CHAPTER 2

Describing Data: Frequency Distributions and Graphic Presentation

GOALS

When you have completed this chapter, you will be able to:

- Organize raw data into a *frequency distribution*
- Produce a *histogram*, a *frequency polygon*, and a *cumulative frequency polygon* from quantitative data
- Develop and interpret a *stem-and-leaf display*
- Present qualitative data using such graphical techniques as a *clustered bar chart*, *a stacked bar chart*, and a *pie chart*
- Detect graphic deceptions and use a graph to present data with clarity, precision, and efficiency.

1498
1548
1598
1648
1698
1748
1898
1948
2000

The tendency to graphically represent information seems to be one of the basic human instincts. As such, identification of the oldest such representation is an elusive- task, the earliest known being the map of Konyo, Turkey, dated 6200 B.C. The earliest known bar chart is the one by Bishop N. Oresme (1350).

Most of the modern forms of statistical graphic techniques were invented between 1780 and 1940. In 1786, William Playfair used time-series graphs to depict the amount of import and export to and from England, and in 1801, he published a pie chart to show graphically that the British paid more tax than other countries. The first stacked bar chart, cumulative frequency polygon and histogram were published, respectively, by A. Humboldt (1811), J.B.J. Fourier (1821), and A.M. Guerry (1833). The same period saw development of non-trivial applications of these techniques to real-world problems. One of the most significant contributors in this regard was *the lady with a lamp*, Florence Nightingale.

Florence Nightingale was born in Florence, Italy in 1820, but was raised mostly in Derbyshire, England. In spite of resistance from society and her mother, her father educated her in Greek, Latin, French, German, Italian, history, philosophy, and, her favourite subject, mathematics.

When she was 17 years old, Florence had a spiritual experience. She felt herself called by God to His service. Since that time, she made up her mind to dedicate her life to some social cause. She refused to marry several suitors and at the age of twenty-five, stunned her parents by informing them that she had decided to be a nurse, a profession considered low class at that time.

During the 1854 British war in Crimea, stirred by the reports of primitive sanitation methods at the British barracks' hospital, she volunteered her services, and set out to Scutari, Turkey with a group of 38 nurses. Here, mainly by improving the sanitary conditions and nursing methods, she managed to bring down the mortality rate at the hospital from 42.7 percent to about 2 percent.

On her return to England after the war as a national hero, she dedicated herself to the task of improving the sanitation, and quality of nursing in military hospitals. In this, she encountered strong opposition from the establishment. But with the support of Queen Victoria, and more importantly, with shrewd use of graphic methods (such as stacked bar charts and a new type of polar bar chart that she developed on her own), she succeeded in bringing forth reforms. She was one of the first to use graphical methods in a prescriptive, rather that merely a descriptive way, to bring about social reform.

Over the subsequent 20 years, she applied statistical methods to civilian hospitals, midwifery, Indian public health, and colonial schools. She briefly served as an adviser to the British war office on medical care in Canada. Her mathematical activities included determining "the average speed of transport by sledge," and "the time to transport the sick over immense distances in Canada."

With her statistical analysis, she revolutionized the idea that social phenomena could be objectively measured and subjected to mathematical analyses. Karl Pearson acknowledged her as "prophetess" in the development of applied statistics.

Nightingale held strong opinions on women's rights, and fought for the removal of restrictions that prevented women from having careers. In 1907 she became the first woman to receive the Order of Merit, an order established by King Edward VII for meritorious service.

INTRODUCTION

Rob Whitner is the owner of Whitner Pontiac. Rob's father founded the dealership in 1964, and for more than 30 years they sold exclusively Pontiacs. In the early 1990s Rob's father's health began to fail, and Rob took over more of the dealership's day-to-day operations. At the same time, the automobile business began to change—dealers began to sell vehicles from several manufacturers—and Rob was faced with some major decisions. The first came when another local dealer, who handled Volvos, Saabs, and Volkswagens, approached Rob about purchasing his dealership. After considerable thought and analysis, Rob made the purchase. More recently, the local Jeep Eagle dealership got into difficulty and Rob bought it out. So now, on the same lot, Rob sells the complete line of Pontiacs; the expensive Volvos and Saabs; Volkswagens; and Chrysler products including the popular Jeep line. Whitner Pontiac employs 83 people, including 23 full-time salespeople. Because of the diverse product line, there is quite a bit of variation in the selling price of the vehicles. A top-of-the-line Volvo sells for more than twice the price of a Pontiac Grand Am. Rob would like to develop some charts and graphs that he could review monthly to see where the selling prices tend to cluster, to see the variation in the selling prices, and to note any trends. In this chapter we present techniques that will be useful to Rob or someone like him in managing his business.

2.1 CONSTRUCTING A FREQUENCY DISTRIBUTION OF QUANTITATIVE DATA

Recall from Chapter 1 that we refer to techniques used to describe a set of data as *descriptive statistics*. To put it another way, we use descriptive statistics to organize data in various ways to point out where the data values tend to concentrate and to help distinguish the largest and the smallest values. The first method we use to describe a set of data is a **frequency distribution**. Here our goal is to summarize the data in a table that reveals the shape of the data.

> *i* **Frequency distribution** A grouping of data into non-overlapping classes (mutually exclusive classes or categories) showing the number of observations in each class. The range of classes includes all values in the data set (collectively exhaustive categories).

How do we develop a frequency distribution? The first step is to tally the data into a table that shows the classes and the number of observations in each class. The steps in constructing a frequency distribution are best described using an example. Remember that our goal is to make a table that will quickly reveal the shape of the data.

Example 2-1

In the introduction to this chapter, we described a case where Rob Whitner, owner of Whitner Pontiac, is interested in collecting information on the selling prices of vehicles sold at his dealership. What is the typical selling price? What is the highest selling price? What is the lowest selling price? Around what value do the selling prices tend to cluster? To answer these questions, we need to collect data. According to sales records, Whitner Pontiac sold 80 vehicles last month. The price paid by the customer for each vehicle is shown in Table 2-1. Summarize the selling prices of the vehicles sold last month. Around what value do the selling prices tend to cluster?

TABLE 2-1: Selling Prices ($) at Whitner Pontiac Last Month

31 373	26 879	31 710	36 442	37 657	21 969	23 132
39 552	42 923	25 544	31 060	50 596	25 026	26 252
32 778	32 839	33 277	39 532	19 320	19 920	25 984
34 266	38 552	33 160	37 642	26 009	26 186	22 109
26 418	34 306	25 699	31 812	36 364	27 558	26 492
31 978	35 085	36 438	45 086	27 169	29 231	32 420
35 110	19 702	23 505	50 719	22 175	23 050	26 728
28 400	28 831	25 149	30 518	25 819	27 154	27 661
30 561	35 859	38 339	40 157	45 417	24 470	28 859
29 836	33 219	34 571	39 018	27 168	31 744	32 678
42 588	29 940	22 932	27 439	35 784	26 865	28 576
28 704	32 795	31 103				

Solution

Table 2-1 contains *quantitative data* (recall from Chapter 1). These data are *raw* or *ungrouped data*. With a little searching, we can find the lowest selling price ($19 320) and the highest selling price ($50 719), but that is about all. It is difficult to get a feel for the shape of the data by mere observation of the raw data. The raw data are more easily interpreted if they are organized into a frequency distribution. The steps for organizing data into a frequency distribution are outlined below.

1. **Decide how many classes you wish to use.** The goal is to use just enough groupings or **classes** to reveal the shape of the distribution. Some judgment is needed here. Too many classes or too few classes might not reveal the basic shape of the set of data. In the vehicle selling price problem, for example, three classes would not give much insight into the pattern of the data (see Table 2-2).

TABLE 2-2: An Example of Too Few Classes

Vehicle Selling Price	Number of Vehicles
19 000 up to 32 900	53
32 900 up to 46 800	25
46 800 up to 60 700	2
Total	80

A useful recipe to determine the number of classes is the "2 to the k rule." This guide suggests you select the smallest number (k) for the number of classes such that 2^k (in words, 2 raised to the power of k) is greater than the number of data points (n).

In the Whitner Pontiac example, there were 80 vehicles sold. So $n = 80$. If we try $k = 6$, which means we would use 6 classes, then $2^6 = 64$, somewhat less than 80. Hence, 6 classes are not enough. If we let $k = 7$, then $2^7 = 128$, which is greater than 80. So the recommended number of classes is 7.

2. **Determine the class width.** Generally, the class width should be the same for all classes. At the end of this section, we shall briefly discuss some situations where unequal class widths may be necessary. All classes taken together must cover at least the distance from the lowest value in the raw data up to the highest value.

STATISTICS IN ACTION

Forestry and the Canadian Economy

Why is Canadian softwood an important commodity? To find the answer, let us look at some statistics from the Statistics Canada Web site (www.statcan.ca).

- Logging and forestry employed 68 000 workers, second only to mining in primary industries in the year 2000
- Canada exported $41 380.8 millions of forestry products on balance of payment basis in the year 2000
- Quebec occupies the most forestland (839 000 km^2)
- PEI covers the least forestland (3000 km^2)
- Canada has 75 800 km^2 of forestland

All the numeric data above are statistics, and allow us to see why logging and forestry is important to the Canadian economy.

Expressing these words in a formula:

$$Class\ width > \frac{H - L}{k}$$

where H is the highest observed value, L is the lowest observed value, and k is the number of classes.

In the Whitner Pontiac case, the lowest value is $19 320 and the highest value is $50 719. If we wish to use 7 classes, the class width should be greater than ($50 719 – $19 320)/7 = $4485 571. In practice, this class width is usually rounded up to some convenient number, such as a multiple of 10 or 100. We round this value up to $4490.

3. **Set up the individual class limits.** We should state class limits very clearly so that each observation falls into only one class. For example, classes such as $19 000–$20 000 and $20 000–$21 000 should be avoided because it is not clear whether $20 000 is in the first or second class. In this text, we will use the format $19 000 up to $20 000 and $20 000 up to $21 000 and so on. With this format it is clear that $19 999 goes into the first class and $20 000 in the second.

   Because we round the class width up to get a convenient class width, we cover a larger than necessary range. For example, seven classes of width $4490 in the Whitner Pontiac case result in a range of ($4490)(7) = $31 430.

   The actual range is $31 399, found by (H – L = 50 719 – 19 320). Comparing this value to $31 430, we have an excess of $31. It is natural to put approximately equal amounts of the excess in each of the two tails. As we have said before, we should also select convenient multiples of 10 for the class limits. We shall use $19 310 as the lower limit of the first class. The upper limit of the first class is then 23 800, found by (19 310 + 4 490 = 23 800). Hence, our first class is from $19 310 upto $23 800. We can determine the other classes (in dollars) similarly, (from $23 800 up to $28 290), (from $28 290 up to $32 780), (from $32 780 up to $37 270), (from $37 270 up to $41 760), (from $41 760 up to $46 250), and (from $46 250 up to $50 740).

4. **Tally the selling prices into the classes.** To begin, the selling price of the first vehicle in Table 2-1 is $31 373. It is tallied in the $28 290 up to $32 780 class. The second selling price in the first column is $39 552. It is tallied in the $37 270 up to $41 760 class. The other selling prices are tallied in a similar manner. When all the selling prices are tallied, we get Table 2-3(a).

TABLE 2-3: Construction of a Frequency Distribution of Whitner Pontiac Data

(a) Tally Count

Classes ($)	Tally
19 310 up to 23 800	𝍸 𝍸
23 800 up to 28 290	𝍸 𝍸 𝍸 𝍸 I
28 290 up to 32 780	𝍸 𝍸 𝍸 𝍸
32 780 up to 37 270	𝍸 𝍸 𝍸
37 270 up to 41 760	𝍸 III
41 760 up to 46 250	IIII
46 250 up to 50 740	II

(b) Frequency Distribution

Selling Prices ($ thousands)	Frequency
19.310 up to 23.800	10
23.800 up to 28.290	21
28.290 up to 32.780	20
32.780 up to 37.270	15
37.270 up to 41.760	8
41.760 up to 46.250	4
46.250 up to 50.740	2
Total	80

5. **Count the number of items in each class.** The number of observations in each class is called the *class frequency*. In the \$19 310 up to \$23 800 class, there are 10 observations; in the \$23 800 up to \$28 290 class there are 21 observations. Therefore, the class frequency in the first class is 10 and the class frequency in the second class is 21. The sum of frequencies of all the classes equals the total number of observations in the entire data set, which is 80.

   Often it is useful to express the data in thousands, or some convenient units, rather than the actual data. Table 2-3(b) reports the frequency distribution for Whitner Pontiac's vehicle selling prices where prices are given in thousands of dollars rather than dollars.

Now that we have organized the data into a frequency distribution, we can summarize the patterns in the selling prices of the vehicles for Rob Whitner. These observations are listed below:

1. The selling prices ranged from about \$19 310 to \$50 740.
2. The largest concentration of selling prices is in the \$23 800 up to \$28 290 class.
3. The selling prices are concentrated between \$23 800 and \$37 270. A total of 56 (70 percent) of the vehicles are sold within this range.
4. Two of the vehicles sold for \$46 250 or more, and 10 sold for less than \$23 800.

By presenting this information to Rob Whitner, we give him a clearer picture of the distribution of the selling prices for the last month.

We admit that arranging the information on the selling prices into a frequency distribution does result in the loss of some detailed information. That is, by organizing the data into a frequency distribution, we cannot pinpoint the exact selling price (such as \$23 820, or \$32 800), and we cannot tell that the actual selling price of the least expensive vehicle was \$19 320 and of the most expensive vehicle was \$50 719. However, the lower limit of the first class and the upper limit of the largest class convey essentially the same meaning. Whitner will make the same judgment if he knows the lowest price is about \$19 310 that he will make if he knows the exact selling price is \$19 320. The advantage of condensing the data into a more understandable form more than offsets this disadvantage.

SELF-REVIEW 2-1

The commissions earned for the first quarter of last year by the 11 members of the sales staff at Master Chemical Company are \$1650, \$1475, \$1510, \$1670, \$1595, \$1760, \$1540, \$1495, \$1590, \$1625, and \$1510.

(a) What are the values such as \$1650 and \$1475 called?
(b) Using \$1400 up to \$1500 as the first class, \$1500 up to \$1600 as the second class, and so forth, organize data on commissions earned into a frequency distribution.
(c) What are the numbers in the right column of your frequency distribution called?
(d) Describe the distribution of commissions earned based on the frequency distribution. What is the largest amount of commission earned? What is the smallest?

CLASS INTERVALS AND CLASS MIDPOINTS

We will use two other terms frequently: **class midpoint** and **class interval**. The midpoint, also called the **class mark**, is halfway between the lower and upper class limits. It can be computed by adding the lower class limit to the upper class limit and dividing by 2. Referring to Table 2-3 for the first class, the lower class limit is \$19 310 and the upper limit is \$23 800. The class midpoint is \$21 555, found by (\$19 310 + \$23 800)/2. The midpoint of \$21 555 best represents, or is typical of, the selling prices of the vehicles in that class.

To determine the class interval, subtract the lower limit of the class from its upper limit. The class interval of the vehicle selling price data is \$4490, which we find by subtracting the lower limit of the first class, \$19 310, from its upper limit; that is, \$23 800 – \$19 310 = \$4490. You can also determine the class interval by finding the distance between consecutive midpoints. The midpoint of the first class is \$21 555 and the midpoint of the second class is \$26 045. The difference is \$4490.

A SOFTWARE EXAMPLE: FREQUENCY DISTRIBUTION USING MEGASTAT

Chart 2-2 shows the frequency distribution of the Whitner Pontiac data produced by MegaStat. The form of the output is somewhat different than the frequency distribution in Table 2-3(b), but overall conclusions are the same.

SELF-REVIEW 2-2

The following table includes the grades of students who took Math 1021 during Fall 2002.

40	55	50	55	28	60	25	55	60	65	70	64
62	70	50	65	55	48	69	25	64	58	55	71

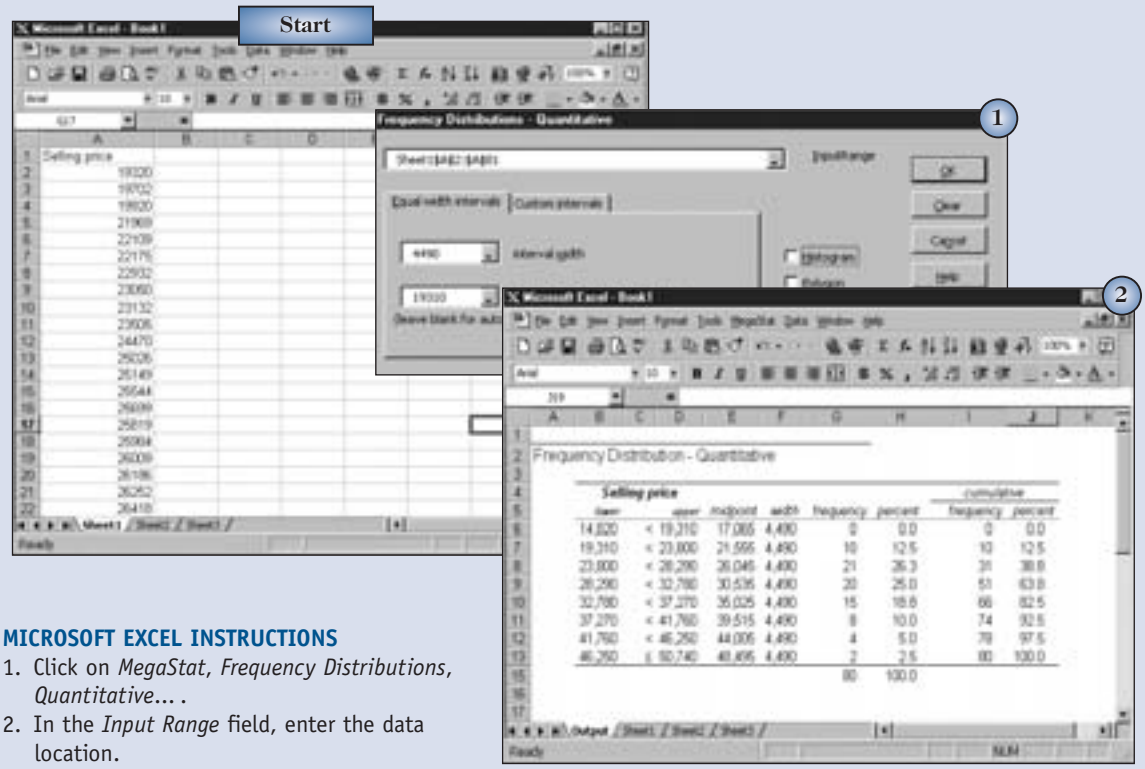
(a) How many classes would you use?
(b) How wide would you make the classes?
(c) Create a frequency distribution table.

RELATIVE FREQUENCY DISTRIBUTION

It may be desirable to convert class frequencies to relative class frequencies to show the fraction of the total number of observations in each class. In our vehicle sales example, we may want to know what percentage of the vehicle prices are in the \$28 290 up to \$32 780 class.

To convert a frequency distribution to a relative frequency distribution, each of the class frequencies is divided by the total number of observations. Using the distribution of vehicle sales again (Table 2-3(b), where the selling prices are reported in thousands of dollars), the relative frequency for the \$19 310 up to \$23 800 class is 0.125, found by dividing 10 by 80. That is, the price of 12.5 percent of the vehicles sold at Whitner Pontiac is between \$19 310 and \$23 800. The relative frequencies for the remaining classes are shown in Table 2-4.

EXCEL CHART 2-2: Frequency Distribution of Data in Table 2-1



MICROSOFT EXCEL INSTRUCTIONS

1. Click on *MegaStat*, *Frequency Distributions*, *Quantitative...* .
2. In the *Input Range* field, enter the data location.
3. Select *Equal Width Interval*, and input *interval size* (= 4490 in our example).
4. Input *value of lower boundary* of the first interval (= 19 310 in our example).
5. Deselect *Histogram*, and click OK.

TABLE 2-4: Relative Frequency Distribution of Selling Prices at Whitner Pontiac Last Month

Selling Price ($ thousands)	Frequency	Relative Frequency	Found by
19.310 up to 23.800	10	0.1250	10/80
23.800 up to 28.290	21	0.2625	21/80
28.290 up to 32.780	20	0.2500	20/80
32.780 up to 37.270	15	0.1875	15/80
37.270 up to 41.760	8	0.1000	8/80
41.760 up to 46.250	4	0.0500	4/80
46.250 up to 50.740	2	0.0250	2/80
Total	80	1.00100	

SELF-REVIEW 2-3

Refer to Table 2-4, which shows the relative frequency distribution for the vehicles sold last month at Whitner Pontiac.

(a) How many vehicles sold for $23 800 up to $28 290?
(b) What percentage of the vehicles sold for a price from $23 800 up to $28 290?
(c) What percentage of the vehicles sold for $37 270 or more?

EXERCISES 2-1 TO 2-8

2-1. A set of data consists of 38 observations. How many classes would you recommend for the frequency distribution?

2-2. A set of data consists of 45 observations. The lowest value is $0 and the highest value is $29. What size would you recommend for the class interval?

2-3. A set of data consists of 230 observations. The lowest value is $235 and the highest value is $567. What class interval would you recommend?

2-4. A set of data contains 53 observations. The lowest value is 42 and the highest is 129. The data are to be organized into a frequency distribution.
(a) How many classes would you suggest?
(b) What would you suggest as the lower limit of the first class?

2-5. The Wachesaw Outpatient Centre, designed for same-day minor surgery, opened last month. Below are the numbers of patients served during the first 16 days.

27	27	23	24	25	28	35	33
34	24	30	30	24	33	23	23

(a) How many classes would you recommend?
(b) What class interval would you suggest?
(c) What lower limit would you recommend for the first class?

2-6. The Quick-Change Oil Company has a number of outlets in Hamilton, Ontario. The numbers of oil changes at the Oak Street outlet in the past 20 days are listed below. The data are to be organized into a frequency distribution.

65	98	55	62	79	59	51	90	72	56
70	62	66	80	94	79	63	73	71	85

(a) How many classes would you recommend?
(b) What class interval would you suggest?
(c) What lower limit would you recommend for the first class?
(d) Organize the number of oil changes into a frequency distribution.
(e) Comment on the shape of the frequency distribution. Also determine the relative frequency distribution.

2-7. The local manager of Food Queen is interested in the number of times a customer shops at her store during a two-week period. The responses of 51 customers were:

5	3	3	1	4	4	5	6	4	2	6	6	6	7	1
1	14	1	2	4	4	4	5	6	3	5	3	4	5	6
8	4	7	6	5	9	11	3	12	4	7	6	5	15	1
1	10	8	9	2	12									

(a) Starting with 0 as the lower limit of the first class and using a class interval of 3, organize the data into a frequency distribution.
(b) Describe the distribution. Where do the data tend to cluster?
(c) Convert the distribution to a relative frequency distribution.

2-8. Moore Travel, a nationwide travel agency, offers special rates on certain Caribbean cruises to senior citizens. The president of Moore Travel wants additional information on the ages of those people taking cruises. A random sample of 40 customers taking a cruise last year revealed these ages:

77	18	63	84	38	54	50	59	54	56	36	26	50	34
44	41	58	58	53	51	62	43	52	53	63	62	62	65
61	52	60	60	45	66	83	71	63	58	61	71		

(a) Organize the data into a frequency distribution, using 7 classes and 15 as the lower limit of the first class. What class interval did you select?
(b) Where do the data tend to cluster?
(c) Describe the distribution.
(d) Determine the relative frequency distribution.

FREQUENCY DISTRIBUTION WITH UNEQUAL CLASS INTERVALS

In constructing frequency distributions of quantitative data, generally, equal class widths are assigned to all classes. This is because unequal class intervals present problems in graphically portraying the distribution and in doing some of the computations, as we will see in later chapters. Unequal class intervals, however, may be necessary in certain situations to avoid a large number of empty, or almost empty, classes. Such is the case in Table 2-5. Canada Customs and Revenue Agency (CCRA) used unequal-sized class intervals to report the adjusted gross income on individual tax returns. Had the CCRA used an equal-sized interval of, say, $1000, more than 1000 classes would have been required to describe all the incomes. A frequency distribution with 1000 classes would be difficult to interpret. In this case, the distribution is easier to understand in spite of the unequal classes. Note also that the number of income tax returns or "frequencies" is reported in thousands in this particular table. This also makes the information easier to digest.

TABLE 2-5: Adjusted Gross Income for Individuals Filing Income Tax Returns

Adjusted Gross Income ($)	Number of Returns (in thousands)
Under 2 000	135
2 000 up to 3 000	3 399
3 000 up to 5 000	8 175
5 000 up to 10 000	19 740
10 000 up to 15 000	15 539
15 000 up to 25 000	14 944
25 000 up to 50 000	4 451
50 000 up to 100 000	699
100 000 up to 500 000	162
500 000 up to 1 000 000	3
1 000 000 and over	1

2.2 STEM-AND-LEAF DISPLAYS

In Section 2.1, we showed how to organize quantitative data into a frequency distribution so we could summarize the raw data into a meaningful form. The major advantage of organizing the data into a frequency distribution is that we get a quick visual picture of the shape of the distribution without doing any further calculation. That is, we can see where the data are concentrated and also determine whether there are any extremely large or small values. However, it has two disadvantages: (1) we lose the exact identity of each value, and (2) we are not sure how the values within each class are distributed. To explain, consider the following frequency distribution of the number of 30-second radio advertising spots purchased by the 45 members of the Toronto Automobile Dealers' Association in 2001. We observe that 7 of the 45 dealers purchased at least 90 but less than 100 spots. However, is the number of spots purchased within this class clustered near 90, spread evenly throughout the class, or clustered near 99? We cannot tell.

Number of Spots Purchased	Frequency
80 up to 90	2
90 up to 100	7
100 up to 110	6
110 up to 120	9
120 up to 130	8
130 up to 140	7
140 up to 150	3
150 up to 160	3
Total	45

For a mid-sized data set, we can eliminate these shortcomings by using an alternative graphic display called the **stem-and-leaf display**. To illustrate the construction of a stem-and-leaf display using the advertising spots data, suppose the seven observations in the 90 up to 100 class are 96, 94, 93, 94, 95, 96, and 97.

Let us sort these values to get: 93, 94, 94, 95, 96, 96, 97. The **stem** value is the leading digit or digits, in this case 9. The **leaves** are the trailing digits. The stem is placed to the left of a vertical line and the leaf values to the right. The values in the 90 up to 100 class would appear in the stem-and-leaf display as follows:

```
9    |    3 4 4 5 6 6 7
```

With the stem-and-leaf display, we can quickly observe that there were two dealers who purchased 94 spots and that the number of spots purchased ranged from 93 to 97. A stem-and-leaf display is similar to a frequency distribution with more information (i.e., data values instead of tallies).

> **Stem-and-leaf display** A statistical technique to present a set of data. Each numerical value is divided into two parts. The leading digit(s) become(s) the stem and the trailing digit(s) become(s) the leaf. The stems are located along the vertical axis and the leaf values are stacked against one another along the horizontal axis.

The following example will explain the details of developing a stem-and-leaf display.

Example 2-2

Table 2-6 lists the number of 30-second radio advertising spots purchased by each of the 45 members of the Toronto Automobile Dealers' Association last year. Organize the data into a stem-and-leaf display. Around what values do the number of advertising spots tend to cluster? What is the smallest number of spots purchased by a dealer and the largest number purchased?

Solution

From the data in Table 2-6 we note that the smallest number of spots purchased is 88. So we will make the first stem value 8. The largest number is 156, so we will have the stem values begin at 8 and continue to 15. The first number in Table 2-6 is 96, which will have a stem value of 9 and leaf value of 6. Moving across the top row, the second value is 93 and the third is 88. After the first three data values are considered, the display is shown opposite.

Stem	Leaf
8	8
9	6 3
10	
11	
12	
13	
14	
15	

TABLE 2-6: Number of Advertising Spots Purchased during 2001 by Members of the Toronto Automobile Dealers' Association

96	93	88	117	127	95	113	96	108	94
148	156	139	142	94	107	125	155	155	103
112	127	117	120	112	135	132	111	125	104
106	139	134	119	97	89	118	136	125	143
120	103	113	124	138					

Organizing all the data, the stem-and-leaf display would appear as shown in Chart 2-3(a).

The usual procedure is to sort the leaf values from smallest to largest. The last line, the row referring to the values in the 150s, would appear as:

15 | 5 5 6

The final table would appear as shown in Chart 2-3(b), where we have sorted all of the leaf values.

CHART 2-3: Stem-and-Leaf Display

a.

Stem	Leaf
8	8 9
9	6 3 5 6 4 4 7
10	8 7 3 4 6 3
11	7 3 2 7 2 1 9 8 3
12	7 5 7 0 5 5 0 4
13	9 5 2 9 4 6 8
14	8 2 3
15	6 5 5

b.

Stem	Leaf
8	8 9
9	3 4 4 5 6 6 7
10	3 3 4 6 7 8
11	1 2 2 3 3 7 7 8 9
12	0 0 4 5 5 5 7 7
13	2 4 5 6 8 9 9
14	2 3 8
15	5 5 6

You can draw several conclusions from the stem-and-leaf display. First, the lowest number of spots purchased is 88 and the highest is 156. Two dealers purchased less than 90 spots, and three purchased 150 or more. You can observe, for example, that the three dealers who purchased more than 150 spots actually purchased 155, 155, and 156 spots. The concentration of the number of spots is between 110 and 139. There were nine dealers who purchased between 110 and 119 spots and eight who purchased between 120 and 129 spots. We can also tell that within the 120 up to 130 group, the actual number of spots purchased was spread evenly throughout. That is, two dealers purchased 120 spots, one dealer purchased 124 spots, three dealers purchased 125 spots, and two dealers purchased 127 spots.

We can also generate this information using Minitab. We have named the variable *Spots*. The Minitab output is given on the next page.

The Minitab stem-and-leaf display provides some additional information regarding cumulative totals. In Chart 2-4, the column to the left of the stem values has numbers such as 2, 9, 15, and so on. The number 9 indicates that there are 9 observations of value less than the upper limit of the current class, which is 100. The number 15 indicates that there are 15 observations less than 110. About halfway down the column the number 9 appears in parentheses. The parentheses indicate that the middle value appears in that row; hence, we call this row the *median row*. In this case, we describe the middle value as the value that divides the total number of observations into two equal parts. There are a total of 45 observations, so the middle value, if the data were arranged from smallest to largest, would be the 23rd observation. After the median row, the values begin to decline. These values represent the "more than" cumulative totals. There are 21 observations of value greater than or equal to the lower limit of this class, which is 120; 13 of 130 or more, and so on. Stem-and-leaf display is useful only for a mid-sized data set. When we use a stem-and-leaf display for a large data set, we produce a large number of stems and/or leaves and are not able to see the characteristics of a large data set.

In the stem-and-leaf display for Example 2-2, the leading digits (stems) take the values from 8 to 15 and thus have 8 stems (8, 9, 10, 11, 12, 13, 14, 15) in units of 10. However, in some data sets, stems assume only two or three values. Generating a stem-and-leaf display in these situations is not as easy as in Example 2-2. Let us look at the sample of marks of 20 students in Math 2010:

50	52	54	53	65	60	45	43	57	62
56	58	51	61	46	44	69	55	64	59

The leading digits (units of 10) in this example assume only three values: 4, 5, and 6. Following the above procedure for drawing a stem-and-leaf display, the stem-and-leaf display of the above data set looks like the one given below.

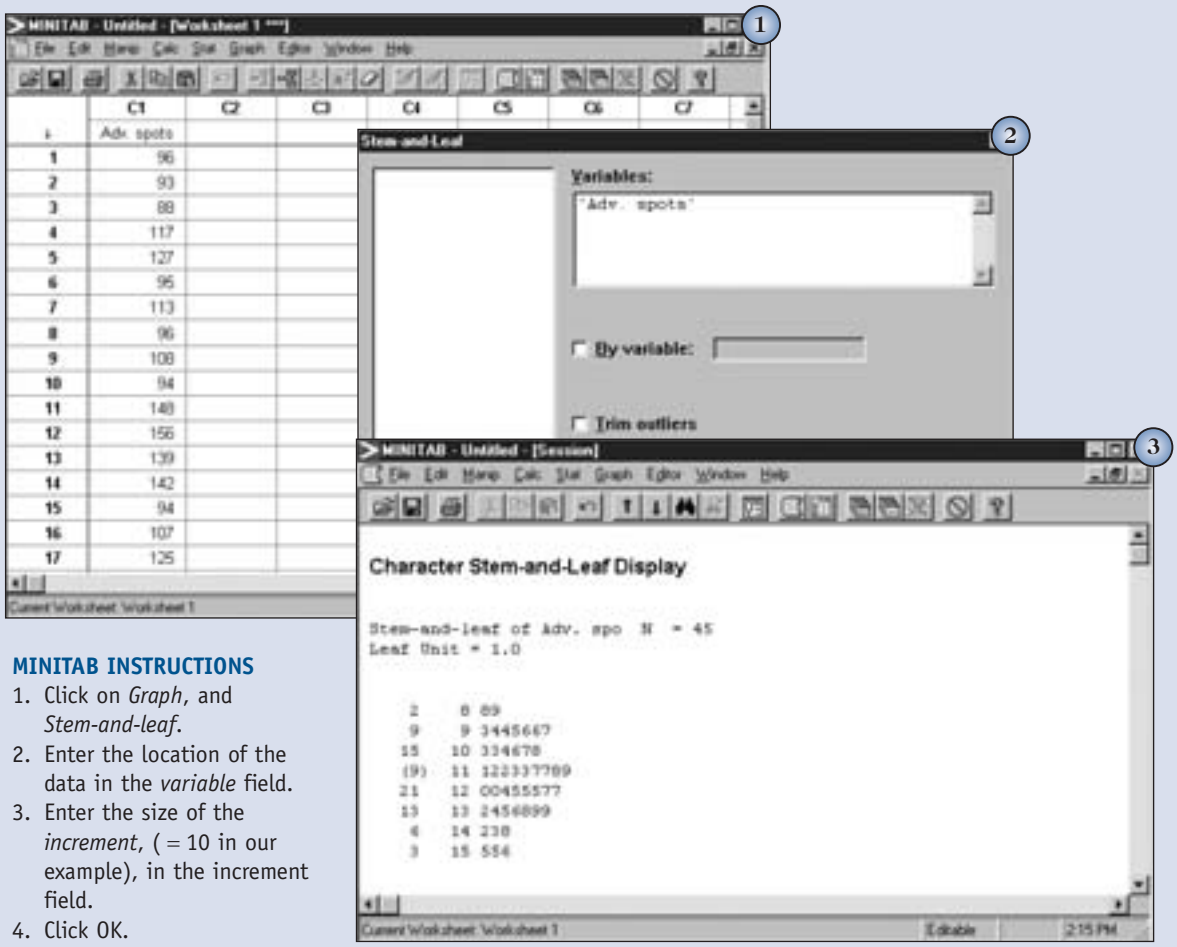
Stem	Leaf
4	3 4 5 6
5	0 1 2 3 4 5 6 7 8 9
6	0 1 2 4 5 9

As we can see, this stem-and-leaf display has only three stems and does not display the characteristics of the data set as well as if there were more stems. We can improve the stem-and-leaf display by *splitting* each stem. For example, stem 4 can be split as

4	|	3 4
4	|	5 6

The first stem 4 contains leaves less than 5 and the second stem 4 contains leaves 5 and above.

MINITAB CHART 2-4: Stem-and-Leaf Display of Data in Table 2-6



MINITAB INSTRUCTIONS

1. Click on *Graph*, and *Stem-and-leaf*.
2. Enter the location of the data in the *variable* field.
3. Enter the size of the *increment*, (= 10 in our example), in the increment field.
4. Click OK.

The revised stem-and-leaf display is given below.

Stem	Leaf
4	3 4
4	5 6
5	0 1 2 3 4
5	5 6 7 8 9
6	0 1 2 4
6	5 9

Other data sets may require even more splitting. The question of how much splitting is necessary can be answered by the rule suggested by Tukey et al.[1] For a sample size ≤ 100, the number of stems should be the integer part of $2\sqrt{n}$, where n is the sample size; for $n \geq 100$, the number of stems should be the integer part of $10 \log_{10} n$. In our example of 20 students' marks, the number suggested by the rule is 8. However, we have 6 stems in our example, which is close to 8. Remember, the rule provides a guideline for selecting the number of stems.

SELF-REVIEW 2-4

The price–earnings ratios for 21 stocks in the retail trade category are:

8.3	9.6	9.5	9.1	8.8	11.2	7.7	10.1	9.9	10.8	10.2
8.0	8.4	8.1	11.6	9.6	8.8	8.0	10.4	9.8	9.2	

Organize this information into a stem-and-leaf display.

(a) How many values are less than 9.0?
(b) List the values in the 10.0 up to 11.0 category.
(c) What are the largest and the smallest price–earnings ratios?

EXERCISES 2-9 TO 2-14

2-9. The first row of a stem-and-leaf display appears as follows: 62 | 1 3 3 7 9. Assume whole number values.
(a) What is the range of the values in this row?
(b) How many data values are in this row?
(c) List the actual values in this row.

2-10. The third row of a stem-and-leaf display appears as follows: 21 | 0 1 3 5 7 9. Assume whole number values.
(a) What is the range of the values in this row?
(b) How many data values are in this row?
(c) List the actual values in this row.

2-11. The following stem-and-leaf display shows the number of units produced per day in a factory.

1	3	8
1	4	
2	5	6
9	6	0 1 3 3 5 5 9
(7)	7	0 2 3 6 7 7 8
9	8	5 9
7	9	0 0 1 5 6
2	10	3 6

(a) How many days were studied?
(b) How many observations are in the first class?
(c) What are the largest and the smallest values in the data set?
(d) List the actual values in the fourth row.
(e) List the actual values in the second row.
(f) How many values are less than 70?
(g) How many values are 80 or more?
(h) How many values are between 60 and 89?

2-12. The following stem-and-leaf display reports the number of movies rented per day at Video Connection.

3	12	6 8 9
6	13	1 2 3
10	14	6 8 8 9
13	15	5 8 9
15	16	3 5
20	17	2 4 5 6 8
23	18	2 6 8
(5)	19	1 3 4 5 6
22	20	0 3 4 6 7 9
16	21	2 2 3 9
12	22	7 8 9
9	23	0 0 1 7 9
4	24	8
3	25	1 3
1	26	
1	27	0

(a) How many days were studied?
(b) How many observations are in the last class?
(c) What are the largest and the smallest values in the entire set of data?
(d) List the actual values in the fourth row.
(e) List the actual values in the next to the last row.
(f) On how many days were fewer than 160 movies rented?
(g) On how many days were 220 or more movies rented?
(h) On how many days were between 170 and 210 movies rented?

2-13. A survey of the number of calls received by a sample of Southern Phone Company subscribers last week revealed the following information. Develop a stem-and-leaf display. How many calls did a typical subscriber receive? What were the largest and the smallest number of calls received?

52	43	30	38	30	42	12	46	39	37	34	46	32
18	41	5										

2-14. Aloha Banking Co. is studying the number of times a particular automated teller machine (ATM) is used each day. The following is the number of times it was used during each of the last 30 days. Develop a stem-and-leaf display. Summarize the data on the number of times the machine was used: How many times was the ATM used on a typical day? What were the largest and the smallest number of times the ATM was used? Around what values did the number of times the ATM was used, tend to cluster?

83	64	84	76	84	54	75	59	70	61	63	80	84
73	68	52	65	90	52	77	95	36	78	61	59	84
95	47	87	60									

2.3 GRAPHIC PRESENTATION OF A FREQUENCY DISTRIBUTION

Sales managers, stock analysts, hospital administrators, and other busy executives often need a quick picture of the trends in sales, stock prices, or hospital costs. These trends can often be depicted by the use of charts and graphs. The charts that depict a frequency distribution graphically are the histogram, the stem-and-leaf display, the frequency polygon, and the cumulative frequency polygon.

HISTOGRAM

One of the most common graphical methods of displaying frequency distribution of a quantitative data is a **histogram**.

> **Histogram** A graph in which classes are marked on the horizontal axis and class frequencies on the vertical axis. The class frequencies are represented by the heights of the rectangles, and the rectangles are drawn adjacent to each other without any space between them.

Thus, a histogram describes a frequency distribution using a series of adjacent rectangles. Since the height of each rectangle equals the frequency of the corresponding class, and all the class widths are equal, the area of each rectangle is proportional to the frequency of the corresponding class.

Example 2-3

Refer to the data in Table 2-7 on life expectancy of males at birth in 40 countries. Construct a frequency distribution and a histogram. What conclusions can you reach based on the information presented in the histogram?

TABLE 2-7: Life Expectancy of Males at Birth

Country	Life Expectancy (years)	Country	Life Expectancy (years)	Country	Life Expectancy (years)
Afghanistan	45	Bhutan	59.5	Egypt	64.7
Albania	69.9	Botswana	46.2	France	74.2
Angola	44.9	Brazil	63.1	Germany	73.9
Argentina	69.6	Bulgaria	67.6	Hungary	66.8
Armenia	67.2	Cambodia	51.5	India	62.3
Australia	75.5	Canada	76.1	Iran	68.5
Austria	73.7	Chad	45.7	Japan	76.8
Bahamas	70.5	Chile	72.3	Kenya	51.1
Bahrain	71.1	China	67.9	Nepal	57.6
Bangladesh	58.1	Congo	48.3	UK	74.5
Barbados	73.7	Cuba	74.2	USA	73.4
Belarus	62.2	Czech Republic	70.3	Venezuela	70
Belgium	73.8	Denmark	73	Zambia	39.5
Bermuda	71.7				

Source: Life Expectancy at Birth (Males), United Nations Statistics Divisions, 1996–2000

Solution

The data in Table 2-7 is a quantitative data. Therefore, the first step is to construct a frequency distribution using the method discussed in Section 2.1 This is given in Table 2-8. (In Table 2-8, we also give relative frequencies. These will be discussed later.)

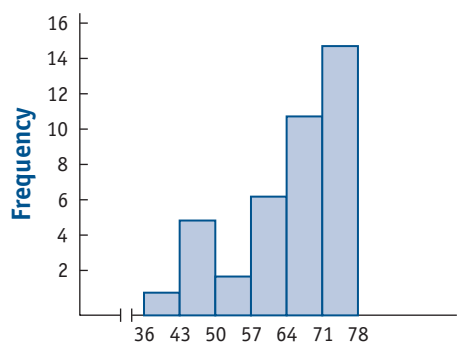
TABLE 2-8: Frequency and Relative Frequency Distribution of Life Expectancy Data

Life Expectancy	Frequency	Relative Frequency		Found by
36 up to 43	1	0.025	→	1/40
43 up to 50	5	0.125	→	5/40
50 up to 57	2	0.050	→	2/40
57 up to 64	6	0.150	→	6/40
64 up to 71	11	0.275	→	11/40
71 up to 78	15	0.375	→	15/40
Total	40	1.000		

To construct a histogram, class frequencies are scaled along the vertical axis (y-axis) and either the class limits or the class midpoints are scaled along the horizontal axis (x-axis).

From the frequency distribution, the frequency of the class 36 up to 43 is 1. Therefore, the height of the column for this class is 1. Make a rectangle whose width spreads from 36 to 43 with the height of one unit. Repeat the process for the remaining classes. The completed histogram should resemble the graph presented in Chart 2-5. The double slant on the x-axis indicates that the class limits did not start at zero. That is, the division between 0 and 36 is not linear. In other words, the distance between 0 and 36 is not the same as the distance between 36 and 43, between 43 and 50, and so on.

CHART 2-5: Histogram of Life Expectancy for Males at Birth



Histogram of Life Expectancy at Birth (Males)

From Chart 2-5, we conclude that:

- the lowest life expectancy is about 36 years and the highest is about 78 years.
- the class with the highest frequency (15) is 71 up to 78. That is, 15 countries have a life expectancy from 71 up to 78 years.
- the class with the lowest frequency (1) is 36 up to 43 years. That is, there is only one country with a life expectancy from 36 up to 43.
- the histogram is j-shaped. There is a tail on the left side of the class with the highest frequency (mode), and no tail on its right side.

COMMON DISTRIBUTION SHAPES

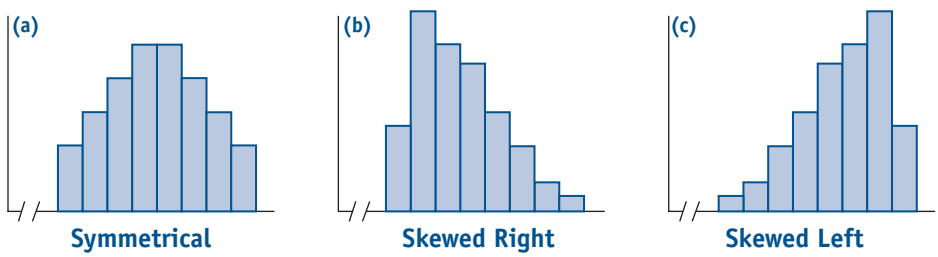
According to the shapes of histograms, distributions can be classified into (i) symmetrical and (ii) skewed.

A symmetrical distribution is one in which, if we divide its histogram into two pieces by drawing a vertical line through its centre, the two halves formed are mirror images of each other. This is displayed in Chart 2-6(a).

A distribution that is not symmetrical is said to be skewed.

For a skewed distribution, it is quite common to have one tail of the distribution longer than the other. If the longer tail is stretched to the right, the distribution is said to be *skewed to the right*. If the longer tail is stretched to the left, it is said to be *skewed to the left*. These are displayed in Charts 2-6(b) and (c) below.

CHART 2-6: Common Distribution Shapes



For a symmetrical distribution, the centre, or the typical value, of the distribution is well defined. For a skewed distribution, however, it is not that easy to define the centre. We shall discuss this in detail in the next chapter.

Another commonly used classification of distributions is according to its number of peaks. When the histogram has a single peak, the distribution is called *unimodal.* A *bimodal distribution* is one in which the histogram has two peaks not necessarily equal in height.

RELATIVE FREQUENCY HISTOGRAM

A relative frequency histogram is a graph in which classes are marked on the horizontal axis and the relative frequencies (frequency of a class/total frequency) on the vertical axis. Let us refer again to the data in Table 2-7 on life expectancy of males at birth in 40 countries. In Table 2-8 we also give a relative frequency distribution corresponding to this data. For example, the relative frequency of the class 43 up to 50 is 0.125 (5/40). We follow the procedure used in drawing a histogram to draw a relative frequency histogram. Chart 2-7 shows the relative frequency histogram of the life expectancy data.

A relative frequency histogram has the following important properties:

- The shape of a relative frequency histogram of a data set is identical to the shape of its histogram. (Verify this for the life expectancy data.)

- It is useful in comparing shapes of two or more data sets with different total frequencies. (Note that when total frequencies of two data sets are different, histograms of these data sets cannot be compared. For example, total frequency of one data set may be 1000, while that of the other data set may be 100. But relative frequencies of any data set add up to 1.0.)
- The area of the rectangle corresponding to a class interval equals (relative frequency of the class) × (class width). For example, the relative frequency of class 43 up to 50 is 0.125 (12.5 percent of the countries listed in Table 2-7 have life expectancy in this class). The area of the corresponding rectangle is (0.125)(50 – 43) = 0.875.

The total area under the entire relative frequency histogram is therefore (class width) × (sum of relative frequencies of all the classes) = class width. (This is because the sum of the relative frequency of all classes equals 1.)

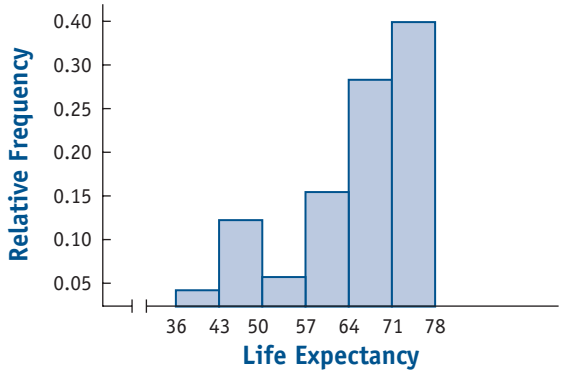
If we scale the height of each rectangle by 1/(class width), then the total area under each rectangle of the scaled relative frequency histogram will be equal to its relative frequency, and the total area under the entire scaled relative frequency histogram will be equal to 1.

A histogram provides an easily interpreted visual representation of the frequency distribution of a given raw data. The shape of the histogram is the same whether we use the actual frequency distribution or the relative frequency distribution. We shall see in later chapters the importance of shapes in determining the appropriate method of statistical analysis.

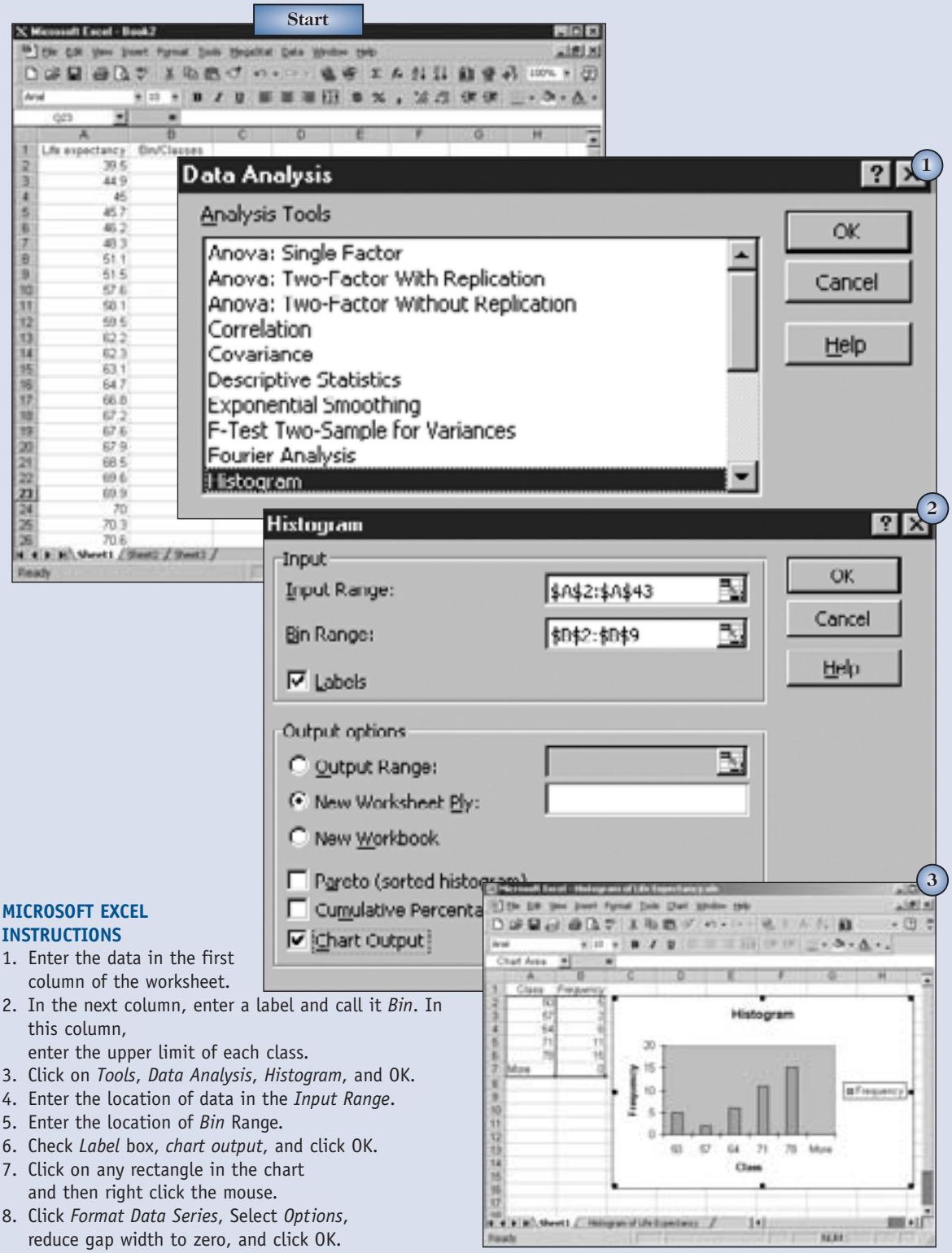
HISTOGRAM USING EXCEL AND MINITAB

We can plot a histogram using MegaStat by following the same instructions as those for the construction of a frequency distribution, except that in this case, we do *not* deselect "histogram." We give below instructions for plotting a histogram using Excel (without Megastat) and Minitab.

CHART 2-7: Relative Frequency Histogram of Life Expectancy at Birth (Males)



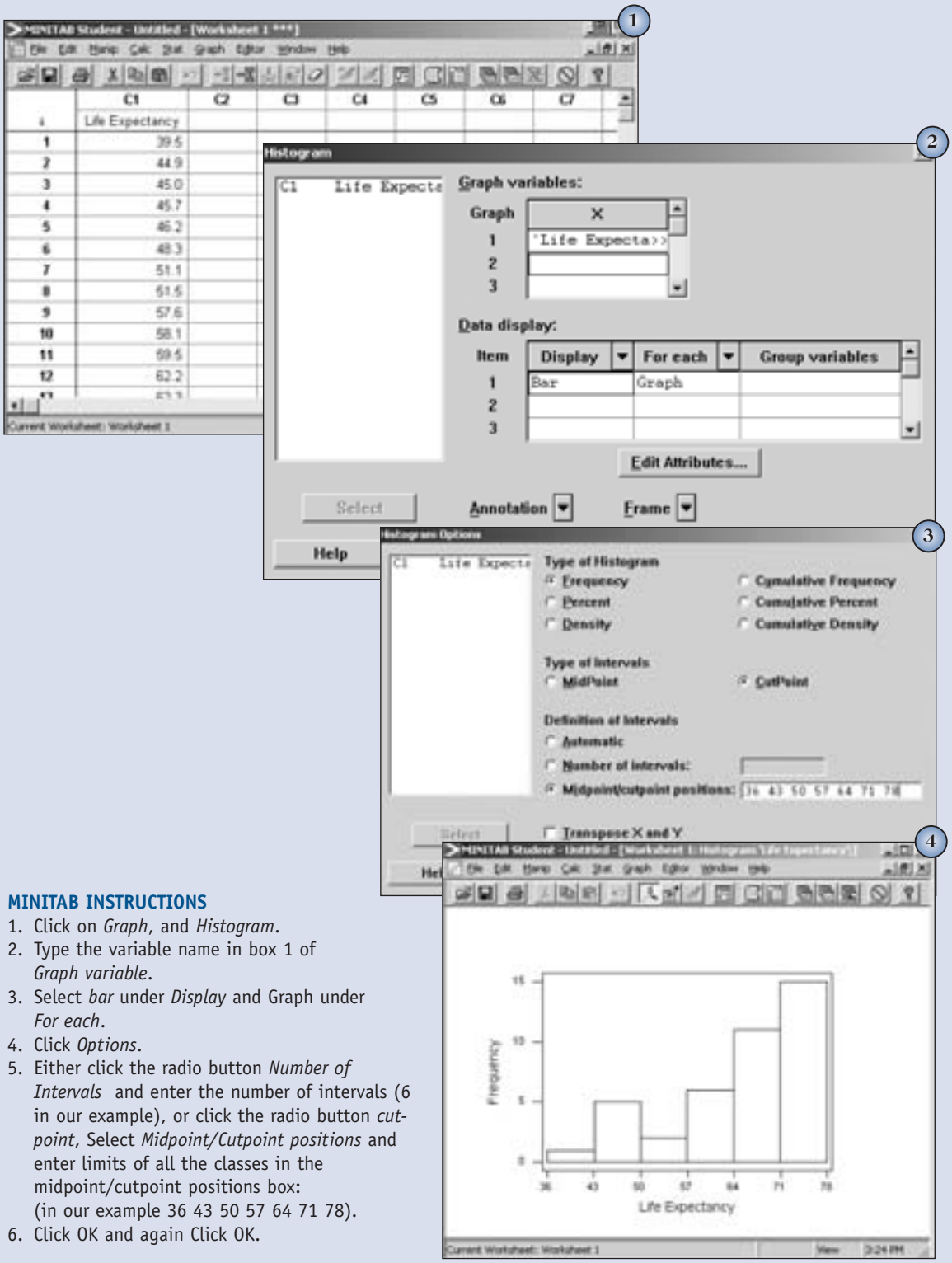
EXCEL CHART 2-8: Histogram of Life Expectancy



MICROSOFT EXCEL INSTRUCTIONS

1. Enter the data in the first column of the worksheet.
2. In the next column, enter a label and call it *Bin*. In this column, enter the upper limit of each class.
3. Click on *Tools*, *Data Analysis*, *Histogram*, and OK.
4. Enter the location of data in the *Input Range*.
5. Enter the location of *Bin* Range.
6. Check *Label* box, *chart output*, and click OK.
7. Click on any rectangle in the chart and then right click the mouse.
8. Click *Format Data Series*, Select *Options*, reduce gap width to zero, and click OK.

MINITAB CHART 2-9: Histogram of Life Expectancy



MINITAB INSTRUCTIONS

1. Click on *Graph*, and *Histogram*.
2. Type the variable name in box 1 of *Graph variable*.
3. Select *bar* under *Display* and Graph under *For each*.
4. Click *Options*.
5. Either click the radio button *Number of Intervals* and enter the number of intervals (6 in our example), or click the radio button *cut-point*, Select *Midpoint/Cutpoint positions* and enter limits of all the classes in the midpoint/cutpoint positions box: (in our example 36 43 50 57 64 71 78).
6. Click OK and again Click OK.

FREQUENCY POLYGON

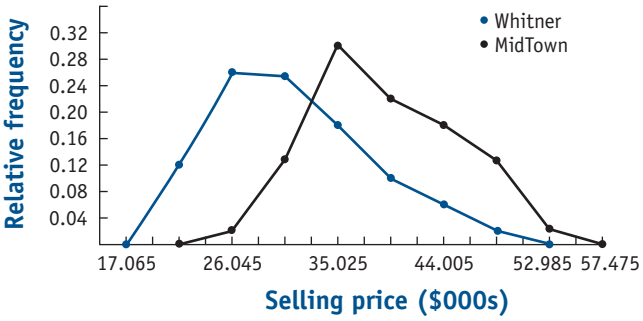
The construction of a frequency polygon is similar to the construction of a histogram. It consists of line segments connecting the points formed by the intersections of the class midpoints and the class frequencies. The construction of frequency polygon is illustrated in Chart 2-10. We use the vehicle prices for the cars sold last month at Whitner Pontiac. The midpoint of each class is scaled on the x-axis and the class frequencies on the y-axis. Recall that the class midpoint is the value at the centre of a class and represents the values in that class. The *class frequency* is the number of observations in a particular class. The vehicle selling prices at Whitner Pontiac are:

Selling Price ($ thousands)	Midpoint	Frequency
19.310 up to 23.800	21.555	10
23.800 up to 28.290	26.045	21
28.290 up to 32.780	30.34	20
32.780 up to 37.270	35.23	15
37.270 up to 41.760	39.12	8
41.760 up to 46.250	44.005	4
46.250 up to 50.740	48.495	2
Total		80

As noted earlier, the 19.310 up to 23.800 class is represented by the midpoint 21.555. To construct a frequency polygon, we move horizontally on the graph to the midpoint 21.555 and then vertically to 10, the class frequency, and place a dot. The x and y values of this point are called the *coordinates*. The coordinates of the next point are $x = 26.045$ and $y = 21$. The process is continued for all classes. Then the points are connected in order. That is, the point representing the lowest class is joined to the one representing the second class, and so on. Note in Chart 2-10 that to complete the frequency polygon, two additional points with x co-ordinates 17.065 and 52.985 and with 0 frequencies (that is, points on the x-axis), are added to *anchor* the polygon. These two values are derived by subtracting the class width of 4.49 from the lowest midpoint (21.555) and adding 4.49 to the highest midpoint (48.495) in the frequency distribution.

CHART 2-10: Frequency Polygon of the Selling Prices of 80 Vehicles at Whitner Pontiac

CHART 2-11: Distribution of Selling Prices at Whitner Pontiac and Midtown Cadillac



Both the histogram and the frequency polygon allow us to get a quick picture of the main characteristics of the data (highs, lows, points of concentration, etc.). Although the two representations are similar in purpose, the histogram has the advantage of depicting each class as a rectangle, with the height of the rectangular bar representing the number of frequencies in each class. The frequency polygon, in turn, has an advantage over the histogram. It allows us to directly compare two or more frequency distributions. Suppose Rob, the owner of Whitner Pontiac, wants to compare the sales last month at his dealership with those at Midtown Cadillac. To do this, two frequency polygons are constructed, one on top of the other, as shown in Chart 2-11. It is clear from the chart that the total sales volume at each dealership is more or less the same.

SELF-REVIEW 2-5

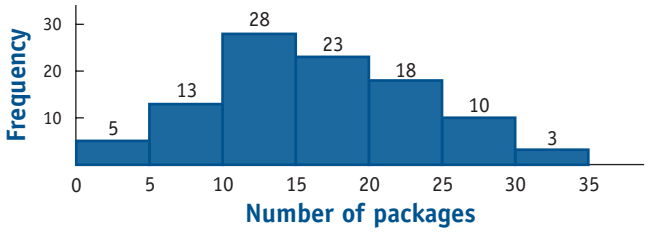
The annual imports of a selected group of electronic suppliers are shown in the following frequency distribution.

Imports ($ millions)	Number of Suppliers
2 up to 5	6
5 up to 8	13
8 up to 11	20
11 up to 14	10
14 up to 17	1

(a) Draw a histogram.
(b) Draw a frequency polygon.
(c) Summarize the important features of the distribution (such as low and high values, concentration, etc.).

EXERCISES 2-15 TO 2-18

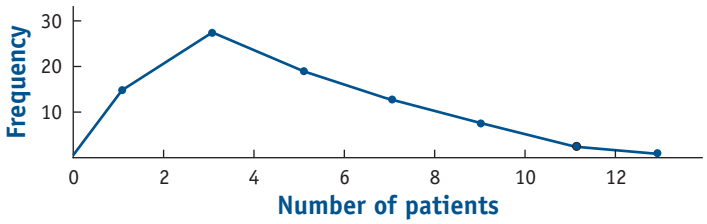
2-15. Molly's Candle Shop has several retail stores in Vancouver. Many of Molly's customers ask her to ship their purchases. The following chart shows the number of packages shipped per day for the last 100 days.



(a) What is this chart called?
(b) What is the total number of frequencies?
(c) What is the class interval?
(d) What is the class frequency for the 10 up to 15 class?

(e) What is the relative frequency for the 10 up to 15 class?
(f) What is the midpoint for the 10 up to 15 class?
(g) On how many days were there 25 or more packages shipped?

2-16. The following chart shows the number of patients admitted daily to Memorial Hospital through the emergency room.



(a) What is the midpoint of the 2 up to 4 class?
(b) On how many days were 2 up to 4 patients admitted?
(c) Approximately how many days were studied?
(d) What is the class interval?
(e) What is this chart called?

2-17. The following frequency distribution represents the number of days during a year that employees at J. Morgan Manufacturing Company were absent from work due to illness.

Number of Days Absent	Number of Employees
0 up to 3	5
3 up to 6	12
6 up to 9	23
9 up to 12	8
12 up to 15	2
Total	50

(a) Construct a relative frequency histogram.
(b) What proportion of the total area under the relative frequency histogram is contained above the interval 3 up to 12?

2-18. A large retailer is studying the lead time (elapsed time between when an order is placed and when it is filled) for a sample of recent orders. The lead times are reported in days.

Lead Time (days)	Frequency
0 up to 5	6
5 up to 10	7
10 up to 15	12
15 up to 20	8
20 up to 25	7
Total	40

(a) How many orders were studied?
(b) What is the midpoint of the first class?
(c) What are the coordinates of the point on the frequency polygon corresponding to the first class?

(d) Draw a histogram.
(e) Draw a frequency polygon.
(f) Interpret the lead times using the two charts.

CUMULATIVE FREQUENCY DISTRIBUTIONS

Let us consider again the distribution of the selling prices of vehicles at Whitner Pontiac. Suppose we were interested in the number of vehicles that sold for less than $28 290. These numbers can be approximated by developing a cumulative frequency distribution and portraying it graphically in a cumulative frequency polygon, which is also called an ogive.

Example 2-4

Refer to Table 2-4 on page 39. Construct a less than cumulative frequency polygon. Fifty percent of the vehicles were sold for less than what amount? Twenty-five of the vehicles were sold for less than what amount?

Solution

As the name implies, a cumulative frequency distribution and a cumulative frequency polygon require cumulative frequencies. The cumulative frequency of a class is the number of observations fewer than the upper limit of that class. For example, in Table 2-10, the frequency distribution of the vehicle selling prices at Whitner Pontiac is repeated from Table 2-4. The cumulative frequency of the class 23.800 up to 28.290 is 31. How did we get it? We added the number of vehicles sold for less than $23 800 (which equals 10) to the 21 vehicles sold in the next higher class. Thus the number of vehicles sold for *less than* $28 290 is 31. Similarly, the cumulative frequency of the next higher class is 10 + 21 + 20 = 51. The process is continued for all the classes.

To plot a cumulative frequency distribution, scale the upper limit of each class along the x-axis and the corresponding cumulative frequencies along the y-axis. We label the vertical axis on the left in units and the vertical axis on the right in percent. In the Whitner Pontiac example, the vertical axis on the left is labelled from 0 to 80 (vehicles sold) and on the right from 0 to 100 percent. The value of 50 percent corresponds to 40 vehicles sold.

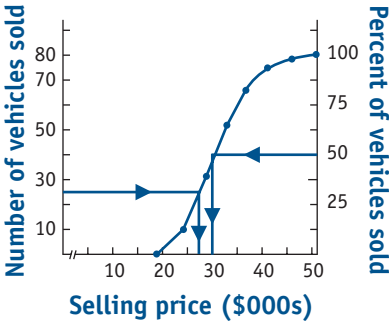
TABLE 2-10: Cumulative Frequency Distribution for Selling Prices at Whitner Pontiac Last Month

Selling Price ($ thousands)	Frequency	Cumulative Frequency		Found by
19.310 up to 23.800	10	10	«——	(10 + 0)
23.800 up to 28.290	21	31	«——	(10 + 21)
28.290 up to 32.780	20	51	«——	(10 + 21 + 20)
32.780 up to 37.270	15	66		
37.270 up to 41.760	8	74		
41.760 up to 46.250	4	78		
46.250 up to 50.740	2	80		
Total	80			

To begin the plotting, 10 vehicles sold for less than $23 800, so the first point in the plot is at $x = 23.80$ and $y = 10$. The coordinates of the next point are $x = 28.29$ and $y = 31$. The rest of the points are plotted and then the dots are connected to form Chart 2-12. Close the lower end of the graph by extending the line to the lower limit of the first class. To find the selling price below which half the cars sold, we draw a line from the 50-percent mark on the right-hand vertical axis over to the polygon, then drop down to the x-axis and read the selling price. The value of the x-axis is about

$29 000. To find the price below which 25 of the vehicles sold, we locate the value of 25 on the left-hand vertical axis. Next, we draw a horizontal line from the value of 25 to the polygon, and then drop down to the x-axis and read the price; it is about 26. So, we estimate that 25 of the vehicles sold for less than $26 000. We can also estimate the percentage of vehicles sold for less than $39 000. We begin by locating the value of 39 on the x-axis, then moving vertically to the polygon and then horizontally to the vertical axis on the right. The value is about 87.5 percent. We therefore conclude that 87.5 percent of the vehicles sold for less than $39 000.

CHART 2-12: Cumulative Frequency Distribution for Vehicle Selling Price



SELF-REVIEW 2-6

The following table provides information on the annual net profits of 34 small companies.

Annual Net Profits ($ thousands)	Number of Companies
65 up to 75	1
75 up to 85	6
85 up to 95	7
95 up to 105	12
105 up to 115	5
115 up to 125	3

(a) What is the table called?
(b) Develop a cumulative frequency distribution and draw a cumulative frequency polygon for the distribution.
(c) Based on the cumulative frequency polygon, find the number of companies with annual net profits of less than $105 000.

EXERCISES 2-19 TO 2-22

2-19. The following table lists the salary distribution of full-time instructors in a community college.

Salary ($)	Number of Instructors
28 000 up to 33 000	5
33 000 up to 38 000	6
38 800 up to 43 000	4
43 000 up to 48 000	3
48 000 up to 53 000	7

(a) Develop a cumulative frequency distribution.
(b) Develop a cumulative relative frequency distribution.
(c) How many instructors earn less than $33 000?
(d) Seventy-two percent of instructors earn less than what amount?

2-20. Active Gas Services mailed statements of payments due to 70 customers. The following amounts are due:

Amount ($)	Number of Customers
70 up to 80	5
80 up to 90	20
90 up to 100	10
100 up to 110	11
110 up to 120	14
120 up to 130	10

(a) Draw a cumulative frequency polygon.
(b) What number of customers owes less than $100?

2-21. Refer to the frequency distribution of the annual number of days the employees at the J. Morgan Manufacturing Company were absent from work due to illness, given in Exercise 2-17.
(a) How many employees were absent less than three days annually? How many were absent less than six days due to illness?
(b) Convert the frequency distribution to a less than cumulative frequency distribution.
(c) Portray the cumulative distribution in the form of a less than cumulative frequency polygon.
(d) From the cumulative frequency polygon, calculate the number of days during which, about three out of four employees were absent, due to illness?

2-22. Refer to the frequency distribution of the lead time to fill an order given in Exercise 2-18.
(a) How many orders were filled in less than 10 days? In less than 15 days?
(b) Convert the frequency distribution to a less than cumulative frequency distribution.
(c) Develop a less than cumulative frequency polygon.
(d) About 60 percent of the orders were filled in fewer than how many days?

2.4 GRAPHICAL METHODS FOR DESCRIBING QUALITATIVE DATA

The histogram, the stem-and-leaf display, the frequency polygon, and the cumulative frequency polygon all are used to display a frequency distribution of quantitative data and all have visual appeal. In this section, we will examine the simple bar chart, the clustered bar chart, the stacked bar chart, the pie chart, and the line chart for depicting frequency distribution of qualitative data.

SIMPLE BAR CHART

A bar chart can be used to depict any level of measurement: nominal, ordinal, interval, or ratio. (Recall our discussion of the levels of measurement of data in Chapter 1.) Let us look at the following example.

Example 2-5

The following table shows the number of students enrolled in each of the five business programs in a certain community college in the year 2000.

Program	Students
Accounting	200
Industrial Relations	150
Financial Planning	250
Marketing	290
Management Studies	275

Represent this data using a bar chart.

Solution

The qualitative variable contains five categories: Accounting, Industrial Relations, Financial Planning, Marketing, and Management Studies. The frequency (number of students) for each category is given. As the variable is qualitative, we select a bar chart to depict the data. To draw the bar chart, we place categories on the horizontal axis at regular intervals. We mark the frequency of each category on the vertical axis. Above each category, we draw a rectangle whose height corresponds to the frequency of the category. With this chart, it is easy to see that the highest enrollment is in Marketing and the lowest is in the Industrial Relations program. This chart is vertical, but a horizontal bar chart can also be drawn by hand or using software such as Excel or Minitab. Horizontal bars are preferred for large category labels.

Chart 2-13 is produced using the data from Example 2-5 in Excel.

CLUSTERED BAR CHART

A clustered bar chart is used to summarize two or more sets of data. Consider Example 2-6.

Example 2-6

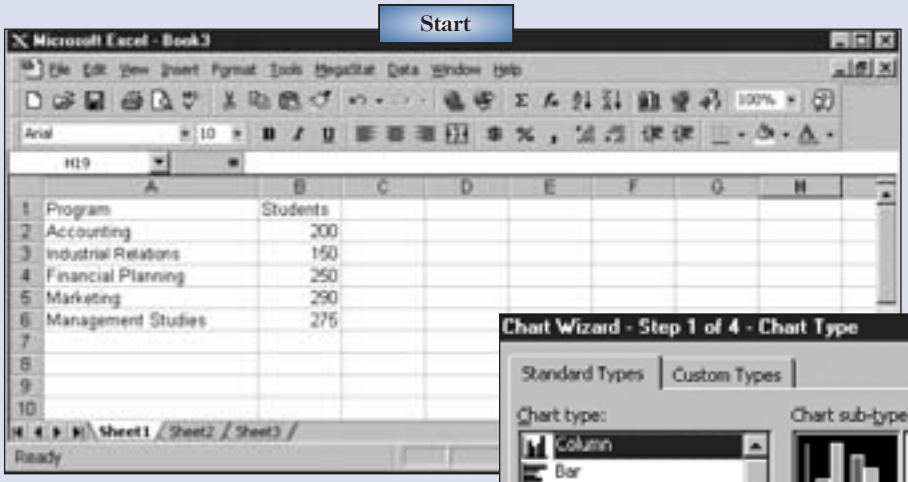
The following table shows the number of students enrolled in five business programs in a community college in 2000 and 2001.

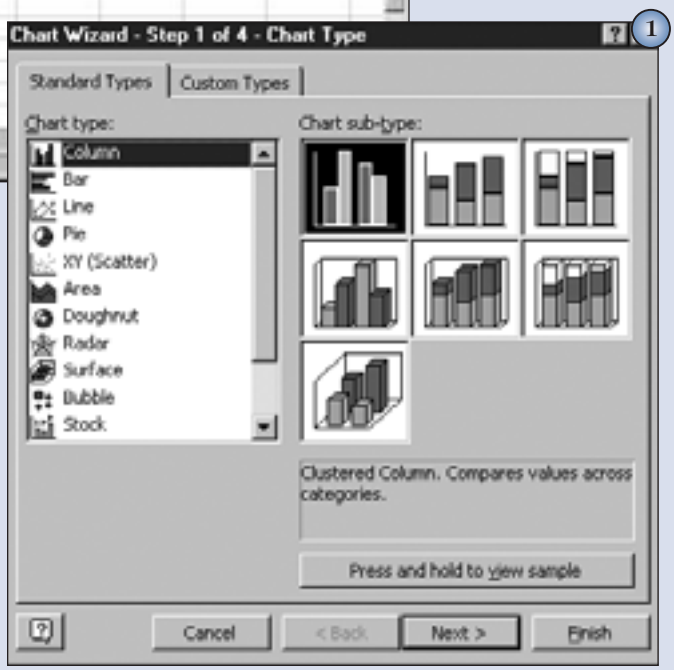
Program	Students (2000)	Students (2001)
Accounting	200	300
Industrial Relations	150	200
Financial Planning	250	230
Marketing	290	230
Management Studies	275	304

Construct a clustered bar chart for this data.

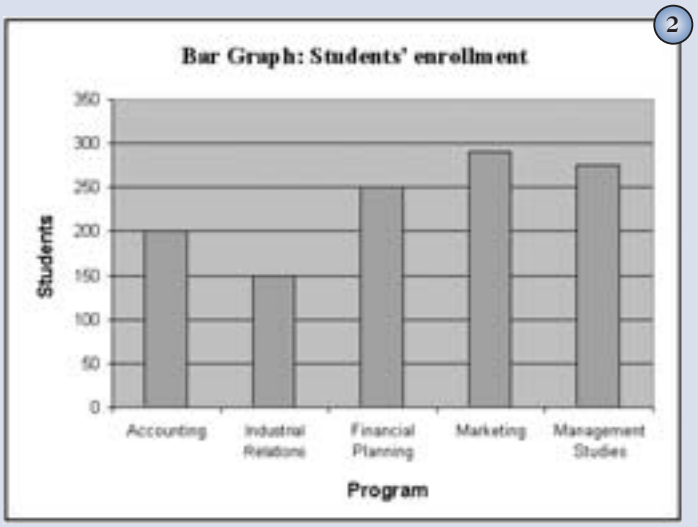
EXCEL CHART 2-13: Bar Chart for Enrollment in Different Programs

Start





1. Click Chart Wizard.
2. Select Chart Type = column and click Next.
3. Enter the location of data (in our case, data on names of five categories and enrollment figures) in the Data Range field.
5. Click Series and enter students in the name field.
6. Click Next.
7. In Chart Title, type *Bar Graph: Students' enrollment*.
8. In Category (X) axis box, type *Program*, and in Value (Y) axis box, type *Students*.
9. Click Finish.



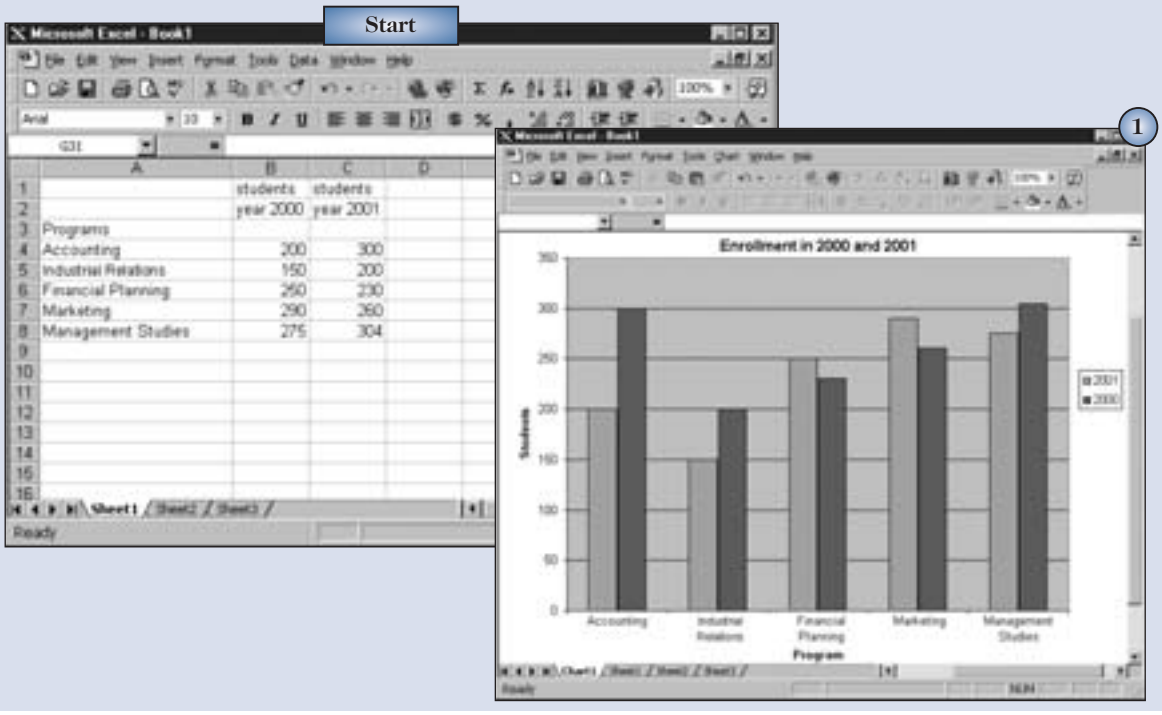
Solution As there are two sets of data (data series) for each category, we can summarize both sets of data simultaneously using a clustered bar chart. Steps to draw a clustered bar chart are the same as for the bar chart, except that for each category we draw two rectangles: one for 2000 and the other for 2001. The height of the Accounting rectangle for 2000 shows the frequency in that program for 2000 and the height of the Accounting rectangle for 2001 shows the frequency in that program for 2001. Both rectangles are side by side without any space between them. We repeat the process for each category.

We use Excel again to draw a clustered bar chart (see Chart 2-14). The instructions are almost the same as those in the case of a simple bar chart. The only difference is that, we enter the location of the entire data (data on names of categories and enrollment figures for 2000 and 2001) in the data range field. Then when we click on Series, we give a name to each of the series (in our case, we give names *year 2000* and *year 2001*). The computer output shows enrollment in 2000 and 2001 in one frame. The frequencies (number of students) in 2000 and 2001 for each program are shown side by side with no space between bars. We can see that enrollment in three programs (Accounting, Industrial Relations, and Management Studies) increased in 2001, while enrollment in Marketing and Financial Planning has decreased in 2001.

STACKED BAR CHART

In a stacked bar chart, the values in different data sets corresponding to the same category are stacked in a single bar. For example, an Excel output for stacked bar chart for data in Example 2-6 is shown in Chart 2-15.

EXCEL CHART 2-14: Clustered Bar Chart of Enrollment in 2000 and 2001

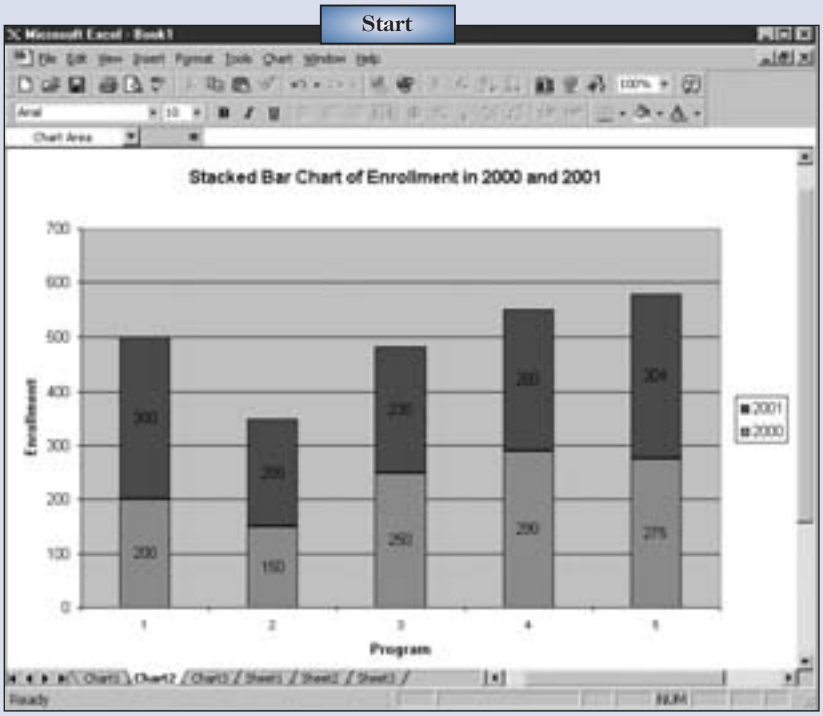


EXCEL CHART 2-15: Stacked Bar Chart Enrollments in 2000 and 2001

MICROSOFT EXCEL INSTRUCTIONS

1. Click Chart Wizard, then select chart subtype *Stacked Column*, and click Next.
2. Enter the data location in the Data Range field.
3. Click Series and type *2000* in the Name field; click Series 2 and type *2001* in the name field.
4. Click Next.
5. In the Chart Title field, type *stacked bar chart of enrollment in 2000 and 2001*.
6. Type *programs* in Category (*x*) field, and *enrollment* in the Value (*y*) field. Click on Data Label and then click the Show Value radio button.
7. Click Finish.

Start



The total height of the Accounting bar is 500, which equals total number of students in Accounting for the years 2000 and 2001 combined. This is divided into two parts: the bottom part (of height 200) shows enrollment during the year 2000, and the top part (of height 300) shows enrollment during 2001. We can compare the enrollments in 2000 and 2001 for each program. We can also compare the enrollments in different programs for 2000 because the baselines of bars representing programs are all anchored to the horizontal axis. For example, the enrollment in 2000 is highest in the Marketing program and lowest in the Industrial Relations program. Due to the floating baselines of bars for 2001, we are not able to visualize the difference in the enrollments for programs in 2001.

In a variation of the stacked bar chart called a 100-percent stacked bar chart, the corresponding data sets for each category are stacked as a percentage of the total. For example, in the data from Example 2-6, the percentage enrollment in Accounting for 2000 is 40 percent (200/500)(100) of the total enrollment in Accounting. The percentage enrollment for Accounting in 2001 is 60 percent.

To produce a 100-percent stacked bar chart, the menu sequence is the same except that we select Chart *Subtype 100-percent stacked column.*

PIE CHART

A pie chart, like a bar chart, is also used to summarize qualitative data. It is used to display the percentage of relative frequency of each category by partitioning a circle into sectors. The size of a sector is proportional to the percentage of relative frequency of the corresponding category.

Example 2-7 Draw a pie chart for the data in Example 2-5 (see page 60).

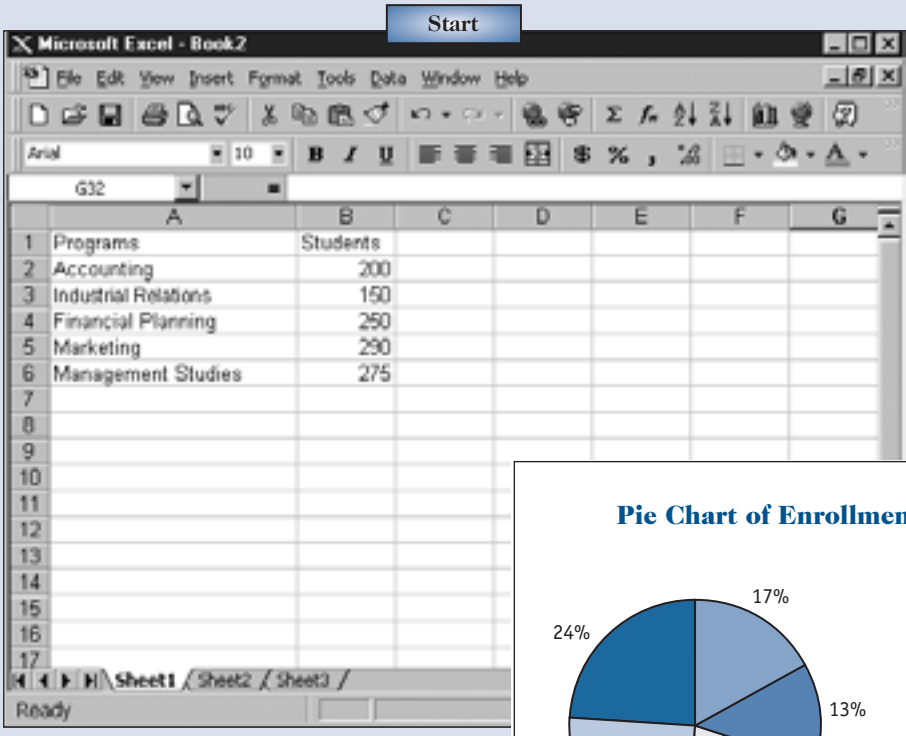
Solution To draw a pie chart, we first calculate the percentage of relative frequency for each category.

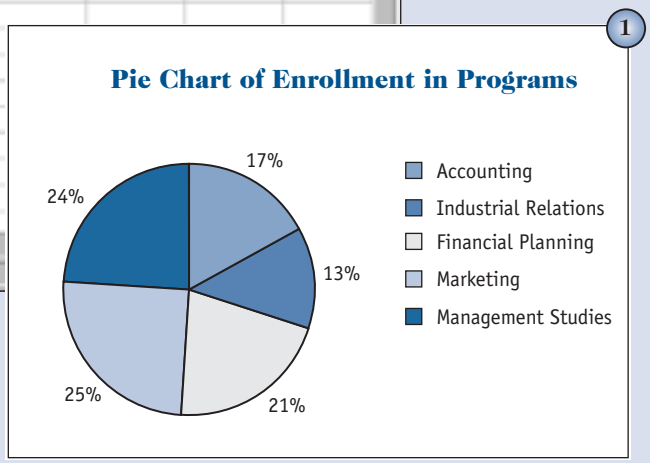
Program	Percentage Relative Frequency
Accounting	(200/1 165)(100) = 17.1
Industrial Relations	(150/1 165)(100) = 12.9
Financial Planning	(250/1 165)(100) = 21.5
Marketing	(290/1 165)(100) = 24.9
Management Studies	(275/1 165)(100) = 23.6

An entire circle corresponds to 360 degrees; therefore, a one-percent relative frequency observation corresponds to 3.6 degrees (360/100). Therefore, the sector angle for the Accounting program is (3.60)(17.1) = 61.6 degrees. Using a protractor, we mark 0 degrees, 90 degrees, 270 degrees, and 360 degrees on a circle. To plot 17.1

EXCEL CHART 2-16: Pie Chart of the Enrollment in Programs

Start





MICROSOFT EXCEL INSTRUCTIONS

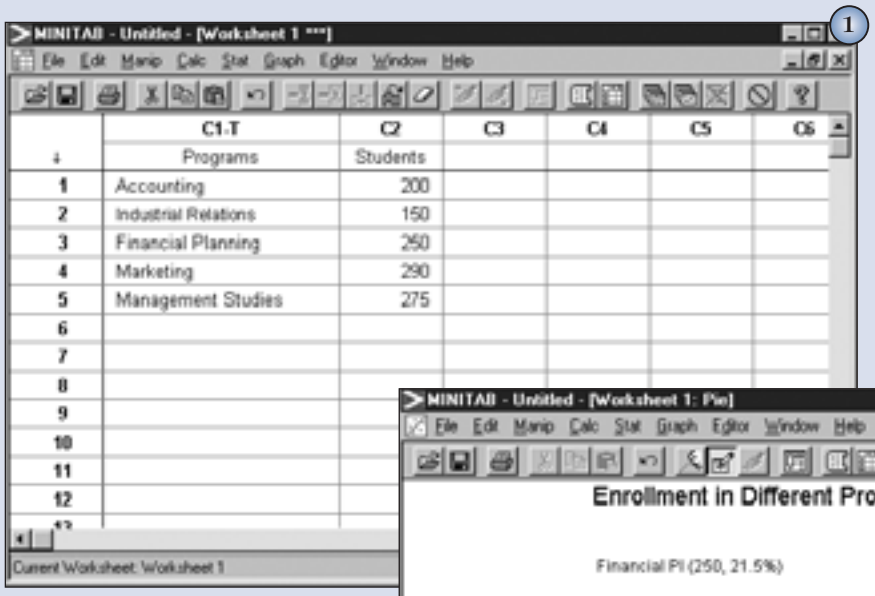
1. Click Chart Wizard, select Pie, and click Next.
2. Enter the location of data (categories and their frequencies) in the Data Range field.
3. Click Next, Title and enter *student enrollment* in the Chart Title field.
4. Click Data Label and the radio button Show Value.
5. Click Next, the radio button As a New Sheet, and then click Finish.

percent for Accounting, we draw a line from the centre of the circle to 0 degrees on the circle and then from the centre of the circle to 61.6 degrees on the circle. The area of this "slice" represents 17.1 percent of total students enrolled in the Accounting program. Next we add 17.1 percent of students enrolled in the Accounting program to 12.9 percent of students enrolled in the Industrial Relations program; the result is 30.0 percent. The angle corresponding to 30.0 percent is (3.60)(3.00) = 108 degrees. We draw a line from the centre of the circle to 108 degrees. Now the sector formed by joining the line from the centre of the circle to 61.8 degrees and from the centre of the circle to 108.2 degrees on the circle represents 12.9 percent of students enrolled in the Industrial Relations program. We continue the process for the other programs.

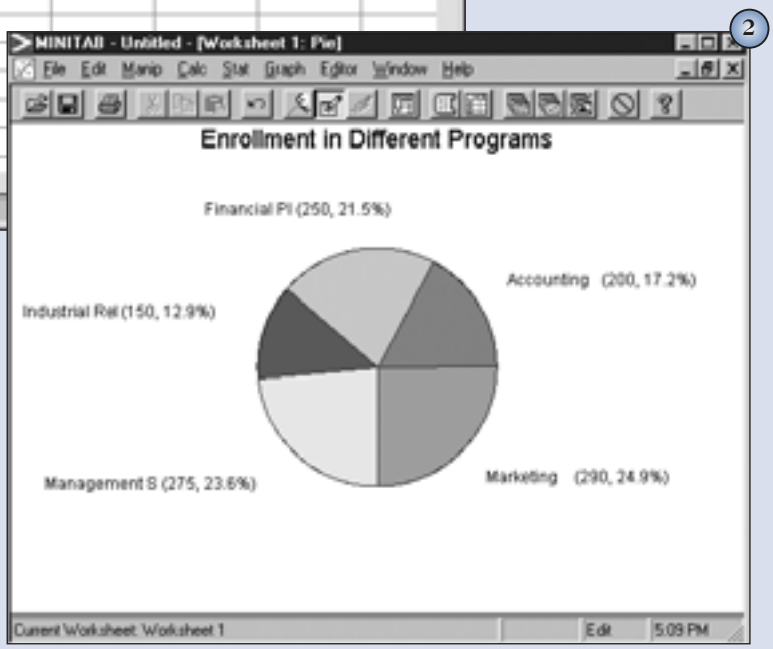
Because the areas of the sectors, or "slices," represent the relative frequencies of the categories, we can quickly compare them.

We can use Excel and Minitab to draw a pie chart. The instructions for this are given in Charts 2-16 and 2-17.

MINITAB CHART 2-17: Pie Chart of the Enrollment in Programs



	C1-T	C2	C3	C4	C5	C6
↓	Programs	Students				
1	Accounting	200				
2	Industrial Relations	150				
3	Financial Planning	250				
4	Marketing	290				
5	Management Studies	275				
6						
7						
8						
9						
10						
11						
12						



MINITAB INSTRUCTIONS

1. Click on Graph, Pie Chart, and Chart Table.
2. Click Categories in and type in the column containing the categories.
3. Click Frequencies in and type in the cell number for the frequencies.
4. Click Title Box and type the title (in our example, we give title. *Enrollment in Different Programs*).
5. Click OK.

It may be noted that in Excel, percentage values are rounded off to the nearest whole number. In Minitab, the percentage values are rounded off to one decimal place.

From the pie charts, we see that the highest enrollment is in the Marketing program and the lowest is in the Industrial Relations program. In addition, we also observe that the enrollment in Marketing is almost twice the enrollment in the Industrial Relations program. (The sector corresponding to Marketing is almost twice as big.)

A pie chart is meaningful when we do not use more than six or seven different data values. If we do, we lose clarity and cannot interpret the pie chart correctly. The other limitation of the pie chart is that we can use it for only one data series.

SELF-REVIEW 2-7

The total consumer credit (excluding mortgages) for the year 2000 is given below.

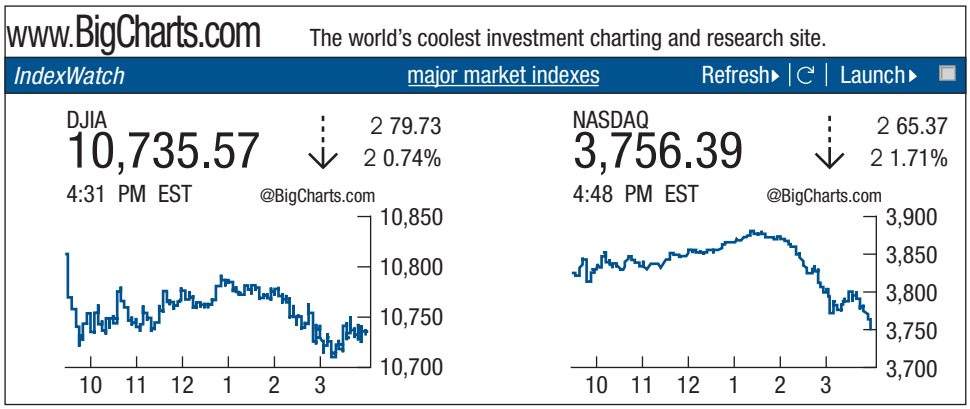
Financial Institution	Consumer Credit ($ millions)
Chartered Banks	119 837
Trust and Mortgage	1 959
Credit Unions	15 345
Life Insurance Companies	4 443
Finance Companies	12 734
Special-Purpose Corporations	29 008

(a) Draw a pie chart. (b) Interpret the pie chart.

LINE CHART

A line chart is often used to depict changes in the value of a variable over a period. Time values are labelled chronologically across the horizontal axis and values of the variable along the vertical axis. A line is drawn through data points. This line chart is also nown as a time-series chart. It is widely used in newspapers and magazines to show the variation of data over a given period, for example to depict the changing values over different periods of the Dow Jones Industrial Average, the Toronto Stock Exchange S&P/TSX composite index, and the NASDAQ. The line chart is also used to display two or more data series simultaneously for a given period, for example share price and price–earning ratios, (the return on shares of a company), versus the S&P/TSX. Chart 2-18 shows the Dow Jones Industrial Average and the NASDAQ, the two most reported measures of business activity, on June 6, 2000.

CHART 2-18: Market Summary on June 6, 2000



2.5 MISLEADING GRAPHS

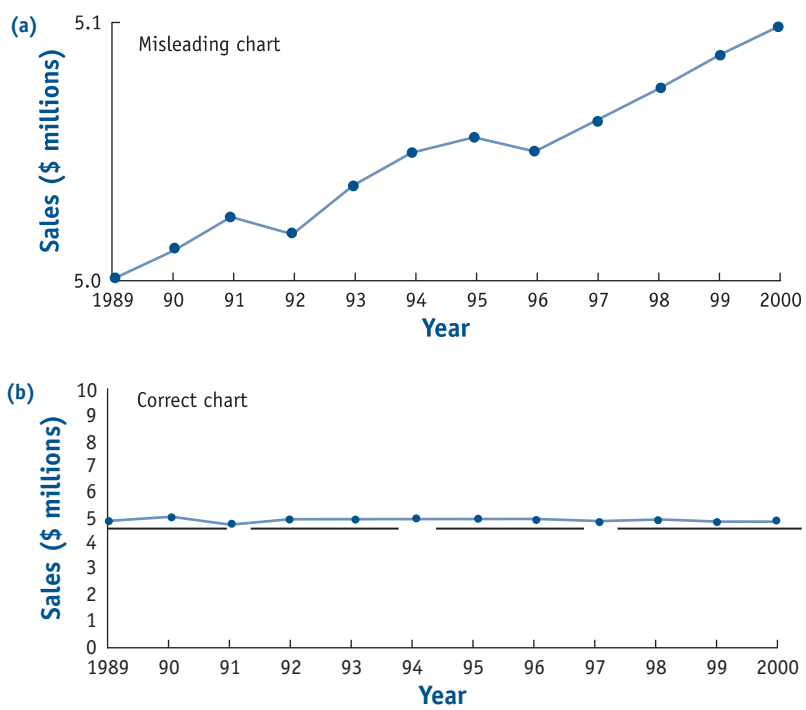
When you purchase a computer for your home or office, it usually includes some graphics and spreadsheet software, such as Excel. This software will produce effective charts and graphs; however, you must be careful not to mislead readers or misrepresent the data. In this section we present several examples of charts and graphs that are misleading. Whenever you see a chart or graph, study it carefully. Ask yourself what the writer is trying to show you. Could the writer have any bias?

One of the easiest ways to mislead the reader is to make the range of the *y*-axis very small in terms of the units. A second method is to begin at some value other than 0 on the *y*-axis. In Chart 2-19(a), it appears there has been a dramatic increase in sales from 1989 to 2000. However, during that period, sales increased only 2 percent (from $5.0 million to $5.1 million)! In addition, observe that the *y*-axis does not begin at 0.

The vertical axis does not have to start at zero. It can start at some value other than zero. If we cannot detect the variation in data with zero as the starting point on the vertical axis, we should consider some value other than zero so that we can see the variation.[2]

Chart 2-19(b) gives the correct impression of the trend in sales. Sales are almost flat from 1989 to 2000; that is, there has been practically no change in sales during the 10-year period.

CHART 2-19: Sales of Matsui Nine-Passenger Vans, 1989–2000

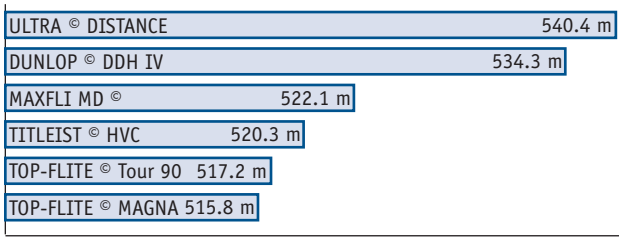


Without much comment, we ask you to look at each of the following scenarios and carefully decide whether the intended message is accurate.

Scenario 1

The following chart was adapted from an advertisement for the new Wilson Ultra Distance golf ball. The chart shows that the new ball gets the longest distance, but what is the scale for the horizontal axis? How was the test conducted?

Maybe everybody can't hit a ball like John Daly. But everybody wants to. That's why Wilson © is introducing the new Ultra © Distance ball. Ultra Distance is the longest, most accurate ball you'll ever hit.

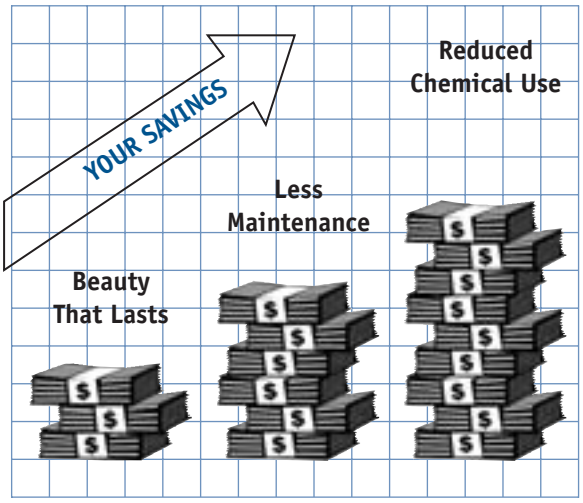


Combined yardage with a driver, #5 iron, and #9 iron, Ultra Distance is clearly measurably longer.

Wilson has totally redesigned this ball from the inside out, making Ultra Distance a major advancement in golf technology.

Scenario 2

Fibre Glass Inc., based in Red Deer, Alberta, makes and installs Fibre Tech, fibreglass coatings for swimming pools. The following chart was included in a brochure. Is the comparison fair? What is the scale for the vertical axis? Is the scale in dollars or in percent?



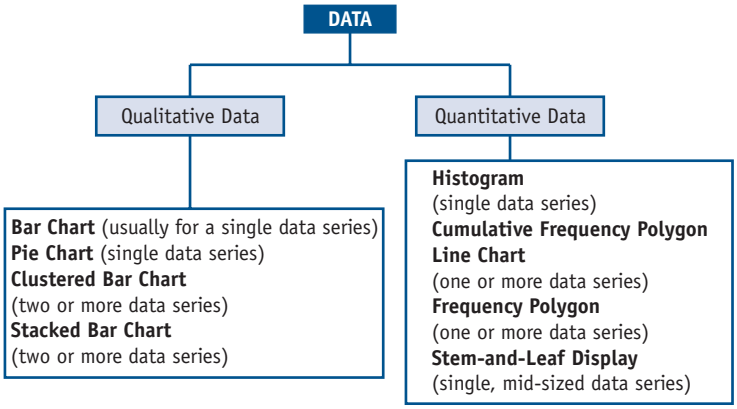
Fibre Tech Reduces Chemical Use, Saving You Time and Money.

- Saves up to 60 percent on chemical costs alone.
- Reduces water loss, which means less need to replace chemicals and up to 10-percent warmer water (reducing heating costs, too).
- Fibre Tech pays for itself in reduced maintenance and chemical costs.

Misleading information may be given by an improper scaling used in a chart or graph where an attempt is made to change all the dimensions simultaneously in response to a change in one-dimensional data.

Again, we caution you. When you see a chart or graph, particularly as part of an advertisement, be careful. Look at the scales used on the x-axis and the y-axis.

Guidelines for Selecting a Graph to Summarize Data



Bar and pie charts both are used to display qualitative data (single data series). Generally, a bar chart is preferred to display a single data series because it is easier to visualize changes within a data set. According to psychologists who have studied visual preferences,[3] it is more complicated to interpret the relative size of angles in a pie chart than to judge the length of the rectangles in a bar chart.

To compare two or more qualitative data sets, both clustered bar charts and stacked bar charts are used; however, in the case of stacked bar charts, it is difficult to compare data *visually* due to the floating baselines of rectangles that are stacked on the bottom rectangles.

The histogram is a more popular graphic to summarize a large, quantitative, single-data set. It is not used to compare two or more quantitative data series. Instead, frequency polygons, cumulative frequency polygons and line charts are used to compare two or more data series in a single graphic frame. The stem-and-leaf display is very convenient for mid-sized quantitative data.

EXERCISES 2-23 TO 2-28

2-23. A small-business consultant is investigating the performance of several companies. The sales in 2000 (in thousands of dollars) for the selected companies are listed below. The consultant wants to include a chart in a report comparing the sales of the six companies. Use a bar chart to compare the fourth-quarter sales of these corporations and write a brief report summarizing the bar chart.

Corporation	Fourth-Quarter Sales ($ thousands)
Hoden Building Products	1 645.2
J & R Printing, Inc.	4 757.0
Long Bay Concrete Construction	8 913.0
Mancell Electric and Plumbing	627.1
Maxwell Heating and Air Conditioning	24 612.0
Mizella Roofing & Sheet Metal	191.9

2-24. The Gentle Corporation in Montreal, Quebec sells fashion apparel for men and women and a broad range of other related products. It serves customers in the United States and Canada by mail. Listed below are the net sales from 1996 to 2001. Draw a line chart depicting the net sales over the period.

Year	Net Sales ($ millions)
1996	525.00
1997	535.00
1998	600.50
1999	625.80
2000	645.70
2001	758.75

2-25. The following are long-term business credit amounts ($millions) from 1996 to 2000 (Canada). Draw a line chart depicting the long-term business credit for the period.

Year	Long-Term Business Credit ($ millions)
1996	357 946
1997	392 846
1998	432 909
1999	470 250
2000	504 850

Source: Adapted from Statistics Canada, Bank of Canada, CANSIM, Matrix 2567

2-26. The following are the unemployment rates in Canada from 1996 to 2000. Draw a line chart for the unemployment rate for the period 1996 to 2000. Describe the trend for the unemployment rate.

Year	Unemployment Rate (%)
1996	9.6
1997	9.1
1998	8.3
1999	7.6
2000	6.8

Source: Adapted from Statistics Canada, CANSIM, Matrix 3472; and Catalogue No. 71-529-XPB

2-27. The following are gross domestic products (GDP) at market prices from 1990 to 2000. Draw a line chart to show the highest and lowest GDP at market prices.

Year	GDP at market prices ($ millions)
1990	705 464
1991	692 247
1992	698 544
1993	714 583
1994	748 350
1995	769 082
1996	780 916
1997	815 013
1998	842 002
1999	880 254
2000	921 485

Source: The Centre for the Study of Living Standards: www.csls.ca

2-28. The following table shows the gross domestic product (GDP) for eight countries in 2000. Develop a bar chart and summarize the results.

Country	GDP ($ trillions)
USA	9.3
Japan	3.9
Germany	2.2
France	1.5
UK	1.4
Italy	1.2
Canada	0.7

CHAPTER OUTLINE

I. A *frequency distribution* is a grouping of data into mutually exclusive categories showing the number of observations in each category.
   A. The steps in constructing a frequency distribution are:
      1. Decide how many classes you need.
      2. Determine the class width or interval.
      3. Set the individual class limits.
      4. Tally the raw data into classes.
      5. Count the number of tallies in each class.
   B. The *class frequency* is the number of observations in each class.
   C. The *class interval* is the difference between the lower limit and the upper limit of a class.
   D. The *class midpoint* is halfway between the lower limit and the upper limit of a class.

II. A *relative frequency distribution* shows the fraction of the observations in each class.

III. A *stem-and-leaf display* provides a frequency distribution of a data set and at the same time shows a graphic similar to a histogram.
   A. The leading digit is the stem and the trailing digits are the leaves.
   B. The advantages of the stem-and-leaf display over a frequency distribution include:
      1. The identity of each observation is not lost.
      2. The digits themselves give a picture of the distribution.

IV. There are two methods for graphically portraying a frequency distribution.
   A. A *histogram* portrays the number of frequencies in each class in the form of rectangles.
   B. A *frequency polygon* consists of line segments connecting the points formed by the intersections of the class midpoints and the class frequencies.

V. A *cumulative frequency polygon* shows the number of observations below a certain value.

CHAPTER EXERCISES 2-29 TO 2-51

2-29. A data set consists of 83 observations. How many classes would you recommend for a frequency distribution?

2-30. A data set consists of 145 observations that range from 56 to 490. What size class interval would you recommend?

2-31. The following is the number of minutes it takes to commute from home to work for a group of automobile executives.

28	25	48	37	41	19	32	26	16	23	23	29	36
31	26	21	32	25	31	43	35	42	38	33	28	

(a) How many classes would you recommend?
(b) What class interval would you suggest?
(c) What would you recommend as the lower limit of the first class?
(d) Organize the data into a frequency distribution.
(e) Comment on the shape of the frequency distribution.

2-32. The following data are the weekly amounts (in dollars) spent on groceries for a sample of households. This data is also found on the accompanying CD in Exercise 2-32.xls.

271	363	159	76	227	337	295	319	250	279	205	279
266	199	177	162	232	303	192	181	321	309	246	278
50	41	335	116	100	151	240	474	297	170	188	320
429	294	570	342	279	235	434	123	325			

(a) How many classes would you recommend?
(b) What class interval would you suggest?
(c) What would you recommend as the lower limit of the first class?
(d) Organize the data into a frequency distribution.

2-33. The following stem-and-leaf display shows the number of minutes spent per week, watching daytime TV for a sample of university students.

2	0	0 5
3	1	0
6	2	1 3 7
10	3	0 0 2 9
13	4	4 9 9
24	5	0 0 1 5 5 6 6 7 7 9 9
30	6	0 2 3 4 6 8
(7)	7	1 3 6 6 7 8 9
33	8	0 1 5 5 8
28	9	1 1 2 2 3 7 9
21	10	0 2 2 3 6 7 8 9 9
12	11	2 4 5 7
8	12	4 6 6 8
4	13	2 4 9
1	14	5

(a) How many students were studied?
(b) How many observations are in the second class?
(c) What is the smallest value? the largest value?
(d) List the actual values in the fourth row.
(e) How many students watched less than 60 minutes of TV?
(f) How many students watched 100 minutes or more of TV?
(g) What is the middle value?
(h) How many students watched at least 60 minutes but less than 100 minutes?

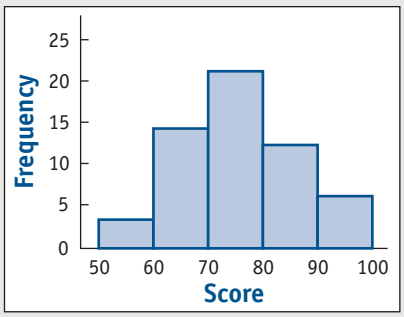
2-34. The following stem-and-leaf display reports the number of orders received per day by a mail-order firm.

1	9	1
2	10	2
5	11	2 3 5
7	12	6 9
8	13	2
11	14	1 3 5
15	15	1 2 2 9
22	16	2 2 6 6 7 7 8
27	17	0 1 5 9 9
(11)	18	0 0 0 1 3 3 4 6 7 9 9
17	19	0 3 3 4 6
12	20	4 6 7 9
8	21	0 1 7 7
4	22	4 5
2	23	1 7

(a) How many days were studied?
(b) How many observations are in the fourth class?
(c) What is the smallest value and what is the largest value?

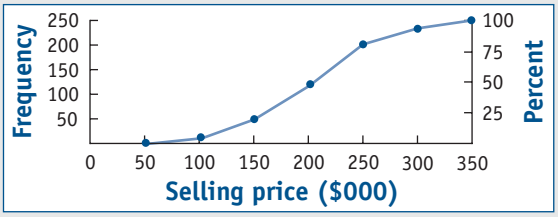
(d) List the actual values in the sixth class.
(e) How many days did the firm receive less than 140 orders?
(f) How many days did the firm receive 200 or more orders?
(g) On how many days did the firm receive 180 orders?
(h) What is the middle value?

2-35. The following histogram shows the scores on the first statistics exam.



(a) How many students took the exam?
(b) What is the class interval?
(c) What is the class midpoint for the first class?
(d) How many students earned a score of less than 70?

2-36. The following chart summarizes the selling price of homes sold last month in Victoria, B.C.



(a) What is the chart called?
(b) How many homes were sold during the last month?
(c) What is the class interval?
(d) About 75 percent of the homes sold for less than what amount?
(e) 175 of the homes sold for less than what amount?

2-37. A chain of ski and sportswear shops catering to beginning skiers, headquartered in Banff, Alberta, plans to conduct a study of how much a beginning skier spends on his or her initial purchase of equipment and supplies. Based on these figures, they want to explore the possibility of offering combinations, such as a pair of boots and a pair of skis, to induce customers to buy more. A sample of their cash-register receipts revealed these initial purchases (in dollars):

140	82	265	168	90	114	172	230	142	86	125	235
212	171	149	156	162	118	139	149	132	105	162	126
216	195	127	161	135	172	220	229	129	87	128	126
175	127	149	126	121	118	172	126				

(a) Arrive at a suggested class interval. Use five classes, and let the lower limit of the first class be $80.
(b) What would be a better class interval?
(c) Organize the data into a frequency distribution.
(d) Interpret your findings.

2-38. The numbers of shareholders for a selected group of large companies (in thousands) are listed in Exercise 2-38.xls on the CD-ROM accompanying the text.

The numbers of shareholders are to be organized into a frequency distribution and several graphs drawn to portray the distribution.

(a) Using seven classes and a lower limit of 130, construct a frequency distribution.
(b) Portray the distribution in the form of a frequency polygon.
(c) Portray the distribution in a less than cumulative frequency polygon.
(d) Based on the polygon, three out of four (75 percent) of the companies have how many shareholders or fewer?
(e) Write a brief analysis of the number of shareholders based on the frequency distribution and graphs.

2-39. The following is the list of top-selling drugs in 2002. Draw an appropriate chart to portray the data.

Products	**Sales ($ billions)**
Lipitor (cholesterol-reducing)	5.7
Zocor (cholesterol-reducing)	5.3
Claritin-family (anti-histamine)	4.2
Norvasc (calcium-antagonist)	4.1
Losec (anti-ulcerant)	3.6

2-40. The Midland National Bank selected a sample of 40 student chequing accounts. Below are their end-of-the-month balances.

404	74	234	149	279	215	123	55	43	321	87	234
68	489	57	185	141	758	72	863	703	125	350	440
37	252	27	521	302	127	968	712	503	489	327	608
358	425	303	203								

(a) Tally the data into a frequency distribution using $100 as a class interval and $0 as the starting point.
(b) Draw a cumulative frequency polygon.
(c) The bank considers any student with an ending balance of $400 or more a "preferred customer." Estimate the percentage of preferred customers.
(d) The bank is also considering a service charge to the lowest 10 percent of the ending balances. What would you recommend as the cut-off point between those who have to pay a service charge and those who do not?

2-41. The following are grades of students in Math 1021 in 2002.

57	81	47	87	21	47	57	64	86	41	84	48	80	58	88
73	30	64	84	77	28	95	40	42	10	72	61	13	56	47
55	48	60	99	88	86	95	49							

(a) Construct a stem-and-leaf display.
(b) Summarize your conclusion.

2-42. A recent study of home technologies reported the number of hours of personal computer usage per week for a sample of 60 persons. Excluded from the study were people who worked out of their homes and used the computer as a part of their work.

9.3	5.3	6.3	8.8	6.5	0.6	5.2	6.6	9.3	4.3	6.3	2.1	2.7	0.4
3.7	3.3	1.1	2.7	6.7	6.5	4.3	9.7	7.7	5.2	1.7	8.5	4.2	5.5
5.1	5.6	5.4	4.8	2.1	10.1	1.3	5.6	2.4	2.4	4.7	1.7	2.0	6.7
1.1	6.7	2.2	2.6	9.8	6.4	4.9	5.2	4.5	9.3	7.9	4.6	4.3	4.5
9.2	8.5	6.0	8.1										

(a) Organize the data into a frequency distribution. How many classes would you suggest? What value would you suggest for a class interval?
(b) Draw a histogram. Interpret your result.

2-43. Merrill Lynch recently completed a study regarding the size of investment portfolios (stocks, bonds, mutual funds, and certificates of deposit) for a sample of clients in the 40 to 50 age group. Listed below are the values of all the investments for the 70 participants in the study.

669.9	7.5	77.2	7.5	125.7	516.9	219.9	645.2
301.9	235.4	716.4	145.3	26.6	187.2	315.5	89.2
36.4	616.9	440.6	408.2	34.4	296.1	185.4	526.3
380.7	3.3	363.2	51.9	52.2	107.5	82.9	63.0
228.6	308.7	126.7	430.3	82.0	227.0	321.1	403.4
39.5	124.3	118.1	23.9	352.8	156.7	276.3	23.5
31.3	301.2	35.7	154.9	174.3	100.6	236.7	171.9
221.1	43.4	212.3	243.3	315.4	5.9	1002.2	171.7
295.7	437.0	87.8	302.1	268.1	899.5		

(a) Organize the data into a frequency distribution. How many classes would you suggest? What value would you suggest for a class interval?
(b) Draw a histogram. Interpret your result.

2-44. The following are gross domestic products (GDP) per head in the following European countries (in dollars). Develop a bar chart depicting this information.

Country	GDP per head ($)
Austria	26 740
Denmark	32 576
France	24 956
Germany	27 337
Greece	11 860
Norway	35 853
Turkey	3 120

2-45. Care Heart Association reported the following percentage breakdown of expenses. Draw a pie chart depicting the information. Interpret the results.

Category	Percent
Research	32.3
Public Health Education	23.5
Community Service	12.6
Fundraising	12.1
Professional and Educational Training	10.9
Management and General	8.6

2-46. In its 2002 annual report, Schering-Plough Corporation reported the income, in millions of dollars, for 1995 to 2002 as listed below. Develop a line chart depicting the results and comment on your findings.

Year	Income ($ millions)
1995	1053
1996	1213
1997	1444
1998	1756
1999	2110
2000	2900
2001	3595
2002	4550

2-47. The following table shows Canada's exports in merchandise trade with its principal trading partners:

Principal Trading Partner	December 1999 ($ millions)	December 2000 ($ millions)
USA	27 243	31 876
Japan	764	824
European Union	1 616	1 896
Other OECD Countries	728	682
All Other Countries	1 510	1 572

Source: Adapted from Statistics Canada, CANSIM, Matrix 3618

(a) Draw a clustered bar graph.
(b) Name the trading partner to whom we exported more than any other trading partner in 2000.

2-48. The following is the population distribution of Canada by sex from 1996 to 2000. Draw a stacked bar graph and comment on your findings.

Year	Male	Female
1996	14 691 777	14 980 115
1997	14 850 874	15 136 340
1998	14 981 482	15 266 467
1999	15 104 717	15 388 716
2000	15 232 909	15 517 178

Source: Adapted from Statistics Canada, CANSIM, Matrix 6213

2-49. Cash receipts from milk and cream sold from farms in six provinces in 2001 are given below. Draw a pie chart to display the data set.

Province	Cash Receipts ($ thousands)
Alberta	318 454
B.C.	336 977
Manitoba	154 029
N.B.	69 041
Nova Scotia	90 368
P.E.I.	50 987

Source: Adapted from Statistics Canada, CANSIM, Matrices 5650–5651; and Catalogue No. 23-001-XIB

2-50. The following are exports of goods to the Organization for Economic Co-operation and Development (OECD[4]) from 1995 to 2000. Draw a line chart and describe the trend in exports of goods to the OECD.

Year	Export of Goods ($ millions)
1995	4563.4
1996	5087.8
1997	8033.5
1998	7560.4
1999	7160.9
2000	8159.3

Source: Adapted from Statistics Canada, CANSIM, Matrices 3651 and 3685

2-51. The following table shows the volume (in kilolitres) of milk and cream sold from farms in 2001 in six Canadian provinces. Draw a simple bar chart to depict the data.

Province	Volume of Milk and Cream Sold from Farms (kL)
NFLD	33 583
P.E.I.	94 472
Nova Scotia	173 985
N.B.	134 428
Manitoba	294 674
Alberta	208 198

Source: Adapted from Statistics Canada, CANSIM, Matrices 5650–5651; and Catalogue No. 23-001-XIB

WWW.exercises.ca 2.52 TO 2.53

2-52. Go to the Web site: www.statcan.ca. Click English, Canadian Statistics, Education, Graduates, and Secondary School Graduates. Draw a bar chart depicting the number of school graduates in each province. Summarize your findings.

2-53. Go to the Statistics Canada Web site (www.statcan.ca) and click English/ Canadian Statistics / Labour, Employment, and Unemployment /Earnings. Select two data series for a given category and draw a clustered bar graph.

COMPUTER DATA EXERCISES 2-54 TO 2-57

2-54. The file Exercise 2-32.xls contains the amount spent on groceries by households.
(a) How many classes would you recommend?
(b) What class interval or width would you suggest?
(c) Organize the data into a frequency distribution.
(d) Use Excel to draw a histogram. Use the number of classes you recommended. Describe the shape of the histogram.
(e) Draw the histogram using Excel. Let Excel decide the number of classes. Describe the shape of the histogram.

2-55. Use the data in the file Exercise 2-32.xls to draw a stem-and-leaf diagram. Use Minitab. Use the same data to draw a histogram. Do you find the stem-and-leaf diagram more informative than the histogram? Explain.

2-56. Refer to the OECD data on the CD, which reports information on census, economic, and business data for 29 countries. Develop a stem-and-leaf diagram for the variable regarding the percentage of the workforce that is over 65 years of age. Are there any outliers? Briefly describe the data.

2-57. The file Exercise 2-57.xls contains the amount of money spent by beginning skiers on the purchase of equipment and supplies.
(a) Draw a cumulative frequency polygon using Excel. Do not specify the Bin.
(b) Estimate the proportion of the amount of money spent on the purchase of equipment and supplies that is less than $143.
(c) How many skiers spent less than $173.50 on the purchase of equipment and supplies?
(d) Draw a suitable graph to summarize the data.

CHAPTER 2 ANSWERS TO SELF-REVIEW

2-1. (a) The raw data.

(b)

Commission	Number of Salespeople
1400 up to 1500	2
1500 up to 1600	5
1600 up to 1700	3
1700 up to 1800	1
Total	11

(c) Class frequencies.

(d) The largest concentration of commissions is in the class \$1500 up to \$1600. The smallest commission is about \$1400 and the largest is about \$1800.

2-2. (a) 5, ($2^4 = 16$, less than 24, and $2^5 = 32$, more than 24. Hence, $k = 5$ is suggested).

(b) 10, found by rounding up $\left[\frac{71-25}{5}\right] = 9.2$.

(c)

Class	Frequency
23 up to 33	3
33 up to 43	1
43 up to 53	3
53 up to 63	9
63 up to 73	8
Total	24

2-3. (a) 21

(b) 26.3 percent

(c) 17.5 percent (found by $(0.1 + 0.05 + 0.025) \times 100$)

2-4.

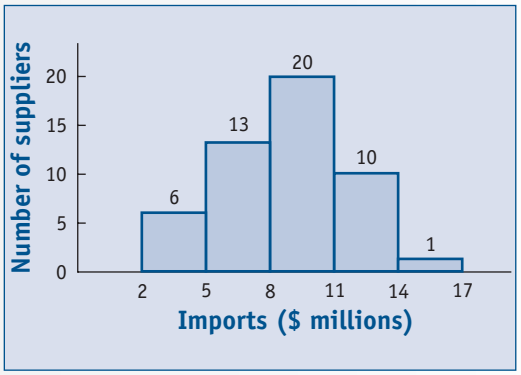
7	7
8	0 0 1 3 4 8 8
9	1 2 5 6 6 8 9
10	1 2 4 8
11	2 6

(a) 8

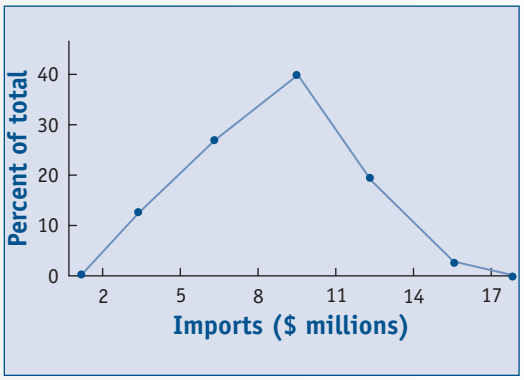
(b) 10.1, 10.2, 10.4, 10.8

(c) 11.6 and 7.7

2-5. (a)



(b)



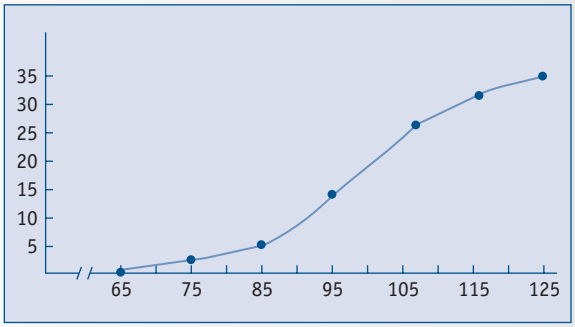
(c) The smallest annual sales volume of imports by a supplier is about \$2 million, the highest about \$17 million. The concentration is between \$8 million and \$11 million.

2-6. (a) A frequency distribution.

(b)

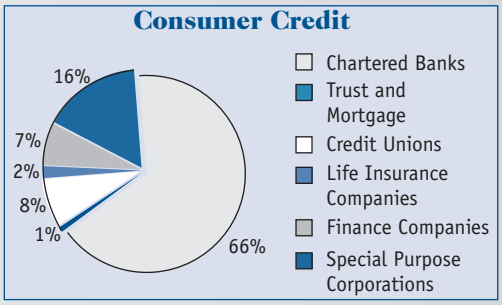
Annual Net Profits (\$000)	Cumulative Number
65 up to 75	1
75 up to 85	7
85 up to 95	14
95 up to 105	26
105 up to 115	31
115 up to 125	34

(c)



The number of companies with annual net profits of less than $105 000 is about 26.

2-7. (a)



(b) Chartered banks provided 66 percent of the total consumer credit, Special Purpose Corporations provided 16 percent of the total consumer credit, and so on.