CHAPTER 4

Descriptive Statistical Measures

Jonny Drake/Shutterstock.com

Learning Objectives

After studying this chapter, you will be able to:

- Explain the difference between a population and a sample.
- Understand statistical notation.
- List different measures of location.
- Compute the mean, median, mode, and midrange of a set of data.
- Use measures of location to make practical business decisions.
- List different measures of dispersion.
- Compute the range, interquartile range, variance, and standard deviation of a set of data.
- Explain Chebyshev's theorem.
- State the Empirical Rules and apply them to practical data.
- Compute a standardized value (z-score) for observations in a data set.
- Define and compute the coefficient of variation.
- Explain the nature of skewness and kurtosis in a distribution.
- Interpret the coefficients of skewness and kurtosis.
- Use the Excel *Descriptive Statistics* tool to summarize data.
- Calculate the mean, variance, and standard deviation for grouped data.
- Calculate a proportion.
- Use PivotTables to compute the mean, variance, and standard deviation of summarized data.
- Explain the importance of understanding relationships between two variables. Explain the difference between covariance and correlation.
- Calculate measures of covariance and correlation.
- Use the Excel *Correlation* tool.
- Identify outliers in data.
- State the principles of statistical thinking.
- Interpret variation in data from a logical and practical perspective.
- Explain the nature of variation in sample data.

As we noted in Chapter 3, frequency distributions, histograms, and cross-tabulations are tabular and visual tools of descriptive statistics. In this chapter, we introduce numerical measures that provide an effective and efficient way of obtaining meaningful information from data. Before discussing these measures, however, we need to understand the differences between populations and samples.

Populations and Samples

A **population** consists of all items of interest for a particular decision or investigation—for example, *all* individuals in the United States who do not own cell phones, *all* subscribers to Netflix, or *all* stockholders of Google. A company like Netflix keeps extensive records on its customers, making it easy to retrieve data about the entire population of customers. However, it would probably be impossible to identify all individuals who do not own cell phones.

A **sample** is a subset of a population. For example, a list of individuals who rented a comedy from Netflix in the past year would be a sample from the population of all customers. Whether this sample is representative of the population of customers—which depends on how the sample data are intended to be used—may be debatable; nevertheless, it is a sample. Most populations, even if they are finite, are generally too large to deal with effectively or practically. For instance, it would be impractical as well as too expensive to survey the entire population of TV viewers in the United States. Sampling is also clearly necessary when data must be obtained from destructive testing or from a continuous production process. Thus, the purpose of sampling is to obtain sufficient information to draw a valid inference about a population. Market researchers, for example, use sampling to gauge consumer perceptions on new or existing goods and services; auditors use sampling to verify the accuracy of financial statements; and quality control analysts sample production output to verify quality levels and identify opportunities for improvement.

Most data with which businesses deal are samples. For instance, the *Purchase Orders* and *Sales Transactions* databases that we used in previous chapters represent samples because the purchase order data include only orders placed within a three-month time period, and the sales transactions represent orders placed on only one day, July 14. Therefore, unless noted otherwise, we will assume that any data set is a sample.

Understanding Statistical Notation

We typically label the elements of a data set using subscripted variables, x_1, x_2, … , and so on. In general, x_i represents the ith observation. It is a common practice in statistics to use Greek letters, such as μ (mu), σ (sigma), and π (pi), to represent population measures and italic letters such as by $\bar{x}$ (x-bar), s, and p to represent sample statistics. We will use N to represent the number of items in a population and n to represent the number of observations in a sample. Statistical formulas often contain a summation operator, Σ (Greek capital sigma), which means that the terms that follow it are added together. Thus, $\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$. Understanding these conventions and mathematical notation will help you to interpret and apply statistical formulas.

Measures of Location

Measures of location provide estimates of a single value that in some fashion represents the "centering" of a set of data. The most common is the *average*. We all use averages routinely in our lives, for example, to measure student accomplishment in college (e.g., grade point average), to measure the performance of sports teams (e.g., batting average), and to measure performance in business (e.g., average delivery time).

Arithmetic Mean

The average is formally called the **arithmetic mean** (or simply the **mean**), which is the sum of the observations divided by the number of observations. Mathematically, the mean of a population is denoted by the Greek letter μ, and the mean of a sample is denoted by $\bar{x}$. If a population consists of N observations $x_1, x_2, \ldots, x_N$, the population mean, μ, is calculated as

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{4.1}$$

The mean of a sample of n observations, $x_1, x_2, \ldots, x_n$, denoted by $\bar{x}$, is calculated as

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4.2}$$

Note that the calculations for the mean are the same whether we are dealing with a population or a sample; only the notation differs. We may also calculate the mean in Excel using the function AVERAGE(*data range*).

One property of the mean is that the sum of the deviations of each observation from the mean is zero:

$$\sum_i (x_i - \bar{x}) = 0 \tag{4.3}$$

This simply means that the sum of the deviations above the mean are the same as the sum of the deviations below the mean; essentially, the mean "balances" the values on either side of it. However, it does not suggest that half the data lie above or below the mean—a common misconception among those who don't understand statistics.

In addition, the mean is unique for every set of data and is meaningful for both interval and ratio data. However, it can be affected by **outliers**—observations that are radically different from the rest—which pull the value of the mean toward these values. We discuss more about outliers later in this chapter.

EXAMPLE 4.1 Computing the Mean Cost per Order

In the *Purchase Orders* database, suppose that we are interested in finding the mean cost per order. Figure 4.1 shows a portion of the data file. We calculate the mean cost per order by summing the values in column G and then dividing by the number of observations. Using formula (4.2), note that x_1 = \$2,700, x_2 = \$19,250, and so on, and n = 94. The sum of these order costs is \$2,471,760. Therefore, the mean cost per order is \$2,471,760/94 = \$26,295.32. We show these calculations in a separate worksheet, *Mean* in the *Purchase Orders* Excel workbook. A portion of this worksheet in split-screen mode is shown in Figure 4.2. Alternatively, we used the Excel function =AVERAGE (B2:B95) in this worksheet to arrive at the same value. We encourage you to study the calculations and formulas used.

	A	B	C	D	E	F	G	H	I	J
1	Purchase Orders									
2										
3	Supplier	Order No.	Item No.	Item Description	Item Cost	Quantity	Cost per order	A/P Terms (Months)	Order Date	Arrival Date
4	Hulkey Fasteners	Aug11001	1122	Airframe fasteners	$ 4.25	19,500	$ 82,875.00	30	08/05/11	08/13/11
5	Alum Sheeting	Aug11002	1243	Airframe fasteners	$ 4.25	10,000	$ 42,500.00	30	08/08/11	08/14/11
6	Fast-Tie Aerospace	Aug11003	5462	Shielded Cable/ft.	$ 1.05	23,000	$ 24,150.00	30	08/10/11	08/15/11
7	Fast-Tie Aerospace	Aug11004	5462	Shielded Cable/ft.	$ 1.05	21,500	$ 22,575.00	30	08/15/11	08/22/11
8	Steelpin Inc.	Aug11005	5319	Shielded Cable/ft.	$ 1.10	17,500	$ 19,250.00	30	08/20/11	08/31/11
9	Fast-Tie Aerospace	Aug11006	5462	Shielded Cable/ft.	$ 1.05	22,500	$ 23,625.00	30	08/20/11	08/26/11
10	Steelpin Inc.	Aug11007	4312	Bolt-nut package	$ 3.75	4,250	$ 15,937.50	30	08/25/11	09/01/11

Figure 4.1

Portion of *Purchase Orders* Database

Figure 4.2

Excel Calculations of Mean Cost per Order

	A	B
1	Observation	Cost per order
2	x1	$2,700.00
3	x2	$19,250.00
4	x3	$15,937.50
5	x4	$18,150.00
93	x92	$74,375.00
94	x93	$72,250.00
95	x94	$6,562.50
96	Sum of cost/order	$2,471,760.00
97	Number of observations	94
98		
99	Mean cost/order	$26,295.32
100		
101	Excel AVERAGE function	$26,295.32

Median

The measure of location that specifies the middle value when the data are arranged from least to greatest is the **median**. Half the data are below the median, and half the data are above it. For an odd number of observations, the median is the middle of the sorted numbers. For an even number of observations, the median is the mean of the two middle numbers. We could use the *Sort* option in Excel to rank-order the data and then determine the median. The Excel function MEDIAN(*data range*) could also be used. The median is meaningful for ratio, interval, and ordinal data. As opposed to the mean, the median is *not* affected by outliers.

EXAMPLE 4.2 Finding the Median Cost per Order

In the *Purchase Orders* database, sort the data in Column G from smallest to largest. Since we have 94 observations, the median is the average of the 47th and 48th observations. You should verify that the 47th sorted observation is $15,562.50 and the 48th observation is $15,750. Taking the average of these two values results in the median value of ($15,562.5 + $15,750)/2 = $15,656.25. Thus, we may conclude that the total cost of half the orders were less than $15,656.25 and half were above this amount. In this case, the median is not very close in value to the mean. These calculations are shown in the worksheet *Median* in the *Purchase Orders* Excel workbook, as shown in Figure 4.3.

Figure 4.3

Excel Calculations for Median Cost per Order

	A	B	C	D
1	Rank	Cost per order		
2	1	$68.75		
3	2	$82.50		
4	3	$375.00		
5	4	$467.50		
45	44	$14,910.00		
46	45	$14,910.00		
47	46	$15,087.50		
48	47	$15,562.50		$15,562.50
49	48	$15,750.00		$15,750.00
50	49	$15,937.50	Average	$15,656.25
51	50	$16,276.75		
52	51	$16,330.00		

Mode

A third measure of location is the **mode**. The mode is the observation that occurs most frequently. The mode is most useful for data sets that contain a relatively small number of unique values. For data sets that have few repeating values, the mode does not provide much practical value. You can easily identify the mode from a frequency distribution by identifying the value having the largest frequency or from a histogram by identifying the highest bar. You may also use the Excel function MODE.SNGL(*data range*). For frequency distributions and histograms of grouped data, the mode is the group with the greatest frequency.

EXAMPLE 4.3 Finding the Mode

In the *Purchase Orders* database, the frequency distribution and histogram for A/P Terms in Figure 3.40 in Chapter 3, we see that the greatest frequency corresponds to a value of 30 months; this is also the highest bar in the histogram. Therefore, the mode is 30 months. For the grouped frequency distribution and histogram of the Cost per order variable in Figure 3.42, we see that the mode corresponds to the group between $0 and $13,000.

Some data sets have multiple modes; to identify these, you can use the Excel function MODE.MULT(*data range*), which returns an array of modal values.

Midrange

A fourth measure of location that is used occasionally is the **midrange**. This is simply the average of the greatest and least values in the data set.

EXAMPLE 4.4 Computing the Midrange

We may identify the minimum and maximum values using the Excel functions MIN and MAX or sort the data and find them easily. For the Cost per order data, the minimum value is $68.78 and the maximum value is $127,500. Thus, the midrange is ($127,500 + $68.78)/2 = $63,784.39.

Caution must be exercised when using the midrange because extreme values easily distort the result, as this example illustrated. This is because the midrange uses only two pieces of data, whereas the mean uses *all* the data; thus, it is usually a much rougher estimate than the mean and is often used for only small sample sizes.

Using Measures of Location in Business Decisions

Because everyone is so familiar with the concept of the average in daily life, managers often use the mean inappropriately in business when other statistical information should be considered. The following hypothetical example, which was based on a real situation, illustrates this.

EXAMPLE 4.5 Quoting Computer Repair Times

The Excel file *Computer Repair Times* provides a sample of the times it took to repair and return 250 computers to customers who used the repair services of a national electronics retailer. Computers are shipped to a central facility, where they are repaired and then shipped back to the stores for customer pickup. The mean, median, and mode are all very close and show that the typical repair time is about 2 weeks (see Figure 4.4). So you might think that if a customer brought in a computer for repair, it would be reasonable to quote a repair time of 2 weeks. What would happen if the stores quoted all customers a time of 2 weeks? Clearly about half the customers would be upset because their computers would not be completed by this time.

Figure 4.5 shows a portion of the frequency distribution and histogram for these repair times (see the *Histogram* tab in the Excel file). We see that the longest repair time took almost 6 weeks. So, should the company give customers a guaranteed repair time of 6 weeks? They probably wouldn't have many customers because few would want to wait that long. Instead, the frequency distribution and histogram provide insight into making a more rational decision. You may verify that 90% of the time, repairs are completed within 21 days; on the rare occasions that it takes longer, it generally means that technicians had to order and wait for a part. So it would make sense to tell customers that they could probably expect their computers back within 2 to 3 weeks and inform them that it might take longer if a special part was needed.

From this example, we see that using frequency distributions, histograms, and percentiles can provide more useful information than simple measures of location. This leads us to introduce ways of quantifying variability in data, which we call *measures of dispersion*.

Figure 4.4

Measures of Location for *Computer Repair Times*

	A	B
1	Computer Repair Times	
2		
3	Sample	Repair Time (Days)
4	1	18
5	2	15
6	3	17
250	247	31
251	248	6
252	249	17
253	250	13
254		
255	Mean	14.912
256	Median	14
257	Mode	15

	A	B	C	D
1	Computer Repair Times			
2				
3			Relative	Cumulative
4	Days	Frequency	Frequency	Percentage
5	0	0	0.000	0.0%
6	1	0	0.000	0.0%
7	2	0	0.000	0.0%
8	3	0	0.000	0.0%
9	4	0	0.000	0.0%
10	5	1	0.004	0.4%
11	6	2	0.008	1.2%
12	7	5	0.020	3.2%
13	8	12	0.048	8.0%
14	9	14	0.056	13.6%
15	10	19	0.076	21.2%
16	11	19	0.076	28.8%
17	12	23	0.092	38.0%
18	13	22	0.088	46.8%

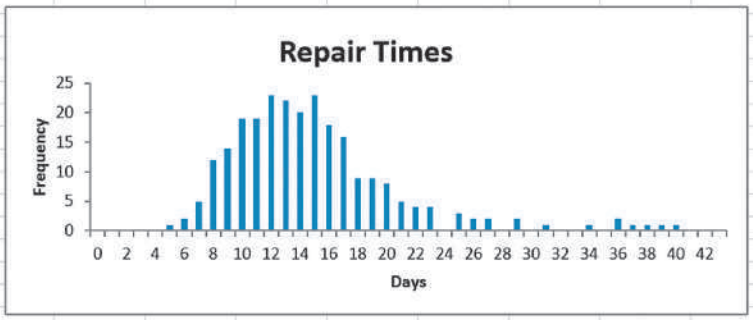


Figure 4.5

Frequency Distribution and Histogram for *Computer Repair Times*

Measures of Dispersion

Dispersion refers to the degree of variation in the data, that is, the numerical spread (or compactness) of the data. Several statistical measures characterize dispersion: the *range*, *variance*, and *standard deviation.*

Range

The **range** is the simplest and is the difference between the maximum value and the minimum value in the data set. Although Excel does not provide a function for the range, it can be computed easily by the formula = MAX(*data range*) − MIN(*data range*). Like the midrange, the range is affected by outliers and, thus, is often only used for very small data sets.

EXAMPLE 4.6 Computing the Range

For the Cost per order data in the *Purchase Orders* database, the minimum value is $68.78 and the maximum value is $127,500. Thus, the range is $127,500 − $68.78 = $127,431.22.

Interquartile Range

The difference between the first and third quartiles, $Q_3 - Q_1$, is often called the **interquartile range (IQR)**, or the **midspread**. This includes only the middle 50% of the data and, therefore, is not influenced by extreme values. Thus, it is sometimes used as an alternative measure of dispersion.

EXAMPLE 4.7 Computing the Interquartile Range

For the Cost per order data, we identified the first and third quartiles as Q_1 = \$6,757.81 and Q_3 = \$27,593.75 in Example 3.25. Thus, IQR = \$27,593.75 − \$6,757.81 = \$20,835.94. Therefore, the middle 50% of the data are concentrated over a relatively small range of \$20,835.94. Note that the upper 25% of the data span the range from \$27,593.75 to \$127,500, indicating that high costs per order are spread out over a large range of \$99,906.25.

Variance

A more commonly used measure of dispersion is the **variance**, whose computation depends on *all* the data. The larger the variance, the more the data are spread out from the mean and the more variability one can expect in the observations. The formula used for calculating the variance is different for populations and samples.

The formula for the variance of a population is

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{4.4}$$

where x_i is the value of the ith item, N is the number of items in the population, and μ is the population mean. Essentially, the variance is the average of the squared deviations of the observations from the mean.

A significant difference exists between the formulas for computing the variance of a population and that of a sample. The variance of a sample is calculated using the formula

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1} \tag{4.5}$$

where n is the number of items in the sample and $\bar{x}$ is the sample mean. It may seem peculiar to use a different denominator to "average" the squared deviations from the mean for populations and samples, but statisticians have shown that the formula for the sample variance provides a more accurate representation of the true population variance. We discuss this more formally in Chapter 6. For now, simply understand that the proper calculations of the population and sample variance use different denominators based on the number of observations in the data.

The Excel function VAR.S(*data range*) may be used to compute the sample variance, s^2, whereas the Excel function VAR.P(*data range*) is used to compute the variance of a population, σ^2.

EXAMPLE 4.8 Computing the Variance

Figure 4.6 shows a portion of the Excel worksheet *Variance* in the *Purchase Orders* workbook. To find the variance of the cost per order using formula (4.5), we first need to calculate the mean, as done in Example 4.1. Then for each observation, calculate the difference between the observation and the mean, as shown in column C. Next, square these differences, as shown in column D. Finally, add these square deviations (cell D96) and divide by $n - 1 = 93$. This results in the variance 890,594,573.82. Alternatively, the Excel function =VAR.S(B2:B95) yields the same result.

Figure 4.6

Excel Calculations for Variance of Cost per Order

	A	B	C	D
1	Observation	Cost per order	(xi - mean)	(xi - mean)^2
2	x1	$2,700.00	-$23,595.32	$556,739,085.74
3	x2	$19,250.00	-$7,045.32	$49,636,521.91
4	x3	$15,937.50	-$10,357.82	$107,284,417.52
5	x4	$18,150.00	-$8,145.32	$66,346,224.04
93	x92	$74,375.00	$48,079.68	$2,311,655,710.74
94	x93	$72,250.00	$45,954.68	$2,111,832,692.12
95	x94	$6,562.50	-$19,732.82	$389,384,151.56
96	Sum of cost/order	$2,471,760.00	Sum of squared deviations	$82,825,295,365.68
97	Number of observations	94		
98				
99	Mean cost/order	$26,295.32	Variance	890,594,573.82
100				
101			Excel VAR.S function	890,594,573.82

Note that the dimension of the variance is the square of the dimension of the observations. So for example, the variance of the cost per order is not expressed in dollars, but rather in dollars squared. This makes it difficult to use the variance in practical applications. However, a measure closely related to the variance that can be used in practical applications is the standard deviation.

Standard Deviation

The **standard deviation** is the square root of the variance. For a population, the standard deviation is computed as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \tag{4.6}$$

and for samples, it is

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}} \tag{4.7}$$

The Excel function STDEV.P(*data range*) calculates the standard deviation for a population (σ); the function STDEV.S(*data range*) calculates it for a sample (s).

EXAMPLE 4.9 Computing the Standard Deviation

We may use the same worksheet calculations as in Example 4.8. All we need to do is to take the square root of the computed variance to find the standard deviation. Thus, the standard deviation of the cost per order is $\sqrt{890{,}594{,}573.82} = \$29{,}842.8312$. Alternatively, we could use the Excel function =STDEV.S(B2:B95) to find the same value.

The standard deviation is generally easier to interpret than the variance because its units of measure are the same as the units of the data. Thus, it can be more easily related to the mean or other statistics measured in the same units.

The standard deviation is a popular measure of risk, particularly in financial analysis, because many people associate risk with volatility in stock prices. The standard deviation

Figure 4.7

Excel File *Closing Stock Prices*

	A	B	C	D	E	F
1	Closing Stock Prices					
2						
3	Date	IBM	INTC	CSCO	GE	DJ Industrials Index
4	9/3/2010	$127.58	$18.43	$21.04	$15.39	10447.93
5	9/7/2010	$125.95	$18.12	$20.58	$15.44	10340.69
6	9/8/2010	$126.08	$17.90	$20.64	$15.70	10387.01
7	9/9/2010	$126.36	$18.00	$20.61	$15.91	10415.24
8	9/10/2010	$127.99	$17.97	$20.62	$15.98	10462.77
9	9/13/2010	$129.61	$18.56	$21.26	$16.25	10544.13
10	9/14/2010	$128.85	$18.74	$21.45	$16.16	10526.49
11	9/15/2010	$129.43	$18.72	$21.59	$16.34	10572.73
12	9/16/2010	$129.67	$18.97	$21.93	$16.23	10594.83
13	9/17/2010	$130.19	$18.81	$21.86	$16.29	10607.85
14	9/20/2010	$131.79	$18.93	$21.75	$16.55	10753.62
15	9/21/2010	$131.98	$19.14	$21.64	$16.52	10761.03
16	9/22/2010	$132.57	$19.01	$21.67	$16.50	10739.31
17	9/23/2010	$131.67	$18.98	$21.53	$16.14	10662.42
18	9/24/2010	$134.11	$19.42	$22.09	$16.66	10860.26
19	9/27/2010	$134.65	$19.24	$22.11	$16.43	10812.04
20	9/28/2010	$134.89	$19.51	$21.86	$16.44	10858.14
21	9/29/2010	$135.48	$19.24	$21.87	$16.36	10835.28
22	9/30/2010	$134.14	$19.20	$21.90	$16.25	10788.05
23	10/1/2010	$135.64	$19.32	$21.91	$16.36	10829.68

measures the tendency of a fund's monthly returns to vary from their long-term average (as *Fortune* stated in one of its issues, ". . . standard deviation tells you what to expect in the way of dips and rolls. It tells you how scared you'll be.").[1] For example, a mutual fund's return might have averaged 11% with a standard deviation of 10%. Thus, about two-thirds of the time the annualized monthly return was between 1% and 21%. By contrast, another fund's average return might be 14% but have a standard deviation of 20%. Its returns would have fallen in a range of −6% to 34% and, therefore, is more risky. Many financial Web sites, such as IFA.com and Morningstar.com, provide standard deviations for market indexes and mutual funds.

For example, the Excel file *Closing Stock Prices* (see Figure 4.7) lists daily closing prices for four stocks and the Dow Jones Industrial Average index over a 1-month period. The average closing prices for Intel (INTC) and General Electric (GE) are quite similar, $18.81 and $16.19, respectively. However, the standard deviation of Intel's price over this time frame was $0.50, whereas GE's was $0.35. GE had less variability and, therefore, less risk. A larger standard deviation implies that while a greater potential of a higher return exists, there is also greater risk of realizing a lower return. Many investment publications and Web sites provide standard deviations of stocks and mutual funds to help investors assess risk in this fashion. We learn more about risk in other chapters.

Chebyshev's Theorem and the Empirical Rules

One of the more important results in statistics is **Chebyshev's theorem**, which states that for *any set of data*, the proportion of values that lie within k standard deviations $(k > 1)$ of the mean is at least $1 - 1/k^2$. Thus, for $k = 2$, at least 3/4, or 75%, of the data lie within two standard deviations of the mean; for $k = 3$, at least 8/9, or 89% of the data lie within three standard deviations of the mean. We can use these values to provide a basic understanding of the variation in a set of data using only the computed mean and standard deviation.

[1] *Fortune* magazine 1999 Investor's Guide (December 21, 1998 issue).

EXAMPLE 4.10 Applying Chebyshev's Theorem

For Cost per order data in the *Purchase Orders* database, a two standard deviation interval around the mean is [−$33,390.34, $85,980.98]. If we count the number of observations within this interval, we find that 89 of 94, or 94.68%, fall within two standard deviations of the mean. A three-standard deviation interval is [−$63,233.17, $115,823.81], and we see that 92 of 94, or 97.9%, fall in this interval. Both are above at least 75% and at least 89% of Chebyshev's Theorem.

For many data sets encountered in practice, such as the Cost per order data, the percentages are generally much higher than what Chebyshev's theorem specifies. These are reflected in what are called the **empirical rules**:

1. Approximately 68% of the observations will fall within one standard deviation of the mean, or between $\bar{x} - s$ and $\bar{x} + s$.
2. Approximately 95% of the observations will fall within two standard deviations of the mean, or within $\bar{x} \pm 2s$.
3. Approximately 99.7% of the observations will fall within three standard deviations of the mean, or within $\bar{x} \pm 3s$.

We see that the Cost per order data reflect these empirical rules rather closely. Depending on the data and the shape of the frequency distribution, the actual percentages may be higher or lower.

Two or three standard deviations around the mean are commonly used to describe the variability of most practical sets of data. As an example, suppose that a retailer knows that on average, an order is delivered by standard ground transportation in 8 days with a standard deviation of 1 day. Using the second empirical rule, the retailer can, therefore, tell a customer with confidence that their package should arrive within 6 to 10 days.

As another example, it is important to ensure that the output from a manufacturing process meets the specifications that engineers and designers require. The dimensions for a typical manufactured part are usually specified by a target, or ideal, value as well as a tolerance, or "fudge factor," that recognizes that variation will exist in most manufacturing processes due to factors such as materials, machines, work methods, human performance, environmental conditions, and so on. For example, a part dimension might be specified as 5.00 ± 0.2 cm. This simply means that a part having a dimension between 4.80 and 5.20 cm will be acceptable; anything outside of this range would be classified as defective. To measure how well a manufacturing process can achieve the specifications, we usually take a sample of output, measure the dimension, compute the total variation using the third empirical rule (i.e., estimate the total variation by six standard deviations), and then compare the result to the specifications by dividing the specification range by the total variation. The result is called the **process capability index**, denoted as *Cp*:

$$Cp = \frac{\text{upper specification} - \text{lower specification}}{\text{total variation}} \tag{4.8}$$

Manufacturers use this index to evaluate the quality of their products and determine when they need to make improvements in their processes.

EXAMPLE 4.11 Using Empirical Rules to Measure the Capability of a Manufacturing Process

Figure 4.8 shows a portion of the data collected from a manufacturing process for a part whose dimensions are specified as 5.00 ± 0.2 centimeters. These are provided in the Excel workbook *Manufacturing Measurements*. The mean and standard deviation are first computed in cells J3 and J4 using the Excel AVERAGE and STDEV.S functions (these functions work correctly whether the data are arranged in a single column or in a matrix form). The total variation is then calculated as the mean plus or minus three standard deviations. In cell J14, *Cp* is calculated using formula (4.8). A *Cp* value less than 1.0 is not good; it means that the variation in the process is wider than the specification limits, signifying that some of the parts will not meet the specifications. In practice, many manufacturers want to have *Cp* values of at least 1.5.

Figure 4.9 shows a frequency distribution and histogram of these data (worksheet *Histogram* in the *Manufacturing Measurements* workbook). Note that the bin values represent the upper limits of the groupings in the histogram; thus, 3 observations fell at or below 4.8, the lower specification limit. In addition, 5 observations exceeded the upper specification limit of 5.2. Therefore, 8 of the 200 observations, or 4%, were actually defective, and 96% were acceptable. Although this doesn't meet the empirical rule exactly, you must remember that we are dealing with sample data. Other samples from the same process would have different characteristics, but overall, the empirical rule provides a good estimate of the total variation in the data that we can expect from any sample.

	A	B	C	D	E	F	G	H	I	J
1	Manufacturing Measurements									
2										
3	5.21	5.87	4.85	4.95	5.07	4.96	4.96	5.11	Mean	4.99
4	5.02	5.33	4.82	4.86	4.82	4.96	5.06	5.11	Standard deviation	0.117
5	4.90	5.11	5.02	5.13	5.03	4.94	4.86	5.08		
6	5.00	5.07	4.90	4.95	4.85	5.19	4.96	5.03	Mean - 3*Stdev	4.640
7	5.16	4.93	4.73	5.22	4.89	4.91	4.99	4.94	Mean + 3*Stdev	5.340
8	5.03	4.99	5.04	4.81	4.82	5.01	4.94	4.88	Total variaton	0.700
9	4.96	5.04	5.07	4.91	5.18	4.93	5.06	4.91		
10	5.04	5.14	4.81	4.95	5.02	5.05	4.95	4.86	Lower Specification	4.8
11	4.98	5.09	5.04	4.94	5.05	4.96	5.02	4.89	Upper Specification	5.2
12	5.07	5.06	5.03	4.81	4.88	4.92	5.01	4.91	Specification range	0.4
13	5.02	4.85	5.01	5.11	5.08	4.95	5.04	4.87		
14	5.08	4.93	5.14	4.81	4.98	5.08	5.01	4.93	Cp	0.57

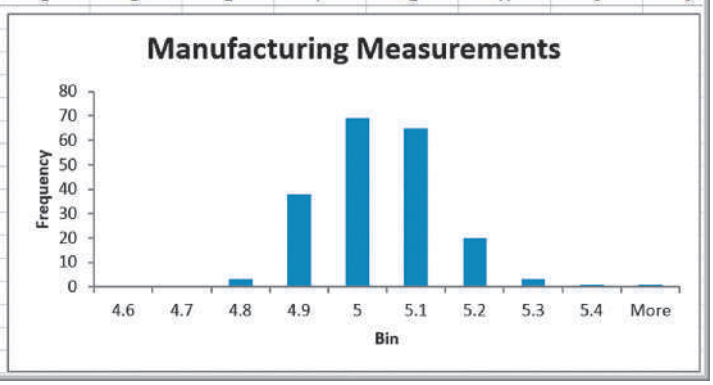
Figure 4.8

Calculation of *Cp* Index

Figure 4.9

Frequency Distribution and Histogram of Manufacturing Measurements

	A	B
1	*Bin*	*Frequency*
2	4.6	0
3	4.7	0
4	4.8	3
5	4.9	38
6	5	69
7	5.1	65
8	5.2	20
9	5.3	3
10	5.4	1
11	More	1



Standardized Values

A **standardized value**, commonly called a ***z*-score**, provides a relative measure of the distance an observation is from the mean, which is independent of the units of measurement. The z-score for the ith observation in a data set is calculated as follows:

$$z_i = \frac{x_i - \bar{x}}{s} \tag{4.9}$$

We subtract the sample mean from the ith observation, x_i, and divide the result by the sample standard deviation. In formula (4.9), the numerator represents the distance that x_i is from the sample mean; a negative value indicates that x_i lies to the left of the mean, and a positive value indicates that it lies to the right of the mean. By dividing by the standard deviation, s, we scale the distance from the mean to express it in units of standard deviations. Thus, a z-score of 1.0 means that the observation is one standard deviation to the right of the mean; a z-score of -1.5 means that the observation is 1.5 standard deviations to the left of the mean. Thus, even though two data sets may have different means and standard deviations, the same z-score means that the observations have the same relative distance from their respective means.

Z-scores can be computed easily on a spreadsheet; however, Excel has a function that calculates it directly, STANDARDIZE(*x, mean, standard_dev*).

EXAMPLE 4.12 Computing *z*-Scores

Figure 4.10 shows the calculations of *z*-scores for a portion of the Cost per order data. This worksheet may be found in the *Purchase Orders* workbook as *z-scores*. In cells B97 and B98, we compute the mean and standard deviation using the Excel AVERAGE and STDEV.S functions. In column C, we could either use formula (4.9) or the Excel STANDARDIZE function. For example, the formula in cell C2 is =(B2 – B97)/B98, but it could also be calculated as =STANDARDIZE(B2,B97,B98). Thus, the first observation $2,700 is 0.79 standard deviations below the mean, whereas observation 92 is 1.61 standard deviations above the mean. Only two observations (x19 and x8) are more than 3 standard deviations above the mean. We saw this in Example 4.10 when we applied Chebyshev's theorem to the data.

Figure 4.10

Computing *z*-Scores for Cost per Order Data

	A	B	C
1	Observation	Cost per order	z-score
2	x1	$2,700.00	-0.79
3	x2	$19,250.00	-0.24
4	x3	$15,937.50	-0.35
5	x4	$18,150.00	-0.27
6	x5	$23,400.00	-0.10
91	x90	$6,750.00	-0.65
92	x91	$16,625.00	-0.32
93	x92	$74,375.00	1.61
94	x93	$72,250.00	1.54
95	x94	$6,562.50	-0.66
96			
97	Mean	$26,295.32	
98	Standard Deviation	$29,842.83	

Figure 4.11

Calculating Coefficients of Variation for *Closing Stock Prices*

	A	B	C	D	E	F
1	Closing Stock Prices					
2						
3	Date	IBM	INTC	CSCO	GE	DJ Industrials Index
4	9/3/2010	$127.58	$18.43	$21.04	$15.39	10447.93
5	9/7/2010	$125.95	$18.12	$20.58	$15.44	10340.69
6	9/8/2010	$126.08	$17.90	$20.64	$15.70	10387.01
22	9/30/2010	$134.14	$19.20	$21.90	$16.25	10788.05
23	10/1/2010	$135.64	$19.32	$21.91	$16.36	10829.68
24	Mean	$130.93	$18.81	$21.50	$16.20	$10,639.98
25	Standard Deviation	$3.22	$0.50	$0.52	$0.35	$171.94
26	Coefficient of Variation	0.025	0.027	0.024	0.022	0.016

Coefficient of Variation

The **coefficient of variation (CV)** provides a relative measure of the dispersion in data relative to the mean and is defined as

$$\text{CV} = \frac{\text{standard deviation}}{\text{mean}} \tag{4.10}$$

Sometimes the coefficient of variation is multiplied by 100 to express it as a percent. This statistic is useful when comparing the variability of two or more data sets when their scales differ.

The coefficient of variation provides a relative measure of risk to return. The smaller the coefficient of variation, the smaller the relative risk is for the return provided. The reciprocal of the coefficient of variation, called **return to risk**, is often used because it is easier to interpret. That is, if the objective is to maximize return, a higher return-to-risk ratio is often considered better. A related measure in finance is the *Sharpe ratio*, which is the ratio of a fund's excess returns (annualized total returns minus Treasury bill returns) to its standard deviation. If several investment opportunities have the same mean but different variances, a rational (risk-averse) investor will select the one that has the smallest variance.[2] This approach to formalizing risk is the basis for modern portfolio theory, which seeks to construct minimum-variance portfolios. As *Fortune* magazine once observed, "It's not that risk is always bad. . . . It's just that when you take chances with your money, you want to be paid for it."[3] One practical application of the coefficient of variation is in comparing stock prices.

EXAMPLE 4.13 Applying the Coefficient of Variation

For example, by examining only the standard deviations in the *Closing Stock Prices* worksheet, we might conclude that IBM is more risky than the other stocks. However, the mean stock price of IBM is much greater than the other stocks. Thus, comparing standard deviations directly provides little information. The coefficient of variation provides a more comparable measure. Figure 4.11 shows the calculations of the coefficients of variation for these variables. For IBM, the CV is 0.025; for Intel, 0.027; for Cisco, 0.024; for GE, 0.022; and for the DJIA, 0.016. We see that the coefficients of variation of the stocks are not very different; in fact, Intel is just slightly more risky than IBM relative to its average price. However, an index fund based on the Dow Industrials would be less risky than any of the individual stocks.

[2]David G. Luenberger, *Investment Science* (New York: Oxford University Press, 1998).
[3]*Fortune* magazine 1999 Investor's Guide (December 21, 1998 issue).

Measures of Shape

Histograms of sample data can take on a variety of different shapes. Figure 4.12 shows the histograms for Cost per order and A/P Terms that we created in Chapter 3 for the *Purchase Orders* data. The histogram for A/P Terms is relatively symmetric, having its modal value in the middle and falling away from the center in roughly the same fashion on either side. However, the Cost per order histogram is asymmetrical, or *skewed*; that is, more of the mass is concentrated on one side, and the distribution of values "tails off" to the other. Those that tail off to the right, like this example, are called *positively skewed*; those that tail off to the left are said to be *negatively skewed.* **Skewness** describes the lack of symmetry of data.

The **coefficient of skewness (CS)** measures the degree of asymmetry of observations around the mean. The coefficient of skewness is computed as

$$\text{CS} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\sigma^3} \tag{4.11}$$

For sample data, replace the population mean and standard deviation with the corresponding sample statistics. Although CS can be computed on a spreadsheet, it can easily be found using the Excel function SKEW(*data range*). If CS is positive, the distribution of values is positively skewed; if negative, it is negatively skewed. The closer CS is to zero, the less the degree of skewness. A coefficient of skewness greater than 1 or less than −1 suggests a high degree of skewness. A value between 0.5 and 1 or between −0.5 and −1 represents moderate skewness. Coefficients between 0.5 and −0.5 indicate relative symmetry.

EXAMPLE 4.14 Measuring Skewness

Using the Excel function in the *Purchase Orders* database SKEW, the coefficients of skewness for the Cost per order and A/P Terms data are calculated as

CS (cost per order) = 1.66
CS (A/P terms) = 0.60

This tells us that the Cost per order data are highly positively skewed, whereas the A/P Terms data have a small positive skewness. These are evident from the histograms in Figure 4.12.

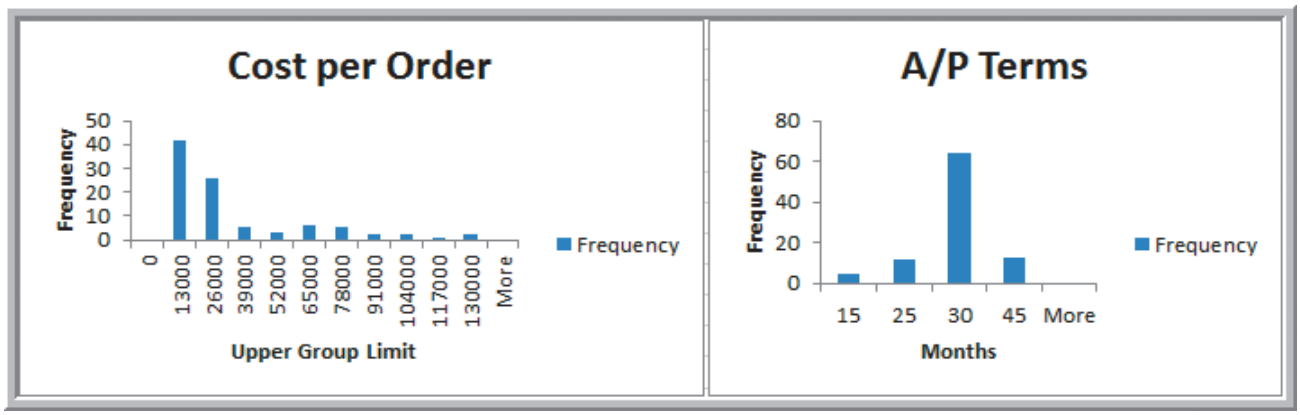
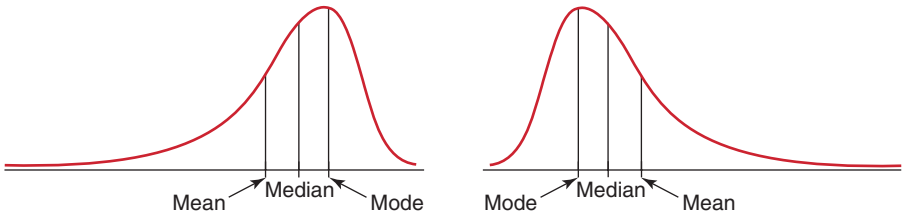


Figure 4.12

Histograms of Cost per Order and A/P Terms

Figure 4.13

Characteristics of Skewed Distributions



Histograms that have only one "peak" are called **unimodal**. (If a histogram has exactly two peaks, we call it **bimodal**. This often signifies a mixture of samples from different populations.) For unimodal histograms that are relatively symmetric, the mode is a fairly good estimate of the mean. For example, the mode for the A/P Terms data is clearly 30 months; the mean is 30.638 months. On the other hand, for the Cost per order data, the mode occurs in the group (0, 13,000). The midpoint of the group, $6,500, which can be used as a numerical estimate of the mode, is not very close at all to the true mean of $26,295.32. The high level of skewness pulls the mean away from the mode.

Comparing measures of location can sometimes reveal information about the shape of the distribution of observations. For example, if the distribution was perfectly symmetrical and unimodal, the mean, median, and mode would all be the same. If it was negatively skewed, we would generally find that mean < median < mode, whereas a positive skewness would suggest that mode < median < mean (see Figure 4.13).

Kurtosis refers to the peakedness (i.e., high, narrow) or flatness (i.e., short, flat-topped) of a histogram. The **coefficient of kurtosis (CK)** measures the degree of kurtosis of a population and can be computed using the Excel function KURT(*data range*). The coefficient of kurtosis is computed as

$$\text{CK} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4} \tag{4.12}$$

(Again, for sample data, use the sample statistics instead of the population measures.) Distributions with values of CK less than 3 are more flat with a wide degree of dispersion; those with values of CK greater than 3 are more peaked with less dispersion.

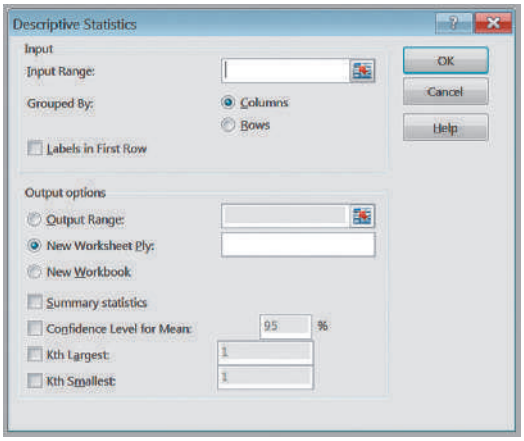
Skewness and kurtosis can help provide more information to evaluate risk than just using the standard deviation. For example, both a negatively and positively skewed distribution may have the same standard deviation, but clearly if the objective is to achieve high return, the negatively skewed distribution will have higher probabilities of larger returns. The higher the kurtosis, the more area the histogram has in the tails rather than in the middle. This can indicate a greater potential for extreme and possibly catastrophic outcomes.

Excel *Descriptive Statistics* Tool

Excel provides a useful tool for basic data analysis, *Descriptive Statistics*, which provides a summary of numerical statistical measures that describe location, dispersion, and shape for sample data (not a population). Click on *Data Analysis* in the *Analysis* group under the *Data* tab in the Excel menu bar. Select *Descriptive Statistics* from the list of tools. The *Descriptive Statistics* dialog shown in Figure 4.14 will appear. You need to enter only the range of the data, which must be in a *single row* or *column.* If the data are in multiple columns, the tool treats each row or column as a separate data set, depending on which you specify. This means that if you have a single data set arranged in a matrix

Figure 4.14

Descriptive Statistics Dialog



format, you would have to stack the data in a single column before applying the *Descriptive Statistics* tool. Check the box *Labels in First Row* if labels are included in the input range. You may choose to save the results in the current worksheet or in a new one. For basic summary statistics, check the box *Summary statistics*; you need not check any others.

EXAMPLE 4.15 Using the Descriptive Statistics Tool

We will apply the *Descriptive Statistics* tool to the Cost per order and A/P Terms data in columns G and H of the *Purchase Orders* database. The results are provided in the *Descriptive Statistics* worksheet in the *Purchase Orders* workbook and are shown in Figure 4.15. The tool provides all the measures we have discussed as well as the standard error, which we discuss in Chapter 6, along with the minimum, maximum, sum, and count.

One important point to note about the use of the tools in the *Analysis Toolpak* versus Excel functions is that while Excel functions dynamically change as the data in the spreadsheet are changed, the results of the *Analysis Toolpak* tools do not. For example, if you compute the average value of a range of numbers directly using the function AVERAGE(*range*), then changing the data in the range will automatically update the result. However, you would have to rerun the *Descriptive Statistics* tool after changing the data.

Figure 4.15

Purchase Orders Data Descriptive Statistics Summary

	A	B	C	D
1	*Cost per order*		*A/P Terms (Months)*	
2				
3	Mean	26295.31915	Mean	30.63829787
4	Standard Error	3078.053014	Standard Error	0.702294026
5	Median	15656.25	Median	30
6	Mode	14910	Mode	30
7	Standard Deviation	29842.8312	Standard Deviation	6.808993205
8	Sample Variance	890594573.8	Sample Variance	46.36238847
9	Kurtosis	2.079637302	Kurtosis	1.512188562
10	Skewness	1.664271519	Skewness	0.599265003
11	Range	127431.25	Range	30
12	Minimum	68.75	Minimum	15
13	Maximum	127500	Maximum	45
14	Sum	2471760	Sum	2880
15	Count	94	Count	94

Descriptive Statistics for Grouped Data

In some situations, data may already be grouped in a frequency distribution, and we may not have access to the raw data. This is often the case when extracting information from government databases such as the Census Bureau or Bureau of Labor Statistics. In these situations, we cannot compute the mean or variance using the standard formulas.

When sample data are summarized in a frequency distribution, the mean of a population may be computed using the formula

$$\mu = \frac{\sum_{i=1}^{N} f_i x_i}{N} \tag{4.13}$$

For samples, the formula is similar:

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{n} \tag{4.14}$$

where f_i is the frequency of observation i. Essentially, we multiply the frequency by the value of observation i, add them up, and divide by the number of observations.

We may use similar formulas to compute the population variance for grouped data,

$$\sigma^2 = \frac{\sum_{i=1}^{N} f_i (x_i - \mu)^2}{N} \tag{4.15}$$

and sample variance,

$$s^2 = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{n - 1} \tag{4.16}$$

To find the standard deviation, take the square root of the variance, as we did earlier.

Note the similarities between these formulas and formulas (4.13) and (4.14). In multiplying the values by the frequency, we are essentially adding the same values f_i times. So they really are the same formulas, just expressed differently.

EXAMPLE 4.16 Computing Statistical Measures from Frequency Distributions

The worksheet *Statistical Calculations* in the *Computer Repair Times* workbook shows the calculations of the mean and variance using formulas (4.14) and (4.16) for the frequency distribution of repair times. A portion of this is shown in Figure 4.16. In column C, we multiply the frequency by the value of the observations [the numerator in formula (4.14)] and then divide by n, the sum of the frequencies in column B, to find the mean in cell C49. Columns D, E, and F provide the calculations needed to find the variance. We divide the sum of the data in column F by $n - 1 = 249$ to find the variance in cell F49.

	A	B	C	D	E	F
1	Computer Repair Times					
2						
3	Days (x)	Frequency (f)	Frequency*Days	Days - Mean	(Days - mean)^2	Frequency*(Days - Mean)^2
4	0	0	0	-14.912	222.368	0.000
5	1	0	0	-13.912	193.544	0.000
6	2	0	0	-12.912	166.720	0.000
7	3	0	0	-11.912	141.896	0.000
43	39	1	39	24.088	580.232	580.232
44	40	1	40	25.088	629.408	629.408
45	41	0	0	26.088	680.584	0.000
46	42	0	0	27.088	733.760	0.000
47	Sum	250	3728			8840.064
48						
49		Mean	14.912		Variance	35.50226506

Figure 4.16

Calculations of Mean and Variance Using a Frequency Distribution

If the data are grouped into k cells in a frequency distribution, we can use modified versions of these formulas to estimate the mean and variance by replacing x_i with a representative value (such as the midpoint) for all the observations in each cell.

EXAMPLE 4.17 Computing Descriptive Statistics for a Grouped Frequency Distribution

Figure 4.17 shows data obtained from the U.S. Census Bureau showing the number of households that spent different percentages of their income on rent. Suppose we wanted to calculate the average percentage and the standard deviation. Because we don't have the raw data, we can only estimate these statistics by assuming some representative value for each group. For the groups that are defined by an upper and lower value, this is easy to do; we can use the midpoints—for instance, 5% for the first group and 12% for the second group. However, it's not clear what to do for the 50 percent or more group. For this group, we have no information to determine what the best value might be. It might be unreasonable to assume the midpoint between 50% and 100%, or 75%; a more rational value might be 58% or 60%. When dealing with uncertain or ambiguous information in business analytics applications, we often have to make the best assumption we can. In this case, we choose 60%. The calculations, shown in Figure 4.18 (worksheet *Calculations* in the *Census Rent Data* workbook), find a mean of close to 30% and a standard deviation of 17.61%.

Figure 4.17

Census Bureau Rent Data

	A	B	C
1	Gross Rent as a Percentage of Household Income in 1999		
2	Source: US Census Bureau		
3			
4	Group	Number of Households	
5	Less than 10 percent	2,239,346	
6	10 to 14 percent	4,130,917	
7	15 to 19 percent	5,037,981	
8	20 to 24 percent	4,498,604	
9	25 to 29 percent	3,666,233	
10	30 to 34 percent	2,585,327	
11	35 to 39 percent	1,809,948	
12	40 to 49 percent	2,364,443	
13	50 percent or more	6,209,568	
14	Not computed	2,657,135	

Figure 4.18

Census Rent Data Calculations

	A	B	C	D	E	F	G
1							
2	Group	Percent (x)	Number (f)	f*x	x - mean	(x - mean)^2	f*(x - mean)^2
3	Less than 10 percent	5%	2,239,346	111967.30	-24.8645%	0.0618	138446.0126
4	10 to 14 percent	12%	4,130,917	495710.04	-17.8645%	0.0319	131834.1452
5	15 to 19 percent	17%	5,037,981	856456.77	-12.8645%	0.0165	83376.1701
6	20 to 24 percent	22%	4,498,604	989692.88	-7.8645%	0.0062	27823.9852
7	25 to 29 percent	27%	3,666,233	989882.91	-2.8645%	0.0008	3008.2636
8	30 to 34 percent	32%	2,585,327	827304.64	2.1355%	0.0005	1179.0089
9	35 to 39 percent	37%	1,809,948	669680.76	7.1355%	0.0051	9215.4310
10	40 to 49 percent	44.50%	2,364,443	1052177.14	14.6355%	0.0214	50645.9048
11	50 percent or more	60%	6,209,568	3725740.80	30.1355%	0.0908	563921.1249
12		Sum	32,542,367	9718613.24			1009450.0462
13							
14			Mean	29.86%		Variance	0.031019565
15						Standard Dev.	17.61%

It is important to understand that because we have not used all the original data in computing these statistics, they are only estimates of the true values.

Descriptive Statistics for Categorical Data: The Proportion

Statistics such as means and variances are not appropriate for categorical data. Instead, we are generally interested in the fraction of data that have a certain characteristic. The formal statistical measure is called the **proportion**, usually denoted by p. Proportions are key descriptive statistics for categorical data, such as defects or errors in quality control applications or consumer preferences in market research.

EXAMPLE 4.18 Computing a Proportion

In the *Purchase Orders* database, column A lists the name of the supplier for each order. We may use the Excel function =COUNTIF(*data range, criteria*) to count the number of observations meeting specified characteristics. For instance, to find the number of orders placed with Spacetime Technologies, we used the function =COUNTIF(A4:A97, "Spacetime Technologies"). This returns a value of 12. Because 94 orders were placed, the proportion of orders placed with Spacetime Technologies is $p = 12/94 = 0.128$.

It is important to realize that proportions are numbers between 0 and 1. Although we often convert these to percentages—for example, 12.8% of orders were placed with Spacetime Technologies in the last example—we must be careful to use the decimal expression of a proportion when statistical formulas require it.

Statistics in PivotTables

We introduced PivotTables in Chapter 3 and applied them to finding simple counts and creating cross-tabulations. PivotTables also have the functionality to calculate many basic statistical measures from the data summaries. If you look at the *Value Field Settings* dialog shown in Figure 4.19, you can see that you can calculate the average, standard deviation, and variance of a value field.

Figure 4.19

Value Field Settings Dialog

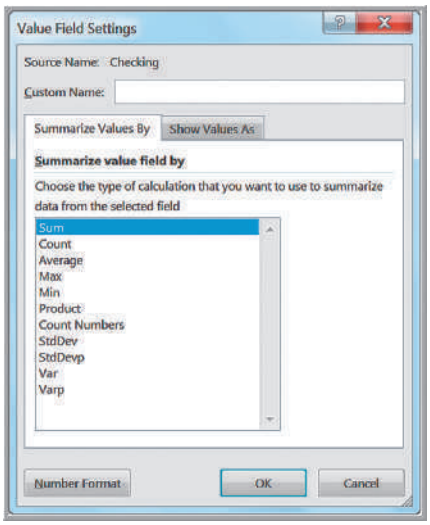


Figure 4.20

PivotTable for Average Checking and Savings Account Balances by Job

	A	B	C
1			
2			
3	Row Labels	Average of Checking	Average of Savings
4	Management	$606.94	$1,616.83
5	Skilled	$1,079.24	$1,836.43
6	Unemployed	$1,697.64	$2,760.91
7	Unskilled	$1,140.27	$1,741.44
8	Grand Total	$1,048.01	$1,812.56

EXAMPLE 4.19 Statistical Measures in PivotTables

In the *Credit Risk Data* Excel file, suppose that we want to find the average amount of money in checking and savings accounts by job classification. Create a PivotTable, and in the *PivotTable Field List*, move Job to the *Row Labels field* and Checking and Savings to the *Values* field. Then change the field settings from "Sum of Checking" and "Sum of Savings" to the averages. The result is shown in Figure 4.20; we have also formatted the values as currency using the *Number Format* button in the dialog. In a similar fashion, you could find the standard deviation or variance of each group by selecting the appropriate field settings.

Measures of Association

Two variables have a strong statistical relationship with one another if they appear to move together. We see many examples on a daily basis; for instance, attendance at baseball games is often closely related to the win percentage of the team, and ice cream sales likely have a strong relationship with daily temperature. We can examine relationships between two variables visually using scatter charts, which we introduced in Chapter 3.

When two variables appear to be related, you might suspect a cause-and-effect relationship. Sometimes, however, statistical relationships exist even though a change in one variable is not *caused* by a change in the other. For example, the *New York Times* reported a strong statistical relationship between the golf handicaps of corporate CEOs and their companies' stock market performance over 3 years. CEOs who were better-than-average golfers

	A	B	C	D	E	F	G
1	Colleges and Universities						
2							
3	School	Type	Median SAT	Acceptance Rate	Expenditures/Student	Top 10% HS	Graduation %
4	Amherst	Lib Arts	1315	22%	$ 26,636	85	93
5	Barnard	Lib Arts	1220	53%	$ 17,653	69	80
6	Bates	Lib Arts	1240	36%	$ 17,554	58	88
7	Berkeley	University	1176	37%	$ 23,665	95	68
8	Bowdoin	Lib Arts	1300	24%	$ 25,703	78	90
9	Brown	University	1281	24%	$ 24,201	80	90
10	Bryn Mawr	Lib Arts	1255	56%	$ 18,847	70	84

Figure 4.21

Portion of Excel File *Colleges and Universities*

were likely to deliver above-average returns to shareholders.[4] Clearly, the ability to golf would not cause better business performance. Therefore, you must be cautious in drawing inferences about causal relationships based solely on statistical relationships. (On the other hand, you might want to spend more time out on the practice range!)

Understanding the relationships between variables is extremely important in making good business decisions, particularly when cause-and-effect relationships can be justified. When a company understands how internal factors such as product quality, employee training, and pricing factors affect such external measures as profitability and customer satisfaction, it can make better decisions. Thus, it is helpful to have statistical tools for measuring these relationships.

The Excel file *Colleges and Universities*, a portion of which is shown in Figure 4.21, contains data from 49 top liberal arts and research universities across the United States. Several questions might be raised about statistical relationships among these variables. For instance, does a higher percentage of students in the top 10% of their high school class suggest a higher graduation rate? Is acceptance rate related to the amount spent per student? Do schools with lower acceptance rates tend to accept students with higher SAT scores? Questions such as these can be addressed by computing statistical measures of association between the variables.

Covariance

Covariance is a measure of the linear association between two variables, X and Y. Like the variance, different formulas are used for populations and samples. Computationally, covariance of a population is the average of the products of deviations of each observation from its respective mean:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N} \tag{4.17}$$

To better understand the covariance, let us examine formula (4.17). The covariance between X and Y is the average of the product of the deviations of each pair of observations from their respective means. Suppose that large (small) values of X are generally associated with large (small) values of Y. Then, in most cases, both x_i and y_i are either above or below their respective means. If so, the product of the deviations from the means will be a positive number and when added together and averaged will give a positive value for the covariance. On the other hand, if small (large) values of X are associated with large (small) values of

[4]Adam Bryant, "CEOs' Golf Games Linked to Companies' Performance," *Cincinnati Enquirer*, June 7, 1998, El.

Y, then one of the deviations from the mean will generally be negative while the other is positive. When multiplied together, a negative value results, and the value of the covariance will be negative. Thus, the larger the absolute value of the covariance, the higher is the degree of linear association between the two variables. The sign of the covariance tells us whether there is a direct relationship (i.e., one variable increases as the other increases) or an inverse relationship (i.e., one variable increases while the other decreases, or vice versa). We can generally identify the strength of any linear association between two variables and the sign of the covariance by constructing a scatter diagram. The Excel function COVARIANCE.P(*array1, array2*) computes the covariance of a population.

The sample covariance is computed as

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \tag{4.18}$$

Similar to the sample variance, note the use of $n - 1$ in the denominator. The Excel function COVARIANCE.S(*array1, array2*) computes the covariance of a sample.

EXAMPLE 4.20 Computing the Covariance

Figure 4.22 shows a scatter chart of graduation rate (*Y*-variable) versus median SAT scores (*X*-variable) for the *Colleges and Universities* data. It appears that as the median SAT scores increase, the graduate rate also increases; thus, we would expect to see a positive covariance. Figure 4.23 shows the calculations using formula (4.18); these are provided in the worksheet *Covariance* in the *Colleges and Universities* Excel workbook. The Excel function =COVARIANCE.S(B2:B50,C2:C50) in cell F55 verifies the calculations.

Correlation

The numerical value of the covariance is generally difficult to interpret because it depends on the units of measurement of the variables. For example, if we expressed the graduation rate as a true proportion rather than as a percentage in the previous example, the numerical value of the covariance would be smaller, although the linear association between the variables would be the same.

Correlation is a measure of the linear relationship between two variables, *X* and *Y*, which does not depend on the units of measurement. Correlation is measured by the

Figure 4.22

Scatter Chart of Graduation Rate versus Median SAT

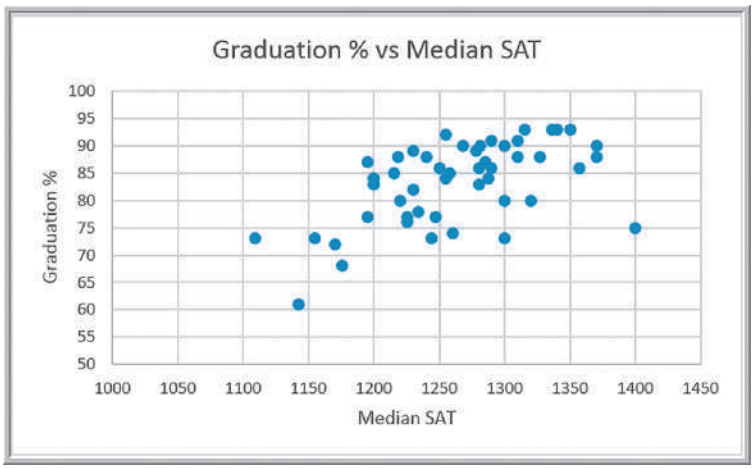


Figure 4.23

Covariance Calculations for Graduation Rate and Median SAT

	A	B	C	D	E	F
1		Graduation % (X)	Median SAT (Y)	X - Mean(X)	Y - Mean(Y)	(X - Mean(X))(Y-Mean(Y))
2		93	1315	9.755	51.898	506.2698875
3		80	1220	-3.245	-43.102	139.8617243
4		88	1240	4.755	-23.102	-109.8525614
47		86	1250	2.755	-13.102	-36.09745939
48		91	1290	7.755	26.898	208.5964182
49		93	1336	9.755	72.898	711.1270304
50		93	1350	9.755	86.898	847.698459
51	Mean	83.245	1263.102		Sum	12641.77551
52					Count	49
53					Covariance	263.3703231
54						
55					COVARIANCE.S	263.3703231

correlation coefficient, also known as the **Pearson product moment correlation coefficient**. The correlation coefficient for a population is computed as

$$\rho_{xy} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \tag{4.19}$$

By dividing the covariance by the product of the standard deviations, we are essentially scaling the numerical value of the covariance to a number between −1 and 1.

In a similar fashion, the **sample correlation coefficient** is computed as

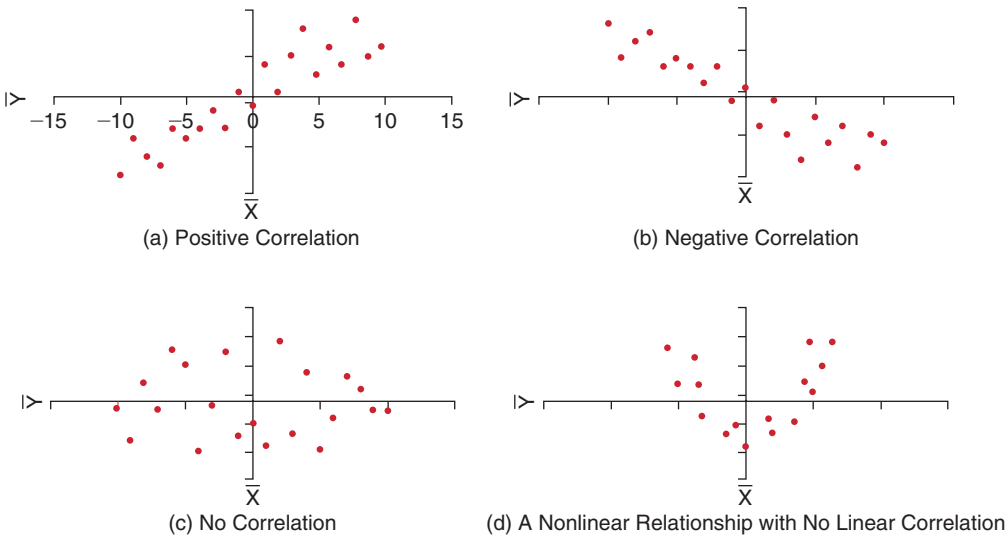
$$r_{xy} = \frac{\text{cov}(X,Y)}{s_x s_y} \tag{4.20}$$

Excel's CORREL function computes the correlation coefficient of two data arrays.

A correlation of 0 indicates that the two variables have no linear relationship to each other. Thus, if one changes, we cannot reasonably predict what the other variable might do. A positive correlation coefficient indicates a linear relationship for which one variable increases as the other also increases. A negative correlation coefficient indicates a linear relationship for one variable that increases while the other decreases. In economics, for instance, a price-elastic product has a negative correlation between price and sales; as price increases, sales decrease, and vice versa. These relationships are illustrated in Figure 4.24. Note that although Figure 4.24(d) has a clear relationship between the variables, the relationship is not linear and the correlation is zero.

Figure 4.24

Examples of Correlation



	A	B	C	D	E	F
1		Graduation % (X)	Median SAT (Y)	X - Mean(X)	Y - Mean(Y)	(X - Mean(X))(Y-Mean(Y))
2		93	1315	9.755	51.898	506.2698875
3		80	1220	-3.245	-43.102	139.8617243
4		88	1240	4.755	-23.102	-109.8525614
47		86	1250	2.755	-13.102	-36.09745939
48		91	1290	7.755	26.898	208.5964182
49		93	1336	9.755	72.898	711.1270304
50		93	1350	9.755	86.898	847.698459
51	Mean	83.245	1263.102		Sum	12641.77551
52	Standard Deviation	7.449	62.676		Count	49
53					Covariance	263.3703231
54					Correlation	0.564146827
55						
56					CORREL Function	0.564146827

Figure 4.25

Correlation Calculations for Graduation Rate and Median SAT

EXAMPLE 4.21 Computing the Correlation Coefficient

Figure 4.25 shows the calculations for computing the sample correlation coefficient for the graduation rate and median SAT variables in the *Colleges and Universities* data file. We first compute the standard deviation of each variable in cells B52 and C52 and then divide the covariance by the product of these standard deviations in cell F54. Cell F56 shows the same result using the Excel function =CORREL(B2:B50,C2:C50).

When using the CORREL function, it does not matter if the data represent samples or populations. In other words,

$$\text{CORREL}(array1, array2) = \frac{\text{COVARIANCE.P}(array1, array2)}{\text{STDEV.P}(array1) \times \text{STDEV.P}(array2)}$$

and

$$\text{CORREL}(array1, array2) = \frac{\text{COVARIANCE.S}(array1, array2)}{\text{STDEV.S}(array1) \times \text{STDEV.S}(array2)}$$

For instance, in Example 4.21, if we assume that the data are populations, we find that the population standard deviation for X is 7.372 and the population standard deviation for Y is 62.034 (using the function STDEV.P). By dividing the population covariance, 257.995 (using the function COVARIANCE.P), by the product of these standard deviations, we find that the correlation coefficient is still 0.564 as computed by the CORREL function.

Excel Correlation Tool

The *Data Analysis Correlation* tool computes correlation coefficients for more than two arrays. Select *Correlation* from the *Data Analysis* tool list. The dialog is shown in Figure 4.26. You need to input only the range of the data (which must be in contiguous columns; if not, you must move them in your worksheet), specify whether the data are grouped by rows or columns (most applications will be grouped by columns), and indicate whether the first row contains data labels. The output of this tool is a matrix giving the correlation between each pair of variables. This tool provides the same output as the CORREL function for each pair of variables.

Figure 4.26

Excel *Correlation* Tool Dialog

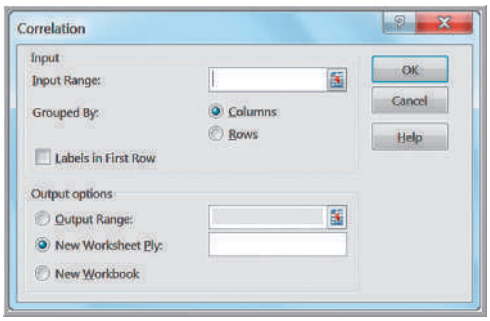


Figure 4.27

Correlation Results for *Colleges and Universities* Data

	A	B	C	D	E	F
1		*Median SAT*	*Acceptance Rate*	*Expenditures/Student*	*Top 10% HS*	*Graduation %*
2	Median SAT	1				
3	Acceptance Rate	-0.601901959	1			
4	Expenditures/Student	0.572741729	-0.284254415	1		
5	Top 10% HS	0.503467995	-0.609720972	0.505782049	1	
6	Graduation %	0.564146827	-0.55037751	0.042503514	0.138612667	1

EXAMPLE 4.22 Using the *Correlation* Tool

The correlation matrix among all the variables in the *Colleges and Universities* data file is shown in Figure 4.27. None of the correlations are very strong. The moderate positive correlation between the graduation rate and SAT scores indicates that schools with higher median SATs have higher graduation rates. We see a moderate negative correlation between acceptance rate and graduation rate, indicating that schools with lower acceptance rates have higher graduation rates. We also see that the acceptance rate is also negatively correlated with the median SAT and Top 10% HS, suggesting that schools with lower acceptance rates have higher student profiles. The correlations with Expenditures/Student also suggest that schools with higher student profiles spend more money per student.

Outliers

Earlier we had noted that the mean and range are sensitive to outliers—unusually large or small values in the data. Outliers can make a significant difference in the results we obtain from statistical analyses. An important statistical question is how to identify them. The first thing to do from a practical perspective is to check the data for possible errors, such as a misplaced decimal point or an incorrect transcription to a computer file. Histograms can help to identify possible outliers visually. We might use the empirical rule and z-scores to identify an outlier as one that is more than three standard deviations from the mean. We can also identify outliers based on the interquartile range. "Mild" outliers are often defined as being between 1.5*IQR and 3*IQR to the left of Q1 or to the right of Q3, and "extreme" outliers, as more than 3*IQR away from these quartiles. Basically, there is no standard definition of what constitutes an outlier other than an unusual observation as compared with the rest. However, it is important to try to identify outliers and determine their significance when conducting business analytic studies.

Figure 4.28

Portion of Home Market Value

	A	B	C
1	Home Market Value		
2			
3	House Age	Square Feet	Market Value
4	33	1,812	$90,000.00
5	32	1,914	$104,400.00
6	32	1,842	$93,300.00
7	33	1,812	$91,000.00
8	32	1,836	$101,900.00
9	33	2,028	$108,500.00
10	32	1,732	$87,600.00
11	33	1,850	$96,000.00

Figure 4.29

Computing *z*-Scores for Examining Outliers

	A	B	C	D	E
1	Home Market Value				
2					
3	House Age	Square Feet	z-score	Market Value	z-score
4	33	1,812	0.5300	$90,000.00	-0.196
5	32	1,914	0.9931	$104,400.00	1.168
6	32	1,842	0.6662	$93,300.00	0.117
7	33	1,812	0.5300	$91,000.00	-0.101
41	27	1,484	-0.9592	$81,300.00	-1.020
42	27	1,520	-0.7957	$100,700.00	0.818
43	28	1,520	-0.7957	$87,200.00	-0.461
44	27	1,684	-0.0511	$96,700.00	0.439
45	27	1,581	-0.5188	$120,700.00	2.713
46	Mean	1,695		92,069	
47	Standard Deviation	220.257		10553.083	

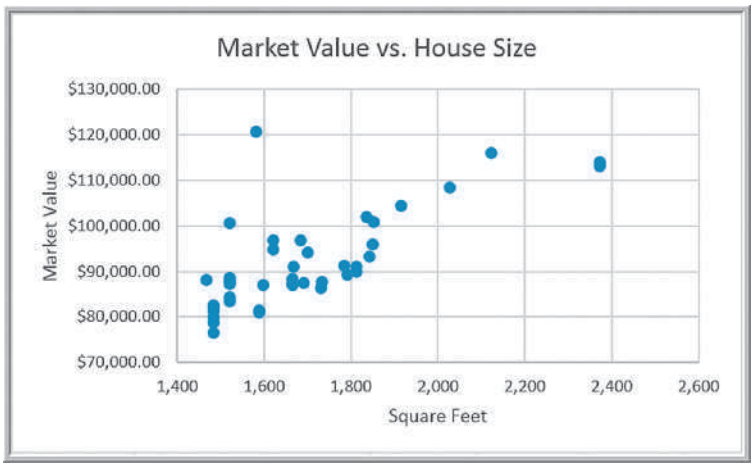
EXAMPLE 4.23 Investigating Outliers

The Excel data file *Home Market Value* provides a sample of data for homes in a neighborhood (Figure 4.28). Figure 4.29 shows *z*-score calculations for the square feet and market value variables. None of the *z*-scores for either of these variables exceed 3 (these calculations can be found in the worksheet *Outliers* in the Excel *Home Market Value* workbook). However, while individual variables might not exhibit outliers, combinations of them might. We see this in the scatter diagram in Figure 4.30. The last observation has a high market value ($120,700) but a relatively small house size (1,581 square feet). The point on the scatter diagram does not seem to coincide with the rest of the data.

The question is what to do with possible outliers. They should not be blindly eliminated unless there is a legitimate reason for doing so—for instance, if the last home in the *Home Market Value* example has an outdoor pool that makes it significantly different from the rest of the neighborhood. Statisticians often suggest that analyses should be run with and without the outliers so that the results can be compared and examined critically.

Figure 4.30

Scatter Diagram of House Size versus Market Value



Statistical Thinking in Business Decisions

The importance of applying statistical concepts to make good business decisions and improve performance cannot be overemphasized. **Statistical thinking** is a philosophy of learning and action for improvement that is based on the principles that

- all work occurs in a system of interconnected processes,
- variation exists in all processes, and
- better performance results from understanding and reducing variation.[5]

Work gets done in any organization through *processes*—systematic ways of doing things that achieve desired results. Understanding business processes provides the context for determining the effects of variation and the proper type of action to be taken. Any process contains many sources of variation. In manufacturing, for example, different batches of material vary in strength, thickness, or moisture content. During manufacturing, tools experience wear, vibrations cause changes in machine settings, and electrical fluctuations cause variations in power. Workers may not position parts on fixtures consistently, and physical and emotional stress may affect workers' consistency. In addition, measurement gauges and human inspection capabilities are not uniform, resulting in measurement error. Similar phenomena occur in service processes because of variation in employee and customer behavior, application of technology, and so on. Reducing variation results in more consistency in manufacturing and service processes, fewer errors, happier customers, and better accuracy of such things as delivery time quotes.

Although variation exists everywhere, many managers often do not recognize it or consider it in their decisions. How often do managers make decisions based on one or two data points without looking at the pattern of variation, see trends in data that aren't justified, or try to manipulate measures they cannot truly control? Unfortunately, the answer is quite often. For example, if sales in some region fell from the previous quarter, a regional manager might quickly blame her sales staff for not working hard enough, even though the drop in sales may simply be the result of uncontrollable variation. Usually, it is simply a matter of ignorance of how to deal with variation in data. This is where business analytics can play a significant role. Statistical analysis can provide better insight into the facts and nature of relationships among the many factors that may have contributed to an event and enable managers to make better decisions.

EXAMPLE 4.24 Applying Statistical Thinking

Figure 4.31 shows a portion of data in the Excel file *Surgery Infections* that document the number of infections that occurred after surgeries over 36 months at one hospital, along with a line chart of the number of infections. (We will assume that the number of surgeries performed each month was the same.) The number of infections tripled in months 2 and 3 as compared to the first month. Is this indicative of trend caused by failure of some health care protocol or simply random variation? Should action be taken to determine a cause? From a statistical perspective, three points are insufficient to conclude that a trend exists. It is more appropriate to look at a larger sample of data and study the pattern of variation.

Over the 36 months, the data clearly indicate that variation exists in the monthly infection rates. The number of infections seems to fluctuate between 0 and 3 with the exception of month 12. However, a visual analysis of the chart cannot necessarily lead to a valid conclusion. So let's apply some statistical thinking. The average number of infections is 1.583 and the standard deviation is 1.180. If we apply the empirical rule that most observations should fall within three standard deviations of the mean, we arrive at the range

(*continued*)

[5]Galen Britz, Don Emerling, Lynne Hare, Roger Hoerl, and Janice Shade, "How to Teach Others to Apply Statistical Thinking," *Quality Progress* (June 1997): 67–79.

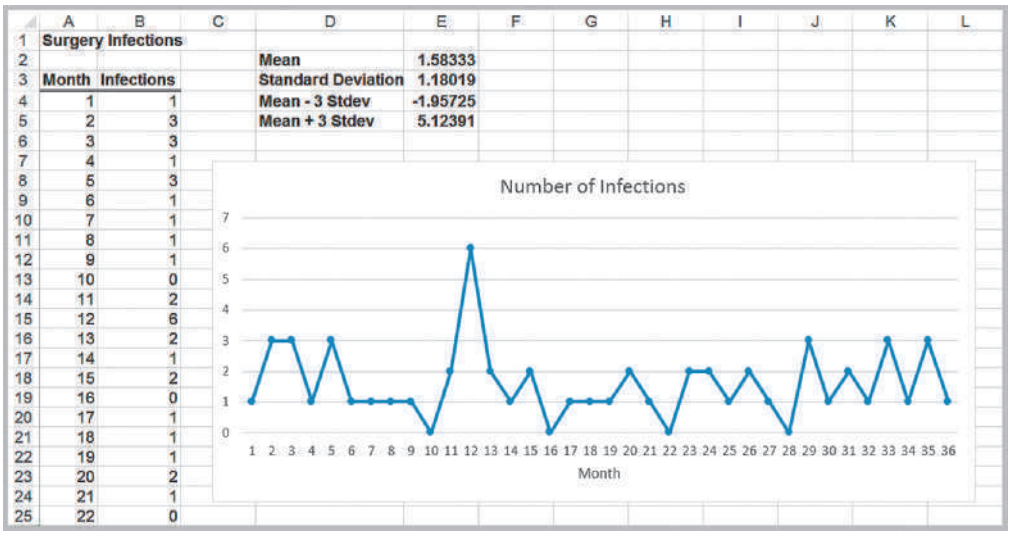
of −1.957 (clearly the number of infections cannot be negative, so let's set this value to zero), and 5.12. This means that, from a statistical perspective, we can expect almost all the observations to fall within these limits. Figure 4.32 shows the chart displaying these ranges. The number of infections for month 12 clearly exceeds the upper range value and suggests that the number of infections for this month is statistically different from the rest. The hospital administrator should seek to investigate what may have happened that month and try to prevent similar occurrences.

Similar analyses are used routinely in quality control and other business applications to monitor performance statistically. The proper analytical calculations depend on the type of measure and other factors and are explained fully in books dedicated to quality control and quality management.

Variability in Samples

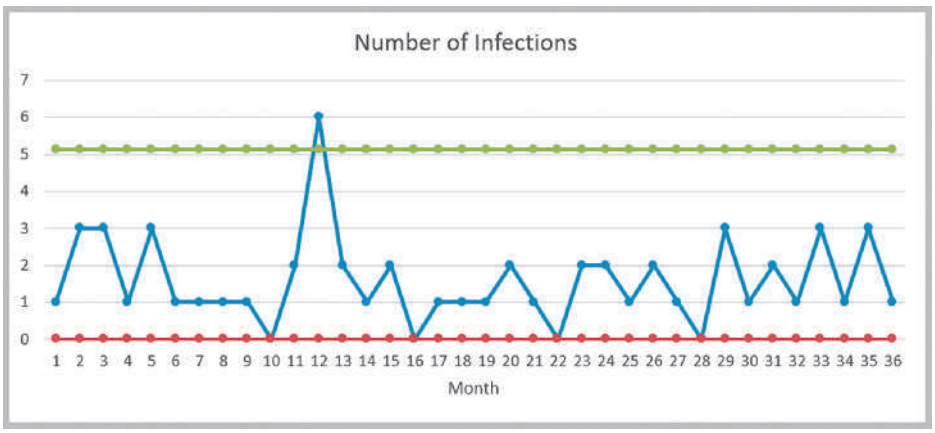
Because we usually deal with sample data in business analytics applications, it is extremely important to understand that different samples from any population will vary; that is, they will have different means, standard deviations, and other statistical measures and will have differences in the shapes of histograms. In particular, samples are extremely sensitive to the sample size—the number of observations included in the samples.

Figure 4.31
Surgery Infections



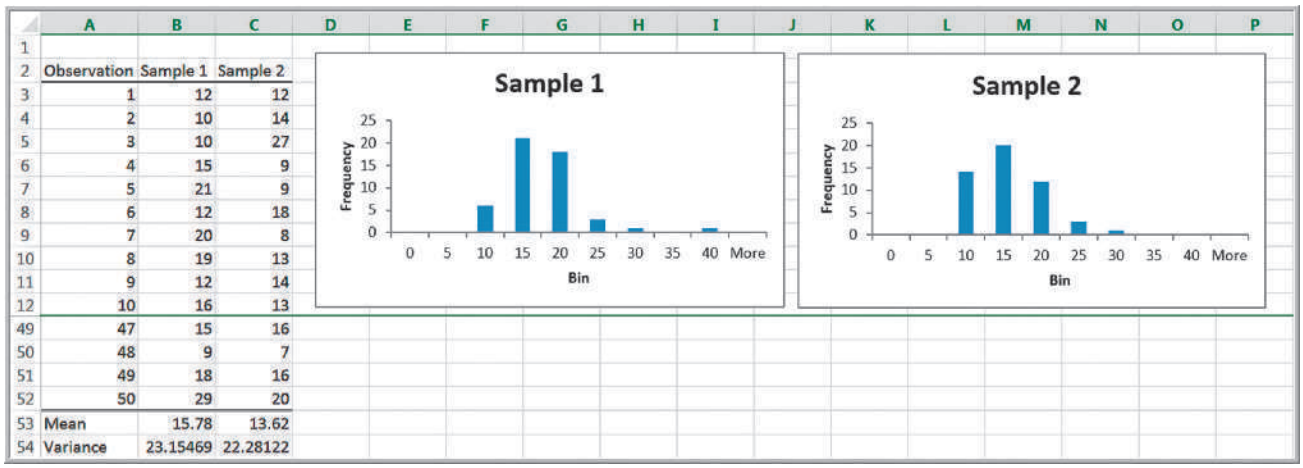
	A	B	C	D	E	F	G	H	I	J	K	L
1	Surgery Infections											
2				Mean	1.58333							
3	Month	Infections		Standard Deviation	1.18019							
4	1	1		Mean - 3 Stdev	-1.95725							
5	2	3		Mean + 3 Stdev	5.12391							
6	3	3										
7	4	1										
8	5	3										
9	6	1										
10	7	1										
11	8	1										
12	9	1										
13	10	0										
14	11	2										
15	12	6										
16	13	2										
17	14	1										
18	15	2										
19	16	0										
20	17	1										
21	18	1										
22	19	1										
23	20	2										
24	21	1										
25	22	0										

Figure 4.32
Infections with Empirical Rule Ranges



EXAMPLE 4.25 Variation in Sample Data

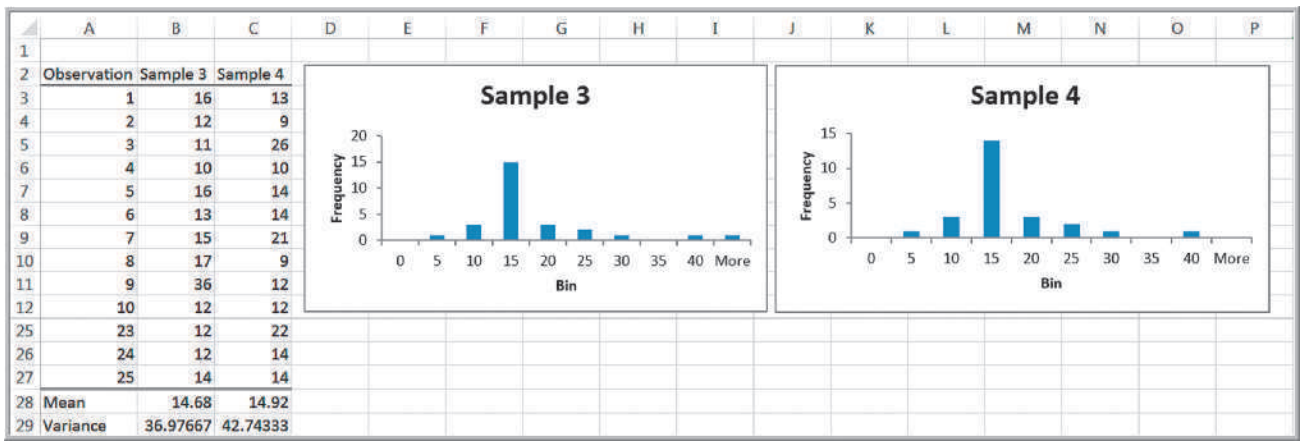
In Example 4.5, we illustrated a frequency distribution for 250 computer repair times. The average repair time is 14.9 days, and the variance of the repair times is 35.50. Suppose we selected some smaller samples from these data. Figure 4.33 shows two samples of size 50 randomly selected from the 250 repair times. Observe that the means and variances differ from each other as well as from the mean and variance of the entire sample shown in Figure 4.5. In addition, the histograms show a slightly different profile. In Figure 4.34 we show the results for two smaller samples of size 25. Here we actually see *more* variability in both the statistical measures and the histograms as compared with the entire data set.



Observation	Sample 1	Sample 2
1	12	12
2	10	14
3	10	27
4	15	9
5	21	9
6	12	18
7	20	8
8	19	13
9	12	14
10	16	13
47	15	16
48	9	7
49	18	16
50	29	20
Mean	15.78	13.62
Variance	23.15469	22.28122

Figure 4.33

Two Samples of Size 50 of *Computer Repair Times*



Observation	Sample 3	Sample 4
1	16	13
2	12	9
3	11	26
4	10	10
5	16	14
6	13	14
7	15	21
8	17	9
9	36	12
10	12	12
23	12	22
24	12	14
25	14	14
Mean	14.68	14.92
Variance	36.97667	42.74333

Figure 4.34

Two Samples of Size 25 of *Computer Repair Times*

This example demonstrates that it is important to understand the variability in sample data and that statistical information drawn from a sample may not accurately represent the population from which it comes. This is one of the most important concepts in applying business analytics. We explore this topic more in Chapter 6.

Analytics in Practice: Applying Statistical Thinking to Detecting Financial Problems[6]

Over the past decade, there have been numerous discoveries of management fraud that have led to the downfall of several prominent companies. These companies had been effective in hiding their financial difficulties, and investors and creditors are now seeking ways to identify financial problems before scandals occur. Even with the passage of the Sarbanes-Oxley Act in July 2002, which helped to improve the quality of the data being disclosed to the public, it is still possible to misjudge an organization's financial strength without analytical evaluation. Several warning signs exist, but there is no systematic and objective way to determine whether a given financial metric, such as a write-off or insider-trading pattern, is high or unusual.

Researchers have proposed using statistical thinking to detect anomalies. They propose an "anomaly detection score," which is the difference between a target financial measure and the company's own past performance or its competitors' current performance using standard deviations. This technique is a variation of a standardized *z*-score. Specifically, their approach involves comparing performance to past performance (within analysis) and comparing performance to the performance of the company's peers over the same period (between analyses). They created two types of exceptional anomaly scores: *z*-between (Z_b) to address the variation between companies and *z*-within (Z_w) to address the variation within the company. These measures quantify the number of standard deviations a company's financial measure deviates from the average. Using these measures, the researchers applied the technique to 25 case studies. These included several high-profile companies that had been charged with financial statement fraud by the SEC or had admitted accounting errors, causing a restatement of their financials. The method was able to identify anomalies for critical metrics known by experts to be warning signs for financial-statement fraud. These warning signs were consistent when compared with expert postmortem commentary on the high-profile fraud cases. More importantly, they signaled anomalous behavior at least six quarters before an SEC investigation announcement with fewer than 5% false negatives and 40% false positives.

NAN728/Shutterstock.com

Key Terms

Arithmetic mean (mean)
Bimodal
Chebyshev's theorem
Coefficient of kurtosis (CK)
Coefficient of skewness (CS)
Coefficient of variation (CV)

[6]Based on Deniz Senturk, Christina LaComb, Radu Neagu, and Murat Doganaksoy, "Detect Financial Problems With Six Sigma," *Quality Progress* (April 2006): 41–47.

Correlation
Correlation coefficient (Pearson product moment correlation coefficient)
Covariance
Dispersion
Empirical rules
Interquartile range (IRQ, or midspread)
Kurtosis
Median
Midrange
Mode
Outlier
Population
Process capability index
Proportion
Range
Return to risk
Sample
Sample correlation coefficient
Skewness
Standard deviation
Standardized value (z-score)
Statistical thinking
Unimodal
Variance

Problems and Exercises

1. Data obtained from a county auditor in the Excel file *Home Market Value* provide information about the age, square footage, and current market value of houses along one street in a particular subdivision. Considering these data as a population of homeowners on this street, compute the mean, variance, and standard deviation for each of these variables using a spreadsheet and formulas (4.1), (4.4), and (4.6). Verify your calculations using the appropriate Excel function.
2. In the Excel file *Facebook Survey*, find the average and median hours online/week and number of friends in the sample using the appropriate Excel functions. Compute the midrange and compare all measures of location.
3. For the Excel file *Tablet Computer Sales*, find the average number, standard deviation, and interquartile range of units sold per week. Show that Chebyshev's theorem holds for the data and determine how accurate the empirical rules are.
4. The Excel file *Atlanta Airline Data* provides arrival and taxi-in time statistics for one day at Atlanta Hartsfield International airport. Find the average and standard deviation of the difference between the scheduled and actual arrival times and the taxi-in time to the gate. Compute the z-scores for each of these variables.
5. Data obtained from a county auditor in the Excel file *Home Market Value* provides information about the age, square footage, and current market value of houses along one street in a particular subdivision.
   a. Considering these data as a sample of homeowners on this street, compute the mean, variance, and standard deviation for each of these variables using formulas (4.2), (4.5), and (4.7). Verify your calculations using the appropriate Excel function.
   b. Compute the coefficient of variation for each variable. Which has the least and greatest relative dispersion?
6. Find 30 days of stock prices for three companies in different industries. The average stock prices should have a wide range of values. Using the data, compute and interpret the coefficient of variation.
7. Compute descriptive statistics for liberal arts colleges and research universities in the Excel file *Colleges and Universities*. Compare the two types of colleges. What can you conclude?
8. Use the *Descriptive Statistics* tool to summarize the mean, median, variance, and standard deviation of the prices of shares in the Excel file *Coffee Shares Data.*
9. The worksheet *Data* in the Excel file *Airport Service Times* lists a large sample of the times in seconds to process customers at a ticket counter. The second worksheet shows a frequency distribution and histogram of the data.
   a. Summarize the data using the *Descriptive Statistics* tool. What can you say about the shape of the distribution of times?
   b. Find the 90th percentile.
   c. How might the airline use these results to manage its ticketing counter operations?

10. The data in the Excel file *Church Contributions* were reported on annual giving for a church. Estimate the mean and standard deviation of the annual contributions of all parishioners by implementing formulas (4.13) and (4.15) on a spreadsheet, assuming these data represent the entire population of parishioners. Second, estimate the mean contribution of families with children in the parish school. How does this compare with all parishioners?

11. The average monthly wages and standard deviations for the two garments manufacturing factories X and Yare given below:

- Factory X: the average monthly wage is $4600, the standard deviation of the wage is $500, and the number of wage-earners is100
- Factory Y: the average monthly wage is $4900, standard deviation is $400, and the number of wage-earners is 80

a. Which factory pays the larger amount as monthly wages?

b. Which factory shows greater variability in the distribution of wages?

12. Consider the Excel file *Mobiles Usage*, which shows the number of people using different kinds of mobile phones in the northern region. Find the proportion of BlackBerry and Android usage in that region.

13. In the Excel file *Bicycle Inventory*, find the proportion of bicycle models that sell for less than $200.

14. In the *Sales Transactions* database, find the proportion of customers who used PayPal and the proportion of customers who used credit cards. Also, find the proportion that purchased a book and the proportion that purchased a DVD.

15. In the Excel file *Economic Poll*, find the proportions of each categorical variable.

16. In the Excel file *Facebook Survey*, use a PivotTable to find the average and standard deviation of hours online/week and number of friends for females and males in the sample.

17. In the Excel file *Cell Phone Survey*, use PivotTables to find the average for each of the numerical variables for different cell phone carriers and gender of respondents.

18. Using PivotTables, find the average and standard deviation of sales in the *Sales Transactions* database. Also, find the average sales by source (Web or e-mail). Do you think this information could be useful in advertising? Explain how and why or why not.

19. For the Excel file *Travel Expenses*, use a PivotTable to find the average and standard deviation of expenses for each sales rep.

20. Using PivotTables, compute the mean and standard deviation for each metric by year in the Excel file *Freshman College Data*. Are any differences apparent from year to year?

21. The Excel file *Freshman College Data* shows data for 4 years at a large urban university. Use PivotTables to examine differences in student high school performance and first-year retention among different colleges at this university. What conclusions do you reach?

22. The Excel file *Cell Phone Survey* reports opinions of a sample of consumers regarding the signal strength, value for the dollar, and customer service for their cell phone carriers. Use PivotTables to find the following:

a. the average signal strength by type of carrier

b. average value for the dollar by type of carrier and usage level

c. variance of perception of customer service by carrier and gender

What conclusions might you reach from this information?

23. Call centers have high turnover rates because of the stressful environment. The national average is approximately 50%. The director of human resources for a large bank has compiled data about 70 former employees at one of the bank's call centers (see the Excel file *Call Center Data*). Use PivotTables to find these statistics:

a. the average length of service for males and females in the sample

b. the average length of service for individuals with and without a college degree

c. the average length of service for males and females with and without prior call center experience

24. In the Excel file *Weddings*, determine the correlation between the wedding costs and attendance.

25. For the data in the Excel file *Rin's Gym*, find the covariances and correlations among height, weight, and BMI calculation.

26. For the Excel file *Test Scores and Sales* made by nine salesmen during the past year, compute the coefficient of correlation between the test scores and sales using Excel's CORREL function.

27. The Excel file *Beverage Sales* lists a sample of weekday sales at a convenience store, along with the daily high temperature. Compute the covariance and correlation between temperature and sales.

28. For the Excel file *Credit Risk Data*, compute the correlation between age and months employed, age and combined checking and savings account balance, and the number of months as a customer and amount of money in the bank. Interpret your results.

29. In the Excel file *Call Center Data*, how strongly is length of service correlated with starting age?

30. A national homebuilder builds single-family homes and condominium-style townhouses. The Excel file *House Sales* provides information on the selling price, lot cost, type of home, and region of the country (M = Midwest, S = South) for closings during 1 month. Use PivotTables to find the average selling price and lot cost for each type of home in each region of the market. What conclusions might you reach from this information?

31. The Excel file *Auto Survey* contains a sample of data about vehicles owned, whether they were purchased new or used, and other types of data. Use the *Descriptive Statistics* tool to summarize the numerical data, find the correlations among each of the numerical variables, and construct PivotTables to find the average miles/gallon for each type of vehicle, and also the average miles/gallon and average age for each type of new and used vehicle. Summarize the observations that you can make from these results.

32. Compute the z-scores for the data in the Excel file *Airport Service Times*. How many observations fall farther than three standard deviations from the mean? Would you consider these as outliers? Why or why not?

33. Use the *Manufacturing Measurements* data to compute sample averages, assuming that each row in the data file represents a sample from the manufacturing process. Plot the sample averages on a line chart, add the control limits, and interpret your results.

34. Find the mean and variance of a deck of 52 cards, where an ace is counted as 11 and a picture card as 10. Construct a frequency distribution and histogram of the card values. Shuffle the deck and deal two samples of 20 cards (starting with a full deck each time); compute the mean and variance and construct a histogram. How does the sample data differ from the population data? Repeat this experiment for samples of 5 cards and summarize your conclusions.

35. Examine the z-scores you computed in Problem 4 for the *Atlanta Airline Data.* Do they suggest any outliers in the data?

36. In the Excel file *Weddings*, find the averages and median wedding cost and the sample standard deviation. What would you tell a newly engaged couple about what cost to expect? Consider the effect of possible outliers in the data.

37. A producer of computer-aided design software for the aerospace industry receives numerous calls for technical support. Tracking software is used to monitor response and resolution times. In addition, the company surveys customers who request support using the following scale:

0—did not exceed expectations
1—marginally met expectations
2—met expectations
3—exceeded expectations
4—greatly exceeded expectations

The questions are as follows:

Q1: Did the support representative explain the process for resolving your problem?
Q2: Did the support representative keep you informed about the status of progress in resolving your problem?
Q3: Was the support representative courteous and professional?
Q4: Was your problem resolved?
Q5: Was your problem resolved in an acceptable amount of time?
Q6: Overall, how did you find the service provided by our technical support department?

A final question asks the customer to rate the overall quality of the product using this scale:

0—very poor
1—poor
2—good
3—very good
4—excellent

A sample of survey responses and associated resolution and response data are provided in the Excel

file *Customer Support Survey*. Use whatever Excel charts and descriptive statistics you deem appropriate to convey the information in these sample data and write a report to the manager explaining your findings and conclusions.

38. A Midwest pharmaceutical company manufactures individual syringes with a self-contained, single dose of an injectable drug.[7] In the manufacturing process, sterile liquid drug is poured into glass syringes and sealed with a rubber stopper. The remaining stage involves insertion of the cartridge into plastic syringes and the electrical "tacking" of the containment cap at a precisely determined length of the syringe. A cap that is tacked at a shorter-than-desired length (less than 4.920 inches) leads to pressure on the cartridge stopper and, hence, partial or complete activation of the syringe. Such syringes must then be scrapped. If the cap is tacked at a longer-than-desired length (4.980 inches or longer), the tacking is incomplete or inadequate, which can lead to cap loss and a potential cartridge loss in shipment and handling. Such syringes can be reworked manually to attach the cap at a lower position. However, this process requires a 100% inspection of the tacked syringes and results in increased cost for the items. This final production step seemed to be producing more and more scrap and reworked syringes over successive weeks.

The Excel file *Syringe Samples* provides samples taken every 15 minutes from the manufacturing process. Develop control limits using the data and use statistical thinking ideas to draw conclusions.

Case: Drout Advertising Research Project

The background for this case was introduced in Chapter 1. This is a continuation of the case in Chapter 3. For this part of the case, summarize the numerical data using descriptive statistics measures, find proportions for categorical variables, examine correlations, and use PivotTables as appropriate to compare average values. Write up your findings in a formal document, or add your findings to the report you completed for the case in Chapter 3 at the discretion of your instructor.

Case: Performance Lawn Equipment

Elizabeth Burke wants some detailed statistical information about much of the data in the PLE database. In particular, she wants to know the following:

- **a.** the mean satisfaction ratings and standard deviations by year and region in the worksheets *Dealer Satisfaction* and *End-User Satisfaction*
- **b.** a descriptive statistical summary for the *2012 customer survey data*
- **c.** how the response times differ in each quarter of the worksheet *Response Time*
- **d.** how defects after delivery (worksheet *Defects after Delivery*) have changed over these 5 years
- **e.** how sales of mowers and tractors compare with industry totals and how strongly monthly product sales are correlated with industry sales

Perform these analyses and summarize your results in a written report to Ms. Burke.

[7]Based on LeRoy A. Franklin and Samar N. Mukherjee, "An SPC Case Study on Stabilizing Syringe Lengths," *Quality Engineering* 12, 1 (1999–2000): 65–71.