2-2 Pie and Bar Charts

As we discussed in Section 2-1, the only allowable calculation on nominal data is to count the frequency of each value of the variable. We can summarize the data in a table that presents the categories and their counts called a **frequency distribution**. A **relative frequency distribution** lists the categories and the proportion with which each occurs. We can use graphical techniques to present a picture of the data. There are two graphical methods we can use: the **bar chart** and the **pie chart**.

EXAMPLE 2.1

Light Beer Preference Survey

[Xm02-01*] In 2012, total light beer sales was approximately $50 billion (*Source:* Bloomberg.com). With this large a market, breweries often need to know more about who is buying their products. The marketing manager of a major brewery wanted to analyze the light beer sales among college and university students who drink light beer. A random sample of 285 graduating students was asked to report which of the following is their favorite light beer:

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

The responses were recorded using the codes 1, 2, 3, 4, 5, 6, and 7, respectively. The data are listed here, and the entire data set is stored on our Web site. The name of the file is listed in the margin. The file also contains each graduate's identification number and gender. The additional data are not needed in this example but will be used later in this book to produce other information for the manager. (Examples and exercises with additional data are indicated with an asterisk next to the file name.)

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

SOLUTION

To extract useful information requires the application of a statistical or graphical technique. To choose the appropriate technique, we must first identify the type of data. In this example, the data are nominal because the numbers represent categories. The only calculation permitted on nominal data is to count the number of occurrences of each category. Hence, we count the number of 1s, 2s, 3s, 4s, 5s, 6s, and 7s. The list of the categories and their counts constitute the frequency distribution. The relative frequency

Light Beer Preferences

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

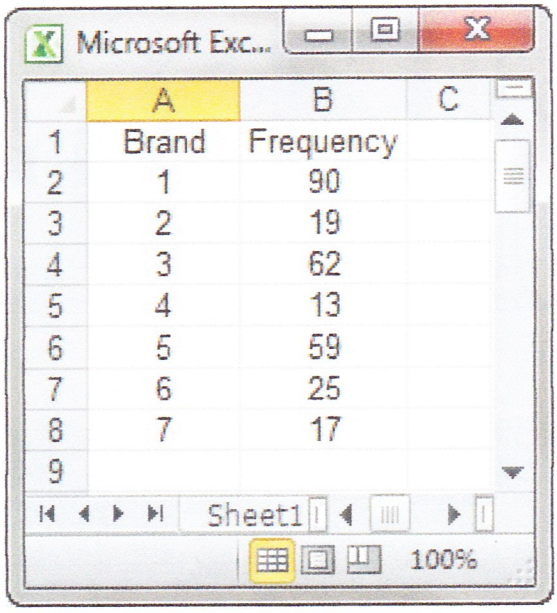
Table 2.1 Frequency and Relative Frequency Distributions for Example 2.1

Light Beer Brand	Frequency	Relative Frequency (%)
Bud Light	90	31.6
Busch Light	19	6.7
Coors Light	62	21.8
Michelob Light	13	4.6
Miller Lite	59	20.7
Natural Light	25	8.8
Other brands	17	6.0
Total	285	100

distribution is produced by converting the frequencies into proportions. The frequency and relative frequency distributions are combined in Table 2.1.

As we promised in Chapter 1 (and the preface) we demonstrate the solution of all examples in this book using three approaches (where feasible): manually, using Excel, and using Minitab.

Excel



INSTRUCTIONS

1. Type or import the data into one or more columns. (Open Xm02-01.)
2. Activate any empty cell and type

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

Input range are the cells that contain the data. In this example, the range is B1:B286. The criteria are the codes you want to count: (1) (2) (3) (4) (5) (6) (7). To count the number of 1s (Bud Light), type

```
=COUNTIF (B1:B286, 1)
```

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

Minitab

Tally for Discrete Variables: Brand

Brand	Count	Percent
1	90	31.58
2	19	6.67
3	62	21.75
4	13	4.56
5	59	20.70
6	25	8.77
7	17	5.96

$N = 285$

INSTRUCTIONS

(Specific commands for this example are highlighted.)

1. Type or import the data into one column. (Open Xm02-01.)
2. Click **Stat**, **Tables**, and **Tally Individual Variables**.
3. Type or use the **Select** button to specify the name of the variable or the column where the data are stored in the **Variables** box (Brand). Under **Display**, click **Counts** and **Percents**.

INTERPRET

Budweiser, Coors, and Miller are by far the most popular light beers among college and university seniors.

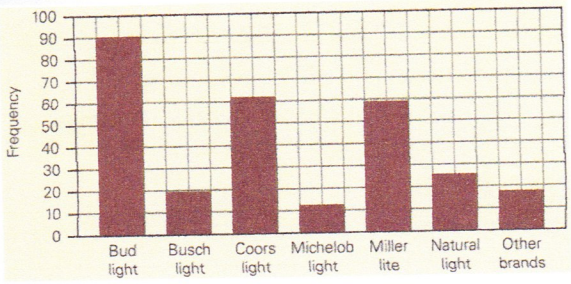
2-2a Bar and Pie Charts

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

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

If we wish to emphasize the relative frequencies instead of drawing the bar chart, we draw the pie chart. A pie chart is simply a circle subdivided into slices that represent the categories. It is drawn so that the size of each

Figure 2.1 Bar Chart for Example 2.1



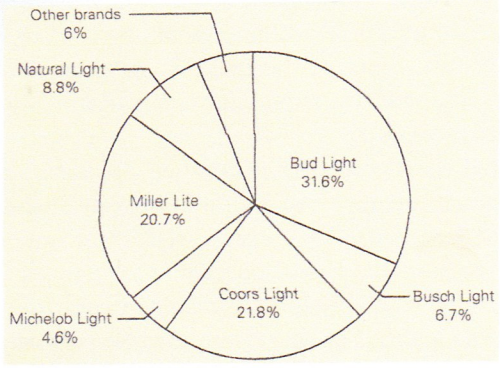
slice is proportional to the percentage corresponding to that category. For example, because the entire circle is composed of 360 degrees, a category that contains 25% of the observations is represented by a slice of the pie that contains 25% of 360 degrees, or 90 degrees. The number of degrees for each category in Example 2.1 is shown in Table 2.2.

Table 2.2 Proportion in Each Category in Example 2.1

Light Beer Brand	Relative Frequency (%)	Slice of the Pie (Degrees)
Bud Light	31.6	113.7
Busch Light	6.7	24.0
Coors Light	21.8	78.3
Michelob Light	4.6	16.4
Miller Lite	20.7	74.5
Natural Light	8.8	31.6
Other brands	6.0	21.5
Total	100.00	360

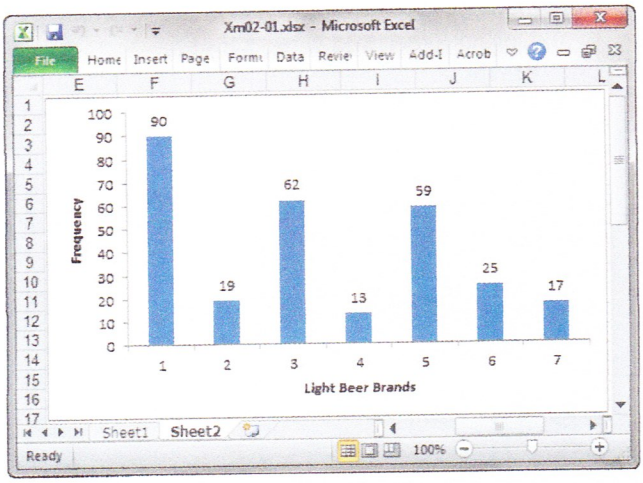
Figure 2.2 was drawn from these results.

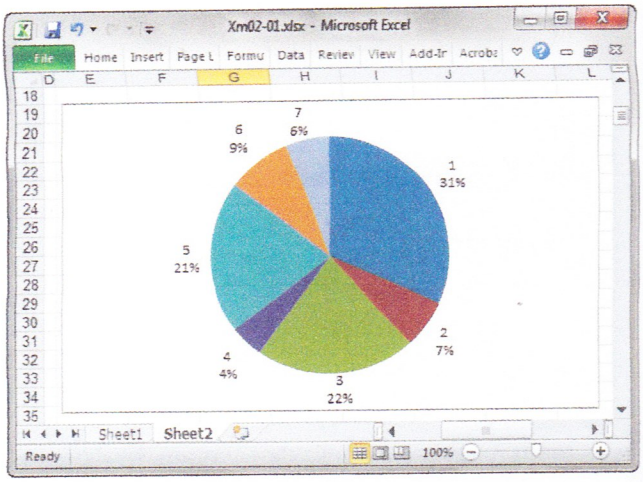
Figure 2.2 Pie Chart for Example 2.1



Excel

Here are Excel's bar and pie charts.

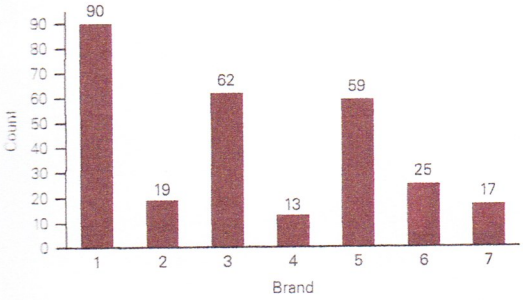


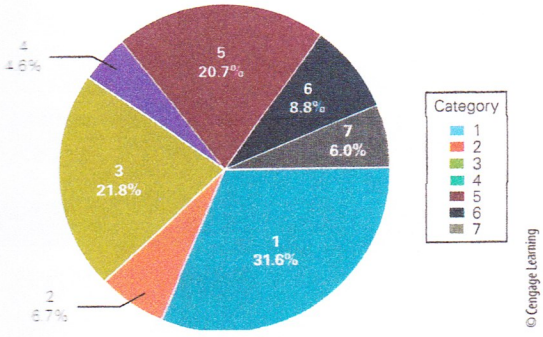


INSTRUCTIONS

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

Minitab





INSTRUCTIONS

1. Type or import the data into one column. (Open Xm02-01.)

For a bar chart:

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

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

For a pie chart:

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

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

INTERPRET

The bar chart focuses on the frequencies. As you can see, Bud Light is the most popular light beer, with 90 college and university seniors selecting it as their favorite. Coors Light and Miller Lite are the second and third most popular light beers.

The pie chart focuses on the proportions. Bud Light is the choice of almost one-third of college seniors.

GENERAL SOCIAL SURVEY

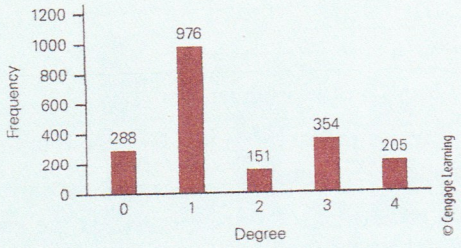
HOW EDUCATED ARE AMERICAN ADULTS: SOLUTION

In this problem, the data are ordinal because the responses are in order from least amount of education to most. However, we will treat the data as nominal. To summarize a set of nominal data, we can use a frequency distribution, a bar chart, or a pie chart. We'll use all three to answer the question.

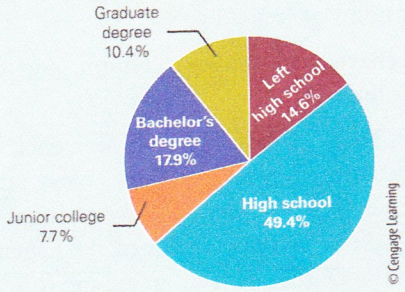
Frequency and Relative Frequency Distribution

Education	Frequency	Relative Frequency (%)
Left high school	288	14.6
Completed high school	976	49.4
Completed junior college	151	7.7
Completed bachelor's degree	354	17.9
Completed graduate degree	205	10.4
TOTAL	1,974	100

Bar Chart



Pie Chart

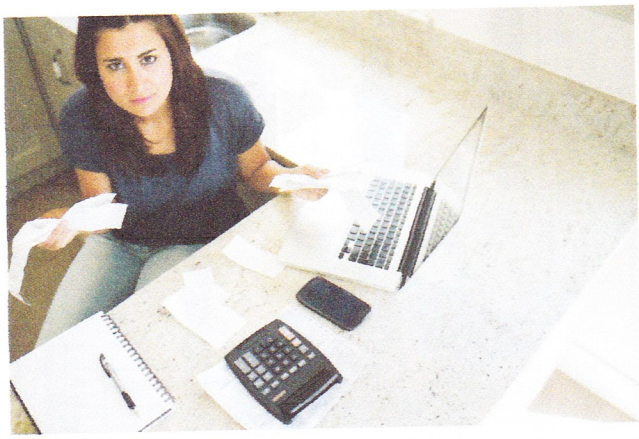


The relative frequencies and the pie chart provide the most information. About 50% of American adults finished high school only, and more than a quarter have a bachelor's degree or more.

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

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

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



© wavebreakmedia/Shutterstock.com

EXAMPLE 2.2

Analysis of Long-Distance Telephone Bills

[Xm02-02] After the telephone industry was deregulated, several new companies were created to compete in the business of providing long-distance telephone service. In almost all cases, these companies competed on price because the service each offered was similar. Pricing a service or product in the face of stiff competition is difficult. Factors to be considered include supply, demand, price elasticity, and the actions of competitors. Long-distance packages may employ per-minute charges, a flat monthly rate, or some combination of the two. Determining the appropriate rate structure is facilitated by acquiring information about the behaviors of customers, especially the size of monthly long-distance bills.

© Glowimages

2-3 Histograms and Stem-and-Leaf Displays

In this section, we introduce two graphical methods that are used when the data are interval. The more important of these graphical methods is the histogram. As you will see, the histogram not only is a powerful graphical technique used to summarize interval data but also is used to help explain an important aspect of probability (see Chapter 7).

Long-Distance Telephone Bills

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

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

SOLUTION

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

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

CLASSES

Amounts that are . . .
Less than or equal to 15

Amounts that are . . .
More than 15 but less than or equal to 30

Amounts that are . . .
More than 30 but less than or equal to 45

Amounts that are . . .
More than 45 but less than or equal to 60

Amounts that are . . .
More than 60 but less than or equal to 75

Amounts that are . . .
More than 75 but less than or equal to 90

Amounts that are . . .
More than 90 but less than or equal to 105

Amounts that are . . .
More than 105 but less than or equal to 120

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

To create the frequency distribution manually, we count the number of observations that fall into each interval. Table 2.3 presents the frequency distribution.

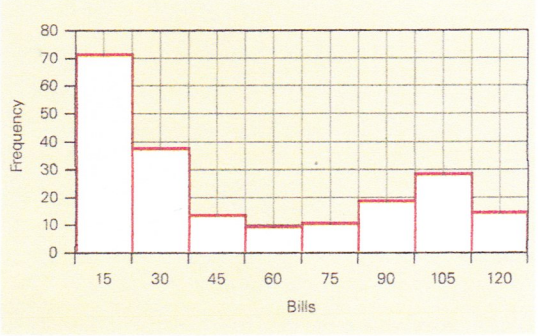
Table 2.3 Frequency Distribution of the Long-Distance Bills in Example 2.2

Class Limits	Frequency
0 to 15*	71
15 to 30	37
30 to 45	13
45 to 60	9
60 to 75	10
75 to 90	18
90 to 105	28
105 to 120	14
Total	200

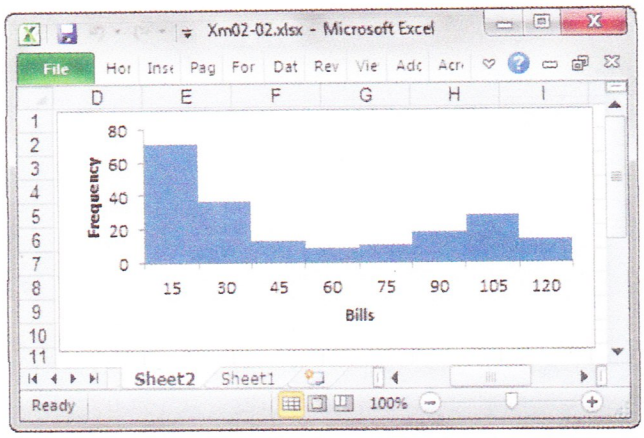
*Classes contain observations greater than their lower limits (except for the first class) and less than or equal to their upper limits.

Although the frequency distribution provides information about how the numbers are distributed, the information is more easily understood and imparted by drawing a picture or graph. The graph is called a **histogram**. A histogram is created by drawing rectangles whose bases are the intervals and whose heights are the **frequencies**. Figure 2.3 exhibits the histogram that was drawn by hand.

Figure 2.3 Histogram for Example 2.2



Excel



INSTRUCTIONS

1. Type or import the data into one column. (Open Xm02-02.) In another column, type the upper limits of the class intervals. Excel calls them *bins*. (You can put any name in the first row; we typed "Bills.")
2. Click **Data**, **Data Analysis**, and **Histogram**. If Data Analysis does not appear in the menu box, see Web site Appendix A1.
3. Specify the **Input Range** (A1:A201) and the **Bin Range** (B1:B9). Click **Chart Output**. Click **Labels** if the first row contains names.
4. To remove the gaps, place the cursor over one of the rectangles and click the right button of the mouse. With the left button, click **Format Data Series . . .** Move the pointer to **Gap Width** and use the slider to change the number from 150 to 0.

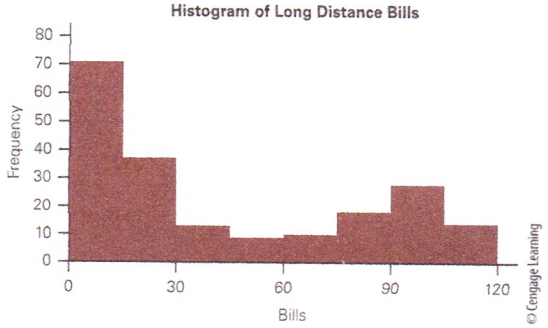
Except for the first class, Excel counts the number of observations in each class that are greater than the lower limit and less than or equal to the upper limit.

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

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

If you have difficulty with this technique, turn to Web site Appendix A2 or A3; both provide step-by-step instructions for Excel as well as troubleshooting tips.

Minitab



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

INSTRUCTIONS

1. Type or import the data into one column. (Open Xm02-02.)
2. Click **Graph, Histogram . . .**, and **Simple**.
3. Type or use the **Select** button to specify the name of the variable in the **Graph variables** box (Bills). Click **Data View**.
4. Click **Data Display** and **Bars**. Minitab will create a histogram using its own choices of class intervals.
5. To choose your own classes, double-click the horizontal axis. Click **Binning**.
6. Under **Interval Type** choose **Cutpoint**. Under **Interval Definition**, choose **Midpoint/Cutpoint positions** and type in your choices (0 15 30 45 60 75 90 105 120).

INTERPRET

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

2-3a Determining the Number of Class Intervals

The number of class intervals we select depends entirely on the number of observations in the data set. The more observations we have, the larger the number of class intervals we need to use to draw a useful histogram. Table 2.4 provides guidelines on choosing the number of classes. In Example 2.2 we had 200 observations. The table tells us to use 7, 8, 9, or 10 classes.

Table 2.4 Approximate Number of Classes in Frequency Distributions

Number of Observations	Number of Classes
Less than 50	5–7
50–200	7–9
200–500	9–10
500–1,000	10–11
1,000–5,000	11–13
5,000–50,000	13–17
More than 50,000	17–20

2-3b Class Interval Widths

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

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

In Example 2.2, we calculated

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

We often round the result to some convenient value. We then define our class limits by selecting a lower limit for the first class from which all other limits are determined. The only condition we apply is that the first class interval must contain the smallest observation. In Example 2.2, we rounded the class width to 15 and set the lower limit of the first class to 0. Thus, the first class is defined as "Amounts that are greater than or equal to 0 but less than or equal to 15." (Minitab users should remember that the classes are defined as the number of observations that are *strictly less* than their upper limits.)