3 Organizing and Displaying Data

CHAPTER OUTLINE

3.1 **Classifying and Organizing Data**
Explains and illustrates numerical scales, and distinguishes among qualitative data, discrete quantitative data, and continuous quantitative data

3.2 **Figures, Tables, and Graphs**
Gives a brief overview of each

3.3 **Creating Tables**
Gives instructions on how to organize data in the form of a frequency table

3.4 **Graphing Data**
Discusses and illustrates various methods of graphing, with an emphasis on those that apply specifically to frequency distributions

✔ LEARNING OBJECTIVES

After studying this chapter, you should be able to

1. Distinguish between
   a. qualitative and quantitative variables
   b. discrete and continuous variables
   c. symmetrical, bimodal, and skewed distributions
   d. positively and negatively skewed distributions
2. Construct and interpret a frequency table that includes class intervals, class frequency, valid percent, and cumulative percent
3. Indicate the appropriate types of graphs for displaying quantitative and qualitative data
4. Distinguish which forms of data presentation are appropriate for different situations

(Continued)

5. Construct a histogram, frequency polygon, ogive, bar chart, and box-and-whisker plot
6. Distinguish among and interpret various graphs
7. Determine and interpret percentiles from an ogive

3.1 CLASSIFYING AND ORGANIZING DATA

To successfully explain your data, one of your first tasks is to classify and organize the data. There are three general ways of organizing and presenting data: tables, graphs, and numerical techniques. Each of these methods will be illustrated by reference to a sample of 100 individuals, selected by systematic random sampling from the Honolulu Heart Study population of 7683 (Phillips, 1972). The data for this sample are presented in Table 3.1 and are also on this book's Web site at www.mhhe.com/Kuzma5e.

Nominal, Ordinal, Interval, and Ratio Scales

Before proceeding, we should discuss the meanings attached to various numbers. One approach is to identify data by measurement scale. There are four commonly used scales: nominal, ordinal, interval, and ratio. The scale used to classify the data to a large extent determines what can be done with the data. This statement is best explained by using specific examples, starting with the **nominal scale.** Nominal scales are used primarily for grouping or categorizing data. Variables that yield nominal-level data are frequently referred to as **qualitative variables** (e.g., zip code, hair color, gender, name of college or university, social security number). In Table 3.1, ID (number) and smoking status (smoker versus nonsmoker) are qualitative variables. Note what these variables have in common. All can be used to group or categorize data; you can organize subjects based on any of these characteristics. Also note that these variables do not typically have a numerical value associated with them. However, to use these variables in some type of statistical analysis, it is necessary to assign them a numerical value. For instance, as in Table 3.1, smoking status might be given the numerical values of 0 and 1 for smoker and nonsmoker, respectively. Hair color might be anything from 1 to however many categories you choose to create. Subject identification number (ID) is typically some unique number for each subject. It can be something as simple as the order in which their individual data are entered into the spreadsheet.

Note that these numbers are used simply to group or categorize data and that mathematical manipulations will yield meaningless information. For example, if your class were organized by gender or hair color, of what value would be the average gender or average hair color? Or, using the data from Table 3.1,

Table 3.1 Data for a Sample of 100 Individuals from the Honolulu Heart Study Population of 7683 Persons, 1969

ID	Educational Level	Weight (kg)	Height (cm)	Age	Smoking Status	Physical Activity at Home	Blood Glucose	Serum Cholesterol	Systolic Blood Pressure	Body Mass Index (BMI)
1	2	70	165	61	1	1	107	199	102	25.7
2	1	60	162	52	0	2	145	267	138	22.9
3	1	62	150	52	1	1	237	272	190	27.6
4	2	66	165	51	1	1	91	166	122	24.2
5	2	70	162	51	0	1	185	239	128	26.7
6	4	59	165	53	0	2	106	189	112	21.7
7	1	47	160	61	0	1	177	238	128	18.4
8	3	66	170	48	1	1	120	223	116	22.8
9	5	56	155	54	0	2	116	279	134	23.3
10	2	62	167	48	0	1	105	190	104	22.2
11	4	68	165	49	1	2	109	240	116	25.0
12	1	65	166	48	0	1	186	209	152	23.6
13	1	56	157	55	0	2	257	210	134	22.7
14	2	80	161	49	0	1	218	171	132	30.9
15	3	66	160	50	0	2	164	255	130	25.8
16	4	91	170	52	0	2	158	232	118	31.5
17	3	71	170	48	1	1	117	147	136	24.6
18	5	66	152	59	0	2	130	268	108	28.6
19	1	73	159	59	0	2	132	231	108	28.9
20	4	59	161	52	0	1	138	199	128	22.8
21	1	64	162	52	1	1	131	255	118	24.4
22	3	55	161	52	1	1	88	199	134	21.2
23	2	78	175	50	1	1	161	228	178	25.5
24	2	59	160	54	0	1	145	240	134	22.8
25	3	51	167	48	1	2	128	184	162	18.3
26	3	83	171	55	0	1	231	192	162	28.4
27	2	66	157	49	1	2	78	211	120	26.8
28	4	61	165	51	0	1	113	201	98	22.4
29	2	65	160	53	0	1	134	203	144	25.4
30	3	75	172	49	0	1	104	243	118	25.4
31	4	61	164	49	0	2	122	181	118	22.7
32	1	73	157	53	1	2	442	382	138	29.6
33	2	66	157	52	0	1	237	186	134	26.8
34	1	73	155	48	0	2	148	198	108	27.8
35	2	61	160	53	0	1	231	165	96	23.8
36	3	68	162	50	0	2	161	219	142	25.9
37	2	52	157	50	0	2	119	196	122	21.1
38	5	73	162	50	0	1	185	239	146	27.8
39	1	52	165	61	1	2	118	259	126	19.1
40	1	56	162	53	1	1	98	162	176	21.3
41	3	67	170	48	1	2	218	178	104	23.2
42	1	61	160	47	0	1	147	246	112	23.8
43	3	52	166	62	1	2	176	176	140	18.9
44	2	61	172	56	1	2	106	157	102	20.6
45	3	62	164	55	1	2	109	179	142	23.1
46	2	56	155	57	1	2	138	231	146	23.3
47	1	55	157	50	0	2	84	183	92	22.3

(*Continued*)

Table 3.1 (*Continued*)

ID	Educational Level	Weight (kg)	Height (cm)	Age	Smoking Status	Physical Activity at Home	Blood Glucose	Serum Cholesterol	Systolic Blood Pressure	Body Mass Index (BMI)
48	3	66	165	48	1	2	137	213	112	24.2
49	1	59	159	51	0	2	139	230	152	23.3
50	3	53	152	53	1	2	97	134	116	22.9
51	5	71	173	52	0	2	169	181	118	23.7
52	2	57	152	49	0	1	160	234	128	24.7
53	2	73	165	50	1	1	123	161	116	26.8
54	3	75	170	49	0	2	130	289	134	26.0
55	3	80	171	50	1	2	198	186	108	27.4
56	4	49	157	53	0	1	215	298	134	19.9
57	4	65	162	52	0	1	177	211	124	24.8
58	2	82	170	56	0	2	100	189	124	28.4
59	3	55	155	52	0	2	91	164	114	22.9
60	3	61	165	58	0	1	141	219	154	22.4
61	2	50	155	54	1	2	139	287	114	20.8
62	5	58	160	56	0	1	176	179	114	22.7
63	1	55	166	50	1	2	218	216	98	20.0
64	5	59	161	47	0	2	146	224	128	22.8
65	2	68	165	53	1	1	128	212	130	25.0
66	2	60	170	53	1	2	127	230	122	20.8
67	1	77	160	47	1	1	76	231	112	30.1
68	5	60	155	52	0	1	126	185	106	25.0
69	3	70	164	54	0	1	184	180	128	26.0
70	2	70	165	46	0	1	58	205	128	25.7
71	3	77	160	58	1	1	95	219	116	30.1
72	5	86	160	53	0	2	144	286	154	33.6
73	2	67	152	49	1	2	124	261	126	29.0
74	3	77	165	53	1	1	167	221	140	28.3
75	3	75	169	57	0	2	150	194	122	26.3
76	2	70	165	52	0	2	156	248	154	25.7
77	2	70	165	49	1	1	193	216	140	25.7
78	1	71	157	53	0	1	194	195	120	28.8
79	1	55	162	49	0	2	73	217	140	21.0
80	2	59	165	53	1	2	98	186	114	21.7
81	3	64	159	50	0	2	127	218	122	25.3
82	1	66	160	54	0	1	153	173	94	25.8
83	4	59	165	60	0	2	161	221	122	21.7
84	3	68	165	57	0	1	194	206	172	25.0
85	5	58	160	52	0	1	87	215	100	22.7
86	1	57	154	65	1	1	188	176	150	24.0
87	2	60	160	65	0	2	149	240	154	23.4
88	2	53	162	62	0	1	215	234	170	20.2
89	2	61	159	62	1	2	163	190	140	24.1
90	1	66	154	62	0	1	111	204	144	27.8
91	1	61	152	67	0	2	198	256	156	26.4
92	2	52	152	66	0	2	265	296	132	22.5
93	1	59	155	62	0	2	143	223	140	24.6
94	1	63	155	62	1	1	136	225	150	26.2
95	2	61	165	63	0	2	298	217	130	22.4

(*Continued*)

Table 3.1 (*Continued*)

ID	Educational Level	Weight (kg)	Height (cm)	Age	Smoking Status	Physical Activity at Home	Blood Glucose	Serum Cholesterol	Systolic Blood Pressure	Body Mass Index (BMI)
96	2	68	155	67	0	2	173	251	118	28.3
97	1	58	170	62	0	1	148	187	162	20.1
98	3	68	160	55	0	1	110	290	128	26.6
99	5	60	159	50	0	2	188	238	130	23.7
100	2	61	160	54	1	1	208	218	208	23.8

Code for variables:
Education: 1 = none, 2 = primary, 3 = intermediate, 4 = senior high, 5 = technical school, 6 = university
Weight: in kilograms
Height: in centimeters
Smoking: 0 = no, 1 = yes
Physical activity: 1 = mostly sitting, 2 = moderate, 3 = heavy
Blood glucose: in milligrams percent
Serum cholesterol: in milligrams percent
Systolic blood pressure: in millimeters of mercury
Body mass index = weight (kg) ÷ height $(m)^2$

of what value would be the average smoking status or average physical activity level. When you use a computer statistical package such as SPSS, nominal or qualitative variables are often used as **grouping variables.** Whereas the average smoking status or average gender is of no value, the difference in systolic blood pressure between those who smoke and those who do not may be quite important. It might also be important to find the difference in serum cholesterol levels between smokers and nonsmokers. Using smoking status as the grouping variable allows you to do this particular analysis. In fact, this is the type of analysis that you may do in the comprehensive exercises in the chapter on two-sample significance testing.

Numbers may also be organized on an **ordinal scale.** Ordinals represent an ordered series of relationships. (e.g., first, second, third, and so on). They may be applied to such diverse situations as the rank order of causes of death or the standings in the Eastern Division of the American League. The five leading causes of death are, in order, heart disease, cancer, cerebrovascular disease, chronic respiratory disease, and unintentional injury. The standings in the American League East typically are New York Yankees, Boston Red Sox, Toronto Blue Jays, Baltimore Orioles, and Tampa Bay Devil Rays. In both examples, what you know is which are first, second, third, fourth, and fifth. In other words, you know the *order.* What you don't know is the *difference.* For example, what is the difference in the numbers of deaths due to heart disease versus cancer? Or how far is New York ahead of Boston? You know which is first and which is second, but you can't quantify the difference with the information provided. Ordinal-level data provide useful information, but without being able to quantify the differences, the information is limited.

The final two measurement scales are **interval** and **ratio,** also considered **quantitative variables.** These variables are measured numerically. Examples from Table 3.1 include weight, height, age, and serum cholesterol. The main distinction between interval and ratio data is whether the number 0 is a true or absolute zero, which means the data are ratio, or an artificial 0, which means the data are interval. True or absolute zero refers to the total *absence* of the characteristic being measured. An example that college students might identify with is $0. If you have $0 in your wallet, you have a total absence of money. The most widely used example of interval data is temperature. Temperature is usually measured using either the Fahrenheit or the centigrade scale. In both scales, the 0 is artificial because it does not represent the total absence of heat.

A key feature of both interval and ratio data is that any differences are measurable and meaningful. For example, if person A has $10 and person B has $20, you know that the difference is exactly $10 and that person B has twice as much money as person A. You also know that 60°F is 30° warmer than 30°F. However, because the Fahrenheit scale is an interval scale, you cannot say that 60°F is twice as warm as 30°F. You can meaningfully compute ratios for ratio data, but not for interval data. This is the only situation in which the distinction between interval and ratio data has any statistical importance. Throughout this text, the various statistical procedures that use interval data are just as effective with ratio data, and vice versa.

Quantitative variables can be further classified as **discrete** (discontinuous) or **continuous.** The number of children per household, the number of times you visit a doctor, and the number of missing teeth are discrete variables; they must always be integers—that is, whole numbers (e.g., 0, 1, 2). Variables such as age, height, and weight may take on fractional values (e.g., 37.8, 138.2, 112.9). They are referred to as continuous variables.

Statisticians often treat discrete variables as continuous variables. An example that you probably have noticed is the number of children per household. You may see a number such as 2.4 children per household. Obviously, you cannot have 0.4 child, yet this is a widely used statistic. The reason for treating discrete variables as continuous variables is that it significantly improves the accuracy or predictability of the data. If a community group such as a school is trying to estimate the number of children who will need services, how should that estimate be made? Suppose a community anticipated that it would have 100 new households in the next 5 years. If the number of children per household is treated strictly as a discrete variable, then the average number of children per household would be 2, and the estimate for 100 new households would be an increase of 200 children. Treating the discrete variable as a continuous variable (2.4 children per household) would yield an estimate of 240 children. In all likelihood, 240 would be the more accurate estimate.

Different types of variables are analyzed differently. Know what type of data you have. This will help you to select the appropriate method of analyis.

Hints for Entering Data into a Spreadsheet

Should you have to enter data into a spreadsheet, here are some hints that may help. If you have to manually enter data, as professors do with grades, remember to verify the accuracy of your data input. Any statistical program will correctly analyze the data provided. If incorrect numbers are entered, the analysis will be mathematically correct, but there will be the infamous "computer error." For example, if you get the wrong grade in the class and the professor claims a computer error, what he or she is really saying is, "I entered the wrong numbers." The computer will correctly compute your grade based on the numbers entered. Throughout this course, you will be either working exercises from this text or using examples from your professor. The most likely reason for incorrect numerical answers will be that you entered one or more wrong numbers. We cannot overemphasize the importance of checking the accuracy of your data. Remember: *Verify, verify, verify.*

Typically, the default setting on a spreadsheet, such as in SPSS, is two decimal places. If you are entering nominal or ordinal data, you may wish to change the setting to zero decimal places. In the spreadsheet that was created from the data in Table 3.1, educational level, physical activity, and smoking status were all changed to zero decimal places. Those three variables are most likely used as grouping variables, not for computations such as average (mean).

Finally, you need to think about subject identification (ID) numbers. If you are doing a study in which the subjects have their own ID numbers, you will need to enter them into a column, probably the first one. You will also want to reset this column to zero decimals. If the subject ID number is simply the order in which the data are entered, you do not actually have to enter ID numbers. The spreadsheet will assign each subject a number that appears to the left of the first column. This is what we did with the Table 3.1 spreadsheet: The numbers 1–100 represent the order in which subjects were entered.

3.2 FIGURES, TABLES, AND GRAPHS

You have undoubtedly seen many examples of figures, tables, and graphs. In this section, we will briefly explain what they are and how they are used. To define and distinguish each of these terms, we draw on the *Publication Manual of the American Psychological Association* (APA), Fifth Edition.

Figures are simply "any type of illustration other than a table" (APA, 2001: 176). Types of figures include charts, graphs, photographs, and drawings. In this chapter, we give examples only of graphs. Because a **graph** is one particular type of figure, the correct APA format is to label a graph as a figure. Typically, a **table** is used to display quantitative data. For example, Table 3.1 displays raw data—that is, data that have not been transformed or analyzed. Again, notice that qualitative data such as smoking status have been assigned numerical values (0 = nonsmoker, 1 = smoker).

The primary purpose of graphs and tables is to visually display information in a manner that makes it easy for readers to comprehend. If the information is quantitative, and so presented as a table, does that table adequately display the data in a way that is easily understood? Table 3.1 is an example of a table that is easy to follow. The title clearly indicates the content of the table, the column heads are short but descriptive, and explanatory notes are given at the bottom. The same basic principles also apply to graphs. A well-done graph or table will enhance the accompanying explanation.

3.3 FREQUENCY TABLES

In this section we will discuss how to create tables using the SPSS statistical program. Any statistical program allows for ease of table construction, but not all computer-generated tables adequately display data. Remember that the primary purpose of a table is to provide a visual representation that makes the data clear and understandable. Perhaps the most convenient way of summarizing or displaying data is by means of a **frequency table.** Table 3.2 is an example of a simple frequency table constructed using the systolic blood pressure readings from the Honolulu Heart Study sample. This table was constructed using SPSS 11.5.

Creating Table 3.2 was a very simple procedure. In the menu bar, we chose Analyze. The second item in the Analyze scroll bar is Descriptive Statistics. From Descriptive Statistics, we selected Frequencies and then chose the column labeled SysBP (for systolic blood pressure). The table was then automatically created. The only further change was to retitle the table and save it.

Table 3.2 may contain some terms that are not familiar. In the first portion, you will see N. Valid, and Missing. The "N" is the number of subjects used in the analysis. In this example, we used all 100 from Table 3.1. "Valid" means we had 100 cases with data entered. Because there were no missing cases, 0 was entered after "Missing." Looking at the main table, you can see the systolic blood pressures listed in ascending order. Should you choose to list the data in descending order, you could change the settings. **Frequency** refers to the number of cases with a particular value. For example, there is one case with a systolic blood pressure of 92 and eight cases with a reading of 128. **Valid percent** is the percentage out of 100, using only those subjects with data. Because all subjects have a systolic blood pressure score listed, in this table, the valid percent is the same as the percentage. **Cumulative percent** is the percentage of all previous cases plus the current interval. To illustrate, look at the cumulative percentage for a reading of 108. It is 15.0. That means there are 15% of the cases beginning with 92 and ending with 108. There are also 4% of the cases at 112. The 4% of these cases, plus the previous 15%, yield the 19.0% you see under Cumulative Percent.

Table 3.2 Systolic Blood Pressure Frequency Distribution (SPSS)

Statistics

SysBP

N	Valid	100
	Missing	0

SysBP

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	92.00	1	1.0	1.0	1.0
	94.00	1	1.0	1.0	2.0
	96.00	1	1.0	1.0	3.0
	98.00	2	2.0	2.0	5.0
	100.00	1	1.0	1.0	6.0
	102.00	2	2.0	2.0	8.0
	104.00	2	2.0	2.0	10.0
	106.00	1	1.0	1.0	11.0
	108.00	4	4.0	4.0	15.0
	112.00	4	4.0	4.0	19.0
	114.00	4	4.0	4.0	23.0
	116.00	5	5.0	5.0	28.0
	118.00	6	6.0	6.0	34.0
	120.00	2	2.0	2.0	36.0
	122.00	6	6.0	6.0	42.0
	124.00	2	2.0	2.0	44.0
	126.00	2	2.0	2.0	46.0
	128.00	8	8.0	8.0	54.0
	130.00	4	4.0	4.0	58.0
	132.00	2	2.0	2.0	60.0
	134.00	7	7.0	7.0	67.0
	136.00	1	1.0	1.0	68.0
	138.00	2	2.0	2.0	70.0
	140.00	6	6.0	6.0	76.0
	142.00	2	2.0	2.0	78.0
	144.00	2	2.0	2.0	80.0
	146.00	2	2.0	2.0	82.0
	150.00	2	2.0	2.0	84.0
	152.00	2	2.0	2.0	86.0
	154.00	4	4.0	4.0	90.0
	156.00	1	1.0	1.0	91.0
	162.00	3	3.0	3.0	94.0
	170.00	1	1.0	1.0	95.0
	172.00	1	1.0	1.0	96.0
	176.00	1	1.0	1.0	97.0
	178.00	1	1.0	1.0	98.0
	190.00	1	1.0	1.0	99.0
	208.00	1	1.0	1.0	100.0
	Total	100	100.0	100.0	

Table 3.3 Systolic Blood Pressure Frequency Distribution with Missing Data (SPSS)

Statistics

SysBP

N	Valid	95
	Missing	5

SysBP

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	92.00	1	1.0	1.1	1.1
	94.00	1	1.0	1.0	2.1
	96.00	1	1.0	1.1	3.2
	98.00	2	2.0	2.1	5.3
	102.00	2	2.0	2.1	7.4
	104.00	2	2.0	2.1	9.5
	106.00	1	1.0	1.1	10.5
	108.00	4	4.0	4.2	14.7
	112.00	4	4.0	4.2	18.9
	114.00	4	4.0	4.2	23.2
	116.00	5	5.0	5.3	28.4
	118.00	6	6.0	6.3	34.7
	120.00	2	2.0	2.1	36.8
	122.00	6	6.0	6.3	43.2
	124.00	2	2.0	2.1	45.3
	126.00	2	2.0	2.1	47.4
	128.00	7	7.0	7.4	54.7
	130.00	3	3.0	3.2	57.9
	132.00	2	2.0	2.1	60.0
	134.00	7	7.0	7.4	67.4
	136.00	1	1.0	1.1	68.4
	138.00	2	2.0	2.1	70.5
	140.00	6	6.0	6.3	76.8
	142.00	1	1.0	1.1	77.9
	144.00	2	2.0	2.1	80.0
	146.00	2	2.0	2.1	82.1
	150.00	2	2.0	2.1	84.2
	152.00	2	2.0	2.1	86.3
	154.00	4	4.0	4.2	90.5
	156.00	1	1.0	1.1	91.6
	162.00	2	2.0	2.1	93.7
	170.00	1	1.0	1.1	94.7
	172.00	1	1.0	1.1	95.8
	176.00	1	1.0	1.1	96.8
	178.00	1	1.0	1.1	97.9
	190.00	1	1.0	1.1	98.9
	208.00	1	1.0	1.1	100.0
	Total	95	95.0	100.0	
Missing	System	5	5.0		
Total		100	100.0		

The table will look slightly different if there are missing data. To create Table 3.3, we arbitrarily deleted five scores from the 100. Note that the first portion now lists 95 as Valid and 5 as Missing. If you look at the main display, the Frequency and Percent columns look similar to Table 3.2 except that an additional row includes the five missing cases. Note that the Valid Percent column is now different from the Percent column. This is because the valid percentage is computed based on the 95 cases that have data. The Percent column still uses all 100, with 5 or 5% identified as Missing. Next, look at the Cumulative Percent column, and notice that this figure is calculated using only the valid cases. The five missing cases are not included in the cumulative percent.

Although Tables 3.2 and 3.3 are clear and relatively easy to create, they have some important shortcomings. Remember, the primary purpose in creating a table is to visually display data such that readers can more easily comprehend the data. One way to improve these tables is to develop a frequency table that includes **class intervals.** Class intervals are usually equal in length, thereby aiding the comparisons between any two intervals. The number of intervals depends on the number of observations, but in general should range from 5 to 15. With too many class intervals, the data are not sufficiently summarized for a clear visualization of how they are distributed. With too few, the data are oversummarized, and some of the details of the distribution may be lost.

In selecting 5–15 class intervals, we can take several simple steps to determine the appropriate **interval width.** The interval width is the number of units between the upper and lower limits, or **class limits.** For example, if we choose an interval of 90–99 mm, the interval width is 10. If we choose an interval of 90–109 mm, the interval width is 20. To determine an appropriate interval width, the first step is to find the **range**—the difference between the highest and lowest numbers in the data set. Using the data from Table 3.2, you can see that the highest systolic blood pressure is 208 and the lowest is 92, which gives a range of 116. Next, divide 116 by 5 to get 23.2, or 23; and divide 116 by 20 to get 7.7, or 8. What this tells us is that, if we are going to have between 5 and 15 class intervals, the interval width should be between 8 and 23.

The next question becomes: What interval width should we choose? In answering this question, remember that the purpose of a table is to visually display data in a manner that readers can more easily comprehend. Earlier, when we were defining class limits, we used examples of 90–99 mm and 90–109 mm. As you can see if you look ahead to Tables 3.4 and 3.5, interval widths of numbers like 10 and 20 are very easy to understand. The interval widths are calculated using the **class boundaries,** or true limits. Class boundaries are points that demarcate the true upper limit of one class and the true lower limit of the next. For example, the class boundary between classes 90–109 and 110–129 is 109.5; it is the upper boundary for the former and the lower boundary for the latter. The reason the interval width for 90–109 is 20 and not 19 is that the interval width is calculated using the class boundaries. And because the class boundaries

Table 3.4 Systolic Blood Pressures of Subjects from Table 3.1

Class Interval (systolic blood pressure, in mm mercury)	Frequency	Percent	Valid Percent	Cumulative Percent
90–99	5	5.0	5.0	5.0
100–109	10	10.0	10.0	15.0
110–119	19	19.0	19.0	34.0
120–129	20	20.0	20.0	54.0
130–139	16	16.0	16.0	70.0
140–149	12	12.0	12.0	82.0
150–159	9	9.0	9.0	91.0
160–169	3	3.0	3.0	94.0
170–179	4	4.0	4.0	98.0
180–189	0	0.0	0.0	98.0
190–199	1	1.0	1.0	99.0
200–209	1	1.0	1.0	100.0

SOURCE: Data from Honolulu Heart Study.

Table 3.5 Systolic Blood Pressures of Subjects from Table 3.1

Class Interval (systolic blood pressure, in mm mercury)	Frequency	Valid Percent	Cumulative Percent
90–109	15	15.0	15.0
110–129	39	39.0	54.0
130–149	28	28.0	82.0
150–169	12	12.0	94.0
170–189	4	4.0	98.0
190–209	1	1.0	100.0

SOURCE: Data from Honolulu Heart Study.

are 89.5–109.5, the interval width is 20, not 19. When choosing an interval width, a good general rule is to use whole numbers and, when possible, multiples of 5. We are much more accustomed to viewing and counting by 5 than by, say, 4 or 6. Tables 3.4 and 3.5 show completed frequency tables.

3.4 GRAPHING DATA

The second way of displaying data is by use of graphs. Graphs give users a nice overview of the essential features of the data. Although such visual aids are even easier to read than tables, they often do not give the same detail.

Graphs are designed to help users obtain at a glance an intuitive feeling for the data. So it is essential that each graph be self-explanatory—that is, have a

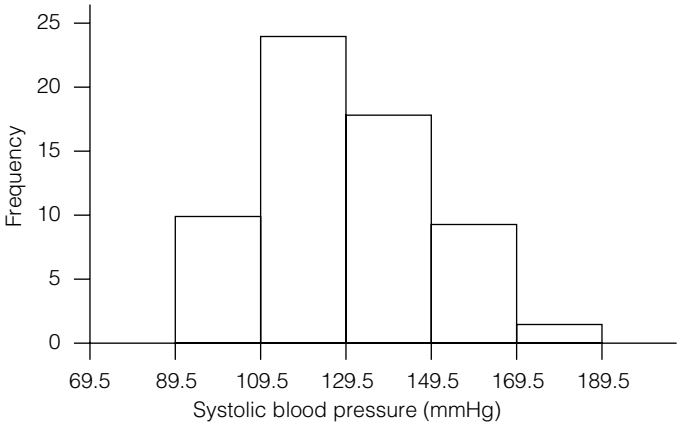


Figure 3.1 Histogram Illustrating the Data of Table 3.5: Systolic Blood Pressure of a Sample of 63 Nonsmokers from the Honolulu Heart Study.

descriptive title, labeled axes, and an indication of the units of observation. An effective graph is simple and clean. It should not attempt to present so much information that it is difficult to comprehend. Seven graphs will be discussed here: histograms, frequency polygons, cumulative frequency polygons, stem-and-leaf displays, bar charts, pie charts, and box-and-whisker plots.

Graphing by hand is often an arduous undertaking, especially for data sets that have more than a few dozen values. The graphs discussed here and other visual representations of data can be easily generated by the computer programs listed at the end of this chapter.

Histograms

Perhaps the most common graph is the **histogram.** A histogram is nothing more than a pictorial representation of the frequency table. It consists of an **abscissa** (horizontal axis), which depicts the class boundaries (not limits), and a perpendicular **ordinate** (vertical axis), which depicts the frequency (or relative frequency) of observations. The vertical scale should begin at zero. A general rule in laying out the two scales is to make the height of the vertical scale equal to approximately three-fourths the length of the horizontal scale. Once the scales have been laid out, a vertical bar is constructed above each class interval equal in height to its class frequency. For our Honolulu Heart Study example, the bar over the first class interval is 10 units high (see Figure 3.1).

Frequencies are represented not only by height but also by the area of each bar. The total area represents 100%. From Figure 3.1, it is possible to measure that 16% of the area corresponds to the 10 scores in the class interval 89.5–109.5 and that 38% of the area corresponds to the 24 observations in the second bar. Because

Table 3.6 Household Income, 1989

Income ($)	Number of Households	Relative Frequency (%)
0–4,999	6,320,400	6.9
5,000–9,999	10,534,000	11.5
10,000–14,999	9,709,600	10.6
15,000–19,999	9,100,000	10.0
20,000–24,999	8,427,200	9.2
25,000–34,999	14,747,600	16.1
35,000–49,999	15,755,200	17.2
50,000–74,999	16,488,000	18.0
75,000 and over	458,000	0.5
Total	91,540,000	100.0

area is proportional to the number of observations, be especially careful when constructing histograms from frequency tables that have unequal class intervals. How this is done is illustrated with the income data shown in Table 3.6.

From Table 3.6 we can see that the first five class intervals are measured in $5000 units while the next two intervals are $10,000 (i.e., two $5000 units) and $15,000 (i.e., three $5000 units), respectively. Because area is an indication of frequency in a histogram, we have to allocate the appropriate amount of area to each bar. The heights of the first five class intervals are their respective relative frequencies—that is, 6.9, 11.5, and so on. The height for the other intervals is obtained using the following formula:

$$\text{Height} = \frac{\text{relative frequency}}{\text{interval width}}$$

The height for the sixth interval is 8.05 (= 16.1/2) and for the seventh, 5.7 (= 17.2/3). For the $50,000–$75,000 interval, the interval will be five times wider than for the $5000 interval [(75,000 − 50,000)/5000 = 5]. Consequently, the height for the last interval will be 3.6 (= 18.0/5).

Using these heights, we can now draw the histogram, as shown in Figure 3.2. From Figure 3.2, we can see that the percent frequencies of households decrease as income increases. Furthermore, we can say that there is a higher percentage of households with low income than with high income. Figure 3.2 is, however, not optimal. Whenever possible, keep the widths of the intervals equal to prevent misleading or awkward-looking histograms.

Frequency Polygons

A second commonly used graph is the **frequency polygon,** which uses the same axes as the histogram. It is constructed by marking a point (at the same height as the histogram's bar) at the **midpoint** of the class interval. These points

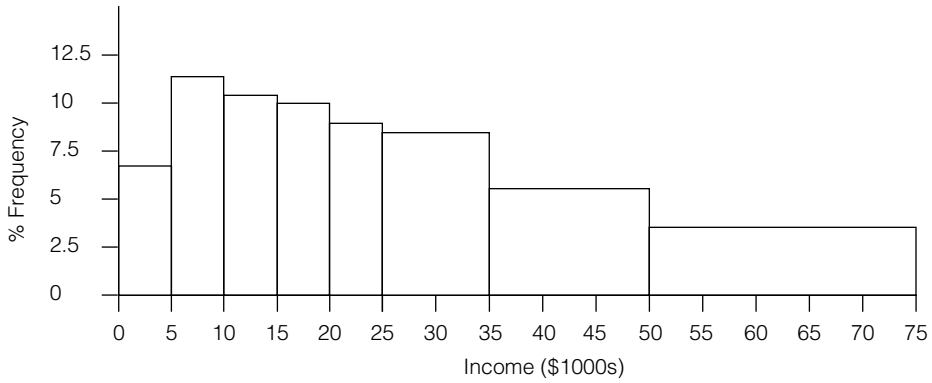


Figure 3.2 Histogram of U.S. Household Income, 1989. (NOTE: The 0.5% of households with income $75,000 and over is not shown.)

are then connected. At the ends, the points are connected to the midpoints of the previous (and succeeding) intervals of zero frequency (see Figure 3.3). Frequency polygons, especially when superimposed, are superior to histograms in providing a means of comparing two frequency distributions. In frequency polygons, the frequency of observations in a given class interval is represented by the area contained beneath the line segment and within the class interval. Frequency polygons should be used to graph only quantitative (numerical) data. Quantitative data have a continuous distribution. Frequency polygons should not be used for qualitative (i.e., nominal or ordinal) data. Qualitative data have an underlying discrete or discontinuous distribution.

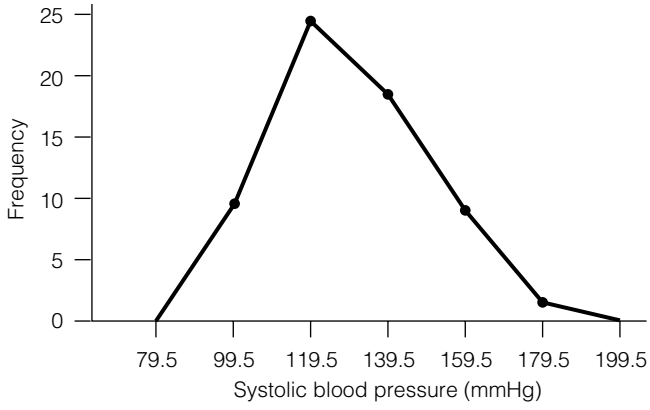


Figure 3.3 Frequency Polygon Illustrating the Data of Table 3.5: Systolic Blood Pressure of a Sample of 63 Nonsmokers from the Honolulu Heart Study.

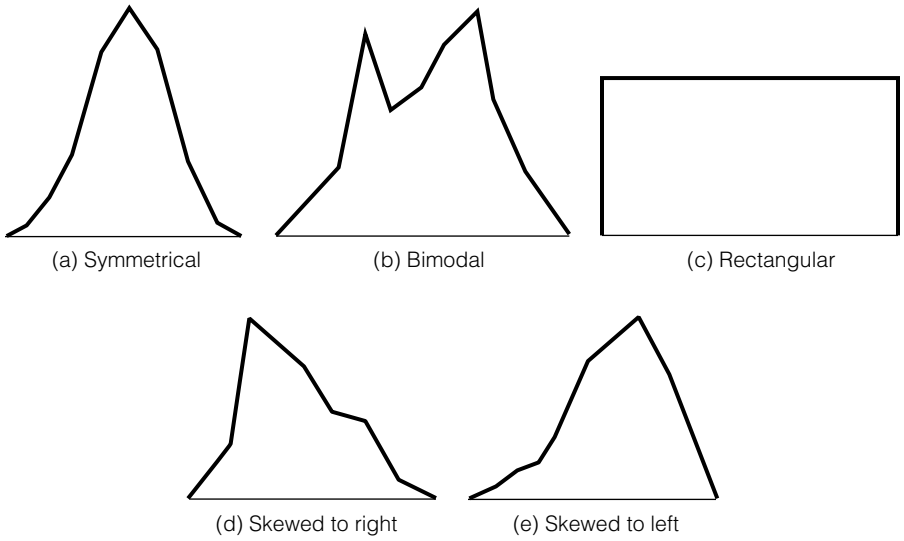


Figure 3.4 Various Shapes of Frequency Polygons.

Frequency polygons may take on a number of different shapes. Some of those most commonly encountered are shown in Figure 3.4. Part (a) of the figure is the classic "bell-shaped" **symmetrical distribution.** Part (b) is a **bimodal** (having two peaks) **distribution** that could represent an overlapping group of males and females. Part (c) is a **rectangular distribution** in which each class interval is equally represented. Parts (a) and (c) are symmetrical, whereas parts (d) and (e) are skewed or asymmetrical. The frequency polygon of part (d) is positively **skewed** because it tapers off in the positive (right-hand) direction, and part (e) is negatively skewed.

Cumulative Frequency Polygons

At times, it is useful to construct a **cumulative frequency polygon,** also called an **ogive,** which is a third type of graph. Although the horizontal scale is the same as that used for a histogram, the vertical scale indicates cumulative frequency or cumulative relative frequency. To construct the ogive, we place a point at the upper class boundary of each class interval. Each point represents the cumulative relative frequency for that class. Note that not until the upper class boundary has been reached have all the data of a class interval been accumulated. The ogive is completed by connecting the points (see Figure 3.5). Ogives are useful in comparing two sets of data—for example, data on healthy and diseased individuals. In Figure 3.5, we can see that 90% of the nonsmokers and 86% of the smokers had systolic blood pressures below 160 mmHg. The ogive gives for each interval the cumulative relative frequency—that is,

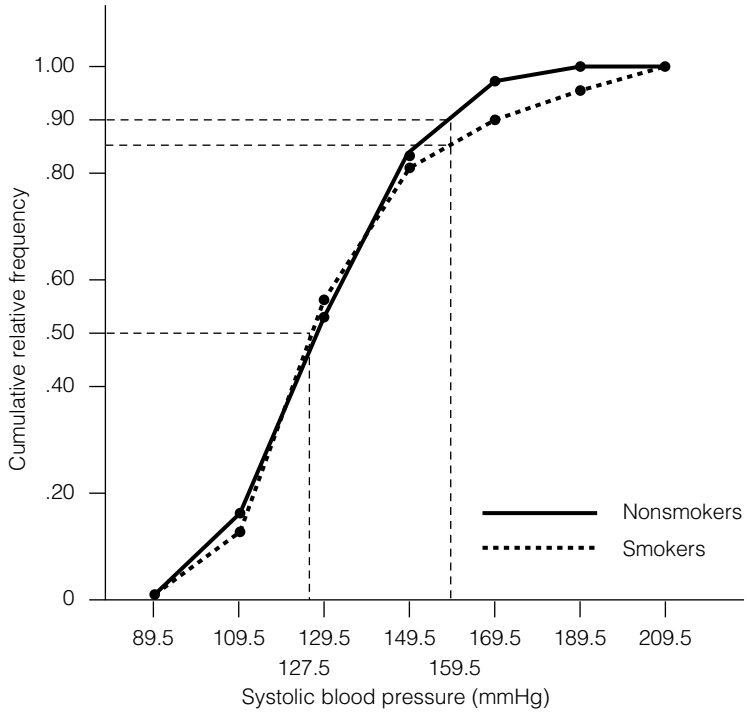


Figure 3.5 Ogives Illustrating the Data on Systolic Blood Pressure of a Sample of 63 Nonsmokers and 37 Smokers from the Honolulu Heart Study.

the percentage of cases having systolic blood pressures in that interval or a lower one.

Percentiles may be obtained from an ogive. The 90th percentile is that observation that exceeds 90% of the set of observations and is exceeded by only 10% of them. Percentiles are readily obtained, as in Figure 3.5. In our example, the 50th percentile, or median, for nonsmokers is a blood pressure of 127.5 mmHg, and the 90th percentile for nonsmokers is 159.5 mmHg.

Stem-and-Leaf Displays

Tukey (1977) suggested an innovative technique for summarizing data that utilizes characteristics of the frequency distribution and the histogram. It is referred to as the **stem-and-leaf display;** in this technique, the "stems" represent the class intervals and the "leaves" are the strings of values within each class interval. Table 3.7 illustrates the usefulness of this technique in helping you develop a better feel for your data. The table is a stem-and-leaf display that utilizes the observations of systolic blood pressures of the 63 nonsmokers of Table 3.1.

Table 3.7 Stem-and-Leaf Display of Data from Table 3.1: Systolic Blood Pressure of 63 Nonsmokers

Stems (intervals)	Leaves (observations)	Frequency (f)
90–99	2 4 6 8	4
100–109	0 4 6 8 8 8	6
110–119	2 2 4 4 8 8 8 8 8	9
120–129	0 2 2 2 2 4 4 8 8 8 8 8 8 8 8	15
130–139	0 0 0 2 2 4 4 4 4 4 4 8	12
140–149	0 0 2 4 4 6	6
150–159	2 2 4 4 4 4 6	7
160–169	2 2	2
170–179	0 2	2
180–189		0
Total		63

For each stem (interval), we arrange the last digits of the observations from the lowest to the highest. This arrangement is referred to as the leaf. The leaves (strings of observations) portray a histogram laid on its side; each leaf reflects the values of the observations, from which it is easy to note their size and frequencies. Consequently, we have displayed all observations and provided a visual description of the shape of the distribution. It is often useful to present the stem-and-leaf display together with a conventional frequency distribution. From the stem-and-leaf display of the systolic blood pressure data, we can see that the range of measurements is 92 to 172. The measurements in the 120s occur most frequently, with 128 being the most frequent. We can also see which measurements are not represented.

Bar Charts

The **bar chart** is a convenient graphical device that is particularly useful for displaying nominal or ordinal data—data like ethnicity, gender, and treatment category. The various categories are represented along the horizontal axis. They may be arranged alphabetically, by frequency within a category, or on some other rational basis. We often arrange bar charts according to frequency, beginning with the least frequent and ending with the most frequent. The height of each bar is equal to the frequency of items for that category. To prevent any impression of continuity, it is important that all the bars be of equal width and separate, as in Figure 3.6.

Note that in a bar chart relative frequencies are shown by *heights,* but in a histogram relative frequencies are shown by the *areas* within the bars.

To avoid misleading readers, it is essential that the scale on the vertical axis begin at zero. If that is impractical, you should employ broken bars (or a similar device), as shown in Figure 3.7. Here is an example of what can happen if neither procedure is followed. The public relations department of a West Coast

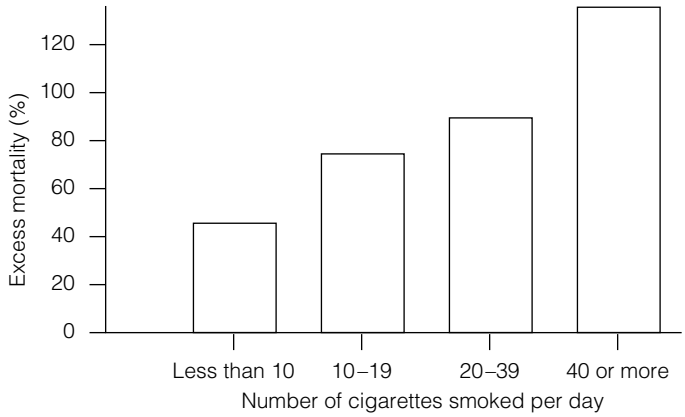


Figure 3.6 Bar Chart of Excess Mortality of Smokers over Nonsmokers According to Number of Cigarettes Smoked. SOURCE: Hammond (1966).

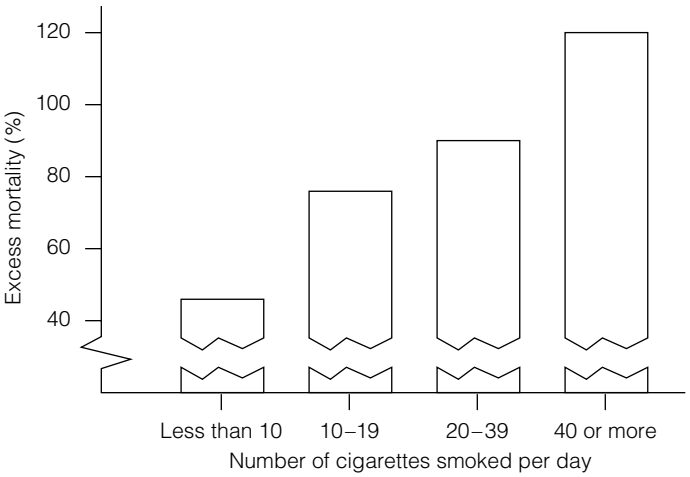
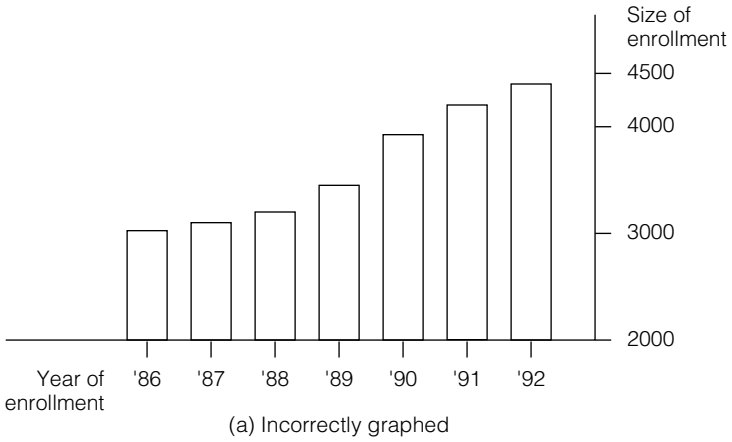
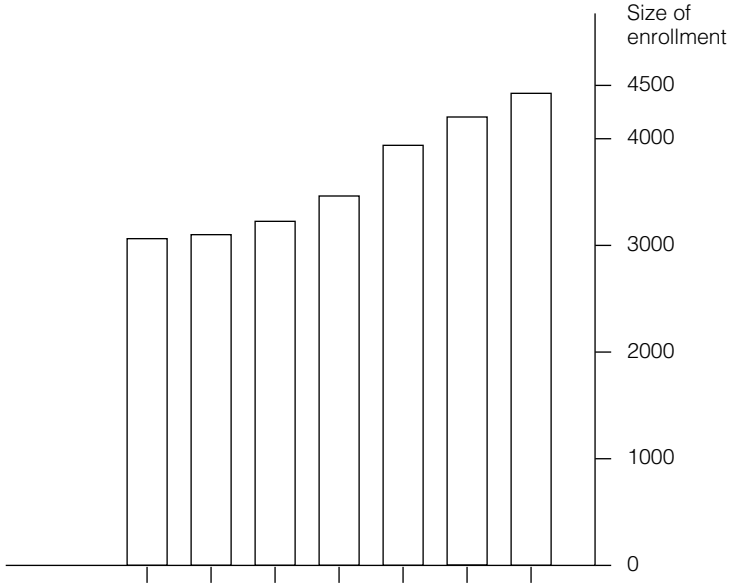


Figure 3.7 Bars Broken to Show Vertical Scale Does Not Begin at Zero. SOURCE: Hammond (1966).

college circulated the graph shown in Figure 3.8a. It gives the clear impression that enrollment doubled between 1986 and 1992. The reason for this is that the bars begin not at zero but at 2000. Persons unskilled in interpreting graphical data may find themselves drawn into one of the many pitfalls that are so well documented in books on the misuse of statistics. Figure 3.8b illustrates the correct way of presenting the same enrollment statistics. This graph makes clear that the enrollment increased by only about 50% over the 7 years.



(a) Incorrectly graphed



(b) Correctly graphed

Figure 3.8 Size of Enrollment of a West Coast College, 1986–1992.

Pie Charts

A common device for displaying data arranged in categories is the **pie chart** — a circle divided into wedges that correspond to the percentage frequencies of the distribution (see Figure 3.9). Pie charts are useful in conveying data that consist of a small number of categories. Because the area is proportional to the frequency, pie charts are rarely generated by hand.

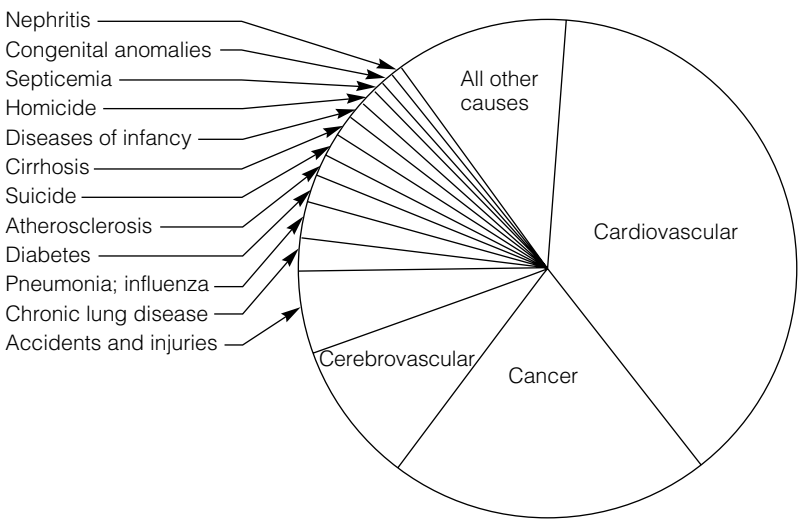


Figure 3.9 Pie Chart of Leading Causes of Death in the United States, 1987. SOURCE: National Center for Health Statistics (1990).

Box-and-Whisker Plots

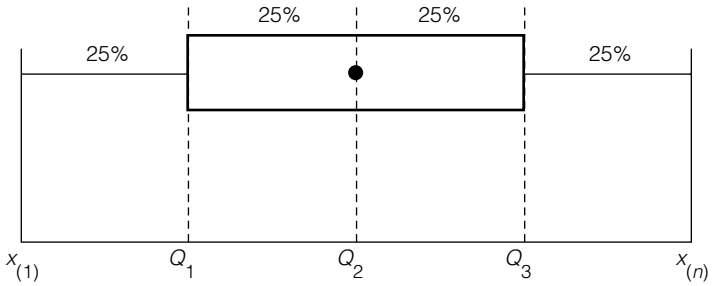
At times we may wish to graphically examine data such as long-distance telephone charges for different cities to get an idea about the typical customer and the range of the billings. We can do this by using a **box-and-whisker plot.** To do so, we need to determine the median and the quartile statistics.

The **median** is the score that divides a ranked series of scores into two equal halves. If there is an equal number of scores, you will need to obtain the average (mean) of the two middle scores. Half of the scores in each sample are less than the median, and half are larger than the median. To determine the quartiles, we need to first locate the median in the ordered list of observations. The first quartile is then the median of the observations below this median, and the third quartile is the median of the observations above the original median.

In Figure 3.10, we see that we use only five values to summarize the data: the two extremes and the three quartiles. Even with such a considerable condensation, the plot provides interesting information about the sample. The two ends of the box show the range within which the middle 50% of all the measurements lie. The median is the center dot of the sample data, and the ends of the whiskers show the spread of the data.

Computerized Graphing

The graphs presented in this chapter are easily generated by a variety of statistical programs. Each program has slightly different rules and default settings. Standard programs such as those found at www.minitab.com, www.JMP.com,



	For sample A the data are:	For sample B the data are:
x_1 =	26	15
Q_1 =	29	17
Median =	35	20
Q_2 =	39.5	25
x_n =	41	30

Figure 3.10 Summary of Telephone Charge Data Using a Box-and-Whisker Plot.

and www.spss.com can produce most of these visual representations quite rapidly.

A spreadsheet program such as Microsoft Excel may also be used to generate simple histograms, bar charts, and pie charts, as well as to quickly generate the necessary statistics (quartiles) that are required for box-and-whisker plots. Scientific calculators such as the TI-83 or TI-84 may be used to generate box-and-whisker plots, histograms, and frequency polygons.

Additionally, there is a considerable amount of free software on the Web that can be adapted to generate these types of graphs. Because of the innate fluidity of the Web, we suggest the use of a search engine such as Google to identify numerous freeware sites that can quickly generate the graphs discussed earlier. There are also sites that list freeware sites, such as www.statsci.org/free.html and www.statistics.com.

◆ CONCLUSION

The principles of tabulating and graphing data are essential if we are to understand and evaluate the flood of data with which we are bombarded. By proper use of these principles, statisticians can present data accurately and lucidly. It is also important to know which method of presentation to choose for each specific type of data. Tables are usually comprehensive, but they do not convey the information as quickly or as impressively as do graphs. Remember that graphs and tables must tell their own story and stand on their own.

They should be complete in themselves and require little (if any) explanation in the text.

◆ VOCABULARY LIST

abscissa
bar chart
bimodal distribution
box-and-whisker plot
class boundaries
class interval
class limits
continuous variable
cumulative frequency polygon (ogive)
cumulative percentage
discrete variable
figure
frequency
frequency polygon
frequency table
graph
grouping variable
histogram
interval scale
interval width
median
midpoint (class midpoint)
nominal scale
ordinal scale
ordinate
percentile
pie chart
qualitative variable
quantitative variable
range
ratio scale
rectangular distribution
skewed distribution
stem-and-leaf display
symmetrical distribution
table
valid percentage

◆ EXERCISES

3.1 Using the data from Table 3.1, construct a simple frequency table for
a. serum cholesterol
b. weight
c. blood glucose

3.2 Name the variables represented in Table 3.1, and state which type each is.

3.3 State the principal difference between a negatively skewed distribution and a positively skewed one.

3.4 a. From the 83 observations of diastolic blood pressure in Table 2.1, prepare a frequency table that includes class interval and frequency.
b. Using the same sheet of graph paper, draw a histogram and a frequency polygon for the same data.
c. Construct an ogive for the same data.
d. Find the following percentiles from the ogive: 20th, 50th (median), and 90th.
e. What percentage of the observations are less than 70? 80? 90?

3.5 For the serum cholesterol values of Table 3.1, perform the same operations as suggested in (a) and (b) of Exercise 3.4. Do this by activity status; that is, for those who reported their physical activity as mostly sitting (code 1) or moderate (code 2), make separate frequency tables, histograms, and frequency polygons for the serum cholesterol values.

3.6 Make a bar graph of the educational levels of Table 3.1.

3.7 With each of the variables listed here, two graphical methods are mentioned. Indicate which method is more appropriate. State why one method is more appropriate than the other.
a. number of dental cavities per person: pie chart, bar graph
b. triglyceride level: frequency polygon, bar graph
c. occupational classification: pie chart, histogram
d. birthrate by year: line graph, histogram

3.8 Prepare a stem-and-leaf display for the weights listed in Table 3.1.
a. Which are the smallest and the largest weights?
b. Which is the most frequent weight?

3.9 a. Prepare a stem-and-leaf display for the systolic blood pressure measurements of smokers in Table 3.1. Use the same stems as in Table 3.7, but put the leaves on the left side of the stem.
b. Combine the stem-and-leaf displays of Exercise 3.9a and Table 3.7 into a back-to-back stem display, and compare the two distributions.

3.10 Prepare a stem-and-leaf display for the heights listed in Table 3.1.
a. Which is the smallest and which is the largest height?
b. Which is the most frequent height?

3.11 For the weight data in Table 3.1, do the following:
a. Construct separate frequency tables for smokers and for nonsmokers. Use six equal class intervals beginning with 45.
b. Construct a histogram for each group.
c. Construct a frequency polygon for each group on the same graph.
d. Compare and discuss the differences in the frequency distributions between smoker and nonsmoker weights.
e. Construct an ogive for each group. Estimate the 50th percentile, and compare them for the two groups.

3.12 Construct a bar chart of educational level using the data in Table 3.1 for
a. smokers
b. nonsmokers
Compare the two bar charts and comment.

3.13 For the serum cholesterol data in Table 3.1, use equal class intervals of 30, beginning with 130, to construct
a. a separate frequency table for each of the two subgroups classified on physical activity
b. a histogram for each subgroup
c. a frequency polygon for each of the three groups. Compare and discuss the differences in the three frequency polygons.
d. an ogive for each group. Estimate the 50th percentile and compare the three.

3.14 Prepare a pie chart of the educational level for the entire sample listed in Table 3.1.

3.15 a. Using the income data from Table 3.6, combine the first two and also the third and fourth class intervals, and prepare a histogram similar to Figure 3.2.
b. Compare your histogram with that of Figure 3.2, and describe your findings.

3.16 The following are weight losses (in pounds) of 25 individuals who enrolled in a 5-week weight-control program:

9	7	10	11	10	2	3	11	5
4	8	10	9	12	5	4	11	
8	3	6	9	7	4	8	9	

a. Construct a frequency table with these six class intervals: 2–3, 4–5, 6–7, 8–9, 10–11, 12–13.
b. Construct a histogram of the weight losses.
c. Construct a frequency polygon and describe the shape of the frequency distribution.
d. What might be a possible interpretation of the particular shape of this distribution?
e. What was the most common weight loss?

3.17 Compare the two frequency distributions that you constructed in Exercise 3.13, and describe them with regard to symmetry, skewness, and modality (most frequently occurring observation).

3.18 Classify the following data as either nominal, ordinal, interval, or ratio.
a. names of students in this class
b. the number of students in this class
c. your ten favorite songs
d. height
e. heads and tails on a coin

3.19 Briefly explain why discrete (discontinuous) variables are treated as continuous variables. Use an example as part of your explanation.

3.20 For the grouped frequency distribution 70–79, 80–89, and 90–99, answer the following questions:
a. What is the class interval?
b. What are the class boundaries for the interval 80–89?

3.21 Determine the median and quartiles necessary to construct a box-and-whisker plot for the following sets of data:
a. 3, 4, 7, 5, 4, 6, 4, 5, 8, 3, 4, 5, 6, 5, 4
b. 18, 14, 17, 22, 16, 26, 33, 27, 35, 28, 44, 40, 31, 53, 70, 73, 62, 74, 93, 103, 75, 86, 84, 90, 79, 99, 73

3.22 Construct the box-and-whisker plots for the data in Exercise 3.21a and b.

3.23 Using the following data from the FBI Uniform Crime Reports, construct a pie chart indicating the weapons used in committing these murders.

11,381 committed with firearms

3,957 committed with personal weapons such as hands or feet

1,099 committed with knives

16,437 all murders

3.24 Construct a box plot for the sample of $n = 100$ systolic blood pressure readings listed in Table 3.1 separately for smokers and nonsmokers, and provide a written comparison of the two groups based on the box plots.

3.25 Construct a box-and-whisker plot for the weight loss data given in Exercise 3.16.