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 5

Studying More Than Two Variables at a Time

IN THIS CHAPTER

- Multidimensional frequency table (or contingency table)
- Studying two quantitative variables within levels of a qualitative variable
- Scatterplot matrix
- Framed rectangles on a map

Is the relationship between female primary school enrollment and contraception use the same for all economic categories? How does the difference between female and male life expectancy vary with overall life expectancy within economic categories? What is a good way to display scatterplots involving birth rate, calorie supply, contraception use, and female primary school enrollment? How should we display quantitative information on a map? We will approach each of these questions with a tool for studying more than two variables at a time. Then we will discuss some general principles for constructing effective graphical displays. Let's begin by examining the relationship among three qualitative variables.

5-1

Multidimensional Frequency Tables for Several Qualitative Variables

We used a two-way frequency table (Table 4-3) to look at the relationship between female primary school enrollment and contraception use among the 89 countries with information on both variables. We found an association between female primary school enrollment and contraception use among these countries. Most of the countries with low enrollments also have low contraception use; more than half of the countries with higher enrollments also have greater contraception use.

Is the relationship between these two variables similar for each economic category? To answer this question, we can look at a three-way frequency table (or three-dimensional contingency table). A three-way frequency table displays counts of cases within each combination of three qualitative variables.

> A **multiway frequency table** or multidimensional contingency table displays the number of cases within each combination of categories of several qualitative variables.

Table 5-1 shows the number of countries within each combination of levels of economic category, contraception use, and female primary school enrollment, as a two-way display of female primary school enrollment and contraception use for each economic category. The data could be arranged differently, but separate displays by economic category seem reasonable here.

We can see a strong association among the three variables in Table 5-1. Lower enrollments and lower contraception use are associated with lower economic categories. Thirty-nine countries have missing information on either female primary school enrollment or contraception use, or both. It might be informative to display missing categories for these two variables in a breakdown by economic category. We will not do that here. However, we will be cautious in our interpretations because of the large number of missing values.

TABLE 5-1 Countries classified by economic category, percentage of married women of childbearing age using contraception in 1984, and number of females in primary school in 1984 as percentage of 6–11-year age group. Thirteen low-income countries, four lower-middle-income countries, six upper-middle-income countries, all four high-income oil exporters, seven industrial market countries, and five nonmembers are excluded because of missing information on female primary school enrollment or contraception use or both.

Economic category	Percent contraception use in 1984	1984 primary school enrollment for females		
		≤ 70%	> 70%	Total
Low-income	≤ 35%	16	6	22
	> 35%	0	2	2
	Total	16	8	24
Lower-middle-income	≤ 35%	7	11	18
	> 35%	1	13	14
	Total	8	24	32
Upper-middle-income	≤ 35%	0	2	2
	> 35%	0	15	15
	Total	0	17	17
Industrial market	≤ 35%	0	0	0
	> 35%	0	12	12
	Total	0	12	12
Nonmembers	≤ 35%	0	1	1
	> 35%	0	3	3
	Total	0	4	4

5-2

Scatterplots for Studying Two Quantitative Variables Within Levels of a Qualitative Variable

Instead of categorizing values of contraception use and female primary school enrollment as either low or high, we might prefer to study these two indicators as quantitative variables. Let's see how they are related within economic categories.

Figure 5-1 is a scatterplot of contraception use and female primary school enrollment, with economic categories designated by different plotting symbols. A difficulty with this approach is that overlapping plotting symbols result when two or more countries have similar values of the two variables. Contrasting colors to designate levels of the qualitative variable can be very effective (Cleveland, 1985, pages 205–207).

We see an overall increasing relationship between female primary school enrollment and contraception use. The lower left-hand portion of the plot corresponds to lower contraception use and lower female primary school enrollments. These countries are predominantly low-income and lower-middle-income. The upper right-hand portion of the plot corresponds to greater con-

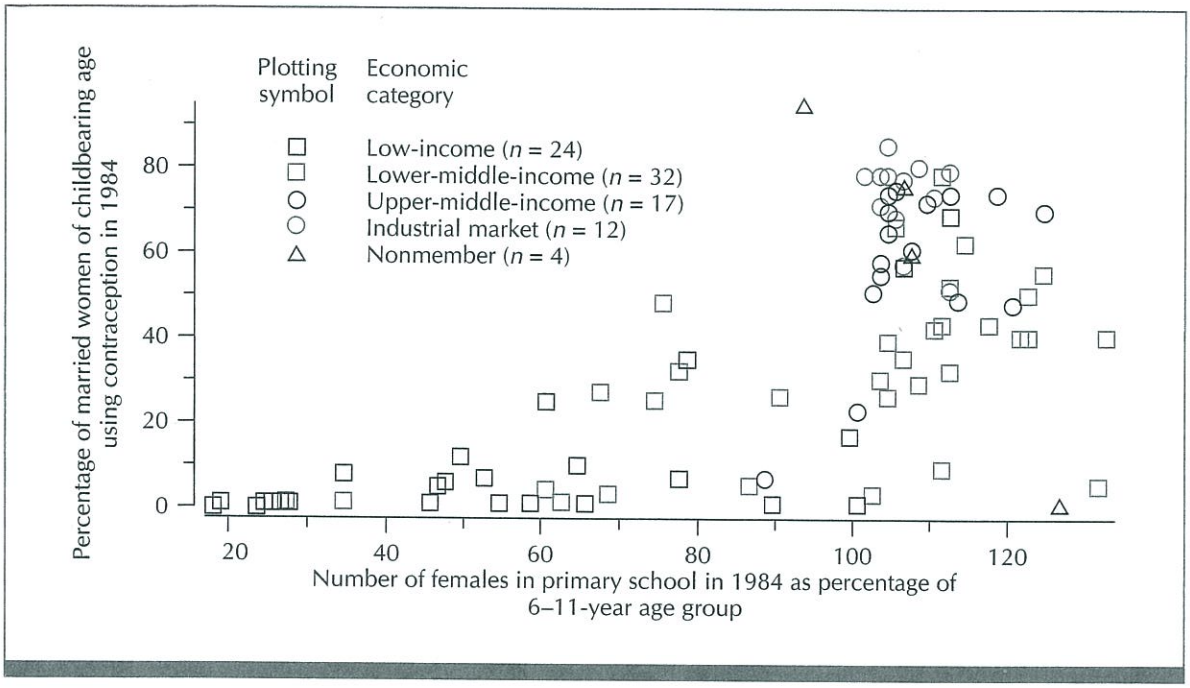


FIGURE 5-1 Scatterplot of percentage of married women of childbearing age using contraception in 1984 and number of females in primary school in 1984 as percentage of 6–11-year age group, with economic categories distinguished by plotting symbols. Thirty-nine countries are excluded because of missing values on at least one of the two indicators: 13 low-income, 4 lower-middle-income, 6 upper-middle-income, all 4 high-income oil exporters, 7 industrial market, and 5 nonmember countries.

traception use and higher female primary school enrollments. These countries are primarily industrial market and upper-middle-income. Figure 5-1 gives us a strong visual impression of an increasing relationship between female primary school enrollment and contraception use. Both these indicators are positively associated with economic category.

A scatterplot such as Figure 5-1 can be confusing if there are many overlapping plotting symbols. An alternative is to construct separate plots of the two quantitative variables for each level of the qualitative variable. This is done in Figure 5-2 for economic category and two variables related to life expectancy.

Figure 5-2 shows scatterplots of the difference between female and male life expectancy (vertical axis) and overall life expectancy (horizontal axis) for each of the six economic categories. The differences between female and male life expectancies over the six graphs range from -2 (life expectancy 2 years longer for males than females) to 9 (life expectancy 9 years longer for females than males). Overall life expectancies range from 40 to 78 years. All scatterplots have the same scales, making visual comparisons easier. Both plotted variables increase with economic category. Also, the relationship between the two plotted variables depends on economic category.

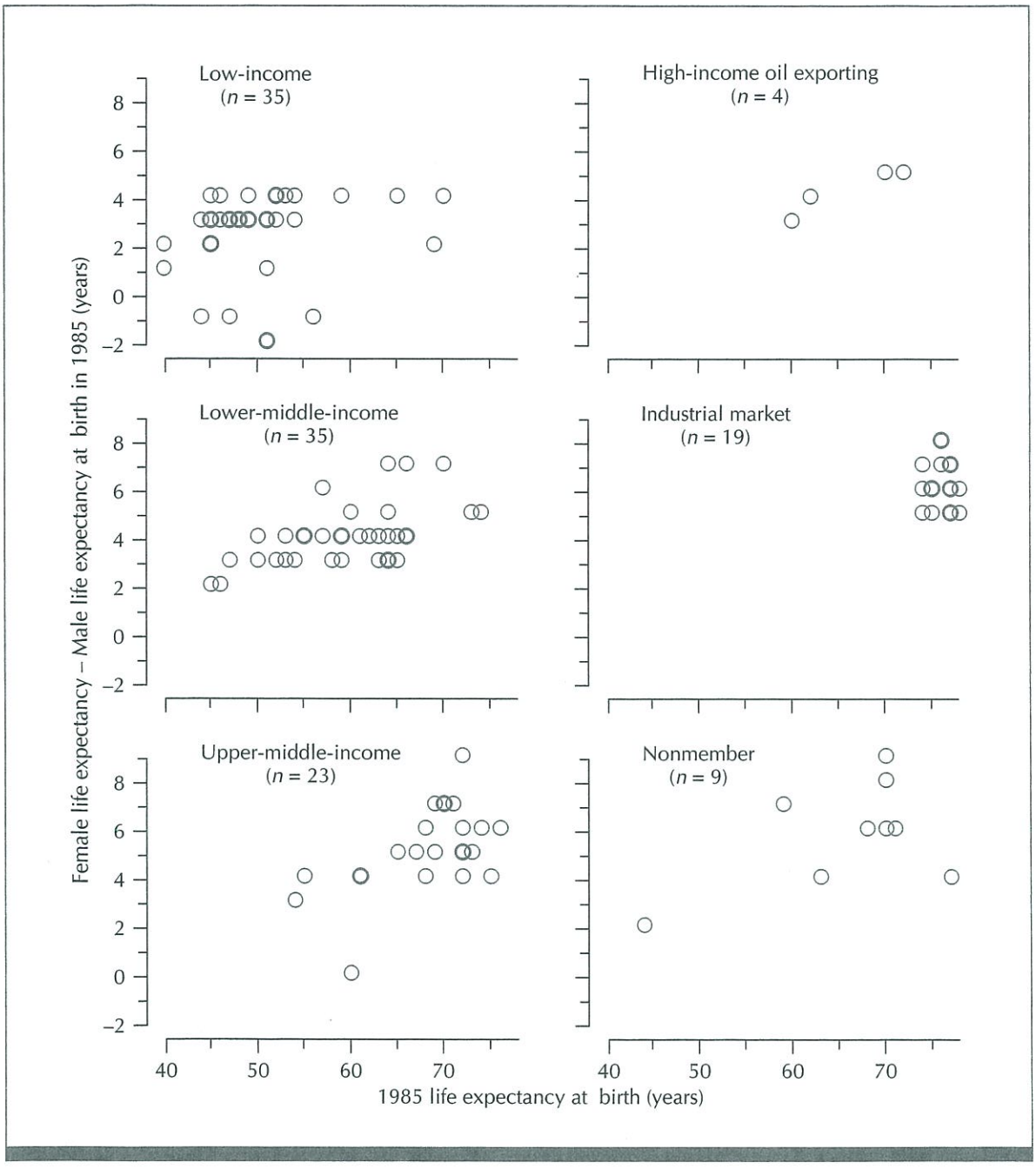


FIGURE 5-2 Scatterplots of the difference between female and male life expectancy at birth and overall life expectancy at birth in 1985 for each economic category. Two low-income countries and one lower-middle-income country are excluded because of missing information on life expectancy.

For low-income countries, life expectancies tend to be short and the differences between female and male life expectancies are relatively small. Among the lower- and upper-middle-income nations, there is more variation in life expectancies. There is also a striking positive relationship between overall life expectancy and the difference between female and male life expectancies. Countries with greater economic development have longer life expectancies and greater differences between female and male life expectancies. However, an increasing relationship between the two plotted variables is not apparent among nations with life expectancies greater than 70 years. (We might hope that the difference between female and male life expectancies will decline with continued economic development!)

In Section 5-3, we group several scatterplots into a scatterplot matrix.

5-3 The Scatterplot Matrix for Several Quantitative Variables

We saw a positive association between female primary school enrollment and contraception use in Figure 5-1. Suppose we would like to consider, in addition, birth rate and calorie supply. We can construct a scatterplot for each pair of variables. If we arrange these plots as in Figure 5-3, we have what is known as a *scatterplot matrix* (Cleveland, 1985; Chambers et al., 1983).

> A **scatterplot matrix** displays scatterplots for pairs of quantitative variables. In the upper right-hand portion are scatterplots for each pair of variables. For each scatterplot in the upper right, there is a corresponding scatterplot in the lower left-hand portion, in which the same two variables are plotted, on opposite axes. The variable names are shown in the body of the scatterplot matrix.

There are six possible pairs of the four indicators. The upper right-hand portion of Figure 5-3 shows six scatterplots resulting from these six pairings. For each plot in the upper right of Figure 5-3, there is a corresponding graph in the lower left-hand portion, with the same two variables plotted on opposite axes. We show variable names in the body of the scatterplot matrix. Scales for the axes lie outside the matrix. Each axis approximately spans the range of values for the corresponding variable.

The 89 World Bank countries with nonmissing values for each of the four indicators are represented in Figure 5-3. Among these 89 nations, female primary school enrollment ranges from a little more than 10% to almost 130%. Contraception use ranges from 0 to over 80%. Birth rate extends from just over 5 to just under 55 births per 1,000 population. Calorie supply per capita ranges from about 1,500 to 3,800 calories per day.

Perhaps the most striking relationship is the strong negative association between contraception use and birth rate. Both contraception use and birth rate are strongly associated with daily calorie supply per capita. Countries with lower calorie supplies tend to have lower levels of contraception use and higher birth rates.

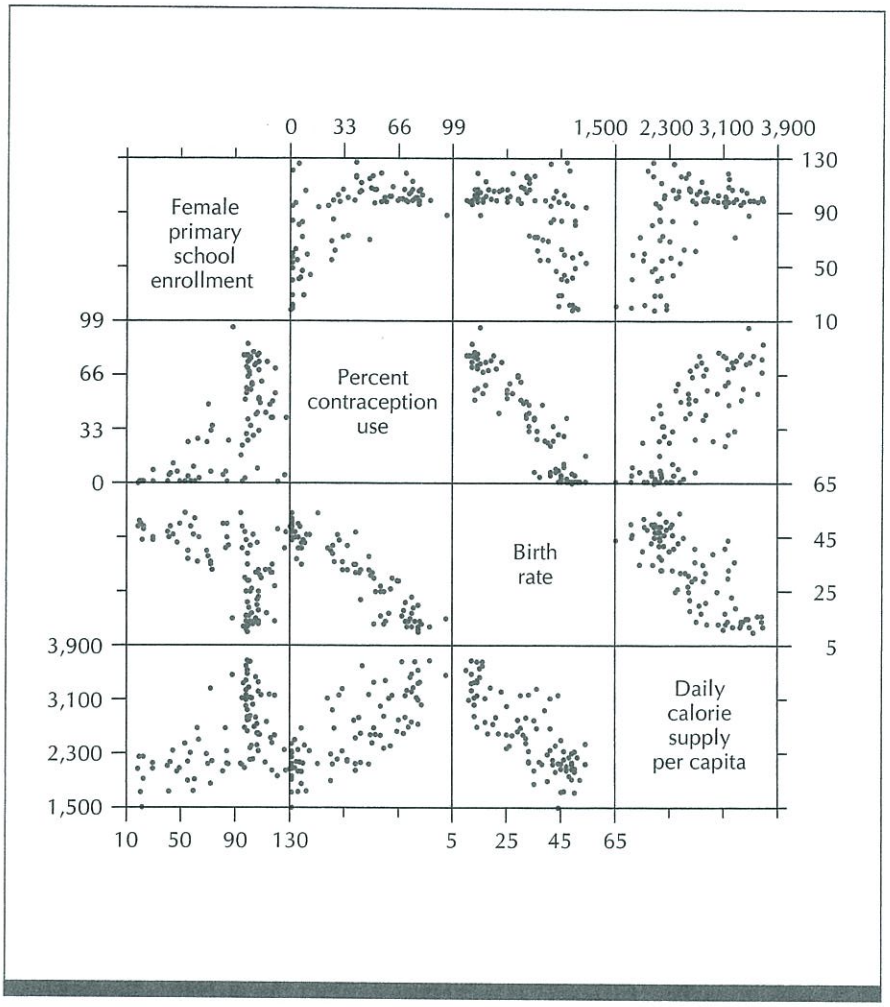


FIGURE 5-3 Scatterplot matrix of number of females enrolled in primary school in 1984 as percentage of 6–11-year age group, percentage of married women of childbearing age using contraception in 1984, birth rate per 1,000 population in 1985, and daily calorie supply per capita in 1985. Plots are based on the 89 World Bank countries with nonmissing information on all four indicators. Thirty-nine countries are excluded because of missing values on at least one of the four indicators: 13 low-income, 4 lower-middle-income, 6 upper-middle-income, all 4 high-income oil exporters, 7 industrial market, and 5 nonmember countries.

Female primary school enrollment has an increasing relationship with contraception use and calorie supply, and a decreasing relationship with birth rate. Countries with lower female primary school enrollments tend to have lower contraception use, higher birth rates, and lower calorie supplies per capita. Among countries with female primary school enrollments above 90%, there is a great deal of variation in values of the other three variables.

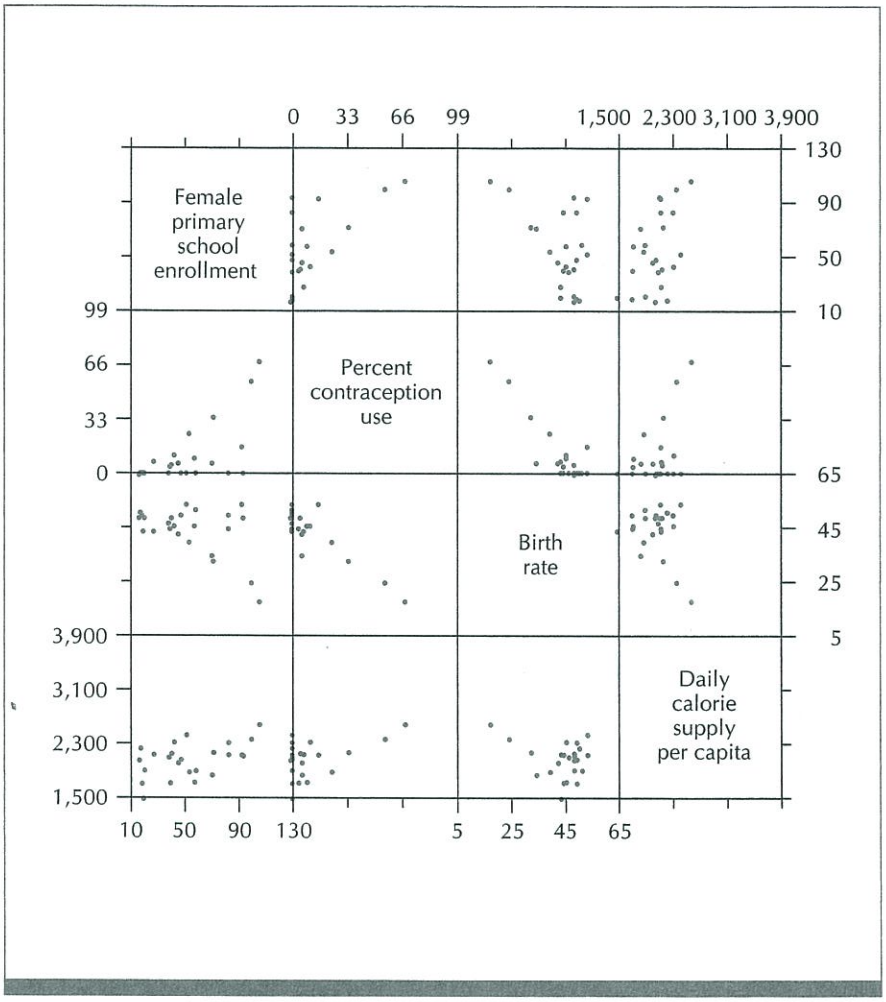


FIGURE 5-4a Scatterplot matrix based on 24 low-income countries with nonmissing information on all four indicators. Thirteen countries are excluded because of missing values.

Now let's look at a separate scatterplot matrix for each of four economic categories: low-income, lower-middle income, upper-middle income, and industrial market. We use the same scales as in Figure 5-3, for easier comparisons.

In Figure 5-4a, two countries have relatively high levels of contraception use and low birth rates. These two countries are China and Sri Lanka, as shown in Table 5-2. Excluding these two nations, we see among the remaining low-income countries low levels of contraception use and calorie supply, high birth rates, and generally low female primary school enrollments.

Figure 5-4d reveals very little variation in values of the four indicators

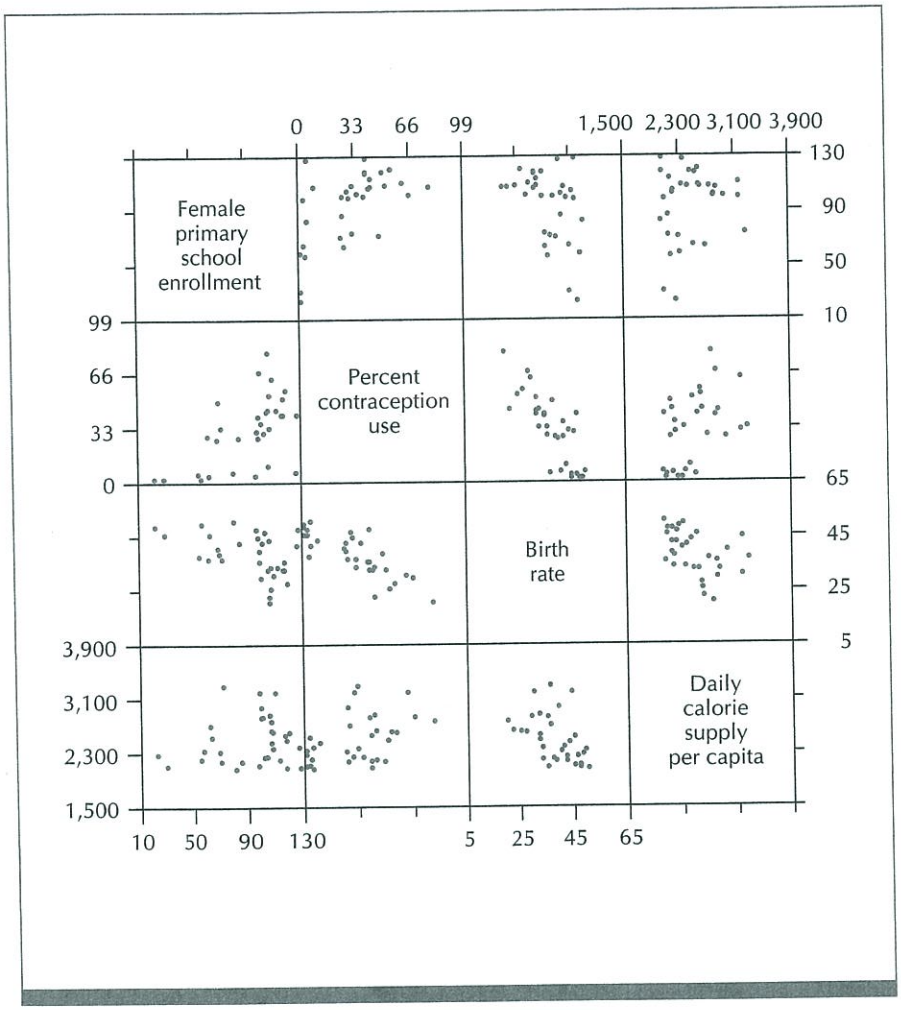


FIGURE 5-4b Scatterplot matrix based on 32 lower-middle-income countries with nonmissing information on all four indicators. Four countries are excluded because of missing values.

TABLE 5-2 Number of females enrolled in primary school in 1984 as percentage of 6–11-year age group, percentage of married women of childbearing age using contraception in 1984, birth rate per 1,000 population in 1985, and daily calorie supply per capita in 1985 for two low-income countries

Country	Female primary school enrollment	Percent contraception use	Birth rate	Calorie supply
China	107	69	18	2,602
Sri Lanka	101	57	25	2,385

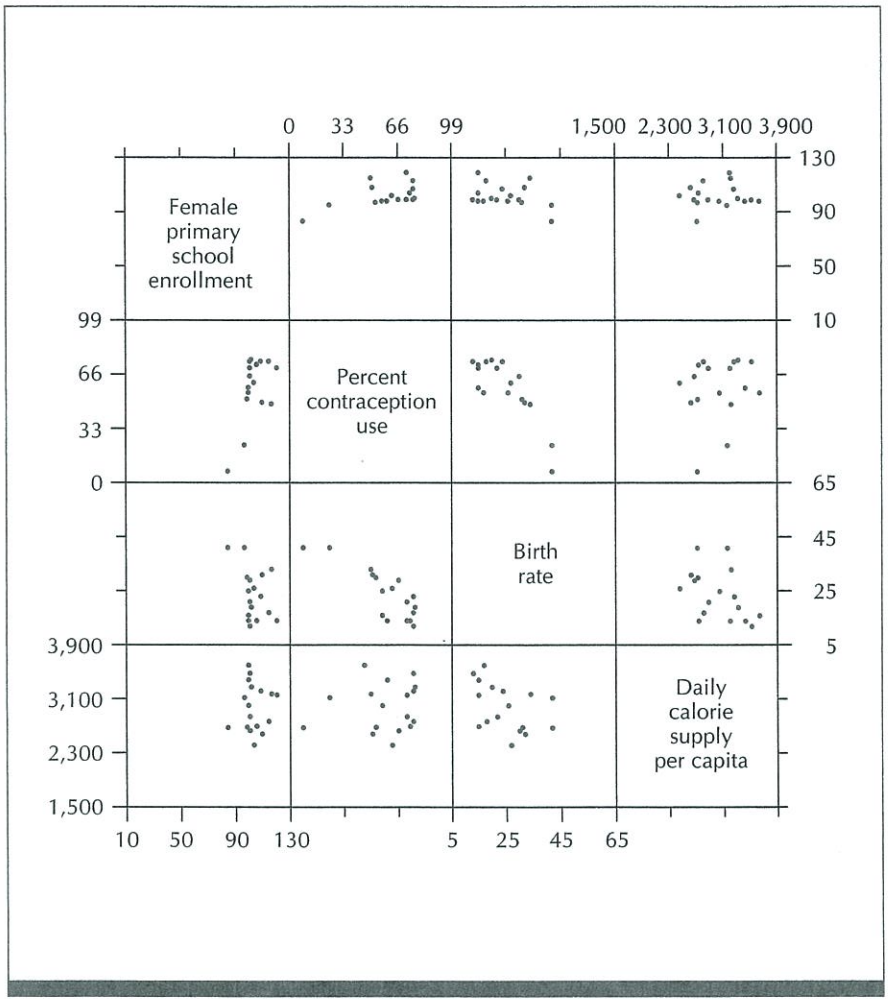


FIGURE 5-4c Scatterplot matrix based on 17 upper-middle-income countries with nonmissing information on all four indicators. Six countries are excluded because of missing values.

among the industrial market countries. These 12 industrial market nations have high female primary school enrollments and daily calorie supplies per capita, relatively high levels of contraception use, and low birth rates.

A comparison of the scatterplot matrices in Figures 5-4a–d shows that the middle-income countries lie between the low-income and industrial market nations as far as these four indicators are concerned. The lower-middle-income countries are closer to the low-income group. The upper-middle-income countries are closer to the industrial market group.

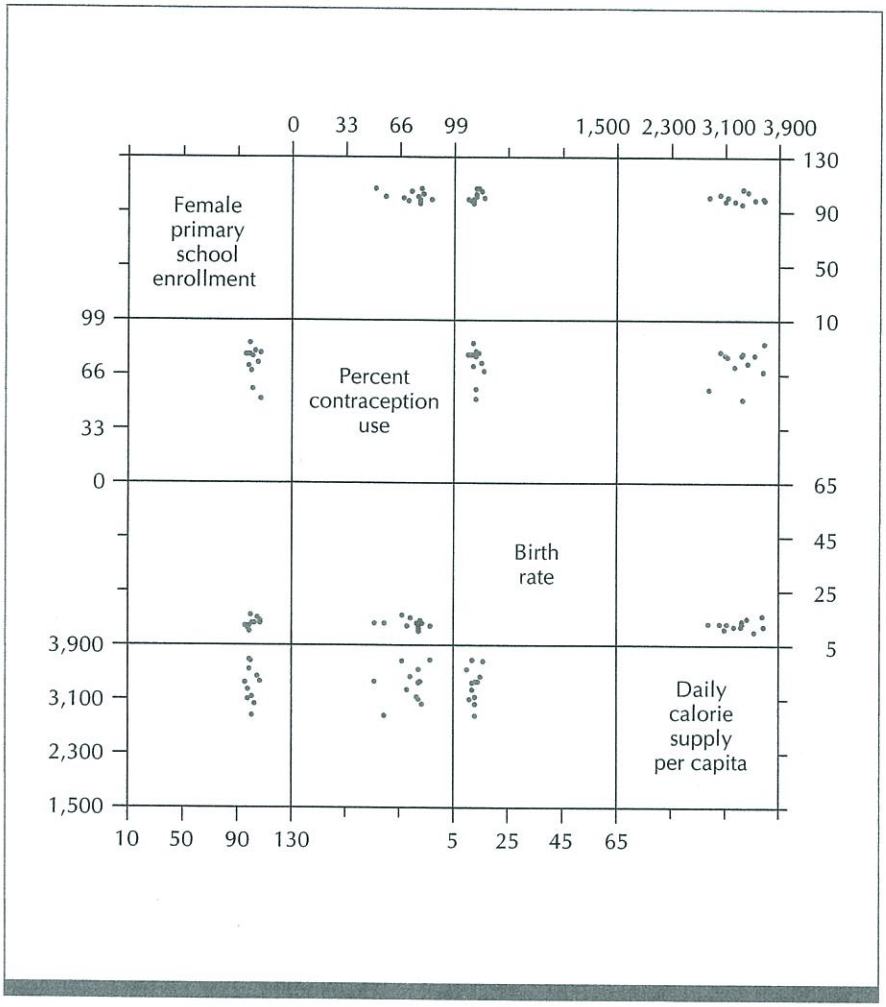


FIGURE 5-4d Scatterplot matrix based on 12 industrial market countries with non-missing information on all four indicators. Seven countries are excluded because of missing values.

we see a strong negative association between contraception use and birth rate. Female primary school enrollment and calorie supply are positively associated with contraception use and negatively associated with birth rate.

In Figure 5-4c, two countries have relatively low levels of contraception use and high birth rates compared with the other upper-middle-income countries represented in the plot. These two countries are Algeria and Iran. Even with these two nations excluded, we still see a decreasing relationship between contraception use and birth rate. There is little association for any other pair of indicators.

All four indicators in Figures 5-3 and 5-4 are related to economic category. Birth rates decrease with economic category. Female primary school enrollment, contraception use, and calorie supply all increase with economic category.

Figure 5-3 is based on the 89 World Bank countries with nonmissing information on all four indicators. Eighty-five countries are included in Figure 5-4. The four extra countries in Figure 5-3 are in the nonmember economic category, which is not included in Figure 5-4. We must be careful when comparing plots, or other tools of data analysis, when varying numbers of cases are included. For this reason we included in each scatterplot matrix only countries with nonmissing information on each of the indicators. Note that in all our figures and tables, we document numbers of countries included and excluded.

5-4 Displaying a Quantitative Variable by Geographic Location: Framed Rectangles on a Map

We often find it useful to convey information related to geographic location by means of a map. In Chapter 1, we displayed a map of the world in Figure 1-1 showing the location and economic category of the World Bank countries. We can think of geographic location as represented by two quantitative variables (corresponding to longitude and latitude, for example.) Figure 1-1 was thus a graphical representation of four variables at a time: the qualitative variable country name, the qualitative variable economic category, and the two quantitative variables corresponding to geographic location.

In this section we discuss how to display a quantitative variable by geographic location. We use framed rectangles on a map to see how calorie supply is related to geographic location among countries on the African continent.

> A quantitative variable may be displayed by geographic location using **framed rectangles on a map**.
>
> A **framed rectangle** is a shaded rectangle within a rectangular frame. The lower and upper ends of the frame define lower and upper bounds for the quantitative variable. Tick marks placed outside the frame mark the halfway point between these extremes. The value of the variable is indicated by the height of the shaded rectangle.

We use framed rectangles on a map in Figure 5-5 to display calorie supply for 39 African nations. We can compare calorie supplies across countries by the heights of the shaded rectangles; the frame makes visual comparisons easy. Framed rectangles allow easier interpretations than the shading commonly used on maps, since gradations in shading are hard to differentiate visually. Also, with shading, larger geographic regions tend to make a stronger visual impact than smaller ones. Both of these problems are avoided when we use framed rectangles on a map. Refer to Cleveland (1985) for further discussion of maps for display of quantitative information by geographic location.

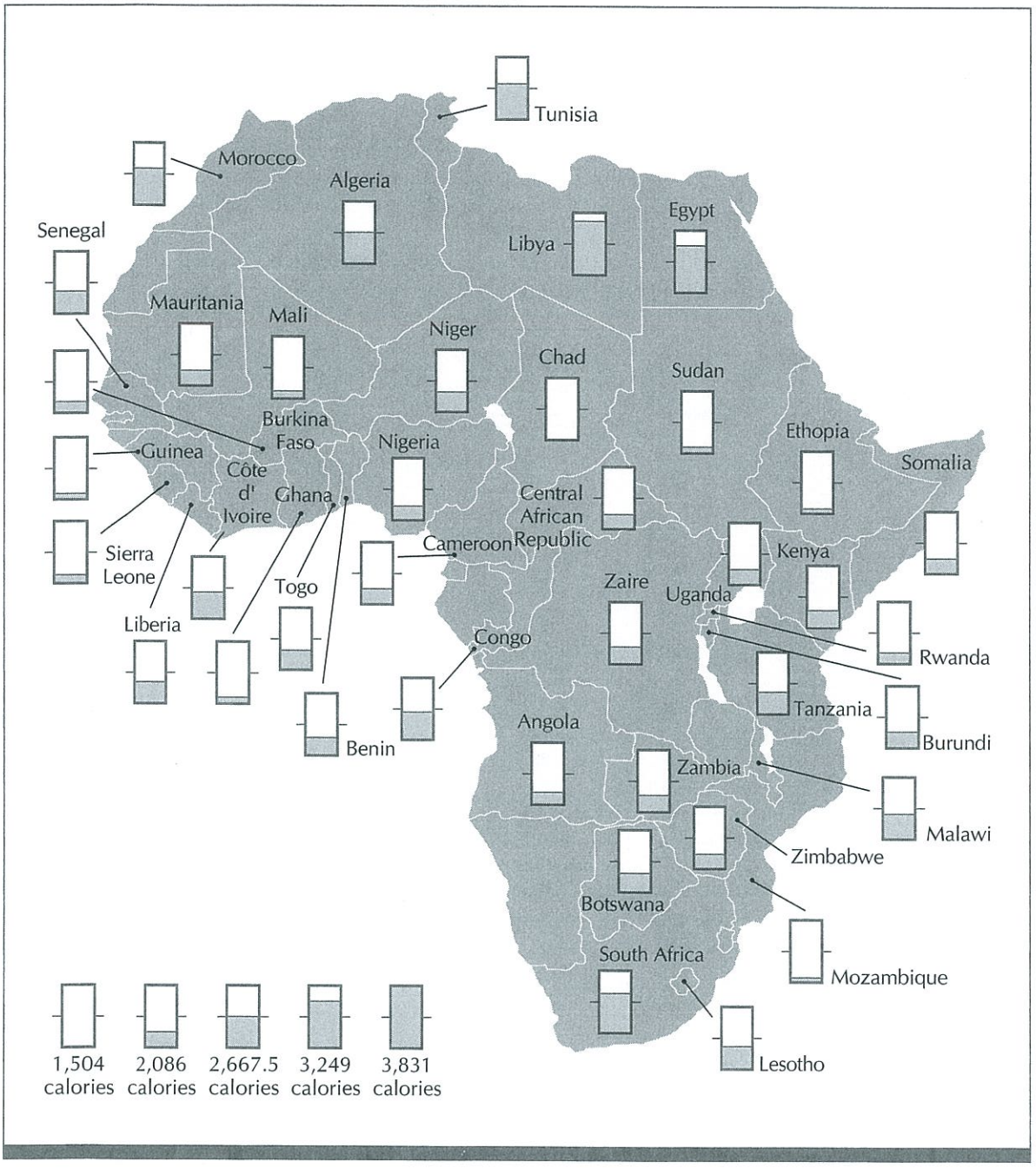


FIGURE 5-5 Framed rectangles on a map of Africa display daily calorie supply per capita in 1985 for 39 African nations. The top of each frame corresponds to the maximum (3,831 calories) and the bottom of each frame to the minimum (1,504 calories) daily calorie supply per capita in 1985 among the 124 World Bank countries with non-missing values.

The bottom of each frame on the map in Figure 5-5 corresponds to the minimum daily calorie supply per capita in 1985 among all 124 World Bank countries with information on this indicator (1,504 calories). The top of each frame corresponds to the maximum daily calorie supply per capita in 1985 among these 124 countries (3,831 calories). The tick marks on the outsides of the frames are halfway between these two values (2,667.5 calories). For reference, the median daily calorie supply per capita in 1985 among the 124 World Bank countries is 2,594 calories, slightly below the tick marks. The mean is 2,675 calories, just above the tick marks. We could have chosen different extremes for the frames. The ones we have used allow comparisons of the African countries with the minimum and maximum over 124 nations.

Looking at the map, we see immediately that Chad has a calorie supply equal to the minimum among World Bank countries. Two neighboring Saharan countries, Sudan and Ethiopia, have calorie supplies very close to this minimum. Most of the 39 countries have low values for calorie supply. Only six countries have calorie supplies above the median (2,594 calories) of 124 World Bank countries. These are the five northernmost countries (Morocco, Algeria, Tunisia, Libya, and Egypt) and the southernmost country (South Africa).

The map in Figure 5-5 gives us a strong overall impression of the relationship between geographic location and calorie supply among these 39 African nations. With framed rectangles, a lot of information is provided in a concise and visually effective fashion. (You might construct a map of Africa, using different shadings to represent relative calorie supplies. Compare the visual effectiveness of your shaded map with the map in Figure 5-5.)

5-5 Effective Graphs

An effective graph summarizes quantitative information in a way that aids visual interpretations. The map in Figure 5-5, for example, helps us see geographical trends in calorie supply for 39 African nations. No extraneous elements that might interfere with visual interpretations are included in the map.

In his book *The Visual Display of Quantitative Information,* Edward Tufte makes a number of suggestions for construction of effective graphical displays. One of these is to maximize the space used for presentation of data and minimize space that is either empty or filled with nonessential elements (such as unnecessary lines, dots, or words). We have tried to follow this advice in graphical displays contained in this book. In scatterplots, for example, the axis for each variable in most cases spans the range of values for the variable. Plotted points then fill the graph as much as possible, minimizing uninformative blank space.

Some types of graphical displays are preferable to others because they are easier to interpret. A dot chart allows more accurate visual comparisons of percentages or proportions than does a pie chart, for example. For excellent

discussions of principles of graphing data, see Edward Tufte's *The Visual Display of Quantitative Information* (Tufte, 1983), Tufte's *Envisioning Information* (Tufte, 1990), and William Cleveland's *The Elements of Graphing Data* (Cleveland, 1985).

We now leave data analysis for a while. In Chapters 6, 7, and 8, we discuss probability. Then in Part III, we will use data analysis and probability as we consider ideas in experimental design and statistical inference.

Summary of Chapter 5

Multiway frequency tables (multidimensional contingency tables) allow us to look at relationships among several qualitative variables. Scatterplots of two quantitative variables at each level of a qualitative variable can be useful for studying the relationship between the two quantitative variables within levels of the qualitative variable. The scatterplot matrix provides a way to display several scatterplots within a single figure. Framed rectangles on a map effectively illustrate how a quantitative variable depends on geographic location; framed rectangles are easier than shadings to interpret visually. Graphical displays should be designed to allow easy visual interpretation, with maximum use of data and minimal use of blank space and extraneous material.

Minitab Appendix for Chapter 5

Creating Multiway Frequency Tables

To produce multidimensional frequency tables, we use the TABLE command. The TABLE command can be followed by up to 10 classification variables. (A classification variable takes integer values between −9999 and +9999 or missing values.) Minitab uses the first variable for rows of the table and the second variable for columns. Minitab prints a separate table for each combination of values for any other variables listed. Consider the example based on Exercise 5-35. In the Minitab Appendix for Chapter 1, we created two classification variables, CALCODE and MORTCODE. We will create another classification variable based on WEIGHT, using the CODE command:

```
MTB> code (0:5)1 (6:30)2 'weight' c30
MTB> name c30 'wtcode'
```

The variable WTCODE in column 30 equals 1 for countries with values of WEIGHT less than or equal to 5, equals 2 for countries with values of WEIGHT greater than or equal to 6. The TABLE command

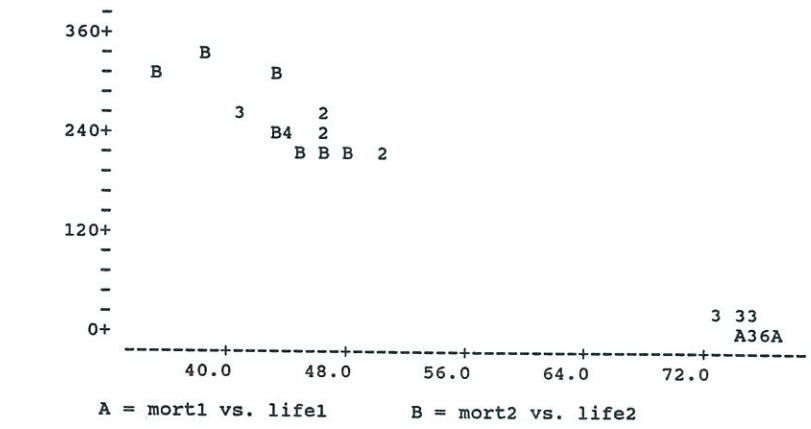


FIGURE M5-2 Scatterplot of child mortality versus life expectancy, with different plotting symbols for low- and high-child-mortality countries, using the MPLOT command

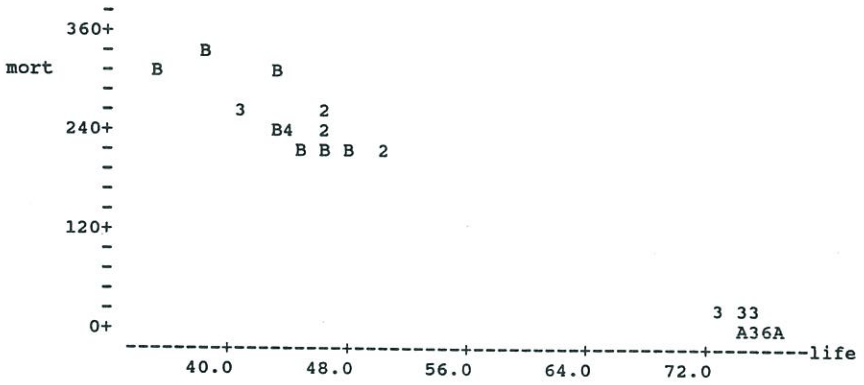


FIGURE M5-3 Scatterplot of child mortality versus life expectancy, with different plotting symbols for low- and high-child-mortality countries, using the LPLOT command

Creating a Scatterplot Matrix

Minitab will not produce a scatterplot matrix directly (some statistical packages will). To construct a scatterplot matrix, we can use the PLOT command to produce scatterplots, controlling the scales using the YSTART and XSTART subcommands. We can then assemble these plots into a scatterplot matrix by hand or by using a word-processing program.

We cannot use Minitab to produce a map with framed rectangles.

Exercises for Chapter 5

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 5-1

How does survival of mice exposed to DDT and urethane compare with survival of unexposed mice? The accompanying table summarizes results of a study designed to address this question (Breslow, 1988; *J. Nat. Cancer Inst.*, volume 52, 1974, pages 233–239). The table shows the percentage of mice alive at five times, by sex and exposure groups.

	Control	2 ppm DDT	10 ppm DDT	50 ppm DDT	250 ppm DDT	Urethane
			Males			
Number of animals	348	362	367	396	372	315
Percentage alive at						
0 weeks	100	100	100	100	100	100
70 weeks	80.7	83.7	81.7	84.6	72.3	72.1
90 weeks	63.5	57.2	57.2	52.8	28.5	39.7
110 weeks	32.5	22.4	25.1	21.5	1.3	8.9
130 weeks	13.8	3.0	8.2	3.5	0.0	0.1
			Females			
Number of animals	363	354	370	349	334	248
Percentage alive at						
0 weeks	100	100	100	100	100	100
70 weeks	79.1	77.4	83.8	80.2	66.5	72.2
90 weeks	60.0	55.1	64.9	56.4	41.3	40.7
110 weeks	35.0	23.7	42.4	26.9	9.9	12.1
130 weeks	14.0	5.6	15.4	6.3	1.5	1.2

a. For the male mice, plot percentage alive versus time for each treatment group. You may want to use different plotting symbols for each treatment group on the same graph. Or you may wish to plot a separate graph for each treatment group.

b. Repeat part (a) for female mice.

c. Separately for males and females, compare survival across treatment groups.

d. Separately for each treatment, compare survival of males and females.

e. Discuss all of your results.

EXERCISE 5-2

When scientists say that two treatments have a synergistic effect, they mean that the effect of the two treatments taken together is greater than the sum of their separate effects. In this experiment, researchers explored possible synergistic antitumor effects of two agents that stimulate the activity of an animal's immune

system. The experimenters injected mice with lymphoma cells and then treated them with interferon or monoclonal antibody or a combination of the two agents. The number of mice surviving 90 days after injection with the tumor cells is shown below, along with the number of animals treated with each treatment combination (Piegorsch, Weinberg, and Margolin, 1988; Basham et al., 1986). The table shows number of animals alive at 90 days/number of animals treated.

Interferon (units/ mouse/ day)	Monoclonal antibody (units/mouse/day)				
	0	.1	1	10	100
0	0/10	0/10	3/10	2/10	9/40
10^4	1/40	1/10	6/10	7/10	8/10

Do the experimental results suggest a synergistic antitumor effect of these two forms of immunotherapy? Graph the data in any way that helps to answer this question.

EXERCISE 5-3

Consider the following information on body weight and kidney weight for 25 normal mice and 9 diabetic mice, from a study by Dr. E. Jones of the Children's Cancer Research Foundation in Boston, Massachusetts (Hill and Padmanabhan, 1984; from Bishop, 1973).

Body weight (g)	Kidney weight (mg)	Body weight (g)	Kidney weight (mg)
	Normal Mice		
34	810	37	780
43	480	38	660
35	680	32	750
33	920	36	780
34	650	32	670
26	650	32	670
30	650	38	700
31	560	42	720
31	620	36	800
27	740	44	830
28	600	33	640
27	640	38	800
30	690		
	Diabetic Mice		
42	1,030	46	1,100
44	1,240	34	1,040
38	1,150	44	1,080
52	1,280	38	870
48	1,240		

a. Construct two dot plots of body weight—one for the normal mice and one for the diabetic mice. Use the same scale for each plot. Describe and compare the two distributions.

b. Repeat part (a) for kidney weight.

c. Plot kidney weight versus body weight for the normal mice. Using the same scales, plot kidney weight versus body weight for the diabetic mice.

d. Is the relationship between body weight and kidney weight the same for the normal and diabetic mice? Discuss your findings.

EXERCISE 5-4

The accompanying table shows forearm tremor frequency (in Hz) for each of five weights (in pounds) applied at the wrist, for six volunteers (Hollander and Wolfe, 1973, page 175; based on Fox and Randall, 1970). Each value is the average of five measurements.

Volunteer	0 lb	1.25 lb	2.5 lb	5 lb	7.5 lb
1	3.01	2.85	2.62	2.63	2.58
2	3.47	3.43	3.15	2.83	2.70
3	3.35	3.14	3.02	2.71	2.78
4	3.10	2.86	2.58	2.49	2.36
5	3.41	3.32	3.08	2.96	2.67
6	3.07	3.06	2.85	2.50	2.43

Construct one or more scatterplots showing forearm tremor frequency versus weight applied to determine whether the relationship between the two variables is the same for each volunteer. Discuss your findings.

EXERCISE 5-5

Percent minority enrollment in grades kindergarten through 12 in 1980, percent minority teachers in 1982, and percent minority new hires in 1982 are shown below for 17 states with high minority enrollments (American Council on Education, 1987, page 27).

State	Percent minority enrollment in public schools in 1980	Percent minority teachers in 1982	Percent minority new hires in 1982
Alabama	33.6	27.0	13.0
Arizona	33.7	11.6	10.9
Arkansas	23.5	19.4	12.8
California	42.9	18.9	25.2
Delaware	28.8	16.8	9.0
Florida	32.2	21.3	11.8
Georgia	34.3	26.4	14.7
Illinois	28.6	6.5	4.8
Louisiana	43.4	30.7	9.1

State	Percent minority enrollment in public schools in 1980	Percent minority teachers in 1982	Percent minority new hires in 1982
Maryland	33.5	26.5	10.4
Mississippi	51.6	38.3	19.8
New Mexico	57.0	28.3	25.4
New York	32.0	9.9	17.8
North Carolina	31.9	21.6	16.6
South Carolina	43.5	25.5	16.7
Texas	45.9	24.0	18.1
Virginia	27.5	20.3	14.7

a. Construct a scatterplot matrix with these three variables. Discuss your findings.

b. Use framed rectangles on a map of the United States to investigate the relationship between these variables and geography. Discuss your findings.

EXERCISE 5-6

Engine displacement, city and expressway gasoline usage, and weight are shown here for ten cars weighing 1,000 kg or less (Ramsay, 1988; from 1986 *Consumer Reports*).

Car	Engine displacement (liters)	City gas (liters/100 km)	Expressway gas (liters/100 km)	Weight (100 kg)
Chevrolet Chevette	1.6	13.3	6.8	10.0
Chevrolet Spectrum	1.5	10.1	5.3	8.7
Dodge Colt	1.5	11.0	5.6	9.9
Dodge Omni	1.6	11.5	5.6	9.5
Honda Civic	1.5	11.5	6.5	9.2
Nissan Sentra	1.6	10.5	5.6	9.5
Renault Alliance	1.4	12.0	5.6	9.1
Toyota Tercel	1.5	11.0	5.5	9.7
Honda Civic CRX	1.5	9.4	5.6	9.0
Nissan Pulsar NX	1.6	9.7	5.1	9.2

Construct a scatterplot matrix using these four variables. What observations can you make about relationships between pairs of variables for these small cars?

EXERCISE 5-7

Estimated radiocarbon dates ($\pm$ 1 standard deviation) are shown by depth of samples taken from sites of two different archeological digs in Tasmania (Cosgrove, 1989). The units for estimated radiocarbon dates are years before A.D. 1950. In this reference, the term *standard deviation* indicates an estimate of errors in counting of the modern radiocarbon standard, background and sample.

Shannon River Valley		Bluff Cave, Florentine River Valley	
Depth (cm)	Estimated radiocarbon date ± 1 SD	Depth (cm)	Estimated radiocarbon date ± 1 SD
5	2,450 ± 70	5	11,630 ± 200
15	10,440 ± 160	10	13,100 ± 110
22	18,480 ± 200	15	13,830 ± 220
23	17,660 ± 250	20	16,120 ± 180
25	19,080 ± 280	30	21,410 ± 240
45	30,840 ± 480	35	24,190 ± 410
50	16,200 ± 590	42	27,770 ± 420
		50	28,000 ± 720
		53	23,640 ± 310
		55	30,750 ± 1,340
		60	30,420 ± 690

a. Construct a scatterplot of estimated radiocarbon date versus depth of the sample for each of the two sites. Use the same scales for each plot.

b. Is the relationship between depth of the sample and radiocarbon age the same for the two sites? Discuss the information provided in your two plots.

EXERCISE 5-8

Investigators studied eight hot springs in the Cascade Range in north central Oregon (Ingebritsen, Sherrod, and Mariner, 1989). They measured discharge temperature (°C) and concentration (mg/liter) of calcium (Ca), sodium (Na), and chlorine (Cl) in hot springs water at each site.

Hot spring	Discharge temperature	Ca	Na	Cl
Austin	86	35	305	390
Bagby	58	3.3	53	14
Breitenbush	84	95	745	1,200
Bigelow	59	195	675	1,250
Belknap	73	210	660	1,200
Foley	79	510	555	1,350
Kahneeta	83	13	400	240
Unnamed (on Rider Creek)	46	215	405	790

a. Construct a dot plot of each of the four variables. Describe each distribution in terms of location, variation, concentrations of values, and symmetry or skewness.

b. Construct a scatterplot matrix using these four variables. Discuss the relationships between pairs of variables you see in the plots.

EXERCISE 5-9

As part of a study of movements and survival of black ducks, U.S. Fish and Wildlife Service workers captured and examined 50 female black ducks in

November and December 1983 (Pollock, Winterstein, and Conroy, 1989). Thirty-one were hatch-year ducks, born during the previous breeding season. The other 19 were after-hatch-year ducks, at least 1 year old. The workers recorded body weight and wing length for each duck. They also calculated a condition index (body weight divided by wing length) for each duck.

Weight (gm)	Wing length (mm)	Condition index (gm/mm)	Weight (gm)	Wing length (mm)	Condition index (gm/mm)
		Hatch-year ducks			
1,140	266	4.29	1,070	267	4.01
1,160	264	4.39	1,270	276	4.60
1,120	262	4.27	1,080	260	4.15
1,070	268	3.99	1,150	271	4.24
940	252	3.73	1,030	265	3.89
1,240	271	4.58	1,160	275	4.22
1,120	265	4.23	1,180	263	4.49
1,010	272	3.71	1,050	271	3.87
1,040	270	3.85	1,050	275	3.82
1,200	276	4.35	1,160	266	4.36
1,280	270	4.74	1,150	263	4.37
1,250	272	4.59	1,220	268	4.55
1,090	275	3.96	1,140	262	4.35
1,040	255	4.08	1,140	270	4.22
1,130	268	4.22	1,120	274	4.09
1,180	259	4.56			
		After-hatch-year ducks			
1,160	277	4.19	1,250	276	4.53
1,260	280	4.50	1,050	275	3.82
1,080	267	4.04	1,320	285	4.63
1,140	277	4.11	1,260	269	4.68
1,200	283	4.24	1,110	270	4.11
1,100	264	4.17	1,280	281	4.55
1,420	270	5.26	1,270	270	4.70
1,120	272	4.12	1,370	275	4.98
1,110	271	4.10	1,220	265	4.60
1,340	275	4.87			

a. For the variable body weight, construct two box plots, one for the hatch-year ducks and another for the after-hatch-year ducks. Display these two box plots on the same graph to allow comparison between the two groups of ducks. In the same way, construct box plot displays to compare wing length and condition index between the two groups of ducks.

b. For each variable, discuss the information provided in the plots for each age group. Use the plots to compare the two age groups.

c. For the 31 hatch-year ducks, construct a scatterplot matrix using the three variables body weight, wing length, and condition index. Discuss the relationships between pairs of variables revealed in the plots.

d. Repeat part (c) for the 19 after-hatch-year ducks.

e. Compare the scatterplot matrices for the two age groups. Are relationships between pairs of variables different for the two groups?

EXERCISE 5-10

Researchers measured crying activity in 38 4–7-day-old babies, 20 females and 18 males. They tested these children 3 years later to measure speech and intellectual development. Sex, infant cry count (units not given), and 3-year Stanford–Binet IQ score are shown below for each child (Hollander and Proschan, 1984, pages 150–151; from Karelitz et al., 1964).

Girls		Boys	
Cry count	IQ score	Cry count	IQ score
10	87	20	90
12	94	17	94
16	100	12	97
19	103	12	103
14	106	9	103
10	109	23	103
15	112	13	104
15	114	16	106
9	119	27	108
12	119	18	109
19	120	18	109
16	124	18	112
20	132	23	113
15	133	21	114
22	135	16	118
31	135	12	120
16	136	17	141
22	157	30	155
33	159		
13	162		

a. Construct a stem-and-leaf plot of the girls' cry counts. Use tens as stems and units as leaves, two rows per stem. Construct a similar plot for the boys. Describe these two distributions of cry counts. Compare the distributions for the girls and boys.

b. Construct a stem-and-leaf plot of the girls' IQ scores. Use the units as leaves, other digits as stems. Construct a similar plot for the boys. Describe the two distributions of IQ scores. Compare the distributions for the girls and boys.

c. For the girls, construct a scatterplot of IQ score versus cry count. Discuss the relationship between the two variables shown in the plot.

d. Construct a scatterplot of IQ score versus cry count for the boys. Use the same scales as for the girls. Describe the information provided in the plot.

e. Compare the two scatterplots. Is the relationship between infant cry count and age-3 IQ score the same for the girls and the boys?

EXERCISE 5-11

Researchers recorded obesity and blood pressure for each person in a random sample of 58 Mexican–American women and a random sample of 44 Mexican–American men aged 35–60 years in a small California town. (A sample is a random sample if each member of the population had an equal and independent chance of being included in the sample.) Obesity is recorded as actual weight divided by ideal weight (based on New York Metropolitan life tables). Blood pressure (BP) is systolic blood pressure in millimeters of mercury (mm Hg). The results are shown here (from a study by J. W. Farquhar and associates discussed in Hollander and Proschan, 1984, pages 147 and 150).

Obesity	BP	Obesity	BP	Obesity	BP
		Women			
1.50	140	1.59	150	1.43	130
1.63	132	2.39	150	1.50	112
.92	138	1.17	116	1.33	124
1.09	112	1.24	116	1.44	110
1.23	160	1.50	140	1.34	124
2.04	138	1.13	118	1.11	104
1.38	114	1.35	138	1.42	170
1.55	144	1.33	108	1.22	108
1.07	98	.97	112	1.26	100
1.65	120	1.01	118	1.54	130
1.43	128	1.74	128	1.51	118
1.36	110	1.37	148	1.67	162
1.03	128	1.32	108	1.56	116
1.33	104	1.56	122	1.25	98
1.24	110	1.27	118	1.57	116
1.30	118	1.32	138	1.41	142
1.21	124	1.20	120	1.15	118
1.43	122	1.24	112	1.28	126
1.75	138	2.20	136	1.64	136
1.73	208				
		Men			
1.31	130	1.31	148	1.19	146
1.11	122	1.34	140	1.17	146
1.56	132	1.18	110	1.04	124
1.03	150	.88	120	1.29	114
1.26	136	1.16	118	1.32	190
1.37	118	1.25	130	1.48	112
1.58	126	.93	162	1.29	124
1.06	126	1.19	134	.96	110
1.13	118	1.19	110	.81	94
1.11	118	1.29	140	1.29	128
1.28	126	1.20	140	1.02	124
1.09	104	1.08	134	1.04	130
1.14	124	1.13	110	1.16	134
1.57	144	1.07	116	1.04	118
1.37	118	1.26	132		

a. Construct box plots of obesity separately for men and women. Display the two box plots in the same graph to allow comparisons. Describe the distri-

bution of obesity values for men and women. Compare the distributions for men and women.

b. Repeat part (a) for blood pressure.

c. Construct a scatterplot of blood pressure versus obesity for women. Construct a similar scatterplot for men, using the same scales. Describe the relationship between obesity and blood pressure for men and for women. Is the relationship the same for the two sexes?

EXERCISE 5-12

The table shows concentrations (nanograms per milliliter) of the brain metabolite homovanillic acid (HVA) in cerebrospinal fluid, full-scale intelligence quotient (IQ), memory quotient (MQ), and IQ − MQ for nine patients with a disorder known as Korsakoff's psychosis (Dietz, 1989; from McEntee and Mair, 1978).

HVA	IQ	MQ	IQ − MQ
21	89	60	29
23	90	59	31
25	122	102	20
25	87	64	23
26	89	61	28
31	106	79	27
40	104	80	24
48	106	80	26
75	127	88	39

The difference between IQ and MQ is a measure of memory impairment; greater impairment is associated with larger differences.

a. Construct a scatterplot matrix using these four variables.

b. Discuss the relationships between pairs of variables shown in the scatterplot matrix.

EXERCISE 5-13

In this study of energy requirements of grazing Merino wether sheep in Australia, researchers determined outdoor maintenance requirements (in Mcal/sheep/day) by radioassay of urinary CO_2. They carried out four separate experiments, in one location. Animal weights (in kg) and energy requirements are shown for each of these experiments (Wallach and Goffinet, 1987; from Young and Corbett, 1972).

Weight	Requirements	Weight	Requirements	Weight	Requirements	Weight	Requirements
			Experiment 1				
22.1	1.31	30.0	1.23	33.8	1.46	49.2	2.53
25.1	1.46	30.2	1.01	34.3	1.14	51.8	1.87
25.1	1.00	30.2	1.12	34.9	1.00	51.8	1.92
25.7	1.20	33.2	1.25	42.6	1.81	52.5	1.65
25.9	1.36	33.2	1.32	43.7	1.73	52.6	1.70
26.2	1.27	33.2	1.47	44.9	1.93	53.3	2.66
27.0	1.21	33.9	1.03	49.0	1.78		

Weight	Require-ments	Weight	Require-ments	Weight	Require-ments	Weight	Require-ments
	Experiment 2				**Experiment 3**		
23.9	1.37	32.1	1.80	46.7	2.21	28.6	2.13
25.1	1.29	32.6	1.75	37.1	2.11	29.2	1.80
26.7	1.26	33.1	1.82	31.8	1.39	26.2	1.05
27.6	1.39	34.1	1.36	36.1	1.79		
28.4	1.27	34.2	1.59				
28.9	1.74	44.4	2.33		**Experiment 4**		
29.3	1.54	44.6	2.25	45.9	2.36	34.4	1.63
29.7	1.44	52.1	2.67	36.8	2.31	26.4	1.27
31.0	1.47	52.4	2.28	34.4	1.85	27.5	.94
31.0	1.50	52.7	3.15				
31.8	1.60	53.1	2.73				
32.0	1.67	52.6	3.73				

a. Plot energy requirement versus animal weight for each of the four experiments. Use the same scales for each plot.

b. What is the relationship between animal weight and outdoor maintenance energy requirements?

c. Do the four experiments demonstrate the same relationship between animal weight and energy requirements? Discuss your findings.

EXERCISE 5-14

Researchers have studied patterns of recovery of stroke patients over time. Such patterns provide useful baselines for evaluating individual patients. In this study, researchers obtained recovery information on 368 patients who survived at least 8 weeks from initial examination. The number (percentage) of these 368 patients past each of three recovery milestones at the initial examination, as well as 1, 2, 4, 6, and 8 weeks later are shown in the table (Partridge, Johnston, and Edwards, 1987).

Recovery milestone	Initial examination	1 week later	2 weeks later	4 weeks later	6 weeks later	8 weeks later
Maintain sitting balance for two minutes	217 (59.0)	280 (76.1)	315 (85.6)	334 (90.8)	337 (91.6)	338 (91.8)
Stand up to free-standing position	105 (28.5)	158 (42.9)	193 (52.4)	230 (62.5)	243 (66.0)	260 (70.7)
Independent walking inside	53 (14.4)	101 (27.4)	138 (37.5)	166 (45.1)	181 (49.2)	196 (53.3)

a. On the same scatterplot, plot the percentage of patients past the milestone versus time, for each of the three milestones. You may want to use different plotting symbols (or colors) for the three milestones.

b. Discuss the recovery patterns over time for these three milestones.

EXERCISE 5-15

Values of maximal oxygen uptake (in ml · kg^{-1} · min^{-1}) are listed here for world-class athletes in their teens and 20's. (Maximal oxygen uptake is a measure of lung function and capacity for work.) Several male and female athletes were tested in each of four sports (Wilmore, 1984).

Basketball	*Females:*	42.3	42.9	49.6		
	Males:	41.9	45.9	50.0		
Speed skating	*Females:*	52.0	46.1			
	Males:	56.1	72.9	64.6		
Cross-country skiing	*Females:*	61.5	68.2	56.9		
	Males:	63.9	73.9	78.3	73.0	
Distance running	*Females:*	63.2	50.8	57.5		
	Males:	65.5	72.2	77.4	78.1	73.2

Display these values in any way(s) that will allow comparisons of maximal oxygen uptake across sports for each sex, and between males and females for each sport.

EXERCISE 5-16

Age, height, weight, and maximal oxygen capacity are listed below for 13 male world-class distance runners (Wilmore, 1984). Maximal oxygen capacity is a measure of the lungs' capacity for work.

Age (years)	Height (cm)	Weight (kg)	Maximal oxygen capacity (ml · kg^{-1} · min^{-1})
10	144.3	31.9	56.6
26	176.1	64.5	72.2
26	178.9	63.9	77.4
26	177.0	66.2	78.1
27	178.7	64.9	73.2
32	177.3	64.3	70.3
35	174.0	63.1	66.6
36	177.3	69.6	65.1
40–49	180.7	71.6	57.5
55	174.5	63.4	54.4
50–59	174.7	67.2	54.4
60–69	175.7	67.1	51.4
70–75	175.6	66.8	40.0

a. Construct a scatterplot matrix using these four variables. (When age is listed as an interval, use a reasonable value such as the midpoint of the interval in the plots that include age.)

b. Discuss the relationships between pairs of variables that you see in these plots.

EXERCISE 5-17 The number of female mosquitos captured coming to bite in a yard with an electrocuting device and the number killed in the device are listed here for each of five different 2-hour sessions at each of two sites (Nasci, Harris, and Porter, 1983).

Trial	1	2	3	4	5
Site 1					
Electrocuting device	31	44	129	15	11
Human bait	94	146	194	54	39
Site 2					
Electrocuting device	49	151	30	12	17
Human bait	90	172	219	60	21

a. Using the same scales, construct four dot plots: for each site, a plot of mosquito numbers killed in the electrocuting device and a plot of numbers captured by the person.

b. How do the numbers captured vary by site? How do the numbers captured vary between the electrocuting device and human bait?

c. For each site, draw a scatterplot of number of mosquitos captured by the person versus number killed in the electrocuting device. What is the relationship between the two variables for each site? Compare sites.

EXERCISE 5-18 The human immunodeficiency virus (HIV) is the virus associated with the acquired immune deficiency syndrome (AIDS). In this study, investigators tested residents of six African countries for presence of HIV (Kanki et al., 1987). They classified residents into three groups. The risk group included prostitutes and people visiting outpatient clinics for sexually transmitted diseases. The disease group consisted of people with tuberculosis and patients hospitalized in infectious disease or internal medicine wards. The control group included healthy adults from the same geographic regions as the risk and disease groups. The numbers testing positive for HIV antibodies in serum samples are shown below by country and group.

Burkina Faso: 1 positive of 22 tested in the disease group, 45 positive of 340 tested in the risk group, 2 positive of 416 tested in the control group

Ivory Coast: 4 positive of 40 tested in the disease group, 46 positive of 232 tested in the risk group, 38 positive of 1,067 tested in the control group

Guinea: 1 positive of 131 tested in the disease group, 0 positive of 13 tested in the risk group, 2 positive of 314 tested in the control group

Guinea Bassau: 0 positive of 273 tested in the disease group, 0 positive of 39 tested in the risk group, 0 positive of 151 tested in the control group

Senegal: 2 positive of 178 tested in the disease group, 3 positive of 422 tested in the risk group, 0 positive of 426 tested in the control group

Mauritania: 2 positive of 35 tested in the disease group, 0 positive of 9 tested in the risk group, 0 positive of 140 tested in the control group

a. Display these results in one or more frequency tables.

b. Discuss any differences between groups with respect to detection of antibodies to HIV.

c. Discuss any differences among countries.

d. Use framed rectangles on a map of Africa to look for geographic trends. Discuss your findings.

EXERCISE 5-19

In 1975–1976: 499,602 men and 418,786 women received bachelor's degrees in the United States, 165,474 men and 143,789 women received master's degrees, 26,010 men and 7,777 women received doctorates, and 52,365 men and 9,720 women received first professional degrees.

In 1984–1985: 482,528 men and 496,949 women received bachelor's degrees, 143,390 men and 142,861 women received master's degrees, 21,700 men and 11,243 women received doctorates, and 50,455 men and 24,608 women received first professional degrees (American Council on Education, 1987, page 187).

a. Arrange this information in one or more frequency tables.

b. Discuss the relationship between sex and degree, separately for the two academic years. Is the relationship the same for the two academic years?

c. Discuss the relationship between year and degree, separately for men and women. Is the relationship the same for men and women?

EXERCISE 5-20

Of 22,632,000 White 18–24-year-olds in the United States in 1985, 18,916,000 had completed high school and 6,500,000 had enrolled in college. Of 3,716,000 Black 18–24-year-olds, 2,810,000 had completed high school and 734,000 had enrolled in college. Of 2,221,000 Hispanic 18–24-year-olds, 1,396,000 had completed high school and 375,000 had enrolled in college (American Council on Education, 1987, page 17).

a. Display these numbers in one or more frequency tables. Plot the data in any helpful way.

b. Discuss the relationship between race and high school completion.

c. Discuss the relationship between race and college entrance.

EXERCISE 5-21

Participants in the 1974 General Social Surveys were asked three questions, each beginning with:

> "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion . . ."

The three questions ended with:

A: ". . . if she is married and does not want any more children."
B: ". . . if the family has a very low income and cannot afford any more children."
C: ". . . if she is not married and does not want to marry the man."

Of 1,060 respondents, 413 said yes to all three questions and 430 said no to all three. Of the remaining 217 respondents, 29 said yes to A and B but no to C; 16 said yes to A and C but no to B; 18 said yes to A but no to B and C; 60 said yes to B and C but no to A; 57 said yes to B but no to A and C; 37 said yes to C but no to A and B (Tanner and Wong, 1987; from Haberman, 1979).

a. Arrange these results in one or more frequency tables.

b. Interpret the attitudes toward abortion reflected by respondents in this survey.

EXERCISE 5-22

Are adults with lifelong exposure to malaria less susceptible to the infection than are children? In a study of malaria in a region of Kenya, researchers treated 83 adults and 62 children (aged 6 months to 5 years) for malaria, achieving what is called a radical cure. By 56 days after radical cure, 57 of the 62 children and 13 of the 83 adults had developed malaria infections. By 84 days after radical cure, all 62 of the children and 48 of the 83 adults had developed malaria infections [*Science,* volume 237, August 7, 1987, pages 639–642).

a. Arrange these observations in a frequency table. Construct any plots that seem helpful.

b. Do adults seem to be less susceptible to malaria infection than children?

EXERCISE 5-23

A study of alcoholism in Sweden included men of known paternity, born to single women and adopted by nonrelatives at a young age. Researchers defined two types of alcoholism. Type I alcoholism is associated with onset after age 25, ability to abstain, infrequent fights when drinking, guilt about alcohol, and psychological dependence. Type II alcoholism is associated with onset before age 25, inability to abstain, frequent fights when drinking, little guilt about drinking, and little dependence. The researchers classified the 862 male adoptees by whether they had severe Type I alcoholism, whether they had a Type I genetic background, and whether they were raised in an environment that would contribute to excessive drinking:

Type I genetic background	Contributing environment	Number with severe Type I abuse/Number in category
No	No	16/376
No	Yes	3/72
Yes	No	22/328
Yes	Yes	10/86

The researchers classified the 862 male adoptees in a similar fashion with respect to Type II alcoholism:

Type II genetic background	Contributing environment	Number with severe Type II abuse/Number in category
No	No	11/567
No	Yes	8/196
Yes	No	12/71
Yes	Yes	5/28

Numbers of adoptees with alcohol abuse were calculated from percentages given in Cloninger (1987).

a. Tabulate and/or plot these results in any way that seems reasonable.

b. Discuss the relative contributions of genetic and environmental background to Type I and Type II alcoholism suggested by this study.

EXERCISE 5-24 Investigators classified sports injuries among children in a French health care district in 1981 and 1982 by age and sex of the child and type of injury. (These frequencies were calculated from percentages reported in Tursz and Crost, 1986.)

Type of injury	Boys 6–11 years	Boys ≥ 12 years	Girls 6–11 years	Girls ≥ 12 years
Contusions	77	119	49	87
Cuts, lacerations	61	32	21	7
Sprains, strains	22	45	12	43
Fractures	43	52	32	36

a. Tabulate and/or plot the results in any way that seems helpful.

b. Discuss the relationships among age, sex, and type of injury in this group of children.

EXERCISE 5-25 The accompanying table shows by species the number of animals testing positive for rabies and the number tested in Maryland, in 1982, 1983, and 1984. (Numbers positive for rabies were calculated from percentages reported in Beck, Felser, and Glickman, 1987.)

	Number positive for rabies/Number tested		
	1982	1983	1984
Raccoon	119/1,484	736/3,134	964/1,691
Skunk	13/80	28/120	32/69
Fox	0/79	5/116	19/91
Bat	17/753	51/1,169	46/1,098
Groundhog	0/72	5/215	13/445
Deer	0/11	1/24	1/36
Rabbit	0/64	0/102	2/202
Mouse/rat	0/100	0/86	1/144

	Number positive for rabies/Number tested		
	1982	1983	1984
Opossum	0/99	0/256	2/510
Chipmunk/squirrel	0/197	0/260	2/597
Ferret/mink	0/24	0/22	0/28
Beaver/muskrat	0/6	2/26	0/12
Horse	0/8	0/19	1/27
Cattle	1/27	3/72	2/103
Cat	0/609	7/1,069	15/1,503
Dog	0/603	0/750	1/801
Goat/sheep/pig	0/12	0/23	0/34

a. Tabulate and/or plot these results in any way that seems helpful.

b. Discuss these results from the point of view of a Maryland public health worker.

EXERCISE 5-26

Schistosomiasis is a parasitic infection common in the tropics. The parasites are carried by snails and passed to humans through contact with water (as in lakes, rivers, and irrigation canals). In a study of schistosomiasis in an Egyptian village, residents were examined and tested for presence of two forms of the parasite, Schistosoma mansoni and Schistosoma haematobium (Ismail et al., 1988). Researchers classified a total of 1,031 villagers by presence of one or both parasites:

Infection with Schistosoma haematobium	Infection with Schistosoma mansoni	
	Yes	No
Yes	119	46
No	213	653

The villagers were also classified by presence of infection and occupation:

Occupation	Number examined	Number infected with Schisto-soma haema-tobium	Number infected with Schisto-soma mansoni
Small child (not yet in school)	169	3	10
Student	382	99	162
Housewife/girl	241	15	34
Farmer	222	47	124
Nonfarm worker	17	0	2

In addition, the villagers were classified by age, sex, and presence of infection:

Age (years)	Number examined	Number infected with Schistosoma haematobium	Number infected with Schistosoma mansoni	Age (years)	Number examined	Number infected with Schistosoma haematobium	Number infected with Schistosoma mansoni
	Females				**Males**		
< 5	63	2	3	< 5	75	0	4
5–15	183	19	48	5–15	235	89	134
16–25	107	13	18	16–25	82	30	57
26–45	101	2	6	26–45	84	5	37
> 45	47	1	1	> 45	54	4	24

a. Tabulate and/or plot these results in any way that seems helpful.

b. Does there appear to be an association between infection with one schistosomiasis parasite and infection with the other?

c. For each of the two parasites, do the infection rates differ for different occupations? Is the association between infection and occupation different for the two parasites?

d. For each of the two parasites, are the infection rates different for males and females? Is the association between infection and sex different for the two parasites?

e. For each of the two parasites, do infection rates depend on age? Is the association between infection and age different for the two parasites?

f. For each of the two parasites, is the age effect on infection different for males and females? This is the same as asking if the sex effect on infection is different for different age groups.

EXERCISE 5-27 Self-reported cigarette smoking among Rhode Island physicians is shown here by specialty and year. The table shows number of smokers/number of physicians (percent smokers). (Numbers of smokers were calculated from percentages reported in Buechner et al., 1986.)

	1963	1968	1973	1978	1983
Internal medicine	31/113 (27.4)	24/158 (15.2)	21/229 (9.2)	31/352 (8.8)	30/496 (6.0)
General and family practice	59/171 (34.5)	59/274 (21.5)	37/215 (17.2)	21/167 (12.6)	19/227 (8.4)
Surgery	32/100 (32.0)	33/130 (25.4)	38/150 (25.3)	23/164 (14.0)	13/154 (8.4)
Pediatrics	14/60 (23.3)	13/78 (16.7)	19/92 (20.7)	12/119 (10.1)	10/140 (7.1)
Obstetrics and gynecology	26/57 (45.6)	21/65 (32.3)	23/82 (28.0)	18/94 (19.1)	16/111 (14.4)
Orthopedic surgery	11/27 (40.7)	12/50 (24.0)	11/58 (19.0)	6/64 (9.4)	9/72 (12.5)

a. Plot percent cigarette smokers versus year for each of the six specialties.

b. What is the trend in percent cigarette smokers over time?

c. Does the time trend vary with specialty?

EXERCISE 5-28

Indicators related to quality of life and children's health are listed for 29 countries (Grant, 1987). Malnutrition is the percentage of children under 5 years of age suffering from mild to moderate malnutrition (60% to 80% of desirable weight), 1980–1984. Water is the percentage of the population with access to drinking water in 1983. Polio is the percentage of 1-year-old children fully immunized against polio, 1984–1985. Low weight is the percentage of infants born weighing less than 2,500 grams (5.5 pounds) in 1982–1983. Breastfeeding is the percentage of mothers wholly or partially breastfeeding their babies for at least 6 months, 1980–1984.

Country	Malnutrition	Water	Polio	Low weight	Breastfeeding
Sierra Leone	24	23	12	14	94
Malawi	30	51	55	10	95
Niger	17	34	4	20	30
Rwanda	29	60	50	17	98
Yemen	54	31	8	9	76
Yemen, Dem.	32	50	14	12	73
Burundi	30	26	29	14	95
Bangladesh	63	42	2	50	97
Sudan	53	48	9	15	86
Bolivia	49	43	46	10	91
Haiti	65	33	24	17	85
Uganda	16	16	8	10	70
Pakistan	62	39	32	27	96
Ghana	23	43	18	17	70
Egypt	46	75	67	7	91
Peru	42	52	48	9	72
Indonesia	27	33	25	14	97
Congo	30	29	59	15	97
Kenya	30	28	57	13	84
Honduras	29	69	75	9	28
Brazil	55	76	86	9	19
Burma	50	25	2	20	90
El Salvador	52	51	55	13	77
Philippines	40	54	53	14	58
Colombia	43	81	61	10	58
Thailand	29	65	65	12	47
Panama	48	62	71	8	48
Chile	10	85	91	9	28
Costa Rica	46	93	74	9	20

a. Construct a stem-and-leaf plot for each of the five variables. Describe the distribution of each variable.

b. Construct a scatterplot matrix with these five variables.

c. Discuss the relationships between pairs of variables that you see in these plots.

EXERCISE 5-29

Baseball statistics summarizing the 1987 season are listed for 26 major-league teams (*USA Today,* October 6, 1987, page 4c; and October 7, 1987, page 5c). St. Louis, San Francisco, Detroit, and Minnesota were division winners in 1987. St. Louis and Minnesota were playoff winners. Minnesota won the World Series. League is a coded variable: 1 = American League, 2 = National League. Div is a coded variable: 1 = Eastern Division, 2 = Western Division. DivWin shows division winners: 1 = yes, 0 = no. POWin shows the playoff winners: 1 = yes, 0 = no. WSWin shows the World Series winner: 1 = yes, 0 = no. Win% is the percentage of games won by the team, to three decimal places. Runs is the number of runs scored by the team. RunAllow is the number of runs allowed by the team. HR is the number of homeruns hit by the team. Walks is the number of walks or bases on balls received by the team. BatAve is the proportion of at-bats that resulted in hits for the team. ERA is the team's earned run average, the average number of runs allowed by the team's pitching staff per nine innings of play. HitA is the number of hits allowed by the team. HRAllow is the number of homeruns allowed by the team. WalkA is the number of walks or bases on balls allowed by the team. SO is the number of batters struck out by the team's pitching staff.

Team	League	Div	DivWin	POWin	WSWin	Win%	Runs	RunAllow
St. Louis	2	1	1	1	0	.586	798	693
NY Mets	2	1	0	0	0	.568	823	698
Montreal	2	1	0	0	0	.562	741	720
San Francisco	2	2	1	0	0	.556	783	669
Cincinnati	2	2	0	0	0	.519	783	752
Philadelphia	2	1	0	0	0	.494	702	749
Pittsburgh	2	1	0	0	0	.494	723	744
Chicago Cubs	2	1	0	0	0	.472	720	801
Houston	2	2	0	0	0	.469	648	678
Los Angeles	2	2	0	0	0	.451	635	675
Atlanta	2	2	0	0	0	.429	747	829
San Diego	2	2	0	0	0	.401	668	763
Detroit	1	1	1	0	0	.605	896	735
Toronto	1	1	0	0	0	.593	845	655
Milwaukee	1	1	0	0	0	.562	862	817
NY Yankees	1	1	0	0	0	.549	788	758
Minnesota	1	2	1	1	1	.525	786	806
Kansas City	1	2	0	0	0	.512	715	691
Oakland	1	2	0	0	0	.500	806	789
Boston	1	1	0	0	0	.481	842	825
Seattle	1	2	0	0	0	.481	760	801
Chicago W. Sox	1	2	0	0	0	.475	748	746
Texas	1	2	0	0	0	.463	823	849
California	1	2	0	0	0	.463	770	803
Baltimore	1	1	0	0	0	.414	729	880
Cleveland	1	1	0	0	0	.377	742	957

a. Select a few of these variables that interest you. Plot these variables overall, by league, and by division within league.

b. Use a scatterplot matrix to examine relationships between pairs of variables.

c. Discuss your findings.

EXERCISE 5-30 Information on racial composition, fires and thefts, age of housing, homeowners insurance availability, and median family income is presented for each of 47 zip code areas of Chicago. The data were collected as part of a study to investigate possible discrimination in homeowners insurance underwriting practices in Chicago. From a report published by the U.S. Commission on Civil Rights (U.S. Commission on Civil Rights, 1979), this data set was contributed by S. E. Fienberg to a collection of problems edited by Andrews and Herzberg (1985, pages 407–411). Zip is the zip code area. Minor is the percentage of the population of minority racial background (provided by the U.S. Bureau of the Census). Fires gives the number of fires per 1,000 housing units during 1975 (provided by the Chicago Fire Department). Thefts gives the number of thefts per 1,000 population in 1975 (provided by the Chicago Police Department). Old is the percentage of housing units built before 1940 (provided by the U.S. Bureau of the Census). Vol is the number of new (voluntary) homeowners policies and renewals less nonrenewals and cancellations, per 100 housing

Team	HR	Walks	BatAve	ERA	HitA	HRAllow	WalkA	SO
St. Louis	94	644	.263	3.91	1,484	129	533	873
NY Mets	192	592	.268	3.84	1,407	135	510	1,032
Montreal	120	501	.265	3.92	1,428	145	446	1,012
San Francisco	205	511	.260	3.68	1,407	146	547	1,038
Cincinnati	192	514	.266	4.24	1,486	170	485	919
Philadelphia	169	587	.254	4.18	1,453	167	587	877
Pittsburgh	131	535	.264	4.20	1,377	164	562	914
Chicago Cubs	209	504	.264	4.55	1,524	159	628	1,024
Houston	122	526	.253	3.84	1,363	141	525	1,137
Los Angeles	125	445	.252	3.72	1,415	130	565	1,097
Atlanta	152	641	.258	4.63	1,529	163	587	837
San Diego	113	577	.260	4.27	1,402	175	602	897
Detroit	225	653	.272	4.02	1,430	180	563	976
Toronto	215	555	.269	3.74	1,323	158	567	1,064
Milwaukee	163	598	.276	4.62	1,548	169	529	1,039
NY Yankees	196	604	.262	4.36	1,475	179	542	900
Minnesota	196	523	.261	4.63	1,465	210	564	990
Kansas City	168	523	.262	3.86	1,424	128	548	923
Oakland	199	593	.260	4.32	1,442	176	531	1,042
Boston	174	606	.278	4.77	1,584	190	517	1,034
Seattle	161	500	.272	4.49	1,503	199	497	919
Chicago W. Sox	173	487	.258	4.30	1,436	189	537	792
Texas	194	567	.266	4.63	1,388	199	760	1,103
California	172	590	.252	4.38	1,481	212	504	941
Baltimore	211	524	.258	5.01	1,555	226	547	870
Cleveland	187	489	.263	5.28	1,566	219	606	849

Zip	Minor	Fires	Thefts	Old	Vol	Invol	Income
60626	10.0	6.2	29	60.4	5.3	.0	11,744
60640	22.2	9.5	44	76.5	3.1	.1	9,323
60613	19.6	10.5	36	73.5	4.8	1.2	9,948
60657	17.3	7.7	37	66.9	5.7	.5	10,656
60614	24.5	8.6	53	81.4	5.9	.7	9,730
60610	54.0	34.1	68	52.6	4.0	.3	8,231
60611	4.9	11.0	75	42.6	7.9	.0	21,480
60625	7.1	6.9	18	78.5	6.9	.0	11,104
60618	5.3	7.3	31	90.1	7.6	.4	10,694
60647	21.5	15.1	25	89.8	3.1	1.1	9,631
60622	43.1	29.1	34	82.7	1.3	1.9	7,995
60631	1.1	2.2	14	40.2	14.3	.0	13,722
60646	1.0	5.7	11	27.9	12.1	.0	16,250
60656	1.7	2.0	11	7.7	10.9	.0	13,686
60630	1.6	2.5	22	63.8	10.7	.0	12,405
60634	1.5	3.0	17	51.2	13.8	.0	12,198
60641	1.8	5.4	27	85.1	8.9	.0	11,600
60635	1.0	2.2	9	44.4	11.5	.0	12,765
60639	2.5	7.2	29	84.2	8.5	.2	11,084
60651	13.4	15.1	30	89.8	5.2	.8	10,510
60644	59.8	16.5	40	72.7	2.7	.8	9,784
60624	94.4	18.4	32	72.9	1.2	1.8	7,342
60612	86.2	36.2	41	63.1	.8	1.8	6,565
60607	50.2	39.7	147	83.0	5.2	.9	7,459
60623	74.2	18.5	22	78.3	1.8	1.9	8,014
60608	55.5	23.3	29	79.0	2.1	1.5	8,177
60616	62.3	12.2	46	48.0	3.4	.6	8,212
60632	4.4	5.6	23	71.5	8.0	.3	11,230
60609	46.2	21.8	4	73.1	2.6	1.3	8,330
60653	99.7	21.6	31	65.0	.5	.9	5,583
60615	73.5	9.0	39	75.4	2.7	.4	8,564
60638	10.7	3.6	15	20.8	9.1	.0	12,102
60629	1.5	5.0	32	61.8	11.6	.0	11,876
60636	48.8	28.6	27	78.1	4.0	1.4	9,742
60621	98.9	17.4	32	68.6	1.7	2.2	7,520
60637	90.6	11.3	34	73.4	1.9	.8	7,388
60652	1.4	3.4	17	2.0	12.9	.0	13,842
60620	71.2	11.9	46	57.0	4.8	.9	11,040
60619	94.1	10.5	42	55.9	6.6	.9	10,332
60649	66.1	10.7	43	67.5	3.1	.4	10,908
60617	36.4	10.8	34	58.0	7.8	.9	11,156
60655	1.0	4.8	19	15.2	13.0	.0	13,323
60643	42.5	10.4	25	40.8	10.2	.5	12,960
60628	35.1	15.6	28	57.8	7.5	1.0	11,260
60627	47.4	7.0	3	11.4	7.7	.2	10,080
60633	34.0	7.1	23	49.2	11.6	.3	11,428
60645	3.1	4.9	27	46.6	10.9	.0	13,731

Adapted from Figure 67.1 of *Data: A Collection of Problems from Many Fields for the Student and Research Worker,* by D. F. Andrews and A. M. Herzberg. Copyright © 1985 by Springer-Verlag New York, Inc. Reprinted by permission.

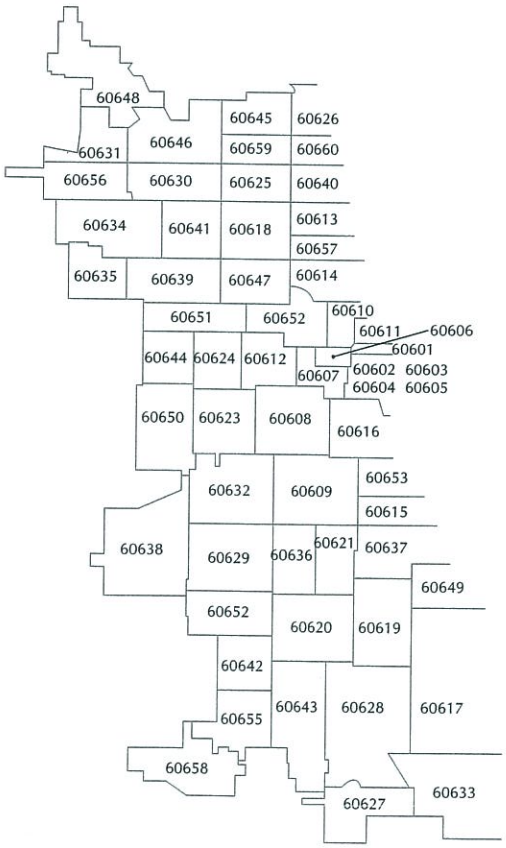
units, from December 1977 through February 1978 (provided by insurance companies to the Illinois Department of Insurance). Invol is the number of new fair-plan (involuntary) homeowners insurance policies and renewals per 100 housing units, from December 1977 through May 1978 (provided by the Illinois Department of Insurance); most fair-plan policyholders were rejected for voluntary policies by homeowners insurance companies. Income is the median family income for the area as estimated by the U.S. Bureau of the Census.

a. Plot each of the variables (other than zip code). Describe the distribution of each variable.

b. Select a few of these variables that interest you. Construct a scatterplot matrix using these variables. Discuss the relationship between pairs of variables revealed in your scatterplot matrix.

c. Using the accompanying map of Chicago zip code areas, use framed rectangles to display by zip code area any variable(s) you wish. You may wish to use separate graphs for different variables. Are there geographic trends?

d. Display these variables in any other way you find useful. Discuss your results.



EXERCISE 5-31 Use framed rectangles on a map of the United States to display the pregnancy rates for teenage girls given in Exercise 2-8. Discuss any geographic trends you see.

EXERCISE 5-32 Use framed rectangles on a map of the United States to display the record cold temperatures listed in Exercise 2-4. Discuss any geographic trends you see.

EXERCISE 5-33 Use framed rectangles on a map of the United States to display the governors' salaries listed in Exercise 2-19. Discuss any geographic trends you see.

EXERCISE 5-34 District scores for third graders in reading, mathematics, and science are shown here for 27 Massachusetts urbanized centers (*The Boston Sunday Globe,* November 30, 1986, page 96).

Urbanized center	Reading	Mathematics	Science
Attleboro	1290	1250	1300
Boston	1190	1210	1180
Brockton	1200	1190	1180
Cambridge	1250	1250	1230
Chelsea	1110	1130	1110
Chicopee	1280	1230	1250
Everett	1190	1220	1190
Fall River	1170	1190	1190
Fitchburg	1260	1240	1270
Gloucester	1240	1250	1260
Greenfield	1300	1340	1360
Haverhill	1260	1240	1250
Holyoke	1150	1170	1150
Lawrence	1100	1110	1110
Lowell	1220	1250	1230
Lynn	1200	1210	1190
Malden	1260	1230	1250
New Bedford	1210	1220	1190
Pittsfield	1270	1270	1280
Quincy	1290	1270	1290
Revere	1260	1220	1220
Salem	1240	1300	1260
Somerville	1190	1190	1190
Springfield	1210	1220	1210
Waltham	1310	1310	1290
Watertown	1340	1360	1330
Worcester	1250	1260	1250

Graph these observations in any ways that you find helpful. Discuss your findings.

EXERCISE 5-35 The United Nations ranks countries according to mortality rate for children under 5 years of age (Grant, 1987). The 20 countries with the highest under-5 mortality rates and the 20 with the lowest under-5 mortality rates are listed, along with values for four variables. Mort is under-5 mortality rate in 1985:

annual number of deaths of children under 5 years of age per 1,000 live births. Life is life expectancy: the number of years newborn children could be expected to live based on prevailing mortality risks in 1985. Weight is percentage of infants born weighing less than 2,500 grams (5.5 pounds) in 1982–1983. Calorie is the daily calorie supply per capita as percentage of requirements in 1983. A −9 denotes a missing value.

Country	Mort	Life	Weight	Calorie
Afghanistan	329	38	20	−9
Mali	302	43	13	68
Sierra Leone	302	35	14	91
Malawi	275	46	10	95
Guinea	259	41	18	84
Ethiopia	257	41	13	93
Somalia	257	41	−9	89
Mozambique	252	46	16	71
Burkina Faso	245	46	21	85
Angola	242	43	19	87
Niger	237	44	20	97
Central African Republic	232	44	23	91
Chad	232	46	11	68
Guinea-Bissau	232	44	15	−9
Senegal	231	44	10	102
Mauritania	223	45	10	97
Kampuchea	216	46	−9	−9
Liberia	215	50	−9	102
Rwanda	214	48	17	98
Yemen	210	50	9	92
Belgium	13	74	5	140
German Dem. Rep.	13	73	6	142
Italy	13	75	7	140
USA	13	75	7	137
Germany, Federal Republic of	12	74	5	130
Ireland	12	73	4	143
Singapore	12	73	8	115
Spain	12	75	−9	132
United Kingdom	12	74	7	128
Australia	11	75	6	115
France	11	75	5	139
Hong Kong	11	76	8	122
Canada	10	76	6	130
Denmark	10	75	6	131
Netherlands	10	76	4	129
Norway	10	76	4	115
Japan	9	77	5	113
Switzerland	9	76	5	129
Finland	8	74	4	114
Sweden	8	76	4	116

Display these observations in any ways that you find helpful. Discuss your findings.