CHAPTER

6

Measures of Dispersion

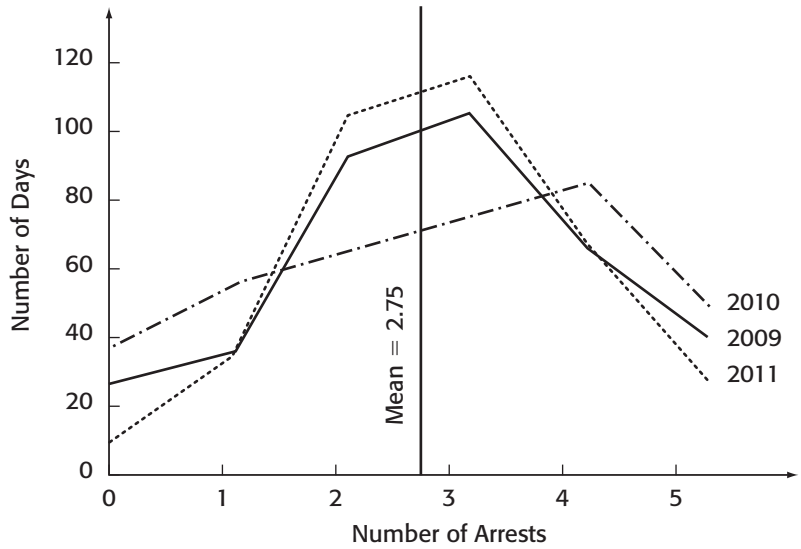
A useful descriptive statistic complementary to the measures of central tendency is a measure of dispersion. A **measure of dispersion** tells how much the data do (or do not) cluster about the mean. For example, the data listed in Table 6.1 show the number of daily arrests in Wheezer, South Dakota, for 2009, 2010, and 2011. The mean number of daily arrests for all 3 years is the same (2.75). How much the daily arrests cluster about the mean, however, varies. In 2010, the numbers cluster less about the mean than they do in 2009. In 2011, the arrests cluster closer to the mean than do either the 2010 or the 2009 arrests. This clustering is illustrated by the frequency polygons in Figure 6.1; the mean is depicted vertically to facilitate interpretation. (If you need to review frequency polygons, please see Chapter 4.) Clearly, the dispersion of the data is a valuable descriptive statistic in analyzing a set of data.

Many statistics texts discuss a variety of dispersion measures, such as the range, the average deviation, the interquartile deviation, and the standard deviation. Of all these measures, only the standard deviation has broad use and great value statistically, and hence this statistic is the only one we will discuss in this chapter. Nevertheless, it is a good idea to know what the other measures of dispersion are because you may encounter them in public and nonprofit management.

Table 6.1 Number of Daily Police Arrests in Wheezer, South Dakota

	Number of Days		
Number of Arrests	2009	2010	2011
0	24	36	10
1	36	54	36
2	95	65	109
3	104	74	118
4	66	84	66
5	40	52	26
	365	365	365

Figure 6.1 **Frequency Polygons for the Data in Table 6.1**



The **range** is the difference between the largest value and the smallest value in a distribution of data. The **interquartile deviation** is the difference between two values in the distribution selected so that the middle 50% of all observations fall between them. The **average deviation** is the average difference between the mean and all other values. If these measures of dispersion seem appropriate for any project, consult a general statistics book to find out how to calculate and use them [for example, see the text by Johnson and Kuby (2006); a list of other suitable texts is given in the annotated bibliography at the end of the book].

The Standard Deviation

The standard deviation is the most common measure of dispersion. The **standard deviation** is the square root of the average squared deviation of the data from the mean; that is, the standard deviation is based on the squared differences between every item in a data set and the mean of that set.

An example can explain this process far better than words. Normal, Oklahoma, has five street-cleaning crews. Because we have data for all five crews, we are working with the population and will calculate the population standard deviation. Later, we will explain the adjustment you would need to make if the five crews were a sample of the Normal, Oklahoma, street-cleaning crews (sample standard deviation). Listed in Table 6.2 are the numbers of blocks of city streets

Table 6.2 Blocks of Streets Cleaned by Work Crews

Work Crew	Number of Blocks
A	126
B	140
C	153
D	110
E	136
	665

cleaned by the five crews. To determine the (population) standard deviation for these data, follow these steps:

Step 1: Calculate the mean from the data. Recall from Chapter 5 that to calculate the mean, you need to add the data values across the cases (here, the sum of the data values is 665 blocks) and divide that sum by the number of cases or **N** (here, the number of work crews, 5). The mean is 133.

Step 2: Subtract the mean from every item in the data set. In this situation, subtract 133 from the number of streets cleaned by each crew. Often a table like Table 6.3 is helpful in performing these calculations. Note that the sum of these differences equals *zero.* In fact, in any distribution of data, the sum of the differences of the data values from the mean will *always* equal zero. The reason is that the positive deviations above the mean will just balance the negative deviations below the mean such that the sum of the deviations is zero. Because the sum of the deviations is zero, we must adjust our calculations to derive a more useful measurement of dispersion in Step 3.

Step 3: Square the difference between each data value and the mean (the third column in Table 6.3). As noted in Step 2, the differences sum to zero,

Table 6.3 Calculating the Differences

Number of Blocks	Subtract the Mean	Difference
126	133	–7
140	133	7
153	133	20
110	133	–23
136	133	3
		0

Table 6.4 Calculating the Difference Squared

Blocks	Mean	Difference	Difference Squared
126	133	–7	49
140	133	7	49
153	133	20	400
110	133	–23	529
136	133	3	9

regardless of how condensed or dispersed the data values are about the mean. The reason is that the positive and negative differences will always balance each other out. Squaring the differences avoids this problem and is instrumental in measuring the actual amount of dispersion in the data (see Table 6.4).

Step 4: Sum the squared differences. You should get a sum of 1,036.

Step 5: Divide the sum by the number of items ($\mathbf{N} = 5$). This number (207.2) is called the **variance**. The variance is the arithmetic average of the squared differences of the data values from the mean. The variance has little descriptive value because it is based on squared (not actual) units of measurement.

Step 6: Take the square root of the variance to find the standard deviation (here, the standard deviation is 14.4). Because we squared the differences between the mean and the data values in Step 3 (so that the differences would not sum to zero), it now makes sense to take the square root. In this manner, the standard deviation converts the variance from squared units to the original units of measurement. The standard deviation is the preferred measure of dispersion for descriptive purposes.

After calculating this relatively simple statistic, you should not be surprised to learn that statisticians have a complex formula for the standard deviation. The formula for σ (the Greek letter sigma, which statisticians use as the symbol for the standard deviation of a population) is

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

This formula is not as formidable as it seems: $(X_i - \mu)$ is Step 2, subtracting the mean of the population from each data value. $(X_i - \mu)^2$ is Step 3, squaring the differences. $\sum_{i=1}^{N}(X_i - \mu)^2$ is Step 4, summing all the squared differences from the first case $i = 1$ to the last case $i = \mathbf{N}$. The entire formula within the square

root sign completes Step 5, dividing by the number of items (**N**). Finally, the square root sign is Step 6.

How would the calculations change if the five street-cleaning crews were a sample of the crews from Normal, Oklahoma, rather than the entire population? You would then need to calculate the sample standard deviation, denoted by the symbol *s*. To do so, you would follow the steps in the formula for the standard deviation, replacing μ with $\bar{X}$, the sample mean, and **N** with $n - 1$, the number of cases in the sample (n) less 1. There are sound statistical reasons for making this adjustment, which we will discuss in Chapter 11, on statistical inference.

There are also good practical reasons. Dividing by $n - 1$ rather than n will yield a slightly larger standard deviation, which is the appropriate measure of care to take when you are using a sample of data to represent the population. In this example, the sample standard deviation would be 16.1 (1,036 ÷ 4 = 259; the square root of 259 = 16.1). When you are working with the population rather than a sample, there is no need to add this extra measure of caution. Thus, if the five Normal, Oklahoma, street-cleaning crews constitute the population, the (population) standard deviation is smaller, 14.4. You need to be aware that statistical package programs, spreadsheets, and hand calculators may assume that you are using the sample standard deviation rather than the population standard deviation and thus make the appropriate adjustment.

The smaller the standard deviation, the more closely the data cluster about the mean. For example, the standard deviations for the Wheezer, South Dakota, police arrests are 1.34 for 2009, 1.55 for 2010, and 1.16 for 2011 (see Figure 6.1). This calculation reinforces our perception that the 2010 data were the most dispersed and the 2011 data were the least dispersed. In general, when the data are closely bunched around the mean (smaller standard deviation), the public or nonprofit manager will feel more comfortable making a decision based on the mean.

Shape of a Frequency Distribution and Measures of Central Tendency

In addition to measuring the central tendency (see Chapter 5) and dispersion of a variable, public and nonprofit managers also need to know something about the "shape" of the frequency distribution of data values. The shape arises from plotting the values of the variable horizontally against their corresponding frequency of occurrence, which is plotted vertically. Several distinctive shapes of data distributions appear regularly in public and nonprofit administration and are important in analysis and interpretation. The shape of the distribution has implications for the usefulness of the different measures of central tendency.

Figure 6.2 shows a **symmetric distribution**. The data are evenly balanced on either side of the center or middle of the distribution. As you can see, each side is a reflection of the other. When a distribution is basically symmetric,

Figure 6.2 **A Symmetric Distribution**

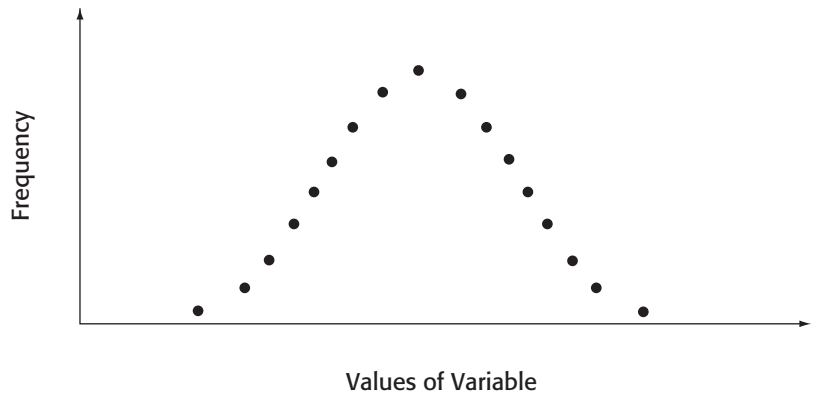
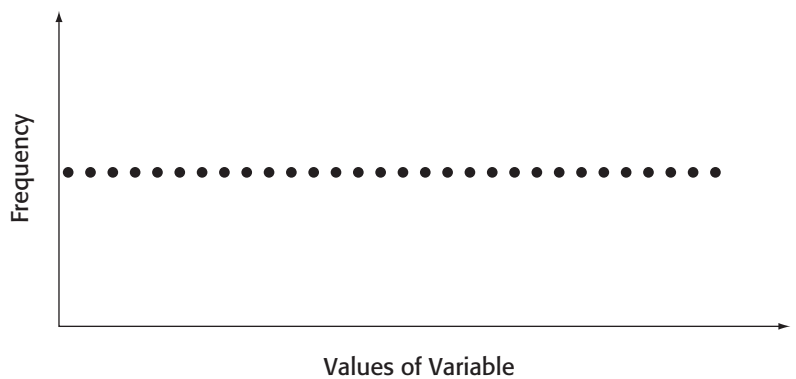


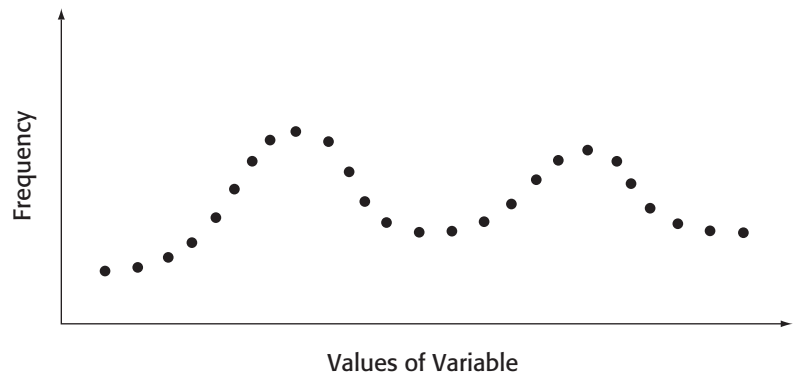
Figure 6.3 **A Uniform Distribution**



the mean, median, and mode have very similar values at or near the center of the distribution. In this situation, the mean is the preferred measure of central tendency.

Figure 6.3 shows a **uniform distribution**. In a uniform distribution, each data value occurs with the same (or nearly the same) frequency. Because the data do not cluster around the middle or center of the distribution but are evenly spread across the variable, the dispersion, as measured by the standard deviation, will be large. The mean and median will be near the center of this distribution. However, because many data values occur with the same

Figure 6.4 **A Bimodal Distribution**



or nearly the same frequency, the mode is generally not very useful in a uniform distribution.

Chapter 5 introduced the idea of a **bimodal distribution**. The shape of such a distribution has two distinct peaks, corresponding to data values that occur with high frequency, separated by other values that occur much less often. Figure 6.4 presents an example of a bimodal distribution. Note that in a bimodal distribution, the mean and median fall near the center of the distribution, but relatively few cases fall there (note the low frequency). As a result, in a bimodal distribution they are generally poor measures of central tendency. By contrast, consider the mode(s). The two modes (bimodal) capture the essence of the distibution and are preferred as measures of central tendency in this situation.

In an **asymmetric distribution**, the data fall more on one side of the center or middle than on the other side. In that case, skewness exists in the data.

Negatively skewed data have a few extremely low numbers that distort the mean. Negatively skewed data form a frequency distribution like the one pictured in Figure 6.5.

Positively skewed data have a few very large numbers that distort the mean. The frequency distribution of positively skewed data resembles the one shown in Figure 6.6.

If data are strongly skewed, the mean is not a good measure of central tendency (see Chapter 5). The reason is that the mean is "pulled," or skewed, in the direction of the skewness away from the center of the distribution. A few very high or low values in a very asymmetric distribution will skew the mean. In that case, the median is the preferred measure of central tendency because it is basically unaffected by skewness. Recall that in calculating the median, the relative positions of the values are what matters, rather than the actual magnitudes.

Figure 6.5 Negatively Skewed Data

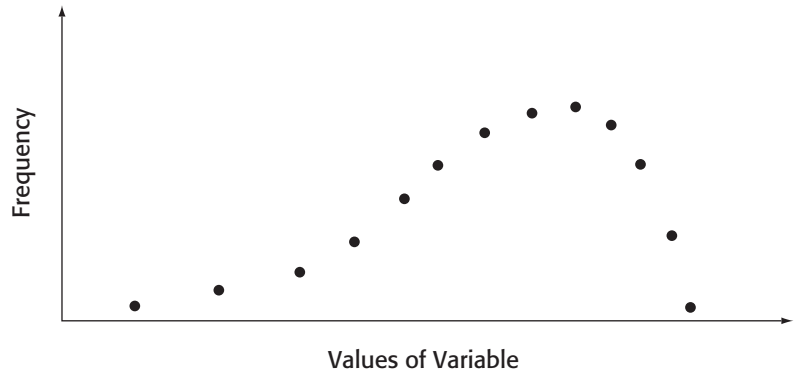
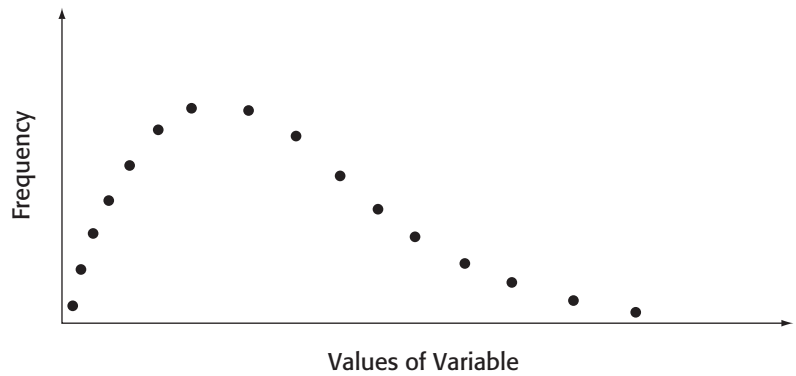


Figure 6.6 Positively Skewed Data



Recall our discussion in Chapter 5 regarding the use of the median as a measure of central tendency. If the standard deviation for a set of data approaches or exceeds the value of the mean itself, the mean may not be a very representative measure of central tendency. When the standard deviation exceeds the mean, extreme values in the data are often to blame. Consider using the median in cases like this, because it is not as sensitive to extreme values in the data.

A measure of "skewness" does exist and should be calculated by computer. Positive values of skewness indicate that the data are positively skewed and, therefore, the mean is artificially high. Skewness figures around zero indicate an unskewed (or symmetric) distribution. Negative numbers indicate negative skewness and, therefore, a mean that is artificially low.

Using Measures of Dispersion and Central Tendency Together

The standard deviation is an important complement to the mean. It's so important, in fact, that an analyst should be cautious in making conclusions about data if she is given only the mean and not the standard deviation as well. In the absence of the standard deviation, the mean has limited value because it is difficult to see how much volatility or variability exists in the data.

For example, let's say that a supervisor is informed that the mean score for a group of employees on a job skills test is 90%. The supervisor is very pleased with the test scores. Without a standard deviation, however, how much does the supervisor really know about the mean score? To illustrate the value of the standard deviation when interpreting the mean, let's look at a couple of different scenarios.

$$\mu = 90 \quad \sigma = 2$$
$$\mu = 90 \quad \sigma = 9$$

In the first scenario, where the standard deviation is 2, most individual test scores are tightly clustered around the mean. In the second scenario, where the standard deviation is 9, the individual test scores are far more dispersed around the mean.

The supervisor might reach far different conclusions about her employees' performance depending on which of these standard deviations was associated with the mean. A standard deviation of 2 suggests that most employees have scores close to the mean, whereas a standard deviation of 9 suggests that the performance of individual employees is much more variable. The two situations are highly different—yet indistinguishable without the standard deviation to accompany the mean.

It is hard to evaluate how useful or "descriptive" a measure of central tendency the mean is without having the value of the standard deviation as well. As a rule of thumb, an analyst should always provide both the mean and the standard deviation when describing data or reporting results.

Chapter Summary

Measures of dispersion indicate how closely a set of data clusters around the midpoint or center of the data distribution. Measures of dispersion include the range, the interquartile deviation, the average deviation, the variance, and the standard deviation. The first three of these measures are not as useful statistically as the last two.

For descriptive purposes, the standard deviation is used most often. The standard deviation is the square root of the average squared difference of the data values from the mean; it is the square root of the variance. This chapter illustrated the calculation and interpretation of the standard deviation. A small standard deviation (and variance) indicates that the data cluster closely around the mean.

Conversely, a large standard deviation indicates much greater dispersion in the data, so that the public or nonprofit manager needs to be more careful in this instance in making decisions based on the mean.

The chapter discussed the characteristic shape of several data distributions common in public and nonprofit administration. These included symmetric, uniform, bimodal, and asymmetric distributions.

In an asymmetric distribution, skewness can present a problem. Negatively skewed data have a few very low values that distort the mean; conversely, positively skewed data have a few very large values that will also distort. Because it is basically unaffected by skewness, the median is the preferred measure of central tendency in this situation.

Problems

6.1 Charles Jones, the local fire chief, wants to evaluate the efficiency of two different water pumps. Brand A pumps an average of 5,000 gallons per minute with a standard deviation of 1,000 gallons. Brand B pumps 5,200 gallons per minute with a standard deviation of 1,500 gallons. What can you say about the two types of pumps that would be valuable to Chief Jones?

6.2 Bobby Gene England, a research analyst for the United Way of Chickasaw, is asked to provide information on the number of daily visits to food banks in Chickasaw for the past 2 weeks. For the accompanying data, calculate the mean, median, and standard deviation for Bobby Gene.

6	9	0
11	12	5
4	7	10
3	3	12
9	8	

median = _____

mean = _____

standard deviation = _____

6.3 Helen Curbside, the chief custodial engineer for Placerville, has entered into the department computer the number of tons of garbage collected per day by all work crews in the city during a 1-week period. One statistic on the computer puzzles her: "Skewness 5 22.46." Interpret this result for Helen. What does it suggest about the performance of the city work crews?

6.4 Refer to Problem 5.4 (Chapter 5) on the Lance missile system. Five shots missed the target by 26, 147, 35, 63, and 51 feet, respectively. What is the standard deviation of the data?

6.5 The head of research and development for the U.S. Army must select for procurement one of the antitank weapons from the accompanying listing. The listed

results indicate the distance away from the intended target that 100 test rounds fell. Which system should the army select, and why?

Weapon	Mean	Standard Deviation
A	22.4	15.9
B	18.7	36.5
C	24.6	19.7

6.6 The Whitehawk Indian Tribe believes the Bureau of Indian Affairs responds faster to grant applications from the Kinsa Tribe than it does to applications from the Whitehawk Tribe. From the accompanying data, what can you tell the Whitehawks?

Days to Respond to Grant Applications

Whitehawk	Kinsa
64	50
58	72
66	74
54	46
70	75
66	81
51	43
56	46

6.7 An audit of the Community Rehabilitation Agency reveals that the average "26 closure" (rehabilitation jargon for a successful effort) takes 193 days with a standard deviation of 49 days and a skewness of 3.15. What does this mean in English?

6.8 The U.S. Postal Service is concerned with the time it takes to deliver mail. It would like not only to deliver mail as quickly as possible but also to have as little variation as possible. Twenty letters are mailed from New York state collection boxes to Cutbank, Montana. From the following number of days for delivery, what can you tell the Postal Service? Calculate all measures of central tendency and dispersion.

2	5	3	4	3	2	6	1	3	3
4	3	8	3	5	2	3	4	4	3

6.9 An employee at the Purchasing Department claims that he and a few other employees do almost all the work. In support of his claim, he collects the accompanying data on the number of purchase orders cleared and processed by each of the 16 members of the department in a typical week. Calculate all measures of central tendency and dispersion for these data and evaluate this employee's claim. Do a few employees do almost all the work?

12	22	8	14	15	32	17	24
20	37	15	23	16	40	19	21

6.10 The director of the South Bloomington YMCA has become concerned about the health of the organization's employees. He issues a directive to all YMCA sites advising all managers to encourage employees to get at least 60 minutes of exercise per day. Following are the data on the number of minutes exercised per day (on the average) by the 10 employees of the South Bloomington YMCA site. Calculate all measures of central tendency and dispersion for these data. How well do these employees meet the standard for exercise recommended by the director of the YMCA?

75 20 15 95 30 100 40 10 90 120

6.11 Complaints have reached the city manager of Normal that it is taking too long to pay bills submitted to the city. You are assigned to check how long it takes by looking at a few bills. Following are the lengths of time in days that it has taken the city to pay seven bills. Calculate the mean, median, and standard deviation. Would you report the mean or the median? Why?

34 27 64 31 30 26 35

6.12 The Texas State Penitentiary is concerned about the number of violent incidents in its prisons. After examining the 10 prisons in the Texas system, a data analyst finds the following data for the numbers of violent incidents last month:

17 21 42 32 16 24 31 15 22 26

Calculate the mean, median, and standard deviation. Should the penitentiary use the mean or the median in its analysis? Why?

6.13 The city manager of Grosse Out Pointe wants to make sure that pay raises are distributed fairly to all employees. He asks his assistant to gather data on raises for a random sample of police and fire employees, because he has received complaints that raises are not being distributed fairly in one of the departments.

Police	Fire
610	570
590	580
650	700
650	600
640	480
580	690
550	740
550	450
$\overline{X} = 603$	$\overline{X} = 601$

The assistant calculates the mean for each group of employees. He tells the city manager that because the average pay raise is just about the same in each department, there is really no fairness issue to worry about. If you were the city manager, would you accept these results, or would you ask the assistant to provide you with more information? Explain.

6.14 Jane Smart, director of the New Directions Youth Center, is planning a marketing campaign designed to attract more corporate donors. Because she has received several inquiries in the past regarding how much it costs to treat each client, Ms. Smart decides to collect these data so she can include this information in future promotional materials. She has data on 100 clients served over the past 6 months. Calculate the mean, median, and standard deviation for these data. Prepare a brief statement summarizing the cost data. *(Note: The data set for this problem is available on the book's Companion Website.)*

6.15 Data on spending per pupil across the entire population of Wisconsin public school districts are provided. Calculate the mean, median, and standard deviation for these data.

Next, use a statistical package program such as Statistical Package for the Social Sciences (SPSS), a statistical software program derived from "statistics" and "data" (STATA), or the like to create a frequency distribution for these data. Use your discretion when determining the number of categories in the distribution. To get started, make sure first to calculate the range for these data. *(Note: The data set for this problem is available on the book's Companion Website.)*

6.16 The Fowlerville, California, Department of Emergency Services is concerned about whether response times to 911 calls are within what the chief regards as an acceptable range. The head of emergency services has her intern collect data on response times for the most recent one hundred 911 calls and asks the intern to calculate the mean, median, mode, and standard deviation for these data. The head feels that average response times should be no greater than 5 minutes, because Fowlerville is a relatively small community (in both population and geographic size). What can the intern tell the head about response times? *(Note: The data set for this problem is available on the book's Companion Website.)*

6.17 The city manager of Grand Rapids, Michigan, has instructed her assistant to compile data on employee travel reimbursement payments for the past year. There were a total of 250 reimbursement claims. As part of her cost management initiatives, the manager would like to get a better sense of what these reimbursement data look like—that is, to understand the important characteristics of the distribution, including range, average, and dispersion. Assist the manager by calculating the mean, median, mode, and standard deviation for these data (reimbursements per trip in dollars).

Next, use a statistical package program such as SPSS, STATA, or the like to create a frequency distribution for the reimbursement data. Use your discretion when determining the number of categories in the distribution. To get started, make sure first to calculate the range for these data. *(Note: The data set for this problem is available on the book's Companion Website.)*

6.18 The director of the Madison County jail is planning to testify before the Wisconsin state legislature with the goal of obtaining more state funds for the county jail system. In preparing his remarks, the director plans to present data on the average number of days that prisoners remain in jail. The director believes that his case will be more compelling if he can show that the average stay for prisoners is greater than 21 days. The director asks you, as chief operations officer for the jail, to calculate descriptive statistics for a random sample of 250 prisoners processed in the last 9 months. Calculate the mean and standard deviation for these data. What can you tell the director about his 21-day target? *(Note: The data set for this problem is available on the book's Companion Website.)*