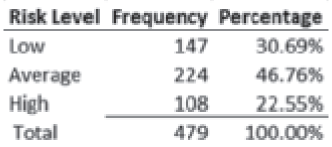
### EXAMPLE 2.1

### Summary Table of Levels of Risk of Retirement Funds

The sample of 479 retirement funds for The Choice *Is* Yours scenario (see page 73) includes the variable Risk Level that has the defined categories low, average, and high. Construct a summary table of the retirement funds, categorized by risk.

**SOLUTION** In Figure 2.1, the percentages for each category are calculated by dividing the number of funds in each category by the total sample size (479). From Figure 2.1, observe that almost half the funds have an average risk, about 30% have low risk, and less than a quarter have high risk.

**FIGURE 2.1**
Frequency and percentage summary table of Risk Level for 485 retirement funds



| Risk Level | Frequency | Percentage |
|---|---|---|
| Low | 147 | 30.69% |
| Average | 224 | 46.76% |
| High | 108 | 22.55% |
| Total | 479 | 100.00% |

## The Contingency Table

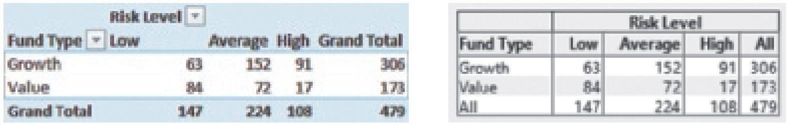
A **contingency table** cross-tabulates, or tallies jointly, the data of two or more categorical variables, allowing you to study patterns that may exist between the variables. Tallies can be shown as a frequency, a percentage of the overall total, a percentage of the row total, or a percentage of the column total. Each tally appears in its own **cell**, and there is a cell for each **joint response**, a unique combination of values for the variables being tallied.

In a contingency table, *both* the rows and the columns represent variables. In the simplest case of a contingency table that summarizes two categorical variables, the rows contain the tallies of one variable and the columns contain the tallies of the other variable. Some use the terms *row variable* and *column variable* to distinguish between the two variables.

For The Choice *Is* Yours scenario, the Fund Type and Risk Levels would be one pair of variables that could be summarized for the sample of 479 retirement funds. Because Fund Type has the defined categories growth and value and the Risk Level has the categories low, average, and high, there are six possible joint responses for this table, forming a two row by three columns contingency table.

Figure 2.2 contains Excel *PivotTable* and JMP versions of this table. (An Excel **PivotTable** generates a table from untallied data.) These summaries show that there are 306 growth and 173 value funds (the row totals) and 147 low risk funds, 224 average risk funds, and 108 high risk funds (the column totals). The tables identify the most frequently encountered joint response in the retirement funds sample as being growth funds with average risk (152).

**FIGURE 2.2**
Excel (PivotTable) and JMP contingency tables of Fund Type and Risk Level for the sample of the 479 retirement funds.



| | Risk Level | | | |
|---|---|---|---|---|
| Fund Type | Low | Average | High | Grand Total |
| Growth | 63 | 152 | 91 | 306 |
| Value | 84 | 72 | 17 | 173 |
| Grand Total | 147 | 224 | 108 | 479 |

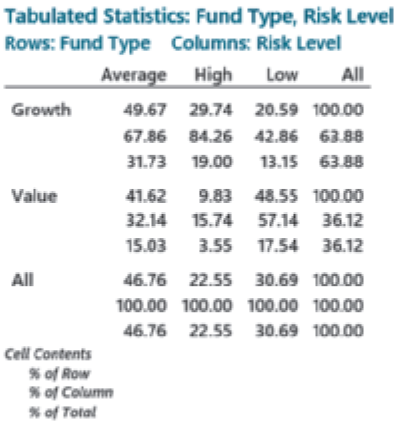
| | Risk Level | | | |
|---|---|---|---|---|
| Fund Type | Low | Average | High | All |
| Growth | 63 | 152 | 91 | 306 |
| Value | 84 | 72 | 17 | 173 |
| All | 147 | 224 | 108 | 479 |

**student TIP**
Remember, each joint response gets tallied into only one cell.

Figure 2.3 presents a Minitab contingency table that expresses tallies as a percentage of the row totals (first line in a row group), as a percentage of the column totals (second line in a row group), and as a percentage of the overall total (third line in a row group). Expressed as percentages, growth funds comprise 63.88% of the funds in the sample (and value funds comprise 36.12%). Of the growth funds, only 20.59% have low risk, whereas 48.55% of the value funds have low risk. As for the funds with high risk, 84.26% are growth funds (and 15.74% are value funds).

From these contingency tables, you conclude that the pattern of risk for growth funds differs from the pattern for value funds.

**FIGURE 2.3**
Minitab contingency table of Fund Type and Risk Level for the sample of the 479 retirement funds, showing row total, column total, and overall total percentages for each joint response.



```
Tabulated Statistics: Fund Type, Risk Level
Rows: Fund Type    Columns: Risk Level

            Average    High     Low     All

Growth        49.67   29.74   20.59  100.00
              67.86   84.26   42.86   63.88
              31.73   19.00   13.15   63.88

Value         41.62    9.83   48.55  100.00
              32.14   15.74   57.14   36.12
              15.03    3.55   17.54   36.12

All           46.76   22.55   30.69  100.00
             100.00  100.00  100.00  100.00
              46.76   22.55   30.69  100.00

Cell Contents
    % of Row
    % of Column
    % of Total
```

## PROBLEMS FOR SECTION 2.1

### LEARNING THE BASICS

**2.1** A categorical variable has three categories, with the following frequencies of occurrence:

| Category | Frequency |
|---|---|
| A | 13 |
| B | 28 |
| C | 9 |

**a.** Compute the percentage of values in each category.
**b.** What conclusions can you reach concerning the categories?

**2.2** The following data represent the responses to two questions asked in a survey of 40 college students majoring in business: What is your gender? (M = male; F = female) and What is your major? (A = Accounting; C = Computer Information Systems; M = Marketing):

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Gender:** | M | M | M | F | M | F | F | M | F | M |
| **Major:** | A | C | C | M | A | C | A | A | C | C |
| **Gender:** | F | M | M | M | M | F | F | M | F | F |
| **Major:** | A | A | A | M | C | M | A | A | A | C |
| **Gender:** | M | M | M | M | F | M | F | F | M | M |
| **Major:** | C | C | A | A | M | M | C | A | A | A |
| **Gender:** | F | M | M | M | M | F | M | F | M | M |
| **Major:** | C | C | A | A | A | A | C | C | A | C |

**a.** Tally the data into a contingency table where the two rows represent the gender categories and the three columns represent the academic major categories.
**b.** Construct contingency tables based on percentages of all 40 student responses, based on row percentages and based on column percentages.

### APPLYING THE CONCEPTS

**2.3** The following table, stored in Smartphone Sales, represents the annual market share of smartphones, by type, for the years 2011, 2012, 2013, 2014, and 2015.

| Type | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|
| Android | 49.2% | 69.0% | 78.8% | 81.5% | 80.7% |
| iOS | 18.8% | 18.7% | 15.1% | 14.8% | 17.7% |
| Microsoft | 1.8% | 2.5% | 3.3% | 2.7% | 1.1% |
| Blackberry | 10.3% | 4.5% | 1.9% | 0.4% | 0.3% |
| OtherOS | 19.8% | 5.4% | 1.0% | 0.6% | 0.2% |

Source: Data extracted from **www.gartner.com/newsroom/id/3215217**.

**a.** What conclusions can you reach about the market for smartphones in 2011, 2012, 2013, 2014, and 2015.
**b.** What differences are there in the 2014 and 2015?

**2.4** The Consumer Financial Protection Bureau reports on consumer financial product and service complaint submissions by state, category, and company. The following table, stored in FinancialComplaints1, represents complaints received from Louisiana consumers by complaint category for 2016.

| Category | Number of Complaints |
|---|---|
| Bank Account or Service | 202 |
| Consumer Loan | 132 |
| Credit Card | 175 |
| Credit Reporting | 581 |
| Debt Collection | 486 |
| Mortgage | 442 |
| Student Loan | 75 |
| Other | 72 |

Source: Data extracted from **bit.ly/2pR7ryO**.

**a.** Compute the percentage of complaints for each category.
**b.** What conclusions can you reach about the complaints for the different categories?

The following table, stored as FinancialComplaints2, summarizes complaints received from Louisiana consumers by most-complained-about companies for 2016.

| Company | Number of Complaints |
|---|---|
| Bank of America | 42 |
| Capital One | 93 |
| Citibank | 59 |
| Ditech Financial | 31 |
| Equifax | 217 |
| Experian | 177 |
| JPMorgan | 128 |
| Nationstar Mortgage | 39 |
| Navient | 38 |
| Ocwen | 41 |
| Synchrony | 43 |
| Trans-Union | 168 |
| Wells Fargo | 77 |

**c.** Compute the percentage of complaints for each company.

**d.** What conclusions can you reach about the complaints for the different companies?

**2.5** In addition to the impact of Big Data, what disruptive technology capability do executives anticipate will have the greatest impact on their firm over the next decade? A survey of 50 Fortune 1000 executives revealed the following:

| Disruptive Capability | Percent |
|---|---|
| Artificial Intelligence/Machine Learning | 44.3 |
| Digital Technologies: mobile/social media/IoT | 26.2 |
| Fin Tech Solutions | 11.5 |
| Cloud Computing | 8.2 |
| Blockchain | 4.9 |
| Other | 4.9 |

What conclusions can you reach concerning the disruptive technology capabilities that executives anticipate will have greatest impact on their firm over the next decade?

SELF TEST **2.6** This table represents the summer power-generating capacity by energy source in the United States as of July 2016.

| Energy Source | Percentage |
|---|---|
| Coal | 26.0 |
| Hydro | 7.5 |
| Natural gas | 42.0 |
| Nuclear | 9.0 |
| Solar | 1.5 |
| Wind | 7.0 |
| Other | 7.0 |

Source: U.S. Department of Energy.

What conclusions can you reach about the source of energy in July 2016?

**2.7** Timetric's 2016 survey of insurance professionals explores the use of technology in the industry. The file Technologies contains the responses to the question that asked what technologies these professionals expected to be most used by the insurance industry in the coming year. Those responses are:

| Technology | Frequency |
|---|---|
| Wearable technology | 9 |
| Blockchain technology | 9 |
| Artificial Intelligence | 17 |
| IoT: retail insurance | 23 |
| IoT: commercial insurance | 5 |
| Social media | 27 |

Source: Data extracted from **bit.ly/2qxMFRj**.

**a.** Compute the percentage of responses for each technology.

**b.** What conclusions can you reach concerning expected technology usage in the insurance industry in the coming year?

**2.8** A survey of 1,520 Americans adults asked "Do you feel overloaded with too much information?" The results indicate that 23% of females feel information overload compared to 17% of males. The results are:

| | GENDER | | |
|---|---|---|---|
| **OVERLOADED** | **Male** | **Female** | **Total** |
| **Yes** | 134 | 170 | 304 |
| **No** | 651 | 565 | 1,216 |
| **Total** | 785 | 735 | 1,520 |

Source: Data extracted from **bit.ly/2pR5bHZ**.

**a.** Construct contingency tables based on total percentages, row percentages, and column percentages.

**b.** What conclusions can you reach from these analyses?

**2.9** A study of selected Kickstarter projects showed that overall a majority were successful, achieving their goal and raising, at a minimum, the targeted amounts. In an effort to identify project types that influence success, selected projects were subdivided into project categories (Film & Video, Games, Music, and Technology). The results are as follows:

| | OUTCOME | | |
|---|---|---|---|
| **CATEGORY** | **Successful** | **Not Successful** | **Total** |
| **Film & Video** | 21,759 | 36,805 | 58,564 |
| **Games** | 9,329 | 18,238 | 27,567 |
| **Music** | 24,285 | 24,377 | 48,662 |
| **Technology** | 5,040 | 20,555 | 25,595 |
| **Total** | 60,413 | 99,975 | 160,388 |

Source: Kickstarter.com, **kickstarter.com/help/stats**.

**a.** Construct contingency tables based on total percentages, row percentages, and column percentages.

**b.** Which type of percentage—row, column, or total—do you think is most informative for these data? Explain.

**c.** What conclusions concerning the pattern of successful Kickstarter projects can you reach?

**2.10** Do social recommendations increase ad effectiveness? A study of online video viewers compared viewers who arrived at an advertising video for a particular brand by following a social media recommendation link to viewers who arrived at the same video by web browsing. Data were collected on whether the viewer could correctly recall the brand being advertised after seeing the video. The results were:

| | CORRECTLY RECALLED THE BRAND | |
|---|---|---|
| ARRIVAL METHOD | Yes | No |
| **Recommendation** | 407 | 150 |
| **Browsing** | 193 | 91 |

Source: Data extracted from "Social Ad Effectiveness: An Unruly White Paper," **www.unrulymedia.com**, January 2012, p. 3.

What do these results tell you about social recommendations?

## 2.2 Organizing Numerical Variables

You create ordered arrays and distribution tables to organize numerical variables. Unless the number of values to organize is very large, you always begin with an **ordered array** that arranges the data for a numerical variable in rank order, from the smallest to the largest value. An ordered array helps you get a better sense of the range of values in your data and is particularly useful when you have more than a few values.

When organizing a numerical variable, you sometimes want to group the data by the value of a categorical variable. For example, in collecting meal cost data as part of a study that reviews the travel and entertainment costs that a business incurs in a major city, you might want to determine if the cost of meals at restaurants located in the center city district differ from the cost at restaurants in the surrounding metropolitan area. As you collect meal cost data for this study, you also note the restaurant location, center city or metro area.

Table 2.2A contains the meal cost data collected from a sample of 50 center city restaurants and 50 metro area restaurants. Table 2.2B presents these two lists of data as two ordered arrays. Note that the ordered arrays in Table 2.2B allow you to make some quick observations about the meal cost data. Using Table 2.2B, you can much more easily see meal costs at center city restaurants range from $23 to $91 and that meal costs at metro area restaurants range from $24 to $81.

**TABLE 2.2A**
Meal Cost Data for 50 Center City and 50 Metro Area Restaurants

| **Center City Restaurants Meal Costs** |
|---|
| 81 28 24 38 45 49 36 60 50 41 84 64 78 57 80 69 89 42 55 32 45 71 50 51 50 |
| 66 49 91 66 58 80 58 50 44 53 62 40 45 23 66 52 47 70 56 55 52 49 26 79 40 |
| **Metro Area Restaurants Meal Costs at** |
| 54 35 29 24 26 31 42 33 25 47 50 59 35 36 43 40 56 34 41 55 42 43 43 64 46 |
| 46 81 33 37 39 54 53 41 39 52 52 42 59 39 69 41 51 36 46 44 75 56 36 33 45 |

**TABLE 2.2B**
Ordered Array of Meal Costs at 50 Center City and 50 Metro Area Restaurants

| **Center City Restaurant Meal Costs** |
|---|
| 23 24 26 28 32 36 38 40 40 41 42 44 45 45 45 47 49 49 49 50 50 50 50 51 52 |
| 52 53 55 55 56 57 58 58 60 62 64 66 66 66 69 70 71 78 79 80 80 81 84 89 91 |
| **Metro Area Restaurant Meal Costs** |
| 24 25 26 29 31 33 33 33 34 35 35 36 36 36 37 39 39 39 40 41 41 41 42 42 42 |
| 43 43 43 44 45 46 46 46 47 50 51 52 52 53 54 54 55 56 56 59 59 64 69 75 81 |

Data for a numerical variable that you intend to group can be stored as stacked or unstacked data, as Section 1.5 discusses. The file Restaurants stores the Table 2.2A data in *both* stacked and unstacked arrangements. As Section 1.5 notes, requirements of specific software procedures often dictate the choice of stacked or unstacked.

When a numerical variable contains a large number of values, using an ordered array to make quick observation or reach conclusions about the data can be difficult. For such a variable, constructing a distribution table would be a better choice. Frequency, relative frequency, percentage, and cumulative distributions are among the types of distribution tables commonly used.

## The Frequency Distribution

A **frequency distribution** tallies the values of a numerical variable into a set of numerically ordered **classes**. Each class groups a mutually exclusive range of values, called a **class interval**. Each value can be assigned to only one class, and every value must be contained in one of the class intervals.

To create a useful frequency distribution, you must consider how many classes would be appropriate for your data as well as determine a suitable *width* for each class interval. In general, a frequency distribution should have at least 5 and no more than 15 classes because having too few or too many classes provides little new information. To determine the **class interval width** [see Equation (2.1)], you subtract the lowest value from the highest value and divide that result by the number of classes you want the frequency distribution to have.

DETERMINING THE CLASS INTERVAL WIDTH

$$\text{Interval width} = \frac{\text{highest value} - \text{lowest value}}{\text{number of classes}} \tag{2.1}$$

For the center city restaurant meal cost data shown in Tables 2.2A and 2.2B, between 5 and 10 classes are acceptable, given the size (50) of that sample. From the center city restaurant meal costs ordered array in Table 2.2B, the difference between the highest value of $91 and the lowest value of $23 is $68. Using Equation (2.1), you approximate the class interval width as follows:

$$\frac{68}{10} = 6.8$$

This result suggests that you should choose an interval width of $6.80. However, your width should always be an amount that simplifies the reading and interpretation of the frequency distribution. In this example, such an amount would be either $5 or $10, and you should choose $10, which creates 8 classes, and not $5, which creates 15 classes, too many for the sample size of 50.

Having chosen a class interval, you examine your data to establish **class boundaries** that properly and clearly define each class. In setting class boundaries, you are looking to establish classes that are simple to interpret and include all values being summarized. With the meal cost data, having decided on $10 as the class interval, you note that the cost of a center city meal ranges from $23 to $91 and the cost of a metro area meal ranges from $24 to $81. You then conclude that the lower class boundary of the first class must be no more than $23 and that the upper boundary of the last class must include $91. You set the lower class boundary of the first class to $20 (for ease of readability) and define the first class as $20 but less than $30, the second class as $30 but less than $40, and so on, ending with the class $90 but less than $100. Table 2.3 uses these class intervals to present frequency distributions for the sample of 50 center city restaurant meal costs and the sample of 50 metro area restaurant meal costs.

Frequency distributions allow you to more easily make observations about your data that support preliminary conclusions about your data. For example, Table 2.3 shows that the cost of center city restaurant meals is concentrated between $40 and $60, while the cost of metro area restaurant meal is concentrated between $30 and $60.

For some charts discussed later in this chapter, class intervals are identified by their **class midpoints**, the values that are halfway between the lower and upper boundaries of each class. For the frequency distributions shown in Table 2.3, the class midpoints are $25, $35, $45, $55, $65, $75, $85, and $95. Note that well-chosen class intervals lead to class midpoints that are simple to read and interpret, as in this example.

**TABLE 2.3**
Frequency Distributions for Cost of a Meal at 50 Center City Restaurants and 50 Metro Area Restaurants

| Meal Cost ($) | Center City Frequency | Metro Area Frequency |
|---|---|---|
| 20 but less than 30 | 4 | 4 |
| 30 but less than 40 | 3 | 14 |
| 40 but less than 50 | 12 | 16 |
| 50 but less than 60 | 14 | 12 |
| 60 but less than 70 | 7 | 2 |
| 70 but less than 80 | 4 | 1 |
| 80 but less than 90 | 5 | 1 |
| 90 but less than 100 | 1 | 0 |
| Total | 50 | 50 |

If the data you have collected do not contain a large number of values, different sets of class intervals can create different impressions of the data. Such perceived changes will diminish as you collect more data. Likewise, choosing different lower and upper class boundaries can also affect impressions.

## EXAMPLE 2.2

### Frequency Distributions of the Three-Year Return Percentages for Growth and Value Funds

As a member of the company task force in The Choice *Is* Yours scenario (see page 73), you are examining the sample of 479 retirement funds stored in Retirement Funds. You want to compare the numerical variable 3YrReturn, the three-year percentage return of a fund, for the two subgroups that are defined by the categorical variable Type (Growth and Value). You construct separate frequency distributions for the growth funds and the value funds.

**SOLUTION** The three-year return for the growth funds is concentrated between 2.5 and 15, while the three-year return for the value funds is concentrated between 2.5 and 10.

**TABLE 2.4**
Frequency Distributions of the Three-Year Return Percentage for Growth and Value Funds

| Three-Year Return Percentage | Growth Frequency | Value Frequency |
|---|---|---|
| −5.00 but less than −2.50 | 1 | 1 |
| −2.50 but less than 0 | 0 | 1 |
| 0 but less than 2.50 | 14 | 8 |
| 2.50 but less than 5.00 | 27 | 20 |
| 5.00 but less than 7.50 | 60 | 69 |
| 7.50 but less than 10.00 | 109 | 67 |
| 10.00 but less than 12.50 | 68 | 7 |
| 12.50 but less than 15.00 | 26 | 0 |
| 15.00 but less than 17.50 | 1 | 0 |
| Total | 306 | 173 |

In the solution for Example 2.2, the total frequency is different for each group (306 and 173). When such totals differ among the groups being compared, you cannot compare the distributions directly as was done in Table 2.3 because of the chance that the table will be misinterpreted. For example, the frequencies for the class interval "5.00 but less than 7.50" look similar—60 and 69—but represent two very different parts of a whole: 60 out of 306 and 69 out of 173 or 19.61% and 39.88%, respectively. When the total frequency differs among the groups being compared, you construct either a relative frequency distribution or a percentage distribution.

## Classes and Excel Bins

Microsoft Excel creates distribution tables using *bins* rather than classes. A **bin** is a range of values defined by a bin number, the upper boundary of the range. Unlike a class, the lower boundary is not explicitly stated but is deduced by the bin number that defines the preceding bin. Consider the bins defined by the bin numbers 4.99, 9.99, and 14.99. The first bin represents all values up to 4.99, the second bin all values greater than 4.99 (the preceding bin number) through 9.99, and the third bin all values greater than 9.99 (the preceding bin number) through 14.99.

Note that when using bins, the lower boundary of the first bin will always be negative infinity, as that bin has no explicit lower boundary. That makes the first Excel bin always much larger than the rest of the bins and violates the rule having equal-sized classes. When you translate classes to bins to make use of certain Excel features, you must include an extra bin number as the first bin number. This extra bin number will always be a value slightly less than the lower boundary of your first class.

You translate your classes into a set of bin numbers that you enter into a worksheet column in ascending order. Tables 2.3 through 2.7 use classes stated in the form "*valueA* but less than *valueB*." For such classes, you create a set of bin numbers that are slightly lower than each *valueB* to approximate each class. For example, you translate the Table 2.4 classes on page 80 as the set of bin numbers −5.01 (the "extra" first bin number that is slightly lower than −5, the lower boundary value of the first class), −2.51 (slightly less than −2.5 the *valueB* of the first class), −0.01, 2.49, 4.99, 7.49, 9.99, 12.49, 14.99, and 17.49 (slightly less than 17.50, the *valueB* of the eighth class).

For classes stated in the form "all values from *valueA* to *valueB*," you can approximate classes by choosing a bin number slightly more than each *valueB*. For example, you can translate the classes stated as 0.0 through 4.9, 5.0 through 9.9, 10.0 through 14.9, and 15.0 through 19.9, as the bin numbers: −0.01 (the extra first bin number), 4.99 (slightly more than 4.9), 9.99, 14.99, and 19.99 (slightly more than 19.9).

## The Relative Frequency Distribution and the Percentage Distribution

Relative frequency and percentage distributions present tallies in ways other than as frequencies. A **relative frequency distribution** presents the relative frequency, or proportion, of the total for each group that each class represents. A **percentage distribution** presents the percentage of the total for each group that each class represents. When you compare two or more groups, knowing the proportion (or percentage) of the total for each group better facilitates comparisons than a table of frequencies for each group would. For example, for comparing meal costs, using Table 2.5 is better than using Table 2.3 on page 80, which displays frequencies.

**TABLE 2.5**
Relative Frequency Distributions and Percentage Distributions of the Meal Costs at Center City and Metro Area Restaurants

| | CENTER CITY | | METRO AREA | |
|---|---|---|---|---|
| MEAL COST ($) | Relative Frequency | Percentage | Relative Frequency | Percentage |
| 20 but less than 30 | 0.08 | 8% | 0.08 | 8% |
| 30 but less than 40 | 0.06 | 6% | 0.28 | 28% |
| 40 but less than 50 | 0.24 | 24% | 0.32 | 32% |
| 50 but less than 60 | 0.28 | 28% | 0.24 | 24% |
| 60 but less than 70 | 0.14 | 14% | 0.04 | 4% |
| 70 but less than 80 | 0.08 | 8% | 0.02 | 2% |
| 80 but less than 90 | 0.10 | 10% | 0.02 | 2% |
| 90 but less than 100 | 0.02 | 2% | 0.00 | 0% |
| Total | 1.00 | 100.0% | 1.00 | 100.0% |

**studentTIP**
Relative frequency columns always sum to 1.00. Percentage columns always sum to 100%.

The **proportion**, or **relative frequency**, in each group is equal to the number of *values* in each class divided by the total number of values. The percentage in each group is its proportion multiplied by 100%.

### COMPUTING THE PROPORTION OR RELATIVE FREQUENCY

The proportion, or relative frequency, is the number of *values* in each class divided by the total number of values:

$$\text{Proportion} = \text{relative frequency} = \frac{\text{number of values in each class}}{\text{total number of values}} \qquad \textbf{(2.2)}$$

If there are 80 values and the frequency in a certain class is 20, the proportion of values in that class is

$$\frac{20}{80} = 0.25$$

and the percentage is

$$0.25 \times 100\% = 25\%$$

You construct a relative frequency distribution by first determining the relative frequency in each class. For example, in Table 2.3 on page 80, there are 50 center city restaurants, and the cost per meal at 14 of these restaurants is between \$50 and \$60. Therefore, as shown in Table 2.5, the proportion (or relative frequency) of meals that cost between \$50 and \$60 at center city restaurants is

$$\frac{14}{50} = 0.28$$

You construct a percentage distribution by multiplying each proportion (or relative frequency) by 100%. Thus, the proportion of meals at center city restaurants that cost between \$50 and \$60 is 14 divided by 50, or 0.28, and the percentage is 28%. Table 2.5 on page 81 presents the relative frequency distribution and percentage distribution of the cost of meals at center city and metro area restaurants. From Table 2.5, you conclude that meal cost is higher at center city restaurants than at metro area restaurants. You note that 14% of the center city

## EXAMPLE 2.3

### Relative Frequency Distributions and Percentage Distributions of the Three-Year Return Percentage for Growth and Value Funds

As a member of the company task force in The Choice *Is* Yours scenario (see page 73), you want to compare the three-year return percentages for the growth and value retirement funds. You construct relative frequency distributions and percentage distributions for these funds.

**SOLUTION** From Table 2.6, you conclude that the three-year return percentage is higher for the growth funds than for the value funds. For example, 19.61% of the growth funds have returns between 5.00 and 7.50 as compared to 39.88% of the value funds, while 22.22% of the growth funds have returns between 10.00 and 12.50 as compared to 4.05% of the value funds.

**TABLE 2.6**
Relative Frequency Distributions and Percentage Distributions of the Three-Year Return Percentage for Growth and Value Funds

| | GROWTH | | VALUE | |
|---|---|---|---|---|
| THREE-YEAR RETURN PERCENTAGE | Relative Frequency | Percentage | Relative Frequency | Percentage |
| −5.00 but less than −2.50 | 0.0033 | 0.33% | 0.0058 | 0.58% |
| −2.50 but less than 0 | 0.0000 | 0.00% | 0.0058 | 0.58% |
| 0 but less than 2.50 | 0.0458 | 4.58% | 0.0462 | 4.62% |
| 2.50 but less than 5.00 | 0.0882 | 8.82% | 0.1156 | 11.56% |
| 5.00 but less than 7.50 | 0.1961 | 19.61% | 0.3988 | 39.88% |
| 7.50 but less than 10.00 | 0.3562 | 35.62% | 0.3873 | 38.73% |
| 10.00 but less than 12.50 | 0.2222 | 22.22% | 0.0405 | 4.05% |
| 12.50 but less than 15.00 | 0.0850 | 8.50% | 0.0000 | 0.00% |
| 15.00 but less than 17.50 | 0.0033 | 0.33% | 0.0000 | 0.00% |
| Total | 1.0000 | 100.00% | 1.0000 | 100.00% |

restaurant meals cost between $60 and $70 as compared to 4% of the metro area restaurant meals and that 6% of the center city restaurant meals cost between $30 and $40 as compared to 28% of the metro area restaurant meals.

## The Cumulative Distribution

The **cumulative percentage distribution** provides a way of presenting information about the percentage of values that are less than a specific amount. You use a percentage distribution as the basis to construct a cumulative percentage distribution.

For example, you might want to know what percentage of the center city restaurant meals cost less than $40 or what percentage cost less than $50. Starting with the Table 2.5 meal cost percentage distribution for center city restaurants on page 81, you combine the percentages of individual class intervals to form the cumulative percentage distribution. Table 2.7 presents the necessary calculations. From this table, you see that none (0%) of the meals cost less than $20, 8% of meals cost less than $30, 14% of meals cost less than $40 (because 6% of the meals cost between $30 and $40), and so on, until all 100% of the meals cost less than $100.

**TABLE 2.7**
Developing the Cumulative Percentage Distribution for Center City Restaurant Meal Costs

| From Table 2.5: | | |
|---|---|---|
| **Class Interval** | **Percentage** | **Percentage of Meal Costs That Are Less Than the Class Interval Lower Boundary** |
| 20 but less than 30 | 8% | 0% (there are no meals that cost less than 20) |
| 30 but less than 40 | 6% | 8% = 0 + 8 |
| 40 but less than 50 | 24% | 14% = 8 + 6 |
| 50 but less than 60 | 28% | 38% = 8 + 6 + 24 |
| 60 but less than 70 | 14% | 66% = 8 + 6 + 24 + 28 |
| 70 but less than 80 | 8% | 80% = 8 + 6 + 24 + 28 + 14 |
| 80 but less than 90 | 10% | 88% = 8 + 6 + 24 + 28 + 14 + 8 |
| 90 but less than 100 | 2% | 98% = 8 + 6 + 24 + 28 + 14 + 8 + 10 |
| 100 but less than 110 | 0% | 100% = 8 + 6 + 24 + 28 + 14 + 8 + 10 + 2 |

Table 2.8 is the cumulative percentage distribution for meal costs that uses cumulative calculations for the center city restaurants (shown in Table 2.7) as well as cumulative calculations for the metro area restaurants (which are not shown). The cumulative distribution shows that the cost of metro area restaurant meals is lower than the cost of meals in center city restaurants. This distribution shows that 36% of the metro area restaurant meals cost less than $40 as compared to 14% of the meals at center city restaurants; 68% of the metro area restaurant meals cost less than $50, but only 38% of the center city restaurant meals do; and 92% of the metro area restaurant meals cost less than $60 as compared to 66% of such meals at the center city restaurants.

**TABLE 2.8**
Cumulative Percentage Distributions of the Meal Costs for Center City and Metro Area Restaurants

| Meal Cost ($) | Percentage of Center City Restaurants Meals That Cost Less Than Indicated Amount | Percentage of Metro Area Restaurants Meals That Cost Less Than Indicated Amount |
|---|---|---|
| 20 | 0 | 0 |
| 30 | 8 | 8 |
| 40 | 14 | 36 |
| 50 | 38 | 68 |
| 60 | 66 | 92 |
| 70 | 80 | 96 |
| 80 | 88 | 98 |
| 90 | 98 | 100 |
| 100 | 100 | 100 |

Unlike in other distributions, the rows of a cumulative distribution do not correspond to class intervals. (Recall that class intervals are mutually *exclusive*. The rows of cumulative distributions are not: The next row "down" *includes* all of the rows above it.) To identify a row, you use the lower class boundaries from the class intervals of the percentage distribution as is done in Table 2.8.

## EXAMPLE 2.4

### Cumulative Percentage Distributions of the Three-Year Return Percentage for Growth and Value Funds

As a member of the company task force in The Choice *Is* Yours scenario (see page 73), you want to continue comparing the three-year return percentages for the growth and value retirement funds. You construct cumulative percentage distributions for the growth and value funds.

**SOLUTION** The cumulative distribution in Table 2.9 indicates that returns are higher for the growth funds than for the value funds. The table shows that 33.33% of the growth funds and 57.23% of the value funds have returns below 7.5%. The table also reveals that 68.95% of the growth funds have returns below 10 as compared to 95.95% of the value funds.

**TABLE 2.9**
Cumulative Percentage Distributions of the Three-Year Return Percentages for Growth and Value Funds

| Three-Year Return Percentages | Growth Percentage Less Than Indicated Value | Value Percentage Less Than Indicated Value |
|---|---|---|
| −5.0 | 0.00% | 0.00% |
| −2.5 | 0.33% | 0.58% |
| 0.0 | 0.33% | 1.16% |
| 2.5 | 4.90% | 5.78% |
| 5.0 | 13.73% | 17.34% |
| 7.5 | 33.33% | 57.23% |
| 10.0 | 68.95% | 95.95% |
| 12.5 | 91.18% | 100.00% |
| 15.0 | 99.67% | 100.00% |
| 17.5 | 100.00% | 100.00% |

## PROBLEMS FOR SECTION 2.2

### LEARNING THE BASICS

**2.11** Construct an ordered array, given the following data from a sample of exam scores in Mathematics:

88 78 78 73 91 78 85

**2.12** Construct an ordered array for 30 students' GPA:

5 4 6 7 9 8 5 6 8 6 5 4 3 6 8
9 7 6 5 9 5 7 8 9 4 7 9 4 5 8

Can you draw any meaningful conclusions? Why or why not?

**2.13** Planning and preparing for the unexpected, especially in response to a security incident, is one of the greatest challenges faced by information technology professionals today. An incident is described as any violation of policy, law, or unacceptable act that involves information assets. Incident Response (IR) teams should be evaluating themselves on metrics, such as incident detection or dwell time, to determine how quickly they can detect and respond to incidents in the environment. In 2016, the SANS Institute surveyed organizations about internal response capabilities. The frequency distribution that summarizes the average time organizations took to detect incidents is:

| Average Dwell Time | Frequency |
|---|---|
| Less than 1 day | 166 |
| Between 1 and less than 2 days | 100 |
| Between 2 and less than 8 days | 124 |
| Between 8 and less than 31 days | 77 |
| Between 31 and less than 90 days | 59 |
| 90 days or more | 65 |

Source: **bit.ly/2oZGXGx**.

**a.** What percentage of organizations took fewer than 2 days, on average, to detect incidents?
**b.** What percentage of organizations took between 2 and 31 days, on average, to detect incidents?
**c.** What percentage of organizations took 31 or more days, on average, to detect incidents?
**d.** What conclusions can you reach about average dwell time of incidents?

**2.14** Data was collected on salaries of compliance specialists in corporate accounting firms. The salaries ranged from $61,000 to $261,000.

**a.** If these salaries were grouped into six class intervals, indicate the class boundaries.
**b.** What class interval width did you choose?
**c.** What are the six class midpoints?

## APPLYING THE CONCEPTS

**2.15** The FIFA World Cup was one of the biggest sporting events of 2018. The file **WC2018TeamAge** contains average age of the players (years, in 2018) of the 32 teams that qualified for the event. These average ages were:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 26.04 | 26.78 | 27.17 | 27.57 | 28.17 | 28.43 | 28.61 | 28.96 |
| 26.09 | 27.09 | 27.26 | 27.78 | 28.22 | 28.43 | 28.78 | 29.17 |
| 26.09 | 27.09 | 27.26 | 27.83 | 28.26 | 28.52 | 28.83 | 29.52 |
| 26.48 | 27.09 | 27.48 | 28.09 | 28.35 | 28.52 | 28.91 | 29.74 |

Source: Data adapted from **https://bit.ly/2zGSWRD**.

**a.** Organize these mean ages as an ordered array.
**b.** Construct a frequency distribution and a percentage distribution for these mean ages.
**c.** Around which class grouping, if any, are these mean ages concentrated? Explain.

✓SELF TEST **2.16** The file **Utility** contains the following data about the cost of electricity (in $) during July 2017 for a random sample of 50 one-bedroom apartments in a large city.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 96 | 171 | 202 | 178 | 147 | 102 | 153 | 197 | 127 | 82 |
| 157 | 185 | 90 | 116 | 172 | 111 | 148 | 213 | 130 | 165 |
| 141 | 149 | 206 | 175 | 123 | 128 | 144 | 168 | 109 | 167 |
| 95 | 163 | 150 | 154 | 130 | 143 | 187 | 166 | 139 | 149 |
| 108 | 119 | 183 | 151 | 114 | 135 | 191 | 137 | 129 | 158 |

**a.** Construct a frequency distribution and a percentage distribution that have class intervals with the upper class boundaries $99, $119, and so on.
**b.** Construct a cumulative percentage distribution.
**c.** Around what amount does the monthly electricity cost seem to be concentrated?

**2.17** How far do commuters in Australia travel for work? The file **CommutingAustralia** contains data about commuting time and distances of the 89 statistical regions of Australia.

Source: Data extracted from Australian Bureau of Statistics, available at https://bit.ly/2Qvtvfu.

For the average commuting distance data,

**a.** Construct a frequency distribution and a percentage distribution.
**b.** Construct a cumulative percentage distribution.
**c.** What conclusions can you reach concerning the average commuting distance of Australians?

**2.18** How does the average annual precipitation differ around the world? The data in **AnnualPrecipitation** contains the average annual precipitation data in millimeters for 4,166 weather stations.

Source: Data extracted from UN Data, available at **https://bit.ly/2DWMYPz**.

**a.** Construct a frequency distribution and a percentage distribution.
**b.** Construct a cumulative percentage distribution.
**c.** What conclusions can you reach concerning the average annual precipitation around the world?

**2.19** One operation of a mill is to cut pieces of steel into parts that will later be used as the frame for front seats in an automobile. The steel is cut with a diamond saw and requires the resulting parts to be within ±0.005 inch of the length specified by the automobile company. Data are collected from a sample of 100 steel parts and stored in **Steel**. The measurement reported is the difference in inches between the actual length of the steel part, as measured by a laser measurement device, and the specified length of the steel part. For example, the first value, −0.002, represents a steel part that is 0.002 inch shorter than the specified length.

**a.** Construct a frequency distribution and a percentage distribution.
**b.** Construct a cumulative percentage distribution.
**c.** Is the steel mill doing a good job meeting the requirements set by the automobile company? Explain.

**2.20** Call centers today play an important role in managing day-to-day business communications with customers. Call centers must be monitored with a comprehensive set of metrics so that businesses can better understand the overall performance of those centers. One key metric for measuring overall call center performance is *service level*, the percentage of calls answered by a human agent within a specified number of seconds. The file **ServiceLevel** contains the following data for time, in seconds, to answer 50 incoming calls to a financial services call center:

16 14 16 19 6 14 15 5 16 18 17 22 6 18 10 15 12 6

19 16 16 15 13 25 9 17 12 10 5 15 23 11 12 14 24 9

10 13 14 26 19 20 13 24 28 15 21 8 16 12

**a.** Construct a frequency distribution and a percentage distribution.
**b.** Construct a cumulative percentage distribution.
**c.** What can you conclude about call center performance if the service level target is set as "80% of calls answered within 20 seconds"?

**2.21** Cycling in cities is getting increasingly popular, which has led to challenges in urban planning. According to the Copenhagenize index, Copenhagen, Denmark, was the most bicycle friendly city in 2017. Assume a new intersection is under construction in your city. The file **BikeTraffic** contains bicycle traffic in your city on 50 different days.

**a.** Construct a frequency distribution and a percentage distribution.
**b.** Construct a cumulative percentage distribution.
**c.** What can you conclude about a planned capacity of 250 people for the intersection?

**2.22** The file **ElectricConsME** contains the electric power consumption data (kWh) of 44 randomly selected four-member households from Saudi Arabia and the United Arab Emirates.

**a.** Construct a frequency distribution and a percentage distribution for each country, using the following class interval widths for each distribution:

Saudi Arabia: 10,000 but less than 20,000; 20,000 but less than 30,000; and so on.
United Arab Emirates: 0 but less than 10,000; 10,000 but less than 20,000; and so on.

**b.** Construct cumulative percentage distributions.
**c.** Which country's families use more electric power–those from Saudi Arabia or United Arab Emirates? Explain.

**2.23** The file Drink contains the following data for the amount of soft drink (in liters) in a sample of fifty 2-liter bottles:

2.109 2.086 2.066 2.075 2.065 2.057 2.052 2.044 2.036 2.038
2.031 2.029 2.025 2.029 2.023 2.020 2.015 2.014 2.013 2.014
2.012 2.012 2.012 2.010 2.005 2.003 1.999 1.996 1.997 1.992
1.994 1.986 1.984 1.981 1.973 1.975 1.971 1.969 1.966 1.967
1.963 1.957 1.951 1.951 1.947 1.941 1.941 1.938 1.908 1.894

**a.** Construct a cumulative percentage distribution.
**b.** On the basis of the results of (a), does the amount of soft drink filled in the bottles concentrate around specific values?

# 2.3 Visualizing Categorical Variables

Visualizing categorical variables involves making choices about how you seek to present your data. When you visualize a single categorical variable, you must think about what you want to highlight about your data and whether your data are concentrated in only a few of your categories. To highlight how categories directly compare to each other, you use a bar chart. To highlight how categories form parts of a whole, you use a pie or doughnut chart. (Figure 2.4 allows you to compare a bar and pie chart for the same data.) To present data that are concentrated in only a few of your categories, you use a Pareto chart.

You can visualize two categorical variables together, again thinking about what you want to highlight. To highlight direct comparisons, you use a side-by-side chart. To highlight how parts form a whole, you use a doughnut chart.

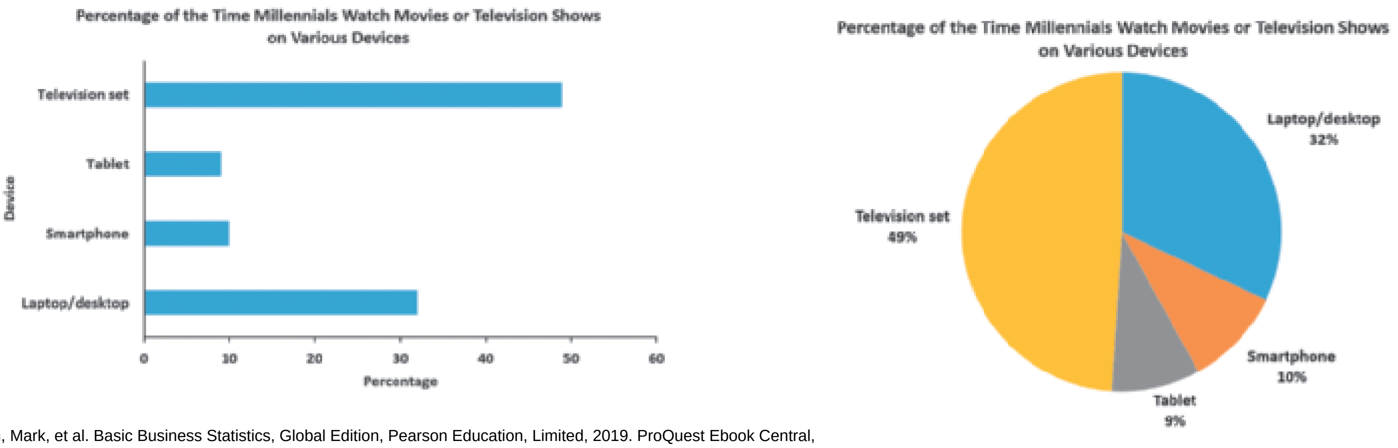
## The Bar Chart

A **bar chart** visualizes a categorical variable as a series of bars, with each bar representing the tallies for a single category. In a bar chart, the length of each bar represents either the frequency or percentage of values for a category and each bar is separated by space, called a gap.

Figure 2.4 includes bar and pie chart visualizations of the Table 2.1 summary table that reports the percentage of the time *millennials*, those born between the years 1983 and 2001, watch movies or television shows on various devices (see page 74). By viewing either of these charts, you can make the same conclusion as reviewing the summary table in the same amount of time: about half of the millennials watch movies and television shows on a television set and half do not. As the complexity of data increases, that equality of time diminishes. With complex data, visualizations will generally allow you to discover relationships among items faster than the equivalent tabular summaries.

**FIGURE 2.4**

Visualizations of the Table 2.1 summary table: bar chart (left) and pie chart (right)



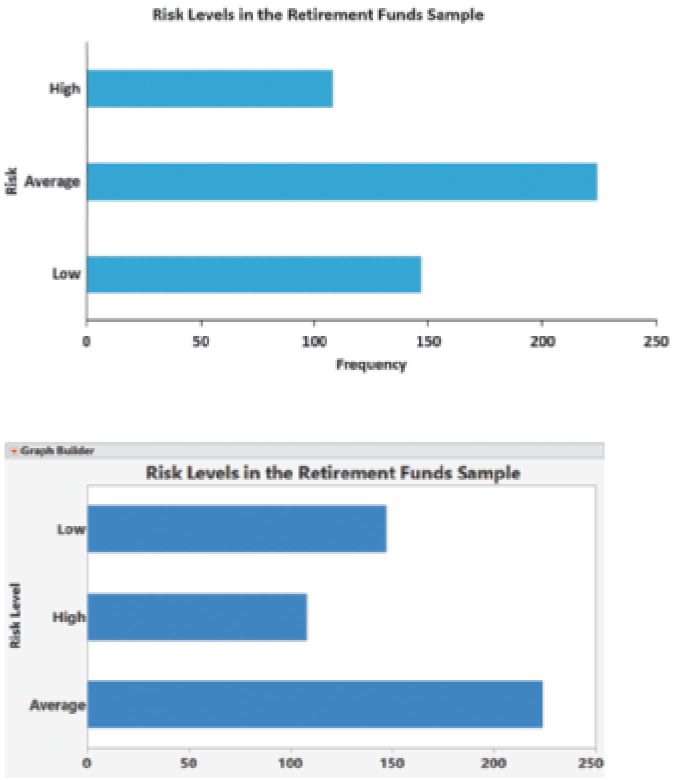
### EXAMPLE 2.5

**Bar Chart of Levels of Risk of Retirement Funds**

As a member of the company task force in The Choice *Is* Yours scenario (see page 73), you want to examine how the Risk Level categories in Figure 2.1 on page 75 compare to each other.

**SOLUTION** You construct the bar chart shown in Figure 2.5. You see that average risk is the largest category, followed by low risk followed by high risk.

**FIGURE 2.5**
Excel and JMP bar chart of the levels of risk of retirement funds



## The Pie Chart and the Doughnut Chart

**Pie** and **doughnut** (or **donut**) charts represent the tallies of each category of a categorical variable as parts of a circle. These parts, or slices, vary by the percentages of the whole for each category. Multiplying category percentages by 360, the number of degrees in a circle, determines the size of each slice, defined as the length of the chord (part of a circle) in degrees. For example, for the Table 2.1 summary table categories, the sizes of the slices would be: desktop/laptop, 115.2 degrees (32% × 360); smartphone, 36 degrees (10% × 360); tablet, 32.4 degrees (9% × 360); and television set, 176.4 degrees (49% × 360). The pie chart in Figure 2.4 displays these slices.

Doughnut charts are pie charts with their centers cut out, creating a hole similar to the holes found in real doughnuts (hence the name). Some believe cutting out centers minimizes a common misperception of pie charts that occurs when people focus on the area of each pie slice and not the length of the chord of each slice. Because most would agree that many pie charts presented together provide an overwhelming visual experience that should be avoided (see reference 2), doughnut charts can be useful when more than one chart is presented together. Doughnut charts can also be used to visualize two variables, as this chapter explains later.

**student TIP**

While using pie or doughnut charts, avoid all "3D" variations or any "exploded" charts in which one or more slices has been pulled away from the center because these forms introduce known visual distortions that can impede understanding of the data.

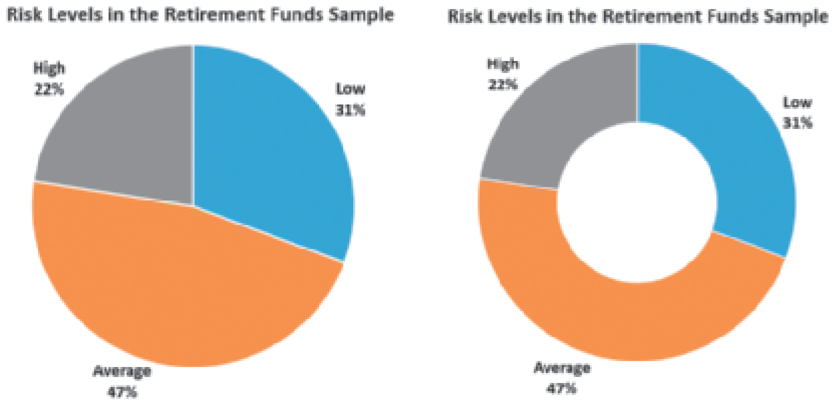
## EXAMPLE 2.6

### Pie Chart and Doughnut Chart of the Risk of Retirement Funds

As a member of the company task force in The Choice *Is* Yours scenario (see page 73), you want to examine how the Risk Level categories in Figure 2.1 on page 75 form parts of a whole.

**SOLUTION** You construct either the Figure 2.6 pie or doughnut chart. You can immediately see that almost half the funds have an average risk and that of the remaining funds, more have low risk than high risk. (A close reading of the labels reveals the actual percentages.)

**FIGURE 2.6**
Excel pie chart and doughnut chart of the risk of retirement funds



## The Pareto Chart

**Pareto charts** help identify the categories that contain the largest tallies from the categories that contain the smallest. Originally developed by the nineteenth-century economist Vilfredo Pareto, these charts help visualize his principle (the **Pareto principle**) that 80% of the consequences result from 20% of the causes. That 20% of the causes are the "vital few" about which one should focus, according to Pareto. While Pareto charts usually do not demonstrate Pareto's 80/20 rule literally, such charts do identify the vital few from the "trivial many" and can be a useful tool today, especially when looking at the frequencies for a large set of categories. In quality management efforts, Pareto charts are very useful tools for prioritizing improvement efforts, such as when data that identify defective or nonconforming items are collected, as in the example that this section uses.

*The vertical scale for each category can be frequency, as Vilfredo Pareto originally used, or percentages of the whole.*

Pareto charts combine two different visualizations: a *vertical* bar chart and a **line graph**, a plot of connected points. The vertical bars represent the tallies for each category, arranged in descending order of the tallies. The line graph represents a cumulative percentage of the tallies from the first category through the last category. The line graph uses a percentage vertical scale, while the bars use either Pareto's original vertical frequency scale or a more recent adaptation that uses a percentage vertical scale line to allow both measurements to share the same scale. In cases with too many categories to display clearly in one chart, categories with the fewest tallies can be combined into a Miscellaneous or Other category and shown as the last (rightmost) bar.

**student TIP**

Excel Pareto charts use the percentage vertical scale for the bars, while JMP and Minitab Pareto charts use the original frequency scale for the bars.

Using Pareto charts can be an effective way to visualize data for studies that seek causes for an observed phenomenon. For example, consider a bank study team that wants to enhance the user experience of automated teller machines (ATMs). During this study, the team identifies incomplete ATM transactions as a significant issue and decides to collect data about the causes of such transactions. Using the bank's own processing systems as a primary data source, causes of incomplete transactions are collected, stored in ATM Transactions, and then organized in the Table 2.10 summary table.

**TABLE 2.10**
Summary Table of Causes of Incomplete ATM Transactions

| Cause | Frequency | Percentage |
|---|---|---|
| ATM malfunctions | 32 | 4.42% |
| ATM out of cash | 28 | 3.87% |
| Invalid amount requested | 23 | 3.18% |
| Lack of funds in account | 19 | 2.62% |
| Card unreadable | 234 | 32.32% |
| Warped card jammed | 365 | 50.41% |
| Wrong keystroke | 23 | 3.18% |
| Total | 724 | 100.00% |

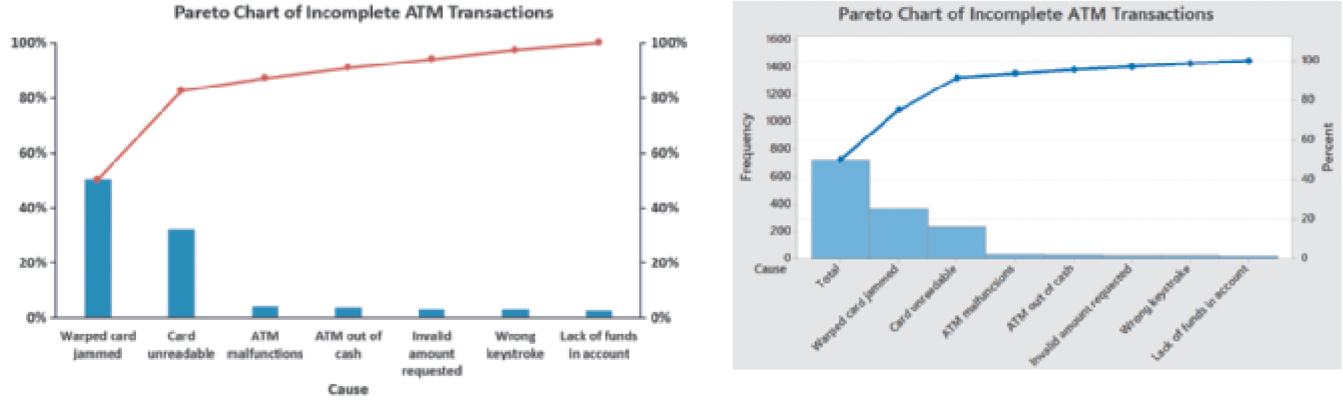
Source: Data extracted from A. Bhalla, "Don't Misuse the Pareto Principle," *Six Sigma Forum Magazine*, May 2009, pp. 15–18.

To separate out the "vital few" causes from the "trivial many" causes, the bank study team creates the Table 2.11 summary table. In this table, causes appear in descending order by frequency, as a Pareto chart requires and the table includes columns for the percentages and cumulative percentages. The team then uses these columns to construct a Figure 2.7 Pareto chart. Note that in Figure 2.7, the left vertical axis represents the percentage due to each cause in the Excel chart, but represents the frequency due to each cause in the Minitab chart. In both charts, the right vertical axis represents the cumulative percentage.

**TABLE 2.11**
Ordered Summary Table of Causes of Incomplete ATM Transactions

| Cause | Frequency | Percentage | Cumulative Percentage |
|---|---|---|---|
| Warped card jammed | 365 | 50.41% | 50.41% |
| Card unreadable | 234 | 32.32% | 82.73% |
| ATM malfunctions | 32 | 4.42% | 87.15% |
| ATM out of cash | 28 | 3.87% | 91.02% |
| Invalid amount requested | 23 | 3.18% | 94.20% |
| Wrong keystroke | 23 | 3.18% | 97.38% |
| Lack of funds in account | 19 | 2.62% | 100.00% |
| Total | 724 | 100.00% | |

**FIGURE 2.7**
Excel and Minitab Pareto charts of incomplete ATM transactions (note the differing left vertical axes)



Because the categories in a Pareto chart are ordered by decreasing frequency of occurrence, the team can quickly see which causes contribute the most to the problem of incomplete transactions. (Those causes would be the "vital few," and figuring out ways to avoid such causes would be, presumably, a starting point for improving the user experience of ATMs.) By following the cumulative percentage line in Figure 2.7, you see that the first two causes, warped card jammed (50.41%) and card unreadable (32.3%), account for 82.7% of the incomplete transactions. Attempts to reduce incomplete ATM transactions due to warped or unreadable cards should produce the greatest payoff.

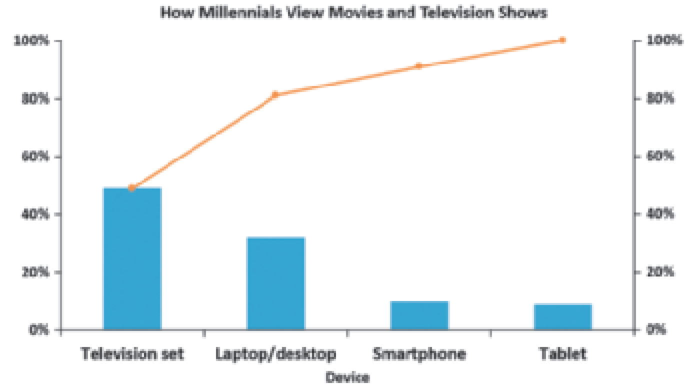
## EXAMPLE 2.7

### Pareto Chart of the Devices Millennials Use to Watch Movies or Television Shows

Construct a Pareto chart from the Table 2.1 summary table that summarizes the devices that millennials, those born between the years 1983 and 2001, use to watch movies or television shows.

**SOLUTION** First, create a new table from Table 2.1 in which the categories are ordered by descending frequency and columns for percentages and cumulative percentages for the ordered categories are included (not shown). From that table, create the Pareto chart in Figure 2.8. From Figure 2.8, observe that about half of the millennials watch movies and television shows on a television and half do not. Also observe that televisions and computers together account for over four-fifths of all such viewing by millennials.

**FIGURE 2.8**
Excel Pareto chart of which devices millennials use to watch movies or television shows

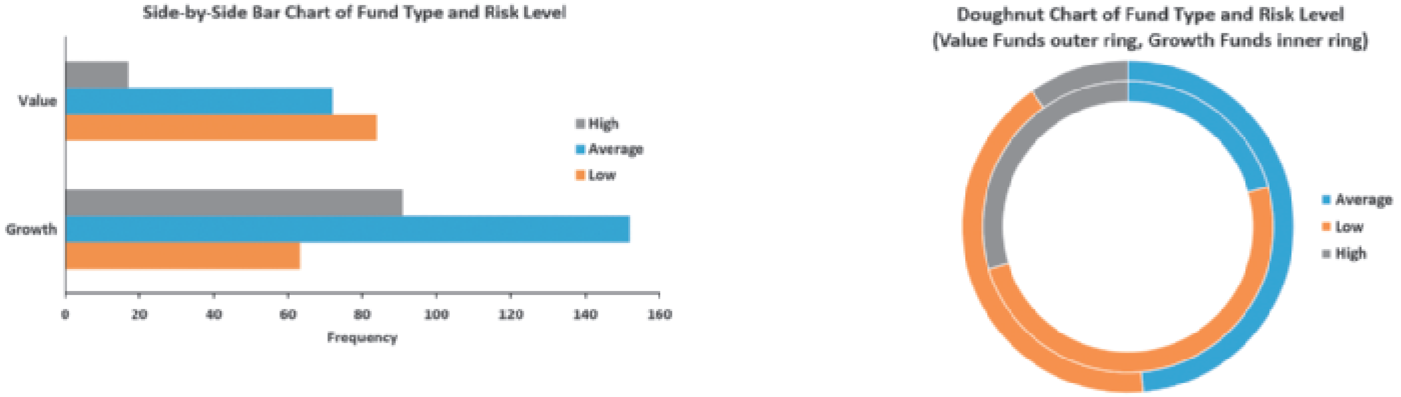


## Visualizing Two Categorical Variables

As when you visualize a single variable, visualizing two categorical variables requires making a choice about what you seek to highlight. To highlight how categories directly compare to each other, you use a side-by-side chart. To highlight how categories form parts of a whole, you use a doughnut chart.

**The side-by-side chart** **A side-by-side chart** visualizes two categorical variables by showing the bars that represent the categories of one variable set grouped by the categories of the second variable. For example, the Figure 2.9 side-by-side chart visualizes the data for the

**FIGURE 2.9**
Side-by-side bar chart and doughnut chart of fund type and risk level



levels of risk for growth and value funds shown in Figure 2.2 on page 75. In Figure 2.9, you see that a substantial portion of the growth funds have average risk. However, more of the value funds have low risk than average or high risk.

**The doughnut chart** When visualizing two variables, the doughnut chart appears as two concentric rings, one inside the other, each ring containing the categories of one variable. In Figure 2.9, the doughnut chart of fund type and risk level highlights that the proportion of funds with average risk (darkest color) is different for growth and value.

## PROBLEMS FOR SECTION 2.3

### APPLYING THE CONCEPTS

**2.24** A survey of online shoppers revealed that in 2015 they bought more of their purchases online than in stores. The data in OnlineShopping reveals how their purchases were made.

**a.** Construct a bar chart, a pie or doughnut chart, and a Pareto chart.

**b.** Which graphical method do you think is best for portraying these data?

**c.** What conclusions can you reach concerning how online shoppers make purchases?

**2.25** How do college students spend their day? The 2016 American Time Use Survey for college students found the following results:

| Activity | Percentage |
|---|---|
| Eating and Drinking | 4% |
| Educational Activities | 14% |
| Grooming | 3% |
| Leisure and Sports | 17% |
| Sleeping | 37% |
| Traveling | 6% |
| Working and Related Activities | 10% |
| Other | 9% |

Source: Data extracted from **bit.ly/2qxIjcH**, accessed February 3, 2017.

**a.** Construct a bar chart, a pie or doughnut chart, and a Pareto chart.

**b.** Which graphical method do you think is best for portraying these data?

**c.** What conclusions can you reach concerning how college students spend their day?

**2.26** The Energy Information Administration reported the following sources of electricity in the United States in 2016:

| Source of Electricity | Percentage |
|---|---|
| Coal | 32% |
| Hydro and renewables | 14% |
| Natural gas | 33% |
| Nuclear power | 19% |
| Other | 2% |

Source: Energy Information Administration, 2016.

**a.** Construct a Pareto chart.

**b.** What percentage of power is derived from coal, nuclear power, or natural gas?

**c.** Construct a pie chart.

**d.** For these data, do you prefer using a Pareto chart or a pie chart? Why?

**2.27** The Consumer Financial Protection Bureau reports on consumer financial product and service complaint submissions by state, category, and company. The following table, stored in FinancialComplaints1, represents complaints received from Louisiana consumers by complaint category for 2016.

| Category | Number of Complaints |
|---|---|
| Bank Account or Service | 202 |
| Consumer Loan | 132 |
| Credit Card | 175 |
| Credit Reporting | 581 |
| Debt Collection | 486 |
| Mortgage | 442 |
| Student Loan | 75 |
| Other | 72 |

Source: Data extracted from **bit.ly/2pR7ryO**.

**a.** Construct a Pareto chart for the categories of complaints.

**b.** Discuss the "vital few" and "trivial many" reasons for the categories of complaints.

The following table, stored in FinancialComplaints2, represents complaints received from Louisiana consumers by most-complained-about companies for 2016.

| Company | Number of Complaints |
|---|---|
| Bank of America | 42 |
| Capital One | 93 |
| Citibank | 59 |
| Ditech Financial | 31 |
| Equifax | 217 |
| Experian | 177 |
| JPMorgan | 128 |
| Nationstar Mortgage | 39 |
| Navient | 38 |
| Ocwen | 41 |
| Synchrony | 43 |
| Trans-Union | 168 |
| Wells Fargo | 77 |

**c.** Construct a bar chart and a pie chart for the complaints by company.
**d.** What graphical method (Pareto, bar, or pie chart) do you think is best for portraying these data?

**2.28** The following table indicates the percentage of residential electricity consumption in the United States, in a recent year organized by type of use.

| Type of Use | Percentage |
|---|---|
| Cooking | 2% |
| Cooling | 15% |
| Electronics | 9% |
| Heating | 15% |
| Lighting | 13% |
| Refrigeration | 10% |
| Water heating | 10% |
| Wet cleaning | 3% |
| Other | 23% |

Source: Department of Energy.

**a.** Construct a bar chart, a pie chart, and a Pareto chart.
**b.** Which graphical method do you think is best for portraying these data?
**c.** What conclusions can you reach concerning residential electricity consumption in the United States?

**2.29** Timetric's 2016 survey of insurance professionals explores the use of technology in the industry. The file **Technologies** contains the responses to the question that asked what technologies these professionals expected to be most used by the insurance industry in the coming year.

| Technology | Frequency |
|---|---|
| Wearable technology | 9 |
| Blockchain technology | 9 |
| Artificial Intelligence | 17 |
| IoT: retail insurance | 23 |
| IoT: commercial insurance | 5 |
| Social media | 27 |

Source: Data extracted from **bit.ly/2qxMFRj**.

**a.** Construct a bar chart and a pie chart.
**b.** What conclusions can you reach concerning expected technology usage in the insurance industry?

**2.30** A survey of 1,520 American adults asked "Do you feel overloaded with too much information?" The results indicate that 23% of females feel information overload compared to 17% of males. The results are:

| | GENDER | | |
|---|---|---|---|
| **OVERLOADED** | **Male** | **Female** | **Total** |
| **Yes** | 134 | 170 | 304 |
| **No** | 651 | 565 | 1,216 |
| **Total** | 785 | 735 | 1,520 |

Source: Data extracted from **bit.ly/2pR5bHZ**.

**a.** Construct a side-by-side bar chart of overloaded with too much information and gender.
**b.** What conclusions can you reach from this chart?

**2.31** A study of selected Kickstarter projects showed that overall a majority were successful, achieving their goal and raising, at a minimum, the targeted amounts. In an effort to identify project types that influence success, selected projects were subdivided into project categories (Film & Video, Games, Music, and Technology). The results are as follows:

| | OUTCOME | | |
|---|---|---|---|
| **CATEGORY** | **Successful** | **Not Successful** | **Total** |
| **Film & Video** | 21,759 | 36,805 | 58,564 |
| **Games** | 9,329 | 18,238 | 27,567 |
| **Music** | 24,285 | 24,377 | 48,662 |
| **Technology** | 5,040 | 20,555 | 25,595 |
| **Total** | 60,413 | 99,975 | 160,388 |

Source: Kickstarter.com, **kickstarter.com/help/stats**.

**a.** Construct a side-by-side bar chart and a doughnut chart of project outcome and category.
**b.** What conclusions concerning the pattern of successful Kickstarter projects can you reach?

**2.32** A research was conducted to find if dogs resemble their owners. The finding of the research was that people tend to select dogs that in some way resemble them and the resemblance increases with the duration of ownership. Assume that this finding is specific to a particular breed of dogs and that the following data has been collected:

| | Resemble Owner | Do Not Resemble Owner |
|---|---|---|
| **Specific Breed** | 20 | 11 |
| **Other Dogs** | 12 | 17 |

**a.** Draw a side-by-side chart to project whether only dogs of a specific breed resemble their owners, or dogs of all breeds do so.
**b.** What conclusions can you draw from the chart?

# 2.4 Visualizing Numerical Variables

You visualize the data for a numerical variable through a variety of techniques that show the distribution of values. These techniques include the stem-and-leaf display, the histogram, the percentage polygon, and the cumulative percentage polygon (ogive), all discussed in this section, as well the boxplot, which requires descriptive summary measures as explained in Section 3.3.

## The Stem-and-Leaf Display

A **stem-and-leaf display** visualizes data by presenting the data as one or more row-wise *stems* that represent a range of values. In turn, each stem has one or more *leaves* that branch out to the right of their stem and represent the values found in that stem. For stems with more than one leaf, the leaves are arranged in ascending order.

Stem-and-leaf displays allow you to see how the data are distributed and where concentrations of data exist. Leaves typically present the last significant digit of each value, but sometimes you round values. For example, suppose you collect the following meal costs (in $) for 15 classmates who had lunch at a fast-food restaurant (stored in FastFood):

7.42 6.29 5.83 6.50 8.34 9.51 7.10 6.80 5.90 4.89 6.50 5.52 7.90 8.30 9.60

To construct the stem-and-leaf display, you use whole dollar amounts as the stems and round the cents to one decimal place as the leaves. For the first value, 7.42, the stem is 7 and its leaf is 4. For the second value, 6.29, the stem is 6 and its leaf 3. The completed stem-and-leaf display for these data with the leaves ordered within each stem is:

| | |
|---|---|
| 4 | 9 |
| 5 | 589 |
| 6 | 3558 |
| 7 | 149 |
| 8 | 33 |
| 9 | 56 |

**student TIP**

A stem-and-leaf display turned sideways looks like a histogram.

## The Histogram

A **histogram** visualizes data as a vertical bar chart in which each bar represents a class interval from a frequency or percentage distribution. In a histogram, you display the numerical variable along the horizontal ($X$) axis and use the vertical ($Y$) axis to represent either the frequency or the percentage of values per class interval. There are never any gaps between adjacent bars in a histogram.

Figure 2.11 visualizes the data of Table 2.3 on page 80, meal costs at center city and metro area restaurants, as a pair of frequency histograms. The histogram for center city restaurants shows that the cost of meals is concentrated between approximately $40 and $60. Ten meals at center city restaurants cost $70 or more. The histogram for metro area restaurants shows that the cost of meals is concentrated between $30 and $60. Very few meals at metro area restaurants cost more than $60.

### EXAMPLE 2.8

### Stem-and-Leaf Display of the Three-Year Return Percentage for the Value Funds

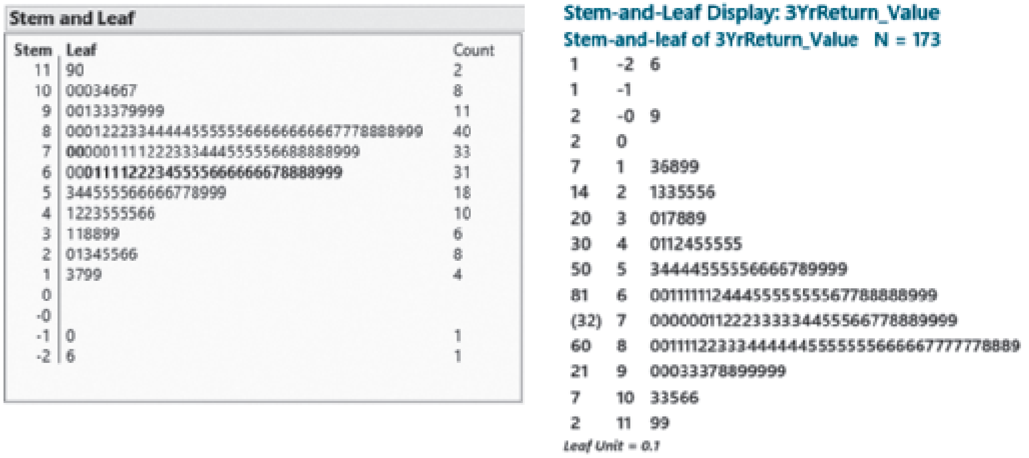
As a member of the company task force in The Choice *Is* Yours scenario (see page 73), you want to study the past performance of the value funds. One measure of past performance is the numerical variable 3YrReturn, the three-year return percentage. Using the data from the 173 value funds, you want to visualize this variable as a stem-and-leaf display.

**SOLUTION** Figure 2.10 presents JMP and Minitab stem-and-leaf displays of the three-year return percentage for value funds. Note that the Minitab display orders percentages from lowest to highest, while the JMP display orders funds from highest to lowest. You observe:

- the lowest three-year return was −2.6.
- the highest three-year return was 11.9.
- the three-year returns were concentrated between 6 and 9.
- very few of the three-year returns were above 11.
- the distribution of the three-year returns appears to have more high values than low values.

▶(*continued*)

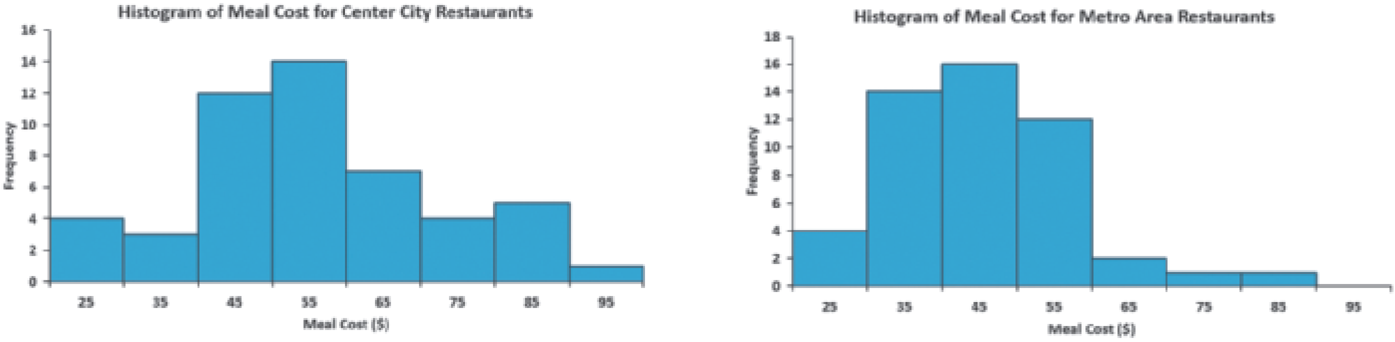
**FIGURE 2.10**
JMP and Minitab stem-and-leaf display of the three-year return percentage for value funds (JMP orders stems from high to low)



```
Stem and Leaf
Stem  Leaf                                        Count
  11 | 90                                          2
  10 | 00034667                                    8
   9 | 00133379999                                 11
   8 | 0001222334444455555566666666667778888999    40
   7 | 000001111222333444555556688888999           33
   6 | 0001111222345555666666678888999             31
   5 | 344555566666778999                          18
   4 | 1223555566                                  10
   3 | 118899                                      6
   2 | 01345566                                    8
   1 | 3799                                        4
   0 |
  -0 |
  -1 | 0                                           1
  -2 | 6                                           1
```

```
Stem-and-Leaf Display: 3YrReturn_Value
Stem-and-leaf of 3YrReturn_Value   N = 173
1     -2  6
1     -1
2     -0  9
2      0
7      1  36899
14     2  1335556
20     3  017889
30     4  0112455555
50     5  34444555556666789999
81     6  0011111124445555555567788888999
(32)   7  00000011222333334455566778889999
60     8  001111223334444445555555566666677777778889
21     9  00033378899999
7     10  33566
2     11  99
Leaf Unit = 0.1
```

**FIGURE 2.11**
Minitab frequency histograms for meal costs at center city and metro area restaurants



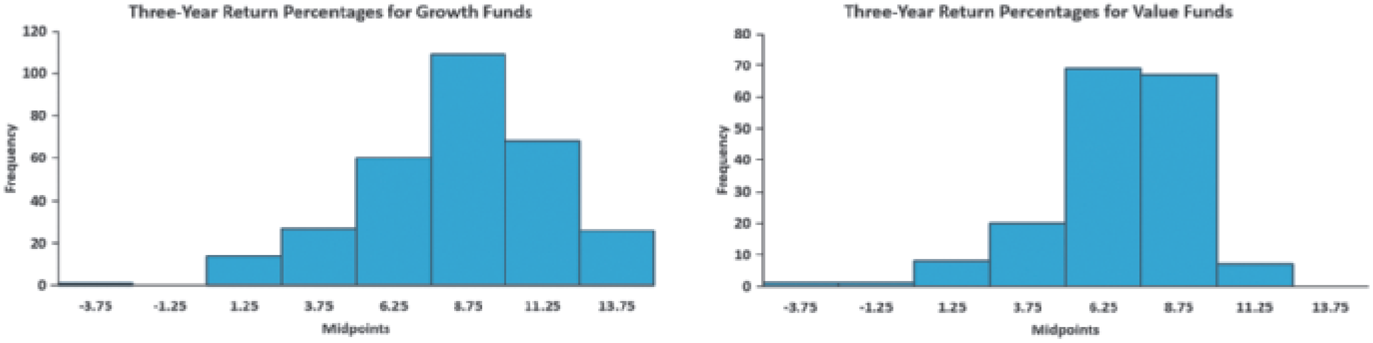
## EXAMPLE 2.9

### Histograms of the Three-Year Return Percentages for the Growth and Value Funds

As a member of the company task force in The Choice *Is* Yours scenario (see page 73), you seek to compare the past performance of the growth funds and the value funds, using the three-year return percentage variable. Using the data from the sample of 479 funds, you construct histograms for the growth and the value funds to create a visual comparison.

**SOLUTION** Figure 2.12 displays frequency histograms for the three-year return percentages for the growth and value funds.

**FIGURE 2.12**
Excel frequency histograms for the three-year return percentages for the growth and value funds



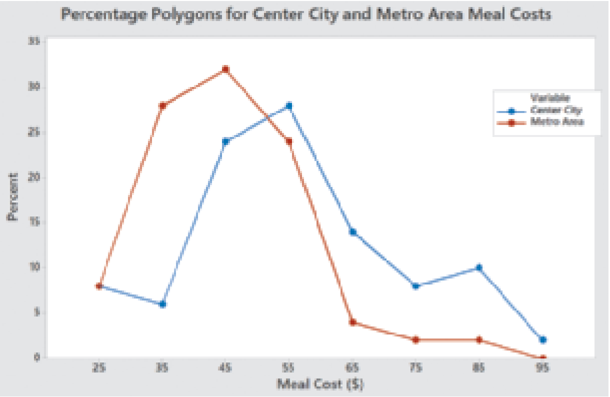
Reviewing the histograms in Figure 2.12 leads you to conclude that the returns were higher for the growth funds than for value funds. The return for the growth funds is more concentrated between 5 and 12.5 while the return for the value funds is more concentrated between 5 and 10.

## The Percentage Polygon

When using a categorical variable to divide the data of a numerical variable into two or more groups, you visualize data by constructing a **percentage polygon**. This chart uses the midpoints of each class interval to represent the data of each class and then plots the midpoints, at their respective class percentages, as points on a line along the *X* axis. While you can construct two or more histograms, as was done in Figures 2.11 and 2.12, a percentage polygon allows you to make a direct comparison that is easier to interpret. (You cannot, of course, combine two histograms into one chart as bars from the two groups would overlap and obscure data.)

Figure 2.13 displays percentage polygons for the cost of meals at center city and metro area restaurants. You can make the same observations from this pair of charts as you made when examining the pair of histograms in Figure 2.11 on page 94. You again note that the center city meal cost is concentrated between $40 and $60 while the metro area meal cost is concentrated between $30 and $60. However, unlike the pair of histograms, the polygons allow you to more easily identify which class intervals have similar percentages for the two groups and which do not.

**FIGURE 2.13**
Minitab percentage polygons of meal costs for center city and metro area restaurants



The polygons in Figure 2.13 have points whose values on the *X* axis represent the midpoint of the class interval. For example, look at the points plotted at $X = 35$ ($35). The point for meal costs at center city restaurants (the lower one) show that 6% of the meals cost between $30 and $40, while the point for the meal costs at metro area restaurants (the higher one) shows that 28% of meals at these restaurants cost between $30 and $40.

When you construct polygons or histograms, the vertical *Y* axis should include zero to avoid distorting the character of the data. The horizontal *X* axis does not need to show the zero point for the numerical variable, but a major portion of the axis should be devoted to the entire range of values for the variable.

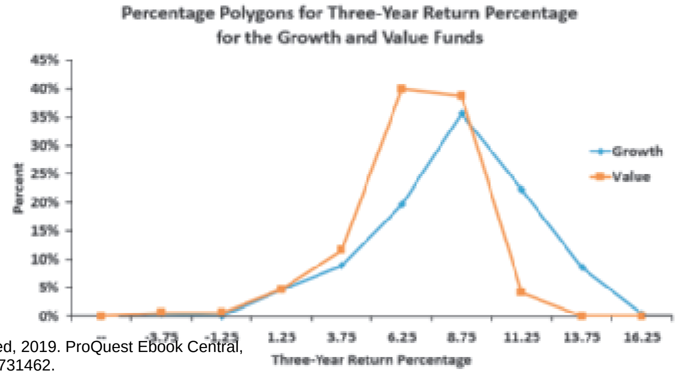
### EXAMPLE 2.10

### Percentage Polygons of the Three-Year Return Percentage for the Growth and Value Funds

As a member of the company task force in The Choice *Is* Yours scenario (see page 73), you seek to compare the past performance of the growth funds and the value funds using the three-year return percentage variable. Using the data from the sample of 479 funds, you construct percentage polygons for the growth and value funds to create a visual comparison.

**SOLUTION** Figure 2.14 displays percentage polygons of the three-year return percentage for the growth and value funds.

**FIGURE 2.14**
Excel percentage polygons of the three-year return percentages for the growth and value funds



▶ (continued)

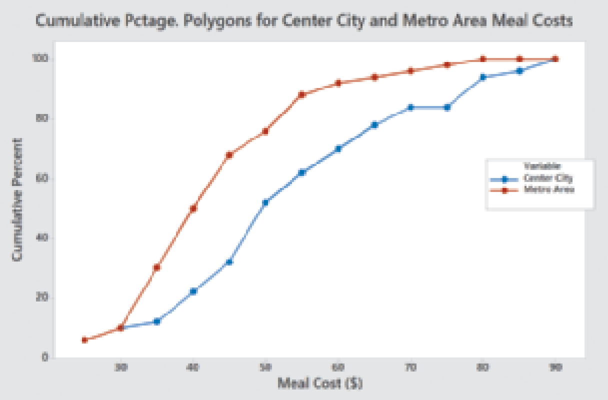
Figure 2.14 shows that the growth funds polygon is to the right of the value funds polygon. This allows you to conclude that the three-year return percentage is higher for growth funds than for value funds. The polygons also show that the return for growth funds is concentrated between 5 and 12.50, and the return for the value funds is concentrated between 5 and 10.

## The Cumulative Percentage Polygon (Ogive)

The **cumulative percentage polygon**, or **ogive**, uses the cumulative percentage distribution discussed in Section 2.2 to plot the cumulative percentages along the *Y* axis. Unlike the percentage polygon, the lower boundaries of the class interval for the numerical variable are plotted, at their respective class percentages as points on a line along the *X* axis.

Figure 2.15 shows cumulative percentage polygons of meal costs for center city and metro area restaurants. In this chart, the lower boundaries of the class intervals (20, 30, 40, etc.) are approximated by the upper boundaries of the previous bins (19.99, 29.99, 39.99, etc.). Reviewing the curves leads you to conclude that the curve of the cost of meals at the center city restaurants is located to the right of the curve for the metro area restaurants. This indicates that the center city restaurants have fewer meals that cost less than a particular value. For example, 38% of the meals at center city restaurants cost less than $50, as compared to 68% of the meals at metro area restaurants.

**FIGURE 2.15**
Minitab cumulative percentage polygons of meal costs for center city and metro area restaurants



## EXAMPLE 2.11

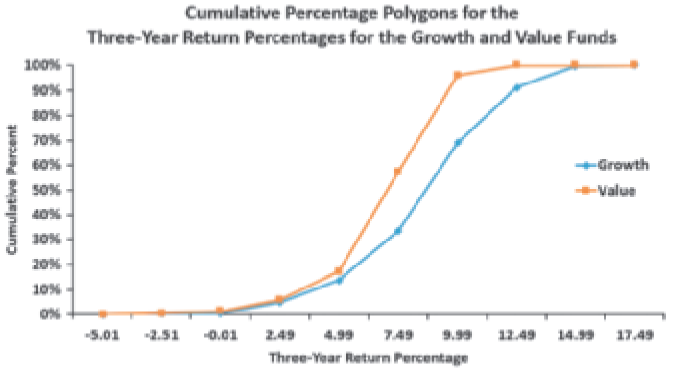
### Cumulative Percentage Polygons of the Three-Year Return Percentages for the Growth and Value Funds

As a member of the company task force in The Choice *Is* Yours scenario (see page 73), you seek to compare the past performance of the growth funds and the value funds using the three-year return percentage variable. Using the data from the sample of 479 funds, you construct cumulative percentage polygons for the growth and the value funds.

**SOLUTION** Figure 2.16 displays cumulative percentage polygons of the three-year return percentages for the growth and value funds.

**FIGURE 2.16**
Excel cumulative percentage polygons of the three-year return percentages for the growth and value funds

*In Microsoft Excel, you approximate the lower boundary by using the upper boundary of the previous bin.*



▶(continued)

The cumulative percentage polygons in Figure 2.16 show that the curve for the three-year return percentage for the growth funds is located to the right of the curve for the value funds. This allows you to conclude that the growth funds have fewer three-year return percentages that are higher than a particular value. For example, 68.95% of the growth funds had three-year return percentages below 10, as compared to 95.95% of the value funds. You can conclude that, in general, the growth funds outperformed the value funds in their three-year returns.

# PROBLEMS FOR SECTION 2.4

## LEARNING THE BASICS

**2.33** Construct a stem-and-leaf display, given the following data from a sample of midterm exam scores in finance:

54 69 98 93 53 74

**2.34** Construct an ordered array, given the following stem-and-leaf display from a sample of $n = 7$ midterm exam scores in information systems:

| | |
|---|---|
| 5 | 0 |
| 6 | |
| 7 | 446 |
| 8 | 19 |
| 9 | 2 |

## APPLYING THE CONCEPTS

**2.35** The following is a stem-and-leaf display representing the amount of gasoline purchased, in gallons (with leaves in tenths of gallons), for a sample of 25 cars that use a particular service station on the New Jersey Turnpike:

| | |
|---|---|
| 9 | 147 |
| 10 | 02238 |
| 11 | 125566777 |
| 12 | 223489 |
| 13 | 02 |

**a.** Construct an ordered array.
**b.** Which of these two displays seems to provide more information? Discuss.
**c.** What amount of gasoline (in gallons) is most likely to be purchased?
**d.** Is there a concentration of the purchase amounts in the center of the distribution?

SELF TEST **2.36** The FIFA World Cup was one of the biggest sporting events of 2018. The file **WC2018TeamAge** contains average age of the players (years, in 2018) of the 32 teams that qualified for the event.

Source: Data adapted from **https://bit.ly/2zGSWRD**.

**a.** Construct a stem-and-leaf display.
**b.** Around what value, if any, are the mean ages of teams concentrated? Explain.

**2.37** The file **MobileSpeed** contains the overall download and upload speeds in mbps for nine carriers in the United States.

Source: Data extracted from "Best Mobile Network 2016", **bit.ly/1KGPrMm**, accessed November 10, 2016.

**a.** Construct an ordered array.
**b.** Construct a stem-and-leaf display.
**c.** Does the ordered array or the stem-and-leaf display provide more information? Discuss.
**d.** Around what value, if any, are the download and upload speeds concentrated? Explain.

**2.38** The file **Utility** contains the following data about the cost of electricity during July of a recent year for a random sample of 50 one-bedroom apartments in a large city:

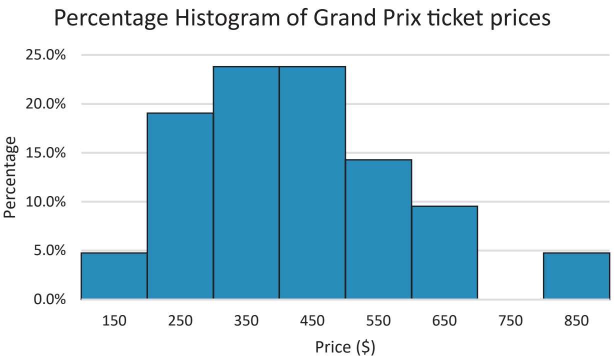
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 96 | 171 | 202 | 178 | 147 | 102 | 153 | 197 | 127 | 82 |
| 157 | 185 | 90 | 116 | 172 | 111 | 148 | 213 | 130 | 165 |
| 141 | 149 | 206 | 175 | 123 | 128 | 144 | 168 | 109 | 167 |
| 95 | 163 | 150 | 154 | 130 | 143 | 187 | 166 | 139 | 149 |
| 108 | 119 | 183 | 151 | 114 | 135 | 191 | 137 | 129 | 158 |

**a.** Construct a histogram and a percentage polygon.
**b.** Construct a cumulative percentage polygon.
**c.** Around what amount does the monthly electricity cost seem to be concentrated?

**2.39** Since its first season in 1950, the FIA Formula One World Championship has become one of the most popular championships of single-seated auto racing. The file **F1Prices2018** contains data about average ticket prices for 21 Grand Prix races.

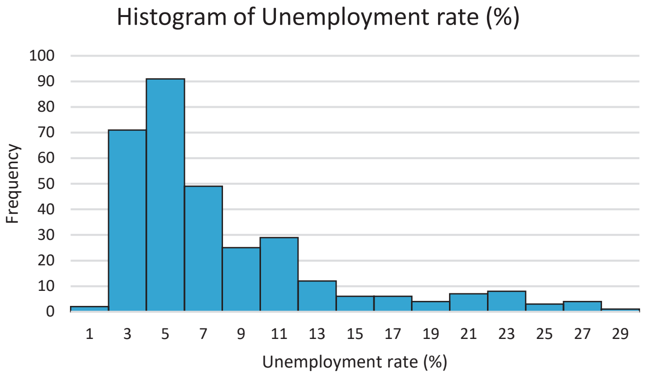
Source: Data adapted from **https://bit.ly/2KTmkZc**.

The following percentage histogram visualizes the total cost (in $) of a ticket at each of the Grand Prix races.



What conclusions can you reach concerning the cost of attending a Grand Prix race?

**2.40** Unemployment is one of the major issues most governments of the world are faced with. The file **EuUnempl2017** contains employment data for 319 European regions in 2017, and the following histogram shows the distribution of unemployment rates.



What conclusions can you reach concerning the unemployment rates in Europe?

**2.41** How far do commuters in Australia travel for work? The file **CommutingAustralia** contains data about commuting time and distances of the 89 statistical regions of Australia.

Source: Data extracted from Australian Bureau of Statistics, available at https://bit.ly/2Qvtvfu.

For the median commuting distance Australians travel for work:

**a.** Construct a percentage histogram.

**b.** Construct a cumulative percentage polygon.

**c.** What conclusions can you reach concerning the median commuting distance Australians travel for work?

**2.42** How does the average annual precipitation differ around the world? The data in **AnnualPrecipitation** contains the average annual precipitation data in millimeters for 4,166 weather stations.

Source: Data extracted from UN Data, available at **https://bit.ly/2DWMYPz**.

**a.** Construct a percentage histogram.

**b.** Construct a cumulative percentage polygon.

**c.** What conclusions can you reach concerning the average annual precipitation around the world?

**2.43** One operation of a mill is to cut pieces of steel into parts that will later be used as the frame for front seats in an automobile. The steel is cut with a diamond saw and requires the resulting parts to be within ±0.005 inch of the length specified by the automobile company. The data are collected from a sample of 100 steel parts and stored in **Steel**. The measurement reported is the difference in inches between the actual length of the steel part, as measured by a laser measurement device, and the specified length of the steel part. For example, the first value, −0.002, represents a steel part that is 0.002 inch shorter than the specified length.

**a.** Construct a percentage histogram.

**b.** Is the steel mill doing a good job meeting the requirements set by the automobile company? Explain.

**2.44** Call centers today play an important role in managing day-to-day business communications with customers. Call centers must be monitored with a comprehensive set of metrics so that businesses can better understand the overall performance of those centers. One key metric for measuring overall call center performance is *service level*, the percentage of calls answered by a human agent within a specified number of seconds. The file **ServiceLevel** contains the following data for time, in seconds, to answer 50 incoming calls to a financial services call center:

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

**a.** Construct a percentage histogram and a percentage polygon.

**b.** Construct a cumulative percentage polygon.

**c.** What can you conclude about call center performance if the service level target is set as "80% of calls answered within 20 seconds"?

**2.45** Cycling in cities is getting increasingly popular, which has led to challenges in urban planning. According to the Copenhagenize index, Copenhagen, Denmark, was the most bicycle friendly city in 2017. Assume a new intersection is under construction in your city. The file **BikeTraffic** contains bicycle traffic in your city on 50 different days.

**a.** Construct a percentage histogram and a percentage polygon.

**b.** Construct a cumulative percentage polygon

**c.** What can you conclude about a planned capacity of 250 people for the intersection?

**2.46** The file **ElectricConsME** contains the electric power consumption data (kWh) of 44 randomly selected four-member households from Saudi Arabia and the United Arab Emirates.

Use the following class interval widths for each distribution:

Saudi Arabia: 10,000 but less than 20,000; 20,000 but less than 30,000; and so on.
United Arab Emirates: 0 but less than 10,000; 10,000 but less than 20,000; and so on.

**a.** Construct percentage histograms on separate graphs and plot the percentage polygons on one graph.

**b.** Plot cumulative percentage polygons on one graph.

**c.** Which country's families use more electric power—Saudi Arabia or the United Arab Emirates? Explain.

**2.47** The data stored in **Drink** represents the amount of soft drink in a sample of fifty 2-liter bottles.

**a.** Construct a histogram and a percentage polygon.

**b.** Construct a cumulative percentage polygon.

**c.** On the basis of the results in (a) and (b), does the amount of soft drink filled in the bottles concentrate around specific values?

# 2.5 Visualizing Two Numerical Variables

Visualizing two numerical variables together can reveal possible relationships between two variables and serve as a basis for applying the methods that Chapters 13 through 16 discuss. To visualize two numerical variables, you use a scatter plot. For the special case in which one of the two variables represents the passage of time, you use a time-series plot.

## The Scatter Plot

A **scatter plot** explores the possible relationship between two numerical variables by plotting the values of one numerical variable on the horizontal, or *X*, axis and the values of a second numerical variable on the vertical, or *Y*, axis. For example, a marketing analyst could study the effectiveness of advertising by comparing advertising expenses and sales revenues of 50 stores by using the *X* axis to represent advertising expenses and the *Y* axis to represent sales revenues.

### EXAMPLE 2.12

**Scatter Plot for NBA Investment Analysis**

Suppose that you are an investment analyst who has been asked to review the valuations of the 30 NBA professional basketball teams. You seek to know if the value of a team reflects its revenues. You collect revenue and valuation data (both in $millions) for all 30 NBA teams, organize the data as Table 2.12, and store the data in NBAValues.

**TABLE 2.12**
Revenues and Values for NBA Teams

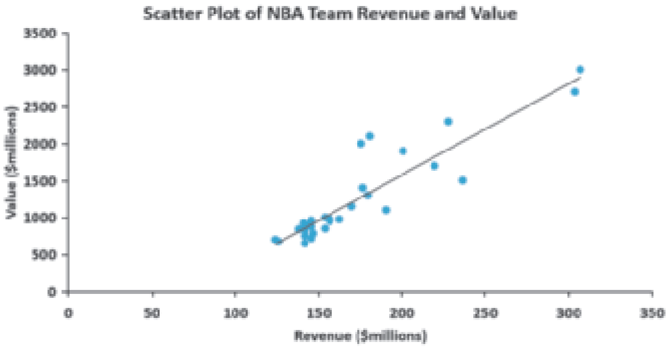
| Team Code | Revenue ($millions) | Current Value ($millions) | Team Code | Revenue ($millions) | Current Value ($millions) | Team Code | Revenue ($millions) | Current Value ($millions) |
|---|---|---|---|---|---|---|---|---|
| ATL | 142 | 825 | HOU | 237 | 1500 | OKC | 157 | 950 |
| BOS | 181 | 2100 | IND | 138 | 840 | ORL | 143 | 900 |
| BKN | 220 | 1700 | LAC | 176 | 2000 | PHI | 124 | 700 |
| CHA | 142 | 750 | LAL | 304 | 2700 | PHX | 154 | 1000 |
| CHI | 228 | 2300 | MEM | 147 | 780 | POR | 157 | 975 |
| CLE | 191 | 1100 | MIA | 180 | 1300 | SAC | 141 | 925 |
| DAL | 177 | 1400 | MIL | 126 | 675 | SAS | 170 | 1150 |
| DEN | 140 | 855 | MIN | 146 | 720 | TOR | 163 | 980 |
| DET | 154 | 850 | NOH | 142 | 650 | UTA | 146 | 875 |
| GSW | 201 | 1900 | NYK | 307 | 3000 | WAS | 146 | 960 |

To quickly visualize a possible relationship between team revenues and valuations, you construct the Figure 2.17 scatter plot, in which you plot the revenues on the *X* axis and the value of the team on the *Y* axis.

**SOLUTION** From Figure 2.17, you see that there appears to be a strong increasing (positive) relationship between revenues and the value of a team. In other words, teams that generate a smaller amount of revenues have a lower value, while teams that generate higher revenues have a higher value. This relationship has been highlighted by the addition of a linear regression prediction line that Chapter 13 explains.

▶(*continued*)

**FIGURE 2.17**
Scatter plot of revenue and value for NBA teams



**learnMORE**
Read the Short Takes for Chapter 2 for an example that illustrates a negative relationship.

Other pairs of variables may have a decreasing (negative) relationship in which one variable decreases as the other increases. In other situations, there may be a weak or no relationship between the variables.

## The Time-Series Plot

A **time-series plot** plots the values of a numerical variable on the *Y* axis and plots the time period associated with each numerical value on the *X* axis. A time-series plot can help you visualize trends in data that occur over time.

### EXAMPLE 2.13

### Time-Series Plot for Movie Revenues

As an investment analyst who specializes in the entertainment industry, you are interested in discovering any long-term trends in movie revenues. You collect the annual revenues (in $billions) for movies released from 1995 to 2016, organize the data as Table 2.13, and store the data in Movie Revenues.

To see if there is a trend over time, you construct the time-series plot shown in Figure 2.18.

**TABLE 2.13**
Movie Revenues (in $billions) from 1995 to 2016

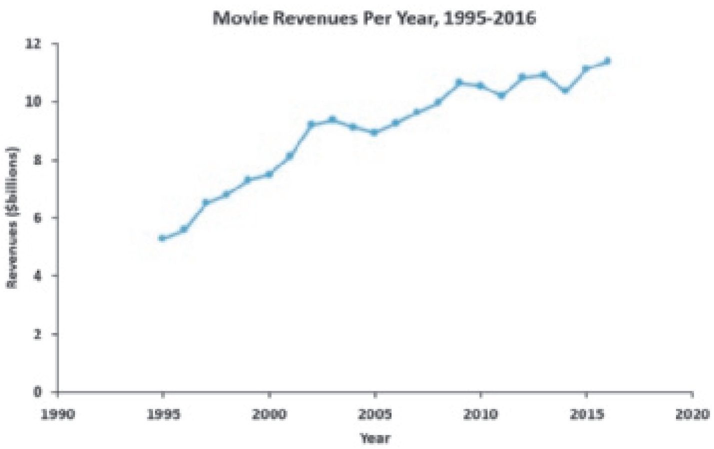
| Year | Revenue ($billions) | Year | Revenue ($billions) | Year | Revenue ($billions) |
|---|---|---|---|---|---|
| 1995 | 5.29 | 2002 | 9.19 | 2009 | 10.65 |
| 1996 | 5.59 | 2003 | 9.35 | 2010 | 10.54 |
| 1997 | 6.51 | 2004 | 9.11 | 2011 | 10.19 |
| 1998 | 6.79 | 2005 | 8.93 | 2012 | 10.83 |
| 1999 | 7.30 | 2006 | 9.25 | 2013 | 10.90 |
| 2000 | 7.48 | 2007 | 9.63 | 2014 | 10.36 |
| 2001 | 8.13 | 2008 | 9.95 | 2015 | 11.13 |
| | | | | 2016 | 11.38 |

Source: Data extracted from **www.the-numbers.com/market**

**SOLUTION** From Figure 2.18, you see that there was a steady increase in the annual movie revenues between 1995 and 2016, followed by an overall upward trend which includes some downturns, reaching new highs in both 2015 and 2016. During that time, the revenues increased from under $6 billion in 1995 to more than $11 billion in 2015 and 2016.

▶(*continued*)

**FIGURE 2.18**
Time-series plot of movie revenues per year from 1995 to 2016



## PROBLEMS FOR SECTION 2.5

### LEARNING THE BASICS

**2.48** The following is a set of data from a sample of $n = 11$ items:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **X:** | 7 | 5 | 8 | 3 | 6 | 0 | 2 | 4 | 9 | 5 | 8 |
| **Y:** | 1 | 5 | 4 | 9 | 8 | 0 | 6 | 2 | 7 | 5 | 4 |

**a.** Construct a scatter plot.
**b.** Is there a relationship between *X* and *Y*? Explain.

**2.49** The following is a series of annual sales (in $millions) over an 11-year period (2007 to 2017):

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Year:** | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
| **Sales:** | 13.0 | 17.0 | 19.0 | 20.0 | 20.5 | 20.5 | 20.5 | 20.0 | 19.0 | 17.0 | 13.0 |

**a.** Construct a time-series plot.
**b.** Does there appear to be any change in annual sales over time? Explain.

### APPLYING THE CONCEPTS

SELF TEST **2.50** Movie companies need to predict the gross receipts of individual movies once a movie has debuted. The following results, stored in PotterMovies, are the first weekend gross, the U.S. gross, and the worldwide gross (in $millions) of the eight Harry Potter movies:

| Title | First Weekend ($millions) | U.S. Gross ($millions) | World-wide Gross ($millions) |
|---|---|---|---|
| *Sorcerer's Stone* | 90.295 | 317.558 | 976.458 |
| *Chamber of Secrets* | 88.357 | 261.988 | 878.988 |
| *Prisoner of Azkaban* | 93.687 | 249.539 | 795.539 |
| *Goblet of Fire/* | 102.335 | 290.013 | 896.013 |
| *Order of the Phoenix* | 77.108 | 292.005 | 938.469 |
| *Half-Blood Prince* | 77.836 | 301.460 | 934.601 |
| *Deathly Hallows Part I* | 125.017 | 295.001 | 955.417 |
| *Deathly Hallows Part II* | 169.189 | 381.011 | 1,328.111 |

Source: Data extracted from **www.the-numbers.com/interactive/comp-Harry-Potter.php**

**a.** Construct a scatter plot with first weekend gross on the *X* axis and U.S. gross on the *Y* axis.
**b.** Construct a scatter plot with first weekend gross on the *X* axis and worldwide gross on the *Y* axis.
**c.** What can you say about the relationship between first weekend gross and U.S. gross and first weekend gross and worldwide gross?

**2.51** Data were collected on the area and population of different states in India. The file IndiaStates contains the vehicle code, zone, area, and population for all 29 states of India.

Source: Data extracted from Population Census 2011, available at **https://bit.ly/1w1BQIG,** and https://bit.ly/1Lc6uDG.

**a.** Construct a scatter plot with area on the *X* axis and population on the *Y* axis.
**b.** What conclusions can you reach about the relationship between area and population?

**2.52** The file MobileSpeed contains the overall download and upload speeds in mbps for nine carriers in the United States.

Source: Data extracted from "Best Mobile Network 2016", **bit.ly/1KGPrMm**, accessed November 10, 2016.

**a.** Do you think that carriers with a higher overall download speed also have a higher overall upload speed?
**b.** Construct a scatter plot with download speed on the *X* axis and upload speed on the *Y* axis.
**c.** Does the scatter plot confirm or contradict your answer in (a)?

**2.53** A Pew Research Center survey found a noticeable rise in smartphone ownership and Internet usage in emerging and developing nations. Once online, adults in these nations are hungry for social interaction. The file GlobalIntenetUsage contains the level of Internet usage, measured as the percentage of adults polled who use the Internet as least occasionally or report owning a smartphone, and the file GlobalSocialMedia contains the level of social media networking, measured as the percentage of Internet users who use social media sites, as well as the GDP at purchasing power parity (PPP, current international $) per capita for each of 28 emerging and developing countries.

Source: Data extracted from Pew Research Center, "Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies," February 22, 2016, **bit.ly/2oRv0rp**.

**a.** Construct a scatter plot with GDP (PPP) per capita on the *X* axis and social media usage on the *Y* axis.

**b.** What conclusions can you reach about the relationship between GDP and social media usage?

**c.** Construct a scatter plot with GDP (PPP) per capita on the *X* axis and Internet usage on the *Y* axis.

**d.** What conclusions can you reach about the relationship between GDP and Internet usage?

**2.54** How have stocks performed in the past? The following table presents the data stored in Stock Performance and shows the performance of a broad measure of stocks (by percentage) for each decade from the 1830s through the 2000s:

| Decade | Perf (%) | Decade | Perf (%) |
|---|---|---|---|
| 1830s | 2.8 | 1920s | 13.3 |
| 1840s | 12.8 | 1930s | −2.2 |
| 1850s | 6.6 | 1940s | 9.6 |
| 1860s | 12.5 | 1950s | 18.2 |
| 1870s | 7.5 | 1960s | 8.3 |
| 1880s | 6.0 | 1970s | 6.6 |
| 1890s | 5.5 | 1980s | 16.6 |
| 1900s | 10.9 | 1990s | 17.6 |
| 1910s | 2.2 | 2000s* | −0.5 |

*Through December 15, 2009.

Source: Data extracted from T. Lauricella, "Investors Hope the '10s Beat the '00s," *The Wall Street Journal*, December 21, 2009, pp. C1, C2.

**a.** Construct a time-series plot of the stock performance from the 1830s to the 2000s.

**b.** Does there appear to be any pattern in the data?

**2.55** The file NewHomeSales contains the number of new homes sold (in thousands) and the median sales price of new single-family houses sold in the United States recorded at the end of each month from January 2000 through December 2016.

Source: Data extracted from **bit.ly/2eEcIBR**, accessed March 19, 2017.

**a.** Construct a times series plot of new home sales prices.

**b.** What pattern, if any, is present in the data?

**2.56** The file Movie Attendance16 contains the yearly movie attendance (in billions) from 2001 through 2016.

| Year | Attendance | Year | Attendance |
|---|---|---|---|
| 2001 | 1.44 | 2009 | 1.41 |
| 2002 | 1.58 | 2010 | 1.34 |
| 2003 | 1.55 | 2011 | 1.28 |
| 2004 | 1.47 | 2012 | 1.36 |
| 2005 | 1.38 | 2013 | 1.34 |
| 2006 | 1.41 | 2014 | 1.27 |
| 2007 | 1.40 | 2015 | 1.32 |
| 2008 | 1.34 | 2016 | 1.32 |

Source: Data extracted from **boxofficemojo.com/yearly**.

**a.** Construct a time-series plot for the movie attendance (in billions).

**b.** What pattern, if any, is present in the data?

**2.57** The Summer Olympics is one of the world's largest sporting events. The file WomenInOlympics contains data about women's sport events and the number of participants from 1900 to 2016.

Source: Data extracted from "Women in the Olympic Movement", 2018, available at **https://bit.ly/2Q9A4VA**.

**a.** Construct a time-series plot on the number of women's events and percentage of participants.

**b.** What pattern, if any, is present in the number of women's events?

**c.** What pattern, if any, is present in the percentage of participants?

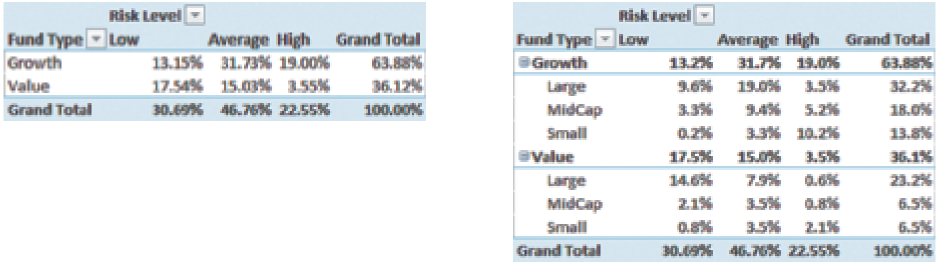
## 2.6 Organizing a Mix of Variables

Earlier sections of this chapter discuss organizing one or two variables of the same type, either categorical or numeric variables. Organizing a mix of many variables into one tabular summary, called a **multidimensional contingency table**, is also possible. Although any number of variables could be theoretically used in multidimensional contingency tables, using many variables together or using a categorical variable that has many categories will produce results that will be hard to comprehend and interpret. As a practical rule, these tables should be limited to no more than three or four variables, which limits their usefulness when exploring sets of data with many variables or analysis that involves big data.[1]

In typical use, these tables either display statistics about each joint response from multiple categorical variables as frequencies or percentages or display statistics about a numerical variable for each joint response from multiple categorical variables. The first form extends contingency tables (see Section 2.1) to two or more row or column variables. The second form replaces the tallies found in a contingency table with summary information about a numeric variable. Figure 2.19 illustrates the first form, adding the variable Market Cap to the Figure 2.2 PivotTable contingency table of Fund Type and Risk Level.

[1] All of the examples in this book follow this rule.

**FIGURE 2.19**
PivotTables of Fund Type and Risk Level (based on Figure 2.2) and Fund Type, Market Cap, and Risk Level for the sample of the 479 retirement funds



| | Risk Level | | | |
|---|---|---|---|---|
| Fund Type | Low | Average | High | Grand Total |
| Growth | 13.15% | 31.73% | 19.00% | 63.88% |
| Value | 17.54% | 15.03% | 3.55% | 36.12% |
| Grand Total | 30.69% | 46.76% | 22.55% | 100.00% |

| | Risk Level | | | |
|---|---|---|---|---|
| Fund Type | Low | Average | High | Grand Total |
| ⊟Growth | 13.2% | 31.7% | 19.0% | 63.88% |
| Large | 9.6% | 19.0% | 3.5% | 32.2% |
| MidCap | 3.3% | 9.4% | 5.2% | 18.0% |
| Small | 0.2% | 3.3% | 10.2% | 13.8% |
| ⊟Value | 17.5% | 15.0% | 3.5% | 36.1% |
| Large | 14.6% | 7.9% | 0.6% | 23.2% |
| MidCap | 2.1% | 3.5% | 0.8% | 6.5% |
| Small | 0.8% | 3.5% | 2.1% | 6.5% |
| Grand Total | 30.69% | 46.76% | 22.55% | 100.00% |

Entries in this new multidimensional contingency table have been formatted as percentages of the whole with one decimal place to facilitate comparisons. The new table reveals patterns in the sample of retirement funds that a table of just Risk Level and Fund Type would not such as:

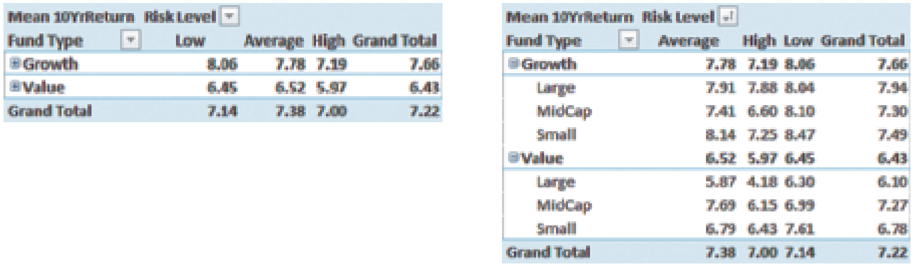
- The pattern of risk for Fund Type when Market Cap is considered can be very different than the summary pattern that Figure 2.2 shows.
- A majority of the large and midcap growth funds have average risk, but most small growth funds have high risk.
- Nearly two-thirds of large market cap value funds have low risk, while a majority of midcap and small value funds have average risk.

**student TIP**
Chapter 3 discusses descriptive statistics for numerical variables, including the mean, also known as the average, that the Figure 2.20 table uses.

Figure 2.20 illustrates the second form of a multidimensional contingency table. To form this table, the numerical variable 10YrReturn has been added to the Figure 2.19 PivotTable of Fund Type, Market Cap, and Risk Level. Note that the numerical variable appears as a statistic that summarizes the variable data, as the mean in these tables. That multidimensional contingency tables can only display a single descriptive statistic for a numerical variable is a limitation of such tables.

Figure 2.20 shows the same PivotTable in two states, with Market Cap *collapsed* into Fund Type (left) and Market Cap *fully expanded* (right). In the collapsed table, funds with high risk have the lowest mean ten-year return percentages. The expanded table discovers that large growth funds with high risk have one the *highest* mean ten-year return percentages, something not suggested by the collapsed table. The expanded table also reveals that midcap value funds with average risk have the highest mean ten-year return percentage among all value funds.

**FIGURE 2.20**
PivotTable (in two states) of Fund Type, Market Cap, and Risk Level, displaying the mean ten-year return percentage for the sample of the 479 retirement funds.



| Mean 10YrReturn | Risk Level | | | |
|---|---|---|---|---|
| Fund Type | Low | Average | High | Grand Total |
| ⊞Growth | 8.06 | 7.78 | 7.19 | 7.66 |
| ⊞Value | 6.45 | 6.52 | 5.97 | 6.43 |
| Grand Total | 7.14 | 7.38 | 7.00 | 7.22 |

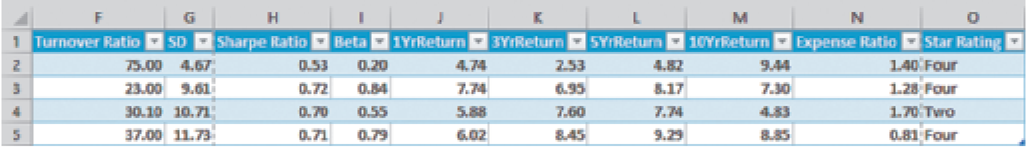
| Mean 10YrReturn | Risk Level | | | |
|---|---|---|---|---|
| Fund Type | Average | High | Low | Grand Total |
| ⊟Growth | 7.78 | 7.19 | 8.06 | 7.66 |
| Large | 7.91 | 7.88 | 8.04 | 7.94 |
| MidCap | 7.41 | 6.60 | 8.10 | 7.30 |
| Small | 8.14 | 7.25 | 8.47 | 7.49 |
| ⊟Value | 6.52 | 5.97 | 6.45 | 6.43 |
| Large | 5.87 | 4.18 | 6.30 | 6.10 |
| MidCap | 7.69 | 6.15 | 6.99 | 7.27 |
| Small | 6.79 | 6.43 | 7.61 | 6.78 |
| Grand Total | 7.38 | 7.00 | 7.14 | 7.22 |

## Drill-down

In addition to their utility to report summaries of variables, multidimensional contingency tables can **drill down** to reveal the data that the table summarizes. When you drill down, you reveal a less summarized form of the data. Expanding a collapsed variable, such as Figure 2.20 demonstrates, is an example of drilling down. In Excel and JMP, you can easily drill down by double-clicking a joint response cell in a multidimensional contingency table. When you double-click a cell, Excel displays the rows of data associated with the joint response in a new worksheet, while JMP highlights those rows in the worksheet data table that is the source for the multidimensional contingency table.

Figure 2.21 shows the drill-down of the small value funds with low risk cell of the Figure 2.20 PivotTables. This drill-down reveals that the ten-year return percentage for this group of four funds ranges from 4.83% to 9.44%, and that the values of some of the other numeric variables also greatly vary.

**FIGURE 2.21**
Drill-down of the Figure 2.20 PivotTable small value funds with low risk cell (some variable columns not shown)



|   | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Turnover Ratio | SD | Sharpe Ratio | Beta | 1YrReturn | 3YrReturn | 5YrReturn | 10YrReturn | Expense Ratio | Star Rating |
| 2 | 75.00 | 4.67 | 0.53 | 0.20 | 4.74 | 2.53 | 4.82 | 9.44 | 1.40 | Four |
| 3 | 23.00 | 9.61 | 0.72 | 0.84 | 7.74 | 6.95 | 8.17 | 7.30 | 1.28 | Four |
| 4 | 30.10 | 10.71 | 0.70 | 0.55 | 5.88 | 7.60 | 7.74 | 4.83 | 1.70 | Two |
| 5 | 37.00 | 11.73 | 0.71 | 0.79 | 6.02 | 8.45 | 9.29 | 8.85 | 0.81 | Four |

# 2.7 Visualizing a Mix of Variables

Earlier sections of this chapter discuss visualizing one or two variables of the same type, either categorical or numeric. Visualizing a mix of many variables is also possible and has the following advantages over multidimensional contingency tables:

- More data and more variables can be presented in a form more manageable to review than a table with many row and column variables
- The data, not summary descriptive statistics, can be shown for numerical variables
- Multiple numerical variables can be presented in one summarization
- Visualizations can reveal patterns that can be hard to see in tables

**learnMORE**

Chapter 17 discusses business analytics and presents additional visualization techniques that also visualize a mix of variables.

These qualities make visualizations of a mix of variables helpful during initial exploratory data analysis and often a necessity in business analytics applications, especially when such techniques are analyzing big data.

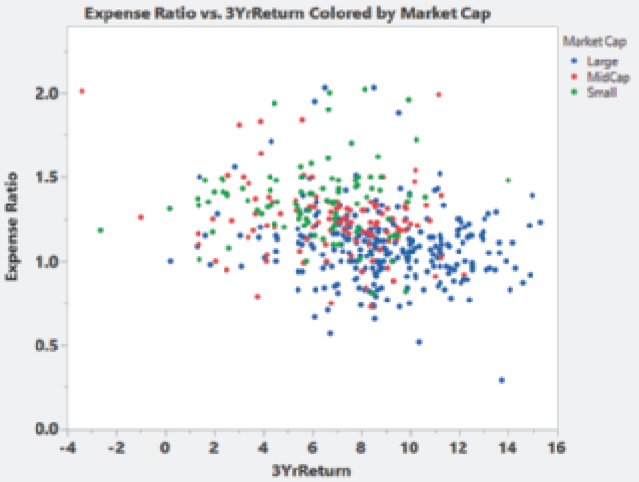
## Colored Scatter Plot

Because of the relative newness of these visualizations, Excel, JMP, and Minitab use different ways to visualize a mix of data. Sometimes these programs confusingly use the *same* name to refer to a visual that works *differently* from program to program. For example, JMP and Minitab can create a **colored scatter plot** that can visualize two (and sometimes more than two) numerical variables and at least one categorical variable.

For example, Figure 2.22 presents a colored scatter plot of the Expense Ratio and 3YrReturn numerical variables and the Market Cap categorical variable for the sample of 479 retirement funds. This visual reveals that for the three-year period, funds with large market capitalizations (red dots) tend to have the best returns and the lowest cost expense ratios (in other words, plot in the lower right quadrant of the chart). However, there are a number of large market cap funds that plot *elsewhere* on the chart, representing funds with relatively high expense ratios or fair to poor three-year returns. For certain types of analyses, the points represented those funds might be drilled down to determine reasons for their different behavior or identify such funds as relative laggards in the set of all large market cap funds.

*The default JMP color theme may prove problematic for those with certain types of color vision deficiencies ("color blindness"). The color theme can be changed, as the JMP Guide explains.*

**FIGURE 2.22**
JMP colored scatter plot of Expense Ratio, 3YrReturn, and Market Cap for the sample of 479 retirement funds.



Because they can compare two numerical variables and one categorical variable, colored scatter plots can be considered an “opposite” of multidimensional contingency tables that summarize the two categorical variables and one numerical variable.

## Bubble Charts

**Bubble charts** extend color scatter plots by using the size of the points, now called bubbles, to represent an additional variable. In Excel and Minitab, that additional variable must be numerical, while in JMP the variable can be either numerical or categorical. JMP also permits coloring and sizing of the bubbles as ways of representing additional variables and can handle time series data in a unique way. (Chapter 17 discusses and presents examples of bubble charts.)

## PivotChart (Excel)

**PivotCharts** pull out and visualize specific categories from a PivotTable summary in a way that would otherwise be hard to do in Excel. For example, Figure 2.23 (left) displays a side-by-side PivotChart based on the Figure 2.20 PivotTables of Fund Type, Market Cap, and Risk Level, that displays the mean ten-year return percentage for the sample of the 479 retirement funds. Filtering the chart to display the mean ten-year return percentages for only low risk funds, Figure 2.23 (right), highlights that small market cap growth funds have the highest mean ten-year return percentage.

## Treemap (Excel, JMP)

**Treemaps** show proportions of the whole of nested categories as colored tiles. In the simplest case (Excel), the size of tiles corresponds to the tallies in a joint response cell in a contingency table. In a more elaborate version (JMP), the tiles can be sized to a numerical variable. Figure 2.24 on page 106 presents Excel and JMP treemaps (simplest case) for Fund Type and Market Cap. Note that Excel can only color the treemap by the categories of first categorical variable (Fund Type), while JMP can color analogous subcategories (the Market Cap categories) while dividing the treemap into parts based on the first categorical variable.

**learnMORE**

Chapter 17 illustrates an application of the more elaborate version of a treemap.

**FIGURE 2.23**
PivotCharts based on the Figure 2.20 PivotTable of Fund Type, Market Cap, and Risk Level, showing the mean ten-year return percentage

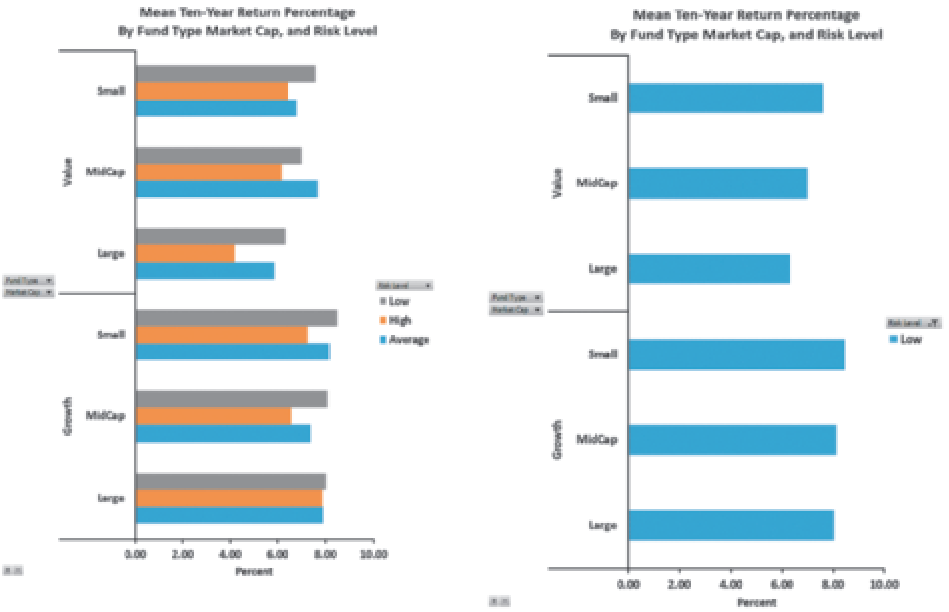
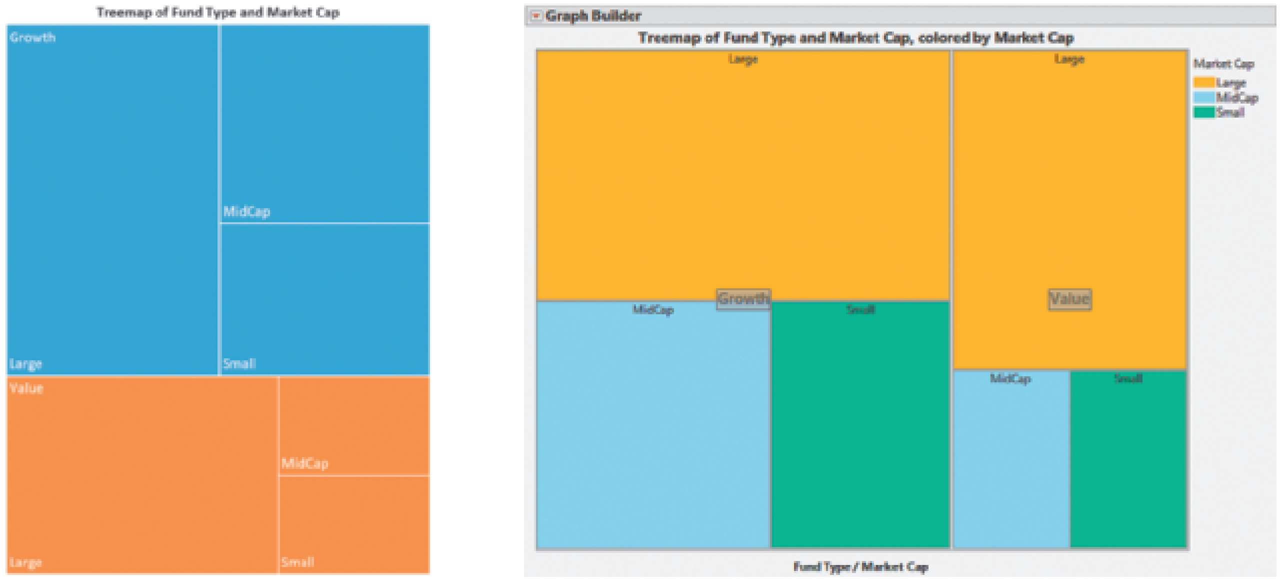


**FIGURE 2.24**
Excel and JMP treemaps for Fund Type and Market Cap

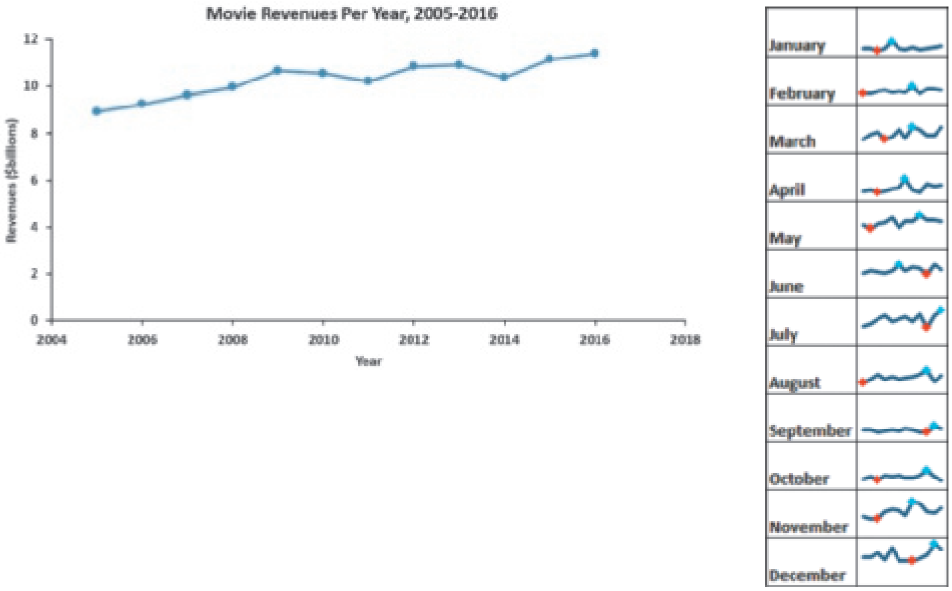


## Sparklines (Excel)

**Sparklines** are compact time-series visualizations of numerical variables. This compact form allows you to view all the visualizations together, which can aid in making comparisons among the variables. Sparklines highlight the trends of the plots over the precise graphing of points found in a time-series plot. Although typically used to plot several independent numerical variables, such as several different business indicators, sparklines can also be used to plot time-series data using smaller time units than a time-series plot to reveal patterns that the time-series plot may not.

For example, Figure 2.25 (left) contains a time-series plot of annual movie revenues (in $billions) from 2005 through 2016, a subset of the data used in Example 2.13 on page 101. Figure 2.25 (right) contains a set of sparklines that plot movie revenues for each month of specific years. The sparklines reveal that movie revenues for the months of February and September do not vary much from year to year, while the monthly revenues for July have rebounded from all-time low to an all-time high for the period 2005–2016.

**FIGURE 2.25**
Time-series plot of movie revenues per year from 2005 to 2016 (left) and sparklines for movie revenues per month for the same time period (right)



# 2.8 Filtering and Querying Data

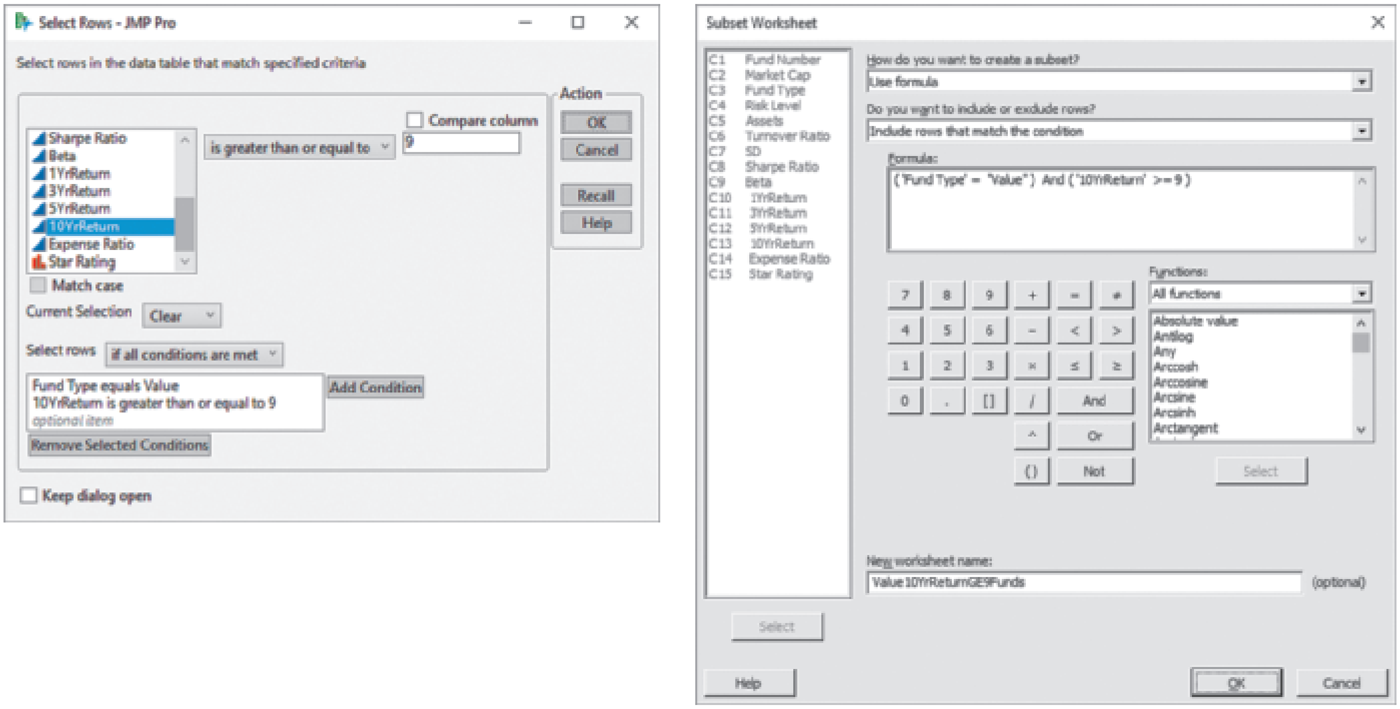
Associated with preparing tabular or visual summaries are two operations that extract subsets of the variables under study. **Data filtering** selects rows of data that match criteria, specified values for specific variables. For example, using the filter that selects all rows in which Fund Type is value, would select 173 rows from the sample of 479 retirement funds that this chapter uses in various examples. In the context of this chapter, **querying** can be a more interactive version of filtering and a method that may not select all of the columns of the matching rows depending how the querying is done.

*Chapter 1 discusses the same JMP Select Rows dialog box in the context of data cleaning.*

Excel, JMP, and Minitab all have data filtering and query features that vary in their implementation and degree of interactivity and JMP has two complementary ways of filtering a data table). Both JMP and Minitab use row-based filtering that can be expressed as a comparison between a variable and a value or value range. Figure 2.26 shows the JMP Select Rows and the Minitab Subset Worksheet dialog boxes with entries that select all rows in value retirement funds that have ten-year return percentages that are greater than or equal to 9.

**FIGURE 2.26**

JMP and Minitab subsetting dialog boxes for data filtering



Select Rows - JMP Pro

Select rows in the data table that match specified criteria

- Sharpe Ratio
- Beta
- 1YrReturn
- 3YrReturn
- 5YrReturn
- 10YrReturn
- Expense Ratio
- Star Rating

is greater than or equal to: 9

□ Compare column

□ Match case

Current Selection: Clear

Select rows: if all conditions are met

Fund Type equals Value
10YrReturn is greater than or equal to 9
optional item

□ Keep dialog open

Subset Worksheet

C1 Fund Number
C2 Market Cap
C3 Fund Type
C4 Risk Level
C5 Assets
C6 Turnover Ratio
C7 SD
C8 Sharpe Ratio
C9 Beta
C10 1YrReturn
C11 3YrReturn
C12 5YrReturn
C13 10YrReturn
C14 Expense Ratio
C15 Star Rating

How do you want to create a subset? Use formula

Do you want to include or exclude rows? Include rows that match the condition

Formula: ('Fund Type' = "Value") And ('10YrReturn' >= 9)

Functions: All functions

New worksheet name: Value10YrReturnGE9Funds (optional)

In Excel, selecting **Data ➔ Filter** displays pull-down menus for each column in row 1 cells. In those menus, check boxes for each unique value in the column appear and check boxes can be cleared or checked to select specific values or ranges. Excel also contains *slicers* that filter and query data from a PivotTable.

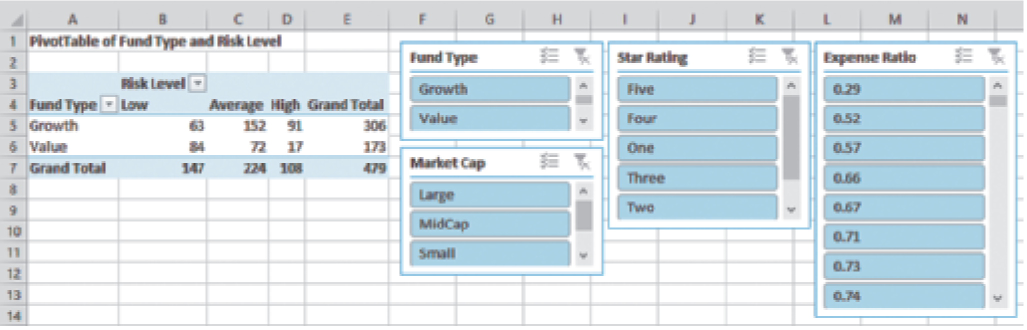
## Excel Slicers

A **slicer** is a panel of clickable buttons that appears superimposed over a worksheet and is unique to a variable in the associated PivotTable. Each button in a slicer represents a unique value of the variable as found in the source data for the PivotTable. You create a slicer for any variable that has been *associated* with a PivotTable, whether or not a variable has been inserted into the PivotTable. This allows you to work with many variables at once in a way that avoids creating an overly complex multidimensional contingency table that would be hard to comprehend and interpret.

By clicking buttons in the slicer panels, you query the data. For example, the Figure 2.27 worksheet contains slicers for the Fund Type, Market Cap, Star Rating, and Expense Ratio variables and a PivotTable that has been associated with the variables stored in the DATA worksheet of the Retirement Funds workbook.

**FIGURE 2.27**
PivotTable and slicers for the retirement funds sample data

*In JMP, selecting* **Rows→Data Filter** *displays the Data Filter window that contains buttons and sliders analogous to the Excel slider panel.*

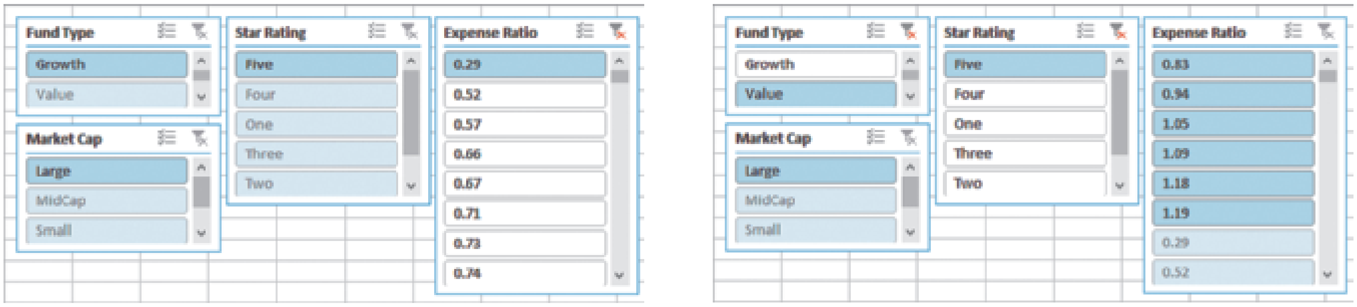


PivotTable of Fund Type and Risk Level

|  | Risk Level |  |  |  |
|---|---|---|---|---|
| Fund Type | Low | Average | High | Grand Total |
| Growth | 63 | 152 | 91 | 306 |
| Value | 84 | 72 | 17 | 173 |
| Grand Total | 147 | 224 | 108 | 479 |

With these four slicers, you can ask questions about the data, for example, "What are the attributes of the fund(s) with the lowest expense ratio?" and "What are the expense ratios associated with large market cap value funds that have a star rating of five?" These questions can be answered by clicking the appropriate buttons of the four slicers. For example, Figure 2.28 displays slicers that answer the two questions. Note that Excel dims, or disables, the buttons representing values that the current data filtering excludes to highlight answers. For example, the answer to the first question is a growth fund with a large market cap and a five-star rating. (The updated PivotTable display, not shown in Figure 2.28, reveals that there is only one such fund.) For the second question, the answer is that 0.83, 0.94, 1.05, 1.09, 1.18, and 1.19 are the expense ratio percentages associated with large market cap value funds that have a star rating of five. (The updated PivotTable display reveals that there are 6 funds with those attributes.)

**FIGURE 2.28**
Slicer displays for answers to questions



## PROBLEMS FOR SECTIONS 2.6 THROUGH 2.8

### APPLYING THE CONCEPTS

**2.58** Using the sample of retirement funds stored in **Retirement Funds**:

**a.** Construct a table that tallies Fund Type, Market Cap, and Star Rating.

**b.** What conclusions can you reach concerning differences among the types of retirement funds (growth and value), based on Market Cap (small, mid-cap, and large) and Star Rating (one, two, three, four, and five)?

**c.** Construct a table that computes the average three-year return for each fund type, market cap, and star rating.

**d.** Drill down to examine the large cap growth funds with a rating of three. How many funds are there? What conclusions can you reach about these funds?

**2.59** Using the sample of retirement funds stored in **Retirement Funds**:

**a.** Construct a table that tallies, Market Cap, Risk Level, and Star Rating.

**b.** What conclusions can you reach concerning differences among the funds based on Market Cap (small, mid-cap, and large), Risk Level (low, average, and high), and Star Rating (one, two, three, four, and five)?

**c.** Construct a table that computes the average three-year return for each market cap, risk level, and star rating.

**d.** Drill down to examine the large cap funds that are high risk with a rating of three. How many funds are there? What conclusions can you reach about these funds?

**2.60** Using the sample of retirement funds stored in **Retirement Funds**:

**a.** Construct a table that tallies Fund Type, Risk Level and Star Rating.
**b.** What conclusions can you reach concerning differences among the types of retirement funds, based on the risk levels and star ratings?
**c.** Construct a table that computes the average three-year return for each fund type, risk level, and star rating.
**d.** Drill down to examine the growth funds with high risk with a rating of three. How many funds are there? What conclusions can you reach about these funds?

**2.61** Using the sample of retirement funds stored in **Retirement Funds**:

**a.** Construct a table that tallies type, market cap, risk, and rating.
**b.** What conclusions can you reach concerning differences among the types of funds based on market cap categories, risk levels, and star ratings?
**c.** Which do you think is easier to interpret: the table for this problem or the ones for problems 2.58 through 2.60? Explain.
**d.** Compare the results of this table with those of Figure 2.19 and problems 2.58 through 2.60. What differences can you observe?

**2.62** In the sample of 479 retirement funds (**Retirement Funds**), what are the attributes of the fund with the highest five-year return?

**2.63** Using the sample of retirement funds stored in **Retirement Funds**:

**a.** Construct a chart that visualizes SD and Assets by Risk Level.
**b.** Construct a chart that visualizes SD and Assets by Fund Type. Rescale the Assets axis, if necessary, to see more detail.
**c.** How do the patterns that you can observe in both charts differ? What data relationships, if any, do those patterns suggest?

**2.64** In the sample of 479 retirement funds (**Retirement Funds**), which funds in the sample have the lowest five-year return?

**2.65** Using the sample of retirement funds stored in **Retirement Funds**:

**a.** Construct one chart that visualizes 10YrReturn and 1YrReturn by Market Cap.
**b.** Construct one chart that visualizes 5YrReturn and 1YrReturn by Market Cap.
**c.** How does the patterns to the points of each market cap category change between the two charts?
**d.** What can you deduce about return percentages in years 6 through 10 included in 10YrReturn but not included in 5YrReturn?

**2.66** In the sample of 479 retirement funds (**Retirement Funds**), what characteristics are associated with the funds that have the lowest five-year return?

**2.67** The data in **NewHomeSales** includes the median sales price of new single-family houses sold in the United States recorded at the end of each month from January 2000 through December 2016.

Source: Data extracted from **bit.ly/2eEcIBR**, March 19, 2017.

**a.** Construct sparklines of new home sales prices by year.
**b.** What conclusions can you reach concerning the median sales price of new single-family houses sold in the United States from January 2000 through December 2016?
**c.** Compare the sparklines in (a) to the time-series plot in Problem 2.55 on p. 102.

**2.68** The file **Natural Gas** includes the monthly average commercial price for natural gas (dollars per thousand cubic feet) in the United States from January 1, 2008, to December 2016.

Source: Data extracted from "U.S. Natural Gas Prices," **bit.ly/2oZIQ5Z**, March 19, 2017.

**a.** Construct a sparkline of the monthly average commercial price for natural gas (dollars per thousand cubic feet) by year.
**b.** What conclusions can you reach concerning the monthly average commercial price for natural gas (dollars per thousand cubic feet)?

# 2.9 Pitfalls in Organizing and Visualizing Variables

The tabular and visual summaries that you create when you organize and visualize your variables of interest can jumpstart the analysis of the variables. However, you must be careful not to produce results that will be hard to comprehend and interpret or to present your data in ways that undercut the usefulness of the methods discussed in this chapter. You can too easily create summaries that obscure the data or create false impressions that would lead to misleading or unproductive analysis. The challenge in organizing and visualizing variables is to avoid these complications.

## Obscuring Data

Management specialists have long known that information overload, presenting too many details, can obscure data and hamper decision making (see reference 4). Both tabular summaries and visualizations can suffer from this problem. For example, consider the Figure 2.29 side-by-side bar chart that shows percentages of the overall total for subgroups formed from combinations of fund type, market cap, risk, and star rating. While this chart highlights that there are more large-cap retirement funds with low risk and a three-star rating than any other combination of risk and star rating, other details about the retirement funds

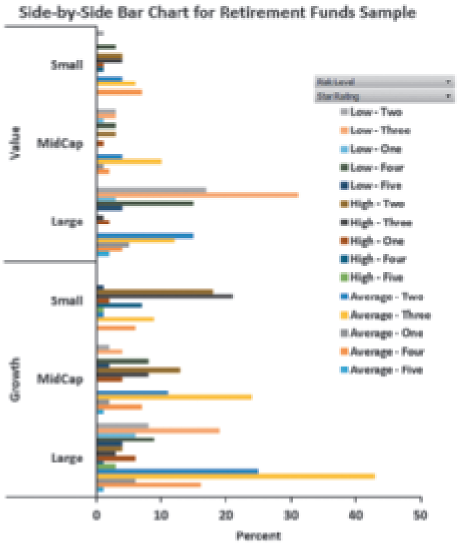


**FIGURE 2.29**
Side-by-side bar chart for the retirement funds sample showing percentage of overall total for fund type, market cap, risk, and star rating

sample are less obvious. The overly complex legend obscures too, and suggests that an equivalent multidimensional contingency table, with 30 joint response cells, would be obscuring, if not overwhelming, for most people.

## Creating False Impressions

As you organize and visualize variables, you must be careful not to create false impressions that could affect preliminary conclusions about the data. Selective summarizations and improperly constructed visualizations often create false impressions.

A *selective summarization* is the presentation of only part of the data that have been collected. Frequently, selective summarization occurs when data collected over a long period of time are summarized as percentage changes for a shorter period. For example, Table 2.14 (left) presents the one-year difference in sales of seven auto industry companies for the month of April. The selective summarization tells a different story, particularly for company G, than does Table 2.14 (right) that shows the year-to-year differences for a three-year period that included the 2008 economic downturn.

**TABLE 2.14**
Left: One-Year Percentage Change in Year-to-Year Sales for the Month of April; Right: Percentage Change for Three Consecutive Years

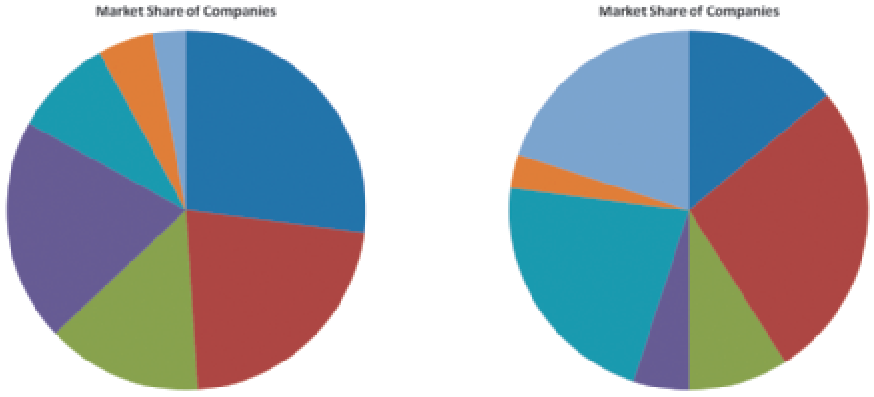
| Company | Change from Prior Year |
|---|---|
| A | +7.2 |
| B | +24.4 |
| C | +24.9 |
| D | +24.8 |
| E | +12.5 |
| F | +35.1 |
| G | +29.7 |

| | Change from Prior Year | | |
|---|---|---|---|
| Company | Year 1 | Year 2 | Year 3 |
| A | −22.6 | −33.2 | +7.2 |
| B | −4.5 | −41.9 | +24.4 |
| C | −18.5 | −31.5 | +24.9 |
| D | −29.4 | −48.1 | +24.8 |
| E | −1.9 | −25.3 | +12.5 |
| F | −1.6 | −37.8 | +35.1 |
| G | +7.4 | −13.6 | +29.7 |

Improperly constructed charts can also create false impressions. Figure 2.30 shows two pie charts that display the market shares of companies in two industries. How quickly did you notice that both pie charts summarize identical data?

**FIGURE 2.30**
Market shares of companies in "two" industries

*If you want to verify that the two pie charts visualize the same data, open the TwoPies worksheet in the Challenging workbook.*



**studentTIP**
Order pie or doughnut slices from the largest to the smallest slice and color pie and doughnut charts meant for comparison in the same way.

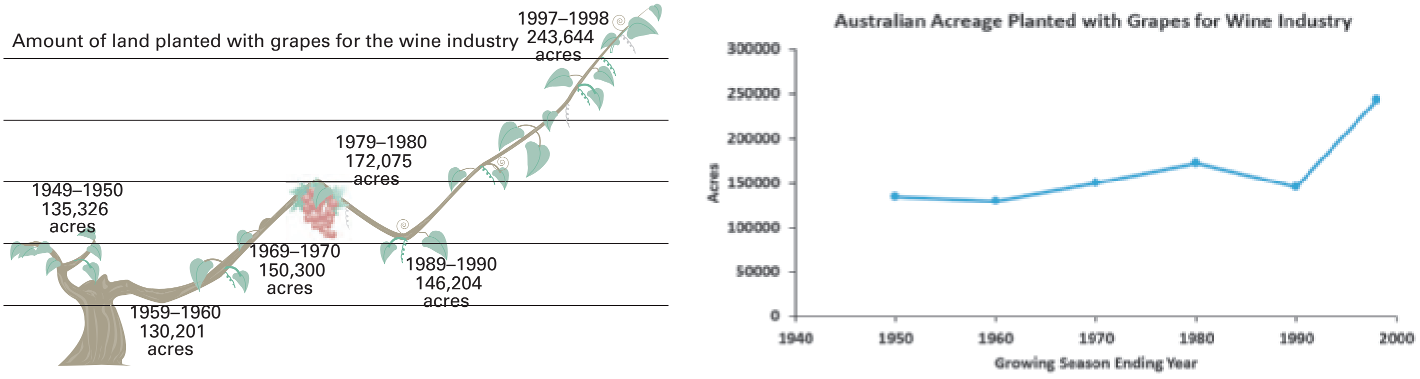
Because of their relative positions and colorings, many people will perceive the dark blue pie slice on the left chart to have a smaller market share than the dark red pie chart on the right chart even though both pie slices represent the company that has 27% market share. In this case, both the ordering of pie slices and the different colorings of the two pie charts contribute to creating the false impression. With other types of charts, improperly scaled axes or a *Y* axis that either does not begin at the origin or is a "broken" axis that is missing intermediate values are other common mistakes that create false impressions.

## Chartjunk

Seeking to construct a visualization that can more effectively convey an important point, some people add decorative elements to enhance or replace the simple bar and line shapes of the visualizations discussed in this chapter. While judicious use of such elements may aid in the memorability of a chart (see reference 1), most often such elements either obscure the data or, worse, create a false impression of the data. Such elements are called **chartjunk**.

Figure 2.31 presents a visualization that illustrates mistakes that are common ways of creating chartjunk unintentionally. The grapevine with its leaves and bunch of grapes adds to the clutter of decoration without conveying any useful information. The chart inaccurately shows the 1949–1950 measurement (135,326 acres) at a *higher* point on the *Y* axis than larger values such as the 1969–1970 measurement, 150,300 acres. The inconsistent scale of the *X* axis distorts the time variable. (The last two measurements, eight years apart, are drawn about as far apart as the 30-year gap between 1959 and 1989.) All of these errors create a very wrong impression that obscures the important trend of accelerating growth of land planted in the 1990s.

**FIGURE 2.31**
Two visualizations of the amount of land planted with grapes for the wine industry



Left illustration adapted from S. Watterson, "Liquid Gold—Australians Are Changing the World of Wine. Even the French Seem Grateful," *Time*, November 22, 1999, p. 68–69.

To help you produce results that will not be hard to comprehend and interpret and that avoid distortions in your data, Exhibit 2.1 summarizes the best practices for creating visual summaries. If you use Excel, be aware that Excel may tempt you to use uncommon chart types and may produce charts that violate some of the best practices that the exhibit lists.

**EXHIBIT 2.1**

### Best Practices for Creating Visual Summaries

- Use the simplest possible visualization
- Include a title and label all axes
- Include a scale for each axis if the chart contains axes
- Begin the scale for a vertical axis at zero and use a constant scale
- Avoid 3D or "exploded" effects and the use of chartjunk
- Use consistent colorings in charts meant to be compared
- Avoid using uncommon chart types including radar, surface, cone, and pyramid charts

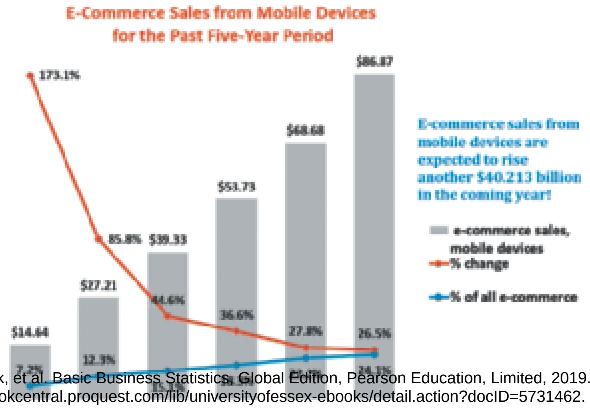
## PROBLEMS FOR SECTION 2.9

### APPLYING THE CONCEPTS

**2.69** **(Student Project)** Bring to class a chart from a website, newspaper, or magazine published recently that you believe to be a poorly drawn representation of a numerical variable. Be prepared to submit the chart to the instructor with comments about why you believe it is inappropriate. Do you believe that the intent of the chart is to purposely mislead the reader? Also, be prepared to present and comment on this in class.

**2.70** **(Student Project)** Bring to class a chart from a website, newspaper, or magazine published this month that you believe to be a poorly drawn representation of a categorical variable. Be prepared to submit the chart to the instructor with comments about why you consider it inappropriate. Do you believe that the intent of the chart is to purposely mislead the reader? Also, be prepared to present and comment on this in class.

**2.71** Examine the following visualization, adapted from one that appeared in a post in a digital marketing blog.



**a.** Describe at least one good feature of this visual display.

**b.** Describe at least one bad feature of this visual display.

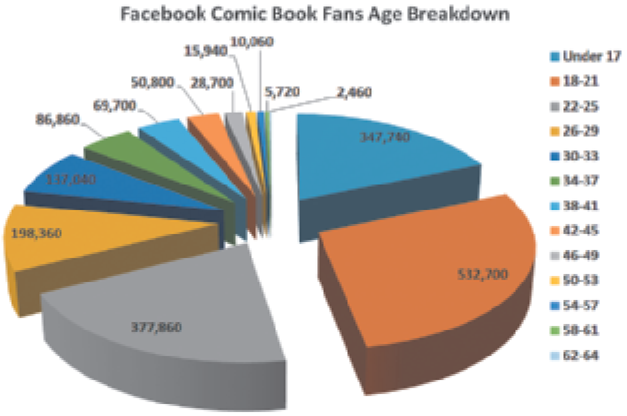
**c.** Redraw the graph, using the guidelines above.

**2.72** Examine the visualization on page 113, adapted from one that appeared in the post "Who Are the Comic Book Fans on Facebook?" on February 2, 2013, as reported by **graphicspolicy.com**.

**a.** Describe at least one good feature of this visual display.

**b.** Describe at least one bad feature of this visual display.

**c.** Redraw the graph, by using best practices given in Exhibit 2.1 above.


Facebook Comic Book Fans Age Breakdown
Under 17