Chapter 1

What Is Big Data and What Do You Do with It?

In This Chapter

- Understanding what big data is all about
- Seeing how data may be analyzed using Exploratory Data Analysis (EDA)
- Gaining insight into some of the key statistical techniques used to analyze big data

Big data refers to sets of data that are far too massive to be handled with traditional hardware. Big data is also problematic for software such as database systems, statistical packages, and so forth. In recent years, data-gathering capabilities have experienced explosive growth, so that storing and analyzing the resulting data has become progressively more challenging.

Many fields have been affected by the increasing availability of data, including finance, marketing, and e-commerce. Big data has also revolutionized more traditional fields such as law and medicine. Of course, big data is gathered on a massive scale by search engines such as Google and social media sites such as Facebook. These developments have led to the evolution of an entirely new profession: the *data scientist,* someone who can combine the fields of statistics, math, computer science, and engineering with knowledge of a specific application.

This chapter introduces several key concepts that are discussed throughout the book. These include the characteristics of big data, applications of big data, key statistical tools for analyzing big data, and forecasting techniques.

Characteristics of Big Data

The three factors that distinguish big data from other types of data are *volume, velocity,* and *variety.*

Clearly, with big data, the *volume* is massive. In fact, new terminology must be used to describe the size of these datasets. For example, one *petabyte* of data consists of 1.0×10^{15} bytes of data. That's 1,000 *trillion* bytes!

A *byte* is a single unit of storage in a computer's memory. A byte is used to represent a single number, character, or symbol. A byte consists of eight *bits,* each consisting of either a 0 or a 1.

Velocity refers to the speed at which data is gathered. Big datasets consist of data that's continuously gathered at very high speeds. For example, it has been estimated that Twitter users generate more than a quarter of a million tweets *every minute.* This requires a massive amount of storage space as well as real-time processing of the data.

Variety refers to the fact that the contents of a big dataset may consist of a number of different formats, including spreadsheets, videos, music clips, email messages, and so on. Storing a huge quantity of these incompatible types is one of the major challenges of big data.

Chapter 2 covers these characteristics in more detail.

Exploratory Data Analysis (EDA)

Before you apply statistical techniques to a dataset, it's important to examine the data to understand its basic properties. You can use a series of techniques that are collectively known as *Exploratory Data Analysis* (EDA) to analyze a dataset. EDA helps ensure that you choose the correct statistical techniques to analyze and forecast the data. The two basic types of EDA techniques are *graphical* techniques and *quantitative* techniques.

Graphical EDA techniques

Graphical EDA techniques show the key properties of a dataset in a convenient format. It's often easier to understand the properties of a variable and the relationships between variables by looking at graphs rather than looking at the raw data. You can use several graphical techniques, depending on the type of data being analyzed. Chapters 11 and 12 explain how to create and use the following:

- Box plots
- Histograms
- Normal probability plots
- Scatter plots

Quantitative EDA techniques

Quantitative EDA techniques provide a more rigorous method of determining the key properties of a dataset. Two of the most important of these techniques are

- Interval estimation (discussed in Chapter 11).
- Hypothesis testing (introduced in Chapter 5).

Interval estimates are used to create a *range* of values within which a variable is likely to fall. *Hypothesis* testing is used to test various propositions about a dataset, such as

- The mean value of the dataset.
- The standard deviation of the dataset.
- The probability distribution the dataset follows.

Hypothesis testing is a core technique in statistics and is used throughout the chapters in Part III of this book.

Statistical Analysis of Big Data

Gathering and storing massive quantities of data is a major challenge, but ultimately the biggest and most important challenge of big data is putting it to good use.

For example, a massive quantity of data can be helpful to a company's marketing research department only if it can identify the key drivers of the demand for the company's products. Political polling firms have access to massive amounts of demographic data about voters; this information must be analyzed intensively to find the key factors that can lead to a successful political campaign. A hedge fund can develop trading strategies from massive quantities of financial data by finding obscure patterns in the data that can be turned into profitable strategies.

Many statistical techniques can be used to analyze data to find useful patterns:

- Probability distributions are introduced in Chapter 4 and explored at greater length in Chapter 13.
- Regression analysis is the main topic of Chapter 15.
- Time series analysis is the primary focus of Chapter 16.
- Forecasting techniques are discussed in Chapter 17.

Probability distributions

You use a *probability distribution* to compute the probabilities associated with the elements of a dataset. The following distributions are described and applied in this book:

- **Binomial distribution:** You would use the binomial distribution to analyze variables that can assume only one of two values. For example, you could determine the probability that a given percentage of members at a sports club are left-handed. See Chapter 4 for details.
- **Poisson distribution:** You would use the Poisson distribution to describe the likelihood of a given number of events occurring over an interval of time. For example, it could be used to describe the probability of a specified number of hits on a website over the coming hour. See Chapter 13 for details.
- **Normal distribution:** The normal distribution is the most widely used probability distribution in most disciplines, including economics, finance, marketing, biology, psychology, and many others. One of the characteristic features of the normal distribution is *symmetry* — the probability of a variable being a given distance below the mean of the distribution equals the probability of it being the same distance above the mean. For example, if the mean height of all men in the United States is 70 inches, and heights are normally distributed, a randomly chosen man is equally likely to be between 68 and 70 inches tall as he is to be between 70 and 72 inches tall. See Chapter 4 and the chapters in Parts III and IV for details.

  The normal distribution works well with many applications. For example, it's often used in the field of finance to describe the returns to financial assets. Due to its ease of interpretation and implementation, the normal distribution is sometimes used even when the assumption of normality is only approximately correct.
- **The Student's t-distribution:** The Student's t-distribution is similar to the normal distribution, but with the Student's t-distribution, extremely small or extremely large values are much more likely to occur. This distribution is often used in situations where a variable exhibits too much variation to be consistent with the normal distribution. This is true when the properties of small samples are being analyzed. With small samples, the variation among samples is likely to be quite considerable, so the normal distribution shouldn't be used to describe their properties. See Chapter 13 for details.

  Note: The Student's t-distribution was developed by W.S. Gosset while employed at the Guinness brewing company. He was attempting to describe the properties of small sample means.

- **The chi-square distribution:** The chi-square distribution is appropriate for several types of applications. For example, you can use it to determine whether a population follows a particular probability distribution. You can also use it to test whether the variance of a population equals a specified value, and to test for the independence of two datasets. See Chapter 13 for details.
- **The F-distribution:** The F-distribution is derived from the chi-square distribution. You use it to test whether the variances of two populations equal each other. The F-distribution is also useful in applications such as regression analysis (covered next). See Chapter 14 for details.

Regression analysis

Regression analysis is used to estimate the strength and direction of the relationship between variables that are *linearly* related to each other. Chapter 15 discusses this topic at length.

Two variables X and Y are said to be *linearly* related if the relationship between them can be written in the form

$$Y = mX + b$$

where

m is the *slope,* or the change in Y due to a given change in X

b is the *intercept,* or the value of Y when $X = 0$

As an example of regression analysis, suppose a corporation wants to determine whether its advertising expenditures are actually increasing profits, and if so, by how much. The corporation gathers data on advertising and profits for the past 20 years and uses this data to estimate the following equation:

$$Y = 50 + 0.25X$$

where

Y represents the annual profits of the corporation (in millions of dollars).

X represents the annual advertising expenditures of the corporation (in millions of dollars).

In this equation, the slope equals 0.25, and the intercept equals 50. Because the slope of the regression line is 0.25, this indicates that on average, for every $1 million increase in advertising expenditures, profits rise by $.25 million, or $250,000. Because the intercept is 50, this indicates that with no advertising, profits would still be $50 million.

This equation, therefore, can be used to forecast future profits based on planned advertising expenditures. For example, if the corporation plans on spending $10 million on advertising next year, its expected profits will be as follows:

$$Y = 50 + 0.25X$$
$$Y = 50 + 0.25(10) = 50 + 2.5 = 52.5$$

Hence, with an advertising budget of $10 million next year, profits are expected to be $52.5 million.

Time series analysis

A *time series* is a set of observations of a single variable collected over time. This topic is talked about at length in Chapter 16. The following are examples of time series:

- The daily price of Apple stock over the past ten years.
- The value of the Dow Jones Industrial Average at the end of each year for the past 20 years.
- The daily price of gold over the past six months.

With time series analysis, you can use the statistical properties of a time series to predict the future values of a variable. There are many types of models that may be developed to explain and predict the behavior of a time series.

One place where time series analysis is used frequently is on Wall Street. Some analysts attempt to forecast the future value of an asset price, such as a stock, based entirely on the history of that stock's price. This is known as *technical analysis.* Technical analysts do not attempt to use other variables to forecast a stock's price — the only information they use is the stock's own history.

Technical analysis can work only if there are inefficiencies in the market. Otherwise, all information about a stock's history should already be reflected in its price, making technical trading strategies unprofitable.

Forecasting techniques

Many different techniques have been designed to forecast the future value of a variable. Two of these are time series regression models (Chapter 16) and simulation models (Chapter 17).

Time series regression models

A *time series regression model* is used to estimate the trend followed by a variable over time, using regression techniques. A *trend line* shows the direction in which a variable is moving as time elapses.

As an example, Figure 1-1 shows a time series that represents the annual output of a gold mine (measured in thousands of ounces per year) since the mine opened ten years ago.

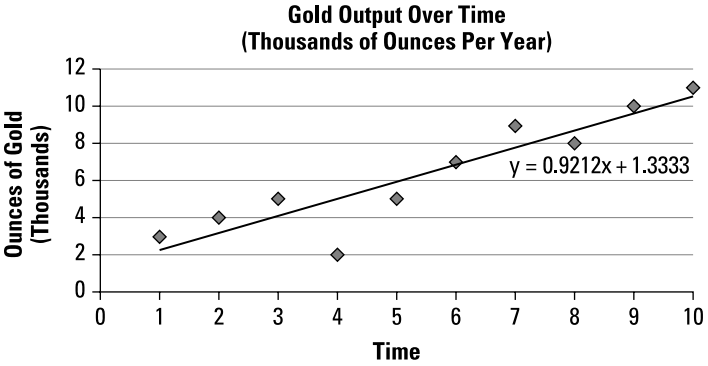

Figure 1-1: A time series showing gold output per year for the past ten years.

The equation of the trend line is estimated to be

$$Y = 0.9212X + 1.3333$$

where

X is the year.

Y is the annual production of gold (measured in thousands of ounces).

This trend line is estimated using regression analysis. The trend line shows that on average, the output of the mine grows by 0.9212 thousand (921.2 ounces) each year.

You could use this trend line to predict the output next year (the 11th year of operation) by substituting 11 for X, as follows:

$$Y = 0.9212X + 1.3333$$
$$Y = 0.9212(11) + 1.3333 = 11.4665$$

Based on the trend line equation, the mine would be expected to produce 11,466.5 ounces of gold next year.

Simulation models

You can use *simulation* models to forecast a time series. Simulation models are extremely flexible but can be extremely time-consuming to implement. Their accuracy also depends on assumptions being made about the time series data's statistical properties.

Two standard approaches to forecasting financial time series with simulation models are historical simulation and Monte Carlo simulation.

Historical simulation

Historical simulation is a technique used to generate a probability distribution for a variable as it evolves over time, based on its past values. If the properties of the variable being simulated remain stable over time, this technique can be highly accurate. One drawback to this approach is that in order to get an accurate prediction, you need to have a lot of data. It also depends on the assumption that a variable's past behavior will continue into the future.

As an example, Figure 1-2 shows a histogram that represents the returns to a stock over the past 100 days.

This histogram shows the probability distribution of returns on the stock based on the past 100 trading days. The graph shows that the most frequent return over the past 100 days was a loss of 2 percent, the second most frequent was a loss of 3 percent, and so on. You can use the information contained within this graph to create a probability distribution for the most likely return on this stock over the coming trading day.

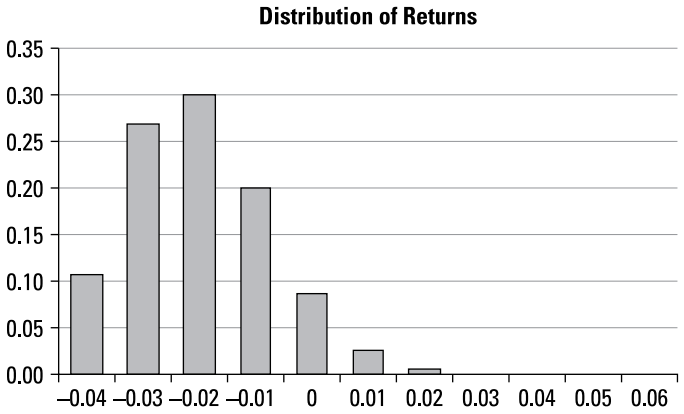

Figure 1-2: A histogram of stock returns.

Monte Carlo simulation

Monte Carlo simulation is a technique in which random numbers are substituted into a statistical model in order to forecast the future values of a variable. This methodology is used in many different disciplines, including finance, economics, and the hard sciences, such as physics. Monte Carlo simulation can work very well but can also be extremely time-consuming to implement. Also, its accuracy depends on the statistical model being used to describe the behavior of the time series.

As you can see, we've got a lot to cover in this book. But don't worry, we take it step by step. In Part I, we look at what big data is. We also build a statistical toolkit that we carry with us throughout the rest of the book. Part II focuses on the (extremely important) process of preparing data for the application of the techniques just described. Then we get to the good stuff in Parts III and IV. Though the equations can appear a little intimidating at times, we have labored to include examples in every chapter that make the ideas a little more accessible. So, take a deep breath and get ready to begin your exploration of big data!