EXERCISES

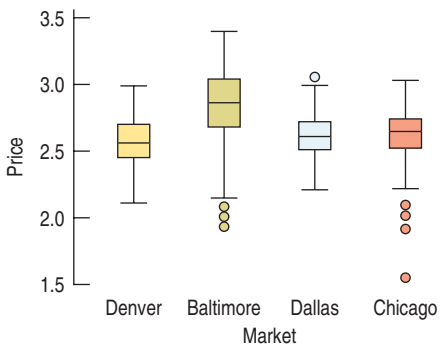
1. **In the news.** Find an article in a newspaper, magazine, or the Internet that compares two or more groups of data.
   a) Does the article discuss the W's?
   b) Is the chosen display appropriate? Explain.
   c) Discuss what the display reveals about the groups.
   d) Does the article accurately describe and interpret the data? Explain.

2. **In the news.** Find an article in a newspaper, magazine, or the Internet that shows a time plot.
   a) Does the article discuss the W's?
   b) Is the timeplot appropriate for the data? Explain.
   c) Discuss what the timeplot reveals about the variable.
   d) Does the article accurately describe and interpret the data? Explain.

3. **Time on the Internet.** Find data on the Internet (or elsewhere) that give results recorded over time. Make an appropriate display and discuss what it shows.

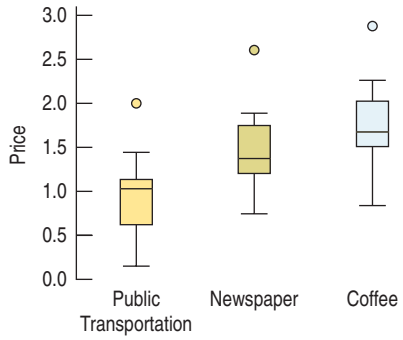
4. **Groups on the Internet.** Find data on the Internet (or elsewhere) for two or more groups. Make appropriate displays to compare the groups, and interpret what you find.

T 5. **Pizza prices.** A company that sells frozen pizza to stores in four markets in the United States (Denver, Baltimore, Dallas, and Chicago) wants to examine the prices that the stores charge for pizza slices. Here are boxplots comparing data from a sample of stores in each market:



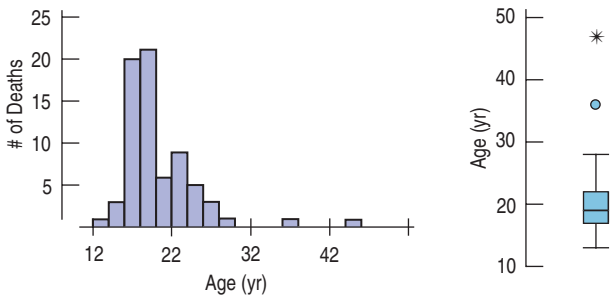
   a) Do prices appear to be the same in the four markets? Explain.
   b) Does the presence of any outliers affect your overall conclusions about prices in the four markets?

T 6. **Costs.** To help travelers know what to expect, researchers collected the prices of commodities in 16 cities throughout the world. Here are boxplots comparing the prices of a ride on public transportation, a newspaper, and a cup of coffee in the 16 cities (prices are all in $US).



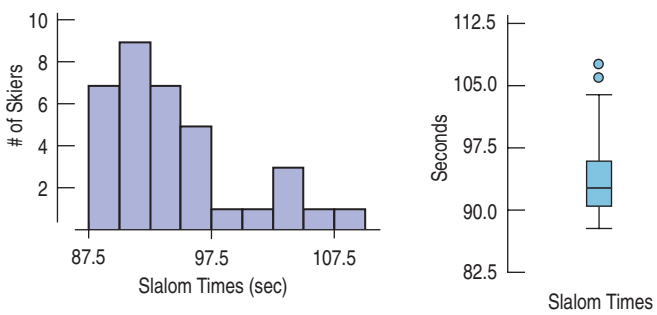
   a) On average, which commodity is the most expensive?
   b) Is a newspaper always more expensive than a ride on public transportation? Explain.
   c) Does the presence of outliers affect your conclusions in a) or b)?

T 7. **Still rockin'.** Crowd Management Strategies monitors accidents at rock concerts. In their database, they list the names and other variables of victims whose deaths were attributed to "crowd crush" at rock concerts. Here are the histogram and boxplot of the victims' ages for data from 1999 to 2000:



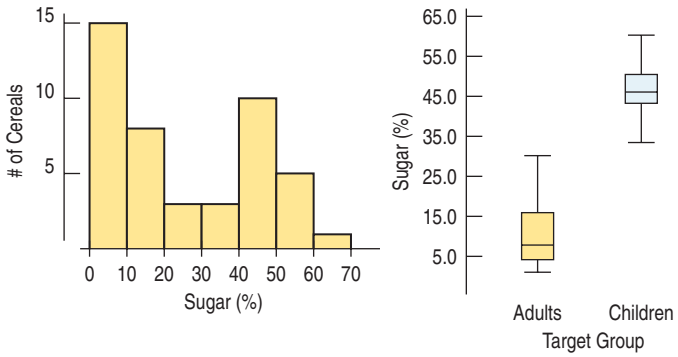
   a) What features of the distribution can you see in both the histogram and the boxplot?
   b) What features of the distribution can you see in the histogram that you could not see in the boxplot?
   c) What summary statistic would you choose to summarize the center of this distribution? Why?
   d) What summary statistic would you choose to summarize the spread of this distribution? Why?

T 8. **Slalom times.** The Men's Combined skiing event consists of a downhill and a slalom. Here are two displays of the slalom times in the Men's Combined at the 2006 Winter Olympics:



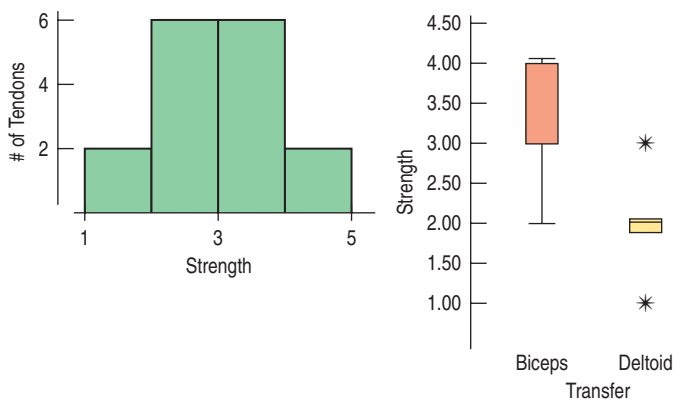
a) What features of the distribution can you see in both the histogram and the boxplot?
b) What features of the distribution can you see in the histogram that you could not see in the boxplot?
c) What summary statistic would you choose to summarize the center of this distribution? Why?
d) What summary statistic would you choose to summarize the spread of this distribution? Why?

9. Cereals. Sugar is a major ingredient in many breakfast cereals. The histogram displays the sugar content as a percentage of weight for 49 brands of cereal. The boxplot compares sugar content for adult and children's cereals.



a) What is the range of the sugar contents of these cereals.
b) Describe the shape of the distribution.
c) What aspect of breakfast cereals might account for this shape?
d) Are all children's cereals higher in sugar than adult cereals?
e) Which group of cereals varies more in sugar content? Explain.

10. Tendon transfers. People with spinal cord injuries may lose function in some, but not all, of their muscles. The ability to push oneself up is particularly important for shifting position when seated and for transferring into and out of wheelchairs. Surgeons compared two operations to restore the ability to push up in children. The histogram shows scores rating pushing strength two years after surgery and boxplots compare results for the two surgical methods. (Mulcahey, Lutz, Kozen, Betz, "Prospective Evaluation of Biceps to Triceps and Deltoid to Triceps for Elbow Extension in Tetraplegia," *Journal of Hand Surgery*, 28, 6, 2003)



a) Describe the shape of this distribution.
b) What is the range of the strength scores?
c) What fact about results of the two procedures is hidden in the histogram?
d) Which method had the higher (better) median score?
e) Was that method always best?
f) Which method produced the most consistent results? Explain.

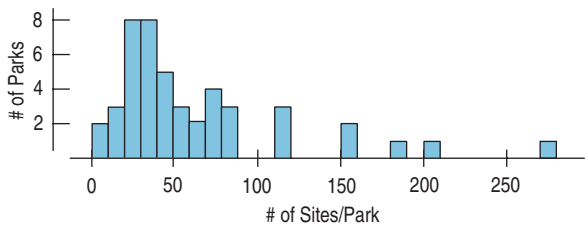
11. Population growth. Here is a "back-to-back" stem-and-leaf display that shows two data sets at once—one going to the left, one to the right. The display compares the percent change in population for two regions of the United States (based on census figures for 1990 and 2000). The fastest growing states were Nevada at 66% and Arizona at 40%. To show the distributions better, this display breaks each stem into two lines, putting leaves 0–4 on one stem and leaves 5–9 on the other.

NE/MW States		S/W States
	6	6
	6	
	5	
	5	
	4	
	4	0
	3	
	3	001
	2	6
	2	001134
	1	578
2100	1	001134444
99998876655	0	6999
4431	0	1

Population Growth rate
(|6| 6 means 66%)

a) Use the data displayed in the stem-and-leaf display to construct comparative boxplots.
b) Write a few sentences describing the difference in growth rates for the two regions of the United States.

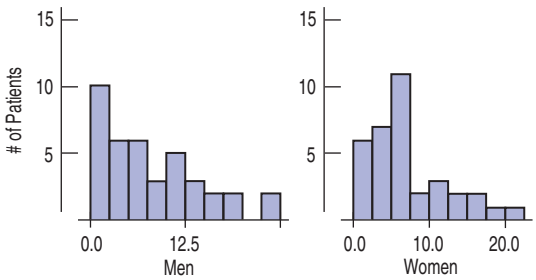
12. Camp sites. Shown below are the histogram and summary statistics for the number of camp sites at public parks in Vermont.



Count	46
Mean	62.8 sites
Median	43.5
StdDev	56.2
Min	0
Max	275
Q1	28
Q3	78

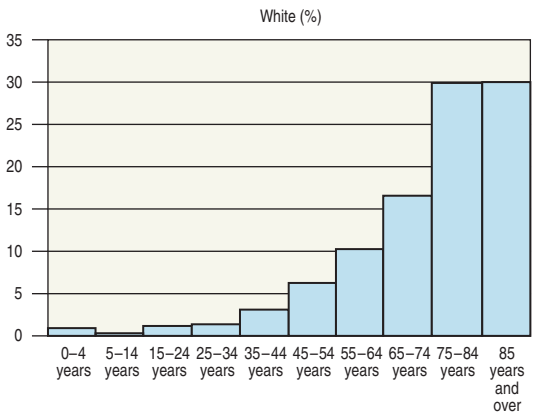
a) Which statistics would you use to identify the center and spread of this distribution? Why?
b) How many parks would you classify as outliers? Explain.
c) Create a boxplot for these data.
d) Write a few sentences describing the distribution.

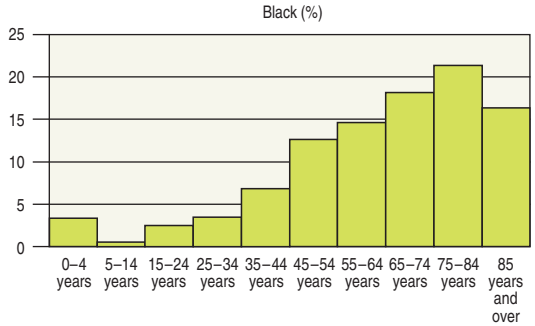
13. Hospital stays. The U.S. National Center for Health Statistics compiles data on the length of stay by patients in short-term hospitals and publishes its findings in *Vital and Health Statistics*. Data from a sample of 39 male patients and 35 female patients on length of stay (in days) are displayed in the histograms below.



a) What would you suggest be changed about these histograms to make them easier to compare?
b) Describe these distributions by writing a few sentences comparing the duration of hospitalization for men and women.
c) Can you suggest a reason for the peak in women's length of stay?

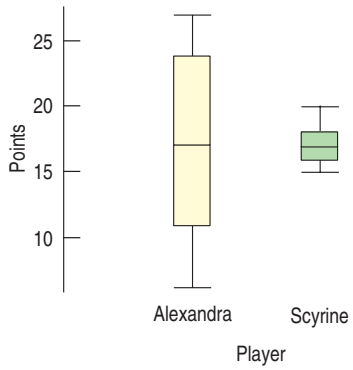
14. Deaths 2003. A National Vital Statistics Report (www.cdc.gov/nchs/) indicated that nearly 300,000 black Americans died in 2003, compared with just over 2 million white Americans. Here are histograms displaying the distributions of their ages at death:





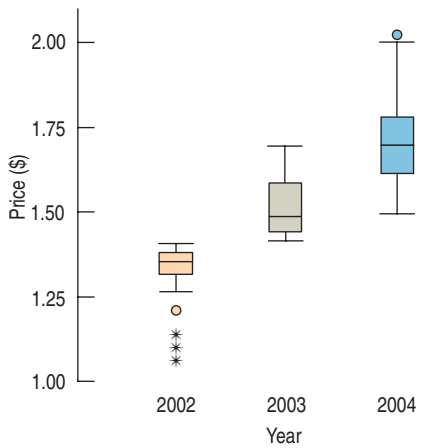
a) Describe the overall shapes of these distributions.
b) How do the distributions differ?
c) Look carefully at the bar definitions. Where do these plots violate the rules for statistical graphs?

T **15. Women's basketball.** Here are boxplots of the points scored during the first 10 games of the season for both Scyrine and Alexandra:



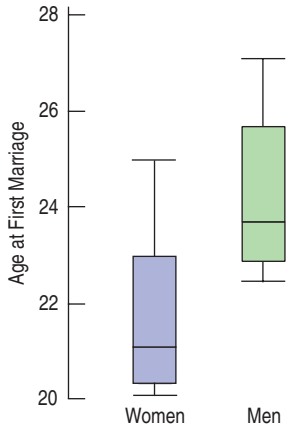
a) Summarize the similarities and differences in their performance so far.
b) The coach can take only one player to the state championship. Which one should she take? Why?

16. Gas prices. Here are boxplots of weekly gas prices at a service station in the midwestern United States (prices in $ per gallon):

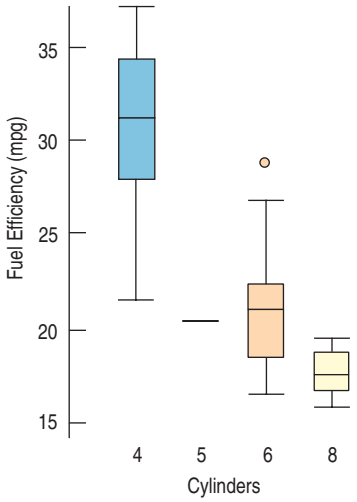


a) Compare the distribution of prices over the three years.
b) In which year were the prices least stable? Explain.

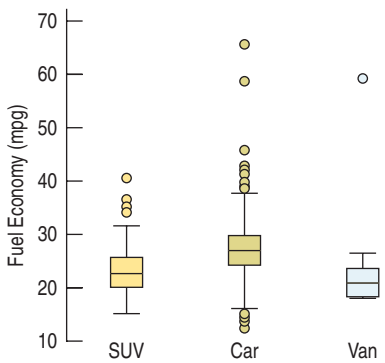
T **17. Marriage age.** In 1975, did men and women marry at the same age? Here are boxplots of the age at first marriage for a sample of U.S. citizens then. Write a brief report discussing what these data show.



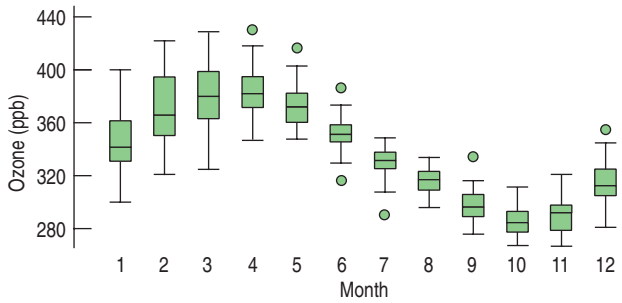
18. Fuel economy. Describe what these boxplots tell you about the relationship between the number of cylinders a car's engine has and the car's fuel economy (mpg):



19. Fuel economy II. The Environmental Protection Agency provides fuel economy and pollution information on over 2000 car models. Here is a boxplot of *Combined Fuel Economy* (using an average of driving conditions) in *miles per gallon* by vehicle *Type* (car, van, or SUV). Summarize what you see about the fuel economies of the three vehicle types.

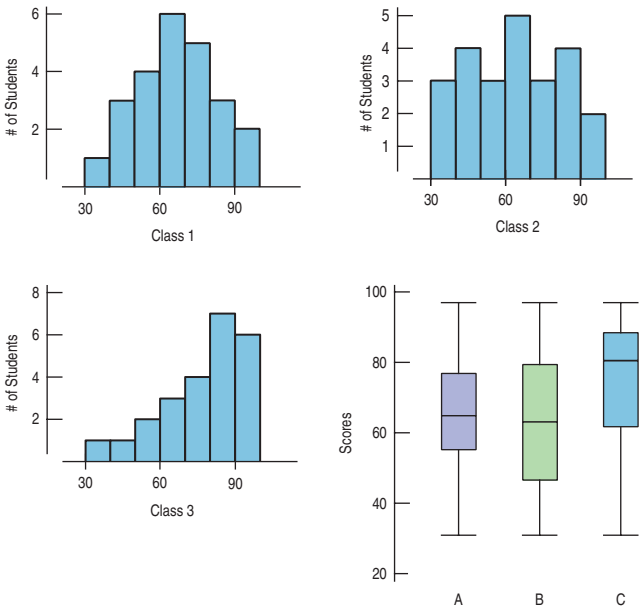


20. Ozone. Ozone levels (in parts per billion, ppb) were recorded at sites in New Jersey monthly between 1926 and 1971. Here are boxplots of the data for each month (over the 46 years), lined up in order (January = 1):



a) In what month was the highest ozone level ever recorded?
b) Which month has the largest IQR?
c) Which month has the smallest range?
d) Write a brief comparison of the ozone levels in January and June.
e) Write a report on the annual patterns you see in the ozone levels.

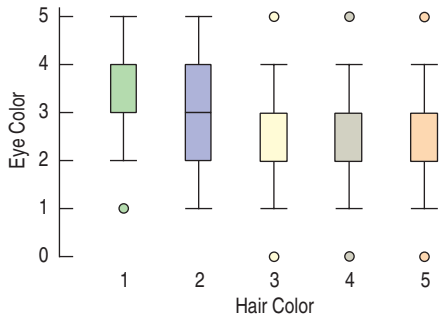
21. Test scores. Three Statistics classes all took the same test. Histograms and boxplots of the scores for each class are shown below. Match each class with the corresponding boxplot.



22. Eye and hair color. A survey of 1021 school-age children was conducted by randomly selecting children from several large urban elementary schools. Two of the questions concerned eye and hair color. In the survey, the following codes were used:

Hair Color	Eye Color
1 = Blond	1 = Blue
2 = Brown	2 = Green
3 = Black	3 = Brown
4 = Red	4 = Grey
5 = Other	5 = Other

The Statistics students analyzing the data were asked to study the relationship between eye and hair color. They produced this plot:



Is their graph appropriate? If so, summarize the findings. If not, explain why not.

23. Graduation? A survey of major universities asked what percentage of incoming freshmen usually graduate "on time" in 4 years. Use the summary statistics given to answer the questions that follow.

	% on Time
Count	48
Mean	68.35
Median	69.90
StdDev	10.20
Min	43.20
Max	87.40
Range	44.20
25th %tile	59.15
75th %tile	74.75

a) Would you describe this distribution as symmetric or skewed? Explain.
b) Are there any outliers? Explain.
c) Create a boxplot of these data.
d) Write a few sentences about the graduation rates.

24. Vineyards. Here are summary statistics for the sizes (in acres) of Finger Lakes vineyards:

Count	36
Mean	46.50 acres
StdDev	47.76
Median	33.50
IQR	36.50
Min	6
Q1	18.50
Q3	55
Max	250

a) Would you describe this distribution as symmetric or skewed? Explain.
b) Are there any outliers? Explain.
c) Create a boxplot of these data.
d) Write a few sentences about the sizes of the vineyards.

25. Caffeine. A student study of the effects of caffeine asked volunteers to take a memory test 2 hours after drinking soda. Some drank caffeine-free cola, some drank regular cola (with caffeine), and others drank a mixture of the two (getting a half-dose of caffeine). Here are the 5-number summaries for each group's scores (number of items recalled correctly) on the memory test:

	n	Min	Q1	Median	Q3	Max
No caffeine	15	16	20	21	24	26
Low caffeine	15	16	18	21	24	27
High caffeine	15	12	17	19	22	24

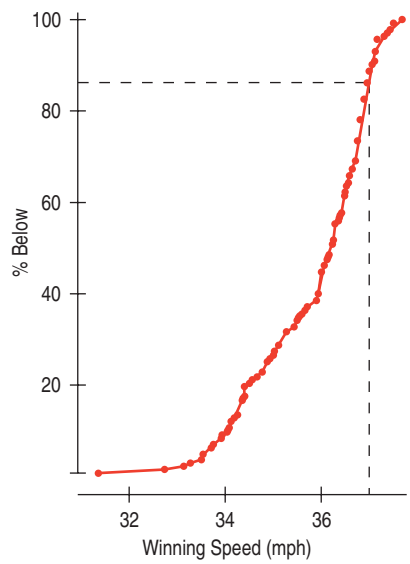
a) Describe the W's for these data.
b) Name the variables and classify each as categorical or quantitative.
c) Create parallel boxplots to display these results as best you can with this information.
d) Write a few sentences comparing the performances of the three groups.

26. SAT scores. Here are the summary statistics for Verbal SAT scores for a high school graduating class:

	n	Mean	Median	SD	Min	Max	Q1	Q3
Male	80	590	600	97.2	310	800	515	650
Female	82	602	625	102.0	360	770	530	680

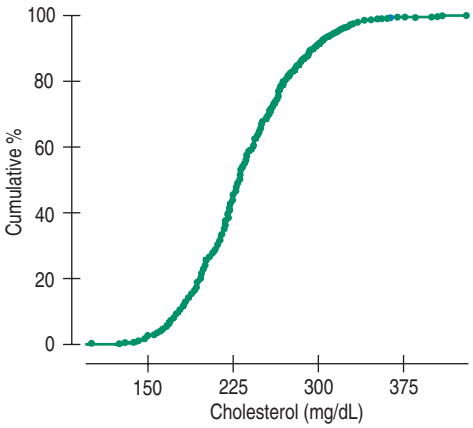
a) Create parallel boxplots comparing the scores of boys and girls as best you can from the information given.
b) Write a brief report on these results. Be sure to discuss the shape, center, and spread of the scores.

27. Derby speeds 2007. How fast do horses run? Kentucky Derby winners top 30 miles per hour, as shown in this graph. The graph shows the percentage of Derby winners that have run *slower* than each given speed. Note that few have won running less than 33 miles per hour, but about 86% of the winning horses have run less than 37 miles per hour. (A cumulative frequency graph like this is called an "ogive.")



a) Estimate the median winning speed.
b) Estimate the quartiles.
c) Estimate the range and the IQR.
d) Create a boxplot of these speeds.
e) Write a few sentences about the speeds of the Kentucky Derby winners.

28. Cholesterol. The Framingham Heart Study recorded the cholesterol levels of more than 1400 men. Here is an ogive of the distribution of these cholesterol measures. (An ogive shows the percentage of cases at or below a certain value.) Construct a boxplot for these data, and write a few sentences describing the distribution.



29. Reading scores. A class of fourth graders takes a diagnostic reading test, and the scores are reported by reading grade level. The 5-number summaries for the 14 boys and 11 girls are shown:

Boys:	2.0	3.9	4.3	4.9	6.0
Girls:	2.8	3.8	4.5	5.2	5.9

a) Which group had the highest score?
b) Which group had the greater range?
c) Which group had the greater interquartile range?
d) Which group's scores appear to be more skewed? Explain.
e) Which group generally did better on the test? Explain.
f) If the mean reading level for boys was 4.2 and for girls was 4.6, what is the overall mean for the class?

30. Rainmakers? In an experiment to determine whether seeding clouds with silver iodide increases rainfall, 52 clouds were randomly assigned to be seeded or not. The amount of rain they generated was then measured (in acre-feet). Here are the summary statistics:

	n	Mean	Median	SD	IQR	Q1	Q3
Unseeded	26	164.59	44.20	278.43	138.60	24.40	163
Seeded	26	441.98	221.60	650.79	337.60	92.40	430

a) Which of the summary statistics are most appropriate for describing these distributions. Why?
b) Do you see any evidence that seeding clouds may be effective? Explain.

31. Industrial experiment. Engineers at a computer production plant tested two methods for accuracy in drilling holes into a PC board. They tested how fast they could set the drilling machine by running 10 boards at each of two different speeds. To assess the results, they measured the distance (in inches) from the center of a target on the board to the center of the hole. The data and summary statistics are shown in the table:

	Distance (in.)	Speed		Distance (in.)	Speed
	0.000101	Fast		0.000098	Slow
	0.000102	Fast		0.000096	Slow
	0.000100	Fast		0.000097	Slow
	0.000102	Fast		0.000095	Slow
	0.000101	Fast		0.000094	Slow
	0.000103	Fast		0.000098	Slow
	0.000104	Fast		0.000096	Slow
	0.000102	Fast		0.975600	Slow
	0.000102	Fast		0.000097	Slow
	0.000100	Fast		0.000096	Slow
Mean	0.000102		Mean	0.097647	
StdDev	0.000001		StdDev	0.308481	

Write a report summarizing the findings of the experiment. Include appropriate visual and verbal displays of the distributions, and make a recommendation to the engineers if they are most interested in the accuracy of the method.

32. Cholesterol. A study examining the health risks of smoking measured the cholesterol levels of people who had smoked for at least 25 years and people of similar ages who had smoked for no more than 5 years and then stopped. Create appropriate graphical displays for both groups, and write a brief report comparing their cholesterol levels. Here are the data:

Smokers				Ex-Smokers		
225	211	209	284	250	134	300
258	216	196	288	249	213	310
250	200	209	280	175	174	328
225	256	243	200	160	188	321
213	246	225	237	213	257	292
232	267	232	216	200	271	227
216	243	200	155	238	163	263
216	271	230	309	192	242	249
183	280	217	305	242	267	243
287	217	246	351	217	267	218
200	280	209		217	183	228

33. MPG. A consumer organization compared gas mileage figures for several models of cars made in the United States with autos manufactured in other countries. The data are shown in the table:

Gas Mileage (mpg)	Country	Gas Mileage (mpg)	Country
16.9	U.S.	26.8	U.S.
15.5	U.S.	33.5	U.S.
19.2	U.S.	34.2	U.S.
18.5	U.S.	16.2	Other
30.0	U.S.	20.3	Other
30.9	U.S.	31.5	Other
20.6	U.S.	30.5	Other
20.8	U.S.	21.5	Other
18.6	U.S.	31.9	Other
18.1	U.S.	37.3	Other
17.0	U.S.	27.5	Other
17.6	U.S.	27.2	Other
16.5	U.S.	34.1	Other
18.2	U.S.	35.1	Other
26.5	U.S.	29.5	Other
21.9	U.S.	31.8	Other
27.4	U.S.	22.0	Other
28.4	U.S.	17.0	Other
28.8	U.S.	21.6	Other

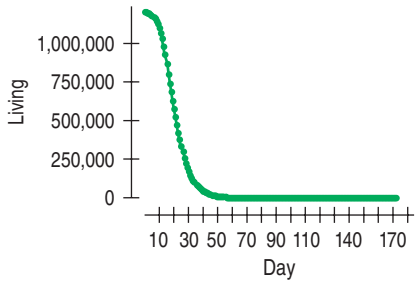
a) Create graphical displays for these two groups.
b) Write a few sentences comparing the distributions.

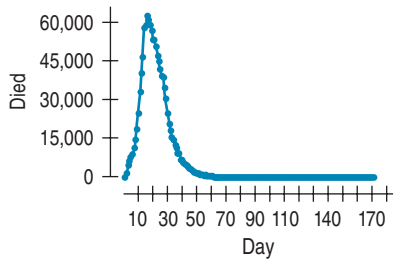
34. Baseball. American League baseball teams play their games with the designated hitter rule, meaning that pitchers do not bat. The League believes that replacing the pitcher, typically a weak hitter, with another player in the batting order produces more runs and generates more interest among fans. Following are the average number of runs scored in American League and National League stadiums for the first half of the 2001 season:

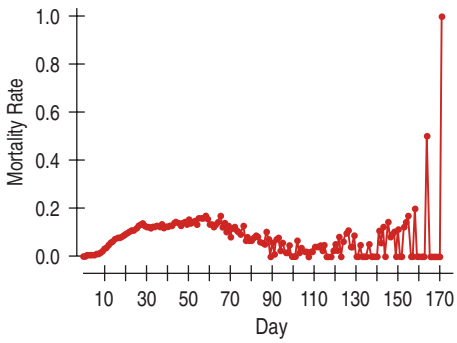
Average Runs	League	Average Runs	League
11.1	American	14.0	National
10.8	American	11.6	National
10.8	American	10.4	National
10.3	American	10.9	National
10.3	American	10.2	National
10.1	American	9.5	National
10.0	American	9.5	National
9.5	American	9.5	National
9.4	American	9.5	National
9.3	American	9.1	National
9.2	American	8.8	National
9.2	American	8.4	National
9.0	American	8.3	National
8.3	American	8.2	National
		8.1	National
		7.9	National

a) Create an appropriate graphical display of these data.
b) Write a few sentences comparing the average number of runs scored per game in the two leagues. (Remember: shape, center, spread, unusual features!)
c) Coors Field in Denver stands a mile above sea level, an altitude far greater than that of any other major league ball park. Some believe that the thinner air makes it harder for pitchers to throw curveballs and easier for batters to hit the ball a long way. Do you see any evidence that the 14 runs scored per game there is unusually high? Explain.

35. Fruit Flies. Researchers tracked a population of 1,203,646 fruit flies, counting how many died each day for 171 days. Here are three timeplots offering different views of these data. One shows the number of flies alive on each day, one the number who died that day, and the third the mortality rate—the fraction of the number alive who died. On the last day studied, the last 2 flies died, for a mortality rate of 1.0.







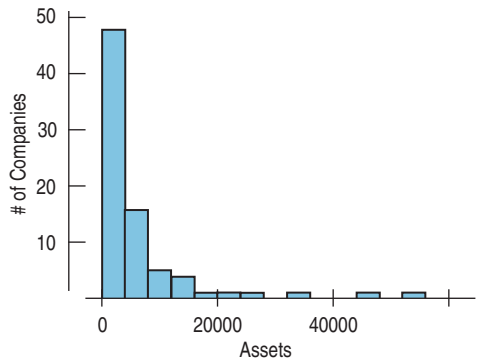
a) On approximately what day did the most flies die?
b) On what day during the first 100 days did the largest *proportion* of flies die?
c) When did the number of fruit flies alive stop changing very much from day to day?

36. Drunk driving 2005. Accidents involving drunk drivers account for about 40% of all deaths on the nation's highways. The table tracks the number of alcohol-related fatalities for 24 years. (www.madd.org)

Year	Deaths (thousands)	Year	Deaths (thousands)
1982	26.2	**1994**	17.3
1983	24.6	**1995**	17.7
1984	24.8	**1996**	17.7
1985	23.2	**1997**	16.7
1986	25.0	**1998**	16.7
1987	24.1	**1999**	16.6
1988	23.8	**2000**	17.4
1989	22.4	**2001**	17.4
1990	22.6	**2002**	17.5
1991	20.2	**2003**	17.1
1992	18.3	**2004**	16.9
1993	17.9	**2005**	16.9

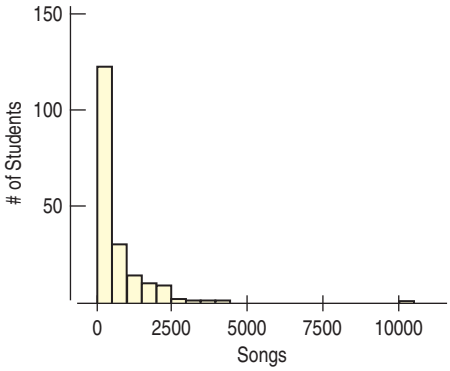
a) Create a stem-and-leaf display or a histogram of these data.
b) Create a timeplot.
c) Using features apparent in the stem-and-leaf display (or histogram) and the timeplot, write a few sentences about deaths caused by drunk driving.

37. Assets. Here is a histogram of the assets (in millions of dollars) of 79 companies chosen from the *Forbes* list of the nation's top corporations:



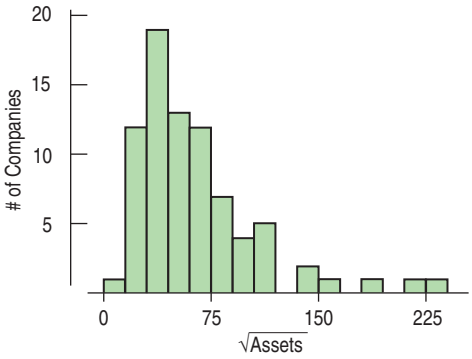
a) What aspect of this distribution makes it difficult to summarize, or to discuss, center and spread?
b) What would you suggest doing with these data if we want to understand them better?

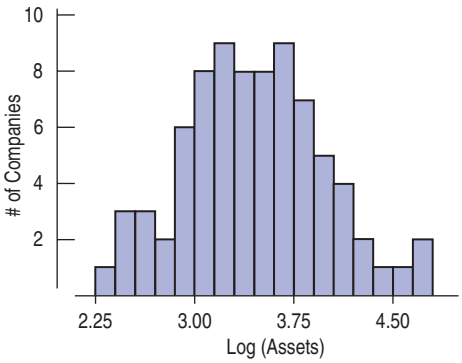
38. Music library. Students were asked how many songs they had in their digital music libraries. Here's a display of the responses:



a) What aspect of this distribution makes it difficult to summarize, or to discuss, center and spread?
b) What would you suggest doing with these data if we want to understand them better?

39. Assets again. Here are the same data you saw in Exercise 37 after re-expressions as the square root of assets and the logarithm of assets:





a) Which re-expression do you prefer? Why?
b) In the square root re-expression, what does the value 50 actually indicate about the company's assets?
c) In the logarithm re-expression, what does the value 3 actually indicate about the company's assets?

40. Rainmakers. The table lists the amount of rainfall (in acre-feet) from the 26 clouds seeded with silver iodide discussed in Exercise 30:

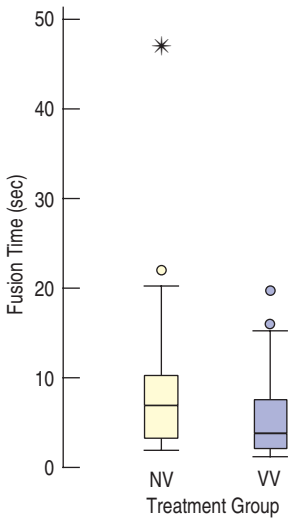
2745	703	302	242	119	40	7
1697	489	274	200	118	32	4
1656	430	274	198	115	31	
978	334	255	129	92	17	

a) Why is acre-feet a good way to measure the amount of precipitation produced by cloud seeding?
b) Plot these data, and describe the distribution.
c) Create a re-expression of these data that produces a more advantageous distribution.
d) Explain what your re-expressed scale means.

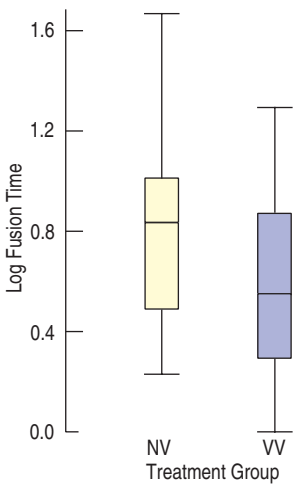
41. Stereograms. Stereograms appear to be composed entirely of random dots. However, they contain separate images that a viewer can "fuse" into a three-dimensional (3D) image by staring at the dots while defocusing the eyes. An experiment was performed to determine whether knowledge of the embedded image affected the

time required for subjects to fuse the images. One group of subjects (group NV) received no information or just verbal information about the shape of the embedded object. A second group (group VV) received both verbal information and visual information (specifically, a drawing of the object). The experimenters measured how many seconds it took for the subject to report that he or she saw the 3D image.

a) What two variables are discussed in this description?
b) For each variable, is it quantitative or categorical? If quantitative, what are the units?
c) The boxplots compare the fusion times for the two treatment groups. Write a few sentences comparing these distributions. What does the experiment show?



T **42. Stereograms, revisited.** Because of the skewness of the distributions of fusion times described in Exercise 41, we might consider a re-expression. Here are the boxplots of the *log* of fusion times. Is it better to analyze the original fusion times or the log fusion times? Explain.



JUST CHECKING Answers

1. The % late arrivals have a unimodal, symmetric distribution centered at about 20%. In most months between 16% and 23% of the flights arrived late.
2. The boxplot of % late arrivals makes it easier to see that the median is just below 20%, with quartiles at about 17% and 22%. It nominates two months as high outliers.
3. The boxplots by month show a strong seasonal pattern. Flights are more likely to be late in the winter and summer and less likely to be late in the spring and fall. One likely reason for the pattern is snowstorms in the winter and thunderstorms in the summer.