Part I

Getting Started with Business Statistics

web extras

Visit www.dummies.com for great Dummies content online.

In this part...

- Use histograms to provide a visual of the distribution of elements in a data set. A histogram can show which values occur most frequently, the smallest and largest values, how spread out these values are.
- Create graphs that reflect non-numerical data, such as colors, flavors, brand names, and so on. Graphs are used where numerical measures are difficult or impossible to compute.
- Identify the center of a data set by using the mean (the average), median (the middle), and mode (the most commonly occurring value). These are known as the measures of central tendencies.
- Use formulas for computing covariance and correlation for both samples and populations; a scatter plot is used to show the relationship (if there is one) between two variables.

Chapter 1

The Art and Science of Business Statistics

In This Chapter

- Looking at the key properties of data
- Understanding probability's role in business
- Sampling distributions
- Drawing conclusions based on results

This chapter provides a brief introduction to the concepts that are covered throughout the book. I introduce several important techniques that allow you to measure and analyze the statistical properties of real-world variables, such as stock prices, interest rates, corporate profits, and so on.

Statistical analysis is widely used in all business disciplines. For example, marketing researchers analyze consumer spending patterns in order to properly plan new advertising campaigns. Organizations use management consulting to determine how efficiently resources are being used. Manufacturers use quality control methods to ensure the consistency of the products they are producing. These types of business applications and many others are heavily based on statistical analysis.

Financial institutions use statistics for a wide variety of applications. For example, a pension fund may use statistics to identify the types of securities that it should hold in its investment portfolio. A hedge fund may use statistics to identify profitable trading opportunities. An investment bank may forecast the future state of the economy in order to determine which new assets it should hold in its own portfolio.

Whereas statistics is a quantitative discipline, the ultimate objective of statistical analysis is to explain real-world events. This means that in addition to the rigorous application of statistical methods, there is always a great deal of room for judgment. As a result, you can think of statistical analysis as both a science and an art; the art comes from choosing the appropriate statistical technique for a given situation and correctly interpreting the results.

Representing the Key Properties of Data

The word *data* refers to a collection of *quantitative* (numerical) or *qualitative* (non-numerical) values. Quantitative data may consist of prices, profits, sales, or any variable that can be measured on a numerical scale. Qualitative data may consist of colors, brand names, geographic locations, and so on. Most of the data encountered in business applications are quantitative.

The word *data* is actually the plural of *datum;* datum refers to a single value, while data refers to a collection of values.

You can analyze data with graphical techniques or numerical measures. I explore both options in the following sections.

Analyzing data with graphs

Graphs are a visual representation of a data set, making it easy to see patterns and other details. Deciding which type of graph to use depends on the type of data you're trying to analyze. Here are some of the more common types of graphs used in business statistics:

- **Histograms:** A histogram shows the distribution of data among different intervals or categories, using a series of vertical bars.
- **Line graphs:** A line graph shows how a variable changes over time.
- **Pie charts:** A pie chart shows how data is distributed between different categories, illustrated as a series of slices taken from a pie.
- **Scatter plots (scatter diagrams):** A scatter plot shows the relationship between two variables as a series of points. The pattern of the points indicates how closely related the two variables are.

Histograms

You can use a histogram with either quantitative or qualitative data. It's designed to show how a variable is distributed among different categories. For example, suppose that a marketing firm surveys 100 consumers to determine their favorite color. The responses are

Red:	23
Blue:	44
Yellow:	12
Green:	21

The results can be illustrated with a histogram, with each color in a single category. The heights of the bars indicate the number of responses for each color, making it easy to see which colors are the most popular (see Figure 1-1).

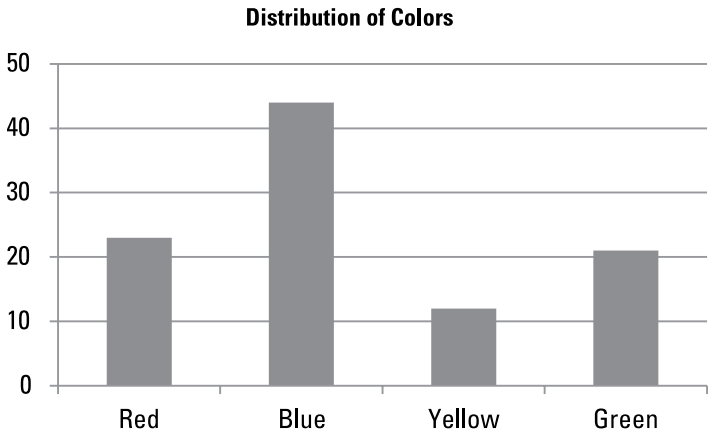

Figure 1-1: A histogram for preferred colors.

Based on the histogram, you can see at a glance that blue is the most popular choice, while yellow is the least popular choice.

Line graphs

You can use a line graph with quantitative data. It shows the values of a variable over a given interval of time. For example, Figure 1-2 shows the daily price of gold between April 14, 2013 and June 2, 2013:

Figure 1-2: A line graph of gold prices.

With a line graph, it's easy to see trends or patterns in a data set. In this example, the price of gold rose steadily throughout late April into mid-May before falling back in late May and then recovering somewhat at the end of the month. These types of graphs may be used by investors to identify which assets are likely to rise in the future based on their past performance.

Pie charts

Use a pie chart with quantitative or qualitative data to show the distribution of the data among different categories. For example, suppose that a chain of coffee shops wants to analyze its sales by coffee style. The styles that the chain sells are French Roast, Breakfast Blend, Brazilian Rainforest, Jamaica Blue Mountain, and Espresso. Figure 1-3 shows the proportion of sales for each style.

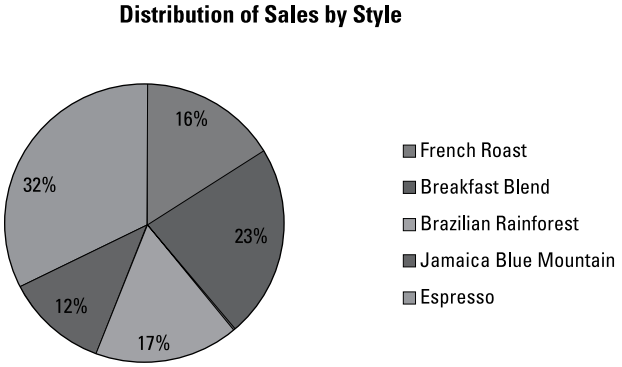

Illustration by Wiley, Composition Services Graphics

Figure 1-3: A pie chart for coffee sales.

The chart shows that Espresso is the chain's best-selling style, while Jamaica Blue Mountain accounts for the smallest percentage of the chain's sales.

Scatter plots

A scatter plot is designed to show the relationship between two quantitative variables. For example, Figure 1-4 shows the relationship between a corporation's sales and profits over the past 20 years.

Each point on the scatter plot represents profit and sales for a single year. The pattern of the points shows that higher levels of sales tend to be matched by higher levels of profits, and vice versa. This is called a positive *trend* in the data.

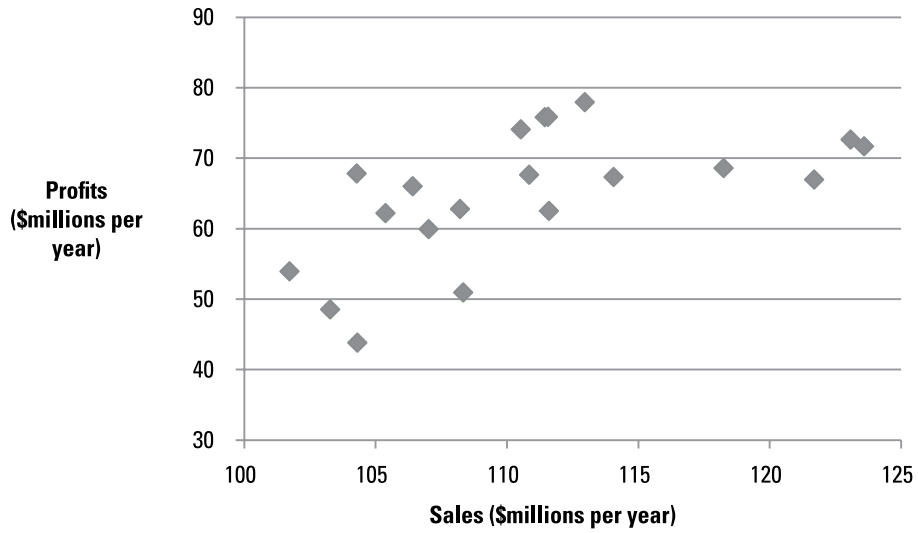

Illustration by Wiley, Composition Services Graphics

Figure 1-4: A scatter plot showing sales and profits.

Defining properties and relationships with numerical measures

A *numerical measure* is a value that describes a key property of a data set. For example, to determine whether the residents of one city tend to be older than the residents in another city, you can compute and compare the average or *mean* age of the residents of each city.

Some of the most important properties of interest in a data set are the *center* of the data and the *spread* among the observations. I describe these properties in the following sections.

Finding the center of the data

To identify the center of a data set, you use measures that are known as *measures of central tendency;* the most important of these are the mean, median, and mode.

The *mean* represents the average value in a data set, while the *median* represents the midpoint. The median is a value that separates the data into two equal halves; half of the elements in the data set are *less than* the median, and the remaining half are *greater than* the median. The *mode* is the most commonly occurring value in the data set.

The mean is the most widely used measure of central tendency, but it can give deceptive results if the data contain any unusually large or small values, known as *outliers*. In this case, the median provides a more representative measure of the center of the data. For example, median household income is usually reported by government agencies instead of mean household income. This is because mean household income is inflated by the presence of a small number of extremely wealthy households. As a result, median household income is thought to be a better measure of how standards of living are changing over time.

The mode can be used for either quantitative or qualitative data. For example, it could be used to determine the most common number of years of education among the employees of a firm. It could also be used to determine the most popular flavor sold by a soft drink manufacturer.

Measuring the spread of the data

Measures of dispersion identify how spread out a data set is, relative to the center. This provides a way of determining if the members of a data set tend to be very close to each other or if they tend to be widely scattered. Some of the most important measures of dispersion are

- Variance
- Standard deviation
- Percentiles
- Quartiles
- Interquartile range (IQR)

The *variance* is a measure of the average squared difference between the elements of a data set and the mean. The larger the variance, the more "spread out" the data is. Variance is often used as a measure of risk in business applications; for example, it can be used to show how much uncertainty there is over the returns on a stock.

The s*tandard deviation* is the square root of the variance, and is more commonly used than the variance (since the variance is expressed in squared units). For example, the variance of a series of gas prices is measured in squared dollars, which is difficult to interpret. The corresponding standard deviation is measured in dollars, which is much more intuitively clear.

Percentiles divide a data set into 100 equal parts, each consisting of 1 percent of the total. For example, if a student's score on a standardized exam is in the 80th percentile, then the student outscored 80 percent of the other students who took the exam. A *quartile* is a special type of percentile; it divides a data

set into four equal parts, each consisting of 25 percent of the total. The first quartile is the 25th percentile of a data set, the second quartile is the 50th percentile, and the third quartile is the 75th percentile. The *interquartile range* identifies the middle 50 percent of the observations in a data set; it equals the difference between the third and the first quartiles.

Determining the relationship between two variables

For some applications, you need to understand the relationship between two variables. For example, if an investor wants to understand the risk of a portfolio of stocks, it's essential to properly measure how closely the returns on the stocks track each other. You can determine the relationship between two variables with two measures of *association:* covariance and correlation.

Covariance is used to measure the tendency for two variables to rise above their means or fall below their means at the same time. For example, suppose that a bioengineering company finds that increasing research and development expenditures typically leads to an increase in the development of new patents. In this case, R&D spending and new patents would have a positive covariance. If the same company finds that rising labor costs typically reduce corporate profits, then labor costs and profits would have a negative covariance. If the company finds that profits are not related to the average daily temperature, then these two variables will have a covariance that is very close to zero.

Correlation is a closely related measure. It's defined as a value between –1 and 1, so interpreting the correlation is easier than the covariance. For example, a correlation of 0.9 between two variables would indicate a very strong positive relationship, whereas a correlation of 0.2 would indicate a fairly weak but positive relationship. A correlation of –0.8 would indicate a very strong negative relationship; a correlation of –0.3 would indicate a weak negative relationship. A correlation of 0 would show that two variables are *independent* (that is, unrelated).

Probability: The Foundation of All Statistical Analysis

Probability theory provides a mathematical framework for measuring uncertainty. This area is important for business applications since all results from the field of statistics are ultimately based on probability theory. Understanding probability theory provides fundamental insights into all the statistical methods used in this book.

Probability is heavily based on the notion of *sets.* A set is a collection of objects. These objects may be numbers, colors, flavors, and so on. This chapter focuses on sets of numbers that may represent prices, rates of return, and so forth. Several mathematical operations may be applied to sets — union, intersection, and complement, for example.

The union of two sets is a new set that contains all the elements in the original two sets. The intersection of two sets is a set that contains only the elements contained in *both* of the two original sets (if any.) The complement of a set is a set containing elements that are *not* in the original set. For example, the complement of the set of black cards in a standard deck is the set containing all red cards.

Probability theory is based on a model of how random outcomes are generated, known as a *random experiment.* Outcomes are generated in such a way that all *possible* outcomes are known in advance, but the *actual* outcome isn't known.

The following rules help you determine the probability of specific outcomes occurring:

- The addition rule
- The multiplication rule
- The complement rule

You use the addition rule to determine the probability of a union of two sets. The multiplication rule is used to determine the probability of an intersection of two sets. The complement rule is used to identify the probability that the outcome of a random experiment will *not* be an element in a specified set.

Random variables

A *random variable* assigns numerical values to the outcomes of a random experiment. For example, when you flip a coin twice, you're performing a random experiment, since:

- All possible outcomes are known in advance
- The actual outcome isn't known in advance

The experiment consists of two *trials*. On each trial, the outcome must be a "head" or a "tail."

Assume that a random variable *X* is defined as the number of "heads" that turn up during the course of this experiment. *X* assigns values to the outcomes of this experiment as follows:

Outcome	***X***
{TT}	0
{HT, TH}	1
{HH}	2

T represents a tail on a single flip

H represents a head on a single flip

TT represents two consecutive tails

HT represents a head followed by a tail

TH represents a tail followed by a head

HH represents two consecutive heads

X assigns a value of 0 to the outcome TT because no heads turned up. *X* assigns a value of 1 to both HT and TH because one head turned up in each case. Similarly, *X* assigns a value of 2 to HH because two heads turned up.

Probability distributions

A *probability distribution* is a formula or a table used to assign probabilities to each possible value of a random variable *X.* A probability distribution may be *discrete,* which means that *X* can assume one of a finite (countable) number of values, or *continuous,* in which case *X* can assume one of an infinite (uncountable) number of different values.

For the coin-flipping experiment from the previous section, the probability distribution of *X* could be a simple table that shows the probability of each possible value of *X*, written as *P*(*X*):

X	***P(X)***
0	0.25
1	0.50
2	0.25

The probability that $X = 0$ (that no heads turn up) equals 0.25 because this experiment has four equally likely outcomes: HH, HT, TH, and TT and in only one of those cases will there be no heads. You compute the other probabilities in a similar manner.

Discrete probability distributions

Several specialized discrete probability distributions are useful for specific applications. For business applications, three frequently used discrete distributions are:

- Binomial
- Geometric
- Poisson

You use the *binomial distribution* to compute probabilities for a process where only one of two possible outcomes may occur on each trial. The *geometric distribution* is related to the binomial distribution; you use the geometric distribution to determine the probability that a specified number of trials will take place before the first success occurs. You can use the *Poisson distribution* to measure the probability that a given number of events will occur during a given time frame.

Continuous probability distributions

Many continuous distributions may be used for business applications; two of the most widely used are:

- Uniform
- Normal

The *uniform distribution* is useful because it represents variables that are evenly distributed over a given interval. For example, if the length of time until the next defective part arrives on an assembly line is equally likely to be any value between one and ten minutes, then you may use the uniform distribution to compute probabilities for the time until the next defective part arrives.

The *normal distribution* is useful for a wide array of applications in many disciplines. In business applications, variables such as stock returns are often assumed to follow the normal distribution. The normal distribution is characterized by a *bell-shaped curve,* and areas under this curve represent probabilities. The bell-shaped curve is shown in Figure 1-5.

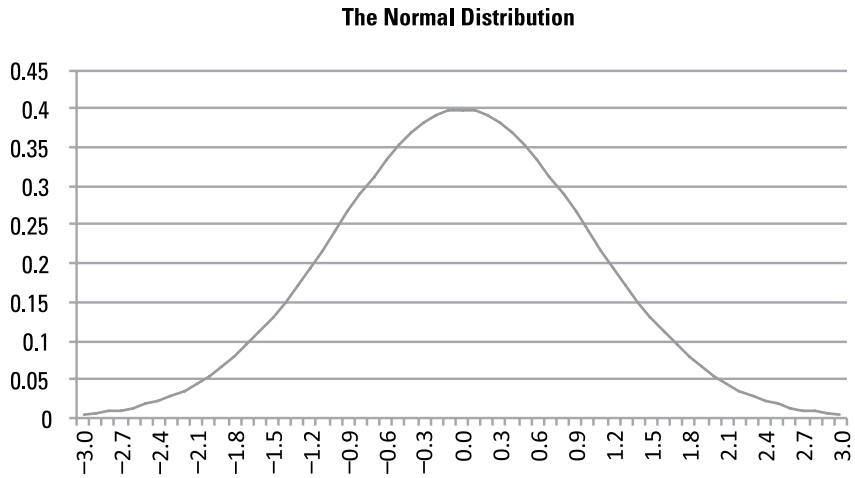

Illustration by Wiley, Composition Services Graphics

Figure 1-5: The bell-shaped curve of the normal distribution.

The normal distribution has many convenient statistical properties that make it a popular choice for statistical modeling. One of these properties is known as *symmetry,* the idea that the probabilities of values below the mean are matched by the probabilities of values that are equally far above the mean.

Using Sampling Techniques and Sampling Distributions

Sampling is a branch of statistics in which the properties of a *population* are estimated from *samples.* A population is a collection of data that someone has an interest in studying. A sample is a selection of data randomly chosen from a population.

For example, if a university is interested in analyzing the distribution of grade point averages (GPAs) among its MBA students, the population of interest would be the GPAs of every MBA student at the university; a sample would consist of the GPAs of a set of randomly chosen MBA students.

Several approaches can be used for choosing samples; a sample is a *subset* of the underlying population.

A *statistic* is a summary measure of a sample, while a *parameter* is a summary measure of a population. The properties of a statistic can be determined with a *sampling distribution* — a special type of probability distribution that describes the properties of a statistic.

The *central limit theorem* (CLT) gives the conditions under which the mean of a sample follows the normal distribution:

- The underlying population is normally distributed.
- The sample size is "large" (at least 30).

A detailed discussion of the central limit theorem can be found at `http://en.wikipedia.org/wiki/Central_limit_theorem`.

Statistical Inference: Drawing Conclusions from Data

Statistical inference refers to the process of drawing conclusions about a population from randomly chosen samples. In the following sections, I discuss two techniques used for statistical inference: confidence intervals and hypothesis testing.

Confidence intervals

A *confidence interval* is a range of values that's expected to contain the value of a population parameter with a specified level of confidence (such as 90 percent, 95 percent, 99 percent, and so on). For example, you can construct a confidence interval for the population mean by following these steps:

1. **Estimate the value of the population mean by calculating the mean of a randomly chosen sample (known as the sample mean).**
2. **Calculate the lower limit of the confidence interval by subtracting a *margin of error* from the sample mean.**
3. **Calculate the upper limit of the confidence interval by adding the same margin of error to the sample mean.**

The margin of error depends on the size of the sample used to construct the confidence interval, whether the population standard deviation is known, and the level of confidence chosen.

The resulting interval is known as a confidence interval. A confidence interval is constructed with a specified level of probability. For example, suppose you draw a sample of stocks from a portfolio, and you construct a 95 percent confidence interval for the mean return of the stocks in the entire portfolio:

(lower limit, upper limit) = (0.02, 0.08)

The returns on the entire portfolio are the population of interest. The mean return in each sample drawn is an *estimate* of the population mean. The sample mean will be slightly different each time a new sample is drawn, as will the confidence interval. If this process is repeated 100 times, 95 of the resulting confidence intervals will contain the true population mean.

Hypothesis testing

Hypothesis testing is a procedure for using sample data to draw conclusions about the characteristics of the underlying population.

The procedure begins with a statement, known as the *null hypothesis.* The null hypothesis is assumed to be true unless strong evidence against it is found. An *alternative hypothesis* — the result accepted if the null hypothesis is rejected — is also stated.

You construct a *test statistic,* and you compare it with a *critical value* (or values) to determine whether the null hypothesis should be rejected. The specific test statistic and critical value(s) depend on which population parameter is being tested, the size of the sample being used, and other factors.

If the test statistic is too extreme (for example, it's too large compared with the critical value[s]) the null hypothesis is rejected in favor of the alternative hypothesis; otherwise, the null hypothesis is *not* rejected.

If the null hypothesis isn't rejected, this doesn't necessarily mean that it's true; it simply means that there is not enough evidence to justify rejecting it.

Hypothesis testing is a general procedure and can be used to draw conclusions about many features of a population, such as its mean, variance, standard deviation, and so on.

Simple regression analysis

Regression analysis uses sample data to estimate the strength and direction of the relationship between two or more variables. *Simple regression analysis* estimates the relationship between a dependent variable (*Y*) and a single independent variable (*X*).

For example, suppose you're interested in analyzing the relationship between the annual returns of the Standard & Poor's (S&P) 500 Index and the annual returns of Apple stock. You can assume that the returns of Apple stock are related to the returns to the S&P 500 because the index is a reflection of the overall strength of the economy. Therefore, the returns of Apple stock are the dependent variable (*Y*) and the returns of the S&P 500 are the independent variable (*X*). You can use regression analysis to measure the numerical relationship between the S&P 500 and Apple stock.

Simple regression analysis is based on the assumption that a linear relationship occurs between *X* and *Y*. A linear relationship takes this form:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Y is the dependent variable, *X* is the independent variable, *m* is the slope, and *b* is the intercept.

The slope tells you how much *Y* changes due to a specific change in *X*; the intercept tells you what the value of *Y* would be if *X* had a value of zero.

The goal of regression analysis is to find a line that best fits or explains the data. The population regression line is written as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

In this equation, Y_i is the dependent variable, X_i is the independent variable, β_0 is the intercept, β_1 is the slope, and ε_i is an error term.

A *sample* regression line, estimated from the data, is written as follows:

$$\hat{Y}_i$$

Here, $\hat{\beta}_0$ is the estimated value of Y_i, $\hat{\beta}_1$ is the estimated value of β_0, and X_i is the estimated value of β_1 and is the independent variable.

The sample regression line shows the estimated relationship between *Y* and *X;* you can use this relationship to determine how much *Y* changes due to a given change in *X*. You can also use it to *forecast* future values of *Y* based on assumed values of *X*.

After estimating the sample regression line, the results are subjected to a series of tests to determine whether the equation is valid. If the equation isn't valid, you reject the results and try a new model.

Multiple regression analysis

With *multiple regression analysis,* you estimate the relationship between a dependent variable (*Y*) and *two or more* independent variables (X_1, X_2, and so on).

For example, suppose that *Y* represents annual salaries (in thousands of dollars) at a corporation. A researcher has reason to believe that the salaries at this corporation depend mainly on the number of years of job experience and the number of years of graduate education for each employee. The researcher may test this idea by running a regression in which salary is the dependent variable (*Y*) and job experience and graduate education are the independent variables (X_1 and X_2, respectively.) The population regression equation in this case would be written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

The sample regression line would be

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$$

Using multiple regression analysis introduces several additional complications compared with simple regression analysis, but you can use it for a much wider range of applications than simple regression analysis.

Forecasting techniques

You can forecast the future values of a variable, using one of several types of models. One approach to forecasting is *time series* models. A time series is a set of data that consists of the values of a single variable observed at different points in time. For example, the daily price of Microsoft stock taken from the past ten years is a time series.

Time series forecasting involves using past values of a variable to forecast future values.

Some forecasting techniques include:

- Trend models
- Moving average models
- Exponential moving average models

A *trend model* is used to estimate the value of a variable as it evolves over time. For example, suppose annual data is used to estimate a trend model that explains the behavior of gasoline prices over time. The price is currently $3.50 per gallon, and you determine that on average, gasoline prices rise by $0.10 per year. A simple trend model that expresses this information would be written as:

$$Y_t = 3.50 + 0.10t + \varepsilon_t$$

In this equation, Y_t represents the estimated gas price at time t, where t represents a specific year. ($t = 0$ represents the present time.) The term 3.50 indicates the current price of gasoline; $0.10t$ indicates that the price of gasoline rises by $0.10 per year. The term ε_t is known as an "error term"; this reflects random fluctuations in the price of gasoline over time.

A *moving average model* shows that the value of a variable evolves over time based on its most recent values. For example, if the price of gasoline over the past three years was:

2010 $3.25

2011 $3.32

2012 $3.42

A three-period moving average model would produce an estimated value of ($3.25 + $3.32 + $3.42) / 3 = $3.33 for 2013.

An *exponential weighted average model* is closely related to a moving average model. The difference is that with an exponential weighted average, older observations aren't given the same "weight" as newer observations. The calculation of an exponential weighted average is more complex, but may give more realistic results.

The appropriate choice of model depends on the properties of the particular time series being used.

Chapter 2

Pictures Tell the Story: Graphical Representations of Data

In This Chapter

- Describing the properties of data with a frequency distribution
- Illustrating frequency distributions with histograms
- Tracking trends with line graphs, pie charts, and scatter diagrams

Much of statistical analysis is based on numerical techniques, such as confidence intervals, hypothesis testing, regression analysis, and so on. (These topics are covered in Chapters 11, 12, and 15, respectively.)

In many cases, these techniques are based on assumptions about the data being used. One way to determine if the data conform to these assumptions is to analyze a graph of the data, as a graph can provide many insights into the properties of a data set. For example, a graph may be used to show:

- How frequently a value occurs in a data set
- The average value of the elements in a data set
- Whether the elements in a data set are increasing or decreasing over time
- Whether the elements in two different data sets are related to each other

Graphs are particularly useful for non-numerical data, such as colors, flavors, brand names, and more, where numerical measures are difficult or impossible to compute.

This chapter explains how to organize data in a convenient form so you can easily analyze it. I introduce charts and graphs — from histograms to line graphs to pie charts and scatter plots — that can help you visualize the most important properties of a data set.

Analyzing the Distribution of Data by Class or Category

To graph *quantitative* (numerical) data, you start by organizing the data into *classes* (also known as *intervals*). For example, suppose the government is conducting a study that measures the salary ranges for employees in the software industry in the United States. Here's one possible set of classes:

$0 to $24,999 per year

$25,000 to $49,999 per year

$50,000 to $74,999 per year

$75,000 to $99,999 per year

$100,000 and more per year

By counting the number of employees that fall into each class, you can easily see how salaries are distributed in the software industry. If you make the data into a graph, you can then easily compare this information with salaries in other industries.

Qualitative (non-numerical) data may be organized into *categories.* For example, suppose that a marketing firm is studying the spending habits of consumers and wants to determine the most popular colors for a new line of watches. In this case, the colors are relevant categories.

What type of graph you use for analyzing a set of data depends on the type of data (quantitative or qualitative) and the type of analysis you are performing. The following sections introduce several important types of graphs.

I also introduce the concept of a *frequency distribution*. This is a list of classes and the number of elements that belong to each class (known as *frequencies*). I cover the steps required to construct a frequency distribution, and I show two related types of distribution: relative frequency distribution and cumulative frequency distribution.

This section covers several widely used types of graphs, including histograms, pie charts, line graphs, and scatter plots. Histograms represent frequency distributions as a series of bars. Pie charts show what proportion of the elements of a data set belongs to various categories. A line graph shows how the value of a variable changes over time. Scatter plots are used to show the *relationship* between two variables.

Frequency distributions for quantitative data

Quantitative data consists of numerical values, such as prices, weights, distances, and so on.

To graphically analyze quantitative data, you first have to organize them into a *frequency distribution* — a table that shows the number of observations that fall into each class within the data set.

For example, suppose that the following values represent the price of gasoline (dollars per gallon) at 20 randomly selected gas stations:

$4.42	$4.34
$4.17	$3.73
$3.92	$3.56
$4.49	$3.65
$3.91	$3.58
$4.46	$4.12
$4.27	$4.21
$3.92	$3.85
$3.57	$4.10
$4.10	$3.63

Now suppose you organize the data into four classes, as follows:

$3.50 to $3.74

$3.75 to $3.99

$4.00 to $4.24

$4.25 to $4.49

Table 2-1 shows the frequency distribution for these.

Table 2-1 Frequency Distribution of Prices for 20 Gas Stations

Gas Prices ($/Gallon)	*Number of Gas Stations*
$3.50–$3.74	6
$3.75–$3.99	4
$4.00–$4.24	5
$4.25–$4.49	5

Table 2-1 shows that the distribution of gas prices among these classes is very nearly equal. Seeing how the prices are distributed with a frequency distribution is much easier than inspecting the raw (original) data, which in this case is a list of 20 gas prices.

When you're constructing a frequency distribution, one of the most important considerations is the *width* of the classes. The class width equals the difference between the largest value that may be included in the class and the smallest. In Table 2-1, the class widths are $0.25. Usually, the class widths will be equal.

Deciding how many classes to use depends on how much data you have and how detailed you need the results to be. For example, if the class width is too large, it can disguise the distribution of values within each class. If the class width is too small, then several classes may contain no elements or very few elements, which makes analyzing the results more cumbersome.

As a rule of thumb, the optimal number of classes in a frequency distribution is between 5 and 15.

Figuring the class width

In the gas station example, each class has a width of $0.25. In general, you can determine the class width by subtracting the smallest value from the largest value and dividing by the total number of desired classes:

$$\text{Class width} = \frac{\text{Largest value in raw data} - \text{Smallest value in raw data}}{\text{Number of classes}}$$

Referring to the raw data (the list of 20 gas prices), you see that the largest price in the sample is $4.49 and the smallest is $3.56. To construct a frequency distribution with four classes, the width of each interval should be

$$\text{Class width} = \frac{\$4.49 - \$3.56}{4} = \frac{\$0.93}{4} = \$0.2325$$

So the class width is equal to approximately $0.25. Although the class width could be kept at $0.2325, using a width of $0.25 is intuitively easier to follow (since prices can't be expressed in quarters of a cent).

When you construct a frequency distribution, remember these key points:

- The classes must not overlap. For example, if the frequency distribution refers to gasoline prices, it would be incorrect to have a class for $1.00 to $2.00 and another class for $2.00 to $3.00, because both contain $2.00. It would be unclear which class contains prices of $2.00.
- The classes must cover all elements in the data set being analyzed.
- Ideally, the classes should have equal widths; otherwise, analyzing the results is much more difficult.
- Class widths should ideally be "round" numbers, such as $0.50, $1.00, $10.00, and so on, compared with numbers such as $0.43, $1.87, and $2.15. These numbers are more difficult to grasp intuitively. For the gas station example, the widths are $0.25, and this is preferable to $0.2325, because $0.2325 isn't a round number.

Observing relative frequency distributions

A frequency distribution shows the number of elements in a data set that belong to each class. In a relative frequency distribution, the value assigned to each class is the *proportion* of the total data set that belongs in the class. For example, suppose that a frequency distribution is based on a sample of 200 supermarkets. It turns out that 50 of these supermarkets charge a price between $8.00 and $8.99 for a pound of coffee. In a relative frequency distribution, the number assigned to this class would be 0.25 (50/200). In other words, that's 25 percent of the total.

Here's a handy formula for calculating the relative frequency of a class:

$$\frac{\text{class frequency}}{n}$$

Class frequency refers to the number of observations in each class; *n* represents the total number of observations in the entire data set. For the supermarket example in this section, the total number of observations is 200.

The relative frequency may be expressed as a proportion (fraction) of the total or as a percentage of the total. See Table 2-2, which gives both types of relative frequency based on the gas station data in Table 2-1.

Table 2-2 Relative Frequencies for Gas Station Prices

Gas Prices ($/ Gallon)	*Number of Gas Stations*	*Relative Frequency (fraction)*	*Relative Frequency (percent)*
$3.50–$3.74	6	6/20 = 0.30	30%
$3.75–$3.99	4	4/20 = 0.20	20%
$4.00–$4.24	5	5/20 = 0.25	25%
$4.25–$4.49	5	5/20 = 0.25	25%

With a sample size of 20 gas stations, the relative frequency of each class equals the actual number of gas stations divided by 20. The result is then expressed as either a fraction or a percentage. For example, you calculate the relative frequency of prices between $3.50 and $3.74 as 6/20 to get 0.30 (30 percent). Similarly, the relative frequency of prices between $3.75 and $3.99 equals 4/20 = 0.20 = 20 percent.

One of the advantages of using a relative frequency distribution is that you can compare data sets that don't necessarily contain an equal number of observations. For example, suppose that a researcher is interested in comparing the distribution of gas prices in New York and Connecticut. Because New York has a much larger population, it also has many more gas stations. The researcher decides to choose 1 percent of the gas stations in New York and 1 percent of the gas stations in Connecticut for the sample. This turns out to be 800 in New York and 200 in Connecticut. The researcher puts together a frequency distribution as shown in Table 2-3.

Table 2-3 Frequency Distribution of Gas Prices in New York and Connecticut

Price	*New York Gas Stations*	*Connecticut Gas Stations*
$3.00–$3.49	210	48
$3.50–$3.99	420	96
$4.00–$4.49	170	56

Based on this frequency distribution, it's awkward to compare the distribution of prices in the two states. By converting this data into a relative frequency distribution, the comparison is greatly simplified, as seen in Table 2-4.

Table 2-4 Relative Frequency Distribution of Gas Prices in New York and Connecticut

Price	*New York Gas Stations*	*Relative Frequency*	*Connecticut Gas Stations*	*Relative Frequency*
$3.00–$3.49	210	210/800 = 0.2625	48	48/200 = 0.2400
$3.50–$3.99	420	420/800 = 0.5250	96	96/200 = 0.4800
$4.00–$4.49	170	170/800 = 0.2125	56	56/200 = 0.2800

The results show that the distribution of gas prices in the two states is nearly identical. Roughly 25 percent of the gas stations in each state charge a price between $3.00 and $3.49; about 50 percent charge a price between $3.50 and $3.99; and about 25 percent charge a price between $4.00 and $4.49.

Frequency distribution for qualitative values

In this section, I use a qualitative data set to illustrate frequency distributions.

Suppose that a data set consists of *qualitative* (non-numerical) values. In this example, consumers were asked to identify their favorite color on a survey. The 20 responses are listed here.

blue	blue	blue	black
black	black	black	black
white	blue	white	blue
red	red	red	red
silver	silver	black	white

In this case, the categories are colors. The frequency distribution of these data is:

Color	*Number of Reponses*
Black	6
Blue	5
Red	4
Silver	2
White	3

Table 2-5 shows the relative frequency distribution.

Table 2-5 Relative Frequency Distribution of Favorite Colors

Color	*Number of Responses*	*Relative Frequency (fraction)*	*Relative Frequency (percent)*
Black	6	6/20 = 0.30	30%
Blue	5	5/20 = 0.25	25%
Red	4	4/20 = 0.20	20%
Silver	2	2/20 = 0.10	10%
White	3	3/20 = 0.15	15%

You can easily see from the table that the most popular choice is black, and the least popular is silver.

Cumulative frequency distributions

Cumulative frequency refers to the total frequency of a given class and all prior classes.

For example, Table 2-6 lists the cumulative frequencies for the gas station data from the earlier section "Frequency distributions for quantitative data."

Table 2-6 Cumulative Frequency of Prices at 20 Gas Stations

Gas Prices ($/Gallon)	*Number of Gas Stations*	*Cumulative Frequency*	*Cumulative Frequency (percent)*
$3.50–$3.74	6	6	30%
$3.75–$3.99	4	6 + 4 = 10	50%
$4.00–$4.24	5	6 + 4 + 5 = 15	75%
$4.25–$4.49	5	6 + 4 + 5 + 5 = 20	100%

To figure out the cumulative frequency of the $3.75 to $3.99 class, you add its class frequency (4) to the frequency of the previous class ($3.50 to $3.74,

which is 6), so 6+4 = 10. This result shows you that ten gas stations' prices are between $3.50 and $3.99. Because 20 gas stations were used in the sample, the percentage of all gas stations with prices between $3.50 and $3.99 is 10/20 or 50 percent of the total.

Histograms: Getting a Picture of Frequency Distributions

You can illustrate a frequency distribution, a relative frequency distribution, or a cumulative frequency with a special type of graph known as a *histogram*. (See the previous section, "Analyzing the Distribution of Data by Class or Category.") With histograms, you list classes or categories on the horizontal axis and frequencies on the vertical axis. A bar represents each class or category.

A histogram's job is to provide a visual of the distribution of elements in a data set. The histogram can show which values in a data set occur most frequently, the smallest and largest values in the data set, how "spread out" these values are, and so on.

Figure 2-1 shows a histogram of the frequency distribution for the gas station prices from the previous section.

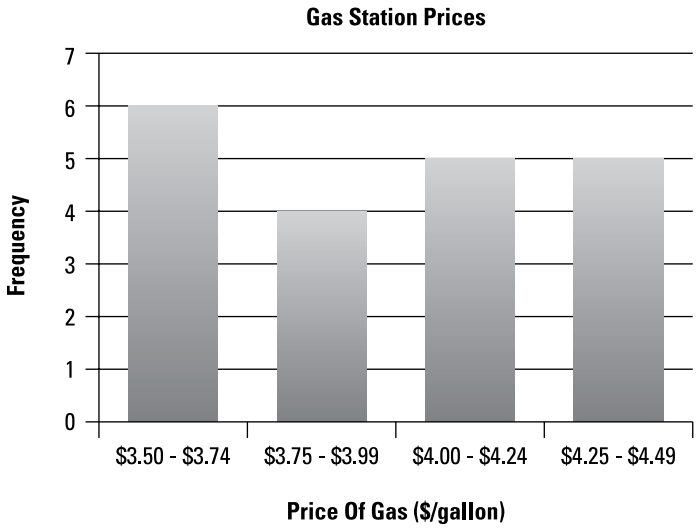

Figure 2-1: Frequency distribution of gas prices.

Figure 2-2 shows the relative frequency distribution.

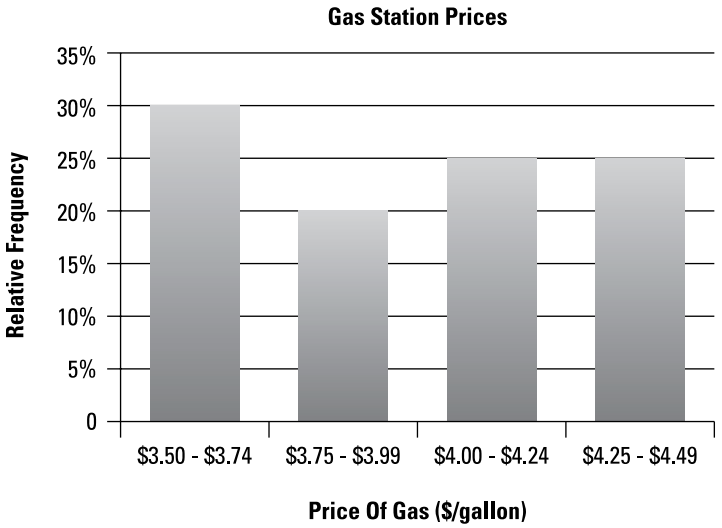

Figure 2-2: Relative frequency distribution of gas station prices.

Figure 2-3 shows the cumulative frequency distribution.

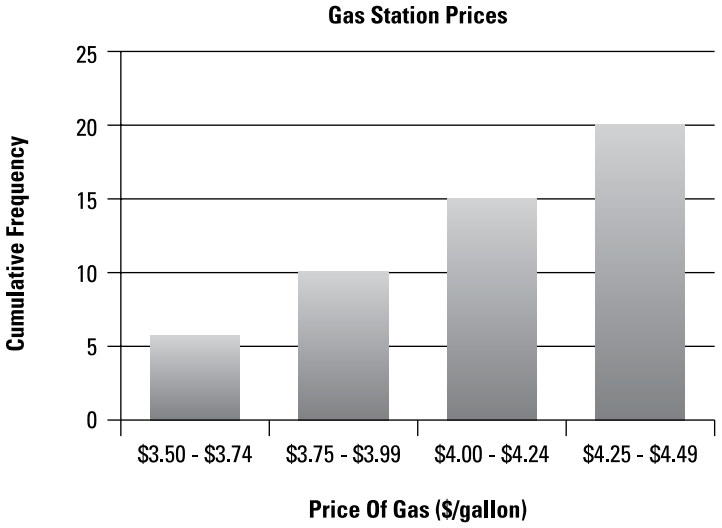

Figure 2-3: Cumulative frequency distribution.

As another example, two restaurants, Pegasus and Orion, each asked 40 customers to estimate how much time they waited to receive their entrees. Figure 2-4 shows the results for the Pegasus survey, and Figure 2-5 shows the results for the Orion survey.

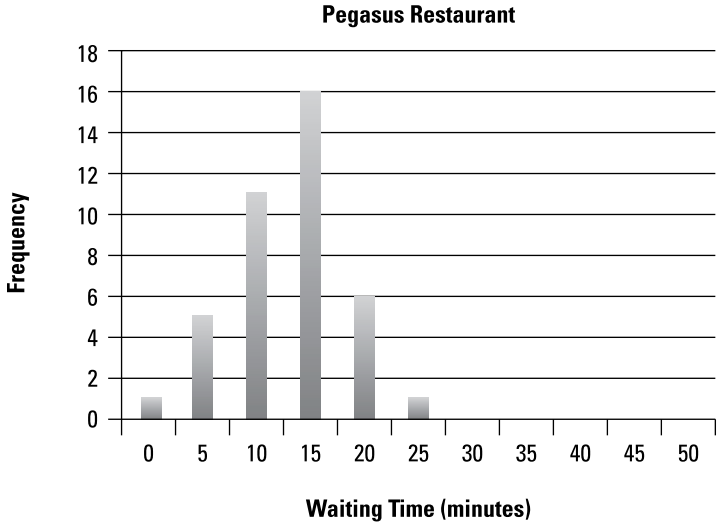

Figure 2-4: Histogram of waiting times at the Pegasus restaurant.

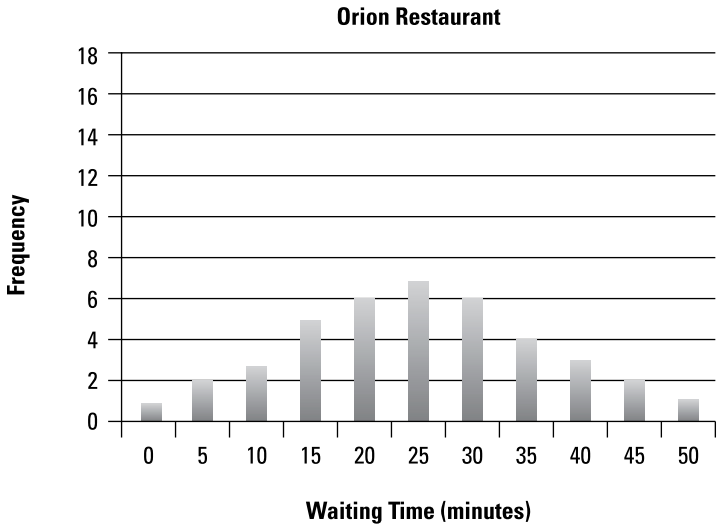

Figure 2-5: Histogram of waiting times at the Orion restaurant.

As you can see, the most common waiting time at Pegasus was 15 to 20 minutes, and at Orion, 25 to 30 minutes. The histograms also show that the waiting times are more spread out at Orion — in other words, the actual waiting time is more uncertain at Orion than at Pegasus.

Checking Out Other Useful Graphs

In addition to histograms, several other types of graphs can illustrate the properties of a data set. This section introduces you to some of the more common types of graphs you're likely to encounter and use.

Line graphs: Showing the values of a data series

A *line graph* is useful for showing how the value of a variable changes over time. With a line graph, the vertical axis represents the value of the variable, and the horizontal axis represents time. Each point on the graph represents the value of the variable at a single point in time, and a line connects the points. This line shows any trends in the data, such as whether the variable increases or decreases over time.

The following shows the price of gold (dollars per ounce) during the first six months of 2012:

Month	***Gold Price ($/Ounce)***
January 2012	$1,652.42
February 2012	$1,723.33
March 2012	$1,676.30
April 2012	$1,646.77
May 2012	$1,567.08
June 2012	$1,602.27

The line chart in Figure 2-6 illustrates how the price of gold changed during this time period, based on the data shown in the table.

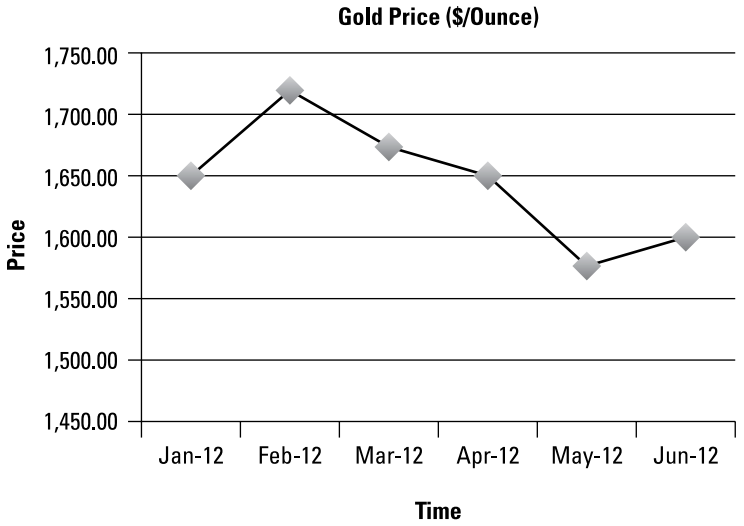

Figure 2-6: Line graph showing how the price of gold fluctuated over six months' time.

Using a line chart to detect patterns in the data is much easier than looking at the original data.

Pie charts: Showing the composition of a data set

A *pie chart* is a circle graph that's divided into slices to represent the distribution of values in a data set. The area of each slice is proportional to the number of values in a given class or category.

For example, suppose a bank has 100 branches throughout the country; the following is the geographical distribution of these branches:

Branch Location	***Number of Branches***
Northeast	44
Northwest	32
Southeast	15
Southwest	9

The pie chart in Figure 2-7 illustrates these results.

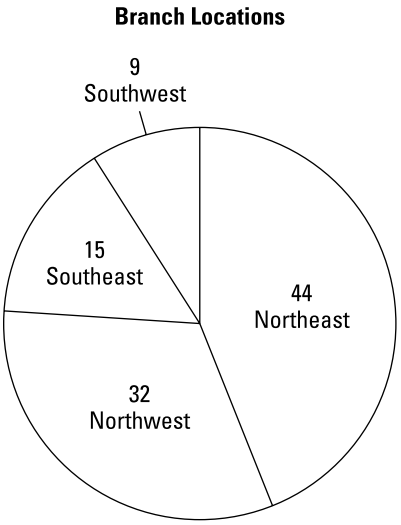

Figure 2-7: Branch location results shown in a pie chart.

The area of each slice in the pie chart indicates the proportional number of branches in each region. With this chart, you can easily see that the majority of the branches are in the northeast, with the fewest in the southwest.

Scatter plots: Showing the relationship between two variables

A *scatter plot* (also known as a *scatter diagram*) shows the relationship between two quantitative (numerical) variables. These variables may be positively related, negatively related, or unrelated:

- **Positively related variables** indicate that

  When one variable increases, the other variable tends to increase.

  When one variable decreases, the other variable tends to decrease.

- **Negatively related variables** indicate that

  When one variable increases or decreases, the other variable tends to do the opposite.

Unrelated variables indicate that

No relationship is seen between the changes in the two variables.

The scatter diagram in Figure 2-8 shows the relationship between the monthly returns to Microsoft stock and the Standard & Poor's (S&P) 500 Index from 2008 to 2012:

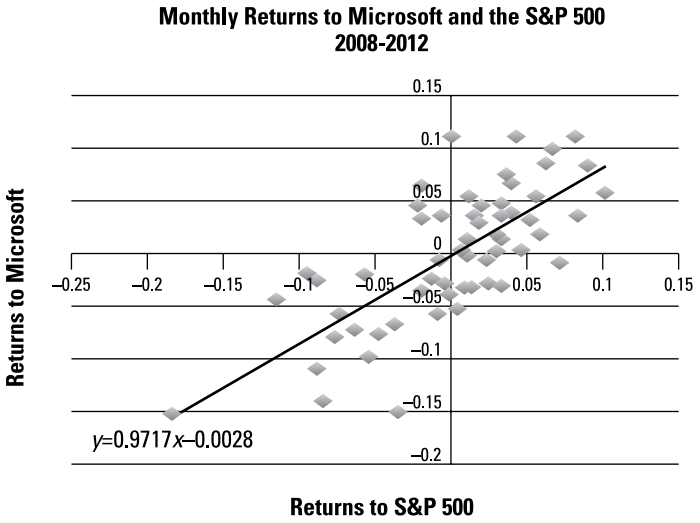

Figure 2-8: Scatter diagram showing relationship of monthly returns.

Each point on the graph represents the return to Microsoft stock and the return to the S&P 500 Index during a single month. The general direction of these points is from the lower-left corner of the graph to the upper-right corner, indicating that the two variables have a positive relationship.

The graph contains a *trend line,* which is a straight line designed to come as close as possible to all the points in the diagram. If two variables are positively related, the trend line has a positive slope; similarly, if two variables are negatively related, the trend line has a negative slope. If two variables are unrelated to each other, the trend line has a zero slope (that is, the trend line will be *flat*).

In the case of Microsoft and the S&P 500 Index, the equation of the trend line is

$$y = -0.0028 + 0.917x$$

Even more types of graphs

In addition to the graphs covered in this chapter — histograms, line graphs, pie charts, and scatter plots — there are many other types of graphs that you can use to analyze statistical data. Many of these have interesting names, such as stemplots, box-and-whisker diagrams, and ogives. These types may be used as alternatives to numerical methods to identify the distribution of elements within a data set, the relationship between the mean and the median of a data set, and several other factors.

In this equation, –0.0028 is the *intercept* (where the trend line crosses the vertical axis) and the *slope* is 0.917 (how much y changes due to a change in x).

Because the slope of the trend line is positive (0.917), the relationship between the returns to Microsoft stock and the S&P 500 Index is positive. The value of the slope also shows that each 1 percent increase in the returns to the S&P 500 Index increases the return to Microsoft by 0.917 percent, and that each 1 percent decrease in the returns to the S&P 500 Index decreases the return to Microsoft by 0.917 percent.

Chapter 3

Finding a Happy Medium: Identifying the Center of a Data Set

In This Chapter

- Computing the mean, median, and mode of a data set
- Noting the specific characteristics of the mean, median, and mode

The center of a data set (sample or population) provides useful information in many business applications. For example, it may be extremely important for a marketing firm to determine the average age of the customers who buy a specific product. Understanding the average household income of a company's customers would also be extremely useful in determining which types of new products to introduce. The portfolio manager at a pension fund is extremely interested in knowing the average rate of return of various stocks that he may be thinking about buying.

This chapter focuses on the techniques you use to find the center of a data set. There are several different ways to define the center: the average value, the middle value, the most frequently occurring value, and so on. Three of the most important measures of the center, formally known as measures of central tendency, are the mean, median, and mode.

The mean is the most commonly used measure of the center; it has the advantage of being easy to compute and interpret. In statistics, the word *mean* is used interchangeably with *average.*

The median and mode are mainly used in situations where the mean is likely to give misleading results. This can happen if the data set contains any extremely large or small values, known as *outliers.*

An *outlier* is a value that's significantly different from the other elements in a data set. Outliers may have a dramatic impact on the accuracy of your calculations.

The *median* is the middle value of a data set (just like a median divides a highway into two equal halves). The *mode* is the most frequently occurring value in a data set. Each of these measures has its own unique set of advantages and disadvantages.

Looking at Methods for Finding the Mean

You can calculate the mean of a data set in several ways; the appropriate choice depends on the type of data and the application. This section explains how to find the three most common types of mean.

Arithmetic mean

The *arithmetic mean* is what most people think of when they hear the word *mean.* This type of mean is the easiest to calculate; it's the sum of the elements in a data set divided by the number of elements.

You use different formulas for computing the arithmetic mean for a *population* and a *sample*. A population is a collection of data that you're interested in studying; a sample is a selection chosen from a population. For example, if a government is interested in the distribution of household incomes, the population of interest would be the incomes of every household. A sample would be a set of incomes for households randomly chosen from the population.

Calculating the sample arithmetic mean

The formula for finding the sample arithmetic mean is

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

The key terms in this formula are:

- $\bar{X}$ (pronounced "X bar") = the sample mean
- n = the number of elements in the sample
- i = an *index,* which assigns a number to each sample element, ranging from 1 to n
- X_i = a single element in the sample
- Σ = the uppercase Greek letter sigma, known as the *summation operator,* which indicates that a sum is being computed

The summation operator is shorthand notation for adding a set of numbers. For example, if a data set contains five elements, the summation operator tells you to perform the following calculations:

$$\sum_{i=1}^{5} X_i = X_1 + X_2 + X_3 + X_4 + X_5$$

Each of the *X*s in this formula is *indexed* by a number ranging from 1 to *n,* where *n* is the size of the data set. In this example, *n* is 5.

Suppose an investor wants to compute the arithmetic mean return of the stock of Omega Airlines, Inc. He or she takes a sample of annual returns — the period from 2008 to 2012.

Year	***Omega Airlines Annual Return (percent)***
2008	2
2009	–1
2010	3
2011	5
2012	1

To find the arithmetic mean, follow these steps:

1. **Assign an index to each return in the sample.**

   $X_1 = 2, X_2 = -1, X_3 = 3, X_4 = 5, X_5 = 1$

   Here, X_1 represents the return in 2008; X_2 is the return in 2009, and so on.

2. **Compute the sum of the returns:**

   $$\sum_{i=1}^{5} X_i = X_1 + X_2 + X_3 + X_4 + X_5 = 2 - 1 + 3 + 5 + 1 = 10$$

3. **Divide the sum of the returns by the number of returns in the sample:**

   $$\bar{X} = \frac{\sum_{i=1}^{5} X_i}{5} = \frac{10}{5} = 2$$

This result shows that the average return of this stock is 2 percent per year.

Calculating the population arithmetic mean

When you calculate the arithmetic mean of a population, the calculation is the same as for arithmetic mean of a sample, but the notation is slightly different. Here's the formula for computing the arithmetic mean of a population:

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n}$$

The new term in this formula is μ, the lowercase Greek letter mu, which replaces $\bar{X}$ from the sample arithmetic mean formula in the previous section. The μ represents the mean of a population.

In statistics, it's common to use Greek letters to represent population measures and Latin letters (that is, the alphabet that you use every day) to represent sample measures.

Geometric mean

The main difference between the arithmetic and geometric means is that the arithmetic mean is based on *sums,* while the geometric mean is based on *products*.

For the Omega Airlines example in the previous section, the arithmetic mean doesn't reflect the fact that the size of an investment in this stock grows over time and so it underestimates the true rate of return during the five-year sample period. This underestimation is one of the major drawbacks of the arithmetic mean. Based on the arithmetic mean return of 2 percent per year, the investor would have earned a cumulative return of 10 percent: 2 + 2 + 2 + 2 + 2 = 10 percent from 2008 to 2012.

In fact, the cumulative return was approximately 10.3 percent. To illustrate this return, assume that an investor started with $100,000 at the beginning of 2008. Table 3-1 shows the value of this investment from 2008 to 2012.

Table 3-1 Computing the Return to Omega Airlines Stock

Year	*Omega Airlines Annual Return (percent)*	*Starting Balance*	*Ending Balance*
2008	2	$100,000.00	$100,000.00(1.02) = $102,000.00
2009	–1	$102,000.00	$102,000.00(0.99) = $100,980.00

Year	*Omega Airlines Annual Return (percent)*	*Starting Balance*	*Ending Balance*
2010	3	$100,980.00	$100,980.00(1.03) = $104,009.40
2011	5	$104,009.40	$104,009.40(1.05) = $109,209.87
2012	1	$109,209.87	$109,209.87(1.01) = $110,301.97

In each year, the starting balance is multiplied by the *gross return* (one plus the rate of return) during the year to get the ending balance. Each year's starting balance equals the previous year's ending balance.

The ending balance in 2012 equals $110,301.97. The cumulative rate of return during this period is the ratio of the ending balance to the starting balance minus one:

$$\begin{aligned}\text{Cumulative rate of return} &= \left(\frac{\text{Ending balance}}{\text{Starting balance}}\right) - 1 \\ &= \left(\frac{110{,}301.97}{100{,}000.00}\right) - 1 \\ &= 1.1030197 - 1 \\ &= 0.1030197 \\ &= 10.30197 \text{ percent}\end{aligned}$$

The cumulative return over period 2008–2012 is 10.30197 percent, more than the 10 percent implied by the arithmetic mean. In this case, the geometric mean provides a more accurate result than the arithmetic mean because the geometric mean takes into account the increasing size of the investment, while the arithmetic mean doesn't.

Because the geometric mean is based on products, for a sample or a population, you multiply the gross returns for each year to get the cumulative five-year return:

$$\begin{aligned}&(1+r_{2008})(1+r_{2009})(1+r_{2010})(1+r_{2011})(1+r_{2012}) \\ &= (1.02)(0.99)(1.03)(1.05)(1.01) = 1.1030197\end{aligned}$$

The returns are multiplied in order to indicate that *each year's return* is applied to the cumulative value of the investment, not the original value.

Because this sample has five returns, the next step is to raise the final result 1.1030197 to the *one-fifth* power:

$$(1.1030197)^{(1/5)} = 1.0198039$$

Raising a number to the one-fifth power is also known as taking the *fifth root* of the number. This corresponds to dividing by five when computing the arithmetic mean.

You can determine any exponent on a calculator with the exponentiation key; for most calculators, this key appears as Y^X or X^Y.

Subtracting 1 from the example's result gives you 1.0198039 – 1 = 0.0198039 = 1.98039 percent per year. If the investor earns this return each year for five years, the five-year return will be computed as follows. First, the annual return plus one is multiplied by itself five times.

$$(1.0198039)(1.0198039)(1.0198039)(1.0198039)(1.0198039)$$

Subtracting one gives the cumulative five year return:

$$= (1.0198039)^5 - 1$$

$$= 0.1030199$$

$$= 10.30199 \text{ percent}$$

(Note that there are slight differences in the results due to rounding.)

You use this process for calculating either the geometric mean of a sample or the geometric mean of a population.

Weighted mean

Sometimes a data set contains a large number of repeated values. In these situations, you can simplify the process of computing the mean by using *weights* — the frequencies of a value in a sample or a population. You can compute both the arithmetic mean and geometric mean as weighted averages.

Calculating the weighted arithmetic mean

The formula for computing a weighted arithmetic mean for a sample or a population is

$$\frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

Here, w_i represents the *weight* associated with element X_i; this weight equals the number of times that the element appears in the data set.

The *numerator* (the top half of the formula) tells you to multiply each element in the data set by its weight and then add the results together, as shown here:

$$\sum_{i=1}^{n} w_i X_i = w_1 X_1 + w_2 X_2 + w_3 X_3 + \ldots + w_n X_n$$

The *denominator* (the bottom half of the formula) tells you to add the weights together:

$$\sum_{i=1}^{n} w_i = w_1 + w_2 + w_3 + \ldots + w_n$$

You find the weighted arithmetic mean by dividing the numerator by the denominator.

As an example, suppose that a marketing firm conducts a survey of 1,000 households to determine the average number of TVs each household owns. The data show a large number of households with two or three TVs and a smaller number with one or four. Every household in the sample has at least one TV and no household has more than four. Here's the sample data for the survey:

Number of TVs per Household	***Number of Households***
1	73
2	378
3	459
4	90

Because many of the values in this data set are repeated multiple times, you can easily compute the sample mean as a weighted mean. Doing so is quicker than summing each value in the data set and dividing by the sample size.

Follow these steps to calculate the weighted arithmetic mean:

1. **Assign a weight to each value in the data set:**

   $X_1 = 1$, $w_1 = 73$

   $X_2 = 2$, $w_2 = 378$

   $X_3 = 3$, $w_3 = 459$

   $X_4 = 4$, $w_4 = 90$

2. **Compute the numerator of the weighted mean formula.**

   Multiply each sample by its weight and then add the products together:

$$\begin{aligned}\sum_{i=1}^{4} w_i X_i &= w_1X_1 + w_2X_2 + w_3X_3 + w_4X_4 \\ &= (1)(73) + (2)(378) + (3)(459) + (4)(90) \\ &= 2{,}566\end{aligned}$$

3. **Compute the denominator of the weighted mean formula by adding the weights together:**

$$\begin{aligned}\sum_{i=1}^{4} w_i &= w_1 + w_2 + w_3 + w_4 \\ &= 73 + 378 + 459 + 90 \\ &= 1{,}000\end{aligned}$$

4. **Divide the numerator by the denominator:**

$$\frac{\sum_{i=1}^{4} w_i X_i}{\sum_{i=1}^{4} w_i} = \frac{2{,}566}{1{,}000} = 2.566$$

The mean number of TVs per household in this sample is 2.566.

Calculating the weighted geometric mean

You can calculate the weighted geometric mean in the same way for both samples and populations. The formula is:

$$\left(\prod_{i=1}^{n} X_i^{w_i} \right)^{1/\sum_{i=1}^{n} w_i}$$

Here's the breakdown of this equation:

- Π = the uppercase Greek letter pi used to indicate that a product is being computed
- X_i = a single element in the sample or population
- w_i = the weight of element X_i
- $\sum_{i=1}^{n} w_i$ = the sum of the weights $w_1, w_2, \ldots, w_n$

You apply an *exponent* to each element in the data set that equals the weight of the element. You then multiply these values together and raise to a power equal to one divided by the sum of the weights.

An exponent is the superscript in an expression such as 3^4; in this case, the *base* is 3 and the *exponent* is 4. This is shorthand for multiplying 3 by itself four times: $3 \times 3 \times 3 \times 3 = 81$. Note that in many formulas and Microsoft Excel, the asterisk (*) represents multiplication. In Excel the carat (^) represents exponentiation.

As an example, a marketing firm conducts a survey of 20 households to determine the average number of cellphones each household owns. Here's the sample data from this survey:

Number of Cell Phones Per Household	***Number of Households***
1	2
2	5
3	6
4	4
5	3

To figure out the weighted geometric mean, follow these steps:

1. **Compute the value of each X_i with an exponent equal to its weight w_i:**

   $X_1^{w_1} = 1^2 = 1$

   $X_2^{w_2} = 2^5 = 32$

   $X_3^{w_3} = 3^6 = 729$

   $X_4^{w_4} = 4^4 = 256$

   $X_5^{w_5} = 5^3 = 125$

2. **Multiply these results together:**

$$\left(\prod_{i=1}^{5} X_i^{w_i}\right) = X_1^{w_1} X_2^{w_2} X_3^{w_3} X_4^{w_4} X_5^{w_5}$$
$$= (1)(32)(729)(256)(125) = 746{,}496{,}000$$

3. **Divide 1 by the sum of the weights:**

$$\frac{1}{\sum_{i=1}^{5} w_i} = \frac{1}{2+5+6+4+3} = \frac{1}{20}$$

4. **Combine these results to find the weighted geometric mean:**

$$\left(\prod_{i=1}^{5} X_i^{w_i}\right)^{1/\sum_{i=1}^{5} w_i} = 746{,}496{,}000^{(1/20)} = 2.77748$$

So on average, each household has approximately 2.78 cellphones.

Getting to the Middle of Things: The Median of a Data Set

The *median* is a value that divides a sample or a population in half. In other words:

- Half of the elements in the data set are *below* the median.
- Half of the elements in the data set are *above* the median.

For example, the sample of returns of Omega Airlines stock from 2008 to 2012 is shown here:

Year	***Omega Airlines Annual Return (percent)***
2008	2
2009	–1
2010	3
2011	5
2012	1

You can compute the median of this sample, using the following steps:

1. **Sort the elements from the smallest to the largest.**

   Original data:

   2, –1, 3, 5, 1

   Sorted data:

   –1, 1, 2, 3, 5

2. **Identify the *middle* observation.**

   Because the sample contains five elements, the median is the third largest element (ensuring that two elements are below the median and two are above). The resulting value of the median is 2.

   –1, 1, ***2***, 3, 5

Note: If the sample contains an even number of elements, then no element exists in the middle of the data. Instead, you calculate the median as the *average* of the middle two elements.

Here's another example. This list is a sample of the returns onto Epsilon Railways stock from 2007 to 2012:

Year	*Epsilon Railways Annual Return (percent)*
2007	0
2008	2
2009	3
2010	6
2011	1
2012	4

1. **Sort the elements from smallest to largest.**

   Original data:

   0, 2, 3, 6, 1, 4

   Sorted data:

   0, 1, 2, 3, 4, 6

2. **Identify the middle observation.**

   In this example, there are six sample elements. Because 6 is an even number, you compute the median as the average of the third and fourth elements:

   0, 1, ***2, 3***, 4, 6

   (2 + 3)/2 = 2.5

   Note that three sample elements are below 2.5, and three elements are above 2.5.

The procedure for computing the median of a sample is the same as for computing the median of a population.

Comparing the Mean and Median

In some data sets, the mean and median may equal each other. When this occurs, the data set is said to be *symmetrical about the mean,* meaning that values below the mean balance the values above the mean. A data set may also be *negatively skewed,* indicating the presence of extreme values below the mean. Likewise, a data set may be *positively skewed,* indicating the presence of extreme values above the mean.

If a data set is skewed, the mean and median won't equal each other; instead, the relationship between them will determine the direction of the skew. I explore the relationship of the mean and median as well as the advantages and disadvantages of each measure in the following sections.

Determining the relationship between mean and median

The relationship between the mean and median of a data set determines whether the data set is symmetrical about the mean, negatively skewed, or positively skewed.

Symmetrical

A data set is symmetrical if the mean equals the median. Mathematically, this is expressed as

mean = median

The histogram in Figure 3-1 shows the frequency distribution for the daily returns of a stock with the following mean and median:

mean = 0.00 percent

median = 0.00 percent

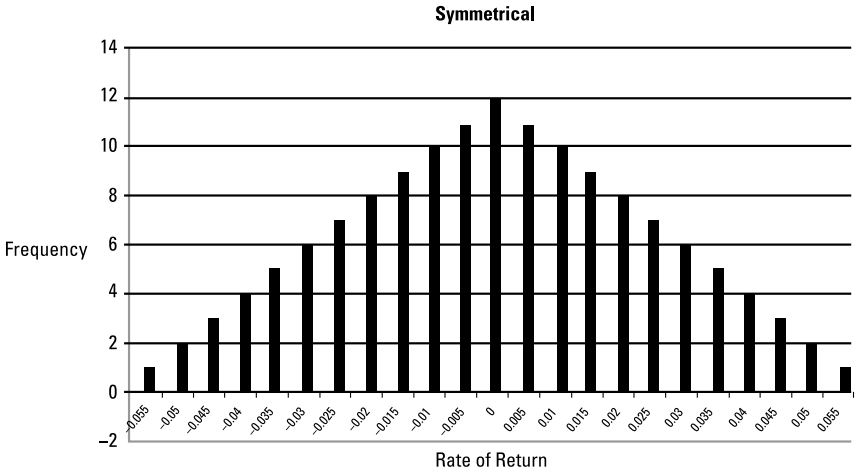

Illustration by Wiley, Composition Services Graphics

Figure 3-1: Symmetrical sample data.

The histogram shows that the left and right *tails* balance each other so that positive and negative values that are equal distances from the center are equally likely. (The left tail represents the smallest observations and the right tail represents the largest observations in the data set.) The left-hand side of this distribution is a mirror image of the right-hand side, showing that this distribution is symmetrical about the mean.

Negatively skewed

A data set is negatively skewed if the mean is less than the median. Mathematically, you can express this relationship as

mean < median

The histogram in Figure 3-2 shows the frequency distribution for the daily returns to a stock with the following mean and median:

mean = –0.95 percent

median = –0.75 percent

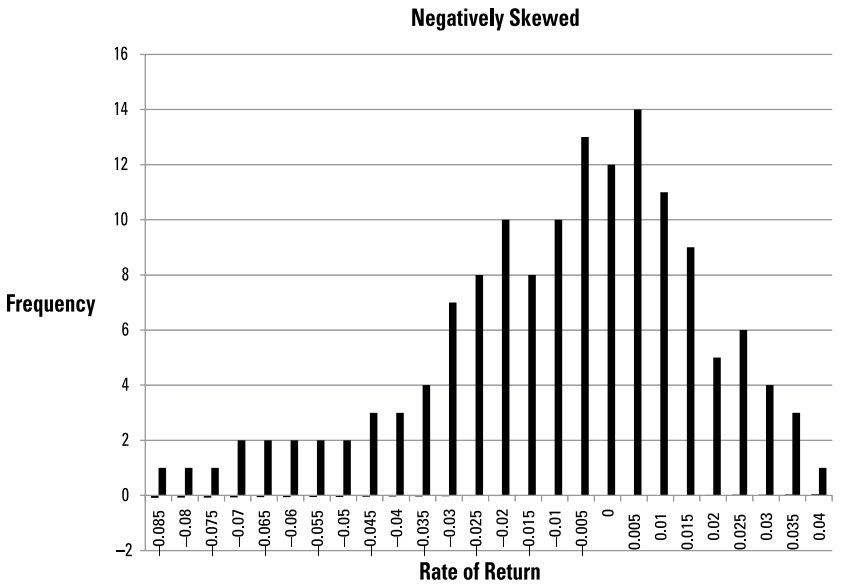

Illustration by Wiley, Composition Services Graphics

Figure 3-2: Negatively skewed sample data.

The histogram shows a long *left tail,* which results from extreme negative values in the data set.

Positively skewed

A data set is positively skewed if the mean is greater than the median. Mathematically, this relationship looks like this:

mean > median

The histogram in Figure 3-3 shows the frequency distribution for the daily returns on a stock with the following mean and median:

mean = 1.55 percent

median = 0.70 percent

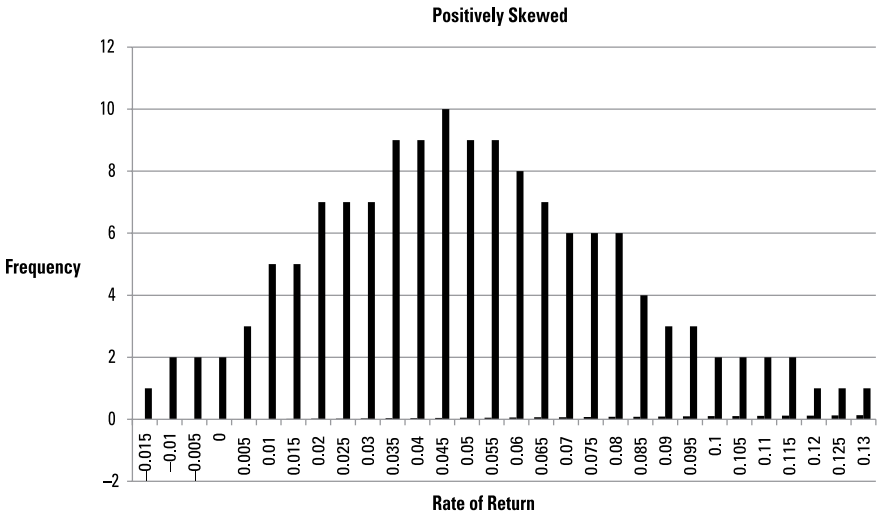

Illustration by Wiley, Composition Services Graphics

Figure 3-3: Positively skewed sample data.

The graph shows a *long right tail*, which results from extreme positive values in the data set.

Acknowledging the relative advantages and disadvantages of the mean and median

The mean is the most commonly used measure of the center of a data set. Under some conditions, though, the median (or even the mode) may be more representative of the center of the data set.

If a data set is symmetrical, the mean and the median are equal, so both are equally useful measures. When a data set is skewed, the median is likely to be a more representative measure of the center of the data than the mean because the median isn't as affected by extreme outcomes as much as the mean.

Discovering the Mode: The Most Frequently Repeated Element

The *mode* is the most frequently occurring value in a sample or a population. For example, suppose a bank chooses a sample of 20 of its branches in New York City, and for each branch, the number of ATMs in the lobby is recorded as follows:

Three branches have two ATMs.

Six branches have three ATMs.

Eight branches have four ATMs.

Three branches have five ATMs.

Because most branches have four ATMs, 4 is the mode in this sample.

One of the most unusual features of the mode is that it isn't necessarily unique; a data set can have two or more modes. It's also possible that a data set has no mode — that is, no values are repeated.

For example, suppose that the same bank chooses a sample of 20 of its branches in Connecticut. For each branch, the number of ATMs in the lobby is recorded. The results are given as follows:

Three branches have two ATMs.

Eight branches have three ATMs.

Eight branches have four ATMs.

One branch has five ATMs.

In this sample, more branches have three or four ATMs than any other number. Because the number of branches with three ATMs equals the number of banks with four ATMs, the mode of this sample is *both* 3 and 4.

The mode is most useful when a data set contains qualitative data (that is, non-numerical data). This type of data can include colors, flavors, brand names, and so on. With qualitative data, calculating a mean or a median is impossible, but you can still find the mode. With quantitative (numerical) data, the mean and the median are typically more useful than the mode.

As an example, suppose that a marketing firm conducts a survey to determine which color consumers would likely choose for a new car. The survey responses are as follows:

blue	red	blue
black	blue	black
blue	blue	black
blue	black	blue
white	silver	blue

Because this data is qualitative, calculating the mean or the median is impossible. But you can determine the mode by tabulating the frequency of the 15 responses. Because blue appears in the survey eight times, black, four times, white, red, and silver, one each, the mode is blue. Consumers in this survey prefer blue to other colors.

The distribution of colors is shown in Figure 3-4. In this example, the histogram shows colors on the horizontal axis and the corresponding frequencies on the vertical axis:

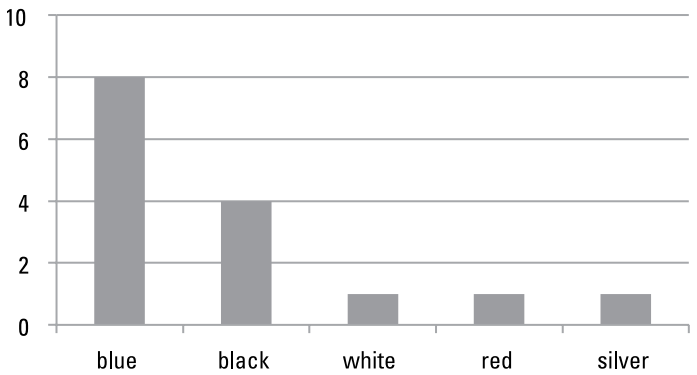

Illustration by Wiley, Composition Services Graphics

Figure 3-4: Distribution of colors chosen by consumers.

Because blue occurs most frequently in this sample, it's the sample's mode.

Chapter 4

Searching High and Low: Measuring Variation in a Data Set

In This Chapter

- Computing variance and standard deviation
- Finding the relative position of data: percentiles and quartiles
- Measuring relative variation: the coefficient of variation

One of the most important properties of a data set (a sample or population) is how "spread out" the data are from the center. (Techniques for measuring the center of a data set are covered in Chapter 3.) You can use several numerical measures, known as *measures of dispersion,* to calculate the spread of a data set.

This chapter covers the techniques used to compute the variance and standard deviation of a sample and a population. (Samples and populations are defined in Chapter 1.) Techniques for determining the relative position of an element within a sample or a population are also explained in detail; these include percentiles and quartiles. Finally, the coefficient of variation is introduced as a measure of *relative variation*; this enables a direct comparison of the properties of two samples or two populations.

Thanks to standard deviation and the mean (covered in Chapter 3), you can calculate relative variation, which has many handy applications.

Determining Variance and Standard Deviation

Variance and *standard deviation* are the two most widely used measures of dispersion in statistics. They're both based on the average squared distance between the elements of a data set and the mean.

Standard deviation and variance are usually better than some other measures of dispersion, such as the range. The range is the difference between the largest and smallest elements in a data set. Interesting, but not that great. The range suffers from the drawback that it's only based on two values, so it doesn't measure the spread among the remaining values.

The variance indicates the size of the average *squared* difference between the elements of a data set and the mean of the data set. And here's what you need to know: A large variance shows a substantial amount of spread among the elements of a data set.

Variance is often used as a measure of uncertainty or risk in business applications. For example, an investor may use variance to determine the degree of risk associated with owning a share of stock. If returns of the stock fluctuate significantly over time, it's a risky investment. Variance provides a method for assigning a numerical value to this fluctuation. The greater the stock's variance, the riskier it is.

Standard deviation is the *square root* of the variance. It's more commonly used than variance as a measure of risk because the variance is expressed in *squared units*. For example, the variance of stock returns is expressed as *percent squared*, which is difficult to visualize. On the other hand, the standard deviation of stock returns is measured as a percentage, which is much easier to interpret.

Finding the sample variance

Use the following formula to figure out the variance of a sample:

$$s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}$$

Here's what each term means:

- s^2 = the sample variance
- $\bar{X}$ (pronounced "X bar"); this is the sample mean (the average value of the sample elements)
- n = the number of elements in the sample
- i = an *index,* assigning a number to each sample element ranging from 1 to n
- X_i = a single element in the sample
- Σ = the uppercase Greek letter sigma, which indicates a sum is being computed

The *numerator* (the top half) of the sample variance formula is:

$$\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2=\left(X_1-\bar{X}\right)^2+\left(X_2-\bar{X}\right)^2+\ldots+\left(X_n-\bar{X}\right)^2$$

This expression tells you to perform the following three calculations:

1. For each sample element, subtract the sample mean.
2. Square the result.
3. Compute the sum of these squares.

The *denominator* (the bottom half) of the sample variance formula is $n - 1$ (the sample size minus 1). Then, you find the sample variance by dividing the numerator by the denominator.

Finding the sample standard deviation

The sample standard deviation is the *square root* of the sample variance:

$$s=\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}}$$

Here's an example: Say you choose sample of coffee prices from 20 stores in 2 supermarket chains: Encore Markets and Pacifica Markets. Figure 4-1a shows the distribution of prices at Encore Markets, and Figure 4-1b shows the distribution of prices at Pacifica Markets. The price of coffee per pound is shown on the horizontal (X) axis, while the number of stores that charge a given price are shown on the vertical (Y) axis.

These graphs show that the prices are much more spread out at Pacifica's stores than at Encore's. In other words, Pacifica has greater *dispersion* among its prices. The range of possible prices at Pacifica's stores is much greater (at least one store charges $20 per pound!), while at Encore, no store charges more than $14. The stores at both chains charge at least $8 per pound. The dispersion among coffee prices is measured by the standard deviation, which is $3.6631 at Pacifica's stores and $2.1637 at Encore's stores. These numbers confirm what Table 4-1 shows: There's more spread among Pacifica's prices than Encore's prices.

Tables 4-1 and 4-2 show the prices at 20 stores in each of the two chains.

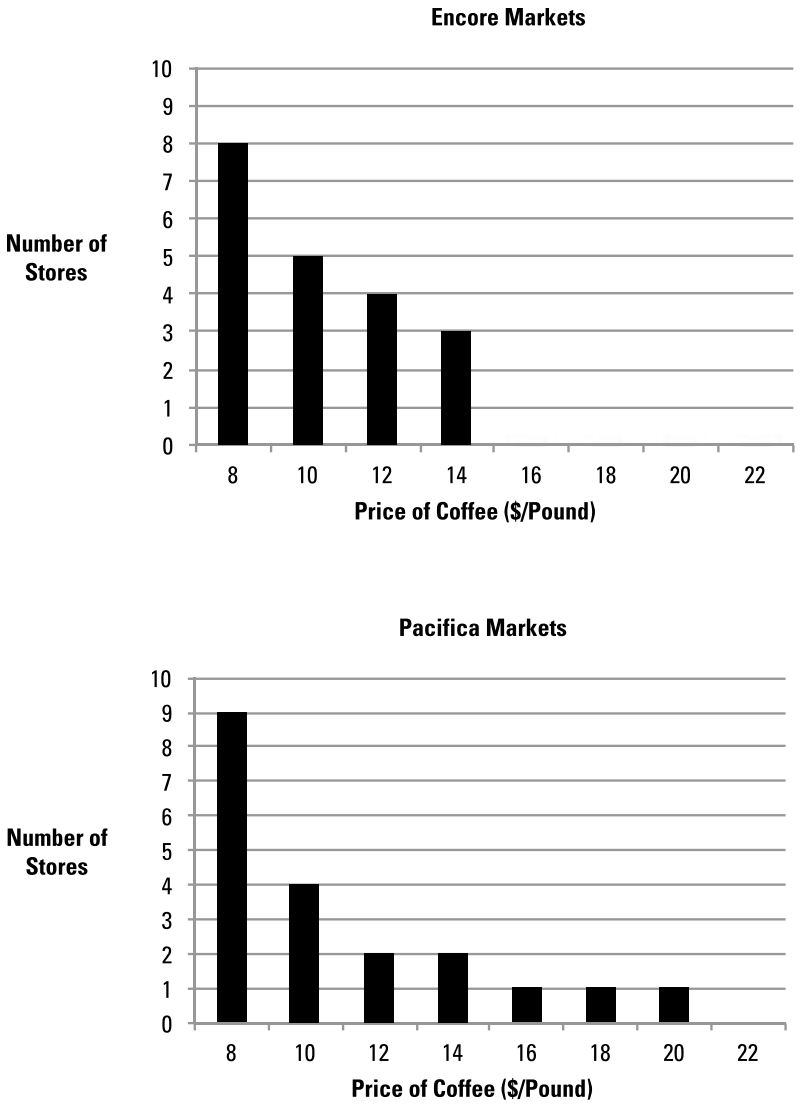

Figure 4-1 (a and b): Distribution of coffee prices at Encore Markets and Pacifica Markets.

Table 4-1	Sample Coffee Prices at Encore Markets		
8	10	11	8
8	9	8	8
13	8	9	14
12	8	12	14
10	12	8	9

Table 4-2	Sample Coffee Prices at Pacifica Markets		
15	17	9	7
13	7	7	9
9	8	7	7
9	13	7	11
19	11	7	7

The first step is to compute the sample mean coffee price. In this example, the sample mean for Encore is computed as follows:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

The numerator is the sum of the coffee prices in the sample, which is 199. The denominator is the sample size, which is 20. The ratio of these two values is the sample mean, $9.95.

To compute the sample variance, subtract the sample mean from each sample coffee price, and square the results. The sum of these terms is the numerator of the sample variance formula. This is shown in the Table 4-3.

Table 4-3	Calculations for the Sample Variance at Encore Markets		
$(8 - 9.95)^2 = 3.8025$	$(10 - 9.95)^2 = 0.0025$	$(11 - 9.95)^2 = 1.1025$	$(8 - 9.95)^2 = 3.8025$
$(8 - 9.95)^2 = 3.8025$	$(9 - 9.95)^2 = 0.9025$	$(8 - 9.95)^2 = 3.8025$	$(8 - 9.95)^2 = 3.8025$
$(13 - 9.95)^2 = 9.3025$	$(8 - 9.95)^2 = 3.8025$	$(9 - 9.95)^2 = 0.9025$	$(14 - 9.95)^2 = 16.4025$
$(12 - 9.95)^2 = 4.2025$	$(8 - 9.95)^2 = 3.8025$	$(12 - 9.95)^2 = 4.2025$	$(14 - 9.95)^2 = 16.4025$
$(10 - 9.95)^2 = 0.0025$	$(12 - 9.95)^2 = 4.2025$	$(8 - 9.95)^2 = 3.8025$	$(9 - 9.95)^2 = 0.9025$

The sum of these terms is 88.95. The sample variance is, therefore:

$$s^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2}{n-1} = \frac{88.95}{19} = 4.6816$$

Now, at last! Take the square root. The sample standard deviation is:

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}} = \sqrt{4.6816} = 2.1637$$

Compute the sample variance and sample standard deviation for Pacifica Markets the same way. Table 4-4 shows the calculations for the numerator of the sample variance formula.

Table 4-4 Calculations for the Sample Variance at Pacifica Markets

$(15 - 9.95)^2 = 25.5025$	$(17 - 9.95)^2 = 49.7025$	$(9 - 9.95)^2 = 0.9025$	$(7 - 9.95)^2 = 8.7025$
$(13 - 9.95)^2 = 9.3025$	$(7 - 9.95)^2 = 8.7025$	$(7 - 9.95)^2 = 8.7025$	$(9 - 9.95)^2 = 0.9025$
$(9 - 9.95)^2 = 0.9025$	$(8 - 9.95)^2 = 3.8025$	$(7 - 9.95)^2 = 8.7025$	$(7 - 9.95)^2 = 8.7025$
$(9 - 9.95)^2 = 0.9025$	$(13 - 9.95)^2 = 9.3025$	$(7 - 9.95)^2 = 8.7025$	$(11 - 9.95)^2 = 1.1025$
$(19 - 9.95)^2 = 81.9025$	$(11 - 9.95)^2 = 1.1025$	$(7 - 9.95)^2 = 8.7025$	$(7 - 9.95)^2 = 8.7025$

The sum of these terms is 254.95. The sample variance is, therefore:

$$s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1} = \frac{254.95}{19} = 13.4184$$

The sample standard deviation is:

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}} = \sqrt{13.4184} = 3.6631$$

These numbers confirm what Figure 4-1a and Figure 4-1b show: There's more spread among Pacifica's prices than Encore's prices. \$2.1637 compared to \$3.6631.

Although you can use graphs to inspect the dispersion of different samples or populations, comparing standard deviations is usually easier, and you don't have to examine the entire data set.

The standard deviation is a more useful measure of dispersion than variance. Again, variance is expressed in *squared* units (percent squared, dollars squared, and so on) because it's taken from the sum of *squared* differences between the elements in a data set and the mean of the data set. That's not as handy as standard deviation.

For example, Table 4-5 compares the variance and standard deviation of the Encore and Pacifica stores.

Table 4-5 Variance and Standard Deviation of Sample Stores

	Encore	*Pacifica*
Standard deviation ($/pound)	2.1637	3.6631
Variance ($^{2}/pound)	4.6816	13.4184

Table 4-5 shows that the variance of coffee prices at Encore is $4.6816 *squared* per pound, while the variance of coffee prices at Pacifica is $13.4184 *squared* per pound. *Dollars squared* is a difficult concept to interpret — prices are never expressed in terms of dollars squared! So people most often use the standard deviation rather than the variance to show dispersion.

Calculating population variance and standard deviation

Unlike the mean, median, and mode, the variance and the standard deviation are calculated slightly differently for *samples* and *populations*. The following section shows the appropriate formulas for computing the variance and standard deviation of a population.

Finding the population variance

When you're calculating the variance for a population, use the following formula:

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

The parameters are:

- σ^2 = population variance (σ is the lowercase Greek letter sigma)
- μ = the population mean (μ is the Greek letter mu)

Σ is the uppercase Greek letter sigma, which represents summation σ is the lowercase sigma, which represents the population standard deviation.

The *numerator* (the top half) of the population variance formula is:

$$\sum_{i=1}^{n}(X_i-\mu)^2=(X_1-\mu)^2+(X_2-\mu)^2+\ldots+(X_n-\mu)^2$$

Use this formula and do the following calculations:

1. **For each population element, subtract the population mean.**
2. **Square the result.**
3. **Compute the sum of the squares.**

The *denominator* (the bottom half) of the population variance formula is *n* (the population size.) You find the population variance by dividing the numerator of the population variance formula by the denominator.

Finding the population standard deviation

After you figure out the population variance, you can get the population standard deviation by taking the square root of the population:

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

For example, suppose an investor wants to analyze the dispersion of Alpha, Inc.'s, sales from one year to the next. Table 4-6 shows the sample of annual profits the investor takes (measured in millions of dollars per year) from 2007 to 2012.

Table 4-6 Alpha, Inc. Sales 2007–2012

Year	*Sales ($ million)*
2007	18
2008	22
2009	31
2010	29
2011	42
2012	50

You find the population variance by following these steps:

1. **Find the population mean.**

   The formula for calculating the sample mean is

   $$\mu = \frac{\sum_{i=1}^{n} X_i}{n}$$

   Plug in the numbers from Table 4-6:

   $$\mu = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{18 + 22 + 31 + 29 + 42 + 50}{6} = 32$$

   The average annual profit during this period was $32 million.

2. **Work through the numerator of the sample variance formula.**

   $$\sum_{i=1}^{n} (X_i - \mu)^2$$

   The calculations are shown in Table 4-7.

Table 4-7 Calculations of Population Variance for Alpha, Inc.

Year	*Alpha, Inc. Sales ($ million)*	$(X_i - \mu)$	$(X_i - \mu)^2$
2007	18	18 – 32 = –14	$(-14)^2 = 196$
2008	22	22 – 32 = –10	$(-10)^2 = 100$
2009	31	31 – 32 = –1	$(-1)^2 = 1$
2010	29	29 – 32 = –3	$(-3)^2 = 9$
2011	42	42 – 32 = 10	$(10)^2 = 100$
2012	50	50 – 32 = 18	$(18)^2 = 324$
		Sum	**730**

   In the third column ($(X_i - \mu)$), subtract the mean return from the actual return for each year. In the fourth column ($(X_i - \mu)^2$), square the result from the third column. The sum of the fourth column is the numerator of the sample variance formula; this equals 730.

3. **Solve the denominator of the population variance formula.**

   The denominator is 6. Because six elements are in this population, $n = 6$.

4. **Substitute these values into the population variance formula.**

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n} = \frac{730}{6} = 121.667$$

The population variance of Alpha's sales is \$121.667 dollars squared.

Finding the population standard deviation

After you figure out the population variance, you get the population standard deviation by taking the square root of the population variance:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}} = \sqrt{121.667} = 11.030$$

The population standard deviation of Alpha's sales is \$11.030 million.

Finding the Relative Position of Data

Identifying the location or position of a value in a data set can be immensely useful, whether you're talking about business profitability, population statistics, or scores on school tests. You use three related measures known as *percentiles*, *quartiles*, and the *interquartile range*.

A percentile is a value that divides a sample or population into two groups, with a specified percentage in each group. For example, on a standardized exam, the 10th percentile is the score such that:

10 percent of the scores are below it

90 percent of the scores are above it

Quartiles are closely related to percentiles; they subdivide a sample or a population into four equal parts. The interquartile range identifies the middle 50 percent.

Percentiles: Dividing everything into hundredths

Percentiles split up a data set into 100 equal parts, each consisting of 1 percent of the values in the data set.

For example, suppose a corporation is analyzing the annual sales of its franchise owners. Those franchises whose sales belong to the 90th percentile will get an award. Being in the 90th percentile means that:

90 percent of the franchises have sales below this value

10 percent of the franchises have sales above this value

As a result, 10 percent of the franchises will receive the award. When you hear someone say that he or she is in the "top 10 percent," you can take that to mean that they are in the 90th percentile.

Percentiles provide a *relative ranking* for an element of a data set. For example, suppose that the corporation's New York franchise has sales of $1 million during the year. Judging whether this franchise is successful without knowing how this value compares with the other franchises is difficult. If it turns out that $1 million places the New York franchise in the 10th percentile, then 90 percent of the other franchises outperformed it this year. On the other hand, if $1 million places the New York franchise in the 80th percentile, then only 20 percent of the other franchises outperformed it this year.

The 50th percentile of a data set is the median because half of the values are below the median, and half are above.

Suppose the Federal Reserve Bank of New York conducts a survey of the assets of the savings banks in its district. A sample of ten banks is chosen; the results (in hundreds of millions of dollars) are:

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

To compute percentiles, first sort the elements from the smallest value to the largest. In this example, the sorted values are:

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

There are several possible approaches to computing percentiles. One of them is to apply the following formula to compute an *index*. This index represents the location of the appropriate percentile.

$$\frac{P}{100}n + 0.5$$

Here, P is the percentile of interest (30th, 40th, and so on), and n is the size of the sample or population. You round the number to the nearest integer (whole number). The percentile equals the corresponding value in the data set.

When rounding a number with a fractional part, if the fractional part is 0.5 or greater, round *up* to the next higher integer; otherwise, round *down* to the next lower integer. For example, you round 3.4 *down* to 3, and 3.5 *up* to 4.

For example, in order to find the 30th percentile of a set of ten, the index is

$$\frac{P}{100}n + 0.5 = \frac{30}{100}(10) + 0.5 = 3.5$$

Round 3.5 up to 4 to see that the fourth smallest value, the number 3 in this example, is the 30th percentile.

1, 2, 2, ***3***, 4, 5, 6, 7, 8, 9

Similarly, you find the 70th percentile of a set of ten as follows:

$$\frac{P}{100}n + 0.5 = \frac{70}{100}(10) + 0.5 = 7.5$$

Don't forget to round 7.5 up to 8, which shows that the eighth smallest value, or the number 7 in this example, is the 70th percentile.

1, 2, 2, 3, 4, 5, 6, ***7***, 8, 9

Microsoft Excel uses a somewhat different approach to computing percentiles. If you use the PERCENTILE function, you will get 2.7 for the 30th percentile and 6.3 for the 70th percentile.

Quartiles: Dividing everything into fourths

Quartiles split up a data set into four equal parts, each consisting of 25 percent of the sorted values in the data set. Quartiles are related to percentiles like so:

First quartile (Q_1) = 25th percentile

Second quartile (Q_2) = 50th percentile

Third quartile (Q_3) = 75th percentile

Because the second quartile is the 50th percentile, it's also the *median* of a data set. The fourth quartile usually isn't used because its value is greater than every element in a data set, so what's the point?

One commonly used approach for calculating quartiles follows these two steps:

1. **Split the data into a lower half and an upper half (leaving out the median).**
2. **Compute the median of the lower half and the upper half.**

   After you've split the data into lower and upper halves, you figure out the quartiles as follows:

Q_1 = the median of the lower half

Q_2 = the median of the entire data set

Q_3 = the median of the upper half

The following data represent a sample of eight stock returns for Gamma Industries:

5, 7, 6, 3, 0, –2, 4, 3

The sorted values are:

–2, 0, 3, 3, 4, 5, 6, 7

In this example, you have eight elements. Because 8 is an even number, the median is the average of the fourth and fifth elements: –2, 0, 3, ***3, 4***, 5, 6, 7

(3 + 4)/2 = 3.5. Therefore, the second quartile (Q_2) is 3.5.

The values below the median constitute the lower half of the sorted sample

–2, 0, 3, 3

The values above the median constitute the upper half of the sorted sample

4, 5, 6, 7

Both the lower and upper halves have four sample elements. Because four is an even number, the median is the average of the second and third elements.

For the lower half, the median is: (0 + 3)/2 = 1.5. This is the *average* value of the two middle elements. Therefore, the first quartile (Q_1) is 1.5.

For the upper half, the median is (5 + 6)/2 = 5.5. Therefore, the third quartile (Q_3) is 5.5.

As with percentiles, Microsoft Excel uses a different approach to computing quartiles; if you use the QUARTILE function, you will get 3.5 for Q_2, but you will also get

2.25 for Q_1 (instead of 1.5)

5.25 for Q_3 (instead of 5.5)

Interquartile range: Identifying the middle 50 percent

The *interquartile range* (IQR) is the difference between the third quartile and the first quartile: IQR = $Q_3 – Q_1$. The IQR represents the middle 50 percent of the data set. For the Gamma Industries example, the IQR is $Q_3 – Q_1$ = 5.5 – 1.5 = 4.

An advantage of the IQR is that it isn't greatly affected by *outliers* — values within a data set that are significantly different than the other elements in the data set. In fact, the IQR can help identify outliers within a data set.

You can find the outliers in a data set in several ways. One of the simpler approaches is to create a *lower bound* and an *upper bound*. What this means is that if any elements are below the lower bound or above the upper bound, they're outliers. You set these bounds based on quartiles and the interquartile range:

lower bound: $Q_1 - 1.5(IQR)$

upper bound: $Q_3 + 1.5(IQR)$

Based on the Gamma Industries data, the lower bound = $1.5 - 1.5(4) = -4.5$, and the upper bound = $5.5 + 1.5(4) = 11.5$.

Because no value in this sample is below –4.5 or above 11.5, the sample has no outliers.

Measuring Relative Variation

Relative variation refers to the spread of a sample or a population as a proportion of the mean. Relative variation is useful because it can be expressed as a percentage, and is independent of the units in which the sample or population data are measured.

For example, you can use a measure of relative variation to compare the uncertainty or variation associated with the temperature in two different countries, even if one country uses Fahrenheit temperatures and the other uses Celsius temperatures. As another example, a measure of relative variation can be useful for comparing the returns earned by two portfolio managers. It wouldn't make any sense to compare the mean returns achieved by two different managers without explicitly considering the levels of risk that they have incurred. A measure of relative variation provides a number that considers both the risk and the return of a portfolio, so that it can be determined which portfolio is riskier relative to the return.

You can use several different types of measures of relative variation. One of the most popular is known as the coefficient of variation.

Coefficient of variation: The spread of a data set relative to the mean

The *coefficient of variation* (CV) indicates how "spread out" the members of a sample or population are relative to the mean. The coefficient of variation is measured as a percentage, so it's independent of the units in which the mean and standard deviation are measured. This enables the relative variation of different samples or populations to be compared directly to each other.

For example, the coefficient of variation can express the risk of an investment portfolio *per unit of return.* This means you can compare the performance of different portfolios to see which one offers the least amount of risk per unit of return.

Here's the formula for finding the coefficient of variation for either samples or populations:

$$CV = \left(\frac{\text{standard deviation}}{\text{mean}} \right) * 100$$

Suppose a corporation requires the services of a consulting firm to improve its accounting systems. The corporation has determined that the two best choices are Superior Accounting, Inc., and Data Services Corp. The corporation has done some research about the pricing practices of these two firms. The average price charged per hour, along with the standard deviation, are shown in Table 4-8:

Table 4-8 Comparative Prices Charged by Superior Accounting and Data Services

	Superior Accounting	*Data Services*
Mean price ($/hour)	$200	$175
Standard deviation ($/hour)	$80	$75

Based on this data, the coefficient of variation for the prices charged by each firm are

Superior Accounting: $CV = \frac{\$80}{\$200} * 100 = 40.00$ percent

Data Services: $CV = \frac{\$75}{\$175} * 100 = 42.86$ percent

These results show that although the prices charged by Superior Accounting have a larger standard deviation than Data Services, the relative variation of Data Services is greater (42.86 percent compared with 40.00 percent.) This indicates that the relative uncertainty associated with Data Services' prices is greater than for Superior Accounting's prices.

Comparing the relative risks of two portfolios

Suppose a portfolio manager is responsible for an insurance company's equity portfolio and bond portfolio. He wants to know which portfolio is riskier in absolute and relative terms. He takes a sample of returns from the past ten years and computes the mean and standard deviation. See Table 4-9 for the results:

Table 4-9 Comparative Performance of Bond and Equity Portfolios

	Bond Portfolio	*Equity Portfolio*
Mean return	8%	20%
Standard deviation of returns	16%	30%

These results show that the equity portfolio offers a higher average (mean) return than the bond portfolio and that the equity portfolio is *riskier* in absolute terms than the bond portfolio.

Because the two portfolios offer different returns and different levels of risk, it's impossible to compare them directly without using a measure of *relative risk,* which shows how risky a portfolio is relative to its return. So you need to find the coefficient of variation for the two portfolios, using the CV formula:

$$\text{Bond: } CV = \frac{16 \text{ percent}}{8 \text{ percent}} * 100 = 200 \text{ percent}$$

$$\text{Equity: } CV = \frac{30 \text{ percent}}{20 \text{ percent}} * 100 = 150 \text{ percent}$$

The bond portfolio offers a level of risk that's 200 percent of the average return, while the equity portfolio offers a level of risk that's 150 percent of the average return. So while the equity portfolio is riskier in *absolute* terms (due to the higher standard deviation) the bond portfolio is riskier in *relative* terms (due to the higher coefficient of variation).

Chapter 5

Measuring How Data Sets Are Related to Each Other

In This Chapter

- Working with measures of association: covariance and correlation
- Determining the correlation coefficient

A *measure of association* is a numerical value that reflects the tendency of two variables to move in the same direction or in opposite directions. For example, it makes sense that corporate profits and sales would both tend to increase when the economy is strong, and decrease when the economy is weak. A measure of association is used to assign a numerical value to the strength and direction of this type of relationship.

Measures of association can help answer questions, such as, "If interest rates fall, do stock prices tend to rise?" or "If oil prices rise, does the unemployment rate tend to rise?" or "Does an increase in advertising expenditures lead to greater revenues?"

The two most widely used measures of association are known as *covariance* and *correlation.*

In this chapter, you see formulas for computing covariance and correlation for both samples and populations. The relationship between two variables is illustrated with a type of graph known as a *scatter plot,* which is useful for seeing the relationship that exists (if any) between two variables. (I cover several types of graphs such as the scatter plot in Chapter 2.) This chapter concludes by illustrating how the risks of a portfolio of stocks may be diversified if the stocks have low or negative correlations between them.

Understanding Covariance and Correlation

Two of the most widely used measures of association are known as *covariance* and *correlation*. These are closely related to each other. You can think

of correlation as a modified version of covariance. Correlation is easier to interpret because its value is always between –1 and 1. For example, a correlation of 0.9 indicates a very strong relationship in which two variables nearly always move in the same direction; a correlation of –0.1 shows a very weak relationship in which there is a slight tendency for two variables to move in opposite directions. With covariance, there is no minimum or maximum value, so the values are more difficult to interpret. For example, a covariance of 50 may show a strong or weak relationship; this depends on the units in which covariance is measured.

Correlation is a measure of the strength and direction of two *linearly related* variables. Two variables are said to be linearly related if they can be expressed with the following equation:

$$Y = mX + b$$

X and Y are variables; m and b are constants. For example, suppose that the relationship between two variables is:

$$Y = 3X + 4$$

3 is the *coefficient* of X; this indicates that an increase of X by 1 causes Y to increase by 3. Equivalently, a decrease of X by 1 causes Y to decrease by 3. The 4 in this equation indicates that Y equals 4 when X equals 0.

Covariance and correlation show that variables can have a positive relationship, a negative relationship, or no relationship at all. With covariance and correlation, there are three cases that may arise:

- **If two variables increase or decrease at the same time, the covariance and correlation between them is *positive*.** For example, the covariance and correlation between the stock prices of two oil companies is positive because many of the same factors affect the stock prices in the same way.
- **If two variables move in opposite directions, the covariance and correlation between them is *negative*.** For example, the covariance and correlation between interest rates and new home sales is negative because rising interest rates increase the cost of purchasing a new home, which in turn reduces new home sales. The opposite occurs with falling interest rates.
- **If two variables are unrelated to each other, the covariance and correlation between them is *zero* (or very close to zero).** For example, the covariance and correlation between gold prices and new car sales is zero because the two have nothing to do with each other.

In the following sections, I introduce formulas for computing sample covariance, sample correlation, population covariance, and population correlation. These measures are illustrated with several examples.

Sample covariance and correlation

Sample covariance measures the strength and the direction of the relationship between the elements of two samples. (Recall from Chapter 1 that a sample is a randomly chosen selection of elements from an underlying population.)

The sample covariance between X and Y is

$$s_{XY} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

Here's what each element in this equation means:

- s_{XY} = the sample covariance between variables X and Y (the two subscripts indicate that this is the sample covariance, not the sample standard deviation).
- $\bar{X}$ ("X bar") = the sample mean for X.
- $\bar{Y}$ ("Y bar") = the sample mean for Y.
- n = the number of elements in both samples.
- i = an *index* that assigns a number to each sample element, ranging from 1 to n.
- X_i = a single element in the sample for X.
- Y_i = a single element in the sample for Y.
- Σ = the uppercase Greek letter sigma that indicates that a sum is being computed.

The sample covariance may have any positive or negative value.

You calculate the *sample correlation* (also known as the sample *correlation coefficient*) between X and Y directly from the sample covariance with the following formula:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

The key terms in this formula are

- r_{XY} = sample correlation between X and Y
- s_{XY} = sample covariance between X and Y
- s_X = sample standard deviation of X
- s_Y = sample standard deviation of Y

The formula used to compute the sample correlation coefficient ensures that its value ranges between –1 and 1.

For example, suppose you take a sample of stock returns from the Excelsior Corporation and the Adirondack Corporation from the years 2008 to 2012, as shown here:

Year	*Excelsior Corp. Annual Return (percent) (X)*	*Adirondack Corp. Annual Return (percent) (Y)*
2008	1	3
2009	–2	2
2010	3	4
2011	0	6
2012	3	0

What are the covariance and correlation between the stock returns? To figure that out, you first have to find the mean of each sample. (The sample mean is discussed in Chapter 3.) In this example, X represents the returns to Excelsior and Y represents the returns to Adirondack.

- The sample mean of X is

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{(1-2+3+0+3)}{5} = \frac{5}{5} = 1$$

You obtain the sample mean by summing all the elements of the sample and then dividing by the sample size. In this case, the sample elements sum to 5 and the sample size is 5. Dividing these numbers gives a sample mean of 1.

- The sample mean of Y is

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{(3+2+4+6+0)}{5} = \frac{15}{5} = 3$$

Table 5-1 shows the remaining calculations for the sample covariance:

Table 5-1 Computing the Sample Covariance

Year	Excelsior Corp Annual Return (percent)	Adirondack Corp Annual Return (percent)	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
2008	1	*3*	1 – 1 = 0	3 – 3 = 0	(0)(0) = 0
2009	–2	*2*	–2 – 1 = –3	2 – 3 = –1	(–3)(–1) = 3
2010	3	*4*	3 – 1 = 2	4 – 3 = 1	(2)(1) = 2
2011	0	*6*	0 – 1 = –1	6 – 3 = 3	(–1)(3)= –3
2012	3	*0*	3 – 1 = 2	0 – 3 = –3	(2)(–3) = –6
Mean	**1**	***3***		**Sum**	**–4**

The $(X_i - \bar{X})$ column represents the differences between each return to Excelsior in the sample and the sample mean; similarly, the $(Y_i - \bar{Y})$ column represents the same calculations for Adirondack. The entries in the $(X_i - \bar{X})(Y_i - \bar{Y})$ column equal the product of the entries in the previous two columns. The sum of the $(X_i - \bar{X})(Y_i - \bar{Y})$ column gives the numerator in the sample covariance formula:

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = -4$$

The denominator equals the sample size minus one, which is 5 – 1 = 4. (Both samples have five elements, n = 5.) Therefore, the sample covariance equals

$$s_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$
$$= \frac{-4}{4} = -1$$

To calculate the sample correlation coefficient, divide the sample covariance by the product of the sample standard deviation of *X* and the sample standard deviation of *Y:*

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

You find the sample standard deviation of *X* by computing the sample variance of *X* and then taking the square root of the result (as I explain in Chapter 4). Table 5-2 shows the calculations for the sample variance of *X.*

Table 5-2 Computing the Sample Variance for Excelsior

Year	Excelsior Corp. Annual Return (percent)	$\left(X_i - \bar{X}\right)$	$\left(X_i - \bar{X}\right)^2$
2008	1	1 – 1 = 0	$(0)^2 = 0$
2009	–2	–2 – 1 = –3	$(-3)^2 = 9$
2010	3	3 – 1 = 2	$(2)^2 = 4$
2011	0	0 – 1 = –1	$(-1)^2 = 1$
2012	3	3 – 1 = 2	$(2)^2 = 4$
Mean	**1**	**Sum**	**18**

The $\left(X_i - \bar{X}\right)$ column represents the differences between each return to Excelsior in the sample and the sample mean; the $\left(X_i - \bar{X}\right)^2$ column represents the *squared* difference between each return to Excelsior and the sample mean. The sum of the $\left(X_i - \bar{X}\right)^2$ column gives the numerator in the sample variance formula. You divide this number by the sample size minus one (5 – 1 = 4) to get the sample variance of *X*:

$$\begin{aligned} s_X^2 &= \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1} \\ &= \frac{18}{4} \\ &= 4.5 \end{aligned}$$

The sample standard deviation of *X* is the square root of 4.5, or $\sqrt{4.5} = 2.1213$.

Table 5-3 shows the calculations for the sample variance of *Y*.

Table 5-3 Computing the Sample Variance for Adirondack

Year	Adirondack Corp. Annual Return (percent)	$\left(Y_i - \bar{Y}\right)$	$\left(Y_i - \bar{Y}\right)^2$
2008	3	3 – 3 = 0	$(0)^2 = 0$
2009	2	2 – 3 = –1	$(-1)^2 = 1$
2010	4	4 – 3 = 1	$(1)^2 = 1$
2011	6	6 – 3 = 3	$(3)^2 = 9$
2012	0	0 – 3 = –3	$(-3)^2 = 9$
Mean	**3**	**Sum**	**20**

Based on the calculations in Table 5-3, the sample variance of Y equals

$$\begin{aligned} s_Y^2 &= \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}{n-1} \\ &= \frac{20}{4} \\ &= 5 \end{aligned}$$

The sample standard deviation of Y equals the square root of 5, or $\sqrt{5} = 2.2361$.

Substituting these values into the sample correlation formula gives you

$$\begin{aligned} r_{XY} &= \frac{s_{XY}}{s_X s_Y} \\ &= \frac{-1}{(2.1213)(2.2361)} \\ &= -0.2108 \end{aligned}$$

The negative result shows that there's a weak negative correlation between the stock returns of Excelsior and Adirondack. If two variables are *perfectly* negatively correlated (they *always* move in opposite directions), their correlation will be –1. If two variables are *independent* (unrelated to each other), their correlation will be 0. The correlation between the returns to Excelsior and Adirondack stock is a –0.2108, which indicates that the two variables show a slight tendency to move in opposite directions.

Population covariance and correlation coefficient

The population covariance measures the strength and the direction of the relationship between the elements of two populations. It's computed in a manner similar to the sample covariance.

You use the following formula to find the population covariance:

$$\sigma_{XY} = \frac{\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)}{n}$$

The key terms here are

- σ_{XY} = the population covariance between variables X and Y
- μ_X = the population mean for X
- μ_Y = the population mean for Y

- n = the number of elements in both populations
- i = an *index* that assigns a number to each population element, ranging from 1 to n
- X_i = a single element in the population for X
- Y_i = a single element in the population for Y
- Σ = the uppercase Greek letter sigma that indicates a sum is being computed

The population correlation coefficient is based on the population covariance. You use the following formula to find the population correlation coefficient:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

The key terms here are

ρ_{XY} = the population correlation coefficient between variables X and Y

σ_{XY} = the population covariance between variables X and Y

σ_X = the population standard deviation of variable X

σ_Y = the population standard deviation of variable Y

For example, suppose that two new companies were created in 2008: Theta Corp. and Eta Corp. The returns to the two companies' stocks from 2008 to 2012 are shown in Table 5-4:

Table 5-4 Annual Returns to Theta and Eta

Year	*Theta Corp. Annual Return (percent) (X)*	*Eta Corp. Annual Return (percent) (Y)*
2008	11	6
2009	9	5
2010	4	1
2011	2	9
2012	5	12

Because these companies have been in business only since 2008, each set of returns represents a *population* (the entire history of returns).

The population covariance and correlation between the returns to these stocks are computed as follows.

- The population mean of X is

$$\mu_X = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{(11+9+4+2+5)}{5} = \frac{31}{5} = 6.2$$

The population mean is obtained by summing all the elements of the population and then dividing by the population size. In this case, the 5 population elements sum to 31, and the population size is 5. Dividing these numbers gives a population mean of 6.2.

- The population mean of Y is

$$\mu_Y = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{(6+5+1+9+12)}{5} = \frac{33}{5} = 6.6$$

Table 5-5 shows the remaining calculations for the population covariance:

Table 5-5 Computing the Population Covariance

Year	Theta Corp. Annual Return (percent) (X)	Eta Corp. Annual Return (percent) (Y)	$(X_i - \mu_X)$	$(Y_i - \mu_Y)$	$(X_i - \mu_X)(Y_i - \mu_Y)$
2008	11	6	11 – 6.2 = 4.8	6 – 6.6 = –0.6	(4.8)(–0.6) = –2.88
2009	9	5	9 – 6.2 = 2.8	5 – 6.6 = –1.6	(2.8)(–1.6) = –4.48
2010	4	1	4 – 6.2 = –2.2	1 – 6.6 = –5.6	(–2.2)(–5.6) = 12.32
2011	2	9	2 – 6.2 = –4.2	9 – 6.6 = 2.4	(–4.2)(2.4) = –10.08
2012	5	12	5 – 6.2 = –1.2	12 – 6.6 = 5.4	(–1.2)(5.4) = –6.48
Mean	**6.2**	**6.6**		**Sum**	**–11.60**

The sum of the $(X_i - \mu_X)(Y_i - \mu_Y)$ column gives the numerator in the population covariance formula:

$$\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) = -11.60$$

The denominator equals the population size, which is 5. Therefore, the population covariance equals

$$\begin{aligned}\sigma_{XY} &= \frac{\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)}{n} \\ &= \frac{-11.60}{5} \\ &= -2.32\end{aligned}$$

To calculate the population correlation coefficient, divide the population covariance by the product of the population standard deviation of *X* and the population standard deviation of *Y:*

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

You find the population standard deviation of *X* by computing the population variance of *X* and then taking the square root of the result (as I explain in Chapter 4). Table 5-6 shows the calculations for the population variance of *X.*

Table 5-6 Computing the Population Variance for Theta

Year	Theta Corp. Annual Return (%) (X)	$(X_i - \mu_X)$	$(X_i - \mu_X)^2$
2008	11	11 – 6.2 = 4.8	$(4.8)^2 = 23.04$
2009	9	9 – 6.2 = 2.8	$(2.8)^2 = 7.84$
2010	4	4 – 6.2 = –2.2	$(-2.2)^2 = 4.84$
2011	2	2 – 6.2 = –4.2	$(-4.2)^2 = 17.64$
2012	5	5 – 6.2 = –1.2	$(-1.2)^2 = 1.44$
Mean	**6.2**	**Sum**	**54.80**

The sum of the $(X_i - \mu_X)^2$ column gives the numerator in the population variance formula. You divide this number by the population size to get the population variance of *X:*

$$\sigma_X^2 = \frac{\sum_{i=1}^{n}(X_i - \mu_X)^2}{n}$$
$$= \frac{54.8}{5}$$
$$= 10.96$$

The population standard deviation of X is the square root of 10.96, or $\sqrt{10.96} = 3.3106$.

Table 5-7 shows the calculations for the population variance of Y.

Table 5-7 Computing the Population Variance for Eta Corporation

Year	Eta Corp. Annual Return (percent) (Y)	$(Y_i - \mu_Y)$	$(Y_i - \mu_Y)^2$
2008	6	6 – 6.6 = –0.6	$(-0.6)^2 = 0.36$
2009	5	5 – 6.6 = –1.6	$(-1.6)^2 = 2.56$
2010	1	1 – 6.6 = –5.6	$(-5.6)^2 = 31.36$
2011	9	9 – 6.6 = 2.4	$(2.4)^2 = 5.76$
2012	12	12 – 6.6 = 5.4	$(5.4)^2 = 29.16$
Mean	**6.6**	**Sum**	**69.2**

Based on the calculations in Table 5-7, the population variance of Y equals

$$\sigma_Y^2 = \frac{\sum_{i=1}^{n}(Y_i - \mu_Y)^2}{n}$$
$$= \frac{69.2}{5}$$
$$= 13.84$$

The population standard deviation of Y equals the square root of 13.84, or $\sqrt{13.84} = 3.7202$.

Substituting these values into the population correlation formula gives you:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$
$$= \frac{-2.32}{(3.3106)(3.7202)}$$
$$= -0.1884$$

The negative result shows that there's a weak negative correlation between the stock returns of Theta and Eta.

Comparing correlation and covariance

When trying to find the relationship between two variables, you see that the correlation coefficient has several advantages over the covariance, including the following:

- The covariance has no lower or upper limits, whereas the correlation coefficient ranges between –1 and 1, making it easier to interpret its meaning.

  In the example with the returns to Excelsior and Adirondack stock (in the earlier section "Sample covariance and correlation"), the covariance is –1. Although this negative number indicates a tendency for the stock returns to move in opposite directions, it's difficult to judge the *strength* of this relationship. On the other hand, the correlation coefficient is –0.2108; because the correlation coefficient ranges from –1 to 1, you can see that the relationship between the stock returns is negative but not very strong.

- Unlike the covariance, the value of the correlation isn't affected by the units in which X and Y are measured. For example, suppose that a sample of tuna is chosen from the catch of two different fishing boats. The covariance between the weights of the tuna caught by the two boats is computed. The value of the covariance is different if the weights are expressed in kilograms or in pounds; however, the correlation is the same whether weights are expressed in kilograms or pounds.

To illustrate the second point further, say you record a sample of the average temperatures (in Celsius and Fahrenheit) in two cities from 2008 to 2012 and come up with the following results.

Year	*City 1 (Celsius)*	*City 2 (Celsius)*	*City 1 (Fahrenheit)*	*City 2 (Fahrenheit)*
2008	0.0°C	–10.0°C	32.0°F	14.0°F
2009	20.0°C	15.0°C	68.0°F	59.0°F
2010	–8.0°C	22.0°C	17.6°F	71.6°F
2011	25.0°C	30.0°C	77.0°F	86.0°F
2012	14.0°C	25.0°C	57.2°F	77.0°F
Mean	10.2°C	16.4°C	50.4°F	61.5°F

Assume that X represents the temperature in City 1 and Y represents the temperature in City 2. Table 5-8 shows the calculations for the covariance between the temperatures in Celsius of both cities.

Table 5-8 Covariance between Celsius Temperatures in City 1 and City 2

Year	City 1 (Celsius)	City 2 (Celsius)	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
2008	0.0°C	–10.0°C	0.0 – 10.2 = –10.2	–10.0 – 16.4 = –26.4	(–10.2)(–26.4) = 269.3
2009	20.0°C	15.0°C	20.0 – 10.2 = 9.8	15.0 – 16.4 = –1.4	(9.8)(–1.4) = –13.7
2010	–8.0°C	22.0°C	–8.0 – 10.2 = –18.2	22.0 – 16.4 = 5.6	(–18.2)(5.6) = –101.9
2011	25.0°C	30.0°C	25.0 – 10.2 = 14.8	30.0 – 16.4 = 13.6	(14.8)(13.6) = 201.3
2012	14.0°C	25.0°C	14.0 – 10.2 = 3.8	25.0 – 16.4 = 8.6	(3.8)(8.6) = 32.7
Mean	**10.2**°C	**16.4**°C		**Sum**	387.6

The $(X_i - \bar{X})$ column represents the differences between each temperature in City 1 and the sample mean. The $(Y_i - \bar{Y})$ column represents the differences between each temperature in City 2 and the sample mean. The $(X_i - \bar{X})(Y_i - \bar{Y})$ column is simply the product of the $(X_i - \bar{X})$ column and the $(Y_i - \bar{Y})$ column. The sum of the $(X_i - \bar{X})(Y_i - \bar{Y})$ column gives the numerator in the sample covariance formula, which is 387.6.

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = 387.6$$

The denominator equals the sample size minus one, which is 5 – 1 = 4 (because both samples have five elements, $n = 5$.) Therefore, the sample covariance equals

$$\begin{aligned} s_{XY} &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \\ &= \frac{387.6}{4} \\ &= 96.9 \end{aligned}$$

You find the sample standard deviation of X by computing the sample variance of X and then taking the square root of the result (see Chapter 4). Table 5-11 shows the calculations for the sample variance of X (Celsius temperatures for City 1):

Table 5-9	Sample Variance of *City 1*		
Year	**City 1 (Celsius)**	$\left(X_i - \bar{X}\right)$	$\left(X_i - \bar{X}\right)^2$
2008	0.0°C	0.0 – 10.2 = –10.2	$(-10.2)^2 = 104.0$
2009	20.0°C	20.0 – 10.2 = 9.8	$(9.8)^2 = 96.0$
2010	–8.0°C	–8.0 – 10.2 = –18.2	$(-18.2)^2 = 331.2$
2011	25.0°C	25.0 – 10.2 = 14.8	$(14.8)^2 = 219.0$
2012	14.0°C	14.0 – 10.2 = 3.8	$(3.8)^2 = 14.4$
Mean	**10.2**°C	**Sum**	**764.8**

To finish the calculation for the sample variance of *X,* you divide the sum of the terms in the $\left(X_i - \bar{X}\right)^2$ column by the sample size minus one, like so:

$$\begin{aligned} s_X^2 &= \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1} \\ &= \frac{764.8}{4} \\ &= 191.2 \end{aligned}$$

The sample standard deviation is the square root of the sample variance, or $\sqrt{191.2} = 13.8275$.

Following the same steps, you can find the sample variance of *Y* with the calculations in Table 5-10.

Table 5-10	Sample Variance of *City 2*		
Year	**City 2 (C)**	$\left(Y_i - \bar{Y}\right)$	$\left(Y_i - \bar{Y}\right)^2$
2008	–10.0	–10.0 – 16.4 = –26.4	$(-26.4)^2 = 697.0$
2009	15.0	15.0 – 16.4 = –1.4	$(-1.4)^2 = 2.0$
2010	22.0	22.0 – 16.4 = 5.6	$(5.6)^2 = 31.4$
2011	30.0	30.0 – 16.4 = 13.6	$(13.6)^2 = 185.0$
2012	25.0	25.0 – 16.4 = 8.6	$(8.6)^2 = 74.0$
Mean	**16.4**	**Sum**	**989.2**

To get the sample variance, divide the sum of the terms in the $\left(Y_i - \bar{Y}\right)^2$ column by the sample size minus one:

$$s_Y^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}{n-1}$$
$$= \frac{989.2}{4}$$
$$= 247.3$$

The sample standard deviation is the square root of the sample variance, or $\sqrt{247.3} = 15.7258$.

Next, substitute these values into the sample correlation formula:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$
$$= \frac{96.9}{(13.8275)(15.7258)}$$
$$= 0.4456$$

Repeating these same calculations for the temperatures in Fahrenheit, the covariance is 313.96 (compared with 96.9 when measured in Celsius) and the correlation remains at 0.4456. The covariance increases with Fahrenheit temperatures because the magnitude of the temperatures is greater, whereas the correlation isn't affected. The fact that the results depend on the units involved is one of the major drawbacks of using covariance instead of correlation.

Interpreting the Correlation Coefficient

Interpreting the correlation coefficient is easier than interpreting the covariance. Consider these examples:

- A correlation of 0.9 (close to the maximum value of 1.0) indicates a strong positive relationship between *X* and *Y;* when *X* increases, *Y* nearly always increases, and vice versa.

  A correlation of 0.2 (close to zero) indicates a weak positive relationship; when *X* increases, *Y* is somewhat more likely to increase than decrease, and vice versa.

- A correlation of –0.9 (close to the minimum value of –1.0) indicates a strong negative relationship between *X* and *Y.* Most of the time, when *X* increases, *Y* decreases; most of the time, when *X* decreases, *Y* increases.

  A correlation of –0.2 (close to zero) indicates a weak negative relationship; when *X* increases, *Y* is somewhat more likely to decrease than increase, and vice versa.

- A correlation of 0 indicates that *X* and *Y* are unrelated. When *X* increases or decreases, it has no direct effect on *Y* increasing or decreasing, and vice versa.

In the Fahrenheit and Celsius temperatures example in the previous section, the covariance was 96.9 for Celsius temperatures and 313.96 for Fahrenheit temperatures. Although the positive values indicate that the temperatures in both cities tend to increase or decrease at the same time, using the covariance measure alone makes it difficult to judge the *strength* of this relationship. On the other hand, the correlation for both Celsius and Fahrenheit temperatures was 0.4456, showing that a moderately strong, positive relationship exists between the temperatures in the two cities, whether measured in Celsius or Fahrenheit degrees.

In the following sections, you see a type of graph known as a *scatter plot* to illustrate the relationship between two different variables. An extremely important application of correlation is introduced; correlation can be used to show the degree of diversification that is present in a portfolio of stocks. In other words, the correlation can be used to determine how much the addition of a stock to a portfolio will affect the overall risk of the portfolio.

Showing the relationship between two variables

As I discuss in detail in Chapter 2, a *scatterplot* is a special type of graph that shows the relationship between two variables *X* and *Y*. The values of *X* are shown on the horizontal axis, and the values of *Y* are shown on the vertical axis.

Suppose that *X* represents a corporation's sales and *Y* represents its profits. Then *X* and *Y* would normally have a positive correlation between them, because higher sales tend to be associated with higher profits and vice versa. Figure 5-1 shows the relationship between two variables with a strong positive correlation.

Each point on the graph represents a corporation's sales (*X*) and its profits (*Y*) during a given year. The graph shows that as *X* increases, there's a strong tendency for *Y* to also increase. The straight line is known as a *trend line*. A trend line shows the direction of the points on a scatter plot. It can have a positive slope, a negative slope, or a zero slope (which means that the line is perfectly flat.) In this example, the trend line is positively sloped, which indicates that the correlation between *X* and *Y* is also positive. Because the points are extremely close to the trend line, the relationship between X and Y is very strong. With a weaker relationship, the points would be more scattered around the trend line.

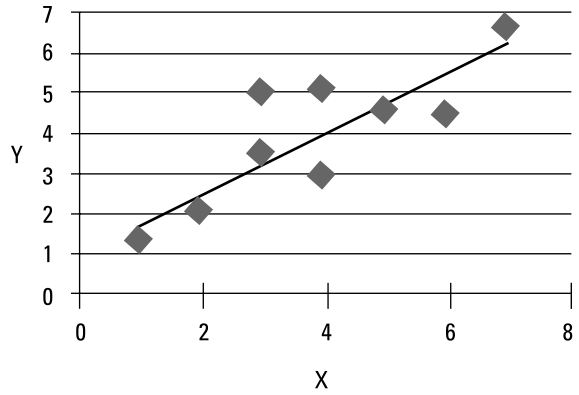

Figure 5-1: Scatterplot showing a strong positive relationship between *X* and *Y*.

Suppose that *X* represents a corporation's costs of production and *Y* represents its profits; then *X* and *Y* would normally have a negative correlation between them, because higher costs tend to be associated with lower profits and vice versa. Figure 5-2 shows the relationship between two variables with a strong negative correlation.

Each point on the graph represents a corporation's costs of production (*X*) and its profits (*Y*) during a given year. The graph shows that as *X* increases, there's a strong tendency for *Y* to decrease. The trend line has a negative slope, which indicates that the correlation between *X* and *Y* is negative.

By contrast, suppose that *X* represents the average daily temperature and *Y* represents a corporation's profits. Unless the corporation produces goods and services with a seasonal demand, these two variables are likely unrelated. Therefore, the correlation between *X* and *Y* will also be close to zero.

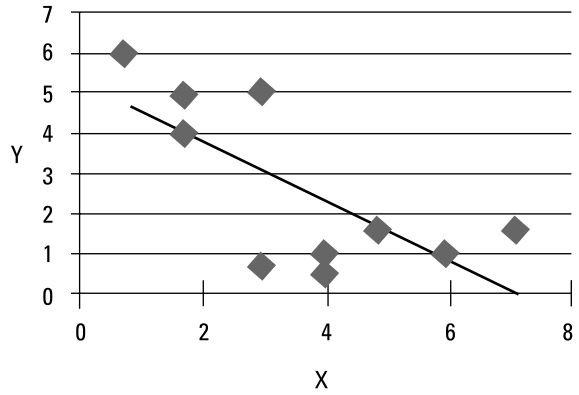

Figure 5-2: Scatterplot showing a strong negative relationship between *X* and *Y*.

Figure 5-3 shows the relationship between two unrelated variables.

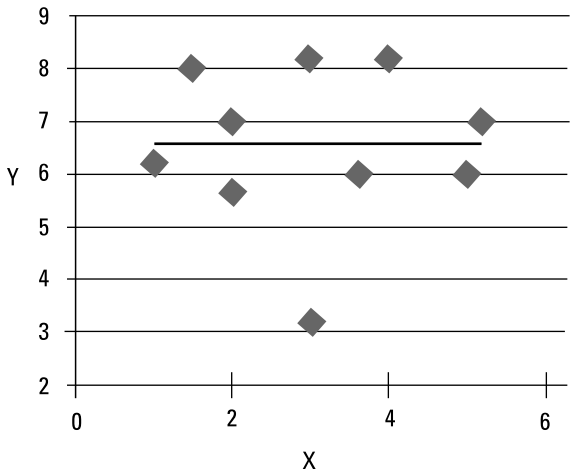

Figure 5-3: Scatterplot showing two unrelated variables.

Each point on the graph represents the average daily temperature (X) and a corporation's profits (Y) during a given year. The graph shows that as X increases, Y sometimes increases and sometimes decreases; no real pattern occurs. The trend line is almost perfectly flat, which indicates that the correlation between X and Y is very close to zero.

Application: Correlation and the benefits of diversification

You can measure the risk of a stock with the standard deviation of its returns. The greater the standard deviation, the further away the returns are from the mean on average (that is, the more "spread out" they are.) This indicates more uncertainty over the actual return during a given year, so the risk is greater. You can measure the diversification benefits of adding a stock to a portfolio with the correlation coefficient. The lower the correlation coefficient between two stocks, the *greater* is the reduction in risk and therefore the greater are the benefits of diversification.

For a portfolio of stocks, the risk depends not only on the standard deviations of the individual stocks but also on the *correlations* between the stocks. With low or negative correlations, the portfolio can experience significant

reductions in risk, which occurs because losses to some stocks tend to be offset by gains by other stocks at any given time. As a result, the variability of the portfolio's returns tends to be lower than the variability of the returns to the individual stocks.

The following data is a sample of returns to the stocks of Hilo, Inc., and Lohi Corp. during the past ten years.

Year	*Hilo*	*Lohi*
2003	0.03	0.10
2004	0.06	0.10
2005	0.07	0.08
2006	0.09	0.05
2007	0.08	0.04
2008	0.10	0.07
2009	0.09	0.01
2010	0.04	0.02
2011	0.02	0.10
2012	0.06	0.13

Table 5-11 summarizes the sample mean, variance, standard deviation, and coefficient of variation of the stock returns.

Table 5-11 Summary Measures for Hilo and Lohi

	Hilo	Lohi
Mean	0.0640	0.0700
Variance	0.0007	0.0015
Standard deviation	0.0272	0.0392
Coefficient of variation (CV)	42.44 percent	55.94 percent

The sample covariance between the stocks is –0.0004, and the sample correlation coefficient is –0.4179.

Assume that an investor purchased $100,000 of each stock for his portfolio at the start of 2003. The returns to the portfolio during this sample period are listed here.

Year	*Portfolio*
2003	0.065
2004	0.080
2005	0.075
2006	0.070
2007	0.060
2008	0.085
2009	0.050
2010	0.030
2011	0.060
2012	0.095

Because the portfolio is composed of 50 percent Hilo stock and 50 percent Lohi stock, you calculate the returns to the portfolio by multiplying the returns to each individual stock by 0.5 and combining the results, like so:

Portfolio return = 0.5(return to Hilo) + 0.5(return to Lohi)

For example, in 2003, the portfolio return is computed as follows:

Portfolio return = 0.5(0.03) + 0.5(0.10) = 0.065. Table 5-12 summarizes the sample mean, variance, standard deviation, and coefficient of variation of the portfolio returns.

Table 5-12 Portfolio Summary Measures

	Portfolio
Mean	0.0670
Variance	0.0003
Standard deviation	0.0186
Coefficient of variation (CV)	27.74 percent

The mean return to the portfolio is halfway between the mean returns to Hilo (0.0640) and Lohi (0.0700). The risk of the portfolio, as measured by the standard deviation of the returns, is only 0.0186 compared with Hilo (0.0272) and Lohi (0.0392). As a result, the portfolio's coefficient of variation is only 27.74 percent compared with Hilo at 42.442 percent and Lohi at 55.94 percent.

This substantial reduction in risk is due to the fact that the portfolio is well diversified, as seen by the negative correlation (–0.4179) between the returns to the two stocks.