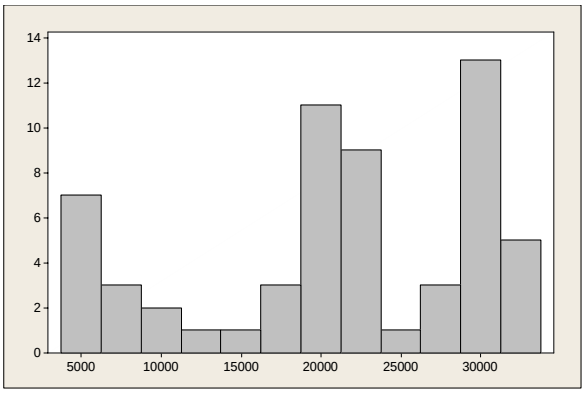
CHAPTER

1



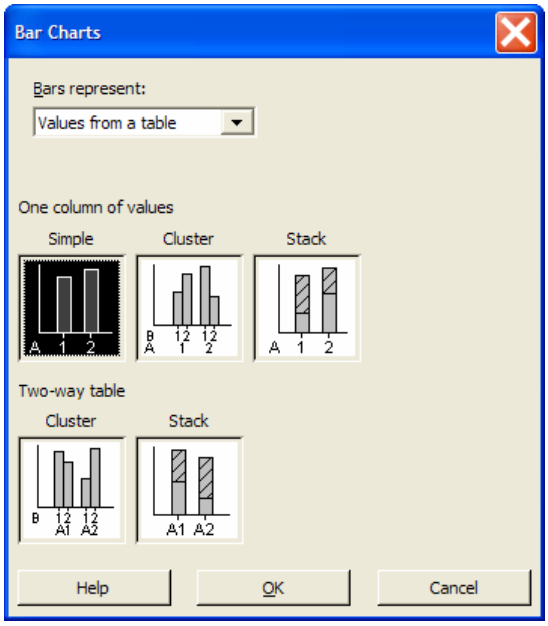
Picturing Distributions with Graphs

Bar Charts

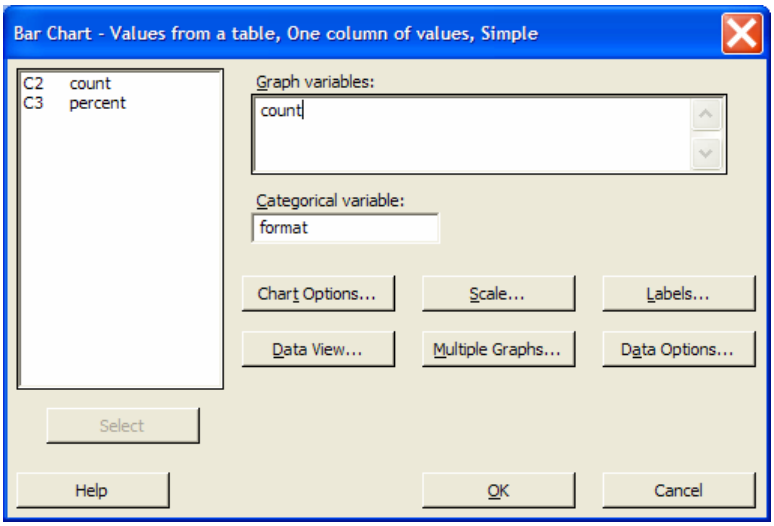
Minitab allows us to examine the distribution of variables with graphs. Bar charts are useful for categorical data. We will use the data found in Example 1.2 in BPS and EG01-02.MTW to show how to make a bar chart with Minitab. The data places the country's 13,838 radio stations into categories that describe the kind of programs they broadcast. The formats are entered in C1 and the counts of stations are given in C2. In addition, the percents of stations are given in C3. To make a bar chart of the data, select

Graph ➤ Bar Chart

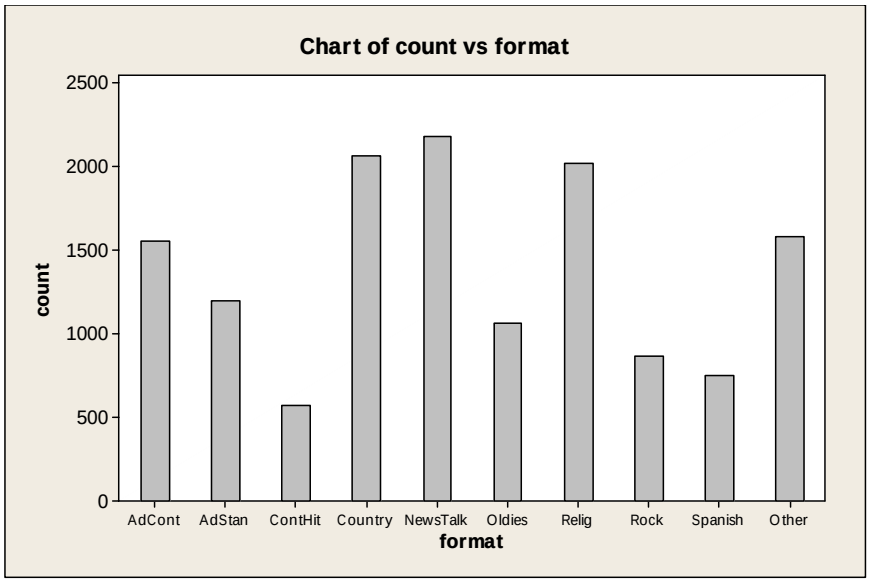
from the menu. Select Values from a table and Simple and then click OK.



As shown in the following dialog box, the graph variable and categorical variables must be filled in. In this example, the categorical variable is "format" and the graph variable is "count." If the graph variable is "percent", then the bar heights would be percents instead of counts.



Clicking on OK will produce the following bar chart.

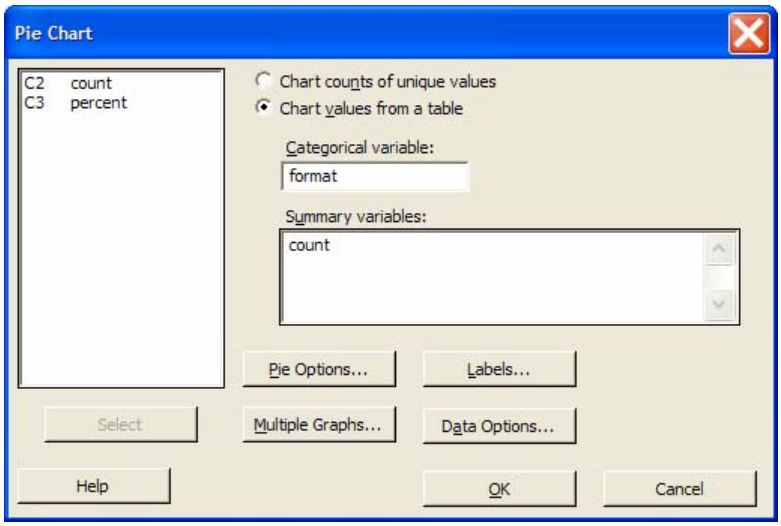


Pie Charts

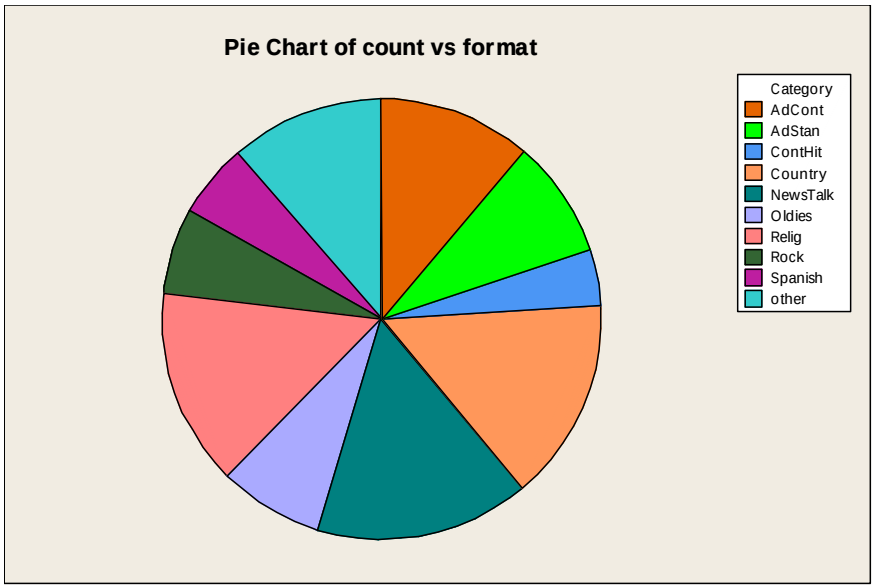
Another way to examine distributions of categorical variables is with a pie chart. We will continue to use the data found in Example 1.2 in BPS to show how to make a pie chart with Minitab. To make a pie chart of the waste data, select

Graph ➤ Pie Chart

from the menu. Because the data are tabulated, click on Chart values from a table on the dialog box and fill in the appropriate variables for the categorical and summary variables.



The pie chart for Example 1.2 follows.

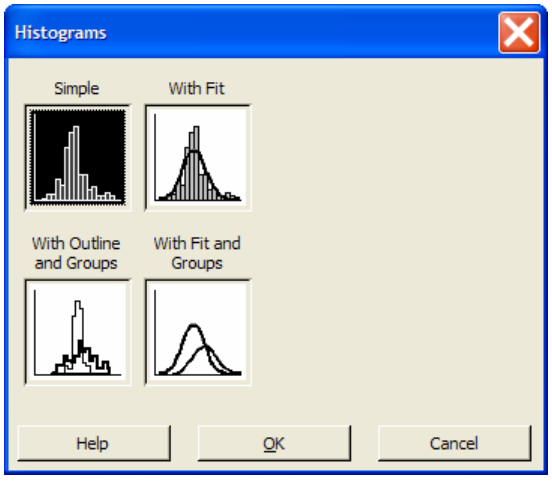


Histograms

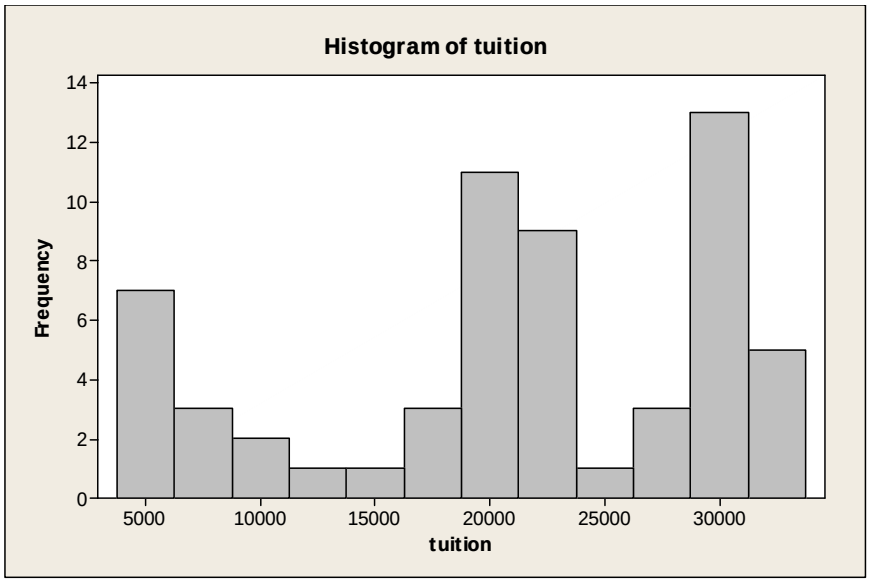
The most common graph for the distribution of a quantitative variable is a histogram. Example 1.8 in BPS and EG01-08.MTW gives the charges for instate students at all 59 four-year colleges and universities in Massachusetts. To create a histogram for these data, select

Graph ➤ Histogram

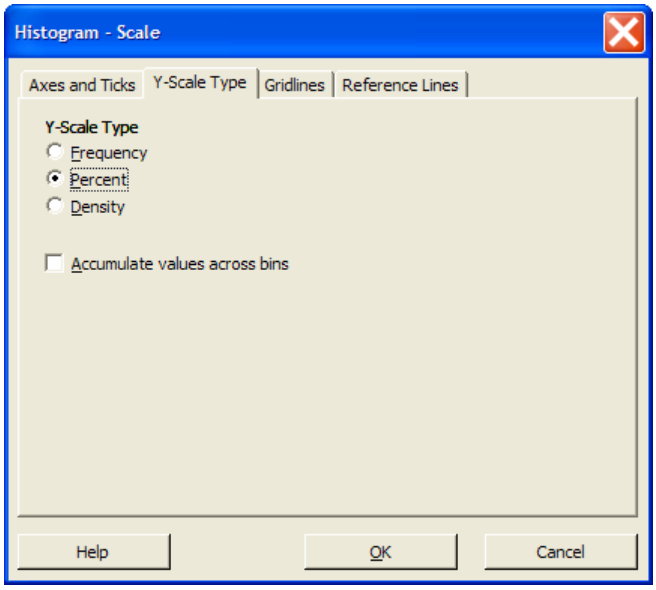
from the menu. Click on Simple and then OK in the first dialog box. The next dialog box will appear as follows. In the dialog box, double click on the variable named tuition and click on OK.



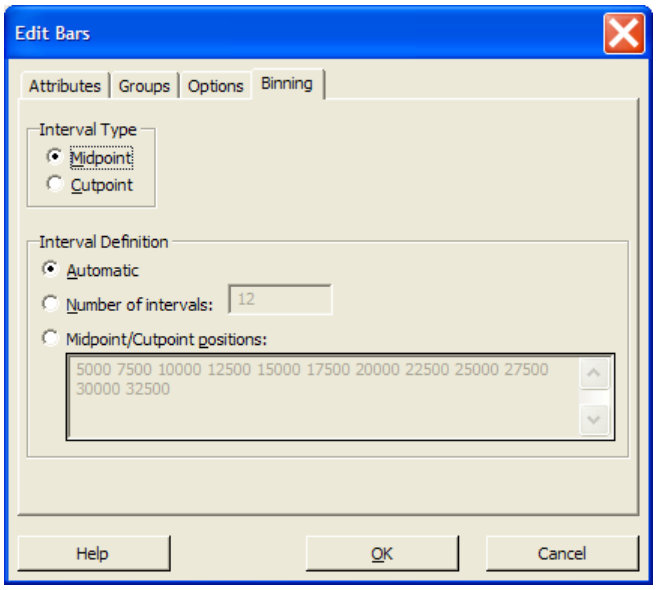
The histogram appears as follows.



To change the type of histogram from frequency to percent, click on the Scales button and then the Y-Scale Type tab and click on Percent.



You may also change from Minitab's automatic choice of intervals. Once you've produced a histogram, double click on the *X*-scale to obtain the Edit Scale dialog box. Click on the Binning tab to change from midpoints to cutpoints and/or specify midpoint/cutpoint positions.

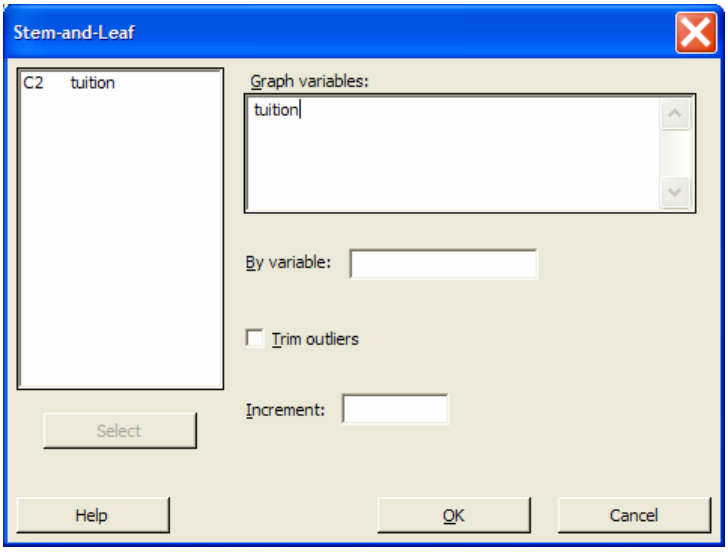


Stemplots

A stem-and-leaf plot (or stemplot) uses the actual data from a quantitative distribution to create the display. To create the stemplot, select

Graph ➤ Stem-and-Leaf

from the menu. The session window will appear as follows.



When you click on the Graph Variables text box, the valid choices appear in the variable list box. Click on the desired variable and then click Select. When you click OK, the following stemplot will appear in the Session window.

Stem-and-Leaf Display: tuition

```
Stem-and-leaf of tuition  N  = 59
Leaf Unit = 1000

 7   0  4444455
 9   0  77
 12  0  899
 12  1
 13  1  3
 14  1  5
 16  1  77
 23  1  8999999
(7)  2  0000011
 29  2  2222233
 22  2  5
 21  2  67
 19  2  8899999
 12  3  0000001111
 2   3  22
```

The first column of a stem-and-leaf display is called the depth, the second column holds the stems, and the rest of the display holds the leaves. Each leaf digit represents one observation. In the stem-and-leaf display shown, the first stem is 0 and the first leaf is 4. The leaf unit at the top of the display tells us where to put the decimal point. In this example, the Leaf Unit = 1000. Therefore the corresponding first observation is 4,000 (which was rounded from 4,316.)

The column on the left gives a cumulative count of values from the top of the figure down and from the bottom of the figure up to the middle. The count for the row containing the median has parentheses around it. Parentheses around the median row are omitted if the median falls between two lines of the display. This occurred in the stemplot shown.

If you wish, you can control the scaling of a stem-and-leaf display by specifying an increment. For example, we can choose an increment of 5000 to obtain the following display.

Stem-and-Leaf Display: tuition

```
Stem-and-leaf of tuition  N  = 59
Leaf Unit = 1000

 5    0  44444
 12   0  5577899
 13   1  3
 23   1  5778999999
(14)  2  00000112222233
 22   2  5678899999
 12   3  000000111122
```

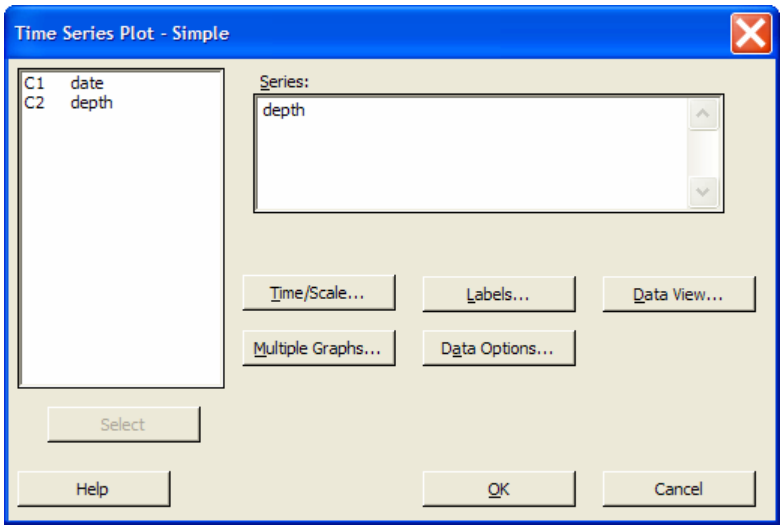
Time Series Plots

When quantitative data are collected over time, it is a good idea to plot the observations in the order they were collected. To create a time plot, select

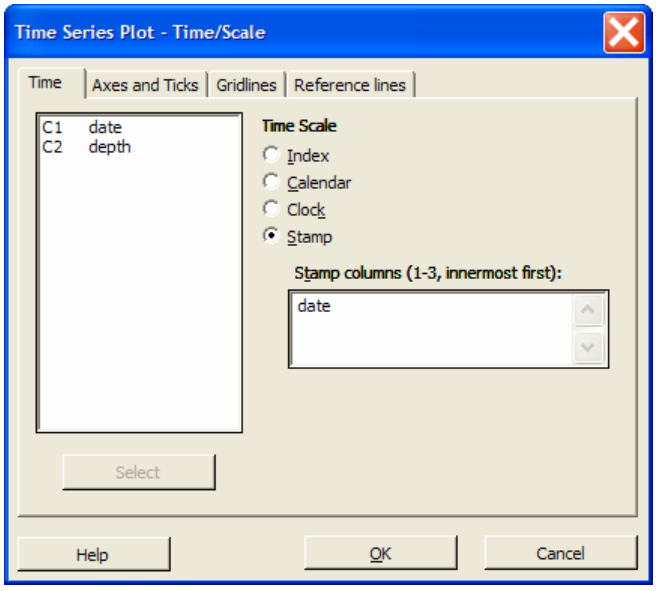
Graph ➤ Time Series Plot

from the menu. This command plots time series data on the *y*-axis versus time on the *x*-axis. For example, the data in EG01-11.MTW give the water levels at a water monitoring station in Shark River Slough, the main path for surface water moving through the "river of grass" that is the Everglades.

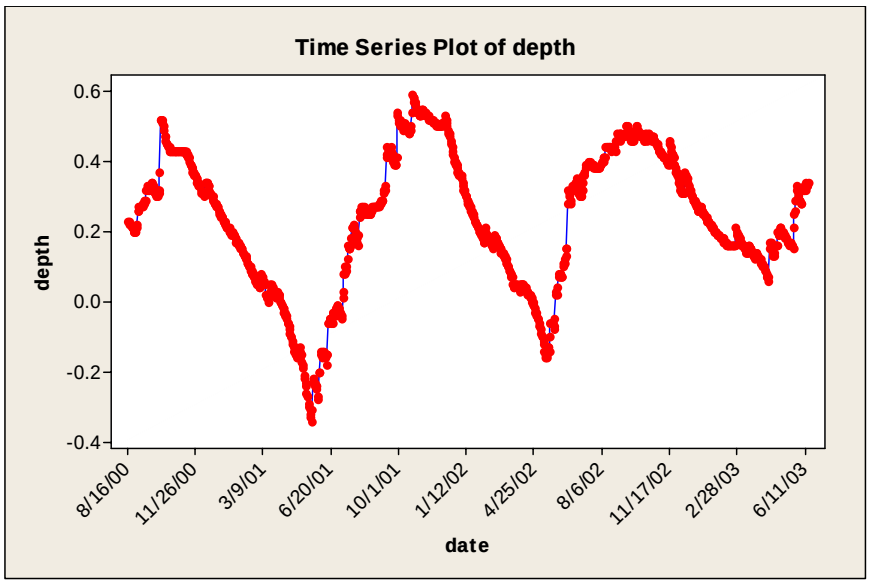
To make a time series plot of this data, select **Graph ➤ Time Series Plot** from the menu. Since there is one time series to plot, choose Simple and then click on OK in the first dialog box. The next dialog box will appear as follows.



Select the series "depth" to be plotted and then click on the Time/Scale button. In the Time/Scale subdialog box, we specify Stamp and use the dates in C1.



Click on OK in the subdialog box and then in the dialog box to obtain the following time series plot.



EXERCISES

1.3 The color of your car. Here is the distribution of the most popular colors for 2005 model luxury cars made in North America.

Color	Percent
Silver	20%
White, pearl	18%
Black	16%
Blue	13%
Light brown	10%
Red	7%
Yellow, gold	6%

(a) Select **Graph ➤ Bar Chart** to make a bar chart of the color data.

(b) What percent of vehicles have some other color? Select **Calc ➤ Column Statistics** and then click on Sum to help you find out. Add an "Other" category to your worksheet. Check that the sum in the Percent column is now 100.

(c) Select **Graph ➤ Pie Chart** from the menu to make a pie chart of the color data.

1.4 Births are not, as you might think, evenly distributed across the days of the week. Here are the average numbers of babies born on each day of the week in 2003.

Day	Births
Sunday	7,564
Monday	11,733
Tuesday	13,001
Wednesday	12,598
Thursday	12,514
Friday	12,396
Saturday	8,605

(a) Select **Graph ➤ Bar Chart** to present these data in a well-labeled bar chart.

(b) Select **Graph ➤ Pie Chart** to make a pie chart of these data. Suggest some possible reasons why there are fewer births on weekends.

1.6 Table 1.2 in BPS and TA01-02.MTW gives the average travel time to work for workers in each state who are at least 16 years old and don't work at home. Select **Graph ➤ Histogram** to make a histogram of the travel times. Is the shape of the distribution closer to symmetric or skewed?

1.10 Table 1.2 in BPS and TA01-02.MTW gives the average travel time to work for workers in each state who are at least 16 years old and don't work at home. Select **Graph ➤ Stem-and-Leaf** to make a stemplot of the travel times. What is the median of the 51 observations?

1.11 People with diabetes must monitor and control their blood glucose level. The goal is to maintain "fasting plasma glucose" between about 90 and 130 milligrams per deciliter (mg/dL). Here and in EX01-11.MTW are the fasting plasma glucose levels for 18 diabetics enrolled in a diabetes control class, five months after the end of the class:

141	158	112	153	134	95	96	78	148
172	200	271	103	172	359	145	147	255

Select **Graph ➤ Stem-and-Leaf** to make a stemplot of these data and describe the main features of the distribution. Are there outliers? How well is the group as a whole achieving the goal for controlling glucose levels?

1.12 Here and in EX01-12.MTW are data on the average tuition and fees charged by public four-year colleges and universities for academic years beginning in 1976 to 2005. Because almost any variable measured in dollars increases over time due to inflation (the falling buying power of a dollar), the values are given in "constant dollars," adjusted to have the same buying power that a dollar had in 2005.

Year	Tuition	Year	Tuition	Year	Tuition	Year	Tuition
1976	$2,059	1984	$2,274	1992	$3,208	2000	$3,925
1977	$2,049	1985	$2,373	1993	$3,396	2001	$4,140
1978	$1,968	1986	$2,490	1994	$3,523	2002	$4,408
1979	$1,862	1987	$2,511	1995	$3,564	2003	$4,890
1980	$1,818	1988	$2,551	1996	$3,668	2004	$5,239
1981	$1,892	1989	$2,617	1997	$3,768	2005	$5,491
1982	$2,058	1990	$2,791	1998	$3,869		
1983	$2,210	1991	$2,987	1999	$3,894		

(a) Select **Graph ➤ Time Series Plot** from the menu to make a time plot of average tuition and fees.

(b) What overall pattern does your plot show?

(c) Some possible deviations from the overall pattern are: outliers; periods of decreasing charges (in 2005 dollars); periods of particularly rapid increase. Which are present in your plot, and during which years?

1.25 Here and in EX01-25.MTW are the colors for luxury cars made in Europe:

Color	Percent
Black	30%
Silver	24%
Gray	19%
Blue	14%
Green	3%
White, pearl	3%

What percent of European luxury cars have other colors? Select **Graph ➤ Pie Chart** to make a graph of these data. What are the most important differences between choice of colors in Europe and North America?

1.26 **The n**umber of deaths among persons aged 15 to 24 years in the United States in 2003 due to the leading causes of death for this age group were: accidents, 14,966; homicide, 5148; suicide, 3921; cancer, 1628; heart disease, 1083; congenital defects. The data are also given in EX01-26.MTW.

(a) Select **Graph ➤ Bar Chart** to make a graph to display these data.

(b) What additional information do you need to make a pie chart?

1.28 Here and in EX01-28.MTW are data on the percent of people in several age groups who attended a movie in the past 12 months:

Age group	Movie attendance
18 to 24 years	83%
25 to 34 years	73%
35 to 44 years	68%
45 to 54 years	60%
55 to 64 years	47%
65 to 74 years	32%
75 years and over	20%

(a) Display these data in a bar graph. What is the main feature of the data?

(b) Would it be correct to make a pie chart of these data? Why?

(c) A movie studio wants to know what percent of the total audience for movies is 18 to 24 years old. Explain why these data do not answer this question.

1.29 Email spam is the curse of the Internet. Here and in EX01-29.MTW is a compilation of the most common types of spam:

Type of spam	Percent
Adult	14.5
Financial	16.2
Health	7.3
Leisure	7.8
Products	21.0
Scams	14.2

Select **Graph ➤ Bar Chart** to make two graphs of these percents, one with bars ordered as in the table (alphabetical) and the other with bars in order from tallest to shortest. Comparisons are easier if you order the bars by height. A bar graph ordered from tallest to shortest bar is sometimes called a **Pareto chart,** after the Italian economist who recommended this procedure. To get the two types of graph, click on Bar Chart Options in the Bar Chart dialog box. Select 'Default' for the bars ordered alphabetically or 'Decreasing Y' to order the bars from tallest to shortest.

1.31 The return on a stock is the change in its market price plus any dividend payments made. Total return is usually expressed as a percent of the beginning price. EX01-31 gives the monthly returns for all stocks listed on U.S. markets from January 1980 to March 2005 (243 months). The extreme low outlier is the market crash of October 1987, when stocks lost 23% of their value in one month. Select **Graph ➤ Histogram** to make a graph of this data.

(a) Ignoring the outliers, describe the overall shape of the distribution of monthly returns.

(b) What is the approximate center of this distribution? (For now, take the center to be the value with roughly half the months having lower returns and half having higher returns.)

(c) Approximately what were the smallest and largest monthly returns, leaving out the outliers? (This is one way to describe the spread of the distribution.)

(d) A return less than zero means that stocks lost value in that month. About what percent of all months had returns less than zero?

1.33 The states differ greatly in the kinds of severe weather that afflict them. Table 1.3 and TA01-03.MTW give the average property damage caused by tornadoes per year over the period from 1950 to 1999 in each of the 50 states and Puerto Rico. (To adjust for the changing buying power of the dollar over time, all damages were restated in 1999 dollars.)

(a) What are the top five states for tornado damage? The bottom five? (Include Puerto Rico, though it is not a state.)

(b) Select **Graph ➤ Histogram** from the menu to make a graph of the data. Double click on the *X*-scale to use the classes: "0 ≤ damage < 10," "10 ≤ damage < 20," and so on. Describe the shape, center, and spread of the distribution. Which states may be outliers? (To understand the outliers, note that most tornadoes in largely rural states such as Kansas cause little property damage. Damage to crops is not counted as property damage.)

(c) Also display the "default" histogram that your software makes when you give it no binning instructions. How does this compare with your graph in (b)?

1.34 Table 1.4 and TA01-04.MTW give the number of active medical doctors per 100,000 people in each state.

(a) Why is the number of doctors per 100,000 people a better measure of the availability of health care than a simple count of the number of doctors in a state?

(b) Select **Graph ➤ Histogram** to make a histogram that displays the distribution of doctors per 100,000 people. Write a brief description of the distribution. Are there any outliers? If so, can you explain them?

1.35 Burning fuels in power plants or motor vehicles emits carbon dioxide (CO_2), which contributes to global warming. Table 1.5 in BPS and TA01-05.MTW give CO_2 emissions per person from countries with populations of at least 20 milion.

(a) Why do you think we choose to measure emissions per person rather than total CO_2 emissions for each country?

(b) Select **Graph ➤ Stem-and-Leaf** from the menu to make a stemplot to display the data of Table 1.5. Describe the shape, center, and spread of the distribution. Which countries are outliers?

1.36 "Recruitment," the addition of new members to a fish population, is an important measure of the health of ocean ecosystems. Here and in EX01-36.MTW are data on the recruitment of rock sole in the Bering Sea from 1973 to 2000:

Year	Recruitment (millions)	Year	Recruitment (millions)	Year	Recruitment (millions)
1973	173	1983	2246	1993	998
1974	234	1984	1793	1994	505
1975	616	1985	1793	1995	304
1976	344	1986	2809	1996	425
1977	515	1987	4700	1997	214
1978	576	1988	1702	1998	385
1979	727	1989	1119	1999	445
1980	1411	1990	2407	2000	676
1981	1431	1991	1049		
1982	1250	1992	505		

Select **Graph ➤ Stem-and-Leaf** from the menu to make a stemplot to display the distribution of yearly rock sole recruitment. Describe the shape, center, and spread of the distribution and any striking deviations that you see.

1.41 The JD Power Initial Quality Survey polls more than 50,000 buyers of new motor vehicles 90 days after their purchase. A two-page questionnaire asks about "things gone wrong." Here and in EX01-41.MTW are data on problems per 100 vehicles for vehicles made by Toyota and by General Motors in recent years. Select **Graph ➤ Time Series Plot** and click on Multiple to make two time plots in the same graph to compare Toyota and GM. What are the most important conclusions you can draw from your graph?

Year	1998	1999	2000	2001	2002	2003	2004
GM	187	179	164	147	130	134	120
Toyota	156	134	116	115	107	115	101

1.43 EX01-43.MTW gives the average price of fresh oranges each month from March 1995 to March 2005. The prices are "index numbers" given as percents of the average price during 1982 to 1984. Select **Graph ➤ Time Series Plot** from the menu to make a graph of the data.

(a) The most notable pattern in this time plot is yearly cycles. At what season of the year are orange prices highest? Lowest? (To read the graph, note that the tick mark for each year is at the beginning of the year.) The cycles are explained by the time of the orange harvest in Florida.

(b) Is there a longer-term trend visible in addition to the cycles? If so, describe it.

1.44 Here and in EX01-44.MTW are data on the number of unprovoked attacks by alligators on people in Florida over a 33-year period:

Year	Attacks	Year	Attacks	Year	Attacks	Year	Attacks
1972	5	1981	5	1990	18	1999	15
1973	3	1982	6	1991	18	2000	23
1974	4	1983	6	1992	10	2001	17
1975	5	1984	5	1993	18	2002	14
1976	2	1985	3	1994	22	2003	6
1977	14	1986	13	1995	19	2004	11
1978	5	1987	9	1996	13		
1979	2	1988	9	1997	11		
1980	4	1989	13	1998	9		

Make two graphs of these data to illustrate why you should always make a time plot for data collected over time.

(a) Select **Graph ➤ Histogram** to make a histogram of the counts of attacks. What is the overall shape of the distribution? What is the median number of alligator attacks per year?

(b) Select **Graph ➤ Time Series Plot** to make a time plot. What overall pattern does your plot show? Why is the median number of attacks from 1972 to 2004 not very useful in (say) 2006? (The main reason for the time trend is the continuing increase in Florida's population.)