2.1 Organizing and Graphing Qualitative Data

This section discusses how to organize and display qualitative (or categorical) data. Data sets are organized into tables and displayed using graphs. First we discuss the concept of raw data.

2.1.1 Raw Data

When data are collected, the information obtained from each member of a population or sample is recorded in the sequence in which it becomes available. This sequence of data recording is random and unranked. Such data, before they are grouped or ranked, are called **raw data**.

Raw Data Data recorded in the sequence in which they are collected and before they are processed or ranked are called *raw data*.

Suppose we collect information on the ages (in years) of 50 students selected from a university. The data values, in the order they are collected, are recorded in Table 2.1. For instance, the first student's age is 21, the second student's age is 19 (second number in the first row), and so forth. The data in Table 2.1 are quantitative raw data.

Table 2.1 Ages of 50 Students

21	19	24	25	29	34	26	27	37	33
18	20	19	22	19	19	25	22	25	23
25	19	31	19	23	18	23	19	23	26
22	28	21	20	22	22	21	20	19	21
25	23	18	37	27	23	21	25	21	24

Suppose we ask the same 50 students about their student status. The responses of the students are recorded in Table 2.2. In this table, F, SO, J, and SE are the abbreviations for freshman, sophomore, junior, and senior, respectively. This is an example of qualitative (or categorical) raw data.

Table 2.2 Status of 50 Students

J	F	SO	SE	J	J	SE	J	J	J
F	F	J	F	F	F	SE	SO	SE	J
J	F	SE	SO	SO	F	J	F	SE	SE
SO	SE	J	SO	SO	J	J	SO	F	SO
SE	SE	F	SE	J	SO	F	J	SO	SO

The data presented in Tables 2.1 and 2.2 are also called **ungrouped data**. An ungrouped data set contains information on each member of a sample or population

individually. If we rank the data of Table 2.1 from lowest to the highest age, they will still be ungrouped data but not raw data.

2.1.2 Frequency Distributions

The Gallup polling agency recently surveyed randomly selected 1015 adults aged 18 and over from all 50 U.S. states and the District of Columbia. These adults were asked, "Please tell me how concerned you are right now about each of the following financial matters, based on your current financial situation—are you very worried, moderately worried, not too worried, or not worried at all." Among a series of financial situations, one such situation was not having enough money to pay their normal monthly bills. Table 2.3 lists the responses of these adults. The Gallup report contained the percent of adults belonging to each category, which we have converted to

numbers in the table. In this table, the variable is how much are adults worried about not having enough money to pay normal monthly bills. The categories representing this variable are listed in the first column of the table. Note that these categories are mutually exclusive. In other words, each of the 1015 adults belongs to one and only one of these categories. The number of adults who belong to a certain category is called the frequency of that category. A **frequency distribution** exhibits how the frequencies are distributed over various categories. Table 2.3 is called a *frequency distribution table* or simply a *frequency table.*

Table 2.3 Worries About Not Having Enough Money to Pay Normal Monthly Bills

Variable → | Response | Number of Adults ← Frequency column

Response	Number of Adults
Very worried	162
Moderately worried	203
Not too worried	305
Not worried at all	325
Others	20
	Sum = 1015

Category → Not too worried; 305 ← Frequency

Source: Gallup Poll.

Frequency Distribution of a Qualitative Variable A *frequency distribution* of a qualitative variable lists all categories and the number of elements that belong to each of the categories.

Example 2-1 illustrates how a frequency distribution table is constructed for a qualitative variable.

Constructing a frequency distribution table for qualitative data.

EXAMPLE 2-1 What Variety of Donuts Is Your Favorite?

© Jack Puccio/iStockphoto

A sample of 30 persons who often consume donuts were asked what variety of donuts is their favorite. The responses from these 30 persons are as follows:

glazed	filled	other	plain	glazed	other
frosted	filled	filled	glazed	other	frosted
glazed	plain	other	glazed	glazed	filled
frosted	plain	other	other	frosted	filled
filled	other	frosted	glazed	glazed	filled

Construct a frequency distribution table for these data.

Solution Note that the variable in this example is *favorite variety of donut*. This variable has five categories (varieties of donuts): glazed, filled, frosted, plain, and other. To prepare a frequency distribution, we record these five categories in the first column of Table 2.4. Then we read each response (each person's favorite variety of donut) from the given information and mark a *tally*, denoted by the symbol |, in the second column of Table 2.4 next to the corresponding category. For example, the first response is *glazed*. We show this in the frequency table by marking a tally in the second column next to the category *glazed*. Note that the tallies are marked in blocks of five for counting convenience. Finally, we record the total of the tallies for each category in the third column of the table. This column is called the *column of frequencies* and is usually denoted by *f*. The sum of the entries in the frequency column gives the sample size or total frequency. In Table 2.4, this total is 30, which is the sample size.

Table 2.4 Frequency Distribution of Favorite Donut Variety

Donut Variety	Tally	Frequency (f)
Glazed	𝍸 |||	8
Filled	𝍸 ||	7
Frosted	𝍸	5
Plain	|||	3
Other	𝍸 ||	7
		Sum = 30

2.1.3 Relative Frequency and Percentage Distributions

The **relative frequency** of a category is obtained by dividing the frequency of that category by the sum of all frequencies. Thus, the relative frequency shows what fractional part or proportion of the total frequency belongs to the corresponding category. A *relative frequency distribution* lists the relative frequencies for all categories.

Calculating Relative Frequency of a Category

$$\text{Relative frequency of a category} = \frac{\text{Frequency of that category}}{\text{Sum of all frequencies}}$$

The **percentage** for a category is obtained by multiplying the relative frequency of that category by 100. A *percentage distribution* lists the percentages for all categories.

Calculating Percentage

$$\text{Percentage} = (\text{Relative frequency}) \cdot 100\%$$

Constructing relative frequency and percentage distributions.

EXAMPLE 2-2 What Variety of Donuts Is Your Favorite?

Determine the relative frequency and percentage distributions for the data in Table 2.4.

Solution The relative frequencies and percentages from Table 2.4 are calculated and listed in Table 2.5. Based on this table, we can state that 26.7% of the people in the sample said that glazed donut is their favorite. By adding the percentages for the first two categories, we can state that 50% of the persons included in the sample said that glazed or filled donut is their favorite. The other numbers in Table 2.5 can be interpreted

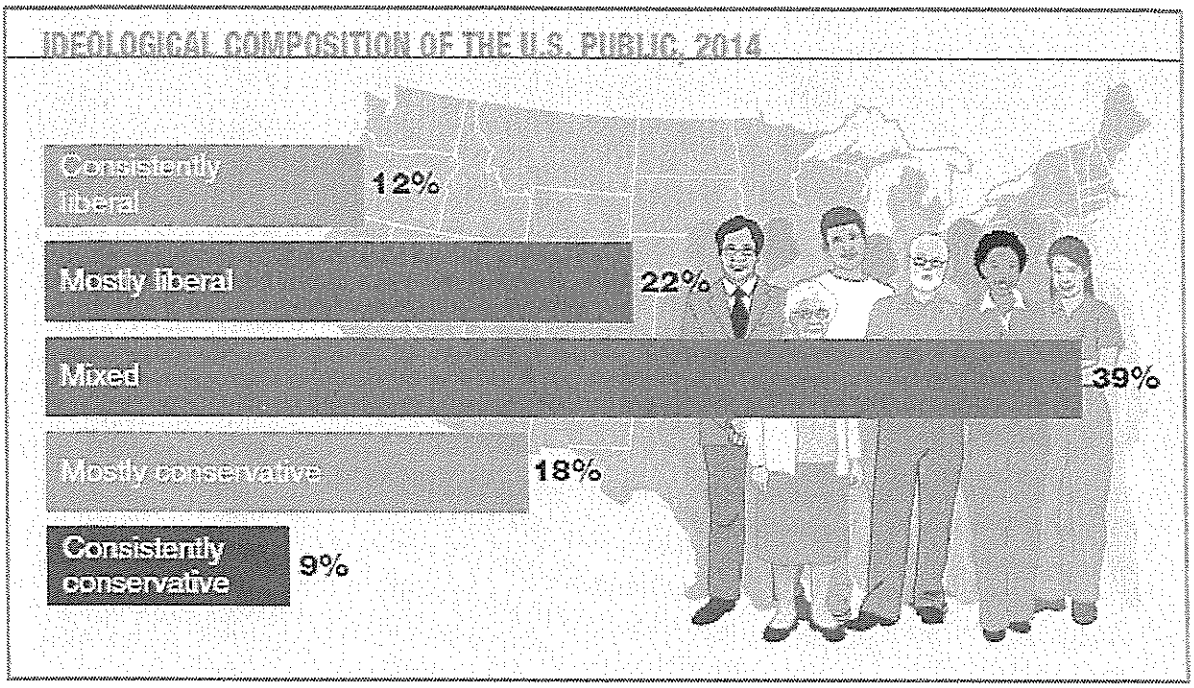
in similar ways.

Table 2.5 Relative Frequency and Percentage Distributions of Favorite Donut Variety

Donut Variety	Relative Frequency	Percentage
Glazed	$8/30 = .267$	$.267(100) = 26.7$
Filled	$7/30 = .233$	$.233(100) = 23.3$
Frosted	$5/30 = .167$	$.167(100) = 16.7$
Plain	$3/30 = .100$	$.100(100) = 10.0$
Other	$7/30 = .233$	$.233(100) = 23.3$
	Sum = 1.000	Sum = 100%

CASE STUDY 2-1

IDEOLOGICAL COMPOSITION OF THE U.S. PUBLIC, 2014



Data source: Pew Research Center

Pew Research Center conducted a national survey of 10,013 adults January 23 to March 16, 2014, to find the political views of adults in the United States. As the above bar chart shows, 12% of the adults polled said that they were consistently liberal, 22% indicated that they were mostly liberal, and so on. In this survey, Pew Research Center also found that, overall, the percentage of Americans who indicated that they were consistently conservative or consistently liberal has increased from 10% to 21% during the past two decades. Note that in this chart, the bars are drawn horizontally.

Source: Pew Research Center, June, 2014 Report: *Political Polarization in the American Public.*

Notice that the sum of the relative frequencies is always 1.00 (or approximately 1.00 if the relative frequencies are rounded), and the sum of the percentages is always 100 (or approximately 100 if the percentages are rounded).

2.1.4 Graphical Presentation of Qualitative Data

All of us have heard the adage "a picture is worth a thousand words." A graphic display can reveal at a glance the main characteristics of a data set. The *bar graph* and the *pie chart* are two types of graphs that are commonly used to display qualitative data.

Bar Graphs

To construct a **bar graph** (also called a *bar chart*), we mark the various categories on the horizontal axis as in Figure 2.1. Note that all categories are represented by intervals of the same width. We mark the frequencies on the vertical axis. Then we draw one bar for each category such that the height of the bar represents the frequency of the corresponding category. We leave a small gap between adjacent bars. Figure 2.1 gives the bar graph for the frequency distribution of Table 2.4.

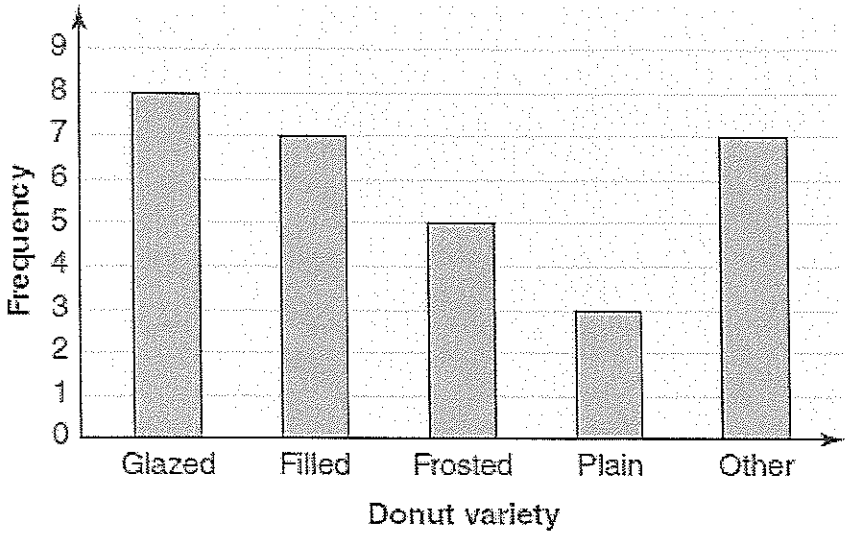
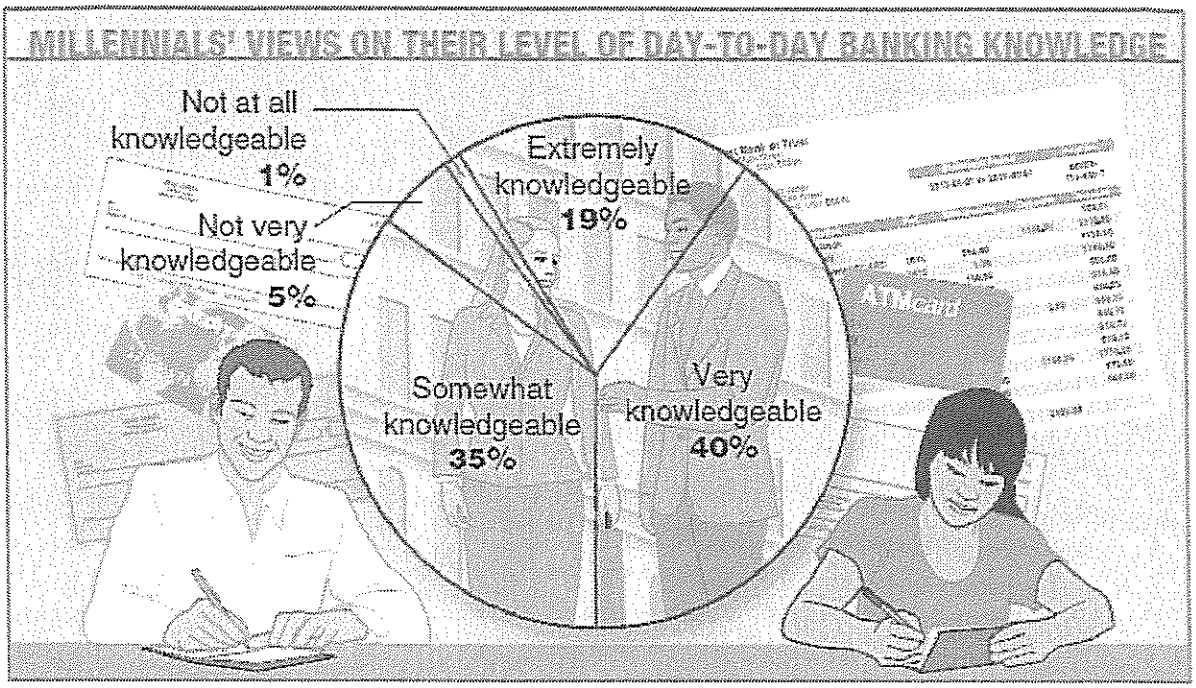


Figure 2.1 Bar graph for the frequency distribution of Table 2.4.

CASE STUDY 2-2

MILLENNIALS' VIEWS ON THEIR LEVEL OF DAY-TO-DAY BANKING KNOWLEDGE



Data source: TD Bank: The Millennial Financial Behaviors & Needs Survey

TD Bank conducted a poll of Millennials (aged 18–34) January 28 to February 10, 2014, with the main focus on understanding their banking behaviors and preferences. As shown in the pie chart, 19% of the Millennials said they were extremely knowledgeable about day-to-day banking, 40% said they were very knowledgeable, 35% said they were somewhat knowledgeable, 5% admitted not to be very knowledgeable, and 1% mentioned that they were not knowledgeable at all.

Source: TD Bank: The Millennial Financial Behaviors & Needs. February 2014.

Bar Graph A graph made of bars whose heights represent the frequencies of respective categories is called a *bar graph*.

The bar graphs for relative frequency and percentage distributions can be drawn simply by marking the relative frequencies or percentages, instead of the frequencies, on the vertical axis.

Sometimes a bar graph is constructed by marking the categories on the vertical axis and the frequencies on the horizontal axis. Case Study 2-1 presents such an example.

Pareto Chart

To obtain a **Pareto chart**, we arrange (in a descending order) the bars in a bar graph

based on their heights (frequencies, relative frequencies, or percentages). Thus, the bar with the largest height appears first (on the left side) in a bar graph and the one with the smallest height appears at the end (on the right side) of the bar graph.

Pareto Chart A Pareto chart is a bar graph with bars arranged by their heights in descending order. To make a Pareto chart, arrange the bars according to their heights such that the bar with the largest height appears first on the left side, and then subsequent bars are arranged in descending order with the bar with the smallest height appearing last on the right side.

Figure 2.2 shows the Pareto chart for the frequency distribution of Table 2.4. It is the same bar chart that appears in Figure 2.1 but with bars arranged based on their heights.

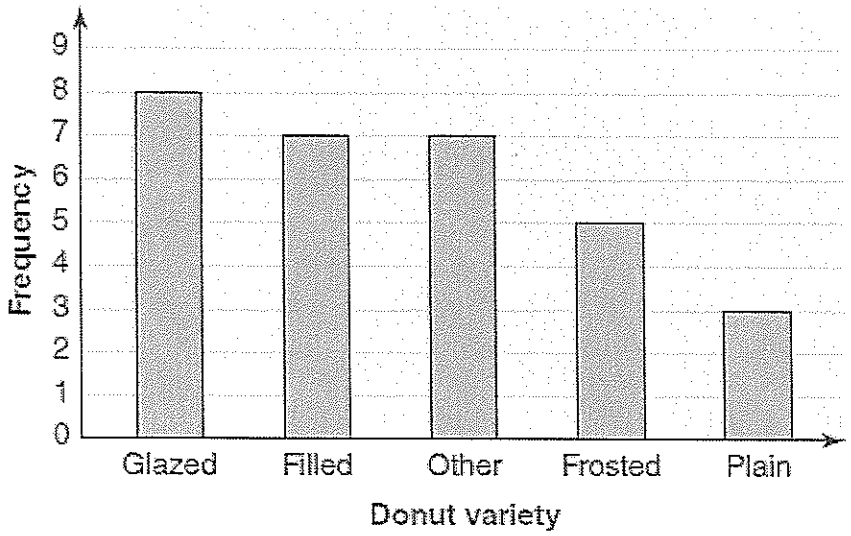


Figure 2.2 Pareto chart for the frequency distribution of Table 2.4.

Pie Charts

A **pie chart** is more commonly used to display percentages, although it can be used to display frequencies or relative frequencies. The whole pie (or circle) represents the total sample or population. Then we divide the pie into different portions that represent the different categories.

Pie Chart A circle divided into portions that represent the relative frequencies or percentages of a population or a sample belonging to different categories is called a *pie chart*.

Figure 2.3 shows the pie chart for the percentage distribution of Table 2.5.

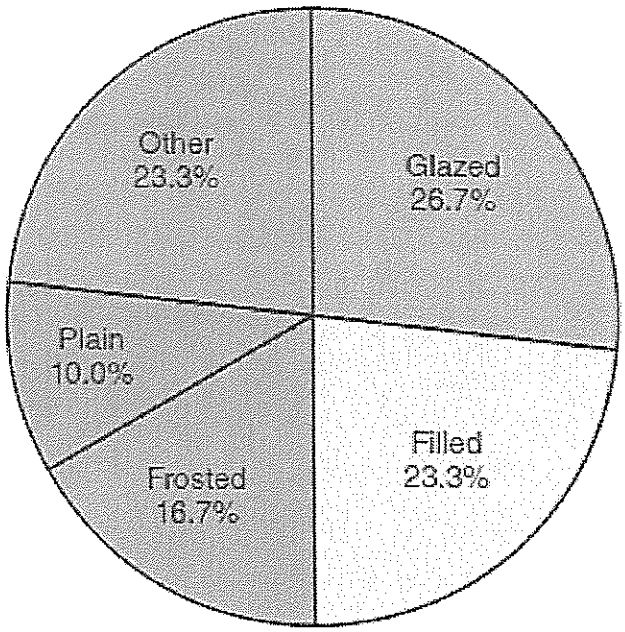


Figure 2.3 Pie chart for the percentage distribution of Table 2.5.

EXERCISES

CONCEPTS AND PROCEDURES

2.1 Why do we need to group data in the form of a frequency table? Explain briefly.

2.2 How are the relative frequencies and percentages of categories obtained from the frequencies of categories? Illustrate with the help of an example.

2.3 The following data give the results of a sample survey. The letters Y, N, and D represent the three categories.

D	N	N	Y	Y	Y	N	Y	D	Y
Y	Y	Y	Y	N	Y	Y	N	N	Y
N	Y	Y	N	D	N	Y	Y	Y	Y
Y	Y	N	N	Y	Y	N	N	D	Y

a.