007, AND 2010 (AND EXCEL: MAC 2011)

MegaStat is a full-featured Excel add-in by J. B. Orris of Butler University that is available with this text. It performs statistical analyses within an Excel workbook. It does basic functions such as descriptive statistics, frequency distributions, and probability calculations, as well as hypothesis testing, ANOVA, and regression.

MegaStat output is carefully formatted. Ease-of-use features include AutoExpand for quick data selection and Auto Label detect. Since MegaStat is easy to use, students can focus on learning statistics without being distracted by the software. MegaStat is always available from Excel's main menu. Selecting a menu item pops up a dialog box. MegaStat works with all recent versions of Excel.

MINITAB®

Minitab® Student Version 17 is available to help students solve the business statistics exercises in the text. This software is available in the student version and can be packaged with any McGraw-Hill business statistics text.

TEGRITY CAMPUS: LECTURES 24/7

Tegrity Campus is a service that makes class time available 24/7. With *Tegrity Campus*, you can automatically capture every lecture in a searchable format for students to review when they study and complete assignments. With a simple one-click start-and-stop process, you capture all computer screens and corresponding audio. Students can replay any part of any class with easy-to-use browser-based viewing on a PC or Mac.

Educators know that the more students can see, hear, and experience class resources, the better they learn. In fact, studies prove it. With *Tegrity Campus*, students quickly recall key moments by using *Tegrity Campus's* unique search feature. This search helps students efficiently find what they need, when they need it, across an entire semester of class recordings. Help turn all your students' study time into learning moments immediately supported by your lecture. To learn more about *Tegrity*, watch a two-minute Flash demo at http://tegritycampus.mhhe.com.

ACKNOWLEDGMENTS

We wish to thank many people who have helped to make this book a reality. As indicated on the title page, we thank Professor Steven C. Huchendorf, University of Minnesota; Dawn C. Porter, University of Southern California; and Patrick J. Schur, Miami University; for major contributions to this book. We also thank Susan Cramer of Miami University for very helpful advice on writing this new edition.

We also wish to thank the people at McGraw-Hill for their dedication to this book. These people include senior brand manager Dolly Womack, who is extremely helpful to the authors; senior development editor Camille Corum, who has shown great dedication to the improvement of this book; content project manager Harvey Yep, who has very capably and diligently guided this book through its production and who has been a tremendous help to the authors; and our former executive editor Steve Scheutz, who always greatly supported our books. We also thank executive editor Michelle Janicek for her tremendous help in developing this new edition; our former executive editor Scott Isenberg for the tremendous help he has given us in developing all of our McGraw-Hill business statistics books; and our former executive editor Dick Hercher, who persuaded us to publish with McGraw-Hill.

We also wish to thank Sylvia Taylor and Nicoleta Maghear, Hampton University, for accuracy checking Connect content; Patrick Schur, Miami University, for developing learning resources; Ronny Richardson, Kennesaw State University, for revising the instructor PowerPoints and developing new guided examples and learning resources; Denise Krallman, Miami University, for updating the Test Bank; and James Miller, Dominican University, and Anne Drougas, Dominican University, for developing learning resources for the new business analytics content. Most importantly, we wish to thank our families for their acceptance, unconditional love, and support.

DEDICATION

Bruce L. Bowerman
To my wife, children, sister, and other family members:
Drena
Michael, Jinda, Benjamin, and Lex
Asa, Nicole, and Heather
Susan
Barney, Fiona, and Radeesa
Daphne, Chloe, and Edgar
Gwyneth and Tony
Callie, Bobby, Marmalade, Randy, and Penney
Clarence, Quincy, Teddy, Julius, Charlie, Sally, Milo, Zeke,
Bunch, Big Mo, Ozzie, Harriet, Sammy, Louise, Pat, Taylor,
and Jamie

Richard T. O'Connell
To my children and grandchildren:
Christopher, Bradley, Sam, and Joshua

Emily S. Murphree
To Kevin and the Math Ladies

CHAPTER-BY-CHAPTER REVISIONS FOR 8TH EDITION

Chapter 1

- Initial example made clearer.
- Two new graphical examples added to better introduce quantitative and qualitative variables.
- How to select random (and other types of) samples moved from Chapter 7 to Chapter 1 and combined with examples introducing statistical inference.
- New subsection on statistical modeling added.
- More on surveys and errors in surveys moved from Chapter 7 to Chapter 1.
- New optional section introducing business analytics and data mining added.
- Sixteen new exercises added.

Chapter 2

- Thirteen new data sets added for this chapter on graphical descriptive methods.
- Fourteen new exercises added.
- New optional section on descriptive analytics added.

Chapter 3

- Twelve new data sets added for this chapter on numerical descriptive methods.
- Twenty-three new exercises added.
- Four new optional sections on predictive analytics added:

  one section on classification trees and regression trees;

  one section on hierarchical clustering and *k*-means clustering;

  one section on factor analysis;

  one section on association rule mining.

Chapter 4

- New subsection on probability modeling added.
- Exercises updated in this and all subsequent chapters.

Chapter 5

- Discussion of general discrete probability distributions, the binomial distribution, the Poisson distribution, and the hypergeometric distribution simplified and shortened.

Chapter 6

- Discussion of continuous probability distributions and normal plots simplified and shortened.

Chapter 7

- This chapter covers the sampling distribution of the sample mean and the sampling distribution of the sample proportion; as stated above, the material on how to select samples and errors in surveys has been moved to Chapter 1.

Chapter 8

- No significant changes when discussing confidence intervals.

Chapter 9

- Discussion of formulating the null and alternative hypotheses completely rewritten and expanded.
- Discussion of using critical value rules and *p*-values to test a population mean completely rewritten; development of and instructions for using hypothesis testing summary boxes improved.
- Short presentation of the logic behind finding the probability of a Type II error when testing a two-sided alternative hypothesis now accompanies the general formula for calculating this probability.

Chapter 10

- Statistical inference for a single population variance and comparing two population variances moved from its own chapter (the former Chapter 11) to Chapter 10.
- More explicit examples of using hypothesis testing summary boxes when comparing means, proportions, and variances.

Chapter 11

- New exercises for one-way, randomized block, and two-way analysis of variance, with added emphasis on students doing complete statistical analyses.

Chapter 12

- No significant changes when discussing chi-square tests.

Chapter 13

- Discussion of basic simple linear regression analysis streamlined, with discussion of r^2 moved up and discussions of t and F tests combined into one section.
- Section on residual analysis significantly shortened and improved.
- New exercises, with emphasis on students doing complete statistical analyses on their own.

Chapter 14

- Discussion of R^2 moved up.
- Discussion of backward elimination added.
- New subsection on model validation and PRESS added.
- Section on logistic regression expanded.
- New section on neural networks added.
- New exercises, with emphasis on students doing complete statistical analyses on their own.

Chapter 15

- Discussion of the Box–Jenkins methodology slightly expanded and moved to Appendix B (at the end of the book).
- New time series exercises, with emphasis on students doing complete statistical analyses on their own.

Chapters 16, 17, and 18

- No significant changes. (These were the former Chapters 17, 18, and 19 on control charts, nonparametrics, and decision theory.)

BRIEF CONTENTS

CONTENTS

Chapter 5
Discrete Random Variables

Chapter 6
Continuous Random Variables

Chapter 7
Sampling Distributions

Chapter 8
Confidence Intervals

Chapter 9
Hypothesis Testing

Chapter 10
Statistical Inferences Based on Two Samples

Business Statistics in Practice
Using Modeling, Data, and Analytics

EIGHTH EDITION

CHAPTER 1

An Introduction to Business Statistics and Analytics

Learning Objectives

When you have mastered the material in this chapter, you will be able to:

LO1-1 Define a variable.

LO1-2 Describe the difference between a quantitative variable and a qualitative variable.

LO1-3 Describe the difference between cross-sectional data and time series data.

LO1-4 Construct and interpret a time series (runs) plot.

LO1-5 Identify the different types of data sources: existing data sources, experimental studies, and observational studies.

LO1-6 Explain the basic ideas of data warehousing and big data.

LO1-7 Describe the difference between a population and a sample.

LO1-8 Distinguish between descriptive statistics and statistical inference.

LO1-9 Explain the concept of random sampling and select a random sample.

LO1-10 Explain the basic concept of statistical modeling.

LO1-11 Explain some of the uses of business analytics and data mining (Optional).

LO1-12 Identify the ratio, interval, ordinal, and nominative scales of measurement (Optional).

LO1-13 Describe the basic ideas of stratified random, cluster, and systematic sampling (Optional).

LO1-14 Describe basic types of survey questions, survey procedures, and sources of error (Optional).

Chapter Outline

The subject of statistics involves the study of how to collect, analyze, and interpret data. **Data are facts and figures from which conclusions can be drawn**. Such conclusions are important to the decision making of many professions and organizations. For example, **economists** use conclusions drawn from the latest data on unemployment and inflation to help the government make policy decisions. **Financial planners** use recent trends in stock market prices and economic conditions to make investment decisions. **Accountants** use **sample data** concerning a company's *actual sales revenues* to assess whether the company's *claimed sales revenues* are valid. **Marketing professionals** and **data miners** help businesses decide which products to develop and market and which consumers to target in marketing campaigns by using data that reveal consumer preferences. **Production supervisors** use manufacturing data to evaluate, control, and improve product quality. **Politicians** rely on data from public opinion polls to formulate legislation and to devise campaign strategies. **Physicians and hospitals** use data on the effectiveness of drugs and surgical procedures to provide patients with the best possible treatment.

In this chapter we begin to see how we collect and analyze data. As we proceed through the chapter, we introduce several case studies. These case studies (and others to be introduced later) are revisited throughout later chapters as we learn the statistical methods needed to analyze them. Briefly, we will begin to study four cases:

The Cell Phone Case: A bank estimates its cellular phone costs and decides whether to outsource management of its wireless resources by studying the calling patterns of its employees.

The Marketing Research Case: A beverage company investigates consumer reaction to a new bottle design for one of its popular soft drinks.

The Car Mileage Case: To determine if it qualifies for a federal tax credit based on fuel economy, an automaker studies the gas mileage of its new midsize model.

The Disney Parks Case: Walt Disney World Parks and Resorts in Orlando, Florida, manages Disney parks worldwide and uses data gathered from its guests to give these guests a more "magical" experience and increase Disney revenues and profits.

1.1 Data

Data sets, elements, and variables

LO1-1

Define a variable.

We have said that data are facts and figures from which conclusions can be drawn. Together, the data that are collected for a particular study are referred to as a **data set.** For example, Table 1.1 is a data set that gives information about the new homes sold in a Florida luxury home development over a recent three-month period. Potential home buyers could choose either the "Diamond" or the "Ruby" home model design and could have the home built on either a lake lot or a treed lot (with no water access).

In order to understand the data in Table 1.1, note that any data set provides information about some group of individual **elements,** which may be people, objects, events, or other entities. The information that a data set provides about its elements usually describes one or more characteristics of these elements.

Any characteristic of an element is called a **variable.**

TABLE 1.1 A Data Set Describing Five Home Sales DS HomeSales

Home	Model Design	Lot Type	List Price	Selling Price
1	Diamond	Lake	$494,000	$494,000
2	Ruby	Treed	$447,000	$398,000
3	Diamond	Treed	$494,000	$440,000
4	Diamond	Treed	$494,000	$469,000
5	Ruby	Lake	$447,000	$447,000

For the data set in Table 1.1, each sold home is an element, and four variables are used to describe the homes. These variables are (1) the home model design, (2) the type of lot on which the home was built, (3) the list (asking) price, and (4) the (actual) selling price. Moreover, each home model design came with "everything included"—specifically, a complete, luxury interior package and a choice (at no price difference) of one of three different architectural exteriors. The builder made the list price of each home solely dependent on the model design. However, the builder gave various price reductions for homes built on treed lots.

The data in Table 1.1 are real (with some minor changes to protect privacy) and were provided by a business executive—a friend of the authors—who recently received a promotion and needed to move to central Florida. While searching for a new home, the executive and his family visited the luxury home community and decided they wanted to purchase a Diamond model on a treed lot. The list price of this home was $494,000, but the developer offered to sell it for an "incentive" price of $469,000. Intuitively, the incentive price's $25,000 savings off list price seemed like a good deal. However, the executive resisted making an immediate decision. Instead, he decided to collect data on the selling prices of new homes recently sold in the community and use the data to assess whether the developer might accept a lower offer. In order to collect "relevant data," the executive talked to local real estate professionals and learned that new homes sold in the community during the previous three months were a good indicator of current home value. Using real estate sales records, the executive also learned that five of the community's new homes had sold in the previous three months. The data given in Table 1.1 are the data that the executive collected about these five homes.

When the business executive examined Table 1.1, he noted that homes on lake lots had sold at their list price, but homes on treed lots had not. Because the executive and his family wished to purchase a Diamond model on a treed lot, the executive also noted that two Diamond models on treed lots had sold in the previous three months. One of these Diamond models had sold for the incentive price of $469,000, but the other had sold for a lower price of $440,000. Hoping to pay the lower price for his family's new home, the executive offered $440,000 for the Diamond model on the treed lot. Initially, the home builder turned down this offer, but two days later the builder called back and accepted the offer. The executive had used data to buy the new home for $54,000 less than the list price and $29,000 less than the incentive price!

LO1-2

Describe the difference between a quantitative variable and a qualitative variable.

TABLE 1.2
2014 MLB Payrolls
DS MLB

Team	2014 Payroll
Los Angeles Dodgers	235
New York Yankees	204
Philadelphia Phillies	180
Boston Red Sox	163
Detroit Tigers	162
Los Angeles Angels	156
San Francisco Giants	154
Texas Rangers	136
Washington Nationals	135
Toronto Blue Jays	133
Arizona Diamondbacks	113
Cincinnati Reds	112
St. Louis Cardinals	111
Atlanta Braves	111
Baltimore Orioles	107
Milwaukee Brewers	104
Colorado Rockies	96
Seattle Mariners	92
Kansas City Royals	92
Chicago White Sox	91
San Diego Padres	90
New York Mets	89
Chicago Cubs	89
Minnesota Twins	86
Oakland Athletics	83
Cleveland Indians	83
Pittsburgh Pirates	78
Tampa Bay Rays	77
Miami Marlins	48
Houston Astros	45

Source: http://baseball.about.com/od/newsrumors/fl/2014-Major-League-Baseball-Team-Payrolls.htm (accessed January 14, 2015).

Quantitative and qualitative variables

For any variable describing an element in a data set, we carry out a **measurement** to assign a value of the variable to the element. For example, in the real estate example, real estate sales records gave the actual selling price of each home to the nearest dollar. As another example, a credit card company might measure the time it takes for a cardholder's bill to be paid to the nearest day. Or, as a third example, an automaker might measure the gasoline mileage obtained by a car in city driving to the nearest one-tenth of a mile per gallon by conducting a mileage test on a driving course prescribed by the Environmental Protection Agency (EPA). If the possible values of a variable are numbers that represent quantities (that is, "how much" or "how many"), then the variable is said to be **quantitative.** For example, (1) the actual selling price of a home, (2) the payment time of a bill, (3) the gasoline mileage of a car, and (4) the 2014 payroll of a Major League Baseball team are all quantitative variables. Considering the last example, Table 1.2 in the page margin gives the 2014 payroll (in millions of dollars) for each of the 30 Major League Baseball (MLB) teams. Moreover, Figure 1.1 portrays the team payrolls as a **dot plot.** In this plot, each team payroll is shown

FIGURE 1.1 A Dot Plot of 2014 MLB Payrolls (Payroll Is a Quantitative Variable)

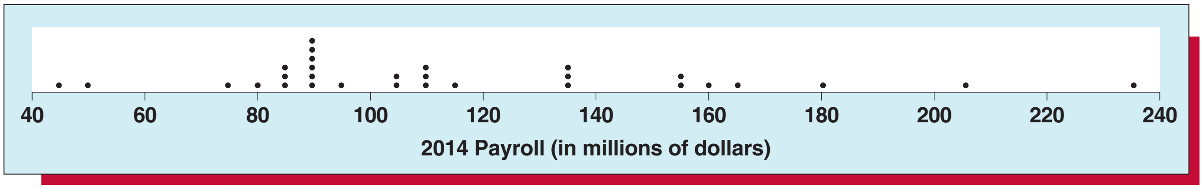
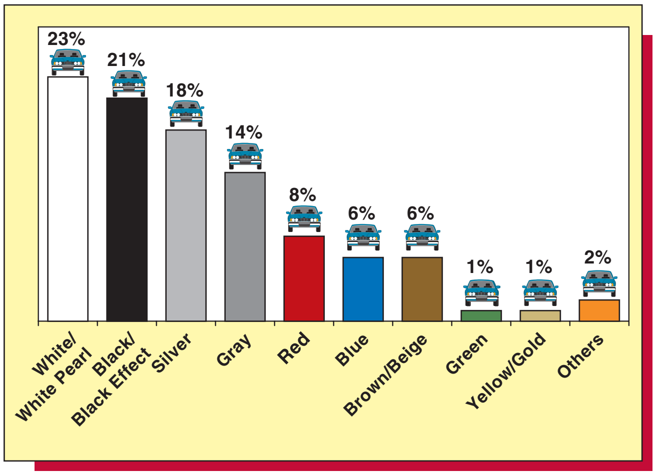


FIGURE 1.2 The Ten Most Popular Car Colors in the World for 2012 (Car Color Is a Qualitative Variable)



Source: http://www.autoweek.com/article/20121206/carnews01/121209911 (accessed September 12, 2013).

FIGURE 1.3 Time Series Plot of the Average Basic Cable Rates in the U.S. from 1999 to 2009 DS BasicCable

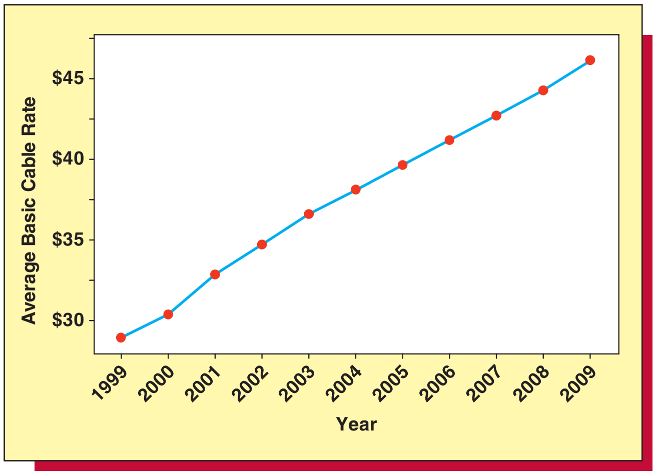


TABLE 1.3 The Average Basic Cable Rates in the U.S. from 1999 to 2009 DS BasicCable

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Cable Rate	$ 28.92	30.37	32.87	34.71	36.59	38.14	39.63	41.17	42.72	44.28	46.13

Source: U.S. Energy Information Administration, http://www.eia.gov/

as a dot located on the real number line—for example, the leftmost dot represents the payroll for the Houston Astros. In general, the values of a quantitative variable are numbers on the real line. In contrast, if we simply record into which of several categories an element falls, then the variable is said to be **qualitative** or **categorical.** Examples of categorical variables include (1) a person's gender, (2) whether a person who purchases a product is satisfied with the product, (3) the type of lot on which a home is built, and (4) the color of a car.[1] Figure 1.2 illustrates the categories we might use for the qualitative variable "car color." This figure is a **bar chart** showing the 10 most popular (worldwide) car colors for 2012 and the percentages of cars having these colors.

Cross-sectional and time series data

LO1-3

Describe the difference between cross-sectional data and time series data.

Some statistical techniques are used to analyze *cross-sectional data,* while others are used to analyze *time series data.* **Cross-sectional data** are data collected at the same or approximately the same point in time. For example, suppose that a bank wishes to analyze last month's cell phone bills for its employees. Then, because the cell phone costs given by these bills are for different employees in the same month, the cell phone costs are cross-sectional data. **Time series data** are data collected over different time periods. For example, Table 1.3 presents the average basic cable television rate in the United States for each of the years 1999 to 2009. Figure 1.3 is a **time series plot**—also called a **runs plot**—of these data. Here we plot each cable rate on the vertical scale versus its corresponding time index (year) on the horizontal scale. For instance, the first cable rate ($28.92) is plotted versus 1999, the second cable rate ($30.37) is plotted versus 2000, and so forth. Examining the time series plot, we see that the cable rates increased substantially from 1999 to 2009. Finally, because the five homes in Table 1.1 were sold over a three-month period that represented a relatively stable real estate market, we can consider the data in Table 1.1 to essentially be cross-sectional data.

LO1-4

Construct and interpret a time series (runs) plot.

[1] Optional Section 1.6 discusses two types of quantitative variables (ratio and interval) and two types of qualitative variables (ordinal and nominative).

LO1-5

Identify the different types of data sources: existing data sources, experimental studies, and observational studies.

1.2 Data Sources, Data Warehousing and Big Data

Primary data are data collected by an individual or business directly through planned **experimentation** or **observation. Secondary data** are data taken from an **existing source.**

Existing sources

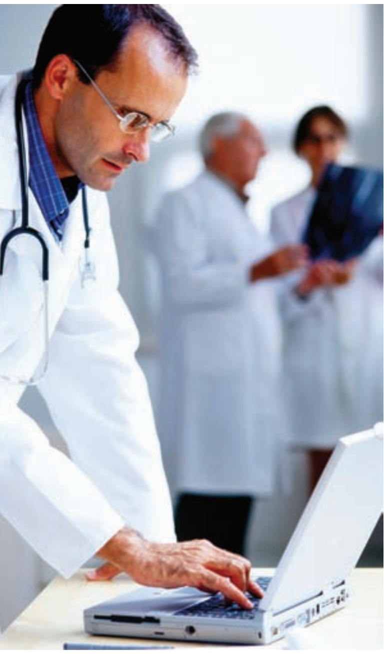
Sometimes we can use data *already gathered* by public or private sources. The Internet is an obvious place to search for electronic versions of government publications, company reports, and business journals, but there is also a wealth of information available in the reference section of a good library or in county courthouse records.

If a business wishes to find demographic data about regions of the United States, a natural source is the U.S. Census Bureau's website at http://www.census.gov. Other useful websites for economic and financial data include the Federal Reserve at http://research.stlouisfed.org/fred2/ and the Bureau of Labor Statistics at http://stats.bls.gov/.

However, given the ease with which anyone can post documents, pictures, blogs, and videos on the Internet, not all sites are equally reliable. Some of the sources will be more useful, exhaustive, and error-free than others. Fortunately, search engines prioritize the lists and provide the most relevant and highly used sites first.

Obviously, performing such web searches costs next to nothing and takes relatively little time, but the tradeoff is that we are also limited in terms of the type of information we are able to find. Another option may be to use a private data source. Most companies keep and use employee records and information about their customers, products, processes (inventory, payroll, manufacturing, and accounting), and advertising results. If we have no affiliation with these companies, however, these data may be difficult to obtain.

Another alternative would be to contact a data collection agency, which typically incurs some kind of cost. You can either buy subscriptions or purchase individual company financial reports from agencies like Bloomberg and Dow Jones & Company. If you need to collect specific information, some companies, such as ACNielsen and Information Resources, Inc., can be hired to collect the information for a fee. Moreover, no matter what existing source you take data from, it is important to assess how reliable the data are by determing how, when, and where the data were collected.



Experimental and observational studies

There are many instances when the data we need are not readily available from a public or private source. In cases like these, we need to collect the data ourselves. Suppose we work for a beverage company and want to assess consumer reactions to a new bottled water. Because the water has not been marketed yet, we may choose to conduct taste tests, focus groups, or some other market research. As another example, when projecting political election results, telephone surveys and exit polls are commonly used to obtain the information needed to predict voting trends. New drugs for fighting disease are tested by collecting data under carefully controlled and monitored experimental conditions. In many marketing, political, and medical situations of these sorts, companies sometimes hire outside consultants or statisticians to help them obtain appropriate data. Regardless of whether newly minted data are gathered in-house or by paid outsiders, this type of data collection requires much more time, effort, and expense than are needed when data can be found from public or private sources.

When initiating a study, we first define our variable of interest, or **response variable.** Other variables, typically called **factors,** that may be related to the response variable of interest will also be measured. When we are able to set or manipulate the values of these factors, we have an **experimental study.** For example, a pharmaceutical company might wish to determine the most appropriate daily dose of a cholesterol-lowering drug for patients having cholesterol levels that are too high. The company can perform an experiment in which one sample of patients receives a placebo; a second sample receives some low dose; a third a higher dose; and so forth. This is an experiment because the company controls the amount of drug each group receives. The optimal daily dose can be determined by analyzing the patients' responses to the different dosage levels given.

When analysts are unable to control the factors of interest, the study is **observational.** In studies of diet and cholesterol, patients' diets are not under the analyst's control. Patients are often unwilling or unable to follow prescribed diets; doctors might simply ask patients what they eat and then look for associations between the factor *diet* and the response variable *cholesterol level*.

Asking people what they eat is an example of performing a **survey.** In general, people in a survey are asked questions about their behaviors, opinions, beliefs, and other characteristics. For instance, shoppers at a mall might be asked to fill out a short questionnaire which seeks their opinions about a new bottled water. In other observational studies, we might simply observe the behavior of people. For example, we might observe the behavior of shoppers as they look at a store display, or we might observe the interactions between students and teachers.

Transactional data, data warehousing, and big data

LO1-6

Explain the basic ideas of data warehousing and big data.

With the increased use of online purchasing and with increased competition, businesses have become more aggressive about collecting information concerning customer transactions. Every time a customer makes an online purchase, more information is obtained than just the details of the purchase itself. For example, the web pages searched before making the purchase and the times that the customer spent looking at the different web pages are recorded. Similarly, when a customer makes an in-store purchase, store clerks often ask for the customer's address, zip code, e-mail address, and telephone number. By studying past customer behavior and pertinent demographic information, businesses hope to accurately predict customer response to different marketing approaches and leverage these predictions into increased revenues and profits. Dramatic advances in data capture, data transmission, and data storage capabilities are enabling organizations to integrate various databases into **data warehouses.** *Data warehousing* is defined as a process of centralized data management and retrieval and has as its ideal objective the creation and maintenance of a central repository for all of an organization's data. The huge capacity of data warehouses has given rise to the term **big data,** which refers to massive amounts of data, often collected at very fast rates in real time and in different forms and sometimes needing quick preliminary analysis for effective business decision making.

C EXAMPLE 1.1 The Disney Parks Case: Improving Visitor Experiences

Annually, approximately 100 million visitors spend time in Walt Disney parks around the world. These visitors could generate a lot of data, and in 2013, Walt Disney World Parks and Resorts introduced the wireless-tracking wristband *MagicBand* in Walt Disney World in Orlando, Florida.

The MagicBands are linked to a credit card and serve as a park entry pass and hotel room key. They are part of the *McMagic*$^+$ system and wearing a band is completely voluntary. In addition to expediting sales transactions and hotel room access in the Disney theme parks, MagicBands provide visitors with easier access to FastPass lines for Disney rides and attractions. Each visitor to a Disney theme park may choose a FastPass for three rides or attractions per day. A FastPass allows a visitor to enter a line where there is virtually no waiting time. The McMagic$^+$ system automatically programs a visitor's FastPass selections into his or her MagicBand. As shown by the photo on the page margin, a visitor simply places the MagicBand on his or her wrist next to a FastPass entry reader and is immediately admitted to the ride or attraction.

In return, the McMagic$^+$ system allows Disney to collect massive amounts of valuable data like real-time location, purchase history, riding patterns, and audience analysis and segmentation data. For example, the data tell Disney the types and ages of people who like specific attractions. To store, process, analyze and visualize all the data, Disney has constructed a gigantic data warehouse and a big data analysis platform. The data analysis allows Disney to improve daily park operations (by having the right numbers of staff on hand for the number of visitors currently in the park); to improve visitor experiences when choosing their "next" ride (by having large displays showing the waiting times for the park's rides); to improve its attraction offerings; and to tailor its marketing messages to different types of visitors.

Finally, although it collects massive amounts of data, Disney is very ethical in protecting the privacy of its visitors. First, as previously stated, visitors can choose not to wear a MagicBand. Moreover, visitors who do decide to wear one have control over the quantities of data collected, stored, and shared. Visitors can use a menu to specify whether Disney can send them personalized offers during or after their park visit. Parents also have to opt-in before the characters in the park can address their children by name or use other personal information stored in the MagicBands.

Exercises for Sections 1.1 and 1.2

CONCEPTS

1.1 Define what we mean by a *variable,* and explain the difference between a quantitative variable and a qualitative (categorical) variable.

1.2 Below we list several variables. Which of these variables are quantitative and which are qualitative? Explain.

a The dollar amount on an accounts receivable invoice.

b The net profit for a company in 2015.

c The stock exchange on which a company's stock is traded.

d The national debt of the United States in 2015.

e The advertising medium (radio, television, or print) used to promote a product.

1.3 **(1)** Discuss the difference between cross-sectional data and time series data. **(2)** If we record the total number of cars sold in 2015 by each of 10 car salespeople, are the data cross-sectional or time series data? **(3)** If we record the total number of cars sold by a particular car salesperson in each of the years 2011, 2012, 2013, 2014, and 2015, are the data cross-sectional or time series data?

1.4 Consider a medical study that is being performed to test the effect of smoking on lung cancer. Two groups of subjects are identified; one group has lung cancer and the other one doesn't. Both are asked to fill out a questionnaire containing questions about their age, sex, occupation, and number of cigarettes smoked per day. **(1)** What is the response variable? **(2)** Which are the factors? **(3)** What type of study is this (experimental or observational)?

1.5 What is a data warehouse? What does the term *big data* mean?

METHODS AND APPLICATIONS

1.6 Consider the five homes in Table 1.1 (page 3). What do you think you would have to pay for a Ruby model on a treed lot?

1.7 Consider the five homes in Table 1.1 (page 3). What do you think you would have to pay for a Diamond model on a lake lot? For a Ruby model on a lake lot?

1.8 The number of Bismark X-12 electronic calculators sold at Smith's Department Stores over the past 24 months have been: 197, 211, 203, 247, 239, 269, 308, 262, 258, 256, 261, 288, 296, 276, 305, 308, 356, 393, 363, 386, 443, 308, 358, and 384. Make a time series plot of these data. That is, plot 197 versus month 1, 211 versus month 2, and so forth. What does the time series plot tell you? DS CalcSale

1.3 Populations, Samples, and Traditional Statistics

LO1-7

Describe the difference between a population and a sample.

We often collect data in order to study a population.

> A **population** is the set of all elements about which we wish to draw conclusions.

Examples of populations include (1) all of last year's graduates of Dartmouth College's Master of Business Administration program, (2) all current MasterCard cardholders, and (3) all Buick LaCrosses that have been or will be produced this year.

We usually focus on studying one or more variables describing the population elements. If we carry out a measurement to assign a value of a variable to each and every population element, we have a *population of measurements* (sometimes called *observations*). If the population is small, it is reasonable to do this. For instance, if 150 students graduated last year from the Dartmouth College MBA program, it might be feasible to survey the graduates and to record all of their starting salaries. In general:

> If we examine all of the population measurements, we say that we are conducting a **census** of the population.

Often the population that we wish to study is very large, and it is too time-consuming or costly to conduct a census. In such a situation, we select and analyze a subset (or portion) of the population elements.

> A **sample** is a subset of the elements of a population.

For example, suppose that 8,742 students graduated last year from a large state university. It would probably be too time-consuming to take a census of the population of all of their starting salaries. Therefore, we would select a sample of graduates, and we would obtain and record their starting salaries. When we measure a characteristic of the elements in a sample, we have a **sample of measurements.**

We often wish to describe a population or sample.

LO1-8

Distinguish between descriptive statistics and statistical inference.

> **Descriptive statistics** is the science of describing the important aspects of a set of measurements.

As an example, if we are studying a set of starting salaries, we might wish to describe (1) what a typical salary might be and (2) how much the salaries vary from each other.

When the population of interest is small and we can conduct a census of the population, we will be able to directly describe the important aspects of the population measurements. However, if the population is large and we need to select a sample from it, then we use what we call **statistical inference.**

> **Statistical inference** is the science of using a sample of measurements to make generalizations about the important aspects of a population of measurements.

For instance, we might use the starting salaries recorded for a sample of the 8,742 students who graduated last year from a large state university to *estimate* the typical starting salary and the variation of the starting salaries for the entire population of the 8,742 graduates. Or General Motors might use a sample of Buick LaCrosses produced this year to estimate the typical EPA combined city and highway driving mileage and the variation of these mileages for all LaCrosses that have been or will be produced this year.

What we might call **traditional statistics** consists of a set of concepts and techniques that are used to describe populations and samples and to make statistical inferences about populations by using samples. Much of this book is devoted to traditional statistics, and in the next section we will discuss **random sampling** (or approximately random sampling). We will also introduce using traditional **statistical modeling** to make statistical inferences. However, traditional statistics is sometimes not sufficient to analyze big data, which (we recall) refers to massive amounts of data often collected at very fast rates in real time and sometimes needing quick preliminary analysis for effective business decision making. For this reason, two related extensions of traditional statistics—**business analytics** and **data mining**—have been developed to help analyze big data. In optional Section 1.5 we will begin to discuss business analytics and data mining. As one example of using business analytics, we will see how Disney uses the large amount of data it collects every day concerning the riding patterns of its visitors. This data is used to keep its visitors informed of the current waiting times for different rides, which helps patrons select the next ride to go on or attraction to attend.