

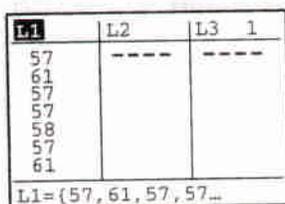
You can also use computer software or a calculator to construct histograms.

TECHNOLOGY TOOLBOX *Making calculator histograms*

1. Enter the presidential age data from Example 1.6 in your statistics list editor.

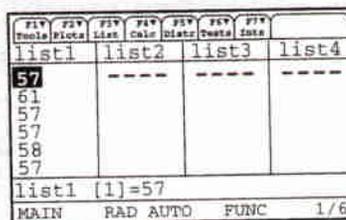
TI-83

- Press **STAT** and choose 1:Edit...
- Type the values into list L₁.



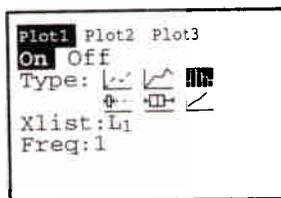
TI-89

- Press **APPS**, choose 1:FlashApps, then select Stats/List Editor and press **ENTER**.
- Type the values into list1.

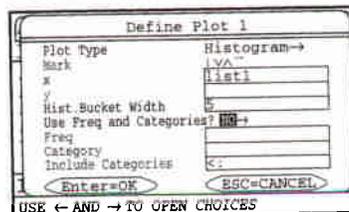


2. Set up a histogram in the statistics plots menu.

- Press **2nd** **Y=** (STAT PLOT).
- Press **ENTER** to go into Plot1.
- Adjust your settings as shown.

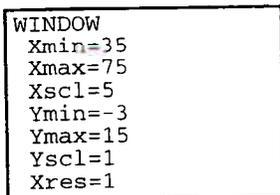


- Press **F2** and choose 1:Plot Setup....
- With Plot 1 highlighted, press **ENTER** to define.
- Change Hist. Bucket Width to 5, as shown.

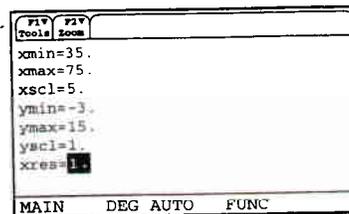


3. Set the window to match the class intervals chosen in Example 1.6.

- Press **WINDOW**.
- Enter the values shown.



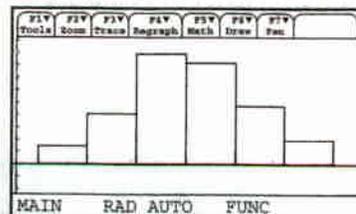
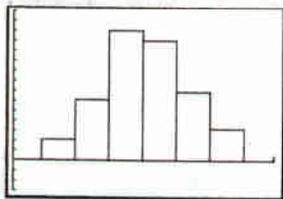
- Press **F2** (WINDOW).
- Enter the values shown.



4. Graph the histogram. Compare with Figure 1.7.

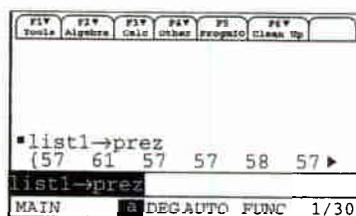
- Press **GRAPH**.

- Press **F3** (GRAPH).

TECHNOLOGY TOOLBOX *Making calculator histograms (continued)*


5. Save the data in a named list for later use.
- From the home screen, type the command $L_1 \rightarrow \text{PREZ}$ ($\text{list1} \rightarrow \text{prez}$ on the TI-89) and press **ENTER**. The data are now stored in a list called PREZ.

```
L1 → PREZ
{57 61 57 57 58...
```


Histogram tips:

- There is no one right choice of the classes in a histogram. Too few classes will give a “skyscraper” graph, with all values in a few classes with tall bars. Too many will produce a “pancake” graph, with most classes having one or no observations. Neither choice will give a good picture of the shape of the distribution.
- Five classes is a good minimum.
- Our eyes respond to the *area* of the bars in a histogram, so be sure to choose classes that are all the same width. Then area is determined by height and all classes are fairly represented.
- If you use a computer or graphing calculator, beware of letting the device choose the classes.

EXERCISES

1.12 WHERE DO OLDER FOLKS LIVE? Table 1.5 gives the percentage of residents aged 65 or older in each of the 50 states.

Construct a histogram for these data. Describe the shape, center, and spread of the distribution of CEO salaries. Are there any apparent outliers?

1.15 CHEST OUT, SOLDIER! In 1846, a published paper provided chest measurements (in inches) of 5738 Scottish militiamen. Table 1.6 displays the data in summary form.

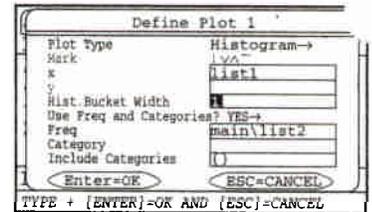
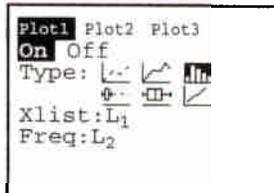
TABLE 1.6 Chest measurements (inches) of 5738 Scottish militiamen

Chest size	Count	Chest size	Count
33	3	41	934
34	18	42	658
35	81	43	370
36	185	44	92
37	420	45	50
38	749	46	21
39	1073	47	4
40	1079	48	1

Source: Data and Story Library (DASL), <http://lib.stat.cmu.edu/DASL/>.

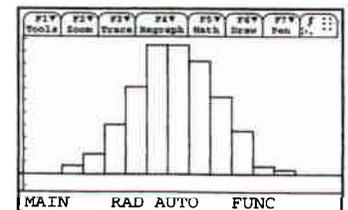
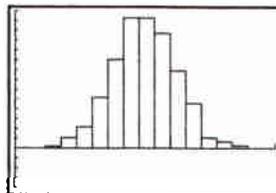
(a) You can use your graphing calculator to make a histogram of data presented in summary form like the chest measurements of Scottish militiamen.

- Type the chest measurements into $L_1/\text{list1}$ and the corresponding counts into $L_2/\text{list2}$
- Set up a statistics plot to make a histogram with x -values from $L_1/\text{list1}$ and y -values (bar heights) from $L_2/\text{list2}$.



- Adjust your viewing window settings as follows: $x_{\min} = 32$, $x_{\max} = 49$, $x_{\text{sc1}} = 1$, $y_{\min} = -300$, $y_{\max} = 1100$, $y_{\text{sc1}} = 100$. From now on, we will abbreviate in this form: $X[32,49]_1$ by $Y[-300,1100]_{100}$. Try using the calculator's built-in ZoomStat/ZoomData command. What happens?

- Graph.



(b) Describe the shape, center, and spread of the chest measurements distribution. Why might this information be useful?

More about shape

When you describe a distribution, concentrate on the main features. Look for major peaks, not for minor ups and downs in the bars of the histogram. Look for clear outliers, not just for the smallest and largest observations. Look for rough *symmetry* or clear *skewness*.

SYMMETRIC AND SKEWED DISTRIBUTIONS

A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.

A distribution is **skewed to the right** if the right side of the histogram (containing the half of the observations with larger values) extends much farther out than the left side. It is **skewed to the left** if the left side of the histogram extends much farther out than the right side.

In mathematics, symmetry means that the two sides of a figure like a histogram are exact mirror images of each other. Data are almost never exactly symmetric, so we are willing to call histograms like that in Exercise 1.15 approximately symmetric as an overall description. Here are more examples.

EXAMPLE 1.7 LIGHTNING FLASHES AND SHAKESPEARE

Figure 1.8 comes from a study of lightning storms in Colorado. It shows the distribution of the hour of the day during which the first lightning flash for that day occurred. The distribution has a single peak at noon and falls off on either side of this peak. The two sides of the histogram are roughly the same shape, so we call the distribution symmetric.

Figure 1.9 shows the distribution of lengths of words used in Shakespeare's plays.⁹ This distribution also has a single peak but is skewed to the right. That is, there are many short words (3 and 4 letters) and few very long words (10, 11, or 12 letters), so that the right tail of the histogram extends out much farther than the left tail.

Notice that the vertical scale in Figure 1.9 is not the *count* of words but the *percent* of all of Shakespeare's words that have each length. A histogram of percents rather than counts is convenient when the counts are very large or when we want to compare several distributions. Different kinds of writing have different distributions of word lengths, but all are right-skewed because short words are common and very long words are rare.

The overall shape of a distribution is important information about a variable. Some types of data regularly produce distributions that are symmetric or skewed. For example, the sizes of living things of the same species (like lengths of cockroaches) tend to be symmetric. Data on incomes (whether of individuals, companies, or nations) are usually strongly skewed to the right. There are many moderate incomes, some large incomes, and a few very large incomes. Do remember that

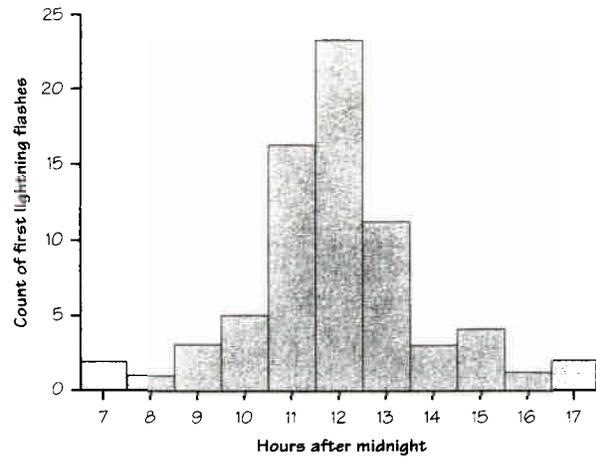


FIGURE 1.8 The distribution of the time of the first lightning flash each day at a site in Colorado, for Example 1.7.

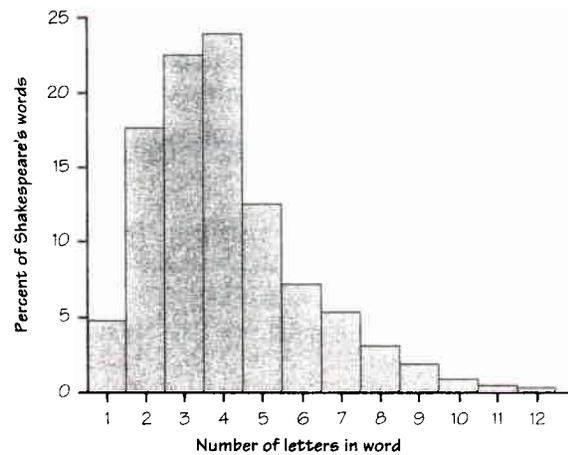


FIGURE 1.9 The distribution of lengths of words used in Shakespeare's plays, for Example 1.7.

many distributions have shapes that are neither symmetric nor skewed. Some data show other patterns. Scores on an exam, for example, may have a cluster near the top of the scale if many students did well. Or they may show two distinct peaks if a tough problem divided the class into those who did and didn't solve it. Use your eyes and describe what you see.

EXERCISES

1.16 STOCK RETURNS The total 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. Figure 1.10 is a histogram of the distribution of total returns for all 1528 stocks listed on the New York Stock Exchange in one year.¹⁰ Like

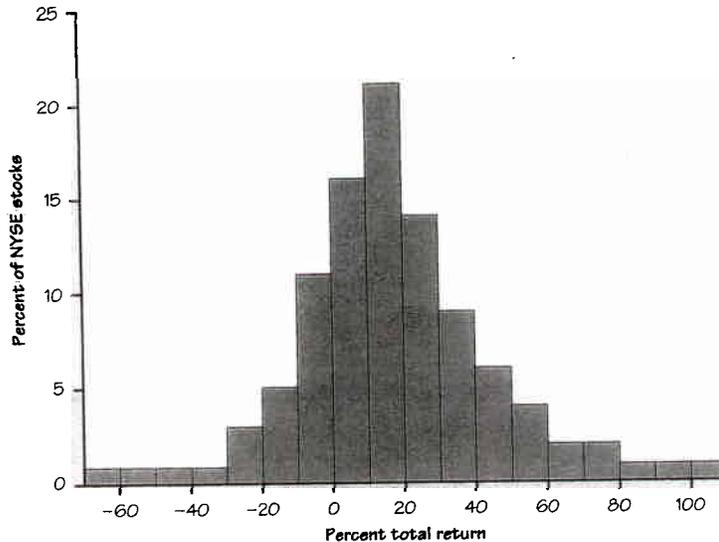


FIGURE 1.10 The distribution of percent total return for all New York Stock Exchange common stocks in one year.

Figure 1.9, it is a histogram of the percents in each class rather than a histogram of counts.

- Describe the overall shape of the distribution of total returns.
- What is the approximate center of this distribution? (For now, take the center to be the value with roughly half the stocks having lower returns and half having higher returns.)
- Approximately what were the smallest and largest total returns? (This describes the spread of the distribution.)
- A return less than zero means that an owner of the stock lost money. About what percent of all stocks lost money?

1.17 FREEZING IN GREENWICH, ENGLAND Figure 1.11 is a histogram of the number of days in the month of April on which the temperature fell below freezing at Greenwich, England.¹¹ The data cover a period of 65 years.

- Describe the shape, center, and spread of this distribution. Are there any outliers?
- In what percent of these 65 years did the temperature never fall below freezing in April?

1.18 How would you describe the center and spread of the distribution of first lightning flash times in Figure 1.8? Of the distribution of Shakespeare's word lengths in Figure 1.9?

Relative frequency, cumulative frequency, percentiles, and ogives

Sometimes we are interested in describing the relative position of an individual within a distribution. You may have received a standardized test score report that said you were in the 80th percentile. What does this mean? Put simply,

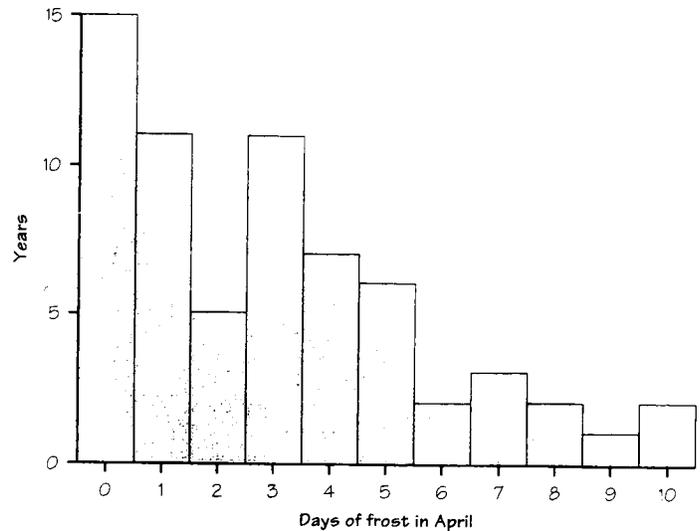


FIGURE 1.11 The distribution of the number of frost days during April at Greenwich, England, over a 65-year period, for Exercise 1.17.

80% of the people who took the test earned scores that were less than or equal to your score. The other 20% of students taking the test earned higher scores than you did.

PERCENTILE

The p th percentile of a distribution is the value such that p percent of the observations fall at or below it.

A histogram does a good job of displaying the distribution of values of a variable. But it tells us little about the relative standing of an individual observation. If we want this type of information, we should construct a **relative cumulative frequency graph**, often called an **ogive** (pronounced O-JIVE).

EXAMPLE 1.8 WAS BILL CLINTON A YOUNG PRESIDENT?

In Example 1.6, we made a histogram of the ages of U.S. presidents when they were inaugurated. Now we will examine where some specific presidents fall within the age distribution.

How to construct an ogive (relative cumulative frequency graph):

Step 1: Decide on class intervals and make a frequency table, just as in making a histogram. Add three columns to your frequency table: relative frequency, cumulative frequency, and relative cumulative frequency.

- To get the values in the *relative frequency* column, divide the count in each class interval by 43, the total number of presidents. Multiply by 100 to convert to a percentage.
- To fill in the *cumulative frequency* column, add the counts in the frequency column that fall in or below the current class interval.
- For the *relative cumulative frequency* column, divide the entries in the cumulative frequency column by 43, the total number of individuals.

Here is the frequency table from Example 1.6 with the relative frequency, cumulative frequency, and relative cumulative frequency columns added.

Class	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
40–44	2	$\frac{2}{43} = 0.047$, or 4.7%	2	$\frac{2}{43} = 0.047$, or 4.7%
45–49	6	$\frac{6}{43} = 0.140$, or 14.0%	8	$\frac{8}{43} = 0.186$, or 18.6%
50–54	13	$\frac{13}{43} = 0.302$, or 30.2%	21	$\frac{21}{43} = 0.488$, or 48.8%
55–59	12	$\frac{12}{43} = 0.279$, or 27.9%	33	$\frac{33}{43} = 0.767$, or 76.7%
60–64	7	$\frac{7}{43} = 0.163$, or 16.3%	40	$\frac{40}{43} = 0.930$, or 93.0%
65–69	3	$\frac{3}{43} = 0.070$, or 7.0%	43	$\frac{43}{43} = 1.000$, or 100%
TOTAL	43			

Step 2: Label and scale your axes and title your graph. Label the horizontal axis “Age at inauguration” and the vertical axis “Relative cumulative frequency.” Scale the horizontal axis according to your choice of class intervals and the vertical axis from 0% to 100%.

Step 3: Plot a point corresponding to the relative cumulative frequency in each class interval at the *left endpoint* of the *next* class interval. For example, for the 40–44 interval, plot a point at a height of 4.7% above the age value of 45. This means that 4.7% of presidents were inaugurated before they were 45 years old. Begin your ogive with a point at a height of 0% at the left endpoint of the lowest class interval. Connect consecutive points with a line segment to form the ogive. The last point you plot should be at a height of 100%. Figure 1.12 shows the completed ogive.

How to locate an individual within the distribution:

What about Bill Clinton? He was age 46 when he took office. To find his relative standing, draw a vertical line up from his age (46) on the horizontal axis until it meets the ogive. Then draw a horizontal line from this point of intersection to the vertical axis. Based on Figure 1.13(a), we would estimate that Bill Clinton’s age places him at the 10% *relative cumulative frequency* mark. That tells us that about 10% of all U.S. presidents were the same age as or younger than Bill Clinton when they were inaugurated. Put another way, President Clinton was younger than about 90% of all U.S. presidents based on his inauguration age. His age places him at the *10th percentile* of the distribution.

How to locate a value corresponding to a percentile:

- What inauguration age corresponds to the 60th percentile? To answer this question, draw a horizontal line across from the vertical axis at a height of 60% until it meets the ogive. From the point of intersection, draw a vertical line down to the horizontal axis.

In Figure 1.13(b), the value on the horizontal axis is about 57. So about 60% of all presidents were 57 years old or younger when they took office.

- Find the center of the distribution. Since we use the value that has half of the observations above it and half below it as our estimate of center, we simply need to find the 50th percentile of the distribution. Estimating as for the previous question, confirm that 55 is the center.

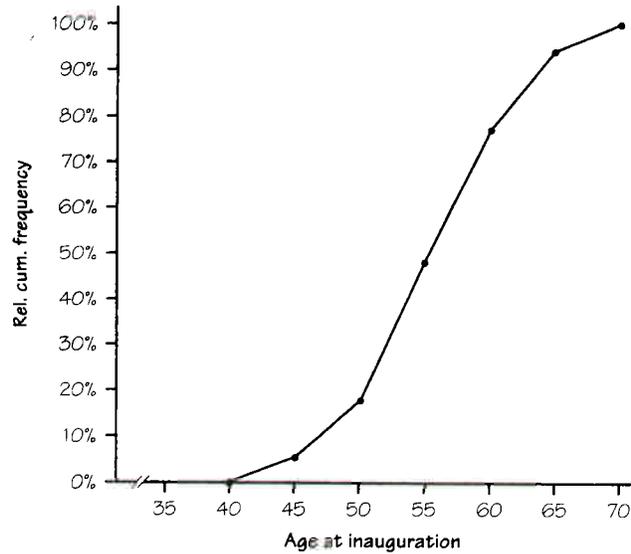


FIGURE 1.12 Relative cumulative frequency plot (ogive) for the ages of U.S. presidents at inauguration*

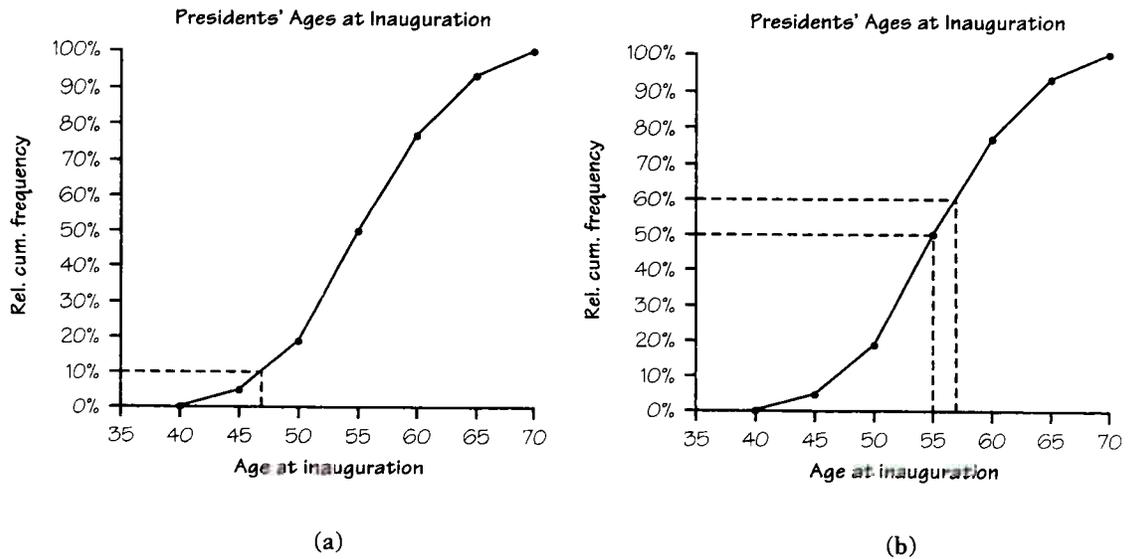


FIGURE 1.13 Ogives of presidents' ages at inauguration are used to (a) locate Bill Clinton within the distribution and (b) determine the 60th percentile and center of the distribution.

EXERCISES

1.19 OLDER FOLKS, II In Exercise 1.12 (page 22), you constructed a histogram of the percentage of people aged 65 or older in each state.

- Construct a relative cumulative frequency graph (ogive) for these data.
- Use your ogive from part (a) to answer the following questions:
 - In what percentage of states was the percentage of “65 and older” less than 15%?
 - What is the 40th percentile of this distribution, and what does it tell us?
 - What percentile is associated with your state?

1.20 SHOPPING SPREE, II Figure 1.14 is an ogive of the amount spent by grocery shoppers in Exercise 1.11 (page 18).

- Estimate the center of this distribution. Explain your method.
- At what percentile would the shopper who spent \$17.00 fall?
- Draw the histogram that corresponds to the ogive.

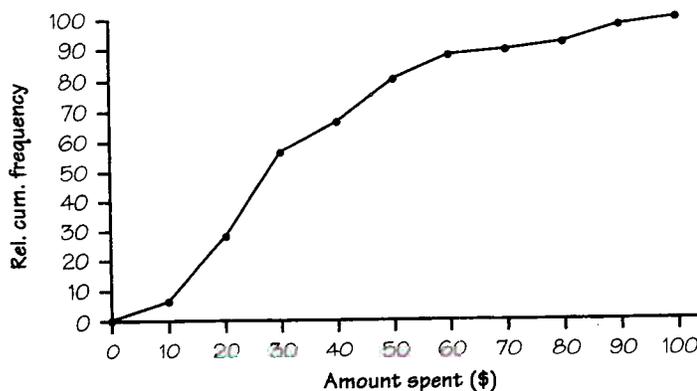


FIGURE 1.14 Amount spent by grocery shoppers in Exercise 1.11.

Time plots

Many variables are measured at intervals over time. We might, for example, measure the height of a growing child or the price of a stock at the end of each month. In these examples, our main interest is change over time. To display change over time, make a time plot.

TIME PLOT

A **time plot** of a variable plots each observation against the time at which it was measured. Always mark the time scale on the horizontal axis and the variable of interest on the vertical axis. If there are not too many points, connecting the points by lines helps show the pattern of changes over time.

trend

seasonal variation

When you examine a time plot, look once again for an overall pattern and for strong deviations from the pattern. One common overall pattern is a *trend*—a long-term upward or downward movement over time. A pattern that repeats itself at regular time intervals is known as *seasonal variation*. The next example illustrates both these patterns.

EXAMPLE 1.9 ORANGE PRICES MAKE ME SOUR!

Figure 1.15 is a time plot of the average price of fresh oranges over the period from January 1990 to January 2000. This information is collected each month as part of the government's reporting of retail prices. The vertical scale on the graph is the orange price index. This represents the price as a percentage of the average price of oranges in the years 1982 to 1984. The first value is 150 for January 1990, so at that time oranges cost about 150% of their 1982 to 1984 average price.

Figure 1.15 shows a clear *trend* of increasing price. In addition to this trend, we can see a strong *seasonal variation*, a regular rise and fall that occurs each year. Orange prices are usually highest in August or September, when the supply is lowest. Prices then fall in anticipation of the harvest and are lowest in January or February, when the harvest is complete and oranges are plentiful. The unusually large jump in orange prices in 1991 resulted from a freeze in Florida. Can you discover what happened in 1999?

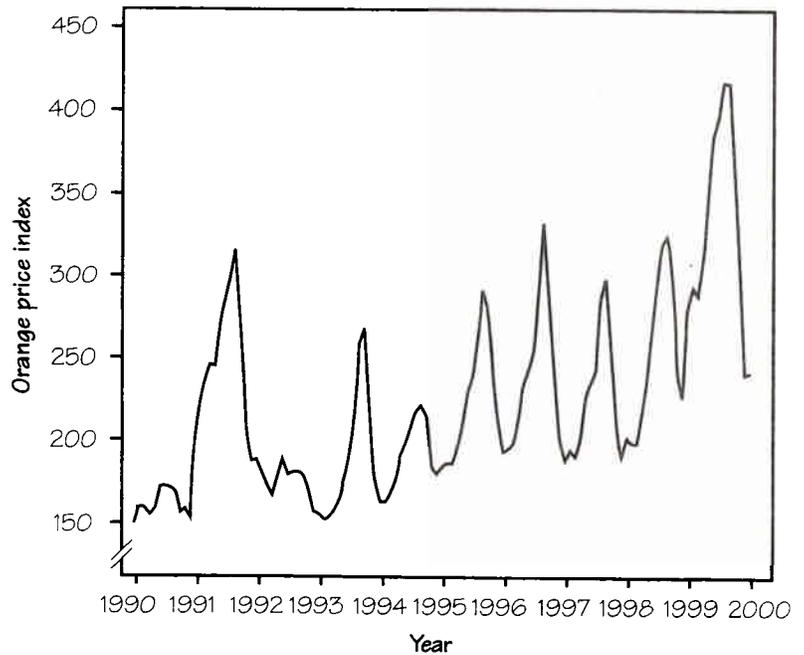


FIGURE 1.15 The price of fresh oranges, January 1990 to January 2000.

EXERCISES

1.21 **CANCER DEATHS** Here are data on the rate of deaths from cancer (deaths per 100,000 people) in the United States over the 50-year period from 1945 to 1995:

Year:	1945	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995
Deaths:	134.0	139.8	146.5	149.2	153.5	162.8	169.7	183.9	193.3	203.2	204.7

- (a) Construct a time plot for these data. Describe what you see in a few sentences.
 (b) Do these data suggest that we have made no progress in treating cancer? Explain.

1.22 **CIVIL UNREST** The years around 1970 brought unrest to many U.S. cities. Here are data on the number of civil disturbances in each three month period during the years 1968 to 1972:

Period	Count	Period	Count
1968 Jan.–Mar.	6	1970 July–Sept.	20
Apr.–June	46	Oct.–Dec.	6
July–Sept.	25	1971 Jan.–Mar.	12
Oct.–Dec.	3	Apr.–June	21
1969 Jan.–Mar.	5	July–Sept.	5
Apr.–June	27	Oct.–Dec.	1
July–Sept.	19	1972 Jan.–Mar.	3
Oct.–Dec.	6	Apr.–June	8
1970 Jan.–Mar.	26	July–Sept.	5
Apr.–June	24	Oct.–Dec.	5

- (a) Make a time plot of these counts. Connect the points in your plot by straight-line segments to make the pattern clearer.
 (b) Describe the trend and the seasonal variation in this time series. Can you suggest an explanation for the seasonal variation in civil disorders?

SUMMARY

A data set contains information on a number of **individuals**. Individuals may be people, animals, or things. For each individual, the data give values for one or more **variables**. A variable describes some characteristic of an individual, such as a person's height, gender, or salary.

Exploratory data analysis uses graphs and numerical summaries to describe the variables in a data set and the relations among them.

Some variables are **categorical** and others are **quantitative**. A categorical variable places each individual into a category, like male or female. A quantitative variable has numerical values that measure some characteristic of each individual, like height in centimeters or annual salary in dollars.

The **distribution** of a variable describes what values the variable takes and how often it takes these values.

To describe a distribution, begin with a graph. Use **bar graphs** and **p. charts** to display categorical variables. **Dotplots**, **stemplots**, and **histograms** graph the distributions of quantitative variables. An **ogive** can help you determine relative standing within a quantitative distribution.

When examining any graph, look for an **overall pattern** and for notable **deviations** from the pattern.

The **center**, **spread**, and **shape** describe the overall pattern of a distribution. Some distributions have simple shapes, such as **symmetric** and **skewed**. Not all distributions have a simple overall shape, especially when there are few observations.

Outliers are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.

When observations on a variable are taken over time, make a **time plot** that graphs time horizontally and the values of the variable vertically. A time plot can reveal **trends**, **seasonal variations**, or other changes over time.

SECTION 1.1 EXERCISES

1.23 GENDER EFFECTS IN VOTING Political party preference in the United States depends in part on the age, income, and gender of the voter. A political scientist selects a large sample of registered voters. For each voter, she records gender, age, household income, and whether they voted for the Democratic or for the Republican candidate in the last congressional election. Which of these variables are categorical and which are quantitative?

1.24 What type of graph or graphs would you plan to make in a study of each of the following issues?

- (a) What makes of cars do students drive? How old are their cars?
- (b) How many hours per week do students study? How does the number of study hours change during a semester?
- (c) Which radio stations are most popular with students?

1.25 MURDER WEAPONS The 1999 *Statistical Abstract of the United States* reports FBI data on murders for 1997. In that year, 53.3% of all murders were committed with handguns, 14.5% with other firearms, 13.0% with knives, 6.3% with a part of the body (usually the hands or feet), and 4.6% with blunt objects. Make a graph to display these data. Do you need an “other methods” category?

1.26 WHAT'S A DOLLAR WORTH THESE DAYS? The buying power of a dollar changes over time. The Bureau of Labor Statistics measures the cost of a “market basket” of goods and services to compile its Consumer Price Index (CPI). If the CPI is 120, goods and services that cost \$100 in the base period now cost \$120. Here are the yearly average values of the CPI for the years between 1970 and 1999. The base period is the years 1982 to 1984.

Year	CPI	Year	CPI	Year	CPI	Year	CPI
1970	38.8	1978	65.2	1986	109.6	1994	148.2
1972	41.8	1980	82.4	1988	118.3	1996	156.9
1974	49.3	1982	96.5	1990	130.7	1998	163.0
1976	56.9	1984	103.9	1992	140.3	1999	166.6

- (a) Construct a graph that shows how the CPI has changed over time.
- (b) Check your graph by doing the plot on your calculator.
- Enter the years (the last two digits will suffice) into $L_1/\text{list1}$ and enter the CPI into $L_2/\text{list2}$.
 - Then set up a statistics plot, choosing the plot type “xyline” (the second type on the TI-83). Use $L_1/\text{list1}$ as X and $L_2/\text{list2}$ as Y. In this graph, the data points are plotted and connected in order of appearance in $L_1/\text{list1}$ and $L_2/\text{list2}$.
 - Use the zoom command to see the graph.
- (c) What was the overall trend in prices during this period? Were there any years in which this trend was reversed?
- (d) In what period during these decades were prices rising fastest? In what period were they rising slowest?

1.27 THE STATISTICS OF WRITING STYLE Numerical data can distinguish different types of writing, and sometimes even individual authors. Here are data on the percent of words of 1 to 15 letters used in articles in *Popular Science* magazine:¹²

Length:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Percent:	3.6	14.8	18.7	16.0	12.5	8.2	8.1	5.9	4.4	3.6	2.1	0.9	0.6	0.4	0.2

- (a) Make a histogram of this distribution. Describe its shape, center, and spread.
- (b) How does the distribution of lengths of words used in *Popular Science* compare with the similar distribution in Figure 1.9 (page 26) for Shakespeare’s plays? Look in particular at short words (2, 3, and 4 letters) and very long words (more than 10 letters).

1.28 DENSITY OF THE EARTH In 1798 the English scientist Henry Cavendish measured the density of the earth by careful work with a torsion balance. The variable recorded was the density of the earth as a multiple of the density of water. Here are Cavendish’s 29 measurements:¹³

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

Present these measurements graphically in a stemplot. Discuss the shape, center, and spread of the distribution. Are there any outliers? What is your estimate of the density of the earth based on these measurements?

1.29 DRIVE TIME Professor Moore, who lives a few miles outside a college town, records the time he takes to drive to the college each morning. Here are the times (in minutes) for 42 consecutive weekdays, with the dates in order along the rows:

8.25	7.83	8.30	8.42	8.50	8.67	8.17	9.00	9.00	8.17	7.92
9.00	8.50	9.00	7.75	7.92	8.00	8.08	8.42	8.75	8.08	9.75
8.33	7.83	7.92	8.58	7.83	8.42	7.75	7.42	6.75	7.42	8.50
8.67	10.17	8.75	8.58	8.67	9.17	9.08	8.83	8.67		

- Make a histogram of these drive times. Is the distribution roughly symmetric, clearly skewed, or neither? Are there any clear outliers?
- Construct an ogive for Professor Moore's drive times.
- Use your ogive from (b) to estimate the center and 90th percentile for the distribution.
- Use your ogive to estimate the percentile corresponding to a drive time of 8.00 minutes.

1.30 THE SPEED OF LIGHT Light travels fast, but it is not transmitted instantaneously. Light takes over a second to reach us from the moon and over 10 billion years to reach us from the most distant objects observed so far in the expanding universe. Because radio and radar also travel at the speed of light, an accurate value for that speed is important in communicating with astronauts and orbiting satellites. An accurate value for the speed of light is also important to computer designers because electrical signals travel at light speed. The first reasonably accurate measurements of the speed of light were made over 100 years ago by A. A. Michelson and Simon Newcomb. Table 1.7 contains 66 measurements made by Newcomb between July and September 1882.

Newcomb measured the time in seconds that a light signal took to pass from his laboratory on the Potomac River to a mirror at the base of the Washington Monument and back, a total distance of about 7400 meters. Just as you can compute the speed of a car from the time required to drive a mile, Newcomb could compute the speed of light from the passage time. Newcomb's first measurement of the passage time of light was 0.000024828 second, or 24,828 nanoseconds. (There are 10^9 nanoseconds in a second.) The entries in Table 1.7 record only the deviation from 24,800 nanoseconds.

TABLE 1.7 Newcomb's measurements of the passage time of light

28	26	33	24	34	-44	27	16	40	-2	29	22	24	21
25	30	23	29	31	19	24	20	36	32	36	28	25	21
28	29	37	25	28	26	30	32	36	26	30	22	36	23
27	27	28	27	31	27	26	33	26	32	32	24	39	28
24	25	32	25	29	27	28	29	16	23				

Source: S. M. Stigler, "Do robust estimators work with real data?" *Annals of Statistics*, 5 (1977), pp. 1055-1078.

- Construct an appropriate graphical display for these data. Justify your choice of graph.
- Describe the distribution of Newcomb's speed of light measurements.