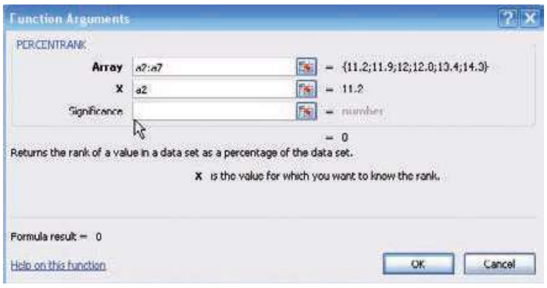
7. Type A2 for the X value, then click [OK].
8. Repeat the procedure above for each data value in column A.

The PERCENTRANK function returns the percentile rank as a decimal. To convert this to a percentage, multiply the function output by 100. Make sure to select a new column before multiplying the percentile rank by 100.



Array a2:a7 = {11.2;11.9;12;12.0;13.4;14.3}

X a2 = 11.2

Significance = number

= 0

Returns the rank of a value in a data set as a percentage of the data set.

X is the value for which you want to know the rank.

Formula result = 0

Descriptive Statistics in Excel

Example XL3–5

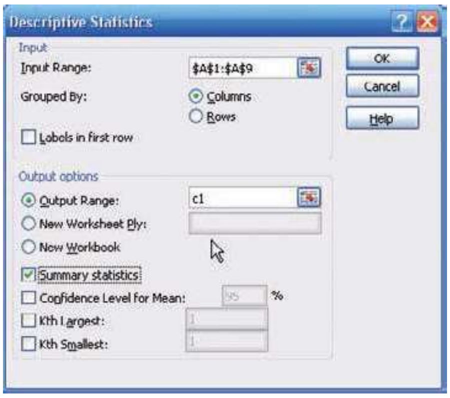
Excel Analysis Tool-Pak Add-in Data Analysis includes an item called Descriptive Statistics that reports many useful measures for a set of data.

1. Enter the data set shown in cells A1 to A9 of a new worksheet.

12 17 15 16 16 14 18 13 10

See the Excel Step by Step in Chapter 1 for the instructions on loading the Analysis Tool-Pak Add-in.

2. Select the Data tab on the toolbar and select Data Analysis.
3. In the Analysis Tools dialog box, scroll to Descriptive Statistics, then click [OK].
4. Type A1:A9 in the Input Range box and check the Grouped by Columns option.
5. Select the Output Range option and type in cell C1.
6. Check the Summary statistics option and click [OK].



Input Range: $A$1:$A$9

Grouped By: ● Columns ○ Rows

□ Labels in first row

● Output Range: c1

○ New Worksheet Ply:

○ New Workbook

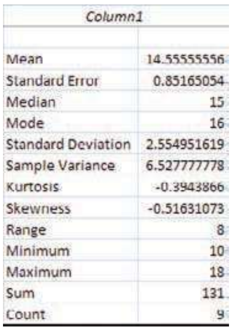
☑ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1

Below is the summary output for this data set.



Column1	
Mean	14.55555556
Standard Error	0.85165054
Median	15
Mode	16
Standard Deviation	2.554951619
Sample Variance	6.527777778
Kurtosis	-0.3943866
Skewness	-0.51631073
Range	8
Minimum	10
Maximum	18
Sum	131
Count	9

3–4 Exploratory Data Analysis

Objective 4

Use the techniques of exploratory data analysis, including boxplots and five-number summaries, to discover various aspects of data.

In traditional statistics, data are organized by using a frequency distribution. From this distribution various graphs such as the histogram, frequency polygon, and ogive can be constructed to determine the shape or nature of the distribution. In addition, various statistics such as the mean and standard deviation can be computed to summarize the data.

The purpose of traditional analysis is to confirm various conjectures about the nature of the data. For example, from a carefully designed study, a researcher might want to know if the proportion of Americans who are exercising today has increased from 10 years ago. This study would contain various assumptions about the population, various definitions such as of exercise, and so on.

In **exploratory data analysis (EDA),** data can be organized using a *stem and leaf plot.* (See Chapter 2.) The measure of central tendency used in EDA is the *median.* The measure of variation used in EDA is the *interquartile range* $Q_3 - Q_1$. In EDA the data are represented graphically using a *boxplot* (sometimes called a box-and-whisker plot). The purpose of exploratory data analysis is to examine data to find out what information can be discovered about the data such as the center and the spread. Exploratory data analysis was developed by John Tukey and presented in his book *Exploratory Data Analysis* (Addison-Wesley, 1977).

The Five-Number Summary and Boxplots

A **boxplot** can be used to graphically represent the data set. These plots involve five specific values:

1. The lowest value of the data set (i.e., minimum)
2. Q_1
3. The median
4. Q_3
5. The highest value of the data set (i.e., maximum)

These values are called a **five-number summary** of the data set.

A **boxplot** is a graph of a data set obtained by drawing a horizontal line from the minimum data value to Q_1, drawing a horizontal line from Q_3 to the maximum data value, and drawing a box whose vertical sides pass through Q_1 and Q_3 with a vertical line inside the box passing through the median or Q_2**.**

Procedure for constructing a boxplot

1. Find the five-number summary for the data values, that is, the maximum and minimum data values, Q_1 and Q_3, and the median.
2. Draw a horizontal axis with a scale such that it includes the maximum and minimum data values.
3. Draw a box whose vertical sides go through Q_1 and Q_3, and draw a vertical line though the median.
4. Draw a line from the minimum data value to the left side of the box and a line from the maximum data value to the right side of the box.

Example 3–38 Number of Meteorites Found

The number of meteorites found in 10 states of the United States is 89, 47, 164, 296, 30, 215, 138, 78, 48, 39. Construct a boxplot for the data.

Source: Natural History Museum.

Solution

Step 1 Arrange the data in order:

30, 39, 47, 48, 78, 89, 138, 164, 215, 296

Step 2 Find the median.

$$30, 39, 47, 48, \underbrace{78, 89}_{\uparrow\ \text{Median}}, 138, 164, 215, 296$$

$$\text{Median} = \frac{78 + 89}{2} = 83.5$$

Step 3 Find Q_1.

$$30, 39, \underset{\substack{\uparrow \\ Q_1}}{47}, 48, 78$$

Step 4 Find Q_3.

$$89, 138, \underset{\substack{\uparrow \\ Q_3}}{164}, 215, 296$$

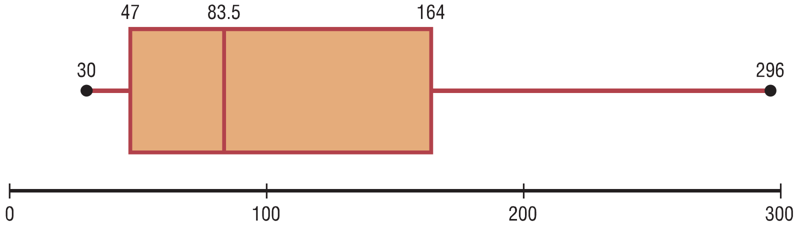
Step 5 Draw a scale for the data on the x axis.

Step 6 Locate the lowest value, Q_1, median, Q_3, and the highest value on the scale.

Step 7 Draw a box around Q_1 and Q_3, draw a vertical line through the median, and connect the upper value and the lower value to the box. See Figure 3–7.

Figure 3–7

Boxplot for Example 3–38



The distribution is somewhat positively skewed.

Information Obtained from a Boxplot

1. *a.* If the median is near the center of the box, the distribution is approximately symmetric.
   b. If the median falls to the left of the center of the box, the distribution is positively skewed.
   c. If the median falls to the right of the center, the distribution is negatively skewed.
2. *a.* If the lines are about the same length, the distribution is approximately symmetric.
   b. If the right line is larger than the left line, the distribution is positively skewed.
   c. If the left line is larger than the right line, the distribution is negatively skewed.

The boxplot in Figure 3–7 indicates that the distribution is slightly positively skewed.

If the boxplots for two or more data sets are graphed on the same axis, the distributions can be compared. To compare the averages, use the location of the medians. To compare the variability, use the interquartile range, i.e., the length of the boxes. Example 3–39 shows this procedure.

Example 3–39

Sodium Content of Cheese

A dietitian is interested in comparing the sodium content of real cheese with the sodium content of a cheese substitute. The data for two random samples are shown. Compare the distributions, using boxplots.

Real cheese				Cheese substitute			
310	420	45	40	270	180	250	290
220	240	180	90	130	260	340	310

Source: *The Complete Book of Food Counts.*

Solution

Step 1 Find Q_1, MD, and Q_3 for the real cheese data.

40 45 90 180 220 240 310 420

↑ ↑ ↑

Q_1 MD Q_3

$$Q_1 = \frac{45 + 90}{2} = 67.5 \qquad \text{MD} = \frac{180 + 220}{2} = 200$$

$$Q_3 = \frac{240 + 310}{2} = 275$$

Step 2 Find Q_1, MD, and Q_3 for the cheese substitute data.

130 180 250 260 270 290 310 340

↑ ↑ ↑

Q_1 MD Q_3

$$Q_1 = \frac{180 + 250}{2} = 215 \qquad \text{MD} = \frac{260 + 270}{2} = 265$$

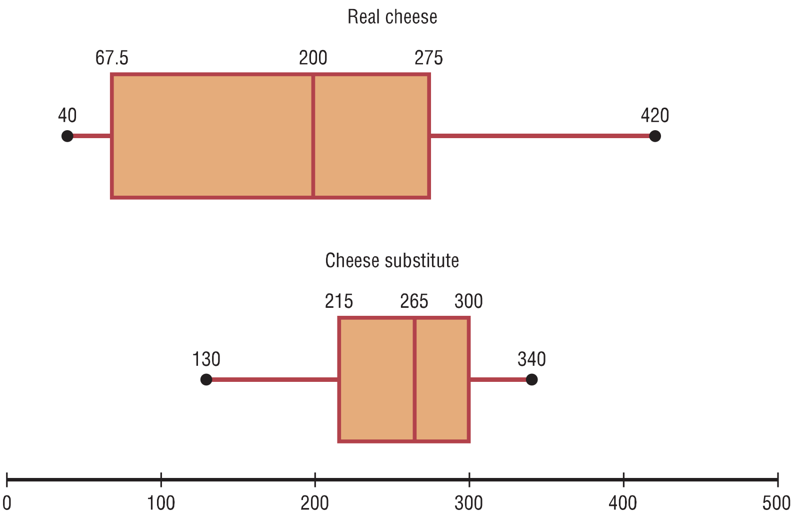
$$Q_3 = \frac{290 + 310}{2} = 300$$

Step 3 Draw the boxplots for each distribution on the same graph. See Figure 3–8.

Step 4 Compare the plots. It is quite apparent that the distribution for the cheese substitute data has a higher median than the median for the distribution for the real cheese data. The variation or spread for the distribution of the real cheese data is larger than the variation for the distribution of the cheese substitute data.

Figure 3–8

Boxplots for Example 3–39



A *modified boxplot* can be drawn and used to check for outliers. See Exercise 18 in Extending the Concepts in this section.

In exploratory data analysis, *hinges* are used instead of quartiles to construct boxplots. When the data set consists of an even number of values, hinges are the same as quartiles. Hinges for a data set with an odd number of values differ somewhat from quartiles. However, since most calculators and computer programs use quartiles, they will be used in this textbook.

Another important point to remember is that the summary statistics (median and interquartile range) used in exploratory data analysis are said to be *resistant statistics.* A **resistant statistic** is relatively less affected by outliers than a *nonresistant statistic.* The mean and standard deviation are nonresistant statistics. Sometimes when a distribution is skewed or contains outliers, the median and interquartile range may more accurately summarize the data than the mean and standard deviation, since the mean and standard deviation are more affected in this case.

Table 3–5 shows the correspondence between the traditional and the exploratory data analysis approach.

Table 3–5 Traditional versus EDA Techniques

Traditional	Exploratory data analysis
Frequency distribution	Stem and leaf plot
Histogram	Boxplot
Mean	Median
Standard deviation	Interquartile range

Applying the Concepts 3–4

The Noisy Workplace

Assume you work for OSHA (Occupational Safety and Health Administration) and have complaints about noise levels from some of the workers at a state power plant. You charge the power plant with taking decibel readings at six different areas of the plant at different times of the day and week. The results of the data collection are listed. Use boxplots to initially explore the data and make recommendations about which plant areas workers must be provided with protective ear wear. The safe hearing level is approximately 120 decibels.

Area 1	Area 2	Area 3	Area 4	Area 5	Area 6
30	64	100	25	59	67
12	99	59	15	63	80
35	87	78	30	81	99
65	59	97	20	110	49
24	23	84	61	65	67
59	16	64	56	112	56
68	94	53	34	132	80
57	78	59	22	145	125
100	57	89	24	163	100
61	32	88	21	120	93
32	52	94	32	84	56
45	78	66	52	99	45
92	59	57	14	105	80
56	55	62	10	68	34
44	55	64	33	75	21

See page 180 for the answers.

Exercises 3–4

For Exercises 1–6, identify the five-number summary and find the interquartile range.

1. 8, 12, 32, 6, 27, 19, 54 6, 8, 19, 32, 54; 24

2. 19, 16, 48, 22, 7 7, 11.5, 19, 35, 48; 23.5

3. 362, 589, 437, 316, 192, 188
188, 192, 339, 437, 589; 245

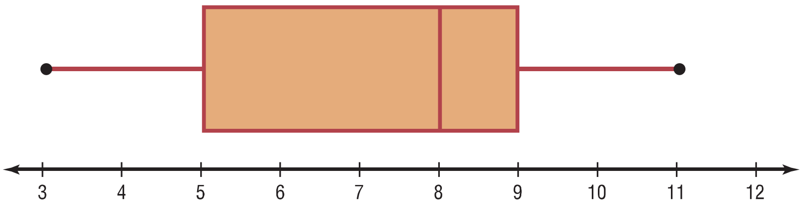
4. 147, 243, 156, 632, 543, 303
147, 156, 273, 543, 632; 387

5. 14.6, 19.8, 16.3, 15.5, 18.2
14.6, 15.05, 16.3, 19, 19.8; 3.95

6. 9.7, 4.6, 2.2, 3.7, 6.2, 9.4, 3.8 2.2, 3.7, 4.6, 9.4, 9.7; 5.7

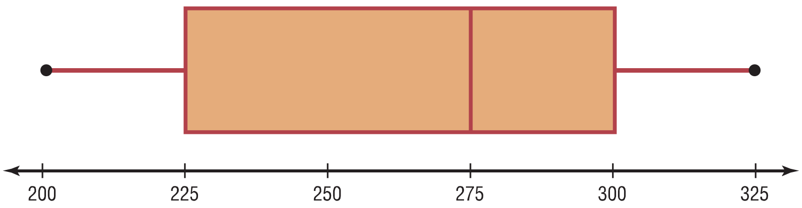
For Exercises 7–10, use each boxplot to identify the maximum value, minimum value, median, first quartile, third quartile, and interquartile range.

7.



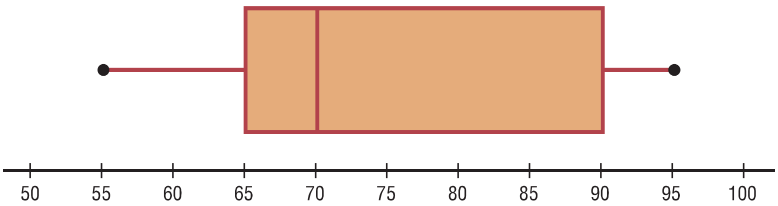
11, 3, 8, 5, 9, 4

8.



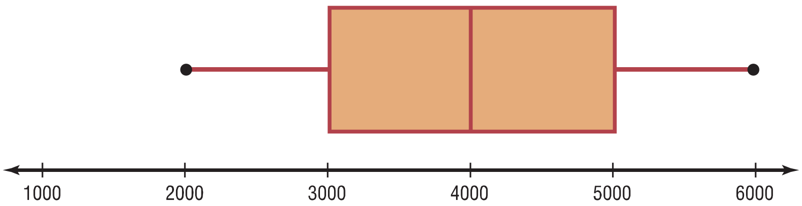
325, 200, 275, 225, 300, 75

9.



95, 55, 70, 65, 90, 25

10.



6000, 2000, 4000, 3000, 5000; 2000

11. **Earned Run Average—Number of Games Pitched** Construct a boxplot for the following data and comment on the shape of the distribution representing the number of games pitched by major league baseball's earned run average (ERA) leaders for the past few years.

30 34 29 30 34 29 31 33 34 27
30 27 34 32

Source: *World Almanac.*

12. **Innings Pitched** Construct a boxplot for the following data which represent the number of innings pitched by the ERA leaders for the past few years. Comment on the shape of the distribution.

192 228 186 199 238 217 213 234 264 187
214 115 238 246

Source: *World Almanac.*

13. **Teacher Strikes** The number of teacher strikes over a 13-year period in Pennsylvania is shown. Construct a boxplot for the data. Is the distribution symmetric?

20	18	7	13
7	14	5	9
9	9	10	17
15			

Source: Pennsylvania School Boards Association.

14. **Visitors Who Travel to Foreign Countries** Construct a boxplot for the number (in millions) of visitors who traveled to a foreign country each year for a random selection of years. Comment on the skewness of the distribution.

4.3	0.5	0.6	0.8	0.5
0.4	3.8	1.3	0.4	0.3

15. **Tornadoes in 2005** Construct a boxplot and comment on its skewness for the number of tornadoes recorded each month in 2005.

33 10 62 132 123 316 138 123 133 18
150 26

Source: Storm Prediction Center.

16. **Size of Dams** These data represent the volumes in cubic yards of the largest dams in the United States and in South America. Construct a boxplot of the data for each region and compare the distributions.

United States	South America
125,628	311,539
92,000	274,026
78,008	105,944
77,700	102,014
66,500	56,242
62,850	46,563
52,435	
50,000	

Source: *New York Times Almanac.*

17. Number of Tornadoes A four-month record for the number of tornadoes in 2003–2005 is given here.

	2005	2004	2003
April	132	125	157
May	123	509	543
June	316	268	292
July	138	124	167

a. Which month had the highest mean number of tornadoes for this 3-year period? May: 391.7

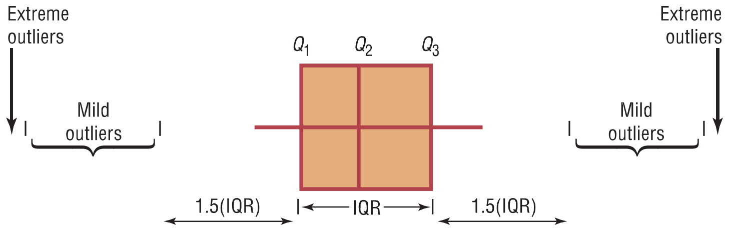
b. Which year has the highest mean number of tornadoes for this 4-month period? 2003: 289.8

c. Construct three boxplots and compare the distributions.

Source: NWS, Storm Prediction Center.

Extending the Concepts

18. Unhealthful Smog Days A *modified boxplot* can be drawn by placing a box around Q_1 and Q_3 and then extending the whiskers to the largest and/or smallest values within 1.5 times the interquartile range (that is, $Q_3 - Q_1$). *Mild outliers* are values between 1.5(IQR) and 3(IQR). *Extreme outliers* are data values beyond 3(IQR).



For the data shown here, draw a modified boxplot and identify any mild or extreme outliers. The data represent the number of unhealthful smog days for a specific year for the highest 10 locations.

97	39	43	66	91
43	54	42	53	39

Source: U.S. Public Interest Research Group and Clean Air Network.

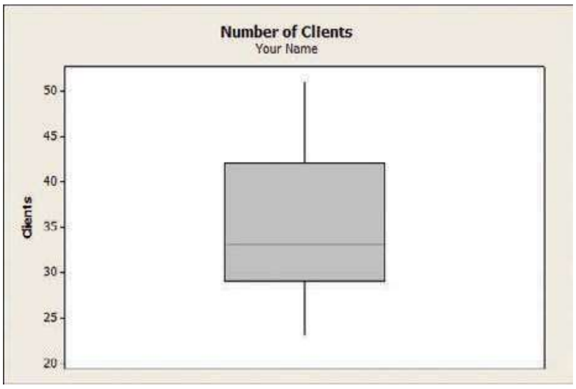
Technology *Step by Step*

MINITAB Step by Step

Construct a Boxplot

1. Type in the data 33, 38, 43, 30, 29, 40, 51, 27, 42, 23, 31. Label the column **Clients.**
2. Select **Stat>EDA>Boxplot.**
3. Double-click Clients to select it for the Y variable.
4. Click on [Labels].
   a) In the Title 1: of the Title/Footnotes folder, type **Number of Clients.**
   b) Press the [Tab] key and type **Your Name** in the text box for Subtitle 1:.

5. Click [OK] twice. The graph will be displayed in a graph window.



Example MT3–2

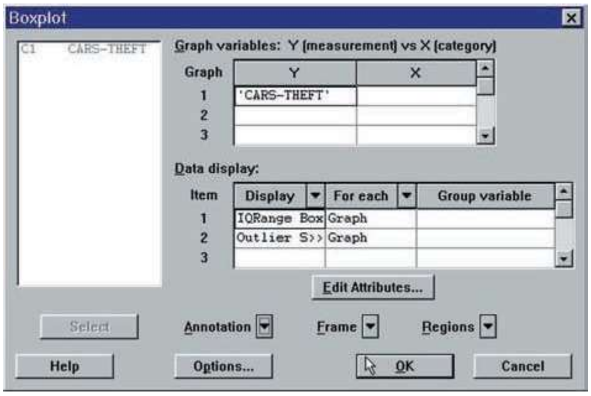
The number of car thefts in a large city over a 30-day period is shown.

52	62	51	50	69
58	77	66	53	57
75	56	65	67	73
79	59	68	65	72
57	51	63	69	75
65	53	78	66	55

1. Enter the data for this example. Label the column **CARS-THEFT.**
2. Select **Stat>EDA>Boxplot.**
3. Double-click CARS-THEFT to select it for the Y variable.
4. Click on the drop-down arrow for Annotation.
5. Click on Title, then enter an appropriate title such as **Car Thefts for Large City, U.S.A.**
6. Click [OK] twice.

A high-resolution graph will be displayed in a graph window.

Boxplot Dialog Box and Boxplot


Boxplot

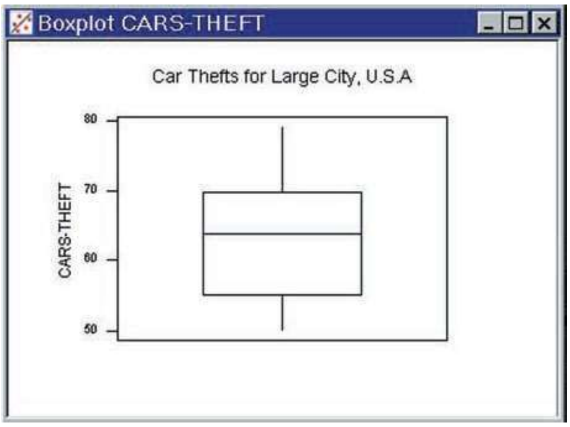
C1 CARS-THEFT

Graph variables: Y (measurement) vs X (category)

Graph	Y	X
1	'CARS-THEFT'	
2		
3		

Data display:

Item	Display	For each	Group variable
1	IQRange Box	Graph	
2	Outlier S>>	Graph	
3			



TI-83 Plus or TI-84 Plus Step by Step

Constructing a Boxplot

To draw a boxplot:

1. Enter data into L_1.
2. Change values in WINDOW menu, if necessary. (*Note:* Make $X_{\min}$ somewhat smaller than the smallest data value and $X_{\max}$ somewhat larger than the largest data value.) Change $Y_{\min}$ to 0 and $Y_{\max}$ to 1.
3. Press **[2nd] [STAT PLOT]**, then **1** for Plot 1.
4. Press **ENTER** to turn Plot 1 on.
5. Move cursor to Boxplot symbol (fifth graph) on the Type: line, then press **ENTER.**
6. Make sure Xlist is L_1.
7. Make sure Freq is 1.
8. Press **GRAPH** to display the boxplot.
9. Press **TRACE** followed by ◀ or ▶ to obtain the values from the five-number summary on the boxplot.

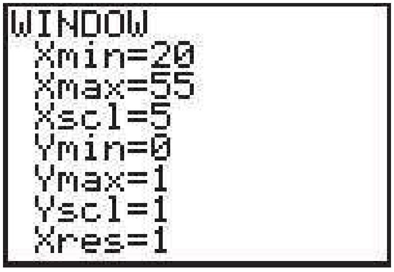
To display two boxplots on the same display, follow the above steps and use the 2: Plot 2 and L_2 symbols.

Example TI3–3

Construct a boxplot for the data values:

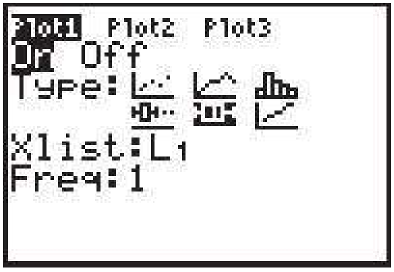
33, 38, 43, 30, 29, 40, 51, 27, 42, 23, 31

Input



WINDOW
Xmin=20
Xmax=55
Xscl=5
Ymin=0
Ymax=1
Yscl=1
Xres=1

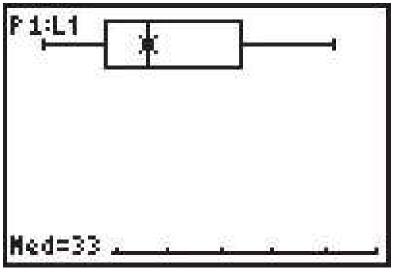
Input



Plot1 Plot2 Plot3
On Off
Type:
Xlist:L_1
Freq:1

Using the **TRACE** key along with the ◀ and ▶ keys, we obtain the five-number summary. The minimum value is 23; Q_1 is 29; the median is 33; Q_3 is 42; the maximum value is 51.

Output



P1:L1

Med=33

Excel Step by Step

Constructing a Stem and Leaf Plot and a Boxplot

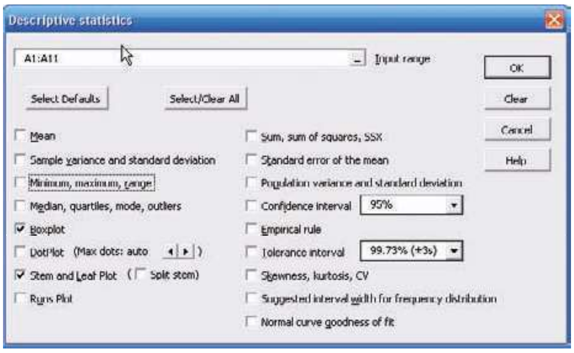
Example XL3–6

Excel does not have procedures to produce stem and leaf plots or boxplots. However, you may construct these plots by using the MegaStat Add-in available on your CD or from the Online

Learning Center. If you have not installed this add-in, refer to the instructions in the Excel Step by Step section of Chapter 1.

To obtain a boxplot and stem and leaf plot:

1. Enter the data values 33, 38, 43, 30, 29, 40, 51, 27, 42, 23, 31 into column A of a new Excel worksheet.
2. Select the Add-Ins tab, then MegaStat from the toolbar.
3. Select Descriptive Statistics from the MegaStat menu.
4. Enter the cell range A1:A11 in the Input range.
5. Check both Boxplot and Stem and Leaf Plot. *Note:* You may leave the other output options unchecked for this example. Click [OK].

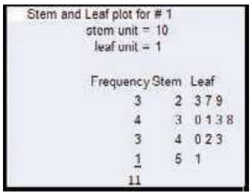


Descriptive statistics

A1:A11 Input range

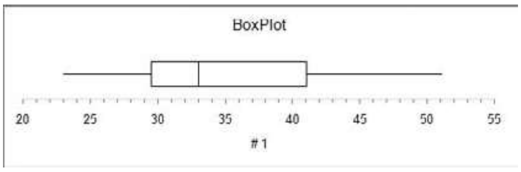
□ Mean
□ Sample variance and standard deviation
□ Minimum, maximum, range
□ Median, quartiles, mode, outliers
☑ Boxplot
□ DotPlot (Max dots: auto)
☑ Stem and Leaf Plot (□ Split stem)
□ Runs Plot
□ Sum, sum of squares, SSX
□ Standard error of the mean
□ Population variance and standard deviation
□ Confidence interval 95%
□ Empirical rule
□ Tolerance interval 99.73% (+3s)
□ Skewness, kurtosis, CV
□ Suggested interval width for frequency distribution
□ Normal curve goodness of fit

The stem and leaf plot and the boxplot are shown below.



```
Stem and Leaf plot for # 1
        stem unit = 10
         leaf unit = 1

Frequency Stem  Leaf
     3       2  3 7 9
     4       3  0 1 3 8
     3       4  0 2 3
     1       5  1
    11
```



Summary

- This chapter explains the basic ways to summarize data. These include measures of central tendency. They are the mean, median, mode, and midrange. The weighted mean can also be used. (3–1)
- To summarize the variation of data, statisticians use measures of variation or dispersion. The three most common measures of variation are the range, variance, and standard deviation. The coefficient of variation can be used to compare the variation of two data sets. The data values are distributed according to Chebyshev's theorem on the empirical rule. (3–2)
- There are several measures of the position of data values in the set. There are standard scores, percentiles, quartiles, and deciles. Sometimes a data set contains an extremely high or extremely low data value, called an outlier. (3–3)
- Other methods can be used to describe a data set. These methods are the five-number summary and boxplots. These methods are called exploratory data analysis. (3–4)

The techniques explained in Chapter 2 and this chapter are the basic techniques used in descriptive statistics.

Important Terms

bimodal 111

boxplot 162

Chebyshev's theorem 134

coefficient of variation 132

data array 109

decile 151

empirical rule 136

exploratory data analysis (EDA) 162

five-number summary 162

interquartile range (IQR) 151

mean 106

median 109

midrange 114

modal class 112

mode 111

multimodal 111

negatively skewed or left-skewed distribution 117

outlier 151

parameter 106

percentile 143

positively skewed or right-skewed distribution 117

quartile 149

range 124

range rule of thumb 133

resistant statistic 165

standard deviation 127

statistic 106

symmetric distribution 117

unimodal 111

variance 127

weighted mean 115

z score or standard score 142

Important Formulas

Formula for the mean for individual data:

$$\bar{X} = \frac{\Sigma X}{n} \qquad \mu = \frac{\Sigma X}{N}$$

Formula for the mean for grouped data:

$$\bar{X} = \frac{\Sigma f \cdot X_m}{n}$$

Formula for the weighted mean:

$$\bar{X} = \frac{\Sigma wX}{\Sigma w}$$

Formula for the midrange:

$$\text{MR} = \frac{\text{lowest value} + \text{highest value}}{2}$$

Formula for the range:

$$R = \text{highest value} - \text{lowest value}$$

Formula for the variance for population data:

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

Formula for the variance for sample data (shortcut formula for the unbiased estimator):

$$s^2 = \frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n - 1)}$$

Formula for the variance for grouped data:

$$s^2 = \frac{n(\Sigma f \cdot X_m^2) - (\Sigma f \cdot X_m)^2}{n(n - 1)}$$

Formula for the standard deviation for population data:

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

Formula for the standard deviation for sample data (shortcut formula):

$$s = \sqrt{\frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n - 1)}}$$

Formula for the standard deviation for grouped data:

$$s = \sqrt{\frac{n(\Sigma f \cdot X_m^2) - (\Sigma f \cdot X_m)^2}{n(n - 1)}}$$

Formula for the coefficient of variation:

$$\text{CVar} = \frac{s}{\bar{X}} \cdot 100 \qquad \text{or} \qquad \text{CVar} = \frac{\sigma}{\mu} \cdot 100$$

Range rule of thumb:

$$s \approx \frac{\text{range}}{4}$$

Expression for Chebyshev's theorem: The proportion of values from a data set that will fall within *k* standard deviations of the mean will be at least

$$1 - \frac{1}{k^2}$$

where *k* is a number greater than 1.

Formula for the *z* score (standard score):

$$z = \frac{X - \mu}{\sigma} \qquad \text{or} \qquad z = \frac{X - \bar{X}}{s}$$

Formula for the cumulative percentage:

$$\textbf{Cumulative \%} = \frac{\textbf{cumulative frequency}}{n} \cdot 100$$

Formula for the percentile rank of a value *X*:

$$\textbf{Percentile} = \frac{\textbf{number of values below } X + 0.5}{\textbf{total number of values}} \cdot 100$$

Formula for finding a value corresponding to a given percentile:

$$c = \frac{n \cdot p}{100}$$

Formula for interquartile range:

$$\text{IQR} = Q_3 - Q_1$$

Review Exercises

1. **Net Worth of Wealthy People** The net worth (in billions of dollars) of a sample of the richest people in the United States is shown. Find the mean, median, mode, midrange, variance, and standard deviation for the data. (3–1) (3–2)

59	52	28	26	19
19	18	17	17	17

Source: *Forbes Magazine.*

2. **Shark Attacks** The number of shark attacks and deaths over a recent 5-year period is shown. Find the mean, median, mode, midrange, variance, and standard deviation for the data. Which data set is more variable? (3–1) (3–2)

Attacks	71	64	61	65	57
Deaths	1	4	4	7	4

3. **Battery Lives** Twelve batteries were tested to see how many hours they would last. The frequency distribution is shown here.

Hours	Frequency
1–3	1
4–6	4
7–9	5
10–12	1
13–15	1

Find each of these. (3–1) (3–2)

a. Mean 7.3
b. Modal class 7–9
c. Variance 10.0
d. Standard deviation 3.2

4. **SAT Scores** The mean SAT math scores for selected states are represented below. Find the mean class, modal class, variance, and standard deviation, and comment on the shape of the data. (3–1) (3–2)

Score	Frequency
478–504	4
505–531	6
532–558	2
559–585	2
586–612	2

Source: *World Almanac.*

5. **Rise in Tides** Shown here is a frequency distribution for the rise in tides at 30 selected locations in the United States.

Rise in tides (inches)	Frequency
12.5–27.5	6
27.5–42.5	3
42.5–57.5	5
57.5–72.5	8
72.5–87.5	6
87.5–102.5	2

Find each of these. (3–1) (3–2)

a. Mean 55.5
b. Modal class 57.5–72.5
c. Variance 566.1
d. Standard deviation 23.8

6. **Fuel Capacity** The fuel capacity in gallons of 50 randomly selected cars is shown here.

Class	Frequency
10–12	6
13–15	4
16–18	14
19–21	15
22–24	8
25–27	2
28–30	1
	50

Find each of these. (3–1) (3–2)

a. Mean 18.5
b. Modal class 19–21
c. Variance 17.7
d. Standard deviation 4.2

7. **Households with Four Television Networks** A survey showed the number of viewers and number of households of four television networks. Find the average number of viewers, using the weighted mean. (3–1) 1.43 viewers

Households	1.4	0.8	0.3	1.6
Viewers (in millions)	1.6	0.8	0.4	1.8

Source: Nielsen Media Research.

8. **Investment Earnings** An investor calculated these percentages of each of three stock investments with payoffs as shown. Find the average payoff. Use the weighted mean. (3–1) $4700.00

Stock	**Percent**	**Payoff**
A	30	$10,000
B	50	3,000
C	20	1,000

9. **Years of Service of Employees** In an advertisement, a transmission service center stated that the average years of service of its employees were 13. The distribution is shown here. Using the weighted mean, calculate the correct average. (3–1) 6

Number of employees	**Years of service**
8	3
1	6
1	30

10. **Textbooks in Professors' Offices** If the average number of textbooks in professors' offices is 16, the standard deviation is 5, and the average age of the professors is 43, with a standard deviation of 8, which data set is more variable? (3–2) 31.25%; 18.6%; the number of books is more variable

11. **Magazines in Bookstores** A survey of bookstores showed that the average number of magazines carried is 56, with a standard deviation of 12. The same survey showed that the average length of time each store had been in business was 6 years, with a standard deviation of 2.5 years. Which is more variable, the number of magazines or the number of years? (3–2) Magazine variance: 0.214; year variance: 0.417; years are more variable

12. **Years of Service of Supreme Court Members** The number of years served by selected past members of the U.S. Supreme Court is listed below. Find the percentile rank for each value. Which value corresponds to the 40th percentile? Construct a boxplot for the data and comment on their shape. (3–3) (3–4)

19, 15, 16, 24, 17, 4, 3, 31, 23, 5, 33

Source: *World Almanac.*

13. **NFL Salaries** The salaries (in millions of dollars) for 29 NFL teams for the 1999–2000 season are given in this frequency distribution. (3–3)

Class limits	**Frequency**
39.9–42.8	2
42.9–45.8	2
45.9–48.8	5
48.9–51.8	5
51.9–54.8	12
54.9–57.8	3

Source: www.NFL.com

a. Construct a percentile graph.
b. Find the values that correspond to the 35th, 65th, and 85th percentiles. 50, 53, 55
c. Find the percentile of values 44, 48, and 54. 10th; 26th; 78th

14. Check each data set for outliers. (3–3)
a. 506, 511, 517, 514, 400, 521 400
b. 3, 7, 9, 6, 8, 10, 14, 16, 20, 12 None
c. 14, 18, 27, 26, 19, 13, 5, 25 None
d. 112, 157, 192, 116, 153, 129, 131 None

15. **Cost of Car Rentals** A survey of car rental agencies shows that the average cost of a car rental is $0.32 per mile. The standard deviation is $0.03. Using Chebyshev's theorem, find the range in which at least 75% of the data values will fall. (3–2) $0.26–$0.38

16. **Average Earnings of Workers** The average earnings of year-round full-time workers 25–34 years old with a bachelor's degree or higher were $58,500 in 2003. If the standard deviation is $11,200, what can you say about the percentage of these workers who earn (3–2)
a. Between $47,300 and $69,700? Nothing because $k = 1$
b. More than $80,900? At most ¼ or 25%
c. How likely is it that someone earns more than $100,000? At most 7.3%

Source: *New York Times Almanac.*

17. **Labor Charges** The average labor charge for automobile mechanics is $54 per hour. The standard deviation is $4. Find the minimum percentage of data values that will fall within the range of $48 to $60. Use Chebyshev's theorem. (3–2) 56%

18. **Costs to Train Employees** For a certain type of job, it costs a company an average of $231 to train an employee to perform the task. The standard deviation is $5. Find the minimum percentage of data values that will fall in the range of $219 to $243. Use Chebyshev's theorem. (3–2) 83%

19. **Delivery Charges** The average delivery charge for a refrigerator is $32. The standard deviation is $4. Find the minimum percentage of data values that will fall in the range of $20 to $44. Use Chebyshev's theorem. (3–2) 88.89%

20. Exam Grades Which of these exam grades has a better relative position? (3–3)

a. A grade of 82 on a test with $\overline{X} = 85$ and $s = 6$ -0.5

b. A grade of 56 on a test with $\overline{X} = 60$ and $s = 5$ -0.8

The test in part *a* is better.

21. Top Movie Sites The number of sites at which the top nine movies (based on the daily gross earnings) opened in a particular week is indicated below.

3017	3687	2525
2516	2820	2579
3211	3044	2330

Construct a boxplot for the data.

The 10th movie on the list opened at only 909 theaters. Add this number to the above set of data and comment on the changes that occur. (3–4)

Source: www.showbizdata.com The range is much larger.

22. Hours Worked The data shown here represent the number of hours that 12 part-time employees at a toy store worked during the weeks before and after Christmas. Construct two boxplots and compare the distributions. (3–4)

Before	38	16	18	24	12	30	35	32	31	30	24	35
After	26	15	12	18	24	32	14	18	16	18	22	12

23. Commuter Times The mean of the times it takes a commuter to get to work in Baltimore is 29.7 minutes. If the standard deviation is 6 minutes, within what limits would you expect approximately 68% of the times to fall? Assume the distribution is approximately bell-shaped. (3–3) 23.7–35.7

Statistics Today

How Long Are You Delayed by Road Congestion?—Revisited

The average number of hours per year that a driver is delayed by road congestion is listed here.

Los Angeles	56
Atlanta	53
Seattle	53
Houston	50
Dallas	46
Washington	46
Austin	45
Denver	45
St. Louis	44
Orlando	42
U.S. average	36

Source: Texas Transportation Institute.

By making comparisons using averages, you can see that drivers in these 10 cities are delayed by road congestion more than the national average.

Data Analysis

A Data Bank is found in Appendix D, or on the World Wide Web by following links from www.mhhe.com/math/stat/bluman/

1. From the Data Bank, choose one of the following variables: age, weight, cholesterol level, systolic pressure, IQ, or sodium level. Select at least 30 values, and find the mean, median, mode, and midrange. State which measurement of central tendency best describes the average and why.
2. Find the range, variance, and standard deviation for the data selected in Exercise 1.
3. From the Data Bank, choose 10 values from any variable, construct a boxplot, and interpret the results.
4. Randomly select 10 values from the number of suspensions in the local school districts in southwestern Pennsylvania in Data Set V in Appendix D. Find the mean, median, mode, range, variance, and standard deviation of the number of suspensions by using the Pearson coefficient of skewness.
5. Using the data from Data Set VII in Appendix D, find the mean, median, mode, range, variance, and standard deviation of the acreage owned by the municipalities. Comment on the skewness of the data, using the Pearson coefficient of skewness.

Chapter Quiz

Determine whether each statement is true or false. If the statement is false, explain why.

1. When the mean is computed for individual data, all values in the data set are used. True
2. The mean cannot be found for grouped data when there is an open class. True
3. A single, extremely large value can affect the median more than the mean. False
4. One-half of all the data values will fall above the mode, and one-half will fall below the mode. False
5. In a data set, the mode will always be unique. False
6. The range and midrange are both measures of variation. False
7. One disadvantage of the median is that it is not unique. False
8. The mode and midrange are both measures of variation. False
9. If a person's score on an exam corresponds to the 75th percentile, then that person obtained 75 correct answers out of 100 questions. False

Select the best answer.

10. What is the value of the mode when all values in the data set are different?

a. 0
b. 1
[✓ *c.*] There is no mode.
d. It cannot be determined unless the data values are given.

11. When data are categorized as, for example, places of residence (rural, suburban, urban), the most appropriate measure of central tendency is the

a. Mean
b. Median
[✓ *c.*] Mode
d. Midrange

12. P_{50} corresponds to *a* and *b*

a. Q_2
b. D_5
c. IQR
d. Midrange

13. Which is not part of the five-number summary?

a. Q_1 and Q_3
[✓ *b.*] The mean
c. The median
d. The smallest and the largest data values

14. A statistic that tells the number of standard deviations a data value is above or below the mean is called

a. A quartile
b. A percentile
c. A coefficient of variation
[✓ *d.*] A *z* score

15. When a distribution is bell-shaped, approximately what percentage of data values will fall within 1 standard deviation of the mean?

a. 50%
[✓ *b.*] 68%
c. 95%
d. 99.7%

Complete these statements with the best answer.

16. A measure obtained from sample data is called a(n) ________. Statistic

17. Generally, Greek letters are used to represent ________, and Roman letters are used to represent ________. Parameters, statistics

18. The positive square root of the variance is called the ________. Standard deviation

19. The symbol for the population standard deviation is ________. σ

20. When the sum of the lowest data value and the highest data value is divided by 2, the measure is called ________. Midrange

21. If the mode is to the left of the median and the mean is to the right of the median, then the distribution is ________ skewed. Positively

22. An extremely high or extremely low data value is called a(n) ________. Outlier

23. Miles per Gallon The number of highway miles per gallon of the 10 worst vehicles is shown.

12 15 13 14 15 16 17 16 17 18

Source: *Pittsburgh Post Gazette.*

Find each of these.

a. Mean 15.3
b. Median 15.5
c. Mode 15, 16, and 17
d. Midrange 15
e. Range 6
f. Variance 3.57
g. Standard deviation 1.9

24. Errors on a Typing Test The distribution of the number of errors that 10 students made on a typing test is shown.

Errors	Frequency
0–2	1
3–5	3
6–8	4
9–11	1
12–14	1

Find each of these.

a. Mean 6.4
b. Modal class 6–8
c. Variance 11.6
d. Standard deviation 3.4

25. Inches of Rain Shown here is a frequency distribution for the number of inches of rain received in 1 year in 25 selected cities in the United States.

Number of inches	Frequency
5.5–20.5	2
20.5–35.5	3
35.5–50.5	8
50.5–65.5	6
65.5–80.5	3
80.5–95.5	3

Find each of these.

a. Mean 51.4
b. Modal class 35.5–50.5
c. Variance 451.5
d. Standard deviation 21.2

26. Shipment Times A survey of 36 selected recording companies showed these numbers of days that it took to receive a shipment from the day it was ordered.

Days	Frequency
1–3	6
4–6	8
7–9	10
10–12	7
13–15	0
16–18	5

Find each of these.

a. Mean 8.2
b. Modal class 7–9
c. Variance 21.6
d. Standard deviation 4.6

27. Best Friends of Students In a survey of third-grade students, this distribution was obtained for the number of "best friends" each had. 1.6

Number of students	Number of best friends
8	1
6	2
5	3
3	0

Find the average number of best friends for the class. Use the weighted mean.

28. Employee Years of Service In an advertisement, a retail store stated that its employees averaged 9 years of service. The distribution is shown here. 4.5

Number of employees	Years of service
8	2
2	6
3	10

Using the weighted mean, calculate the correct average.

29. Newspapers for Sale The average number of newspapers for sale in an airport newsstand is 12, and the standard deviation is 4. The average age of the pilots is 37 years, with a standard deviation of 6 years. Which data set is more variable? 0.33; 0.162; newspapers

30. Brands of Toothpaste Carried A survey of grocery stores showed that the average number of brands of toothpaste carried was 16, with a standard deviation of 5. The same survey showed the average length of time each store was in business was 7 years, with a standard deviation of 1.6 years. Which is more variable, the number of brands or the number of years? 0.3125; 0.229; brands

31. Test Scores A student scored 76 on a general science test where the class mean and standard deviation were 82 and 8, respectively; he also scored 53 on a psychology test where the class mean and standard deviation were 58 and 3, respectively. In which class was his relative position higher? −0.75; −1.67; science

32. Which score has the highest relative position?

a. $X = 12$ $\overline{X} = 10$ $s = 4$ 0.5
b. $X = 170$ $\overline{X} = 120$ $s = 32$ 1.6
c. $X = 180$ $\overline{X} = 60$ $s = 8$ 15, *c* is highest

33. Sizes of Malls The number of square feet (in millions) of eight of the largest malls in southwestern Pennsylvania is shown.

1	0.9	1.3	0.8
1.4	0.77	0.7	1.2

Source: International Council of Shopping Centers.

a. Find the percentile for each value.
b. What value corresponds to the 40th percentile?
c. Construct a boxplot and comment on the nature of the distribution.

34. Exam Scores On a philosophy comprehensive exam, this distribution was obtained from 25 students.

Score	Frequency
40.5–45.5	3
45.5–50.5	8
50.5–55.5	10
55.5–60.5	3
60.5–65.5	1

a. Construct a percentile graph.
b. Find the values that correspond to the 22nd, 78th, and 99th percentiles. 47; 55; 64
c. Find the percentiles of the values 52, 43, and 64. 56th, 6th, 99th percentiles

35. Gas Prices for Rental Cars The first column of these data represents the prebuy gas price of a rental car, and the second column represents the price charged if the car is returned without refilling the gas tank for a selected car rental company. Draw two boxplots for the data and compare the distributions. (*Note:* The data were collected several years ago.)

Prebuy cost	No prebuy cost
$1.55	$3.80
1.54	3.99
1.62	3.99
1.65	3.85
1.72	3.99
1.63	3.95
1.65	3.94
1.72	4.19
1.45	3.84
1.52	3.94

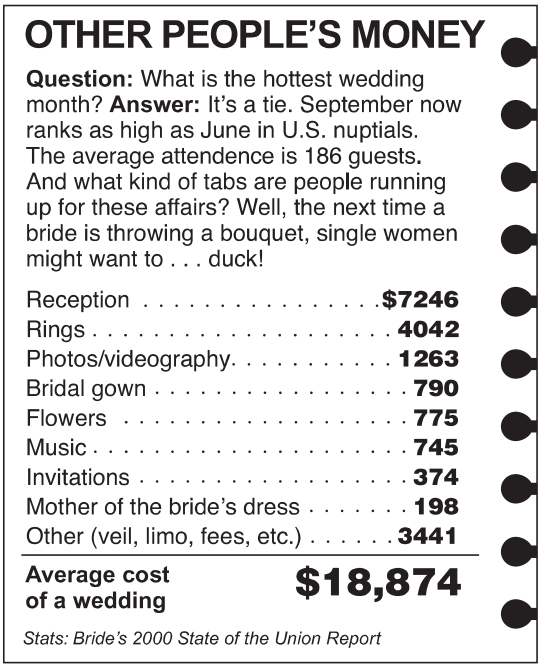
Source: *USA TODAY.*

36. **SAT Scores** The average national SAT score is 1019. If we assume a bell-shaped distribution and a standard deviation equal to 110, what percentage of scores will you expect to fall above 1129? Above 799? 16%, 97.5%

Source: *New York Times Almanac,* 2002.

Critical Thinking Challenges

1. **Average Cost of Weddings** Averages give us information to help us to see where we stand and enable us to make comparisons. Here is a study on the average cost of a wedding. What type of average—mean, median, mode, or midrange—might have been used for each category?



OTHER PEOPLE'S MONEY

Question: What is the hottest wedding month? **Answer:** It's a tie. September now ranks as high as June in U.S. nuptials. The average attendence is 186 guests. And what kind of tabs are people running up for these affairs? Well, the next time a bride is throwing a bouquet, single women might want to . . . duck!

Reception	**$7246**
Rings	**4042**
Photos/videography	**1263**
Bridal gown	**790**
Flowers	**775**
Music	**745**
Invitations	**374**
Mother of the bride's dress	**198**
Other (veil, limo, fees, etc.)	**3441**
Average cost of a wedding	**$18,874**

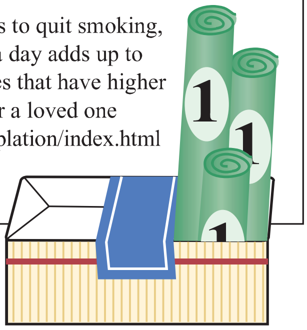
Stats: Bride's 2000 State of the Union Report

Source: Reprinted with permission from the September 2001 Reader's Digest. Copyright © 2001 by The Reader's Digest Assn., Inc.

2. **Average Cost of Smoking** This article states that the average yearly cost of smoking a pack of cigarettes a day is $1190. Find the average cost of a pack of cigarettes in your area, and compute the cost per day for 1 year. Compare your answer with the one in the article.

Burning Through the Cash

Everyone knows the health-related reasons to quit smoking, so herés an economic argument: A pack a day adds up to $1190 a year on average; it’s more in states that have higher taxes on tobacco. To calculate what you or a loved one spends, visit ashline.org/ASH/quit/contemplation/index.html and try out their smoker’s calculator. You’ll be stunned.



Source: Reprinted with permission from the April 2002 Reader’s Digest. Copyright © 2002 by The Reader’s Digest Assn., Inc.

3. **Ages of U.S. Residents** The table shows the median ages of residents for the 10 oldest states and the 10 youngest states of the United States including Washington, D.C. Explain why the median is used instead of the mean.

10 Oldest			10 Youngest		
Rank	**State**	**Median age**	**Rank**	**State**	**Median age**
1	West Virginia	38.9	51	Utah	27.1
2	Florida	38.7	50	Texas	32.3
3	Maine	38.6	49	Alaska	32.4
4	Pennsylvania	38.0	48	Idaho	33.2
5	Vermont	37.7	47	California	33.3
6	Montana	37.5	46	Georgia	33.4
7	Connecticut	37.4	45	Mississippi	33.8
8	New Hampshire	37.1	44	Louisiana	34.0
9	New Jersey	36.7	43	Arizona	34.2
10	Rhode Island	36.7	42	Colorado	34.3

Source: U.S. Census Bureau.

Data Projects

Where appropriate, use MINITAB, the TI-83 Plus, the TI-84 Plus, or a computer program of your choice to complete the following exercises.

1. **Business and Finance** Use the data collected in data project 1 of Chapter 2 regarding earnings per share. Determine the mean, mode, median, and midrange for the two data sets. Is one measure of center more appropriate than the other for these data? Do the measures of center appear similar? What does this say about the symmetry of the distribution?

2. **Sports and Leisure** Use the data collected in data project 2 of Chapter 2 regarding home runs. Determine the mean, mode, median, and midrange for the two data sets. Is one measure of center more appropriate than the other for these data? Do the measures of center appear similar? What does this say about the symmetry of the distribution?

3. **Technology** Use the data collected in data project 3 of Chapter 2. Determine the mean for the frequency table created in that project. Find the actual mean length of all 50 songs. How does the grouped mean compare to the actual mean?

4. **Health and Wellness** Use the data collected in data project 6 of Chapter 2 regarding heart rates. Determine the mean and standard deviation for each set of data. Do the means seem very different from one another? Do the standard deviations appear very different from one another?

5. **Politics and Economics** Use the data collected in data project 5 of Chapter 2 regarding delegates. Use the formulas for population mean and standard deviation to compute the parameters for all 50 states. What is the *z* score associated with California? Delaware? Ohio? Which states are more than 2 standard deviations from the mean?

6. **Your Class** Use your class as a sample. Determine the mean, median, and standard deviation for the age of students in your class. What *z* score would a 40-year-old have? Would it be unusual to have an age of 40? Determine the skew of the data, using the Pearson coefficient of skewness. (See Exercise 48, page 141.)

Answers to Applying the Concepts

Section 3–1 Teacher Salaries

1. The sample mean is $22,921.67, the sample median is $16,500, and the sample mode is $11,000. If you work for the school board and do not want to raise salaries, you could say that the average teacher salary is $22,921.67.
2. If you work for the teachers' union and want a raise for the teachers, either the sample median of $16,500 or the sample mode of $11,000 would be a good measure of center to report.
3. The outlier is $107,000. With the outlier removed, the sample mean is $15,278.18, the sample median is $16,400, and the sample mode is still $11,000. The mean is greatly affected by the outlier and allows the school board to report an average teacher salary that is not representative of a "typical" teacher salary.
4. If the salaries represented every teacher in the school district, the averages would be parameters, since we have data from the entire population.
5. The mean can be misleading in the presence of outliers, since it is greatly affected by these extreme values.
6. Since the mean is greater than both the median and the mode, the distribution is skewed to the right (positively skewed).

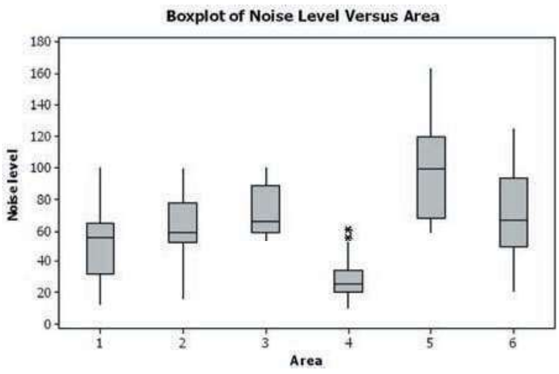
Section 3–2 Blood Pressure

1. Chebyshev's theorem does not work for scores within 1 standard deviation of the mean.
2. At least 75% (900) of the normotensive men will fall in the interval 105–141 mm Hg.
3. About 95% (1330) of the normotensive women have diastolic blood pressures between 62 and 90 mm Hg. About 95% (1235) of the hypertensive women have diastolic blood pressures between 68 and 108 mm Hg.
4. About 95% (1140) of the normotensive men have systolic blood pressures between 105 and 141 mm Hg. About 95% (1045) of the hypertensive men have systolic blood pressures between 119 and 187 mm Hg. These two ranges do overlap.

Section 3–3 Determining Dosages

1. The quartiles could be used to describe the data results.
2. Since there are 10 mice in the upper quartile, this would mean that 4 of them survived.
3. The percentiles would give us the position of a single mouse with respect to all other mice.
4. The quartiles divide the data into four groups of equal size.
5. Standard scores would give us the position of a single mouse with respect to the mean time until the onset of sepsis.

Section 3–4 The Noisy Workplace



From this boxplot, we see that about 25% of the readings in area 5 are above the safe hearing level of 120 decibels. Those workers in area 5 should definitely have protective ear wear. One of the readings in area 6 is above the safe hearing level. It might be a good idea to provide protective ear wear to those workers in area 6 as well. Areas 1–4 appear to be "safe" with respect to hearing level, with area 4 being the safest.