AN INTRODUCTION TO STATISTICS

WITH

DATA ANALYSIS

by

SHELLEY RASMUSSEN

Department of Mathematical Sciences
Olney 428T
University of Massachusetts/Lowell
Lowell, MA 01854

Originally published by Brooks/Cole Publishing Company,
Division of Wadsworth, Inc.

ISBN 0-534-13578-1

Shelley_Rasmussen@uml.edu

Minitab is a statistical package, a computer program that performs many statistical procedures. The versions of Minitab now available for use on personal computers are menu-driven and much easier to use than the main-frame version originally discussed in this text. Those sections are not included in this online edition of the text. At this time, the most recent version is Minitab 15, available at very reasonable prices for purchase or rental from:

www.e-academy.com/minitab

System Requirements

Processor:	PC with a 1 GHz 32- or 64-bit processor
Memory:	512 MB or more of available RAM
Disk Space:	125 MB free space available
Operating System:	Microsoft Windows 2000, XP, or Vista.
Display:	A display capable of 1024 X 768 or higher resolution
Software:	Adobe Acrobat Reader 5.0 or higher for Meet Minitab

CHAPTER 2

Studying One Variable at a Time: Lists, Tables, and Plots

IN THIS CHAPTER

Data list

Ordered (ranked) data

Dot plot

Stem-and-leaf plot

Frequency table

Frequency plot, histogram

Quantiles

Box plot, box graph

Symmetrical, positively skewed, and negatively skewed distributions

Unimodal, bimodal, and multimodal distributions

What can we learn about per capita calorie supplies in the World Bank countries? How do life expectancies vary across these countries? Can we identify different groups of countries based on birth rates? What is a good way to summarize primary school enrollments?

All these questions are about individual variables. We consider ways to study individual variables in Chapters 2 and 3. Chapter 3 discusses numbers, called descriptive statistics, used to summarize location and spread in a set of values. Chapter 2 looks at lists, tables, and plots. We use these tools to learn about the distribution of a variable—that is, how the values of the variable are distributed along the number line.

2-1 Lists and Dot Plots

A *list* can be useful if the number of cases is not too large: we simply list all of the values of a variable. We will see many data lists in the examples and exercises throughout this book.

A **data list** is a listing of the values of a variable in a data set.

Consider the life expectancies listed in Table 2-1 for each of the World Bank countries. This alphabetical list is useful for looking up the life expectancy for a country. Three countries (Afghanistan, Kampuchea, and Lebanon) have no 1985 life expectancy information available; these three countries must be excluded from analyses involving the life expectancies.

When looking at Table 2-1, you may find yourself wondering which countries have the longest and shortest life expectancies. Another (perhaps more interesting) way to look at the life expectancies is to order their values from smallest to largest, as in Table 2-2. We say the observations are *sorted* or *ordered* by life expectancy. The values of the ordered variables in a sorted list may go from smallest to largest or from largest to smallest. Either way, extreme values are readily apparent, and we can scan intermediate values more easily than in an unordered list. We see from Table 2-2 that the shortest life expectancy is 40 years, for Guinea and Sierra Leone. Australia and France share the longest life expectancy, 78 years. Afghanistan, Kampuchea, and Lebanon are shown at the bottom of the table as having no 1985 life expectancy information available.

Values are **ordered, sorted,** or **ranked** if they are listed in order of magnitude.

Suppose we just want to get a feel for life expectancies; for the moment we do not care which life expectancies go with which countries. Then the ordered list of life expectancies in Table 2-2 is much more useful than the scrambled list of numbers in Table 2-1. Creating such an ordered list of values is often the first step in data analysis.

Can we do better than simply listing the values in order of magnitude? We may be able to interpret a plot or graph of some kind more easily than a

TABLE 2-1 Life expectancy at birth in 1985 for 128 countries. Countries are listed alphabetically.

Country	Life expectancy at birth (years)	Country	Life expectancy at birth (years)	Country	Life expectancy at birth (years)
Afghanistan	Missing	Guatemala	60	Oman	54
Albania	70	Guinea	40	Pakistan	51
Algeria	61	Haiti	54	Panama	72
Angola	44	Honduras	62	Papua New Guinea	52
Argentina	70	Hong Kong	76	Paraguay	66
Australia	78	Hungary	71	Peru	59
Austria	74	India	56	Philippines	63
Bangladesh	51	Indonesia	55	Poland	72
Belgium	75	Iran	60	Portugal	74
Benin	49	Iraq	61	Romania	72
Bhutan	44	Ireland	74	Rwanda	48
Bolivia	53	Israel	75	Saudi Arabia	62
Botswana	57	Italy	77	Senegal	47
Brazil	65	Jamaica	73	Sierra Leone	40
Bulgaria	71	Japan	77	Singapore	73
Burkina Faso	45	Jordan	65	Somalia	46
Burma	59	Kampuchea	Missing	South Africa	55
Burundi	48	Kenya	54	Spain	77
Cameroon	55	Korea, North	68	Sri Lanka	70
Canada	76	Korea, South	69	Sudan	48
Central African Republic	49	Kuwait	72	Sweden	77
Chad	45	Lao PDR	45	Switzerland	77
Chile	70	Lebanon	Missing	Syria	64
China	69	Lesotho	54	Tanzania	52
Colombia	65	Liberia	50	Thailand	64
Congo	58	Libya	60	Togo	51
Costa Rica	74	Madagascar	52	Trinidad and Tobago	69
Côte d'Ivoire	53	Malawi	45	Tunisia	63
Cuba	77	Malaysia	68	Turkey	64
Czechoslovakia	70	Mali	46	Uganda	49
Denmark	75	Mauritania	47	USSR	70
Dominican Republic	64	Mauritius	66	United Arab Emirates	70
Ecuador	66	Mexico	67	United Kingdom	75
Egypt	61	Mongolia	63	United States	76
El Salvador	64	Morocco	59	Uruguay	72
Ethiopia	45	Mozambique	47	Venezuela	70
Finland	76	Nepal	47	Viet Nam	65
France	78	Netherlands	77	Yemen Arab Republic	45
Germany, East	59	New Zealand	74	Yemen, PDR	46
Germany, West	75	Nicaragua	59	Yugoslavia	72
Ghana	53	Niger	44	Zaire	51
Greece	68	Nigeria	50	Zambia	52
		Norway	77	Zimbabwe	57

TABLE 2-2 Life expectancy at birth in 1985 for 128 countries. Countries are ordered from smallest to largest life expectancy.

Country	expectancy at birth (years)	Country	expectancy at birth (years)	Country	expectancy at birth (years)
Guinea	40	Cameroon	55	Albania	70
Sierra Leone	40	Indonesia	55	Argentina	70
Angola	44	South Africa	55	Chile	70
Bhutan	44	India	56	Czechoslovakia	70
Niger	44	Botswana	57	Sri Lanka	70
Burkina Faso	45	Zimbabwe	57	USSR	70
Chad	45	Congo	58	United Arab Emirates	70
Ethiopia	45	Burma	59	Venezuela	70
Lao PDR	45	Germany, East	59	Bulgaria	71
Malawi	45	Morocco	59	Hungary	71
Yemen Arab Republic	45	Nicaragua	59	Kuwait	72
Mali	46	Peru	59	Panama	72
Somalia	46	Guatemala	60	Poland	72
Yemen, PDR	46	Iran	60	Romania	72
Mauritania	47	Libya	60	Uruguay	72
Mozambique	47	Algeria	61	Yugoslavia	72
Nepal	47	Egypt	61	Jamaica	73
Senegal	47	Iraq	61	Singapore	73
Burundi	48	Honduras	62	Austria	74
Rwanda	48	Saudi Arabia	62	Costa Rica	74
Sudan	48	Mongolia	63	Ireland	74
Benin	49	Philippines	63	New Zealand	74
Central African Republic	49	Tunisia	63	Portugal	74
Uganda	49	Dominican Republic	64	Belgium	75
Liberia	50	El Salvador	64	Denmark	75
Nigeria	50	Syria	64	Germany, West	75
Bangladesh	51	Thailand	64	Israel	75
Pakistan	51	Turkey	64	United Kingdom	75
Togo	51	Brazil	65	Canada	76
Zaire	51	Colombia	65	Finland	76
Madagascar	52	Jordan	65	Hong Kong	76
Papua New Guinea	52	Viet Nam	65	United States	76
Tanzania	52	Ecuador	66	Cuba	77
Zambia	52	Mauritius	66	Italy	77
Bolivia	53	Paraguay	66	Japan	77
Côte d'Ivoire	53	Mexico	67	Netherlands	77
Ghana	53	Greece	68	Norway	77
Haiti	54	Korea, North	68	Spain	77
Kenya	54	Malaysia	68	Sweden	77
Lesotho	54	China	69	Switzerland	77
Oman	54	Korea, South	69	Australia	78
		Trinidad and Tobago	69	France	78

Missing: Afghanistan, Kampuchea, Lebanon

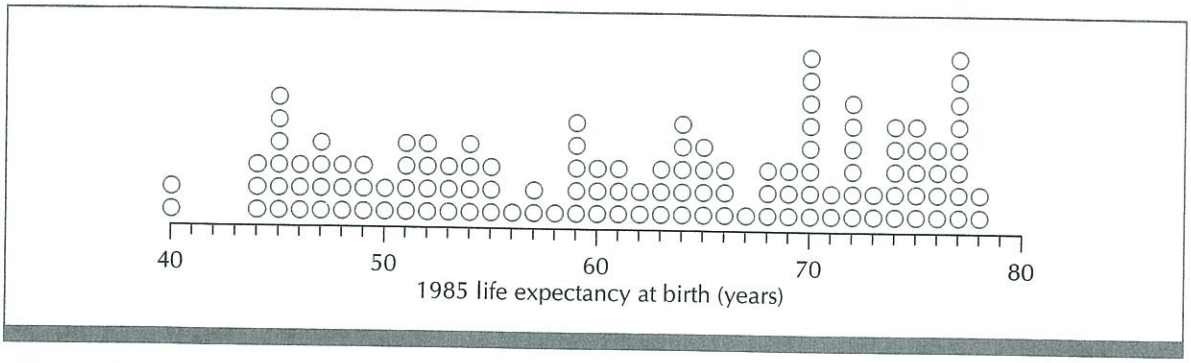

FIGURE 2-1 Dot plot of 1985 life expectancy at birth for 125 countries. Three countries are excluded because of missing values.

list of numbers. Often we can *see* more in a plot than in a list. In fact, much of modern data analysis (or exploratory data analysis) involves use of graphical techniques to examine a data set. One simple graphical technique is the dot plot.

A *dot plot* uses dots or circles to represent values, arranged along a line or axis showing the scale and units of measurement. The dot plot shown in Figure 2-1 is a picture representation of the ordered life expectancies in Table 2-2. Only the 125 countries with nonmissing values are included in the plot.

> A **dot plot** is a graphical display, with dots representing values positioned along an axis or number line.

We stack dots corresponding to cases with the same value. Eight stacked dots, for instance, indicate that eight countries have a life expectancy of 70 years. Figure 2-1 shows us that 11 countries have life expectancies of 45 years or less. There is a fairly even distribution of countries with life expectancies in the range from 46 to 69 years. A large group of countries have life expectancies from 70 to 78 years.

We will consider life expectancies again later. In Section 2-2, we use stem-and-leaf plots to look at fertility rate and calorie supply.

2-2

Stem-and-Leaf Plots

A *stem-and-leaf plot* is especially useful for getting a quick picture of a set of values when graphing by hand rather than by computer. The plot gets its name from the way it is constructed: A fixed number of the leftmost digits of a value form the stem; one or more digits to the right of the stem form the leaf.

> A **stem-and-leaf plot** uses the values taken on by a variable in a graphical display showing the frequency of values in specified intervals.

TABLE 2-3 Total fertility rate in 1985 for 128 countries.

Fertility rate	Country	Fertility rate	Country
1.3	Germany (West)	4.7	Ecuador, Egypt, Haiti
1.4	Denmark	4.9	Mongolia, Morocco, South Africa
1.5	Italy, Netherlands, Switzerland	5.2	El Salvador, Kuwait
1.6	Belgium	5.4	Papua New Guinea
1.7	Canada, Finland, Hungary, Norway, Singapore, Sweden	5.6	Central African Republic, Iran, Nicaragua
1.8	Germany (East), Hong Kong, Japan, United Kingdom, United States	5.7	Bangladesh, Chad, Guatemala
		5.8	Lesotho
		5.9	Bolivia, United Arab Emirates
2.0	Australia, Bulgaria, Cuba, France, Greece, Portugal, Spain	6.0	Guinea, Honduras, Yemen PDR
		6.1	Pakistan, Zaire
2.1	Austria, Czechoslovakia, New Zealand, Romania, Yugoslavia	6.2	Bhutan, Ethiopia, Jordan, Zimbabwe
		6.3	Algeria, Congo, Mauritania, Mozambique, Nepal
2.3	China, Poland, USSR	6.4	Angola, Ghana, Lao PDR
2.4	Korea (South)		
2.5	Chile, Mauritius	6.5	Benin, Burkina Faso, Burundi, Côte d'Ivoire, Madagascar, Mali, Sierra Leone, Togo
2.6	Ireland, Uruguay		
2.8	Jamaica, Trinidad and Tobago		
2.9	Israel		
3.2	Panama, Sri Lanka, Thailand	6.6	Sudan
		6.7	Botswana, Iraq, Oman, Senegal, Syria
3.3	Argentina, Colombia, Costa Rica		
3.4	Albania	6.8	Cameroon, Somalia, Yemen Arab Republic, Zambia
3.6	Brazil		
3.7	Malaysia		
3.8	Korea (North)	6.9	Liberia, Nigeria, Uganda
3.9	Burma, Turkey, Venezuela	7.0	Niger, Tanzania
4.0	Dominican Republic	7.1	Saudi Arabia
4.1	Indonesia	7.2	Libya
4.3	Mexico, Peru, Philippines	7.6	Malawi
		7.8	Kenya
4.4	Paraguay	8.0	Rwanda
4.5	India		
4.6	Tunisia, Viet Nam		

Missing: Afghanistan, Kampuchea, Lebanon

Let's construct a stem-and-leaf plot for the variable fertility rate. Fertility rate is an estimate of the average number of children born per woman in a country. As we can see from the listing in Table 2-3, three countries (Afghanistan, Kampuchea, and Lebanon) have no information available on fertility rate. We must exclude them from analyses involving this variable.

Since each fertility rate has just two digits, the choice of stem and leaf is easy: The number to the left of the decimal point is the stem; the number to the right of the decimal point forms the leaf. Fertility rates in Table 2-3 range from 1.3 to 8.0. Thus, the stems go from 1 to 8, as shown in Figure 2-2a.

The first fertility rate in Table 2-3 is 1.3, for West Germany. This number has a stem of 1 and a leaf of 3 and is plotted in Figure 2-2b. The next fertility rate is 1.4, for Denmark. This number has a stem of 1 and a leaf of 4, added to the plot in Figure 2-2c. By the time we add the value 2.0 for Spain, the plot looks like Figure 2-2d.

A completed stem-and-leaf plot of fertility rates is displayed in Figure 2-3. The legend of the plot explains how to interpret the stems and leaves. We see a peak in the distribution at the row for stem 6; there is a concentration of countries with fertility rates near 6. The plot shows another concentration of countries around fertility rate 2.

FIGURE 2-2 Steps in constructing a stem-and-leaf plot of 1985 total fertility rates. The stem is an integer and the leaf is a decimal value.

Stem	Leaf
1	
2	
3	
4	
5	
6	
7	
8	

a. The stems go from 1 to 8.

Stem	Leaf
1	3
2	
3	
4	
5	
6	
7	
8	

b. The value 1.3 for West Germany is plotted.

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

c. The value 1.4 for Denmark is added to the plot.

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

d. After the value 2.0 for Spain is added, the plot looks like this.

FIGURE 2-3 Stem-and-leaf plot of 1985 total fertility rates for 125 countries. The stem is an integer and the leaf is a decimal value. Three countries are excluded because of missing values.

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

The purpose of a stem-and-leaf plot is to provide insight into the distribution of a set of values. We might decide that the rows are too long in Figure 2-3, that the plot is too condensed. Variations of the stem-and-leaf plot use two or more rows of leaves for each stem, in effect stretching the plot. One such variation of the stem-and-leaf plot of fertility rates is shown in Figure 2-4, with two rows for each stem: one for leaves 0–4 and the other for leaves 5–9. We can see that there is a large group of countries with fertility rates between 1.5 and 2.5. Another large group has values between 5.5 and 7.0. Three countries have fertility rates greater than 7.5 children per woman.

Another variation on the plot in Figure 2-3 might use five rows for each stem: for leaves 0 and 1, leaves 2 and 3, leaves 4 and 5, leaves 6 and 7, leaves 8 and 9. You might try constructing this variation yourself for the fertility rate values, and compare it with the plots in Figures 2-3 and 2-4.

A stem-and-leaf plot is more condensed than a dot plot. Each stack of dots in a dot plot corresponds to a single value. Each row or stem in a stem-and-leaf plot corresponds instead to a range of values. The stem-and-leaf plot summarizes the information in a dot plot, while still showing us every data value. We determine the stems, leaves, and number of rows per stem to give a plot that is more informative than a simple list, but not as stretched out as a dot plot.

How we define stems and leaves depends in large part on the values of a variable. Figure 2-5a shows a stem-and-leaf plot of calorie supplies. Estimated daily calorie supply ranges from 1,504 to 3,831 calories per person. The stems in the plot are in hundreds of calories, ranging from 15 hundred to 38 hundred calories. Leaves are in calories. The smallest value, 1,504, has a stem of 15 and a leaf of 04. The largest value, 3,831, has a stem of 38 and a leaf of 31.

Figure 2-5b is a variation of the stem-and-leaf plot in Figure 2-5a, with single digits as leaves. To obtain a leaf for Figure 2-5b, we *cut* the rightmost (units) digit from the corresponding calorie supply value in Figure 2-5a, so leaves in Figure 2-5b are in tens of calories. The two stem-and-leaf plots convey almost the same information. However, in his book *Exploratory Data Analysis* (1977), John Tukey suggests that single-digit leaves are generally preferable, making plots easier to look at and interpret. Do you find the plot in Figure 2-5b easier to examine and interpret than the one in Figure 2-5a?

FIGURE 2-4 Stem-and-leaf plot of 1985 total fertility rates for 125 countries. The stem is an integer and the leaf is a decimal value. Each stem takes up two rows, one for leaves 0–4 and the other for leaves 5–9. Three countries are excluded because of missing values.

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

FIGURE 2-5 Stem-and-leaf plots of 1985 calorie supply per capita for 124 countries. Four countries are excluded because of missing values.

Stem	Leaf
15	04
16	78 81
17	28 37 47 88
18	17 55 99
19	19 24 69
20	34 38 50 54 54 72 78 83 89
21	16 37 46 48 51 54 59 71 73 81 89
22	11 19 28 36 40 50 50 94
23	11 35 37 41 42 58 85
24	19 25 48 61 62 69
25	05 33 47 49 71 74 83 85
26	02 02 33 77 78 84 95 98
27	26 40 71 96
28	03 07 36 41 56
29	26 47 79
30	06 26 60 97
31	22 22 28 31 38 51 61 67 68 77
32	21 39 63 80
33	43 58 59 85 86 89
34	32 32 40 65 74 82
35	14 38 47
36	02 12 25 63 63 79
37	21 91
38	31

a. The stem is in hundreds of calories; the leaf is in calories.

Stem	Leaf
15	0
16	78
17	2348
18	159
19	126
20	335557788
21	13445557788
22	11234559
23	1334458
24	124666
25	03447788
26	00377899
27	2479
28	00345
29	247
30	0269
31	2223356667
32	2368
33	455888
34	334678
35	134
36	012667
37	29
38	3

b. The stem is in hundreds of calories; the leaf is in tens of calories.

The stem-and-leaf plots in Figure 2-5 show a peak and large concentration of values near the stem for 21 hundred calories. There is another, somewhat smaller, peak and concentration of values near 31 hundred calories. The World Bank countries seem to fall into two major groups: countries with adequate per capita calorie supplies and countries with inadequate per capita calorie supplies.

With stems representing hundreds of calories in Figure 2-5, we get a plot with 24 rows or stems. Different choices for stems and leaves are possible. Figure M2-6 in the Minitab Appendix for Chapter 2 shows a stem-and-leaf plot for these calorie supplies with five rows for each stem value. The single-digit stems are in thousands of calories and the single-digit leaves are in hundreds of calories (obtained after cutting off the two rightmost digits of each calorie supply value).

The way a stem-and-leaf plot is constructed makes it easy to read the actual data values. The plot could be constructed with stems across the bottom and leaves as columns, but the numbers are not as easy to read that way.

So far we have considered lists and displays of all the values of a variable. If the number of cases is large, such lists and displays can get unwieldy. An important goal of data analysis is data reduction: We want fewer numbers to look at. A good summary provides insights into the data, highlighting features and characteristics that we might otherwise have missed. We will discuss some simple ways to summarize information about a variable in Sections 2-3 and 2-4.

2-3 Frequency Tables, Frequency Plots, and Histograms

We might like to summarize the values of a variable, without necessarily listing each value. A *frequency table* provides such a summary. If a variable takes on relatively few values, a frequency table shows the number of occurrences of each value. Otherwise, a frequency table displays the number of occurrences of values within specified intervals. We will use the first type of frequency table for number of cities of over 500,000 persons, and the second type of frequency table for primary school enrollments.

> When a variable takes on relatively few values, a **frequency table** lists the number of occurrences of each of these values.
>
> When a variable takes on relatively many values, a **frequency table** lists the number of occurrences of values over specified intervals.

The number of cities of over 500,000 persons (large cities) in 1980 is shown in Table 2-4. One way to summarize this information is to count the countries with no large cities, with one large city, and so on. Then we can display these counts in a frequency table, such as Table 2-5. Table 2-5 is based on the 123 countries with nonmissing information, the last column showing the percentage of these 123 countries with each number of large cities. Twenty-seven (22.0%) countries had no large cities; 51 (41.5%) countries had one such

TABLE 2-4 Number of cities of over 500,000 persons in 1980 for 128 countries.

Number of large cities	Countries
0	Albania, Bhutan, Burkina Faso, Burundi, Central African Republic, Chad, Congo, El Salvador, Honduras, Kuwait, Lao PDR, Lesotho, Liberia, Malawi, Mali, Mauritania, Mongolia, Nepal, Niger, Papua New Guinea, Rwanda, Sierra Leone, Somalia, Togo, Trinidad and Tobago, Yemen Arab Republic, Yemen PDR
1	Afghanistan, Algeria, Angola, Austria, Benin, Bolivia, Bulgaria, Cameroon, Canada, Chile, Costa Rica, Côte d'Ivoire, Cuba, Czechoslovakia, Denmark, Dominican Republic, Ethiopia, Finland, Guatemala, Guinea, Haiti, Hong Kong, Hungary, Ireland, Israel, Jamaica, Jordan, Kenya, Lebanon, Libya, Madagascar, Malaysia, Mozambique, New Zealand, Nicaragua, Panama, Paraguay, Portugal, Romania, Senegal, Singapore, Sri Lanka, Sudan, Switzerland, Tanzania, Thailand, Tunisia, Uganda, Uruguay, Zambia, Zimbabwe
2	Belgium, Burma, Ecuador, Egypt, Ghana, Greece, Korea (North), Peru, Philippines, Saudi Arabia, Syria, Zaire
3	Bangladesh, Germany (East), Iraq, Netherlands, Sweden, Yugoslavia
4	Colombia, Morocco, Turkey, Venezuela, Viet Nam
5	Argentina, Australia
6	France, Iran, Spain
7	Korea (South), Mexico, Pakistan, South Africa
8	Poland
9	Indonesia, Italy, Japan, Nigeria, Norway
11	Germany (West)
14	Brazil
17	United Kingdom
36	India
50	USSR
65	United States
78	China
Missing:	Botswana, Kampuchea, Mauritius, Oman, United Arab Emirates

city. Three (2.4%) countries had 50 or more cities of over 500,000 persons in 1980. Note that because of rounding, the listed percentages do not sum to exactly 100%.

A *frequency plot* is a graphical presentation of the number of occurrences of each value of a variable. The frequency plot in Figure 2-6 displays the same information as Table 2-5. The horizontal axis corresponds to values of the variable, number of large cities (the left-hand column of Table 2-5). The

TABLE 2-5 Number of cities of over 500,000 persons (large cities) in 1980, summarized for 123 countries. Five countries have missing values.

Number of large cities	Number of countries	Percentage of countries
0	27	22.0
1	51	41.5
2	12	9.8
3	6	4.9
4	5	4.1
5	2	1.6
6	3	2.4
7	4	3.3
8	1	.8
9	5	4.1
11	1	.8
14	1	.8
17	1	.8
36	1	.8
50	1	.8
65	1	.8
78	1	.8
Total	123	100.0

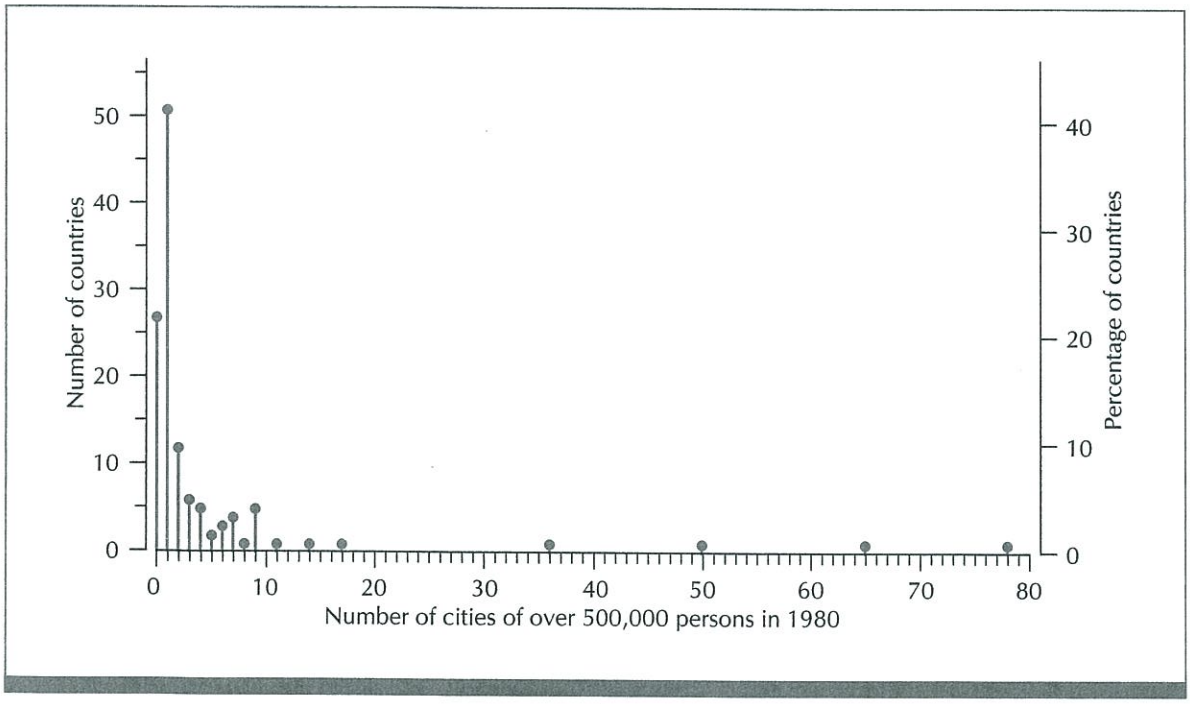

FIGURE 2-6 Frequency plot of number of cities of over 500,000 persons in 1980, summarized for 123 countries. Five countries are excluded because of missing values.

vertical axis on the left refers to the frequency or number of countries with any given value (the center column of Table 2-5). The vertical axis on the right refers to the percentage of countries with any given value (the right-hand column of Table 2-5). The heights of the vertical lines show the relative frequencies of the data values. Other names for a frequency plot are *bar chart* and *bar graph*. Frequency plots are related to histograms, to be discussed shortly.

> When a variable takes on relatively few values, a **frequency plot** is a graphical way of displaying the number of occurrences of each of these values.

Values for primary school enrollment are listed in Table 2-6. Primary school enrollment is the number of children enrolled in primary school as the percentage of 6- to 11-year-olds in the country. Some countries have children outside of this age interval in primary school, so percentages may be greater than 100. The list in Table 2-6 is ordered from smallest to largest primary school enrollment. Certainly this ordered list is easier to study than an unordered list (such as enrollments for countries listed in alphabetical order). But with so many different values, we might prefer a table that has fewer numbers to look at, a summary of the enrollment figures in Table 2-6.

How can we summarize the numbers in Table 2-6? In a summary, we may not feel the need to discriminate between individual enrollment figures that are very close, such as 76 and 77. Perhaps we can look at intervals of numbers without losing too much information. Such a grouping of values is shown in Table 2-7, a frequency table that groups values of the variable into intervals. Primary school enrollments are grouped into intervals of length 10: 20–29, 30–39, and so on. The table shows that four countries have enrollments in the range from 20 to 29, for instance. Primary school enrollments range from values in the 20's to well above 100%. A majority of countries have primary school

TABLE 2-6 Number enrolled in primary school in 1984 as percentage of 6–11-year age group, for 119 countries. Nine countries have missing information.

25	62	80	97	99	102	106	113
25	62	83	97	99	102	106	113
28	62	83	97	99	103	107	114
29	64	84	97	99	103	107	115
32	66	87	97	99	103	107	116
32	67	87	97	99	103	107	116
37	67	90	97	100	104	107	116
38	68	90	98	100	105	107	118
42	70	91	98	101	105	107	118
45	76	92	98	101	105	108	119
49	76	94	98	101	105	108	120
49	76	95	98	101	106	109	121
55	77	96	98	101	106	109	131
57	77	97	98	101	106	111	134
61	77	97	98	102	106	112	

TABLE 2-7 Number enrolled in primary school in 1984 as percentage of 6–11-year age group for 119 countries. Nine countries have missing information.

Primary school enrollment	Number of countries	Percentage of countries
20–29	4	3.4
30–39	4	3.4
40–49	4	3.4
50–59	2	1.7
60–69	9	7.6
70–79	7	5.9
80–89	6	5.0
90–99	30	25.2
100–109	37	31.1
110–119	12	10.1
120–129	2	1.7
130–139	2	1.7
Total	119	100.0

enrollments from 90% to 110%. The percentages listed in the right-hand column of Table 2-7 do not sum to 100 exactly, because of rounding.

It is clear in Table 2-7 where each value should go. A value of 29% goes in the first interval, a value of 30% goes in the second interval, and so on. However, such intervals are commonly labeled somewhat differently, as shown in Table 2-8. The intervals designated in Table 2-8 have overlapping endpoints, which can be convenient for presenting frequency information. But how does the reader interpret the endpoints of the intervals? With the labeling in Table 2-8, for instance, the reader does not know whether a value of 30% should go in the interval 20–30 or in the interval 30–40. For compiling the frequency table, it does not matter whether an endpoint is counted in the corresponding upper or lower interval, just so the counting is consistent (to ensure that the intervals all have the same length and are comparable). If 30% goes in the interval 30–40, for example, then 40% goes in the interval 40–50. The convention should be clearly noted somewhere in the legend. The legend for Table 2-8 explains that intervals include the lower endpoint, not the upper endpoint. Therefore, the intervals in Table 2-8 are the same as those in Table 2-7. Note that in Table 2-8, the missing observations are included in the tabulation. Also, the column for percentage of countries is not included in Table 2-8.

We can display the information in a frequency table such as Table 2-7 graphically via a histogram. A *histogram* displays frequency information for a variable, summarized over intervals. Figure 2-7 is a histogram displaying the same information as Table 2-7. The horizontal axis shows the scale and units for the variable. The vertical axes show frequencies on the left and percentages on the right. It is helpful to show both frequencies and percentages, but many people draw histograms showing just one or the other. Most histograms are

TABLE 2-8 Number enrolled in primary school in 1984 as percentage of 6–11-year-olds for 128 countries. Intervals include the lower endpoint, not the upper endpoint.

Primary school enrollment	Number of countries
20–30	4
30–40	4
40–50	4
50–60	2
60–70	9
70–80	7
80–90	6
90–100	30
100–110	37
110–120	12
120–130	2
130–140	2
Missing	9
Total	128

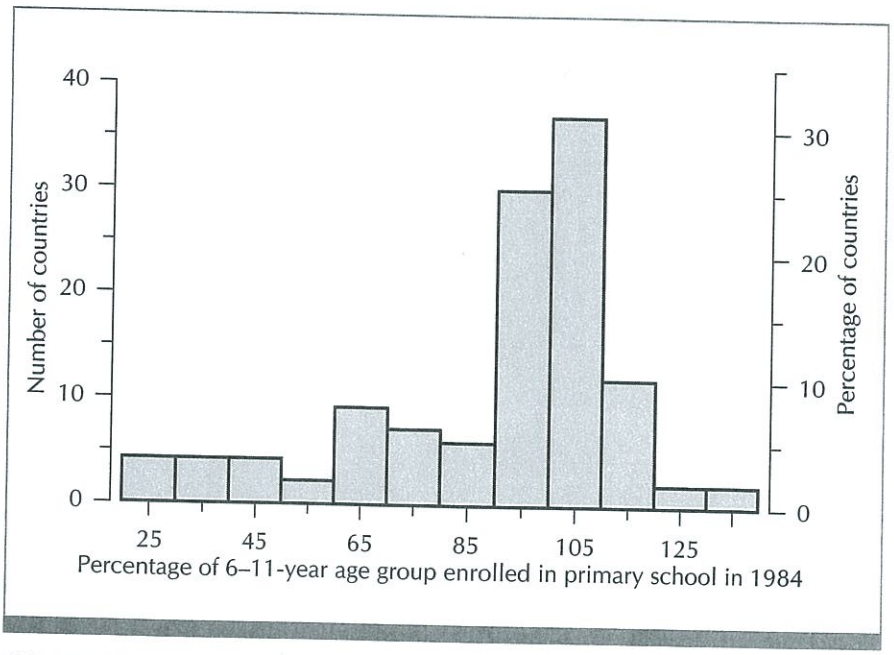

FIGURE 2-7 Histogram of number enrolled in primary school in 1984 as percentage of 6–11-year age group, for 119 countries. Intervals include the lower endpoint, not the upper endpoint. Nine countries are excluded because of missing values.

set up like the one in Figure 2-7. However, they can be oriented differently. Bars might go from left to right, for instance, as a stem-and-leaf plot is usually drawn.

> When a variable takes on relatively many values, a **histogram** is a graphical way of displaying the number of occurrences of values over specified intervals.

Figure 2-8 is a histogram with intervals of length 5: 25–30, 30–35, and so on. The endpoints of the intervals are labeled in this histogram; in Figure 2-7, we labeled the midpoints of the intervals. Either way is acceptable.

The histograms in Figures 2-7 and 2-8 show us that a majority of the countries had primary school enrollments between 90% and 110%. A few countries had enrollments greater than 110%. A large portion of the countries had relatively low primary school enrollments.

Since these two histograms provide essentially the same visual information, how can we decide between them? In general, how do we choose the

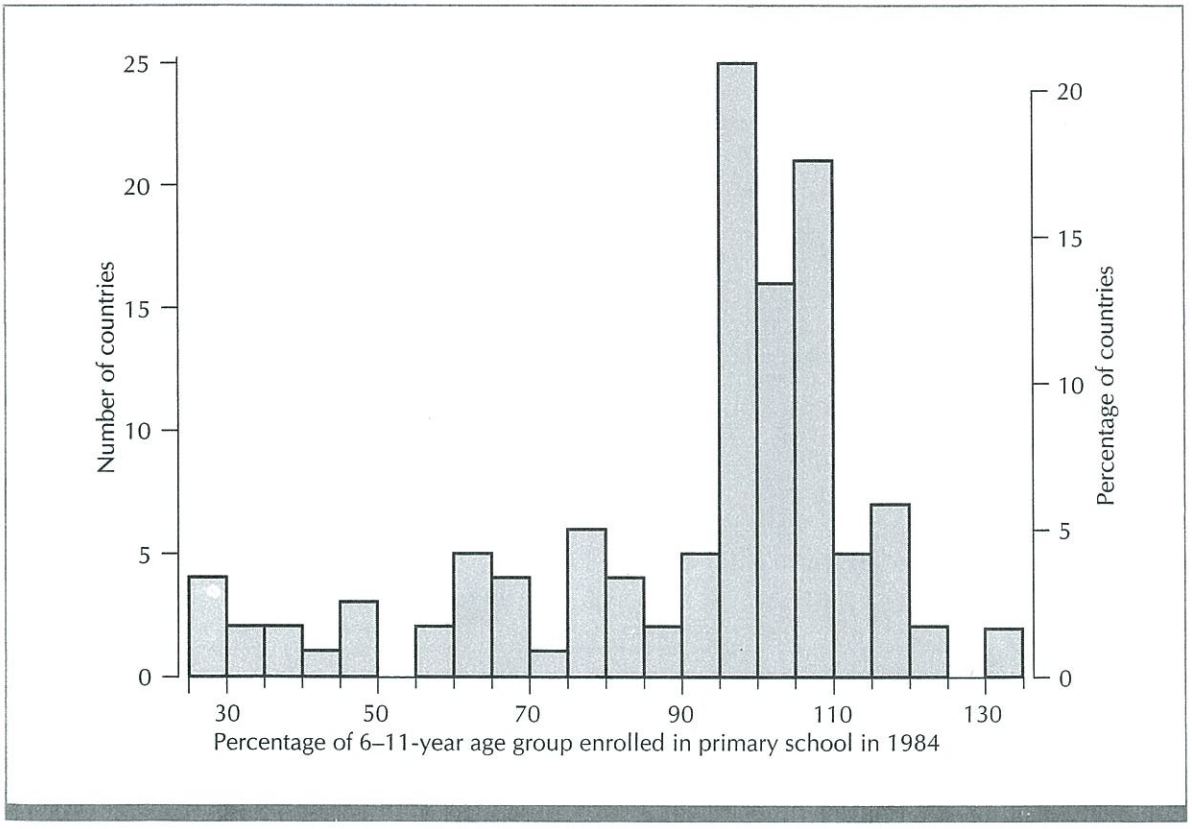

FIGURE 2-8 Histogram of number enrolled in primary school in 1984 as percentage of 6–11-year age group, for 119 countries. Intervals include the lower endpoint, not the upper endpoint. Nine countries are excluded because of missing values.

interval widths for a frequency table or histogram? If the interval widths are too great, there are few intervals and the histogram is of limited use in summarizing the variable. The extreme case is when the intervals are so wide that all values are contained in a single interval. The other extreme is when intervals are so narrow that at most a single value is contained in any interval; the resulting histogram conveys the same information as a dot plot. There is no set way to choose interval widths. We must choose a satisfactory middle ground between too many intervals and too few.

2-4

Describing the Shape of a Distribution

The *distribution* of a variable describes how the values of the variable are positioned along the number line. A dot plot shows the exact distribution of a set of values. A stem-and-leaf plot and a histogram each provide an abbreviation or summary of a distribution.

> The **distribution** of a variable is a description of how the values of the variable are positioned along an axis or number line.

As we will see later when we discuss statistical inference, the shape of the distribution of a set of observations can help us decide on a method of formal analysis. There are many shapes that a distribution can have. Three major types are symmetric distributions, distributions that are skewed to the left (negatively skewed), and distributions that are skewed to the right (positively skewed).

A distribution is *symmetric* if the values to the right of some center point form a mirror image of the values to the left of that point. Seldom will we find a real set of data values that is exactly symmetrical. But we will often encounter sets of values that are roughly or approximately symmetrical.

> A distribution is **symmetric** if the values to the right of some center point form a mirror image of the values to the left of that point.

Figure 2-9 shows a dot plot of gross national product for 19 industrial market countries. These 19 values are roughly symmetrical about a center point somewhere between 10,000 and 11,000 dollars.

Primary school enrollments are displayed in Figure 2-10 for 22 upper-

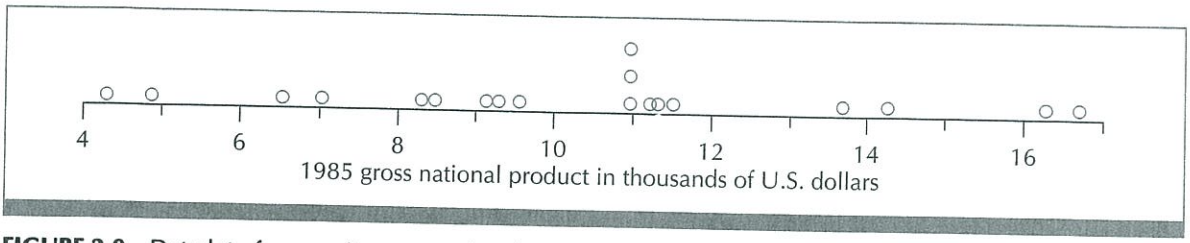

FIGURE 2-9 Dot plot of per capita gross national product in 1985 for 19 industrial market countries.

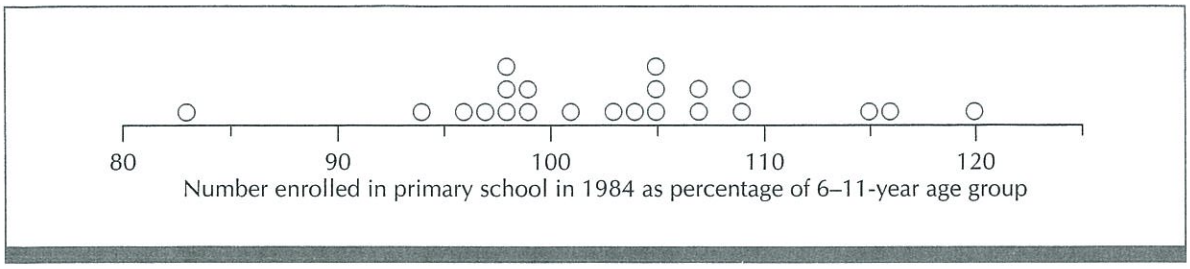

FIGURE 2-10 Dot plot of number enrolled in primary school in 1984 as percentage of 6–11-year age group for 22 upper-middle-income countries. One country is excluded with a missing value.

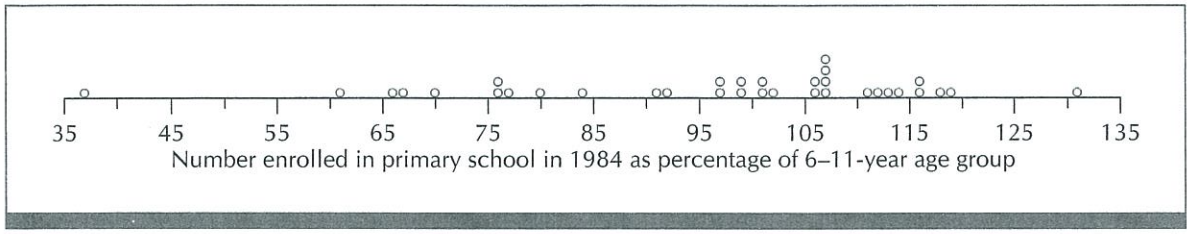

FIGURE 2-11 Dot plot of number enrolled in primary school in 1984 as percentage of 6–11-year age group for 34 lower-middle-income countries. Two countries are excluded because of missing values.

middle-income countries. These 22 values are approximately symmetrical about a center point between 100% and 105%.

Consider the histogram of primary school enrollments for 119 countries in Figure 2-8. There is a concentration of relatively high values in the 95–110 range. The distribution tails off to the left across a wide range of relatively low values. We say such a distribution is *skewed to the left* or *negatively skewed.*

> A distribution is **negatively skewed** or **skewed to the left** if there is a concentration of relatively large values, with some scatter over a range of smaller values.

Another negatively skewed distribution is illustrated in Figure 2-11. This dot plot shows primary school enrollments for 34 lower-middle-income countries. Again, there is a concentration of values on the right, and a tailing off of values to the left.

Figure 2-12 shows higher education enrollments for 119 countries. There is a concentration of relatively low values in the range from 0 to 5 or 10. The distribution tails off to the right across a wide range of relatively high values. We say this distribution is *skewed to the right* or *positively skewed.*

> A distribution is **positively skewed** or **skewed to the right** if there is a concentration of relatively small values, with some scatter over a range of larger values.

Child death rates are plotted in Figure 2-13 for 124 countries. A value of 0 indicates a country with a death rate among 1–4-year-olds of less than .5 per

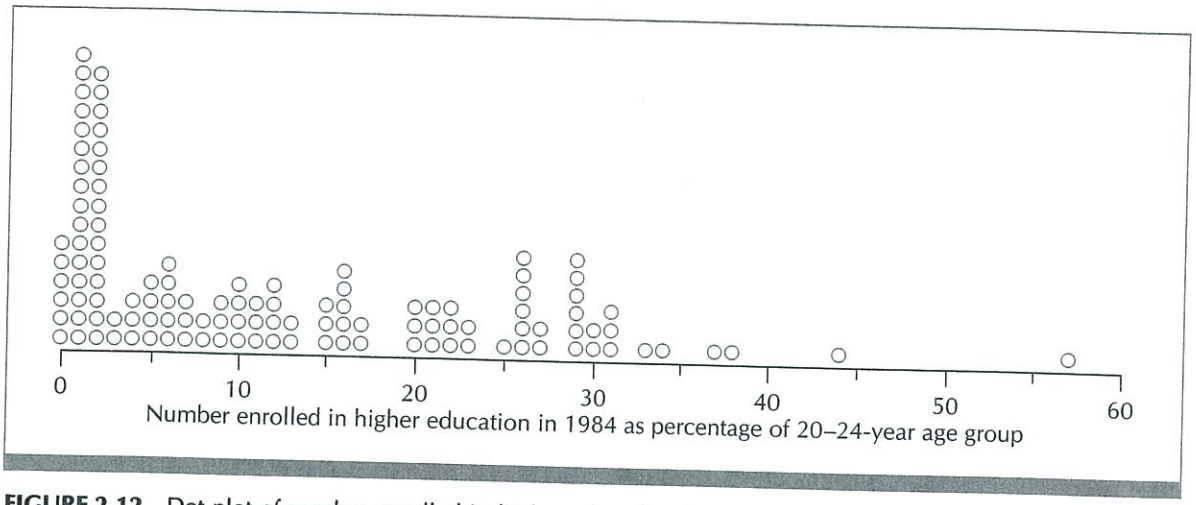

FIGURE 2-12 Dot plot of number enrolled in higher education in 1984 as percentage of 20–24-year age group for 119 countries. Nine countries are excluded because of missing values.

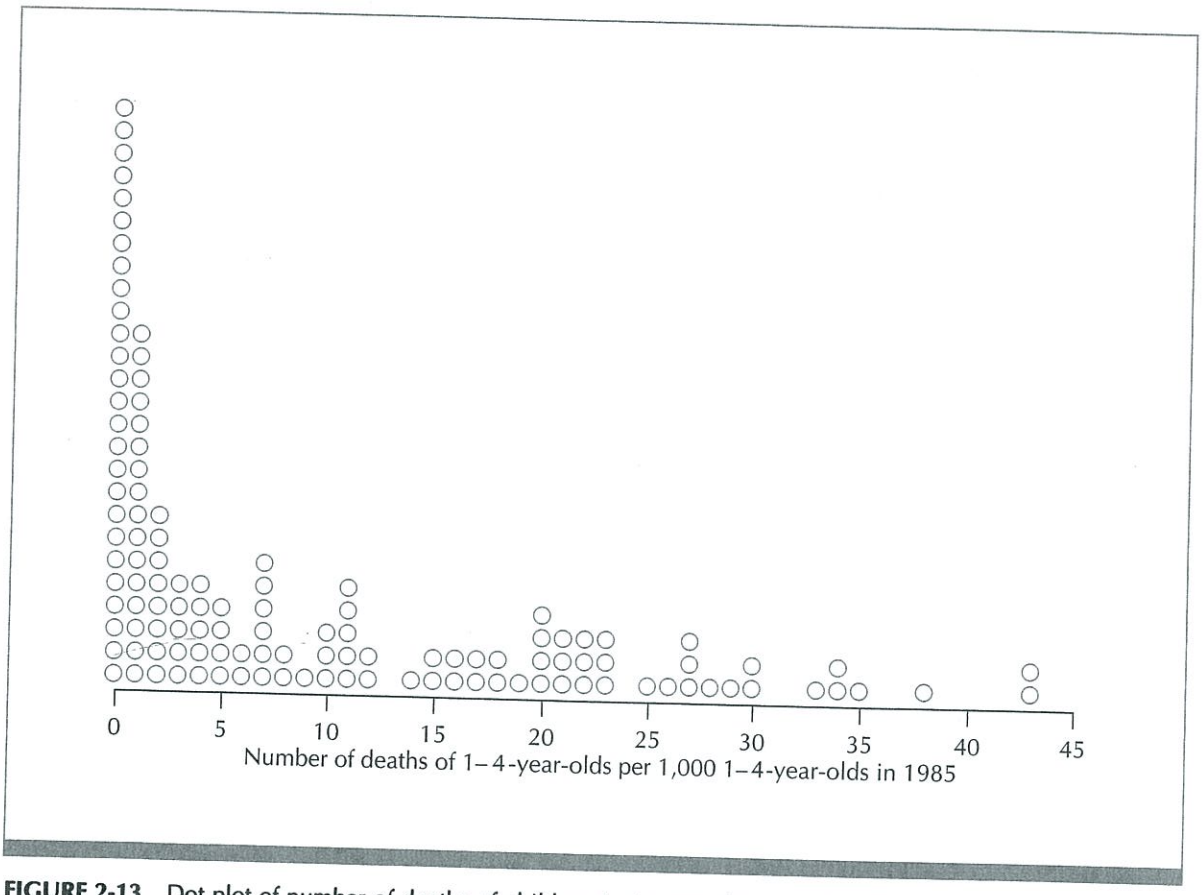

FIGURE 2-13 Dot plot of number of deaths of children 1–4 years of age per 1,000 children in this age group in 1985 for 124 countries. Four countries are excluded because of missing values.

1,000. This distribution is skewed to the right or positively skewed. A majority of the values are less than 10. The others range from 10 to 43 deaths per 1,000 1–4-year-olds.

The term *mode* or *peak* refers to a major concentration of values in a distribution. Describing the number and locations of such modes or peaks is an important part of studying a distribution, because we may be able to identify subgroups based on the modes that we see.

A **mode** or **peak** in a distribution is a major concentration of values.

Consider the plot of death rates in Figure 2-14. This distribution has a major concentration of values (a mode or peak) in the interval from 7 to 10 deaths per 1,000 population. Because there is only one major peak, this is a *unimodal distribution.* The primary school enrollments in Figure 2-8 form a unimodal distribution, as do the higher education enrollments in Figure 2-12 and the child death rates in Figure 2-13.

A **unimodal distribution** has one major peak or concentration of values.

A distribution may have more than one major peak or concentration of values. Consider the dot plot of fertility rates in Figure 2-15. (We saw stem-and-leaf plots of these values in Figures 2-3 and 2-4.) There is a major peak or mode in the interval from 1.7 to 2.1 children per woman. Another major peak is in the interval from 6.2 to 6.8 children per woman. Because there are two major peaks, this is a *bimodal distribution.* The two peaks suggest that we can consider two distinct groups of countries, differentiated by high and low values of fertility rate. The stem-and-leaf plot of per capita calorie supplies in Figure 2-5 also looks bimodal. As we noted when discussing that plot, there seem to

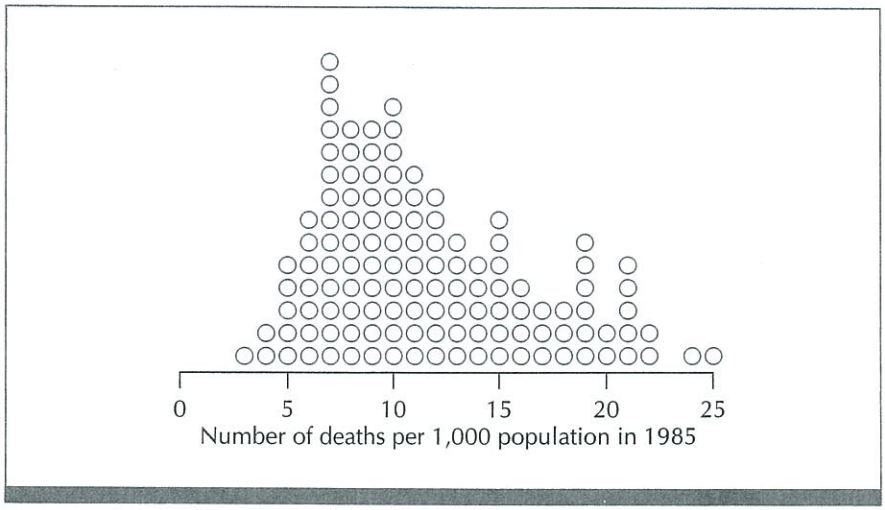

FIGURE 2-14 Dot plot of the number of deaths per 1,000 population (death rate) in 1985 for 125 countries. Three countries are excluded because of missing values. This is a unimodal distribution.

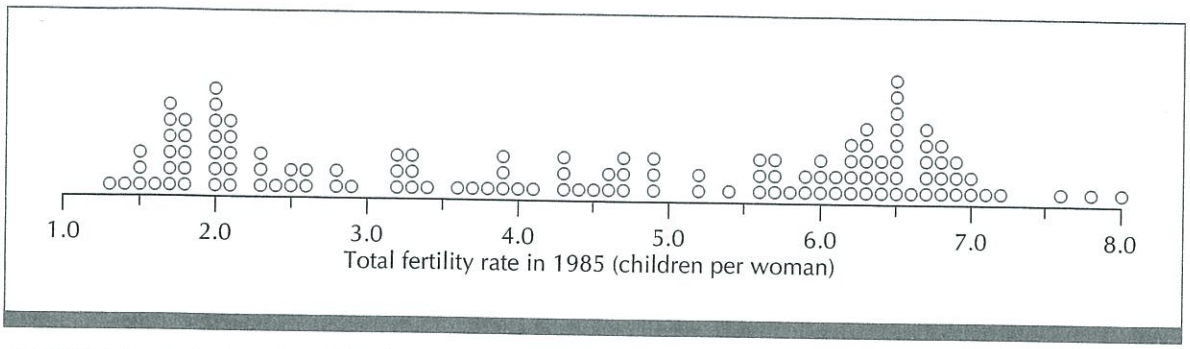

FIGURE 2-15 Dot plot of total fertility rate in 1985 for 125 countries. Three countries are excluded because of missing values. This is a bimodal distribution.

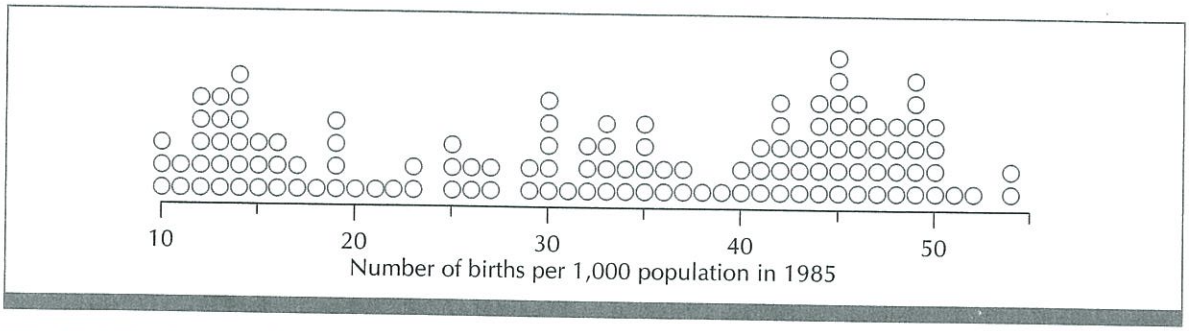

FIGURE 2-16 Dot plot of number of births per 1,000 population (birth rate) in 1985 for 125 countries. Three countries are excluded because of missing values. This is a multimodal distribution.

be two groups of countries, corresponding to adequate and inadequate per capita calorie supplies.

> A distribution is **bimodal** if it has two major peaks or concentrations of values.

We say a distribution with several peaks or modes is *multimodal.* The birth rates illustrated in Figure 2-16 form such a distribution. Values are concentrated in three intervals: from 12 to 16, from 30 to 35, and from 42 to 50 births per 1,000 population. We might classify the countries into three distinct groups based on birth rates. The life expectancies in Figure 2-1 also form a multimodal distribution; countries could be separated into several groups based on life expectancy.

> A distribution is **multimodal** if it has several major peaks or concentrations of values.

In Section 2-5 we use box plots and box graphs to summarize the distribution of a variable and to compare distributions. To discuss box plots and box graphs, we first have to introduce quantiles.

2-5

Quantiles, Box Plots, and Box Graphs

Numbers that divide an ordered list of values into equal or approximately equal sized groups are called *quantiles.* Because they summarize the information in a set of values, quantiles are descriptive statistics. Descriptive statistics are the subject of the next chapter, so you might wonder why we are discussing them here. We do so because quantiles are used in the construction of box plots and box graphs, graphical tools useful for summarizing the distribution of a variable and for comparing two or more distributions. Since graphical tools are the main subject of Chapter 2, box plots and box graphs (and therefore quantiles) are introduced here rather than in Chapter 3.

> **Quantiles** divide an ordered list of values into equal or approximately equal sized groups. If two values satisfy the definition of a quantile, we let the quantile equal the average of the two values.

Commonly used quantiles include the *median, quartiles, deciles,* and *percentiles.* The median divides an ordered list of values in half. Quartiles, deciles, and percentiles divide an ordered list of values into 4, 10, and 100 groups, respectively (if the data set is large enough).

> The **median** divides an ordered list of values in half: At least half the values are less than or equal to the median and at least half are greater than or equal to the median.
>
> **Quartiles** divide an ordered list of values into 4 groups of equal or approximately equal size. At least one-fourth of the values are less than or equal to the **first quartile** and at least three-fourths are greater than or equal to the first quartile. The **second quartile** is the same as the median. At least three-fourths of the values are less than or equal to the **third quartile** and at least one-fourth are greater than or equal to the third quartile.
>
> **Deciles** divide an ordered list of values into 10 groups of equal or approximately equal size, provided there are enough values. At least $10x\%$ of the values are less than or equal to the xth decile and at least $(100 - 10x)\%$ of the values are greater than or equal to the xth decile.
>
> **Percentiles** divide an ordered list of values into 100 groups of equal or approximately equal size, provided there are enough values. At least $x\%$ of the values are less than or equal to the xth percentile and at least $(100 - x)\%$ of the values are greater than or equal to the xth percentile.

The median, which is the same as the 50th percentile, the fifth decile, and the second quartile, divides a set of ordered values into two groups. The median of an odd number of values is the middle value. For the three numbers 1, 2, and 5, the median is the middle value, 2. The median of an even number of values is typically defined to be the average of the two middle values. For instance, the median of the four numbers 1, 2, 5, and 8 is 3.5, the average of the middle values 2 and 5.

Box plots and *box graphs* are graphical displays based on quantiles (Tukey, 1977). The *box plot* or *box-and-whisker plot* is defined by five numbers: the minimum, the first quartile, the median, the third quartile, and the maximum value for a variable. The first and third quartiles define the extremes of the box. The median is indicated within the box. The minimum and maximum values determine the whiskers. Figure 2-17 shows the structure of a box plot.

> A **box plot** is a graphical summary of the distribution of a variable. The box plot is constructed from five descriptive statistics: the minimum (or smallest value), the first quartile, the median, the third quartile, and the maximum (or largest value).

The box plot provides a brief summary of location, variation, and extreme values of a distribution. Box plots are especially useful for comparing two or more distributions. For example, Figure 2-18 shows box plots of life expectancy for 35 low-income countries and 19 industrial market countries. Before discussing the information contained in this figure, let's find the numbers we need to construct these two box plots.

Consider first the 35 nonmissing values of life expectancy for low-income countries: 40, 40, 44, 44, 45, 45, 45, 45, 45, 46, 46, 47, 47, 47, 48, 48, 48, 49, 49, 49, 51, 51, 51, 51, 52, 52, 52, 53, 54, 54, 56, 59, 65, 69, 70. To find the median, we can start counting at the smallest value and continue until we come to the middle value. (The middle in this case is the 18th value. To find how far to count, divide the sample size by 2 and then round up to the nearest integer. In this case, the sample size is 35, and 35/2 = 17.5, which we round up to 18.) Alternatively, we could start counting at the largest value, until we come to the middle. Either way, we find that the median of these 35 life expectancies is 49.

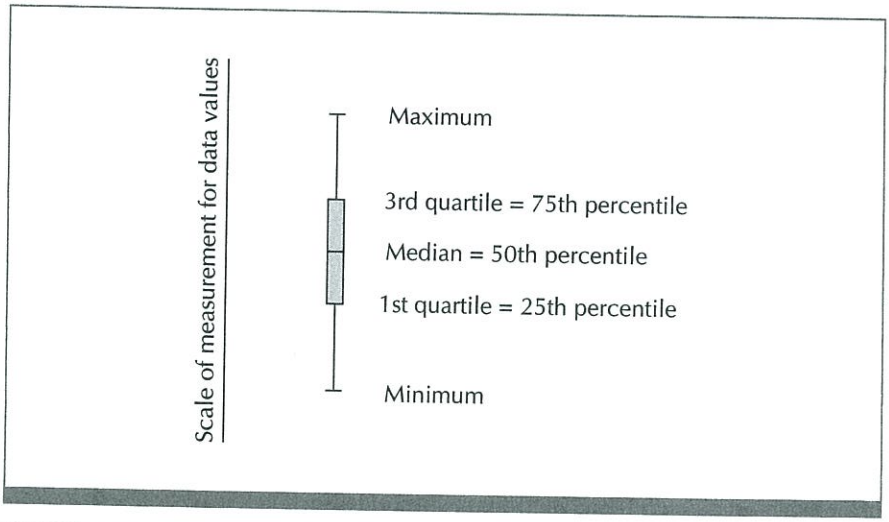

FIGURE 2-17 Structure of a box plot

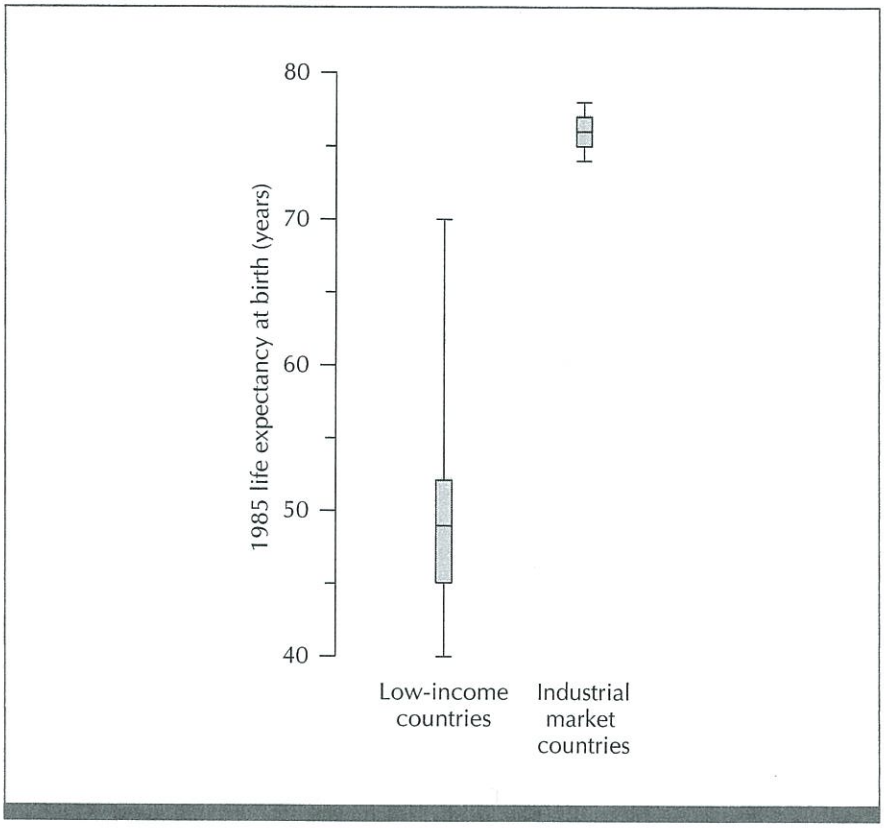

FIGURE 2-18 Box plots of 1985 life expectancy at birth, for 35 low-income countries and 19 industrial market countries. Two low-income countries are excluded because of missing values.

The first quartile, or 25th percentile, is the number with at least one-quarter of the values below and at least three-quarters above. Since there are 35 life expectancies for low-income countries, we start counting at the smallest and continue until we get to the 9th smallest value. (To find how far to count, divide the sample size by 4 and then round up to the nearest integer. In this case, 35/4 = 8.75, which we round to 9.) Doing this, we find the first quartile to be 45. The third quartile, or 75th percentile, is the number with at least three-quarters of the values below and at least one-quarter above. Because there are 35 values, we can start counting at the largest and continue until we get to the 9th largest value. If we do this, we find the third quartile to be 52.

The extreme life expectancies for the low-income countries are easy to see in the ordered list above. The five summary numbers for constructing the box plot of life expectancy for the low-income countries are the minimum, 40; the first quartile, 45; the median, 49; the third quartile, 52; and the maximum, 70. Since these summary numbers are based on quartiles, we know that approximately one-fourth of the life expectancies for these low-income countries are in each of these intervals: 40–45, 45–49, 49–52, and 52–70.

The life expectancies for the 19 industrial market countries are: 74, 74, 74, 75, 75, 75, 75, 76, 76, 76, 77, 77, 77, 77, 77, 77, 77, 78, 78. You can show for yourself that the five summary numbers for constructing the box plot of life expectancy for the industrial countries are the minimum, 74; the first quartile, 75; the median, 76; the third quartile, 77; and the maximum, 78.

The simple plots in Figure 2-18 illustrate that life expectancies are very different for the low-income and industrial market countries. Not only are the centers of the two sets of values widely separated, but also there is no overlap between the two ranges of values. The longest life expectancy among the low-income countries (maximum = 70 years) is 4 years less than the shortest life expectancy among the industrial market countries (minimum = 74 years). Life expectancies for the industrial market countries appear to be fairly symmetrical about the median. The distribution of life expectancies for the low-income countries appears less symmetrical. Dot plots (or stem-and-leaf plots or histograms) would provide more information on shape and symmetry of these two distributions.

Box Graphs

A *box graph* is similar to a box plot, with more information (Tukey, 1977; Cleveland, 1985). As illustrated in Figure 2-19, the box is the same as in a box

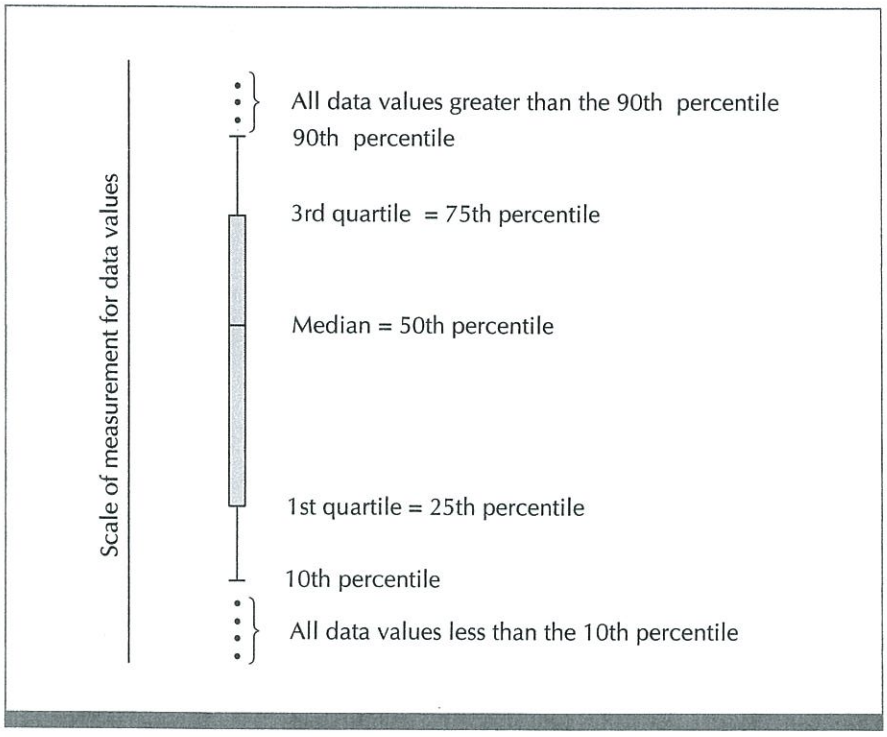

FIGURE 2-19 Structure of a box graph

plot. The 10th and 90th percentiles are the endpoints of the whiskers. In addition, values greater than the 90th percentile and less than the 10th percentile are plotted.

> A **box graph** is a graphical summary of the distribution of a variable. The box graph is constructed using the 10th percentile, the first quartile, the median, the third quartile, and the 90th percentile, plus all values less than the 10th percentile and all values greater than the 90th percentile.

Box graphs of population growth are displayed in Figure 2-20 for 35 low-income countries and 19 industrial market countries. The 35 nonmissing values of population growth for the low-income countries are: 1.2, 1.4, 1.8, 2.0, 2.0, 2.2, 2.2, 2.2, 2.3, 2.3, 2.4, 2.4, 2.5, 2.5, 2.6, 2.6, 2.6, 2.6, 2.7, 2.7, 2.9, 2.9, 3.0, 3.0, 3.0, 3.1, 3.1, 3.1, 3.2, 3.2, 3.3, 3.3, 3.5, 3.5, 4.1. For the box-and-whisker portion of the graph, we find the 10th percentile, 2.0; the first quartile, 2.3; the

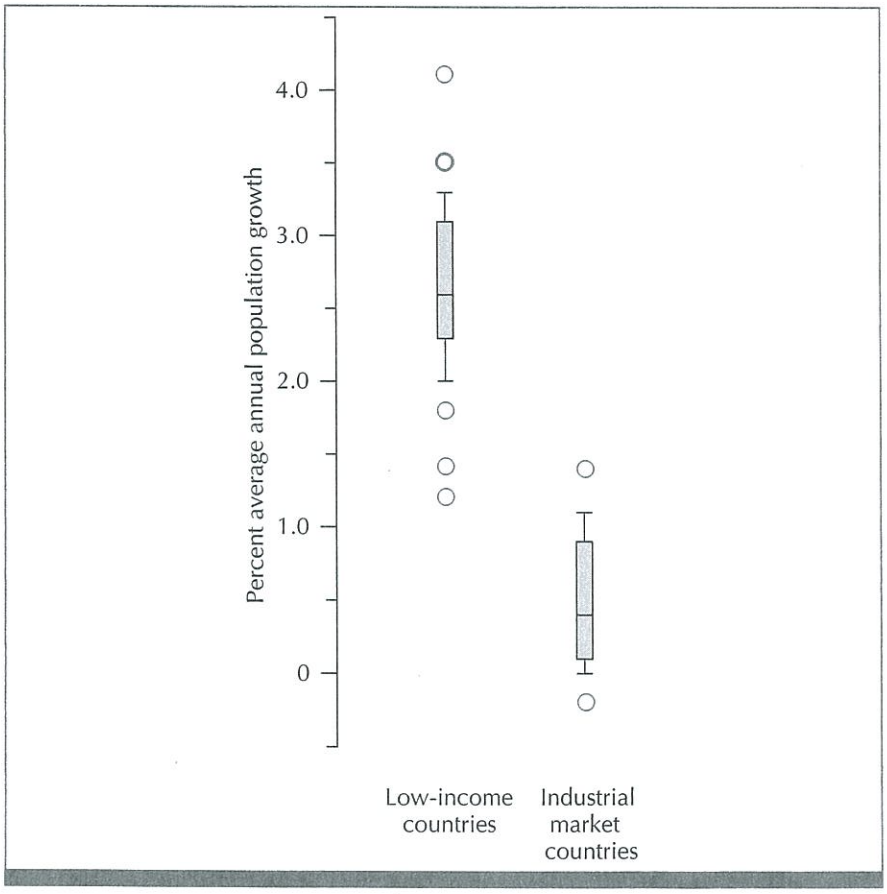

FIGURE 2-20 Box graphs of percent average annual growth of population over the period 1980–1985, for 35 low-income countries and 19 industrial market countries. Two low-income countries are excluded because of missing values.

Summary of Chapter 2

One way to study the values of a variable is to list all of them. A list in which the values are ordered from smallest to largest (or vice versa) is especially useful. A dot plot represents values as points along a number line, providing a clear picture of the distribution of values of the variable. A stem-and-leaf plot is more compact than a dot plot; it uses the actual values taken on by a variable in a graphical display that summarizes the shape of the distribution of values.

It is often useful to summarize the values of a variable rather than to list or plot all values. Frequency tables, frequency plots, and histograms are useful for this purpose.

Plots such as dot plots, stem-and-leaf plots, and histograms help us see the symmetry or skewness of a distribution of values. We also look for modes or peaks, which are concentrations of values in a distribution.

Box plots and box graphs are simple graphical displays that provide information about the location, spread, and extremes of a set of values. These plots are very simple and are especially useful for comparing distributions.

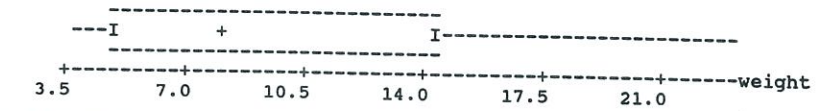

FIGURE M2-9 Box plot of the variable WEIGHT for 40 countries in Exercise 5-35

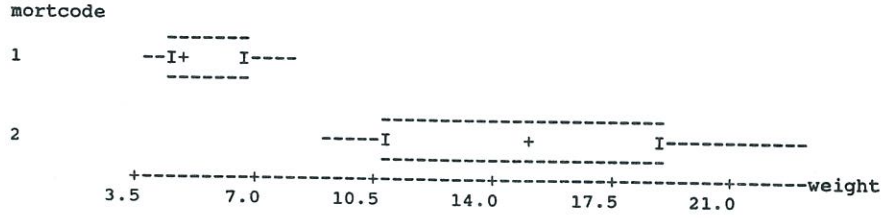

FIGURE M2-10 Box plots of WEIGHT for the 20 countries with MORTCODE = 1 and for the 20 countries with MORTCODE = 2

We can use the BY subcommand, as follows:

```
MTB>  boxplot 'weight';
SUBC> by 'mortcode'.
```

Two box plots will be displayed, one for countries with MORTCODE = 1 and one for countries with MORTCODE = 2, as in Figure M2-10.

Exercises for Chapter 2

In the exercises, answer the following questions: What would you need to know about the sample to be willing to use it to make inferences about a larger population? What is that larger population (if any)? What limitations do you see in the sample?

For each figure and table, include a legend that completely describes its contents. Note the number of cases included and the number excluded if there are missing values.

EXERCISE 2-1 The numbers of doctoral degrees awarded in the United States in 1984–1985 are listed here by race/ethnic group (American Council on Education, 1987):

Group	Number of doctoral degrees	Percent of total
American Indian	119	.4
Asian/Pacific Islander	1,106	3.4
Black	1,154	3.6
Hispanic	677	2.1
Nonresident Alien	5,317	16.5
White	23,934	74.1
Total	32,307	100.0

Construct a frequency plot of this information. Do you think you can evaluate the information more easily using the table or the plot, or do you think there is no big difference?

EXERCISE 2-2

In a study of domestic water usage conducted from July 1981 to June 1982 in Perth, Western Australia, average toilet flush volume (in liters) was determined for 147 households (James and Knuiman, 1987):

4.0	4.0	4.5	5.4	6.0	6.0	6.1	6.1	6.4	6.5	6.5
6.6	6.9	7.0	7.0	7.0	7.0	7.0	7.0	7.0	7.2	7.3
7.4	7.5	7.5	7.6	7.7	7.8	8.0	8.0	8.0	8.0	8.0
8.0	8.1	8.2	8.2	8.2	8.2	8.4	8.5	8.5	8.6	8.7
8.7	8.8	8.9	8.9	9.0	9.0	9.0	9.0	9.0	9.2	9.2
9.2	9.3	9.4	9.5	9.5	9.6	9.7	9.7	9.7	9.8	9.8
9.9	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
10.0	10.0	10.0	10.0	10.0	10.1	10.1	10.1	10.4	10.5	10.5
10.5	10.6	10.6	10.6	10.6	10.6	10.7	10.8	10.9	11.0	11.0
11.0	11.0	11.0	11.0	11.0	11.0	11.0	11.0	11.0	11.1	11.2
11.2	11.4	11.4	11.4	11.4	11.4	11.5	11.5	11.5	11.5	11.6
11.8	11.8	11.8	11.9	12.0	12.0	12.0	12.0	12.0	12.2	12.3
12.4	12.5	12.5	13.0	13.0	13.0	13.1	13.4	13.5	14.1	14.6
15.0	15.0	16.0	20.0							

a. Construct a stem-and-leaf plot of these 147 toilet flush volumes.

b. How many peaks does this distribution have? (Is the distribution unimodal, bimodal, or multimodal?)

c. Would you describe this distribution as symmetrical, negatively skewed, positively skewed, or none of these?

d. What can you say about this sample of toilet flush volumes based on the stem-and-leaf plot?

e. Why do you think there are so many 0 leaves in this sample? (*Hint:* Volumes were recorded by human chart readers.)

EXERCISE 2-3

In a study of domestic water usage conducted in Perth, Western Australia, from July 1981 to June 1982, average shower flow rate (in liters/minute) was determined for 129 houses (James and Knuiman, 1987):

2.2	2.3	3.2	3.3	3.4	3.4	3.5	3.6	3.7	3.7	3.8
3.9	4.0	4.1	4.3	4.5	4.6	4.8	4.8	4.9	5.0	5.0
5.0	5.0	5.1	5.1	5.1	5.4	5.4	5.5	5.5	5.6	5.6
5.6	5.7	5.8	5.9	6.0	6.0	6.0	6.0	6.1	6.2	6.2
6.2	6.2	6.3	6.3	6.4	6.4	6.4	6.5	6.6	6.6	6.6
6.7	6.7	6.8	6.9	6.9	6.9	6.9	7.0	7.0	7.0	7.1
7.2	7.2	7.3	7.3	7.4	7.5	7.5	7.5	7.5	7.5	7.6
7.6	7.8	8.0	8.2	8.2	8.3	8.3	8.4	8.4	8.8	9.0
9.1	9.2	9.2	9.3	9.3	9.3	9.3	9.5	9.6	9.6	9.6

9.7	9.8	9.8	10.2	10.3	10.4	10.4	10.4	10.5	10.5	10.6
10.8	10.8	11.2	11.3	11.3	11.5	11.9	11.9	11.9	12.3	12.7
13.8	14.3	14.6	15.0	15.0	15.3	15.5	18.9			

a. Construct a stem-and-leaf plot of these 129 shower flow rates.

b. How many peaks does this distribution have? (Is the distribution unimodal, bimodal, or multimodal?)

c. Would you describe this distribution as symmetrical, negatively skewed, positively skewed, or none of these?

d. What can you say about this sample of shower flow rates from what you see in the stem-and-leaf plot?

e. Can you use this sample of shower flow rates to infer characteristics of a larger population of households? What would you need to know about the sample to feel comfortable doing this?

EXERCISE 2-4

The coldest temperature on record (°F) is shown here for each state (*USA Today,* October 22, 1987, page 10A; from National Climatic Data Center):

Alabama	−27	Louisiana	−16	Ohio	−39
Alaska	−80	Maine	−48	Oklahoma	−27
Arizona	−40	Maryland	−40	Oregon	−54
Arkansas	−29	Massachusetts	−35	Pennsylvania	−42
California	−45	Michigan	−51	Rhode Island	−23
Colorado	−61	Minnesota	−59	South Carolina	−19
Connecticut	−32	Mississippi	−19	South Dakota	−58
Delaware	−17	Missouri	−40	Tennessee	−32
Florida	−2	Montana	−70	Texas	−23
Georgia	−17	Nebraska	−47	Utah	−69
Hawaii	12	Nevada	−50	Vermont	−50
Idaho	−60	New Hampshire	−46	Virginia	−30
Illinois	−35	New Jersey	−34	Washington	−48
Indiana	−35	New Mexico	−50	West Virginia	−37
Iowa	−47	New York	−52	Wisconsin	−54
Kansas	−40	North Carolina	−34	Wyoming	−63
Kentucky	−34	North Dakota	−60		

a. Construct a frequency table and draw a histogram using each of three widths (say, widths 5, 10, and 20). Which interval width do you prefer for the histogram? Why?

b. Is the distribution symmetrical, negatively skewed, positively skewed, or none of these?

c. Is the distribution unimodal, bimodal, or multimodal?

EXERCISE 2-5

Stress loads at which 41 graphite beams fractured (bend stress, MPa $\times 10^{-2}$) are listed below (Cheng, 1987; from Cooper, 1984):

2.7555	2.9890	3.0065	3.0649	3.1233	3.1525	3.1525
3.1817	3.2225	3.2284	3.2692	3.2984	3.3276	3.3276
3.3743	3.3743	3.3860	3.3860	3.3860	3.4152	3.4152
3.4152	3.4444	3.4619	3.4736	3.4736	3.5028	3.5028
3.5320	3.5436	3.5611	3.5611	3.5728	3.5903	3.6195
3.6779	3.7071	3.7363	3.7363	3.7363	4.0282	

a. Construct a frequency table and histogram using each of three widths (say, widths of .1, .2, and .5). Which interval width do you prefer for the histogram? Why?

b. Is the distribution unimodal, bimodal, or multimodal?

c. Construct a box plot of these observations.

d. Would you describe this distribution as symmetrical, negatively skewed, positively skewed, or none of these?

e. Compare the information about these stress loads that you obtain from the histograms and the box plot.

EXERCISE 2-6

A stem-and-leaf plot can provide a quick way to look at a distribution when you are doing some data analysis by hand, rather than on the computer. Consider the following unordered list of strength measurements (in pounds per square inch) on 100 samples of yarn (Duncan, 1974, page 67; from U.S. Department of Agriculture, 1945):

66	117	132	111	107	85	89	79	91	97	138	103	111
86	78	96	93	101	102	110	95	96	88	122	115	92
137	91	84	96	97	100	105	104	137	80	104	104	106
84	92	86	104	132	94	99	102	101	104	107	99	85
95	89	102	100	98	97	104	114	111	98	99	102	91
95	111	104	97	98	102	109	88	91	103	94	105	103
96	100	101	98	97	97	101	102	98	94	100	98	99
92	102	87	99	62	92	100	96	98				

a. Start a stem-and-leaf plot with stems in tens and leaves in units. Use two rows per stem. Your stems will go from 6 to 13 (two rows each).

b. Now write in the leaves as you come to them in the list of yarn strengths. The first leaf is 6, in the second row for stem 6, since the first yarn strength is 66. Reading across, the next yarn strength is 117. The leaf will be 7, in the second row for stem 11. And so on. You will end up with a stem-and-leaf plot in which the leaves are not listed in order of magnitude. From this plot you can easily rearrange the leaves in order to get a stem-and-leaf plot with ordered leaves.

c. Is this distribution of yarn strengths unimodal, bimodal, or multimodal?

d. Would you describe the distribution as symmetrical, negatively skewed, positively skewed, or none of these?

EXERCISE 2-7

Survival times (in days from diagnosis) are listed below for 43 patients with chronic granulocytic leukemia (Hollander and Wolfe, 1973, pages 251–252; Siddiqui and Gehan, 1966):

7	47	58	74	177	232	273	285	317
429	440	445	455	468	495	497	532	571
579	581	650	702	715	779	881	900	930
968	1,077	1,109	1,314	1,334	1,367	1,534	1,712	1,784
1,877	1,886	2,045	2,056	2,260	2,429	2,509		

a. What was the median survival time for this group of patients?

b. Construct a box graph for these survival times.

c. Describe the distribution in terms of shape, location, variation, and symmetry or skewness.

EXERCISE 2-8

Numbers of pregnancies per 1,000 girls 15–19 years of age in 1980 are listed below by state (*USA Today,* December 11, 1986, page 7A; from Alan Guttmacher Institute):

Alabama	117.3	Louisiana	118.1	Ohio	101.3
Alaska	124.2	Maine	86.9	Oklahoma	119.5
Arizona	123.2	Maryland	122.5	Oregon	118.7
Arkansas	117.2	Massachusetts	85.7	Pennsylvania	90.3
California	140.2	Michigan	102.4	Rhode Island	83.1
Colorado	113.7	Minnesota	77.0	South Carolina	113.7
Connecticut	80.7	Mississippi	125.0	South Dakota	86.4
Delaware	105.6	Missouri	106.4	Tennessee	113.0
Florida	131.2	Montana	93.3	Texas	137.0
Georgia	130.9	Nebraska	80.7	Utah	94.6
Hawaii	105.6	Nevada	144.0	Vermont	94.8
Idaho	96.4	New Hampshire	80.7	Virginia	107.4
Illinois	100.6	New Jersey	95.8	Washington	122.3
Indiana	101.9	New Mexico	125.6	West Virginia	103.6
Iowa	79.0	New York	100.7	Wisconsin	84.8
Kansas	101.0	North Carolina	110.3	Wyoming	126.6
Kentucky	110.7	North Dakota	74.8		

a. Construct a frequency table and draw a histogram summarizing these pregnancy rates.

b. Is the distribution symmetrical, negatively skewed, positively skewed, or none of these?

c. Is the distribution unimodal, bimodal, or multimodal?

EXERCISE 2-9

Consider the following list of shear strengths of welds of stainless steel, in pounds per weld (Duncan, 1974, page 67):

2,385	2,280	2,330	2,360	2,350	2,350	2,370	2,310	2,280
2,310	2,310	2,330	2,280	2,290	2,190	2,280	2,270	2,260
2,250	2,260	2,270	2,270	2,305	2,310	2,340	2,330	2,340
2,350	2,360	2,360	2,340	2,280	2,290	2,350	2,330	2,280
2,285	2,250	2,340	2,330	2,350	2,275	2,190	2,240	2,230
2,210	2,220	2,190	2,230	2,160	2,270	2,400	2,350	2,360
2,360	2,300	2,350	2,340	2,290	2,250	2,270	2,340	2,310
2,360	2,300	2,430	2,340	2,440	2,370	2,340	2,360	2,340
2,330	2,380	2,350	2,360	2,390	2,360	2,400	2,320	2,360
2,350	2,340	2,320	2,350	2,330	2,320	2,300	2,280	2,230
2,290	2,270	2,290	2,270	2,270				

a. Construct a stem-and-leaf plot of these weld strengths. Use the first two digits as the stem and the last two digits as the leaf. Your stems will go from 21 to 24. Also, use two rows per stem. Construct the plot from the weld strengths as they are listed. You can always rearrange the leaves later to get a stem-and-leaf plot with ordered leaves.

b. Use the information in the stem-and-leaf plot to construct a histogram. Compare the information provided by the two graphs.

c. Is the distribution of weld strengths unimodal, bimodal, or multimodal?

d. Would you describe this distribution as symmetrical, negatively skewed, positively skewed, or none of these?

EXERCISE 2-10

Our ability to read a sign at night depends on the light intensity near the sign. The following are measurements of light intensity (in candela per square meter) for 30 highway signs in a city (Milton and Arnold, 1986, page 249; based on "Use of Retroreflectors in the Improvement of Nighttime Highway Visibility," by H. Waltman in *Color*):

10.9	1.7	9.5	2.9	9.1	3.2	9.1	7.4	13.3	13.1
6.6	13.7	1.5	7.4	9.9	13.6	17.3	3.6	4.9	13.1
7.8	10.3	10.3	9.6	5.7	6.3	2.6	15.1	2.9	16.2

a. Construct a stem-and-leaf plot of these intensity measurements.

b. Describe the distribution in terms of location, variability, peaks, and symmetry or skewness.

EXERCISE 2-11

Consider the following measurements of rainfall (in acre-feet) from 26 seeded clouds (Devore, 1987, page 30; from "A Bayesian Analysis of a Multiplicative Treatment Effect in Weather Modification," *Technometrics,* 1975, pages 161–166):

4.1	7.7	17.5
31.4	32.7	40.6
92.4	115.3	118.3
119.0	129.6	198.6

200.7	242.5	255.0
274.7	274.7	302.8
334.1	430.0	489.1
703.4	978.0	1,656.0
1,697.8	2,745.6	

a. Construct a box graph of these rainfall measurements.

b. Discuss the information provided in this box graph.

EXERCISE 2-12 The following are measurements of sulfur dioxide concentrations (in micrograms per cubic meter) from a damaged Bavarian forest (Milton and Arnold, 1986, page 226; based on Roberts, 1983):

33.4	38.6	41.7	43.9	44.4	45.3	46.1	47.6	50.0	52.4	52.7	53.9
54.3	55.1	56.4	56.5	60.7	61.8	62.2	63.4	65.5	66.6	70.0	71.5

a. Construct a dot plot and a box plot of these measurements. Compare the information provided by the two plots.

b. Is the distribution unimodal, bimodal, or multimodal?

c. Would you describe this distribution as symmetric, negatively skewed, positively skewed, or none of these?

d. The average concentration of sulfur dioxide in undamaged parts of the country was 20 micrograms per cubic meter. Does the evidence here suggest that the damage in the forest might be related to acid rain? Or would you hesitate to suggest such a connection? Explain your answer.

EXERCISE 2-13 Shear strengths of welds in pounds per weld are shown below for 95 samples of the same material (Duncan, 1974, page 69):

146	148	146	140	150	146	146	142	152	148	158	152
146	156	152	150	150	146	148	152	156	162	160	146
156	150	146	154	150	148	148	148	150	152	150	152
152	152	148	152	146	152	150	144	152	154	138	154
160	106	152	150	150	150	152	158	152	156	156	152
154	154	144	148	154	152	152	154	154	158	158	156
160	162	154	148	152	146	158	162	160	160	158	154
160	162	164	150	150	152	144	150	152	154	152	

a. Construct a dot plot and a box graph of these observations. Compare the information provided in these two plots.

b. Is this distribution unimodal, bimodal, or multimodal?

c. Would you describe the distribution as symmetrical, negatively skewed, positively skewed, or none of these?

EXERCISE 2-14 The following list shows percentages of tanks with leaks, from daily inspections of fuel tanks (Duncan, 1974, page 68; from Kauffman, 1945, page 17):

32.6	35.5	44.0	43.3	40.8	40.0	49.4	45.5	46.8	45.9	45.9
49.5	46.4	50.5	50.5	50.0	53.8	52.8	53.3	53.6	57.6	55.7
56.4	58.7	59.3	59.5	59.6	55.9	57.8	55.1	62.7	61.4	60.5
67.3	72.0	72.4	73.3	70.1	72.7	73.8	75.0	75.5	82.3	80.5
80.0	80.5	86.8	54.6							

a. Construct a dot plot and box graph for these observations. Compare the information provided by the two plots.

b. Describe the distribution with respect to location, variation, peaks, and symmetry or skewness.

EXERCISE 2-15 Errors (in meters) made using a lightweight handheld laser range finder to locate an object from 500 meters are shown below (Milton and Arnold, 1986, page 252; from *Civil Engineering,* February 1983, page 52). A "+" indicates an overestimate and a "−" indicates an underestimate.

−.10	−.05	+.01	+.03	+.06	−.07
−.03	+.01	+.03	+.09	−.06	−.02
+.02	+.05	+.10			

a. Construct a dot plot and box plot of these errors. Compare the information provided by these two plots.

b. Describe the distribution of errors in terms of location, variation, and symmetry or skewness.

EXERCISE 2-16 In a study of X-ray microanalysis as a method of chemical analysis, a chemical was analyzed that was in theory 26.6% potassium by weight. Listed here are the percentages of potassium found by X-ray microanalysis of 27 samples (Milton and Arnold, 1986, pages 249–250; based on Kiss, 1983).

21.9	23.1	24.0	24.6	24.9	25.4	26.7	22.0	23.4	24.1	24.7
25.1	25.5	27.2	22.1	23.7	24.2	24.8	25.2	26.5	27.8	22.1
23.8	24.5	24.8	25.3	26.5						

a. Construct a stem-and-leaf plot and box plot of these observations. Compare the information provided in these two plots.

b. Describe the distribution with respect to location, spread, peaks, and symmetry or skewness.

c. How well did X-ray microanalysis seem to do in this experiment?

EXERCISE 2-17 Lifespan (in kilometers driven) was determined for each of 70 aluminum air batteries being developed for electric cars (Milton and Arnold, 1986, page 547; from "Aluminum-Air Battery Development: Toward an Electric Car," *Energy and Technology Review,* June 1983, pages 20–33):

1,625	1,726	2,498	1,942	2,216	1,631	2,101	1,820	2,239
2,037	2,618	2,173	1,902	2,415	1,698	2,587	1,810	2,612
1,733	2,245	1,947	1,622	2,016	1,867	2,831	2,357	1,747

2,417	2,021	2,639	1,650	2,702	1,929	2,381	1,719	2,291
2,093	1,603	3,150	2,109	1,913	1,727	1,672	2,071	2,815
1,871	2,750	2,280	1,763	2,470	2,353	1,893	2,897	2,539
2,150	1,702	1,802	2,925	2,918	1,635	3,070	1,988	2,306
1,616	2,178	1,750	3,200	1,690	2,592	2,072		

a. Construct a frequency table and histogram for these battery lifespans.

b. Describe the distribution in terms of location, spread, peaks, and symmetry or skewness.

EXERCISE 2-18 Numbers of defects found in 29 100-yard pieces of wool cloth are listed below (Duncan, 1974, page 42; data provided by the Bendix Radio Division of Bendix Aviation Corporation):

2	0	0	1	1	2	2	1	0	1	1	0	2	1	
1	5	0	3	1	1	1	2	2	1	0	0	0	1	4

Construct a frequency table and a frequency plot summarizing this information. Discuss the results.

EXERCISE 2-19 Base pay of the state's governor in 1986 is shown here (in dollars) for each state (*USA Today,* December 11, 1986, page 9C; data supplied by Council of State Governments).

Alabama	63,839	Louisiana	73,400	Ohio	65,000
Alaska	85,728	Maine	35,000	Oklahoma	70,128
Arizona	62,500	Maryland	75,000	Oregon	72,000
Arkansas	35,000	Massachusetts	75,000	Pennsylvania	75,000
California	49,100	Michigan	85,800	Rhode Island	49,500
Colorado	60,000	Minnesota	84,560	South Carolina	60,000
Connecticut	65,000	Mississippi	63,000	South Dakota	55,120
Delaware	70,000	Missouri	81,000	Tennessee	68,200
Florida	78,757	Montana	50,542	Texas	94,350
Georgia	79,356	Nebraska	40,000	Utah	60,009
Hawaii	59,400	Nevada	65,000	Vermont	60,000
Idaho	50,000	New Hampshire	62,880	Virginia	75,000
Illinois	58,000	New Jersey	85,000	Washington	63,000
Indiana	65,988	New Mexico	60,000	West Virginia	72,000
Iowa	64,000	New York	100,000	Wisconsin	75,337
Kansas	65,000	North Carolina	98,196	Wyoming	70,000
Kentucky	61,200	North Dakota	60,862		

a. Construct a frequency table and draw a histogram summarizing these governors' salaries.

b. Is the distribution symmetrical, negatively skewed, positively skewed, or none of these?

c. Is the distribution unimodal, bimodal, or multimodal?

EXERCISE 2-20

Forty-six U.S. universities accepted less than half of applicants in 1985 and had an average freshman SAT score of at least 1200. The average yearly cost (in dollars; tuition, room and board, books, supplies, any out-of-state surcharge) for each of these universities is shown below (*USA Today,* December 15, 1986, page 2D).

U.S. Naval Academy	0	Tufts	17,060
U.S. Military Academy	0	Virginia	9,320
Stanford	17,458	Penn	17,210
Harvard/Radcliffe	17,395	Lafayette	14,600
Princeton	17,555	Haverford	15,930
Yale	17,400	Wesleyan (Connecticut)	16,565
Brown	17,264	William & Mary	10,084
Cooper Union	1,300	Colgate	15,680
Amherst	15,920	Trinity (Connecticut)	15,370
Dartmouth	17,285	Bates	15,070
USAF Academy	0	Colby	16,000
Bowdoin	15,620	Bucknell	14,965
Georgetown	15,830	Chicago	17,310
Duke	14,340	Cal Tech	16,385
Williams	15,498	Northwestern	16,175
Columbia	17,175	Hamilton (New York)	15,100
Middlebury	14,440	Claremont/McKenna	15,300
Cornell	16,490	Notre Dame	12,240
Rice	9,770	Carleton	13,575
Swarthmore	16,200	Wellesley	15,980
MIT	17,700	Harvey Mudd	16,030
Davidson	12,470	UNC, Chapel Hill	7,470
Washington & Lee	11,780	Vassar	15,498

a. Construct a frequency plot and histogram of these college costs.

b. Describe the distribution in terms of peaks and skewness.

EXERCISE 2-21

In the 1980 Wisconsin Restaurant Survey, information on number of full-time employees was obtained on 265 restaurants (Ryan, Joiner, and Ryan, 1985, pages 77, 321–328). Listed below are number of full-time employees (number of restaurants): 0 (58), 1 (25), 2 (29), 3 (24), 4 (17), 5 (11), 6 (16), 7 (10), 8 (8), 9 (1), 10 (16), 11 (1), 12 (3), 13 (3), 14 (1), 15 (4), 16 (1), 18 (3), 20 (10), 25 (5), 26 (1), 28 (1), 30 (6), 32 (1), 35 (2), 36 (1), 40 (3), 42 (1), 51 (1), 80 (1), 250 (1).

a. Construct a frequency table and frequency plot for number of full-time employees. For the frequency plot, you may wish to exclude the largest value or show a break in the scale on the horizontal axis.

b. Describe the distribution of number of full-time employees for these 265 restaurants.

c. These 265 restaurants were part of 1,000 restaurants surveyed. With only 26.5% response, would you be willing to use this sample to draw inferences about all Wisconsin restaurants?

EXERCISE 2-22

Information was obtained on the length of stay (in days) for all patients voluntarily committed to the acute psychiatric unit of a Wisconsin health care center during the first half of 1981 (Ryan, Joiner, and Ryan, 1985, page 82). Listed below are the lengths of stay (number of patients): 0 (3), 1 (11), 2 (5), 3 (2), 4 (3), 5 (5), 6 (2), 7 (2), 8 (3), 9 (2), 10 (1), 11 (3), 12 (1), 13 (2), 14 (1), 15 (1), 18 (1), 19 (2), 25 (4), 35 (2), 45 (1), 75 (1).

a. Construct a frequency table and frequency plot for these lengths of stay.

b. Describe the distribution of length of stay for this group of patients.

EXERCISE 2-23

Numbers of cancers diagnosed in Western Australia in 1982 are shown below for each of the ten leading sites, separately for males and females (Hatton and Clarke-Hundley, 1984, page 21).

Site	Number of males	Site	Number of females
Lung	360	Breast	384
Prostate	234	Melanoma	158
Colon	147	Colon	143
Melanoma	142	Lung	109
Bladder	129	Cervix	98
Stomach	102	Rectum	81
Rectum	100	Uterus	73
All leukemias	46	Ovary	66
Kidney	44	Bladder	49
Pancreas	40	All leukemias	48

Calculate the percentage of the 1,344 males in each site category. Calculate the percentage of 1,209 females in each site category. Construct two frequency plots using this information, one plot for males and one for females. Note that each of these frequency plots summarizes frequency information about a qualitative or categorical variable: site of cancer. You may want to construct each frequency plot with the cancer sites listed in the order shown above, along the vertical axis (for easier reading). Frequencies and percentages can then be indicated on horizontal axes, say frequencies on the bottom horizontal axis and percentages on the top. A frequency plot arranged in this way for a qualitative variable is a form of dot chart, discussed in Chapter 4. If you want to compare the plots for males and females, how should you construct the scales for the two frequency plots?

EXERCISE 2-24 In a study of preovulatory estrogen levels and basal body temperature, investigators estimated the mid-cycle fertile period for each of 24 women (Carter and Blight, 1981).

Length of fertile period (days)	Number of women
4	1
5	4
6	3
7	3
8	6
9	5
10	1
11	1
Total	24

a. Construct a frequency plot based on this frequency table.

b. Calculate any descriptive statistics you think are appropriate.

c. Discuss your findings.

EXERCISE 2-25 The percentage of dwelling units with lead paint is shown here for 23 Massachusetts communities (*The Boston Sunday Globe,* January 25, 1987, page 29).

Community	Percentage of dwelling units with lead paint	Community	Percentage of dwelling units with lead paint
Arlington	63.6	Malden	79.4
Boston	80.7	Medford	83.6
Brockton	63.6	New Bedford	83.5
Brookline	72.6	Newton	72.8
Chelsea	93.6	Pittsfield	71.1
Chicopee	52.8	Quincy	76.3
Fall River	86.4	Somerville	48.6
Framingham	33.8	Springfield	69.4
Haverhill	81.9	Waltham	60.6
Lawrence	82.9	Weymouth	49.9
Lowell	76.1	Worcester	77.3
Lynn	82.1		

Display and/or summarize these observations in any ways that you find useful.