[illegible] 434 – 77

CHAPTER FOUR

4

DESCRIPTIVE STATISTICS

Figures won't lie, but liars will figure—Gen. C. H. Grosvenor

OBJECTIVES

When you have completed this chapter, you will be able to:

△

Compute the following measures of central tendency: mean, median, mode, and weighted mean.

△

Explain the advantages and disadvantages of each measure of central tendency.

△

Compute the following measures of variability: variance, standard deviation, and coefficient of variation.

△

Explain the advantages and disadvantages of using each measure of variability.

Financial Executive for May/June 1989 contained an article entitled, "Are Performance Fees Justified?" The *Yes* portion of this article argued that financial advisors should be paid based on the performance of their clients' portfolio accounts. The *No* position argued that advisors should be paid for their best efforts since portfolio performance depends on many factors other than the quality of advice △

The last two chapters described various ways of collecting data and displaying them. Frequency distributions and graphs are important for showing the essential properties of data collections to assist in the decision-making process. These methods are especially important if the collected data are qualitative, that is, measured on either a nominal or an ordinal scale.

WHY DESCRIPTIVE STATISTICS IS IMPORTANT TO MANAGERS

The methods for collecting and graphically presenting data discussed in Chapters 2 and 3 provide a starting point for data analysis. But managers also need to become acquainted with numerical descriptive measures that provide very brief and easy-to-understand summaries of a data collection. There are two broad categories into which these measures fall: measures of central tendency and measures of variability. Measures of central tendency describe the central location in a set of numerical observations. Measures of variability describe the spread or dispersion of the data values.

A possible way to resolve the issue posed by the *Financial Executive* article is to collect data on the performance of many different portfolio accounts, both those involving financial advisors and those not. These data values would need to be summarized in some way so that their essential characteristics become obvious. Important numerical methods of summarizing data collections are discussed in this chapter.

SITUATION 4.1

The Hearn Rope Company makes various lengths of rope and twine for industrial use. There have been some recent complaints from customers about the breaking strength of one of its main products, a heavy twine used to bind hay bales. Management of the Hearn Company is very concerned about quality control. It is easy to randomly sample sections of this twine as it comes off the weaving machine, and it is decided that this will be done over the next two months. The sections will be tested until they break, and the breaking points will be recorded. According to the sampling plan, all machines, shifts, and machine operators will be sampled, resulting in a total sample size of 350 twine sections. Management wants a brief summary of the sample results so that these data may be used to assess the extent of the twine problem.

MEASURES OF CENTRAL TENDENCY

In calculating summary values for a data collection, the first consideration is to find a central, or typical, value for the data. Four important measures of central tendency are presented in this section: mean, median, mode, and weighted mean.

Mean

The **arithmetic mean,** or simply the **mean,** is a summary value calculated by summing the numerical data values and dividing by the number of values. The mean is also referred to as the *calculated average*.

> The **arithmetic mean** of a collection of numerical values is the sum of these values divided by the number of values. The symbol for the population mean is the Greek letter μ (mu), and the symbol for a sample mean is $\overline{X}$ (X-bar).

Measurements of a data set are denoted

$$X_1, X_2, X_3, \ldots, X_N$$

where X_1 is the first measurement in the data set, X_2 is the second measurement in the data set, and so on up to X_N, the last or Nth measurement in the data set. For four measurements, 3, 6, 4, and 9, the data set is

$$X_1 = 3, X_2 = 6, X_3 = 4, X_4 = 9$$

There are usually two different data collections of interest in any statistical study: the population and the sample. Equation 4.1 is the formula for computing the mean of a population.

$$\mu = \frac{\Sigma X}{N} \tag{4.1}$$

where μ = Population mean
ΣX = Sum of all population data values
N = Population size

To simplify the computations in this text, some shorthand notation is used. In the simplified notation for summing all X values, ΣX, the summations are understood to extend from 1 to N. A more formal and complete notation system for this procedure is

$$\sum_{i=1}^{N} X_i$$

where the subscript i varies from its initial value of 1 to N in increments of 1. Since almost all sums run from 1 to N, the starting ($i = 1$) and ending (N) indices will be suppressed, and the simpler notation will be used.

Any measurable characteristic of a population, for example, the population mean (μ), is called a **parameter.**

> A population **parameter** is any measurable characteristic of a population.

Equation 4.2 is used to compute the mean of a sample.

$$\overline{X} = \frac{\Sigma X}{n} \tag{4.2}$$

where $\overline{X}$ = Sample mean
ΣX = Sum of all sample data values
n = Sample size

Any measurable characteristic of a sample, for example, the sample mean ($\overline{X}$), is called a **statistic.** A sample statistic is frequently used to estimate a population parameter.

A sample **statistic** is any measurable characteristic of a sample.

Characteristics of the arithmetic mean

1. Every data set measured on an interval or ratio level has a mean.
2. The mean has valuable mathematical properties that make it convenient to use in further computations.
3. The mean is sensitive to extreme values.
4. The sum of the deviations of the numbers in a data set from the mean is zero: $\Sigma(X - \mu) = 0$, and $\Sigma(X - \overline{X}) = 0$.
5. The sum of the squared deviations of the numbers in a data set from the mean is a minimum value: $\Sigma(X - \mu)^2$ is a minimum value, and $\Sigma(X - \overline{X})^2$ is a minimum value.

EXAMPLE 4.1 There are five people in a small class, and the instructor is interested in computing their mean age. Since the only people of interest are those in the class, this group constitutes a population, so Equation 4.1 is used to calculate the mean. The ages of the five people are 21, 19, 25, 19 and 23. Their mean is calculated as:

$$\mu = \frac{\Sigma X}{N} = \frac{21 + 19 + 25 + 19 + 23}{5} = 21.4$$

The mean or average age of the students in the class is 21.4 years. Note that the sum of deviations from the mean, $\Sigma(X - \mu)$, is equal to zero: $[(21 - 21.4) + (19 - 21.4) + (25 - 21.4) + (19 - 21.4) + (23 - 21.4)] = 0$.

EXAMPLE 4.2 The Atlas Welding & Sharpening Shop has 10 employees. Personnel records are checked, and the number of sick days used during the past month is recorded for each employee. Equation 4.1 is used to calculate their mean. The data values and the mean calculation are:

$$\mu = \frac{\Sigma X}{N} = \frac{3 + 0 + 5 + 6 + 1 + 0 + 11 + 8 + 0 + 4}{10} = 3.8$$

The computation reveals that an average of 3.8 sick days per hourly employee were taken during the past month.

Note that the sum of the squared deviations from the mean is equal to 127.6: $\Sigma(X - \mu)^2 = [(3 - 3.8)^2 + (0 - 3.8)^2 + (5 - 3.8)^2 + (6 - 3.8)^2 + (1 - 3.8)^2 +$

$(0 - 3.8)^2 + (11 - 3.8)^2 + (8 - 3.8)^2 + (0 - 3.8)^2 + (4 - 3.8)^2] = 127.6$. This is a minimum value because the population mean is the mathematical center of the distribution of 10 population values. If any value other than 3.8 is subtracted from the data values and the resulting deviations are squared and summed, the result will be a number larger than 127.6. Note that the results will be exactly the same if the data collection is treated as a sample.

The advantage of the arithmetic mean is that it is easy to compute, is understood by almost everyone, and is a good central value to use in summarizing a data collection, no matter how many values the collection contains. The disadvantage of the mean is that extreme values distort it. For this reason, the mean is not the best summary statistic for all data collections.

EXAMPLE 4.3 The sales of the six largest restaurant chains are presented in Table 4.1. A mean sales amount of $5,280 million is computed using Equation 4.2.

$$\overline{X} = \frac{\Sigma X}{n} = \frac{14,110 + 5,590 + 3,700 + 3,030 + 2,800 + 2,450}{6} = 5,280$$

Note that this mean has been distorted upward by the large sales of McDonald's, $14,110 million.

TABLE 4.1 Sales for the Six Largest Restaurant Chains, 1987 (Example 4.3).

Company	Sales ($ millions)
McDonald's	14,110
Burger King	5,590
Kentucky Fried Chicken	3,700
Hardee's	3,030
Wendy's	2,800
Pizza Hut	2,450

Source: C. Bovee and W. Arens, *Contemporary Advertising* (Homewood, Ill.: Richard D. Irwin, 1989).

EXAMPLE 4.4 The mean can also be distorted downward. The mean age of students in a pottery-making class might be 70 years if the people taking the class were all retired senior citizens. However, if a grade school student were to join the class, the mean might drop to 56 years. This would be a very misleading value to use to summarize the ages of the class.

Median

In cases where a typical, central value that does not suffer the distorting effects of extreme values is desired, the **median** is used to summarize the data. Approximately half the data values in a set are less than the median, and approximately half are greater. The median is sometimes referred to as the *counting average*.

The **median** of a data collection is the middle item in a set of observations that are arranged in order of magnitude.

Equation 4.3 is used to compute the item number of the median in a data set that is arranged in ascending or descending order.

$$\text{Median item number} = \frac{n + 1}{2} \tag{4.3}$$

Characteristics of the median

1. Every ordinal-level, interval-level, and ratio-level data set has a median.
2. The median is not sensitive to extreme values.
3. The median does not have valuable mathematical properties for use in further computations.

EXAMPLE 4.5 Table 4.2 shows 1987 sales for a random sample of 11 of the top 50 retail chains. Clint Stone wants to compute a summary value to indicate average sales. The mean of this sample, calculated using Equation 4.2, is $6,864.5 million. This is considered a misleading summary statistic because it is distorted upward by two large values in the data collection.

TABLE 4.2 Sales for 11 of the Top 50 Retail Chains, 1987 (Example 4.5)

Retail chain	Sales ($ millions)
Sears	28,085
J. C. Penney	15,332
A&P	9,532
Albertson's	5,869
Walgreen	4,282
Toys "R" Us	3,137
Tandy	2,452
Circle K	2,289
Nordstrom	1,920
Costco	1,370
Petrire Stores	1,242

Source: *Industry Surveys*, October 5, 1989, p. R79.

The median is the middle item in the set. Since the data values are arranged from highest to lowest, the median can be easily found. In this data collection, the median is $3,137 million: half the values are greater than this value, and half are smaller. The median item number can also be computed using Equation 4.3:

$$\text{Median item number} = \frac{n + 1}{2} = \frac{11 + 1}{2} = 6$$

The median, $3,137 million, is the sixth item in the array. This is a central, summary value that is not distorted by the sales of Sears and J. C. Penney, which are now weighted the same as the other values in the collection.

EXAMPLE 4.6 The median number of people treated daily at the emergency room of St. Luke's Hospital must be determined from the following data for the last six days:

25, 26, 45, 52, 65, 78

Since the number of values is even, the two values in the center are used to compute the median; their average represents the median of the collection. The calculated average of the two central values, 45 and 52, is 48.5, so this is the median. Half of the values in the data array are less than 48.5, and half are greater. The median item can also be computed using Equation 4.3:

$$\text{Median item number} = \frac{n+1}{2} = \frac{6+1}{2} = 3.5$$

Since the median is item 3.5 in the array, the third and fourth elements need to be averaged: (45 + 52)/2 = 48.5. Therefore, 48.5 is the median number of patients treated in St. Luke's emergency room during the six-day period.

The median separates a data array into two equal sections. If each section is subdivided by a new median, the result is four equal sections. Each of the three separating values is called a *quartile*; the middle quartile is the original median. An extension of this idea, which is often used on very large data arrays, is to separate the data into 100 sections, each with the same number of data elements. The separating values are then called *percentiles*.

Mode

It is sometimes important to know the most prevalent value in a data collection. The value that occurs most frequently is known as the **mode.** The mode is sometimes referred to as the *observed average*.

> The **mode** of a data collection is the value that occurs most frequently.

> **Characteristics of the mode**
> 1. Some data sets do not have a mode.
> 2. Some data collections have more than one mode.
> 3. The mode does not have valuable mathematical properties for use in further computations.

EXAMPLE 4.7 A data collection consists of the values 2, 3, 3, 5, 6, 4, 3, 6, 7, 9, 3, 2, and 6. The mode of this collection is 3, since there are more 3s (i.e., four of them) than any other number.

EXAMPLE 4.8 The mode is to be determined for the following data values:

12, 14, 15, 16, 15, 18, 19, 20, 14

In this data array, two values occur with a frequency of two: 14 and 15. Therefore, the collection can be said to be bimodal, with modes 14 and 15. If no value appeared more than once, the data collection would have no mode.

Table 4.3 compares the advantages and disadvantages of the mean, median, and mode.

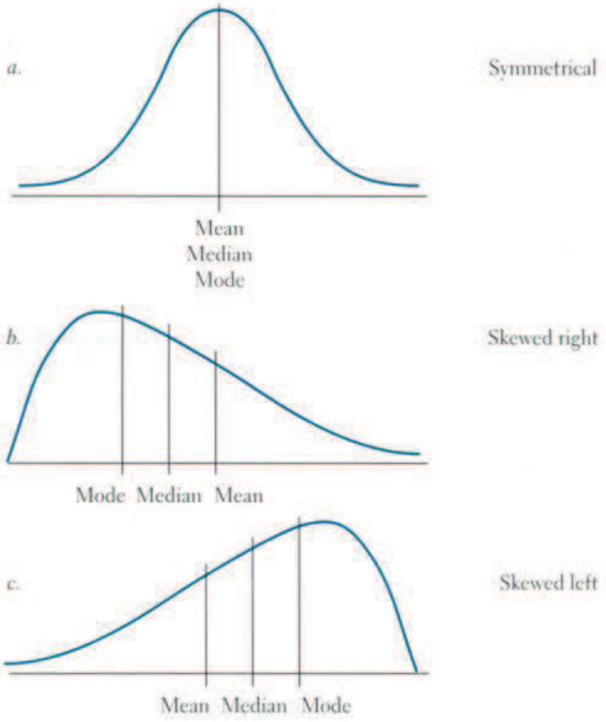
Analysts are frequently concerned about how data values are distributed, that is, how values in the data collection are spread out between the extremes. Differences among the mean, median, and mode can be easily seen from graphs of **symmetrical**

TABLE 4.3 A Comparison of the Mean, Median, and Mode

Average	Advantages	Disadvantages
Mean	Reflects the value of every data point Easy to compute and understand Has valuable mathematical properties; useful for further computations	Unduly influenced by extreme values
Median	Not distorted by extreme values	Lacking certain mathematical properties
Mode	Value that appears most frequently	Lacking certain mathematical properties Some data sets have no mode

and **skewed distributions.** Figure 4.1 shows three curves, one that is symmetrical (curve *a*) and two that are skewed (curves *b* and *c*). Curve *a* depicts a symmetrical distribution because a vertical line drawn from the peak of the curve to the horizontal axis will divide the area of the curve into two equal, symmetrical parts. Note that the mean, median, and mode are all located at the peak value of curve *a*.

FIGURE 4.1 Symmetrical and Skewed Distributions



A **symmetrical distribution** is represented by a curve that can be divided by a vertical line into two parts that are mirror images.

Curves *b* and *c* in Figure 4.1 are referred to as skewed curves because they lack symmetry. Values in such distributions are concentrated at either the low end or the high end of the scale along the horizontal axis. Curve *b* is skewed to the right, or the high end of the scale (such a distribution is sometimes referred to as *positively skewed*). The mean is drawn away from the highest point of the curve toward the skewed end. This is because the mean is sensitive to a few extreme values at that end of the curve. The mode (the value that occurs most frequently) is the X value corresponding to the highest point of the curve, and the median (the most typical value) is located between the mean and mode.

When the distribution is skewed to the left, as shown in curve *c* of Figure 4.1, the mean is pulled away from the highest point of the curve toward the low end of the scale. The median is pulled down, but not as much, and the mode remains at the highest point on the curve.

> A **skewed distribution** is represented by a curve that lacks symmetry.

Figure 4.1 illustrates the effects of skewness on the three averages discussed in this section. In particular, this figure demonstrates the importance of determining the extent of skewness of a distribution before choosing among the available averaging methods for a summary value. The median is generally the best measure of central tendency to use when a distribution is skewed.

Weighted Mean

The final measure of central tendency to be discussed in this chapter is the **weighted mean,** so named because in its calculation some data values are given more weight than others.

> The **weighted mean** assigns more weight to some data values than to others.

In calculating the arithmetic mean of a data array using Equation 4.1 or 4.2, it is implicitly assumed that each data value carries the same weight. If, for some reason, certain data values are more important than others, different weights can be assigned to the values in the calculation of the mean.

Equation 4.4 is used to calculate the weighted mean for either a population or a sample.

$$\overline{X}_w = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} \tag{4.4}$$

where $\overline{X}_w$ = Weighted mean
X_i = Data values to be averaged
w_i = Weights applied to the X values

Note from Equation 4.4 that each selected weight is multiplied by the corresponding data value. Next these weighted values are summed. Finally, this summation of weighted values is divided by the summation of the weights. The result is the computation of a mean to which some data values contribute more than others.

EXAMPLE 4.9 Professor Chin gives three regular exams, a midterm, and a final exam in his statistics class each semester. These exams are averaged to determine each class member's final grade. The three regular exams each account for 15% of the grade, the midterm accounts for 25%, and the final accounts for 30%. Thus, the weights for the five exams are: .15, .15, .15, .25, .30. One class member achieves the following scores during the quarter: 75, 82, 84, 79, 91. This person's final average score using the weighted mean is:

$$\overline{X}_w = \frac{w_1X_1 + w_2X_2 + w_3X_3 + w_4X_4 + w_5X_5}{w_1 + w_2 + w_3 + w_4 + w_5}$$

$$\overline{X}_w = \frac{(.15)75 + (.15)82 + (.15)84 + (.25)79 + (.30)91}{.15 + .15 + .15 + .25 + .30}$$

$$\overline{X}_w = \frac{83.2}{1} = 83.2$$

EXAMPLE 4.10 Table 4.4 contains data on the percentage of the civilian labor force unemployed in three counties of Eastern Washington. Barbara McWilliams has been asked to make a presentation to the Spokane Economic Development Council showing the unemployment rate for this region. The unemployment rates and labor force sizes for the three counties are shown in Table 4.4. She calculates the mean percentage unemployed for the three counties to be $(15.1 + 13.4 + 7.8)/3 = 12.1\%$. However, since Spokane County is much larger than the other two counties; Barbara feels that this figure does not truly reflect the unemployment rate for the region. She decides that a weighted mean is more appropriate and calculates it using the civilian labor force for each county as the weights:

$$\overline{X}_w = \frac{(.151)(7,360) + (.134)(3,670) + (.078)(162,300)}{7,360 + 3,670 + 162,300}$$

$$= \frac{14,263}{173,330} = .082$$

The weighted average of the percentage of the labor force unemployed for the region is equal to 8.2%. This value is much more representative of the region because it properly reflects the size of the labor force in the largest county.

TABLE 4.4 Civilian Labor Force (Example 4.10)

County	Percentage unemployed	Civilian labor force
Adams	15.1	7,360
Pend Oreille	13.4	3,670
Spokane	7.8	162,300

Source: Washington State Employment Security Department, Labor Market and Economic Analysis Branch, July 1986.

The analyst must carefully choose which of the four methods of summarizing central tendency is most appropriate for a given data collection. The mean is commonly used but may not be appropriate if there are extreme values in the collection. In these cases, the median may provide a more accurate summary statistic. The mode is used when the value that occurs most frequently is desired. Its disadvantage is that this value may not accurately represent the entire collection. The weighted mean is used when certain data values are more important than others.

SITUATION 4.1
Resolved

The management of the Hearn Rope Company selects a random sample of 350 twine sections from their manufacturing process during a two-month period and records their breaking points in pounds. Hearn's analyst, Ted Mack, decides that there are no extreme values among these breaking points. Therefore, the mean is a good summary value. The mean or average breaking point of the twine sections is computed to be 127.5 pounds. This value is given to management so they can deal with the complaints of their farming customers. A new concern voiced by management involves the extent to which the breaking points vary around this mean. Ted agrees that it would make a big difference whether the breaking points were all about 127 pounds or if some were a lot more and some a lot less.

EXERCISES

1. Which measure of central tendency (mean, median, mode) is most sensitive to extreme values?
2. Which measure of central tendency is used to indicate the value with the greatest frequency?
3. Which measure of central tendency places differing amounts of importance on the values in a data collection?
4. When a data collection contains extreme values, which measure of central tendency should be used?
5. Which measure of central tendency takes into account the value of every item in a data collection in its computation?
6. Which measure of central tendency is useful for performing statistical procedures such as comparing central tendencies of several data sets?
7. Which measure of central tendency has mathematical properties that enable it to be easily used in further computations?
8. Which measure of central tendency allows different weights to be assigned to the values being averaged?
9. Which measure of central tendency would be a good choice for an average for a collection containing many small values and one very large value?
10. If one of the values slightly larger than the mean in a data collection is replaced by a very large value, does the mean go up, go down, or stay the same? How does this replacement affect the median?
11. If you wanted an average to be proportional to the total income of a community, which measure of central tendency would you use?
12. If you wanted an average to represent the income received by the most people in a community, which measure of central tendency would you use?
13. If you wanted an average to represent the incomes of a community, in the sense that it would differ as little as possible from those incomes, which measure of central tendency would you use?
14. If you were to manufacture a new aluminum window screen and wished to produce only one size, which measure of central tendency would you use to assess market demand?
15. What is the difference between a sample statistic and a population parameter?
16. What is the difference between a symmetrical distribution and a skewed distribution?

17. Indicate whether the distribution for each of the following variables is symmetrical, skewed right, or skewed left.
    a. Annual household income
    b. Lengths of rolls of wallpaper
    c. Scores on a very easy statistics exam
    d. Family size
    e. Mileages of tires before wearout
18. Indicate where the mean, median, and mode are located on each of the following types of distribution:
    a. Skewed right
    b. Symmetrical
    c. Skewed left
    d. Values concentrated at the upper end of the scale
    e. Symmetry lacking in the upper end
19. What is the shape of the distribution described by the following measures of central tendency: mean = 46, median = 42, mode = 39?
20. What is the shape of the distribution described by the following measures of central tendency: mean = 3.1, median = 3.1, mode = 3.1?
21. What is the shape of the distribution described by the following measures of central tendency: mean = 105, median = 110, mode = 115?
22. The following data show a population consisting of the number of Snickers candy bars purchased from a cafeteria vending machine on the first 10 days of operation: 7, 3, 0, 5, 8, 6, 7, 10, 1, 3.
    a. Calculate the mode, median, and mean for this data collection.
    b. Which measure of central tendency would you use to estimate monthly sales of Snickers from this vending machine?
    c. Compute the sum of the deviations from the mean: $\Sigma(X - \mu)$.
    d. Compute the sum of the squared deviations from the mean: $\Sigma(X - \mu)^2$. Is it possible to obtain a smaller sum of squared deviations by using any number other than the mean?
23. Ten bear markets have ravaged investors in the past four decades. In the rate of their descent, none of these declining markets rivaled what occurred in the fall of 1987. The most important question being asked at that time was how long it would last. Here are data on the ten bear markets:

Bear market	Months long
1948–1949	8
1953	8
1957	3
1960	10
1961–1962	7
1966	8
1968–1970	18
1973–1974	21
1976–1978	17
1981–1982	16

Source: "The Months Ahead," *Changing Times*, December 1987.

a. What is the modal length of the bear market?
b. Which value divides the data collection into two equal parts?
c. Calculate the mean for this data collection.

d. Which of these measures of central tendency would you use to estimate the duration of the 1987 bear market?

24. The manager of Ted's Corner Grocery, Ted Mitchell, decides to investigate the average amount spent by consumers on groceries during a one-week period. The following data represent the amounts spent last week by a randomly selected sample of 12 customers.

\$ 65	\$ 75	\$ 85
153	250	99
80	191	55
131	93	182

a. Calculate the mode, median, and mean for this data collection.
b. Which measure of central tendency would you use to indicate the typical amount of groceries purchased?

25. The following data represent several forecasts of the number of passenger cars in the United States for 1983, in millions of units:

Forecaster	Number of cars (millions)
Chase	10.9
D.R.I.	9.5
General Motors	12.8
Wharton	10.3
K.E.D.I.	8.3
U.S. Market Research	7.5

Source: "Sociometrics," in *An Executive's Guide to Econometric Forecasting*, ed. A. Migliaro and C. L. Jain, Graceway Publications, 1983.

a. Calculate the mode, median, and mean for this data collection.
b. Which measure of central tendency would you use to forecast passenger cars in the United States for 1983?

26. Anderson Motors sold 53 Honda Civics in 1988 for the regular price of \$8,250 (standard price, with options costing extra). In October, when the new 1989 models arrived, the standard price was reduced to \$7,350. Anderson sold 15 Civics at this reduced price. In December, Anderson had a closeout sale and sold 7 Civics for \$6,650. What was the average price that Anderson received for Honda Civics in 1988 (not including the cost of optional extras)?

27. The Burt Distribution Corporation has a main office in Madison, Wisconsin, and a branch office in Spokane, Washington. The branch manager, Julie Pearson, is concerned about the amount of money being spent on sending 1- to 2-pound packages to the main office. The following quantities indicate the volumes of packages sent at different postal rates for the past year:

Type of mailing	Number of packages	Rate
Fourth class	2,023	\$2.13
Third class	5,478	1.38
First class	8,457	2.40
Special delivery	1,023	2.95
Registered	423	3.60

What was the average cost of sending 1- to 2-pound packages to the main office?

28. The Perfection Tire Company wants to determine the average mileage for a particular tire before wearout so that a warranty policy can be established. A sample is selected, and the following mileages are recorded to the nearest thousand miles:

33	41	55	47	38	45	47	46	48	39	40	40	41	42
38	48	50	49	36	44	44	45	42	35	46	43	47	47

a. Calculate the mode, median, and mean for this data collection.
b. Is the data collection symmetrical or skewed?
c. Which measure of central tendency would you use to help determine the warranty policy?

29. Alton's Traction Headquarters sells four types of Goodyear tires. The following table lists the volume and price for each type of tire sold during a recent month:

Type of tire	Number sold	Price
F-32 radial	288	$49.95
Tiempo radial	940	29.95
Arriva radial	348	39.95
Vector radial	456	44.95

What was Alton's average revenue per Goodyear tire sold?

30. Jim Donaldson, Marketing Director of the Clear Soft Drink Company, wants to determine the average selling price (in cents) for 8-ounce cans of soft drinks in Chicago supermarkets. He samples 44 brands and finds the following prices:

55	52	62	78	41	45	45	65	72	49	55
65	54	55	77	54	65	45	48	70	60	50
40	42	56	81	49	61	63	66	69	48	42
67	50	59	70	59	68	41	41	77	65	54

a. Calculate the mode, median, and mean for this data collection.
b. Is the data collection symmetrical or skewed?
c. Which measure of central tendency should Jim Donaldson use if he is interested in determining the typical price of an 8-ounce can of pop?

SITUATION 4.2

The Rockford Grain Growers Company has selected 50 of its acres for inclusion in a sample for a study designed to investigate the desirability of introducing peas into its product mix. During the growing season these acres will be planted with peas and will receive the same water and fertilizer treatment as its other acreage. Gail Holden, the president of Rockford, believes the sampled acres will provide a good indication of the yield that could be anticipated if large amounts of peas were planted. Because extreme yields per acre are not anticipated, Gail decides that the mean yield will be a good summary statistic. By knowing the average yield per acre, Rockford will be in a good position to determine if peas will be a profitable crop. However, the variability of yield is also of interest. Gail indicates to her staff that they need a way of measuring the extent to which the yields of the 50 sampled acres vary.

Ko[illegible] 437 – [illegible]

MEASURES OF VARIABILITY

Computing a measure of central tendency for a data collection is a valuable way to summarize the numerical values, especially if there are a large number of them. There is another measurement of equal importance, however. It is frequently necessary to know the extent of variability of the numbers in a data collection or distribution. The best descriptions of variability concern the deviations of the data values from some measure of central tendency. Since the mean has beneficial mathematical properties, it is the measure of central tendency most commonly used. Two important measures of variability are presented in this section: *variance* and *standard deviation*.

The variance and standard deviation are measures of how the data tend to vary around the mean. If the data are widely scattered around the mean, the variance and standard deviation will be relatively large. If the data are tightly clustered around the mean, both measures will be relatively small.

Variance

The **variance** is the average of the squared differences between the data values and the mean. It has certain mathematical properties that make it quite useful in other statistical applications. Some of these uses will be presented in later chapters. However, the interpretation of the variance as the average of the squared differences is not useful as a descriptive measure. Equation 4.5 shows how the variance for a population of data values is computed.

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N} \tag{4.5}$$

where σ^2 = Population variance
X = Population values
μ = Population mean
N = Number of observations in the population

> The **variance** is the average of the squared differences between the data values and the mean.

TABLE 4.5 Data Collection of Ages (Examples 4.11 and 4.12)

	X	$X - \mu$	$(X - \mu)^2$
	20	−20	400
	30	−10	100
	40	0	0
	50	10	100
	60	20	400
Sums:	200	0	1,000

$\Sigma X = 200$; $\Sigma(X - \mu) = 0$
$\Sigma(X - \mu)^2 = 1{,}000$
Mean $= \mu = 200/5 = 40$
Variance $= \sigma^2 = 1{,}000/5 = 200$
Standard deviation $= \sigma = \sqrt{200} = 14.1$

EXAMPLE 4.11 Table 4.5 presents a population of five ages. The mean of this data array is found by summing the values and dividing by N, so $\mu = 200/5 = 40$. The mean age of 40 accurately describes the central tendency of the data collection. But what about variability? To what extent do the data values differ from their mean? In column 2 of Table 4.5, the deviations from the mean are calculated and summed. However, taking the average of these deviations provides no indication of variability. The summation is equal to zero since the mean is at the mathematical center of the array, and the negative values cancel out the positive values; this concept was illustrated in Example 4.1

One mathematical approach to eliminating the minus signs is to square each of the deviations. This has been done in column 3 of Table 4.5. After the deviations are squared, Equation 4.5 is applied, and the variance of ages is found to be 200:

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N} = \frac{1{,}000}{5} = 200$$

Unfortunately, knowing that the average squared deviation in Example 4.11 is equal to 200 is not very meaningful. One solution to this problem is to return to the original units of measure by taking the square root of the variance. Equation 4.6 shows how the square root of the variance, called the population **standard deviation,** is computed. The standard deviation is the standard amount by which the values in a data collection differ from the mean.

Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}} \tag{4.6}$$

The **standard deviation** measures the amount by which the values in a data collection differ from the mean.

EXAMPLE 4.12 In the last column of Table 4.5, values are calculated for use in Equation 4.6 to compute the standard deviation of the population data. This column shows the squared difference between each data value and the mean of the distribution, 40. The sum of these squared deviations is 1,000, and their average is 200. The square root of 200 is 14.1, so this is the standard deviation of the data array. The data values have a mean of 40 and a standard deviation of 14.1. This means that the standard amount by which the values in the array differ from their mean (40) is about 14.1.

The standard deviation is commonly used to describe the extent to which a collection of data values is dispersed around its mean. A small standard deviation means that the values tend to be close to their mean. A large standard deviation means that the values are widely scattered about their mean.

Standard Score

The standard deviation is also useful for describing how far individual values are located from the mean. A measure called the **standard score** indicates the number of standard deviations by which a particular value lies above or below the mean. Equation 4.7 shows how a population standard score is calculated.

$$\text{Standard score} = \frac{X - \mu}{\sigma} \tag{4.7}$$

where X = Value of interest in the population
μ = Population mean
σ = Population standard deviation

> The **standard score** indicates the number of standard deviations by which a particular value lies above or below the mean.

Example 4.13 Example 4.12 presented a distribution that had a mean of 40 and a standard deviation of 14.1. The standard score for the observation 50 is:

$$\text{Standard score} = \frac{X - \mu}{\sigma} = \frac{50 - 40}{14.1} = 0.71$$

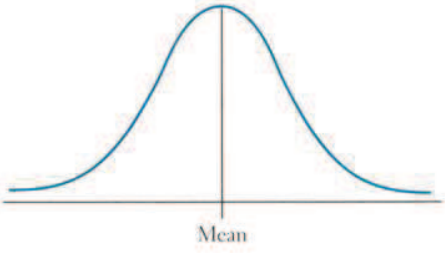
A value of 50 from this distribution has a standard score of 0.71. The standard score indicates that the value 50 deviates from the mean, 40, by 0.71 standard deviation units: 0.71 (14.1) = 10. Standard scores will be used extensively in Chapter 6.

Since the standard deviation is a measure of standard distance from the mean, knowing the mean and the standard deviation provides the analyst with great insight about the data from which they have been computed. Two important statistical concepts are often used in this process: Chebyshev's theorem and the empirical rule.

Chebyshev's Theorem

Chebyshev's theorem (sometimes referred to as *Tchebysheff's theorem*), credited to the Russian mathematician P. L. Chebyshev, states that no matter what the shape of a distribution, at least 75% of the values will fall within ±2 standard deviations of the mean of the distribution, and at least 89% of the values will lie within ±3 standard deviations of the mean. Chebyshev's theorem will be discussed in more detail in Chapter 6.

Figure 4.2 Bell-Shaped Distribution



Empirical Rule

A popular rule of thumb in statistics called the *empirical rule* involves a collection of data values that is symmetrical about its mean with most of the values situated close to the mean. Such a bell-shaped distribution appears in Figure 4.2 and has the following characteristics:

1. Approximately 68% of the values are within 1 standard deviation of the mean.
2. Approximately 95% of the values are within 2 standard deviations of the mean.
3. Approximately 99.7% of the values are within 3 standard deviations of the mean.

Note that Chebyshev's theorem is extremely conservative compared to the empirical rule (75% of the values instead of 95% lie within 2 standard deviations). The reason is that Chebyshev's theorem applies to any population or sample, without regard to the shape of the distribution. A detailed discussion of the empirical rule will be presented in Chapter 6.

The data of Table 4.5 were defined as a population, and Equations 4.5 and 4.6 were used to calculate the population variance and standard deviation. If the measured data constitute a sample, the calculations differ slightly. Equations 4.8 and 4.9 are used to calculate the sample variance and standard deviation. Note that the denominator is 1 less than the sample size ($n - 1$), as opposed to the entire population size (N) used in Equations 4.5 and 4.6. Note also that whereas the Greek letter σ (sigma) is used to represent the population parameter, the letter s is used to represent the sample statistic.

Sample variance:

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1} \tag{4.8}$$

Sample standard deviation:

$$s = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n - 1}} \tag{4.9}$$

where s^2 = Sample variance
s = Sample standard deviation
X = Sample values
$\overline{X}$ = Sample mean
n = Number of observations in the sample

Degrees of Freedom

The denominator in Equations 4.8 and 4.9, ($n - 1$), represents the **degrees of freedom.** This term appears frequently in statistical applications and refers to the number of data values in the sample that are free of each other in the sense that they carry unique information.

Degrees of freedom refers to the number of data elements that are free to vary.

The calculation of Equation 4.9 uses the sample mean ($\overline{X}$) as an estimate of the population mean (μ). If the sum of the squared deviations in the numerator of Equation 4.9 were divided by the sample size, n, a biased standard deviation would result. That is, the value of s, which is an estimate of the unknown population standard deviation, σ, would tend over many trials to be slightly too small. This is because the $\Sigma(X - \overline{X})^2$ computation provides a minimum value, as was illustrated in Example 4.2. If the actual population mean were used in Equation 4.9, the numerator would tend to be slightly larger.

Mathematicians have discovered that the small-size bias of the numerator in Equation 4.9 is compensated for by reducing the denominator. By using $(n - 1)$ as the denominator of the sample standard deviation calculation, the bias is removed, and an unbiased estimate of the unknown population standard deviation results.

In general, a piece of sample information is lost each time a sample statistic is used to estimate an unknown population parameter in an equation. In Equation 4.9, it would be preferable to measure the variability of the sampled items around the true population mean, but since this value is not known, an estimate—the sample mean—is used in its place.

A short-cut formula has been derived for calculating the sample variance and standard deviation. This is handy when the data being evaluated number more than a few items. Equation 4.10 is the short-cut formula used to compute the sample variance. The sample standard deviation is computed by taking the square root of this variance.

$$s^2 = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n - 1} \qquad (4.10)$$

Both the sum of the sample X values and the sum of their squared values are needed for the calculation. It is important to note that the ΣX^2 does not equal $(\Sigma X)^2$, as pointed out in Chapter 1.

Is is easy to become confused by all the equations used to compute the variance and standard deviation. However, Equation 4.10 can be used in most cases since real-world situations usually involve samples.

TABLE 4.6 Jarms Appliances Sold (Example 4.14)

4	5	12	9	10	8
7	4	5	3	0	1
8	2	15	7	9	11
9	8	7	8	6	12

$n = 24$
$\Sigma X = 170$
$\bar{X} = 170/24 = 7.08$
$\Sigma X^2 = 1{,}512$

EXAMPLE 4.14 Table 4.6 shows a data collection representing the number of units sold per day in a random sample of selling days for Jarms, an appliance dealer. As shown, the mean of the sample is 7.08 units. Since the data represent a sample, Equation 4.10 can be used to calculate the standard deviation. As shown in Table 4.6, the X values sum to 170, and the sum of the squared X values is 1,512. Following are the squared X values used to compute this sum:

16	25	144	81	100	64
49	16	25	9	0	1
64	4	225	49	81	121
81	64	49	64	36	144

The variance can now be computed using Equation 4.10:

$$s^2 = \frac{\Sigma X^2 - \dfrac{(\Sigma X)^2}{n}}{n - 1} = \frac{1{,}512 - \dfrac{(170)^2}{24}}{24 - 1} = \frac{307.8}{23} = 13.38$$

The square root of the variance is the standard deviation:

$$s = \sqrt{13.38} = 3.66$$

Jarms Appliances sells an average (mean) of 7.08 units per day, and the typical amount of variation in these sales is 3.66 units from the mean of 7.08.

Many handheld calculators automatically figure the mean and standard deviation. In calculating standard deviation, it is necessary to know whether the data constitute a population or a sample so that the proper denominator can be used. Calculators commonly have two keys for this purpose: the N key for populations and the $(n - 1)$ key for samples.

In addition, almost all data analysis programs for either mainframe or personal computers include the mean and standard deviation as summary measures of data collections. The MINITAB commands for computing descriptive statistics for the Jarms Appliance Store example are:

```
MTB> SET INTO C1
DATA> 4 5 12 9 10 8 7 4 5 3 0 1 8 2 15 7 9 11 9
DATA> 8 7 8 6 12
DATA> END
MTB> DESCRIBE C1

C1         N         MEAN        MEDIAN       STDEV
          24        7.083         7.50        3.658

C1       MIN          MAX           Q1          Q3
         0.0         15.0         4.25         9.0

MTB> STOP
```

The variability of the mean and standard deviation for different data distributions is shown in Figure 4.3. In part *a*, three distributions appear. Since they all have the same shape, it is apparent that these three symmetrical distributions have the same standard deviation. That is, the data values of all three distributions are approximately the same distance from their respective means. However, their means are different; each is at a different point on the horizontal axis.

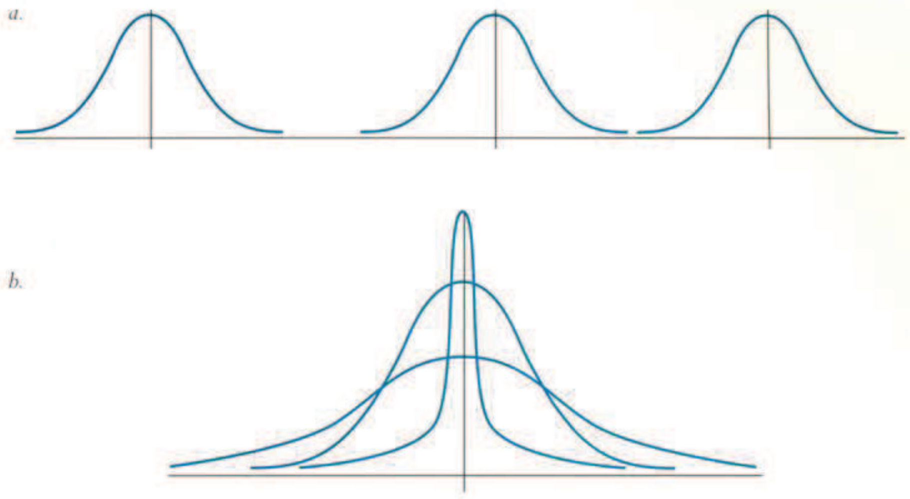
By contrast, the three symmetrical curves of part *b* all have the same mean. However, their standard deviations are different. The data values of the widest curve are typically far from the mean compared with those of the middle curve, and the inner-curve data values are even more tightly packed about the mean. Figure 4.3 illustrates the importance of knowing both the mean and standard deviation of a data distribution for an accurate summary of the distribution.

Coefficient of Variation

Another method sometimes used to measure the variability of a population or sample of data values is the **coefficient of variation.** This statistic specifies the size of the standard deviation as a percentage of the mean. It is calculated by dividing the standard deviation by the mean. The coefficient of variation indicates the *relative* amount of variability in a distribution.

The **coefficient of variation** for a data collection expresses the standard deviation as a percentage of the mean.

FIGURE 4.3 Data Distributions



Equation 4.11 is used to calculate the coefficient of variation for a sample. The abbreviation CV is used for this statistic.

$$CV = \frac{s}{\overline{X}}(100) \tag{4.11}$$

Note that for a population, s in the numerator would be replaced by σ, and $\overline{X}$ in the denominator would be replaced by μ.

The coefficient of variation is used by decision makers to: (1) determine how reliable the mean is as a measure of central tendency, (2) assess whether the standard deviation is large, small, or somewhere in between, or (3) compare the variability of two or more distributions.

EXAMPLE 4.15 The president of First Federal Savings Bank has been disturbed by the inaccuracy of statistics received from the federal government. She has been shown an article in *The Wall Street Journal* (August 31, 1989) indicating that many statistics issued by the government are subsequently revised. She decides to begin a more intensive collection of data within the bank system rather than depend on outside sources.

Kim Horns, a new analyst for the bank, is asked to compare the average account balances for the downtown office and a suburban branch. A random sample of savings account balances is drawn from both locations. For the main office, the mean account balance is \$1,248.50, and the standard deviation is \$537.93. At the branch, the mean is \$743.84, and the standard deviation is \$325.10. It is obvious to Kim that the suburban branch has a lower average account balance, but she finds it difficult to compare the variability of the two branches due to the difference in means. For this reason, Kim decides to compute the coefficient of variation for each branch:

Main office:

$$CV = \frac{s}{\overline{X}}(100) = \frac{537.93}{1,248.5}(100) = 43.1\%$$

Suburban office:

$$CV = \frac{s}{\overline{X}}(100) = \frac{325.10}{743.84}(100) = 43.7\%$$

It is obvious that both the mean and standard deviation of account balances at the suburban branch are smaller than at the main office. However, as a percentage of their means, the standard deviations of the two offices are about equal. In this sense, the variability of account balances at the two locations is relatively the same.

SITUATION 4.2
Resolved

The 50 sample acres at the Rockford Grain Growers Company are planted with peas and harvested at the end of the season. Following are the results of the harvest:

$$\overline{X} = 1{,}438 \text{ pounds per acre}$$
$$s = 85.7 \text{ pounds per acre}$$
$$CV = 6\%$$

Management can now evaluate the desirability of planting several hundred acres of peas next year. It is important to note the average yield per acre for this purpose, but also of great interest is the small standard deviation of yield. This gives the Rockford Grain Growers Company confidence that the actual yield will not vary substantially from the forecasted amount. The coefficient of variation is only 6%, indicating a small standard deviation relative to the mean.

EXERCISES

31. What is the difference between a measure of central tendency and a measure of variability?
32. What is the difference between standard deviation and variance?
33. What is the difference between absolute variation and relative variation?
34. Why are the deviations from the mean *squared* in the computation of standard deviation?
35. In the formula for sample standard deviation, why is the sum of squared deviations divided by $(n - 1)$ instead of by n?
36. Why is the concept of degrees of freedom important?
37. Supposing you live in an area where the standard deviation for the average amount of rainfall is zero. Describe your climate.
38. Jim Perez, customer service manager for the National Tune-up Corporation, recently collected the following data, which represent the number of complaints received by his department on each of eight randomly selected days.

    10, 12, 8, 5, 11, 10, 9, 14

    a. Compute the variance.
    b. Compute the standard deviation.
    c. Compute a standard score for the day when 10 complaints were received.
    d. Compute the coefficient of variation.
    e. Interpret each of these measures of variability.

t balances
ever, as a
are about
locations

vith peas
vest:

red acres
purpose,
Rockford
stantially
g a small

of vari-

ard devi-

s divided

nount of

recently
d by his

39. Phoenix Body & Frame employs eight workers. The following data collection shows the years of experience for each worker: 1, 7, 9, 15, 9, 1, 7, 15. Treat this data collection as a population.
    a. Compute the standard deviation.
    b. If a worker with 8 years of experience is added, how will this affect the standard deviation?
    c. If a worker with 15 years of experience replaces a worker with 7 years of experience, how will this affect the standard deviation?
    d. If a worker with no years of experience replaces a worker with 7 years of experience, how will this affect the standard deviation?

40. As a manufacturer of some product requiring great uniformity (interchangeability of parts), would you be interested in a product whose pertinent characteristics included a large or small standard deviation?

41. The ages of some of the people in an office are: 23, 25, 34, 35, 37, 41, 42, 56. Treat this data collection as a sample.
    a. Compute the variance.
    b. Compute the standard deviation.
    c. Compute a standard score for a person who is 23 years old.
    d. Compute the coefficient of variation.

42. Last year at this time, personal loan data at Farmers and Merchants Bank showed a mean of $650 and a standard deviation of $300. Recently the mean was calculated to be $1,000 and the standard deviation $350. Did loans last year show more or less relative variation than recent loans?

43. Michelle Simmons wants to determine the variability of the amounts of checks she writes during a typical month. The following data collection represents 11 checks randomly selected from last month's personal checking account: $8.63, $102.36, $45.00, $50.12, $75.65, $9.87, $224.56, $78.95, $78.98, $15.62, $20.00. What is the average amount per check? What is the typical amount of variability per check?

44. The annual salaries of the seven workers employed by Lake City Transmission are: $15,000, $22,000, $25,000, $17,500, $14,500, $32,500, $13,250. A competitor, the Transmission Exchange Company, has workers who earn a mean annual salary of $21,000 with a standard deviation of $3,000. Compare the means and relative variability of the two companies.

45. Dick Hoover, owner of Modern Office Equipment, is concerned about freight costs and clerical costs incurred on small orders. In an effort to reduce expenditures in this area, he has decided to introduce a discount policy for orders over $40 in the hope that this will cause customers to consolidate a number of small orders into large orders. The following data show the mean amount per transaction for a sample of 28 customers:

    10, 15, 20, 25, 15, 17, 41, 50, 5, 9, 12, 14, 35, 18,
    19, 17, 28, 29, 11, 11, 43, 54, 7, 8, 16, 13, 37, 18

    a. Compute the variance.
    b. Compute the standard deviation.
    c. Compute the coefficient of variation.
    d. Is the distribution symmetrical, positively skewed, or negatively skewed?
    e. If the policy is successful, will the mean of the distribution increase, decrease, or remain the same?
    f. If the policy is successful, will the standard deviation of the distribution increase, decrease, or remain the same?

46. Describe what the standard deviation measures.

SUMMARY

This chapter has presented various quantitative measurements designed to summarize the essential characteristics of data collections. The two key elements of a data collection were introduced: central tendency and variability. Methods were described to measure and summarize each of these important components.

The mean, median, mode, and weighted mean were presented as methods of indicating a central summary value for a data collection. The variance, standard deviation, and coefficient of variation were presented as methods to indicate the degree of variability. It was stressed that the analyst must carefully choose among these methods in computing a summary statistic so that it fairly and accurately summarizes the underlying data collection.

APPLICATIONS OF STATISTICAL CONCEPTS IN THE BUSINESS WORLD

Few, if any, quantitative techniques are more widely used throughout business than those presented in this chapter. Quantitative data are used everywhere in modern organizations, and methods of summarizing their essential features are universally employed.

The two essential components of any data collection, measures of central tendency and variability, are commonly used as summary values. Accountants, for example, constantly deal with numerical values and employ summary figures to describe the current condition of various accounts and to make comparisons over time.

Financial data abound in the business world and are routinely summarized so that their essential properties can be analyzed and used by decision makers. Production or manufacturing functions in businesses measure such things as output, temperature, and pressure, and these data values must be briefly described using summary statistics if they are to be useful to managers.

Personnel records contain a large number of quantitative variables that must be summarized for use in intelligent decision making. Overtime hours per week, years of education, months with the company, and hourly rate of pay are examples of such variables.

Although several measurements made in marketing result in categorical data, many quantitative measurements are used as well. Examples are units sold per day, dollar volume per day for each department in a retail store, number of credit card uses per month by customers, and number of units sold for each type and style of product.

A key aspect of any business in this highly competitive age is quality control. Summary measures such as those discussed in this chapter are commonly used to measure such indications of quality as defective units per batch, variability of a manufacturing process, and sizes and weights of parts.

GLOSSARY

Arithmetic mean The sum of the numerical values in a collection divided by the number of values.

Parameter Any measurable characteristic of a population.

Statistic Any measurable characteristic of a sample.

Median The middle item in a set of observations that are arranged in order of magnitude.

Mode The value that occurs most frequently in a data collection.

Symmetrical distribution A curve that can be divided by a vertical line into two parts that are mirror images.

Skewed distribution A curve that lacks symmetry.

Weighted mean A mean computation in which some of the data values are given more weight than others.

Variance The average of the squared differences between the data values and the mean.

Standard deviation A measure of the amount by which the values in a data collection differ from the mean.

Standard score A measure of the number of standard deviations by which a particular value lies above or below the mean.

Degrees of freedom The number of data elements that are free to vary.

Coefficient of variation The standard deviation expressed as a percentage of the mean.

KEY FORMULAS

Arithmetic mean: population

$$\mu = \frac{\Sigma X}{N} \tag{4.1}$$

Arithmetic mean: sample

$$\overline{X} = \frac{\Sigma X}{n} \tag{4.2}$$

Median

$$\text{Median item number} = \frac{n + 1}{2} \tag{4.3}$$

Weighted mean

$$\overline{X}_w = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} \tag{4.4}$$

Variance: population

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N} \tag{4.5}$$

Standard deviation: population

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}} \tag{4.6}$$

Standard score

$$\text{Standard score} = \frac{X - \mu}{\sigma} \tag{4.7}$$

Variance: sample

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1} \tag{4.8}$$

Standard deviation: sample

$$s = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n - 1}} \tag{4.9}$$

Variance: sample (short-cut formula)

$$s^2 = \frac{\Sigma X^2 - \dfrac{(\Sigma X)^2}{n}}{n - 1} \tag{4.10}$$

Coefficient of variation

$$CV = \frac{s}{\overline{X}}(100) \tag{4.11}$$

SOLVED EXERCISES

1. CHOOSING AMONG THE MEAN, MEDIAN, MODE, AND WEIGHTED MEAN

Which measure of central tendency should be used in each of the following situations?

a. You want to determine the average annual percentage rate of net profit to sales for the General Electric Company for the last seven years.

b. You want to determine the average amount each worker receives per month to ensure a fair distribution in a profit-sharing plan.

c. You want to determine a representative wage value for use in later arbitration for Precision Landscape Systems. The company employs 200 workers, including several highly paid specialists.

Solution:

a. The weighted mean should be used to weight the profit rates by the sales values for each year.

b. The mean should be used to divide the total profits to be shared by the workers.

c. The median should be used so that the highly paid specialists' salaries do not distort the representative wage value.

2. COMPUTATION OF MEAN, MEDIAN, AND MODE

June Shapiro is thinking about starting an advertising agency and is interested in analyzing firms that operate in the Spokane area. The following data collection shows a sample of the local agencies, 1986 total billings, and the number of full-time employees:

Agency	Billings ($ millions)	Employees
Wendt Advertising	7.4	29
Clark, White & Associates	5.0	31
Coons, Corker & Associates	3.5	10
Elgee Corporation	2.1	3
Pierce-Stuart & Associates	1.5	8
Robideaux & Associates	1.2	15
Pacific Advertising	1.0	4
Bright Ideas, Inc.	1.0	5
Creative Consultants	.4	1
Degerness & Associates	.3	3
Rasor & Associates	.2	4

Source: "Advertising Marketing & Public Relations," *Journal of Business*, Sept 17, 1987.

a. Compute the mean, median, and mode for 1986 billings per agency.

b. Which measure of central tendency should June use if she wants the typical amount of 1988 billings?

c. Compute the mean, median, and mode for the number of employees per agency.

d. Which measure of central tendency should June use if she wants to measure variability as well?

Solution:

X	X^2	Y	Y^2
7.4	54.76	29	841
5.0	25.00	31	961
3.5	12.25	10	100
2.1	4.41	3	9
1.5	2.25	8	64
1.2	1.44	15	225
1.0	1.00	4	16
1.0	1.00	5	25
.4	.16	1	1
.3	.09	3	9
.2	.04	4	16
23.6	102.40	113	2,267

a. Mean $= \overline{X} = \dfrac{\Sigma X}{n} = \dfrac{23.6}{11} = 2.1455$

$$\text{Median item number} = \frac{n+1}{2} = \frac{11+1}{2} = \frac{12}{2} = 6$$

Median = 1.2
Mode = 1.0

b. June should use the median, 1.2, if she wants the typical amount. This value is not unduly influenced by the billings of $7.4 million for Wendt Advertising.

c. Mean $= \overline{X} = \dfrac{\Sigma X}{n} = \dfrac{113}{11} = 10.273$

$$\text{Median item number} = \frac{n+1}{2} = \frac{11+1}{2} = \frac{12}{2} = 6$$

Median = 5
Mode = 3, 4

d. Since June wants to perform further computations, such as finding the standard deviation, she should use the mean, 10.273.

3. COMPUTATION OF VARIANCE AND STANDARD DEVIATION

Refer to the data collected by June Shapiro in Solved Exercise 2.

a. Compute the variance and standard deviation for billings per agency.

b. Compute the variance and standard deviation for number of employees per agency.

Solution:

a. Variance:

$$s^2 = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n-1} = \frac{102.4 - \frac{(23.6)^2}{11}}{11-1} = \frac{51.77}{10} = 5.18$$

Standard deviation:

$$s = 2.276$$

b. Variance:

$$s^2 = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n-1} = \frac{2{,}267 - \frac{(113)^2}{11}}{11-1} = \frac{1{,}106}{10} = 110.6$$

Standard deviation:

$$s = 10.52$$

4. COMPUTATION OF THE COEFFICIENT OF VARIATION

Refer to the data collected by June Shapiro in Solved Exercise 2.

a. Compute the coefficient of variation for billings per agency.

b. Compute the coefficient of variation for number of employees per agency.

Solution:

a. $\text{CV} = \frac{s}{\overline{X}}(100) = \frac{2.276}{2.1455}(100) = 106\%$

b. $\text{CV} = \frac{s}{\overline{X}}(100) = \frac{10.52}{10.273}(100) = 102.4\%$

EXERCISES

47. Which measure of central tendency is a good choice for an average if further computations are needed?

48. For a certain operation, a factory supervisor is told to set a standard time that "differs as little as possible from the time now taken by all the employees in the shop." The supervisor should use which measure of central tendency?

49. The Adams Tractor Company employs six workers. The following data collection shows the age of each worker: 21, 27, 19, 35, 31, 29. Treat this data collection as a population.
    a. Compute the mean and median.
    b. Compute the standard deviation.
    c. If a new worker of age 27 is added, how will this affect the standard deviation?
    d. (This question refers to the original six workers.) If a worker of age 20 replaces the worker who is 35, how will this affect the mean and standard deviation?
    e. (This question refers to the original six workers.) If a worker of age 38 replaces the worker who is 27, how will this affect the standard deviation?

50. The largest 12 U.S. daily newspapers have daily circulations as listed below (as of March 31, 1988). Treat this data collection as a population. Compute the mean and median.

Kourica 437-779

Newspapers	Circulation
The Wall Street Journal	2,025,176
USA Today	1,345,271
New York Daily News	1,283,302
Los Angeles Times	1,132,920
New York Times	1,078,443
Washington Post	810,011
Chicago Tribune	774,045
Detroit News	688,218
Newsday	665,218
Detroit Free Press	647,763
Chicago Sun Times	625,035
San Francisco Chronicle	569,185

Source: Audit Bureau of Circulation.

51. The following data collection shows the current number of enrollees, the number of companies or groups served, and the number of primary physicians for the top Spokane-area health care plans. Treat this data collection as a population.

Organization	Enrollees	Companies	Physicians
HMOs:			
HMO Washington	161	14	46
Foundation Health Plan	3,569	139	145
Group Health Northwest	30,000	350	26
HealthPlus	10,500	125	47
Maxicare Washington	173	11	12
PPOs:			
United Northwest Services	50,000	10	170
Inland Health Associates	29,000	9	86
Medical Services Corp.	17,373	580	541
First Choice Health Plan	3,836	291	43
Blue Cross Prudent Buyer	3,128	33	114

Source: "Hospitals Health Care & Insurance," *Journal of Business*, Nov. 25–Dec. 9, 1987.

a. Compute the mean and the standard deviation of the number of enrollees in the 10 health plans.

b. Compute the mean and the standard deviation of the number of companies or groups served by the 10 health plans.

c. Compute the mean and the standard deviation of the number of primary physicians used by the 10 health plans.

d. Compute the median number of enrollees in the 10 health plans.

e. Compute the median number of companies or groups served by the 10 health plans.

f. Compute the median number of primary physicians used by the 10 health plans.

g. Is there any application for the weighted mean in this data collection?

h. Which average would you choose (mean or median) for each data collection if you wanted to describe the typical health care plan?

i. Compute the coefficient of variation for the five HMOs and the five preferred provider organizations (PPOs) for each of the three data collections. Compare the relative variability of the two types of health care plans.

52. The Hord Automobile Manufacturing Company is considering two brands of batteries for their latest model. The Telco battery has a mean lifetime of 55 months with a standard deviation of 5 months. The Long-Life battery has a mean lifetime of 45 months with a standard deviation of 3 months.

a. If the decision criterion for selecting a brand of battery is maximum lifetime, which brand should be selected?

b. Which brand should be selected if consistency of service is the decision criterion?

53. The following data represent net income as a percentage of sales (rounded to the nearest full percent) during 1988 for a random sample of 70 of the 500 largest industrial corporations:

5	6	8	10	4	9	7
9	6	4	9	10	9	8
3	9	5	9	9	8	7
10	2	7	4	8	5	10
9	6	8	8	8	7	8
6	11	9	11	7	7	11
10	8	8	5	9	8	8
8	9	10	7	7	7	5
8	7	9	9	8	6	9
5	8	8	7	9	13	8

Compute summary measures for this distribution. Write a memo to management comparing net income for your company, which is 9, to this distribution.

54. Before making a decision on purchasing stock in the Electronic Research & Development Corporation, the management of Fidelity Investments, a mutual fund, wants information concerning the price movements of the firm's stock during the past year. Thirty-five days of the past year were randomly selected, and the closing price (to the nearest dollar) was recorded for each day:

43	29	42	35	32	28	22
52	34	35	32	28	50	33
34	37	29	30	28	29	24
39	27	40	43	48	33	48
29	28	39	36	49	26	47

a. Fidelity has decided not to purchase the stock unless the mean closing price for last year is $34 or more. Is further analysis of these data necessary?

b. Fidelity is also very interested in the variability of this stock. They will not invest in this stock if the usual variation from the mean price is more than $10. Would they be interested?

c. Finally, Fidelity needs to compare the Electronic Research & Development Corporation stock to that of Innovative Technology Systems. Fidelity is satisfied with both stocks and will purchase the one with less relative variability in price. If the Innovative Technology Systems stock has a mean price of $61 and a standard deviation of $12, which stock should Fidelity purchase?

55. The accompanying table analyzes the demand for both covered and vacant boat slips on Lake Pend Oreille in Idaho.

	Covered		Open		
Marina	Occupied	Vacant	Occupied	Vacant	Total
The Captn's Table	0	0	22	3	25
Lee Peters Moorage	44	0	19	2	65
Sunset Resort	0	0	18	10	28
Bottle Bay Resort	0	0	20	2	22
Sandpoint Marina	26	0	109	0	135
Windbag Marina	0	0	60	0	60
Holiday Shores	0	0	33	2	35

(continued)

	Covered		Open		
Marina	Occupied	Vacant	Occupied	Vacant	Total
Ellisport Marina	0	0	75	0	75
Pend Oreille Shores	0	0	33	32	65
Unknown name	0	0	63	7	70
Scenic Bay Marina	58	0	116	3	177
Vista Bay Resort	30	0	36	4	70
Bitter End Marina	0	0	103	2	105
Bayview Marina	65	0	13	2	80
Boileau's	119	1	20	4	144
McDonald's Hudson Bay Resort	90	0	90	0	180
Totals	432	1	830	73	1,336

Source: *Development Plan for Harbor View Marina in Garfield Bay on Lake Pend Oreille*, J & H Research Service, November 1987.

Metropolitan Mortgage & Securities Company has repossessed a vacant marina and is trying to decide how many slips to develop. Write a memo to Metropolitan summarizing the number of covered, vacant, and total boat slips presently on the lake.

56. This exercise refers to the company data base in Appendix C. Select a sample of 10 workers.

a. Compute the mean and standard deviation for each of the following variables:

X_1 = Number of years with the company
X_2 = Number of overtime hours worked during the last six months
X_4 = Number of continuing education courses completed
X_5 = Number of sick days taken during the last six months
X_6 = Score on company aptitude test
X_8 = Annual base salary
X_9 = Employee age

b. Indicate whether each of the variables is symmetrical, positively skewed, or negatively skewed.

EXTENDED EXERCISES

57. WORDAN WINE BOTTLING

Mike Wordan, a grape grower, is considering buying a wine-bottling plant and producing a line of quality table wines. Mike would like to know the capacity of the bottling plant and works out an arrangement with Rick Roig, the current owner, to sample a number of production days. The number of bottles produced per day by the current process will be recorded, and these data will help Mike decide if a purchase is feasible.

The test days are randomly selected during the busy bottling season, and the measurements are made. For each of the 25 days selected for the sample, the production line is observed, and the number of bottles produced is recorded. Since there is a backlog of wine waiting to be bottled, Mike Wordan believes that the output is a function of the capabilities of the process, not of raw material availability.

After the sample values are recorded, the following statistics are calculated:

$$n = 25 \text{ days} \qquad \bar{X} = 584 \text{ bottles} \qquad s = 253 \text{ bottles}$$

Mike has not yet determined an appropriate price for the bottling facility and has not yet talked with bankers about financing. But Mike thinks the sample statistics will help him make a decision about the feasibility of the purchase.

1. What can Mike conclude after considering the statistics that emerged from the sampled days?
2. Was the sample size sufficiently large to make the statistics useful?
3. Are any problem areas revealed by the sample?

58. PIERONE'S CLOTHING COMPANY

Pierone's, a men's clothing company, sends salespeople to small retail establishments around the country. Every two years the company buys a large number of cars to be used by the salespeople in their travels. Bob Pierone, the owner, has just read an article in *Fortune* (September 25, 1989) indicating that automakers are trying to develop cars that will run on methanol. Such a car would be much cleaner and would accelerate faster. Bob is very interested in this concept. but meanwhile he must replace his current fleet.

The company has narrowed the choice of car to two models, and since the price and estimated upkeep costs for each are about the same, Bob is interested in determining the miles per gallon of each car. If one car has substantially better mileage than the other, the choice will be made in its favor.

Pierone's arranges to test-drive a number of cars of each model for a period of one week. Each car will be driven about 1,000 miles to produce a fair estimate of mileage. Following are the sample statistics that result from this test:

Model 1	Model 2
$\overline{X}$ = 19.4 mpg	$\overline{X}$ = 20.1 mpg
s = 1.7 mpg	s = 5.3 mpg
n = 12 cars	n = 15 cars

The company plans to use these data in making a decision about which model to buy for their fleet. They will be purchasing between 500 and 600 cars.

1. What do you think about the sample sizes used for this test?
2. The standard deviation for model 2 is much larger than that for model 1. In view of the number of cars to be purchased, is this a problem?
3. Overall, what direction do the sample results give Pierone's Clothing Company?

59. SECOND AVENUE CAR STEREOS

Second Avenue Radio, a company that produces car stereos, is concerned about the number of units produced during the past several weeks. Owner Greg Dempsey decides that a random sample of observation times will be selected, and the number of units produced during each selected hour will be recorded. Following are the numbers recorded for the sample:

5	6	8	4	5	9	7	5	8	4	9	8	7	6	8
4	6	5	7	5	7	6	8	9	3	4	8	2	9	5
5	1	5	6	8	6	6	9	5	9	0	5	7	8	6

1. What useful statistics can be calculated from the sample data?
2. Summarize the results of the sampling effort in a memo to Dempsey.

60. OUR LADY OF LOURDES HOSPITAL

One of the variables recorded for Our Lady of Lourdes Hospital employees is family size. Family sizes for the population of 200 employees are presented in the accompanying table. The customer benefits director, Curtis Huff, has asked you to quickly estimate the average family size per employee.

[illegible] 437 – [illegible]

1. Select a simple random sample of 30 family sizes and compute the mean.
2. You have decided to supply Curtis with additional information, so also compute the sample standard deviation. (Save these answers, since they will be used in future exercises.)

(1) 3
(2) 2
(3) 7
(4) 3
(5) 4
(6) 2
(7) 3
(8) 1
(9) 5
(10) 3
(11) 2
(12) 3
(13) 4
(14) 1
(15) 2
(16) 2
(17) 4
(18) 4
(19) 3
(20) 2
(21) 1
(22) 5
(23) 2
(24) 1
(25) 4
(26) 3
(27) 2
(28) 3
(29) 6
(30) 1
(31) 2
(32) 4
(33) 3
(34) 2
(35) 1
(36) 2
(37) 4
(38) 1
(39) 4
(40) 2
(41) 1
(42) 3
(43) 5
(44) 2
(45) 1
(46) 4
(47) 3
(48) 5
(49) 2
(50) 4
(51) 1
(52) 6
(53) 2
(54) 5
(55) 4
(56) 1
(57) 2
(58) 1
(59) 5
(60) 2
(61) 7
(62) 1
(63) 2
(64) 6
(65) 4
(66) 1
(67) 2
(68) 1
(69) 2
(70) 4
(71) 3
(72) 7
(73) 2
(74) 6
(75) 2
(76) 7
(77) 3
(78) 6
(79) 4
(80) 2
(81) 3
(82) 5
(83) 2
(84) 1
(85) 3
(86) 3
(87) 2
(88) 4
(89) 1
(90) 2
(91) 3
(92) 3
(93) 2
(94) 4
(95) 1
(96) 2
(97) 4
(98) 3
(99) 2
(100) 6
(101) 4
(102) 1
(103) 2
(104) 5
(105) 3
(106) 2
(107) 1
(108) 2
(109) 2
(110) 1
(111) 4
(112) 1
(113) 1
(114) 2
(115) 2
(116) 1
(117) 4
(118) 2
(119) 1
(120) 3
(121) 5
(122) 1
(123) 2
(124) 3
(125) 4
(126) 3
(127) 2
(128) 1
(129) 6
(130) 1
(131) 2
(132) 5
(133) 2
(134) 1
(135) 5
(136) 2
(137) 1
(138) 4
(139) 2
(140) 4
(141) 1
(142) 2
(143) 4
(144) 1
(145) 2
(146) 2
(147) 5
(148) 3
(149) 1
(150) 2
(151) 6
(152) 2
(153) 5
(154) 1
(155) 2
(156) 1
(157) 4
(158) 2
(159) 2
(160) 7
(161) 4
(162) 2
(163) 1
(164) 7
(165) 2
(166) 7
(167) 4
(168) 6
(169) 3
(170) 2
(171) 3
(172) 4
(173) 2
(174) 2
(175) 1
(176) 5
(177) 3
(178) 2
(179) 4
(180) 3
(181) 5
(182) 3
(183) 1
(184) 2
(185) 4
(186) 3
(187) 2
(188) 5
(189) 3
(190) 4
(191) 3
(192) 2
(193) 3
(194) 2
(195) 5
(196) 3
(197) 3
(198) 2
(199) 5
(200) 1

MICRO COMPUTER PACKAGE

You can use the micro package *Computerized Business Statistics* to compute measures of central tendency and variability.

In the exercise involving the Electronic Research & Development Corporation (Exercise 54), you were asked to analyze the price movements of the firm's stock for 35 days. The computation of both the mean and standard deviation of this data collection was necessary.

Computer Solution:

On the main menu of *Computerized Business Statistics*, a **3** is selected, which indicates Descriptive Statistics. The descriptive statistics menu will appear on the screen.

Since the data for this problem need to be entered on the keyboard, 1 is selected. The following message and question appear on the screen:

```
Descriptive Statistics-Define New Problem
Raw or Group Data: Enter R/G, press ↵ R
```

Since the data are in raw form and have not been grouped, **R** is selected. The next question asked by the CBS program is:

```
Population or Sample Data: Enter P/S, press ↵ S
```

Since the data constitute a sample, choose **S**. Next, the program asks:

```
Number of Data Points: Enter 1 - 125, press ↵ 35
```

Since the data collection consists of 35 closing stock prices, **35** is the correct entry. Next, the program instructs you to:

```
Variable Name Enter 0-5 Char. Press ↵ Price
```

The variable name used in this problem is **Price.** Next, the program asks:

```
Problem Definition Correct? Enter Y/N/Q, press ↵ Y
```

If the problem has been set up correctly, the answer is **Y.**

Next, the program provides spaces for the raw data to be entered. The data values are numbered from 1 to 35. The cursor allows you to replace the 0.0's on the screen with the actual data values.

```
Table Commands        Enter Raw Data            File: None

                   Price
1.                  43
2.                  29
3.                  42
4.                  35
5.                  32
.                   .
.                   .
.                   .
35.                 47

Press F when Finished
```

After the data are entered and **F** pressed, you are asked:

```
Save data? Enter Y/N & press ↵ N
```

If you want to save these data in a disk file, answer **Y**; otherwise enter **N.**

The program options menu then reappears, and you are instructed to:

```
Enter number (1 - 9) for your selection from the menu & press ↵ 7
```

You are now ready to run the problem, so enter 7. You are now asked:

```
Convert raw data to group data? Y/N & press ↵ N
```

If you want to convert these raw data to group data, answer **Y**; otherwise answer **N.** Next, you are given output selections. Since a hard copy of the results is needed, **P,** or printer, is selected.