

Introduction to Statistics and Business Analytics

LEARNING OBJECTIVES

The primary objective of Chapter 1 is to introduce you to the world of statistics and analytics, thereby enabling you to:

1. List quantitative and graphical examples of statistics within a business context.
2. Define important statistical terms, including population, sample, and parameter, as they relate to descriptive and inferential statistics.
3. Explain the difference between variables, measurement, and data.
4. Compare the four different levels of data: nominal, ordinal, interval, and ratio.
5. Define important business analytics terms including big data, business analytics, data mining, and data visualization.
6. List the four dimensions of big data and explain the differences between them.
7. Compare and contrast the three categories of business analytics.

Decision Dilemma

Statistics Describe the State of Business in India's Countryside

India is the second largest country in the world with more than 1.25 billion people. More than 70% of the people live in rural areas scattered about the countryside in 650,000 villages. In fact, it can be said that 1 in every 10 people in the world live in rural India. While it has a per capita income of less than \$1 (U.S.) per day, rural India, which has been described in the past as poor and semi-illiterate, now contributes to about one-half of the country's gross national product (GNP). However, rural India still has the most households in the world without electricity, over 300,000.

Despite its poverty and economic disadvantages, there are compelling reasons for companies to market their goods and services to rural India. The market of rural India has been growing at five times the rate of the urban Indian market. There is increasing agricultural productivity leading to growth in disposable income, and there is a reduction in the gap between the tastes of urban and rural customers. The literacy level is increasing, and people are becoming more conscious about their lifestyles and opportunities for a better life.



Hemis/Alamy Stock Photo

Around 60% of all middle-income households in India are in rural areas and more than one-third of all rural households in India now have a main source of income other than farming. Virtually every home has a radio, about one-third have a television, and more than one-half of rural households benefit from banking services. Forty-two percent of the people living in India's villages and small towns use toothpaste and that proportion is increasing as rural income rises and as there is greater awareness about oral hygiene.

In rural India, consumers are gaining more disposable income due to the movement of manufacturing jobs to rural areas. It is estimated that nearly 75% of the factories that opened in India in the past decade were built in rural areas. Products that are doing well in sales to people in rural India include televisions, fans, bicycles, bath soap, two- and three-wheelers, cars, and many others. According to MART, a New Delhi-based research organization, rural India buys 46% of all soft drinks and 49% of motorcycles sold in India. Because of such factors, many U.S. and Indian firms such as Microsoft, General Electric, Kellogg's, Colgate-Palmolive, Idea Cellular, Hindustan Lever, Godrej, Nirma Chemical Works, Novartis, Dabur, Tata Motors, and Vodafone India have entered the rural Indian market with enthusiasm. Marketing to rural customers often involves persuading them to try and to adopt products that they may not have used before. Rural India is a huge, relatively untapped market for businesses. However, entering such a market is not without risks and obstacles. The dilemma facing

companies is whether to enter this marketplace and, if so, to what extent and how.

Managerial, Statistical, and Analytical Questions

1. Are the statistics presented in this report exact figures or estimates?
2. How and where could business analysts have gathered such data?
3. In measuring the potential of the rural Indian marketplace, what other statistics could have been gathered?
4. What levels of data measurement are represented by data on rural India?
5. How can managers use these and other data to make better decisions about entering this marketplace?
6. What big data might be available on rural India?

Source: Adapted from "Marketing to Rural India: Making the Ends Meet," March 8, 2007, in *India Knowledge@Wharton*, at: <http://knowledge.wharton.upenn.edu/india/article.cfm?articleid=4172>; "Rural Segment Quickly Catching Up," September 2015, IBEF (India Brand Equity Foundation), at: www.ibef.org/industry/indian-rural-market.aspx; "Unlocking the Wealth in Rural Markets," June 2014, *Harvard Business Review*, at: <https://hbr.org/2014/06/unlocking-the-wealth-in-rural-markets>; "Much of Rural India Still Waits for Electricity," October 2013, University of Washington, at: <https://artsci.washington.edu/news/2013-10/much-rural-india-still-waits-electricity>

Every minute of the working day, decisions are made by businesses around the world that determine whether companies will be profitable and growing or whether they will stagnate and die. Most of these decisions are made with the assistance of information gathered about the marketplace, the economic and financial environment, the workforce, the competition, and other factors. Such information usually comes in the form of data or is accompanied by data. Business statistics and business analytics provide the tools through which such data are collected, analyzed, summarized, and presented to facilitate the decision-making process, and both business statistics and business analytics play an important role in the ongoing saga of decision-making within the dynamic world of business.

Virtually every area of business uses statistics in decision-making. Here are some recent examples:

- According to Truckinginfo.org, only 2.4% of all car accidents involve a commercial truck. Of truck problems that occur on the road and which result in a road call, 51.3% are due to tire problems. According to a study by Atlas Ergonomics, only 8% of truck drivers are female. One in six truck drivers is age 55 or older.
- A survey of 1465 workers by Hotjobs reports that 55% of workers believe that the quality of their work is perceived the same when they work remotely as when they are physically in the office.
- A survey of 477 executives by the Association of Executive Search Consultants determined that 48% of men and 67% of women say they are more likely to negotiate for less business travel compared with five years ago.
- According to usadata.com, the average age of an employee in the manufacturing industry is 43.7. The average salary of an employee in the manufacturing industry is \$59,351, with \$64,099 for males and \$47,683 for females.
- A Deloitte Retail "Green" survey of 1080 adults revealed that 54% agreed that plastic, noncompostable shopping bags should be banned.
- According the U.S. Bureau of Labor Statistics, the average annual expenditure per consumer on apparel is \$1,803; on entertainment is \$2,913; on reading is \$118; and on food at home is \$4,049.



Nick M Do/Photodisc/
Getty Images

William King/The Image Bank/
Getty Images

Comstock Images/Stockbyte/Getty Images

You can see from these few examples that there is a wide variety of uses and applications of statistics in business. Note that in several of these examples, business analysts have conducted a study and provided us rich and interesting information.

In this text we will examine several types of graphs for visualizing data as we study ways to arrange or structure data into forms that are both meaningful and useful to decision makers. We will learn about techniques for sampling from a population that allow studies of the business world to be conducted more inexpensively and in a more timely manner. We will explore various ways to forecast future values and examine techniques for predicting trends. This text also includes many statistical and analytics tools for testing hypotheses and for estimating population values. These and many other exciting statistics and statistical techniques await us on this journey through business statistics and analytics. Let us begin.

1.1 Basic Statistical Concepts

Business statistics, like many areas of study, has its own language. It is important to begin our study with an introduction of some basic concepts in order to understand and communicate about the subject. We begin with a discussion of the word *statistics*. The word *statistics* has many different meanings in our culture. *Webster's Third New International Dictionary* gives a comprehensive definition of **statistics** as *a science dealing with the collection, analysis, interpretation, and presentation of numerical data*. Viewed from this perspective, statistics includes all the topics presented in this text. **Figure 1.1** graphically displays the key elements of statistics.

The study of statistics can be organized in a variety of ways. One of the main ways is to subdivide statistics into two branches: descriptive statistics and inferential statistics. To understand the difference between descriptive and inferential statistics, definitions of *population* and *sample* are helpful. *Webster's Third New International Dictionary* defines **population** as *a collection of persons, objects, or items of interest*. The population can be a widely defined category, such as “all automobiles,” or it can be narrowly defined, such as “all Ford Mustang cars produced from 2017 to 2019.” A population can be a group of people, such as “all workers presently employed by Microsoft,” or it can be a set of objects, such as “all dishwashers produced on February 3, 2019, by the General Electric Company at the Louisville plant.” The researcher defines the population to be whatever he or she is studying. When researchers gather data from the whole population for a given measurement of interest, they call it a **census**.

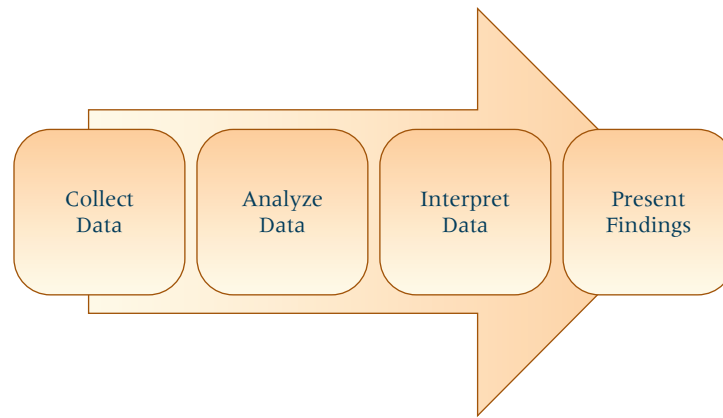


FIGURE 1.1 The Key Elements of Statistics

Most people are familiar with the U.S. Census. Every 10 years, the government attempts to count all persons living in this country.

A **sample** is a *portion of the whole* and, if properly taken, is representative of the whole. For various reasons (explained in Chapter 7), researchers often prefer to work with a sample of the population instead of the entire population. For example, in conducting quality-control experiments to determine the average life of lightbulbs, a lightbulb manufacturer might randomly sample only 75 lightbulbs during a production run. Because of time and money limitations, a human resources manager might take a random sample of 40 employees instead of using a census to measure company morale.

If a business analyst is *using data gathered on a group to describe or reach conclusions about that same group*, the statistics are called **descriptive statistics**. For example, if an instructor produces statistics to summarize a class's examination effort and uses those statistics to reach conclusions about that class only, the statistics are descriptive.

Many of the statistical data generated by businesses are descriptive. They might include number of employees on vacation during June, average salary at the Denver office, corporate sales for 2019, average managerial satisfaction score on a company-wide census of employee attitudes, and average return on investment for the Lofton Company for the years 2000 through 2019.

Another type of statistics is called **inferential statistics**. If a researcher *gathers data from a sample and uses the statistics generated to reach conclusions about the population from which the sample was taken*, the statistics are inferential statistics. The data gathered from the sample are used to infer something about a larger group. Inferential statistics are sometimes referred to as *inductive statistics*. The use and importance of inferential statistics continue to grow.

One application of inferential statistics is in pharmaceutical research. Some new drugs are expensive to produce, and therefore tests must be limited to small samples of patients. Utilizing inferential statistics, researchers can design experiments with small randomly selected samples of patients and attempt to reach conclusions and make inferences about the population.

Market researchers use inferential statistics to study the impact of advertising on various market segments. Suppose a soft drink company creates an advertisement depicting a dispensing machine that talks to the buyer, and market researchers want to measure the impact of the new advertisement on various age groups. The researcher could stratify the population into age categories ranging from young to old, randomly sample each stratum, and use inferential statistics to determine the effectiveness of the advertisement for the various age groups in the population. The advantage of using inferential statistics is that they enable the researcher to study effectively a wide range of phenomena without having to conduct a census. Most of the topics discussed in this text pertain to inferential statistics.

A *descriptive measure of the population* is called a **parameter**. Parameters are usually denoted by Greek letters. Examples of parameters are population mean (μ), population

variance (σ^2), and population standard deviation (σ). A *descriptive measure of a sample* is called a **statistic**. Statistics are usually denoted by Roman letters. Examples of statistics are sample mean (\bar{x}), sample variance (s^2), and sample standard deviation (s).

Differentiation between the terms *parameter* and *statistic* is important only in the use of inferential statistics. A business analyst often wants to estimate the value of a parameter or conduct tests about the parameter. However, the calculation of parameters is usually either impossible or infeasible because of the amount of time and money required to take a census. In such cases, the business analyst can take a random sample of the population, calculate a statistic on the sample, and infer by estimation the value of the parameter. The basis for inferential statistics, then, is the ability to make decisions about parameters without having to complete a census of the population.

For example, a manufacturer of washing machines would probably want to determine the average number of loads that a new machine can wash before it needs repairs. The parameter is the population mean or average number of washes per machine before repair. A company researcher takes a sample of machines, computes the number of washes before repair for each machine, averages the numbers, and estimates the population value or parameter by using the statistic, which in this case is the sample average. **Figure 1.2** demonstrates the inferential process.

Inferences about parameters are made under uncertainty. Unless parameters are computed directly from the population, the statistician never knows with certainty whether the estimates or inferences made from samples are true. In an effort to estimate the level of confidence in the result of the process, statisticians use probability statements. For this and other reasons, part of this text is devoted to probability (Chapter 4).

Business statistics is about measuring phenomena in the business world and organizing, analyzing, and presenting the resulting numerical information in such a way that better, more informed business decisions can be made. Most business statistics studies contain variables, measurements, and data.

In business statistics, a **variable** is a *characteristic of any entity being studied that is capable of taking on different values*. Some examples of variables in business include return on investment, advertising dollars, labor productivity, stock price, historical cost, total sales, market share, age of worker, earnings per share, miles driven to work, time spent in store shopping, and many, many others. In business statistics studies, most variables produce a measurement that can be used for analysis. A **measurement** is taken *when a standard process is used to assign numbers to particular attributes or characteristics of a variable*. Many measurements are obvious, such as time spent in a store shopping by a customer, age of the worker,

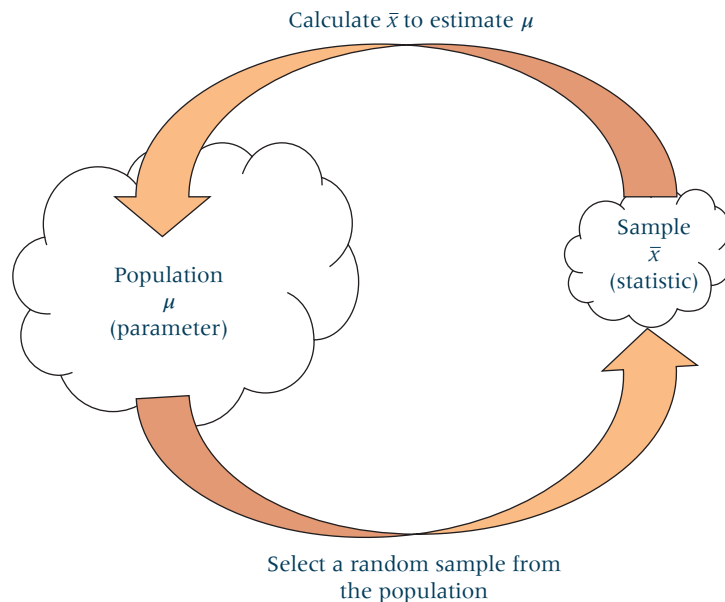


FIGURE 1.2 The Inferential Process

or the number of miles driven to work. However, some measurements, such as labor productivity, customer satisfaction, and return on investment, have to be defined by the business analyst or by experts within the field. Once such measurements are recorded and stored, they can be denoted as “data.” It can be said that **data** are *recorded measurements*. The processes of measuring and data gathering are basic to all that we do in business statistics and analytics. It is data that are analyzed by business statisticians and analysts in order to learn more about the variables being studied. Sometimes, sets of data are organized into databases as a way to store data or as a means for more conveniently analyzing data or comparing variables. Valid data are the lifeblood of business statistics and business analytics, and it is important that the business analyst give thoughtful attention to the creation of meaningful, valid data before embarking on analysis and reaching conclusions.

1.2 Data Measurement



Immense volumes of numerical data are gathered by businesses every day, representing myriad items. For example, numbers represent dollar costs of items produced, geographical locations of retail outlets, weights of shipments, and rankings of subordinates at yearly reviews. All such data should not be analyzed the same way statistically because the entities represented by the numbers are different. For this reason, the business analyst needs to know the *level of data measurement* represented by the numbers being analyzed.

The disparate use of numbers can be illustrated by the numbers 40 and 80, which could represent the weights of two objects being shipped, the ratings received on a consumer test by two different products, or football jersey numbers of a fullback and a wide receiver. Although 80 pounds is twice as much as 40 pounds, the wide receiver is probably not twice as big as the fullback! Averaging the two weights seems reasonable, but averaging the football jersey numbers makes no sense. The appropriateness of the data analysis depends on the level of measurement of the data gathered. The phenomenon represented by the numbers determines the level of data measurement. Four common levels of data measurement follow.

1. Nominal
2. Ordinal
3. Interval
4. Ratio

Nominal is the lowest level of data measurement followed by ordinal, interval, and ratio. Ratio is the highest level of data measurement, as shown in **Figure 1.3**.

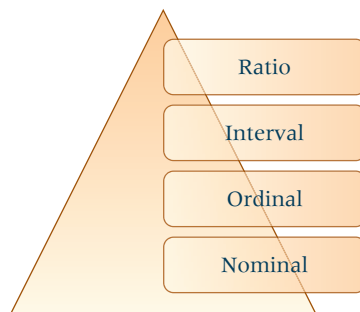
Nominal Level

The *lowest level of data measurement* is the **nominal level**. Numbers representing nominal-level data (the word *level* often is omitted) can be *used only to classify or categorize*. Employee identification numbers are an example of nominal data. The numbers are used only to differentiate employees and not to make a value statement about them. Many demographic questions in surveys result in data that are nominal because the questions are used for classification only. The following is an example of such a question that would result in nominal data:

Which of the following employment classifications best describes your area of work?

1. Educator
2. Construction worker
3. Manufacturing worker
4. Lawyer
5. Doctor
6. Other

Highest Level of Data Measurement



Lowest Level of Data Measurement

FIGURE 1.3 Hierarchy of Levels of Data

Suppose that, for computing purposes, an educator is assigned a 1, a construction worker is assigned a 2, a manufacturing worker is assigned a 3, and so on. These numbers should be used only to classify respondents. The number 1 does not denote the top classification. It is used only to differentiate an educator (1) from a lawyer (4).

Some other types of variables that often produce nominal-level data are sex, religion, ethnicity, geographic location, and place of birth. Social Security numbers, telephone numbers, employee ID numbers, and Zip codes are further examples of nominal data. Statistical techniques that are appropriate for analyzing nominal data are limited. However, some of the more widely used statistics, such as the chi-square statistic, can be applied to nominal data, often producing useful information.

Ordinal Level

Ordinal-level data measurement is higher than the nominal level. In addition to the nominal-level capabilities, ordinal-level measurement can be used to rank or order people or objects. For example, using ordinal data, a supervisor can evaluate three employees by ranking their productivity with the numbers 1 through 3. The supervisor could identify one employee as the most productive, one as the least productive, and one as somewhere in between by using ordinal data. However, the supervisor could not use ordinal data to establish that the intervals between the employees ranked 1 and 2 and between the employees ranked 2 and 3 are equal; that is, she could not say that the differences in the amount of productivity between workers ranked 1, 2, and 3 are necessarily the same. With ordinal data, the distances or spacing represented by consecutive numbers are not always equal.

Some questionnaire Likert-type scales are considered by many researchers to be ordinal in level. The following is an example of one such scale:

This computer tutorial is	—	—	—	—	—
	not	somewhat	moderately	very	extremely
	helpful	helpful	helpful	helpful	helpful
	1	2	3	4	5

When this survey question is coded for the computer, only the numbers 1 through 5 will remain, not the adjectives. Virtually everyone would agree that a 5 is higher than a 4 on this scale and that ranking responses is possible. However, most respondents would not consider the differences between not helpful, somewhat helpful, moderately helpful, very helpful, and extremely helpful to be equal.

Mutual funds as investments are sometimes rated in terms of risk by using measures of default risk, currency risk, and interest rate risk. These three measures are applied to investments by rating them as having high, medium, or low risk. Suppose high risk is assigned a 3, medium risk a 2, and low risk a 1. If a fund is awarded a 3 rather than a 2, it carries more risk, and so on. However, the differences in risk between categories 1, 2, and 3 are not necessarily equal. Thus, these measurements of risk are only ordinal-level measurements. Another example of the use of ordinal numbers in business is the ranking of the top 50 most admired companies in *Fortune* magazine. The numbers ranking the companies are only ordinal in measurement. Certain statistical techniques are specifically suited to ordinal data, but many other techniques are not appropriate for use on ordinal data. For example, it does not make sense to say that the average of “moderately helpful” and “very helpful” is “moderately helpful and a half.”

Because nominal and ordinal data are often derived from imprecise measurements such as demographic questions, the categorization of people or objects, or the ranking of items, *nominal and ordinal data* are **nonmetric data** and are sometimes referred to as *qualitative data*.

Interval Level

Interval-level data measurement is the *next to the highest level of data in which the distances between consecutive numbers have meaning and the data are always numerical*. The distances represented by the differences between consecutive numbers are equal; that is, interval data

have equal intervals. An example of interval measurement is Fahrenheit temperature. With Fahrenheit temperature numbers, the temperatures can be ranked, and the amounts of heat between consecutive readings, such as 20°, 21°, and 22°, are the same.

In addition, with interval-level data, the zero point is a matter of convention or convenience and not a natural or fixed zero point. Zero is just another point on the scale and does not mean the absence of the phenomenon. For example, zero degrees Fahrenheit is not the lowest possible temperature. Some other examples of interval-level data are the percentage change in employment, the percentage return on a stock, and the dollar change in stock price.

Ratio Level

Ratio-level data measurement is *the highest level of data measurement*. Ratio data have the same properties as interval data, but ratio data have an *absolute zero*, and *the ratio of two numbers is meaningful*. The notion of absolute zero means that zero is fixed, and *the zero value in the data represents the absence of the characteristic being studied*. The value of zero cannot be arbitrarily assigned because it represents a fixed point. This definition enables the statistician to create *ratios* with the data.

Examples of ratio data are height, weight, time, volume, and kelvin temperature. With ratio data, a researcher can state that 180 pounds of weight is twice as much as 90 pounds or, in other words, make a ratio of 180:90. Many of the data measured by valves or gauges in industry are ratio data.

Other examples in the business world that are ratio level in measurement are production cycle time, work measurement time, passenger miles, number of trucks sold, complaints per 10,000 fliers, and number of employees.

Because interval- and ratio-level data are usually gathered by precise instruments often used in production and engineering processes, in national standardized testing, or in standardized accounting procedures, they are called **metric data** and are sometimes referred to as *quantitative data*.

Comparison of the Four Levels of Data

Figure 1.4 shows the relationships of the usage potential among the four levels of data measurement. The concentric squares denote that each higher level of data can be analyzed by any of the techniques used on lower levels of data but, in addition, can be used in other statistical techniques. Therefore, ratio data can be analyzed by any statistical technique applicable to the other three levels of data plus some others.

Nominal data are the most limited data in terms of the types of statistical analysis that can be used with them. Ordinal data allow the researcher to perform any analysis that can be done with nominal data and some additional analyses. With ratio data, a statistician can make ratio comparisons and appropriately do any analysis that can be performed on nominal, ordinal,

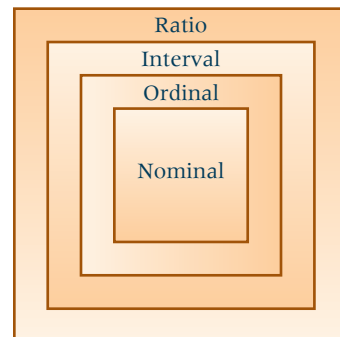


FIGURE 1.4 Usage Potential of Various Levels of Data

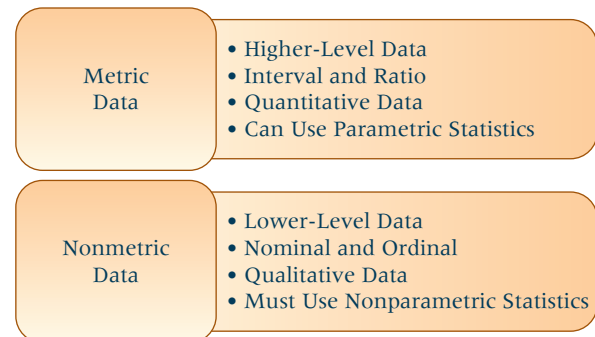


FIGURE 1.5 Metric vs. Nonmetric Data

or interval data. Some statistical techniques require ratio data and cannot be used to analyze other levels of data.

Statistical techniques can be separated into two categories: parametric statistics and nonparametric statistics. **Parametric statistics** require that data be interval or ratio. If the data are nominal or ordinal, **nonparametric statistics** must be used. Nonparametric statistics can also be used to analyze interval or ratio data. This text focuses largely on parametric statistics, with the exception of Chapter 16 and Chapter 17, which contain nonparametric techniques. Thus much of the material in this text requires that data be interval or ratio data.

Figure 1.5 contains a summary of metric data and nonmetric data.

DEMONSTRATION PROBLEM 1.1

Because of increased competition for patients among providers and the need to determine how providers can better serve their clientele, hospital administrators sometimes administer a quality satisfaction survey to their patients after the patient is released. The following types of questions are sometimes asked on such a survey. These questions will result in what level of data measurement?

1. How long ago were you released from the hospital?
2. Which type of unit were you in for most of your stay?
 - Coronary care
 - Intensive care
 - Maternity care
 - Medical unit
 - Pediatric/children's unit
 - Surgical unit
3. In choosing a hospital, how important was the hospital's location?

(circle one)

Very	Somewhat	Not Very	Not at All
Important	Important	Important	Important

4. What was your body temperature when you were admitted to the hospital?
5. Rate the skill of your doctor:

Excellent Very Good Good Fair Poor

Solution Question 1 is a time measurement with an absolute zero and is therefore a ratio-level measurement. A person who has been out of the hospital for two weeks has been out twice as long as someone who has been out of the hospital for one week.

Question 2 yields nominal data because the patient is asked only to categorize the type of unit he or she was in. This question does not require a hierarchy or ranking of the type of unit. Questions 3 and 5 are likely to result in ordinal-level data. Suppose a number is assigned to the descriptors in these two questions. For question 3, “very important” might be assigned a 4, “somewhat important” a 3, “not very important” a 2, and “not at all important” a 1. Certainly, the higher the number, the more important is the hospital's location. Thus, these responses can be ranked by selection. However, the increases in importance from 1 to 2 to 3 to 4 are not necessarily equal. This same logic applies to the numeric values assigned in question 5. In question 4, body temperature, if measured on a Fahrenheit or Celsius scale, is interval in measurement.

1.3 Introduction to Business Analytics

In the current world of business, there is exponential growth in the data available to decision makers to assist them in producing better business outcomes. Growing sources of data are available from the Internet, social media, governments, transportation systems, healthcare,

environmental organizations, and a plethora of business data sources, among others. Business data include but are not limited to consumer information, labor statistics, financials, product and resource tracking information, supply chain information, operations, and human resource information. According to vCloud, every day 2.5 quintillion bytes of data are created. As a business example, from its millions of products and hundreds of millions of customers, Walmart alone collects multi-terabytes of new data every day that are then added to its petabytes of historical data.¹

Big Data

The advent of such growth in the numbers and types of data available to researchers, data scientists, and business decision makers has resulted in a new term, **big data**. Big data has been defined to be *a collection of large and complex datasets from different sources that are difficult to process using traditional data management and processing applications.*² In addition, big data can be seen as *a large amount of either organized or unorganized data that is analyzed to make an informed decision or evaluation.*³ All data are not created in the same way, nor do they represent the same things. Thus, researchers recognize that there are at least four characteristics or dimensions associated with big data.⁴ These are:

1. variety
2. velocity
3. veracity
4. volume

Variety refers to the many different forms of data based on data sources. A wide variety of data is available from such sources as mobile phones, videos, text, retail scanners, Internet searches, government documents, multimedia, empirical research, and many others. Data can be structured (such as databases or Excel sheets) or unstructured (such as writing and photographs). **Velocity** refers to the speed with which the data are available and with which the data can be processed. The velocity characteristic of data is important in ensuring that data are current and updated in real time.⁵ **Veracity** of data has to do with data quality, correctness, and accuracy.⁶ Data lacking veracity may be imprecise, unrepresentative, inferior, and untrustworthy. In using such data to better understand business decisions, it might be said that the result is “garbage in, garbage out.” Veracity indicates reliability, authenticity, legitimacy, and validity in the data. **Volume** has to do with the ever-increasing size of the data and databases. Big data produces vast amounts of data as exemplified by the Walmart example mentioned above. A fifth characteristic or dimension of data that is sometimes considered is Value. Analysis of data which does not generate value offers no contribution to an organization.⁷

Along with this unparalleled explosion of data, major advances have been made in computing performance, network functioning, data handling, device mobility, and others. Capitalizing on these developments, new opportunities are becoming available for discovering greater and deeper insights into making significant improvements in the way businesses operate and function. So how do businesses go about capitalizing on this potential new information?

Business Analytics

Rising to meet the challenges, opportunities, and potentialities presented to business decision makers through big data, a new field of business has emerged, **business analytics**. There are different definitions of business analytics in the literature. However, one that more accurately describes the intent of the approach here is that *business analytics is the application of processes and techniques that transform raw data into meaningful information to improve decision-making.*⁸ Because big data sources are too large and too complex, new questions have arisen that cannot always be effectively answered with traditional analysis methods.⁹ In light of this,

new methodologies and processing techniques have been developed, giving birth to a new era in decision-making referred to as the business analytics period.¹⁰ There are other names sometimes given to business analytics, such as business intelligence, data science, data analytics, analytics, and even big data, but the goal in all cases is to convert data into actionable insight for more timely and accurate decision-making.¹¹ For example, when applied to the business environment, data analytics becomes synonymous with business analytics.¹² Business analytics provides the added value to data, resulting in deeper and broader business insights that can be used to improve decision makers' insights and understandings in all aspects of business, as shown in **Figure 1.6**.

There are many new opportunities for employment in business analytics, yet there is presently a shortage of available talent in the field. In fact, an IBM survey of 900 business and information technology executives cited the lack of business analytics skills as a top business challenge.¹³ Executives of today say that they are seeking data-driven leaders. They are looking for people who can work comfortably with both data and people—managers who can build useful models from data and also lead the team that will put them into practice.¹⁴ A paper on big data released by the McKinsey Global Institute said that by the year 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the knowledge and skill to use the outcomes of big data analysis to make effective decisions.¹⁵

Categories of Business Analytics

It might be said that the mission of business analytics is to apply processes and techniques to transform raw data into meaningful information. A plethora of such techniques are available and used to carry out this mission, drawing from such areas as statistics, operations research, mathematical modeling, data mining, and artificial intelligence, just to name a few. The various techniques and approaches provide different information to decision makers. Accordingly, the business analytics community has organized and classified business analytic tools into three main categories: descriptive analytics, predictive analytics, and prescriptive analytics.

Descriptive Analytics The simplest and perhaps the most commonly used of the three categories of business analytics is **descriptive analytics**. Often the first step in the analytic process, descriptive analytics takes traditional data and describes what has or is happening in a business. It can be used to condense big data into smaller, more useful, data.¹⁶ Also referred to sometimes as reporting analytics, descriptive analytics can be used to discover hidden relationships in the data and identify previously undiscovered patterns.¹⁷ Descriptive analytics drills down into the data to uncover useful and important details and mines data for trends.¹⁸ At this step, visualization can be a key player in presenting information.

Much of what is taught in a traditional introductory statistics course could be classified as descriptive analytics, including descriptive statistics, frequency distributions, discrete distributions, continuous distributions, sampling distributions, and statistical inference.¹⁹ One could probably add correlation and various clustering techniques to the mix along with data mining and data visualization techniques.

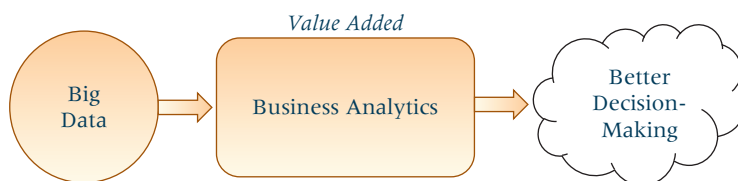


FIGURE 1.6 Business Analytics Adds Value to Data

Predictive Analytics The next step in data reduction is **predictive analytics**. It finds relationships in the data that are not readily apparent with descriptive analytics. With predictive analytics, patterns or relationships are extrapolated forward in time and the past is used to make predictions about the future.²⁰ Also, it provides answers that move beyond using the historical data as the principal basis for decisions.²¹ Predictive analytics builds and assesses algorithmic models aimed at making empirical rather than theoretical predictions and is designed to predict future observations.²² It can help managers develop likely scenarios.²³

Topics in predictive analytics can include regression, time-series, forecasting, simulation, data mining, statistical modeling, machine learning techniques, and others. They can also include classifying techniques, such as decision tree models and neural networks.²⁴

Prescriptive Analytics The final stage of business analytics, still in its early stages of development, is **prescriptive analytics**.²⁵ Prescriptive analytics follows descriptive and predictive analytics in an attempt to find the best course of action under certain circumstances.²⁶ The goal is to examine current trends and likely forecasts and use that information to make better decisions.²⁷ Prescriptive analytics takes uncertainty into account, recommends ways to mitigate risks, and tries to see what the effect of future decisions will be in order to adjust the decisions before they are actually made.²⁸ It does this by exploring a set of possible actions based on descriptive and predictive analyses of complex data and suggesting courses of action.²⁹ It evaluates data and determines new ways to operate while balancing all constraints.³⁰ Prescriptive analytics can continually and automatically process new data to improve recommendations and provide better decision options.³¹ It not only foresees what will happen and when but also why it will happen and provides recommendations as to how to act in order to take advantage of the predictions.³² Predictive analytics involves using a set of mathematical techniques that computationally determine the optimal action of decision-making given a complex set of objectives, requirements, and constraints.³³

Topics in prescriptive analytics traditionally come from the fields of management science or operations research and are generally those aimed at optimizing the performance of a system,³⁴ such as mathematical programming, simulation, network analysis, and others.

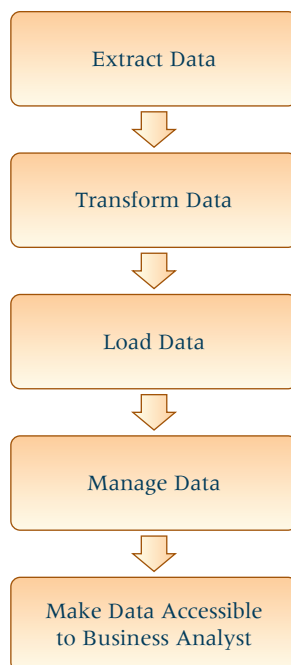


FIGURE 1.7 Process of Data Mining

Data Mining

The dawning of the big data era has given rise to new and promising prospects for improving business decisions and outcomes. A main component in the process of transforming the mountains of data now available into useful business information is **data mining**. In the field of business, data mining is *the process of collecting, exploring, and analyzing large volumes of data in an effort to uncover hidden patterns and/or relationships that can be used to enhance business decision-making*. In short, data mining is a process used by companies to turn raw data into useful information. A main objective of data mining is to uncover patterns or relationships in data that are meaningful in that they can lead to some business advantage. Data mining allows businesspeople to discover and interpret useful information, thereby helping them to make more knowledgeable decisions and better serve their customers and clients.

Figure 1.7 displays the process of data mining, which involves finding the data, converting the data into useful forms, loading the data into a holding or storage location, managing the data, and making the data accessible to business analytics users. The first three steps of this process are often referred to ETL (extract, transform, and load) by data scientists. Extracting encompasses locating the data by asking the question, “Where can it be found?” Countless data are unearthed the world over in a great variety of forms and in multiple disparate sources. Because of this, extracting data can be the most time-consuming step.³⁵ After a set of data has been located and extracted, it must be transformed or converted into a usable

form. Included in this transformation may be a sorting process to determine which data are useful and which are not. In addition, data are typically “cleaned” by removing corrupt or incorrect records and identifying incomplete, incorrect, or irrelevant parts of the data.³⁶ Often the data are sorted into columns and rows to improve usability and searchability.³⁷ After the set of data is extracted and transformed, it is loaded into an end target, which is often a database. Data are often managed through a database management system, a software system that enables users to define, create, maintain, and control access to the database.³⁸ Lastly, the ultimate goal of the process of data mining is to make the data accessible and usable to the business analyst.

Data Visualization

As business organizations amass large reservoirs of data, one swift and easy way to obtain an overview of the data is through data visualization. It has been said that data visualization is perhaps the most useful component of business analytics and what makes it truly unique.³⁹ So what is data visualization? Generally speaking, **data visualization** is any attempt made by data analysts to help individuals better understand data by putting it in a visual context. Specifically, data visualization is *the study of the visual representation of data and is employed to convey data or information by imparting it as visual objects displayed in graphics*.

Interpretation of data is vital to unlocking the potential value held within and to making the most informed decisions.⁴⁰ Utilizing such visual techniques to convey information hidden in the data allows for a broader audience with a wider range of backgrounds to view and understand its meaning. Data visualization tools aid in making data-driven insights accessible to people at all levels throughout an organization and can reveal surprising patterns and connections, resulting in improved decision-making. To communicate information clearly and efficiently, data visualization uses statistical graphics, plots, information graphics, and other tools. Numerical data may be encoded using dots, lines, or bars to visually communicate a quantitative message, thereby making complex data more accessible, understandable, and usable.⁴¹

Visualization Example As an example of data visualization, consider the top five contractors to the U.S. Treasury and their respective dollars obligated in a recent year displayed in **Table 1.1**.

Global Vendor Name	Dollars Obligated
Coins 'N Things Inc.	\$529,070,983.45
Spectrum Group International Inc.	\$415,013,005.82
Sunshine Minting Inc.	\$406,300,921.45
Deloitte LLP	\$205,655,400.82
Crane & Co. Inc.	\$173,888,697.30

One of the leading companies in the development of data visualization software is Tableau Software. Shown in **Figure 1.8** is a bubble chart of the Table 1.1 data developed by using Tableau. One of the advantages of using visualization to display data is the variety of types of graphs and charts. To give the user a different perspective, **Figure 1.9** contains a Tableau-produced bar graph of the same information.



FIGURE 1.8 Bubble Chart of the Top Five Government Contractors to the U.S. Treasury

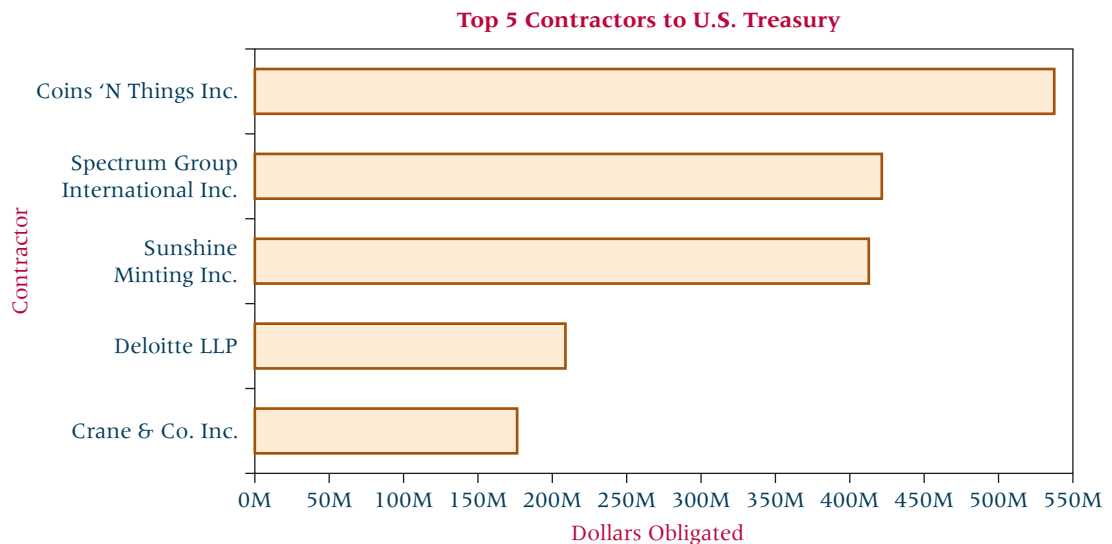


FIGURE 1.9 Bar Chart of the Top Five Government Contractors to the U.S. Treasury

Statistical Analysis Using the Computer: Excel, Minitab, and Tableau

The advent of the modern computer opened many new opportunities for statistical analysis. The computer allows for storage, retrieval, and transfer of large data sets. Furthermore, computer software has been developed to analyze data by means of sophisticated statistical techniques. Some widely used statistical techniques, such as multiple regression, are so tedious and cumbersome to compute manually that they were of little practical use to researchers before computers were developed.

Business statisticians use a variety of popular statistical software packages, including Minitab, SAS, and SPSS. Many computer spreadsheet software packages also have the capability of analyzing data statistically. In this text, the computer statistical output presented is from Excel, Minitab, and Tableau.

Thinking Critically About Statistics in Business Today

Cellular Phone Use in Japan

The Communications and Information Network Association of Japan (CIAJ) conducts an annual study of cellular phone use in Japan. In a recent year a survey was taken of 1200 mobile phone users split evenly between men and women and almost equally over six age brackets of people residing in the larger Tokyo and Osaka metropolitan areas. The study produced several interesting findings. Of the respondents, 76.8% said that their main-use terminal was a smartphone while 23.2% said it was a feature phone, a mobile phone that incorporates features such as the ability to access the Internet and store and play music but lacks the advanced functionality of a smartphone. Of the smartphone users, 49.2% reported that their second device was a tablet (with a telecom subscription) compared to the previous year where 36.7% said that a tablet was their second device. Survey participants were asked what the decisive factors in purchasing a smartphone are. The number one decisive factor was battery life, selected by over

70%, followed by manufacturer/brand, selected by 67.5%. Survey participants who owned feature phones were asked the same question and the number one decisive factor was monthly cost (84.7%) followed by purchase price of handset (84.4%). In terms of usage of smartphone features and services, the number-one most popular use was “searching the Internet” (90.2%) followed by “shooting still photos with camera” (88.7%).

Things to Ponder

1. In what way was this study an example of inferential statistics?
2. What is the population of this study?
3. What are some of the variables being studied?
4. How might a study such as this yield information that is useful to business decision makers?

Source: “CIAJ Releases Report on the Study of Mobile Phone Use,” July 29, 2015, at <https://www.ciaj.or.jp/en/news/news2015/92.html>

Decision Dilemma Solved

Statistics Describe the State of Business in India’s Countryside

Several statistics were reported in the Decision Dilemma about rural India. The authors of the sources from which the Decision Dilemma was drawn never stated whether the reported statistics were based on actual data drawn from a census of rural Indian households or on estimates taken from a sample of rural households. If the data came from a census, then the totals, averages, and percentages presented in the Decision Dilemma are parameters. If, on the other hand, the data were gathered from samples, then they are statistics. Although governments especially do conduct censuses and at least some of the reported numbers could be parameters, more often than not, such data are gathered from samples of people or items. For example, in rural India, the government, academicians, or business analysts could have taken random samples of households, gathering consumer statistics that were then used to estimate population parameters, such as percentage of households with televisions and so forth.

In conducting research on a topic like consumer consumption in rural India, there is potential for a wide variety of statistics to be gathered that represent several levels of data. For example, ratio-level measurements on items such as income, number of children, age of household heads, number of livestock, and grams of toothpaste consumed per year might be obtained. On the other

hand, if researchers use a Likert scale (1 to 5 measurements) to gather responses about the interests, likes, and preferences of rural Indian consumers, an ordinal-level measurement would be obtained, as would the ranking of products or brands in market research studies. Other variables, such as geographic location, sex, occupation, or religion, are usually measured with nominal data.

The decision to enter the rural Indian market is not just a marketing decision. It involves production capacity and schedule issues, transportation challenges, financial commitments, managerial growth or reassignment, accounting issues (accounting for rural India may differ from techniques used in traditional markets), information systems, and other related areas. With so much on the line, company decision makers need as much relevant information to be available as possible. In this Decision Dilemma, it is obvious to the decision maker that rural India is still quite poor and illiterate. Its capacity as a market is great. The statistics on the increasing sales of a few personal-care products look promising. What are the future forecasts for the earning power of people in rural India? Will major cultural issues block the adoption of the types of products that companies want to sell there? The answers to these and many other interesting and useful questions can be obtained by the appropriate use of statistics. The 800 million people living in rural India represent the second largest group of people in the world. It certainly is a market segment worth studying further.

Ethical Considerations

With the abundance and proliferation of statistical data, potential misuse of statistics in business dealings is a concern. It is, in effect, unethical business behavior to use statistics out of context. Unethical businesspeople might use only selective data from studies to underscore their point, omitting statistics from the same studies that argue against their case. The results of statistical studies can be misstated or overstated to gain favor.

This chapter noted that if data are nominal or ordinal, then only nonparametric statistics are appropriate for analysis. The use of

parametric statistics to analyze nominal and/or ordinal data is wrong and could be considered under some circumstances to be unethical.

In this text, each chapter contains a section on ethics that discusses how businesspeople can misuse the techniques presented in the chapter in an unethical manner. As both users and producers, business students need to be aware of the potential ethical pitfalls that can occur with statistics.

Summary

Statistics is an important decision-making tool in business and is used in virtually every area of business. In this course, the word *statistics* is defined as the science of gathering, analyzing, interpreting, and presenting numerical data.

The study of statistics can be subdivided into two main areas: *descriptive statistics* and *inferential statistics*. Descriptive statistics result from gathering data from a body, group, or population and reaching conclusions only about that group. Inferential statistics are generated by gathering sample data from a group, body, or population and reaching conclusions about the larger group from which the sample was drawn.

Most business statistics studies contain variables, measurements, and data. A *variable* is a characteristic of any entity being studied that is capable of taking on different values. Examples of variables might include monthly household food spending, time between arrivals at a restaurant, and patient satisfaction rating. A *measurement* occurs when a standard process is used to assign numbers to particular attributes or characteristics of a variable. Measurements on monthly household food spending might be taken in dollars, time between arrivals might be measured in minutes, and patient satisfaction might be measured using a 5-point scale. *Data* are recorded measurements. It is data that are analyzed by business statisticians in order to learn more about the variables being studied.

The appropriate type of statistical analysis depends on the level of data measurement, which can be (1) *nominal*, (2) *ordinal*, (3) *interval*, or (4) *ratio*. Nominal is the lowest level, representing classification only of such data as geographic location, sex, or Social Security number. The next level is ordinal, which provides rank-ordering measurements in which the intervals between consecutive numbers do not necessarily represent equal distances. Interval is the next to highest level of data measurement in which the distances represented by consecutive numbers are equal. The highest level of data measurement is ratio, which has all the qualities of interval measurement but ratio data contain an absolute zero and ratios between numbers are meaningful. Interval and ratio data sometimes are called *metric* or *quantitative* data. Nominal and ordinal data sometimes are called *nonmetric* or *qualitative* data.

Two major types of inferential statistics are (1) *parametric statistics* and (2) *nonparametric statistics*. Use of parametric statistics requires interval or ratio data and certain assumptions about the distribution of the data. The techniques presented in this text are largely parametric. If data are only nominal or ordinal in level, nonparametric statistics must be used.

The emergence of exponential growth in the number and type of data existing has resulted in a new term, *big data*, which is a *collection of large and complex datasets from different sources that are difficult to process using traditional data management and process applications*. There are four dimensions of big data: (1) variety, (2) velocity, (3) veracity, and (4) volume. Variety refers to the many different forms of data, velocity refers to the speed with which the data are available and can be processed, veracity has to do with data quality and accuracy, and volume has to do with the ever-increasing size of data.

Business analytics is a relatively new field of business dealing with the challenges, opportunities, and potentialities available to business analysts through big data. Business analytics is *the application of processes and techniques that transform raw data into meaningful information to improve decision making*. There are three categories of business analytics: (1) descriptive analytics, (2) predictive analytics, and (3) prescriptive analytics.

A main component in the process of transforming the mountains of data now available is data mining. Data mining is *the process of collecting, exploring, and analyzing large volumes of data in an effort to uncover hidden patterns and/or relationships that can be used to enhance business decision-making*.

One effective way of communicating data to a broader audience of people is data visualization. Data visualization is any attempt made by data analysts to help individuals better understand data by putting it in a visual context. Specifically, data visualization is *the study of the visual representation of data and is employed to convey data or information by imparting it as visual objects displayed in graphics*.

Key Terms

big data
business analytics
census
data
data mining
data visualization
descriptive analytics
descriptive statistics
inferential statistics
interval-level data

measurement
metric data
nominal-level data
nonmetric data
nonparametric statistics
ordinal-level data
parameter
parametric statistics
population
predictive analytics

prescriptive analytics
ratio-level data
sample
statistic
statistics
variable
variety
velocity
veracity
volume

Supplementary Problems

1.1. Give a specific example of data that might be gathered from each of the following business disciplines: accounting, finance, human resources, marketing, information systems, production, and management. An example in the marketing area might be “number of sales per month by each salesperson.”

1.2 State examples of data that can be gathered for decision-making purposes from each of the following industries: manufacturing, insurance, travel, retailing, communications, computing, agriculture, banking, and healthcare. An example in the travel industry might be the cost of business travel per day in various European cities.

1.3 Give an example of *descriptive* statistics in the recorded music industry. Give an example of how *inferential* statistics could be used in the recorded music industry. Compare the two examples. What makes them different?

1.4 Suppose you are an operations manager for a plant that manufactures batteries. Give an example of how you could use *descriptive* statistics to make better managerial decisions. Give an example of how you could use *inferential* statistics to make better managerial decisions.

1.5 There are many types of information that might help the manager of a large department store run the business more efficiently and better understand how to improve sales. Think about this in such areas as sales, customers, human resources, inventory, suppliers, etc., and list five variables that might produce information that could aid the manager in his or her job. Write a sentence or two describing each variable, and briefly discuss some numerical observations that might be generated for each variable.

1.6 Suppose you are the owner of a medium-sized restaurant in a small city. What are some variables associated with different aspects of the business that might be helpful to you in making business decisions about the restaurant? Name four of these variables, and for each variable, briefly describe a numerical observation that might be the result of measuring the variable.

1.7 Classify each of the following as nominal, ordinal, interval, or ratio data.

- The time required to produce each tire on an assembly line
- The number of quarts of milk a family drinks in a month
- The ranking of four machines in your plant after they have been designated as excellent, good, satisfactory, or poor
- The telephone area code of clients in the United States
- The age of each of your employees
- The dollar sales at the local pizza shop each month
- An employee’s identification number
- The response time of an emergency unit

1.8 Classify each of the following as nominal, ordinal, interval, or ratio data.

- The ranking of a company in the *Fortune* 500
- The number of tickets sold at a movie theater on any given night
- The identification number on a questionnaire
- Per capita income
- The trade balance in dollars
- Profit/loss in dollars
- A company’s tax identification
- The Standard & Poor’s bond ratings of cities based on the following scales:

Rating	Grade
Highest quality	AAA
High quality	AA
Upper medium quality	A
Medium quality	BBB
Somewhat speculative	BB
Low quality, speculative	B
Low grade, default possible	CCC
Low grade, partial recovery possible	CC
Default, recovery unlikely	C

1.9 The Rathburn Manufacturing Company makes electric wiring, which it sells to contractors in the construction industry. Approximately 900 electric contractors purchase wire from Rathburn annually. Rathburn’s director of marketing wants to determine electric contractors’ satisfaction with Rathburn’s wire. He developed a questionnaire that yields a satisfaction score between 10 and 50 for participant responses. A random sample of 35 of the 900 contractors is asked to complete a satisfaction survey. The satisfaction scores for the 35 participants are averaged to produce a mean satisfaction score.

- What is the population for this study?
- What is the sample for this study?
- What is the statistic for this study?
- What would be a parameter for this study?

Exploring the Databases with Business Analytics

See www.wiley.com/college/black

Twenty databases are available with this text, providing additional opportunities to apply business analytics using the statistics presented in this course. These databases are located in WileyPLUS, and each is available in either Minitab or Excel format for your convenience. These

20 databases represent a wide variety of business areas, such as agribusiness, consumer spending, energy, finance, healthcare, international labor, manufacturing, top worldwide companies, leading U.S. market research firms, top U.S. retailers, largest banks in the world, U.K. job

change data, top U.S. companies experiencing growth, EPA emissions data, personal savings rates, numbers of furniture stores, and the stock market. Altogether, these databases contain over 100 variables and over 68,000 observations. The data are gathered from such reliable sources as the U.S. government's Bureau of Labor, the U.S. Department of Agriculture, the American Hospital Association, the Energy Information Administration, the Environmental Protection Agency, the Federal Bank of St. Louis, Moody's Handbook of Common Stocks, Fortune, Inc. Magazine, American Marketing Association, National Retail Federation, Standard & Poor, Emsi, and the U.S. Census Bureau. Five of the 20 databases contain time-series data. These databases are:

12-Year Gasoline Database

The 12-year time-series gasoline database contains monthly data for four variables: U.S. Gasoline Prices, OPEC Spot Price, U.S. Finished Motor Gasoline Production, and U.S. Natural Gas Wellhead Price. There are 137 data entries for each variable. U.S. Gasoline Prices are given in cents, the OPEC Spot Price is given in dollars per barrel, U.S. Finished Motor Gasoline Production is given in 1000 barrels per day, and U.S. Natural Gas Wellhead Price is given in dollars per 1000 cubic feet.

Consumer Food Database

The consumer food database contains five variables: Annual Food Spending per Household, Annual Household Income, Non-Mortgage Household Debt, Geographic Region of the U.S. of the Household, and Household Location. There are 200 entries for each variable in this database representing 200 different households from various regions and locations in the United States. Annual Food Spending per Household, Annual Household Income, and Non-Mortgage Household Debt are all given in dollars. The variable Region tells in which one of four regions the household resides. In this variable, the Northeast is coded as 1, the Midwest is coded 2, the South is coded as 3, and the West is coded as 4. The variable Location is coded as 1 if the household is in a metropolitan area and 2 if the household is outside a metro area. The data in this database were randomly derived and developed based on actual national norms.

Manufacturing Database

This database contains eight variables taken from 20 industries and 140 subindustries in the United States. Some of the industries are food products, textile mill products, furniture, chemicals, rubber products, primary metals, industrial machinery, and transportation equipment. The eight variables are Number of Employees, Number of Production Workers, Value Added by Manufacture, Cost of Materials, Value of Industry Shipments, New Capital Expenditures, End-of-Year Inventories, and Industry Group. Two variables, Number of Employees and Number of Production Workers, are in units of 1000. Four variables, Value Added by Manufacture, Cost of Materials, New Capital Expenditures, and End-of-Year Inventories, are in million-dollar units. The Industry Group variable consists of numbers from 1 to 20 to denote the industry group to which the particular subindustry belongs. Value of Industry Shipments has been recoded to the following 1 to 4 scale.

- 1 = \$0 to \$4.9 billion
- 2 = \$5 billion to \$13.9 billion
- 3 = \$14 billion to \$28.9 billion
- 4 = \$29 billion or more

International Labor Database

This time-series database contains the civilian unemployment rates in percent from seven countries presented yearly over a 40-year period.

The data are published by the Bureau of Labor Statistics of the U.S. Department of Labor. The countries are the United States, Canada, Australia, Japan, France, Germany, and Italy.

Financial Database

The financial database contains observations on eight variables for 100 companies. The variables are Type of Industry, Total Revenues (\$ millions), Total Assets (\$ millions), Return on Equity (%), Earnings per Share (\$), Average Yield (%), Dividends per Share (\$), and Average Price per Earnings (P/E) ratio. The companies represent seven different types of industries. The variable Type displays a company's industry type as:

- 1 = apparel
- 2 = chemical
- 3 = electric power
- 4 = grocery
- 5 = healthcare products
- 6 = insurance
- 7 = petroleum

Energy Database

The time-series energy database consists of data on five energy variables over a period of 26 years. The five variables are U.S. Energy Consumption, World Crude Oil Production, U.S. Nuclear Electricity Generation, U.S. Coal Production, and U.S. Natural Dry Gas Production. U.S. Energy Consumption is given in quadrillion BTUs per year, World Crude Oil Production is given in million barrels per day, U.S. Nuclear Electricity Generation is given in billion kilowatt-hours, U.S. Coal Production is given in million short tons, and U.S. Natural Dry Gas Production is given in million cubic feet.

U.S. and International Stock Market Database

This database contains seven variables—three from the U.S. stock market and four from international stock markets—with data representing monthly averages of each over a period of five years resulting in 60 data points per variable. The U.S. stock market variables include the Dow Jones Industrial Average, the NASDAQ, and Standard & Poor's 500. The four international stock market variables of Nikkei 225, Hang Seng, FTSE 100, and IPC represent Japan, Hong Kong, United Kingdom, and Mexico.

Agribusiness Time-Series Database

The agribusiness time-series database contains the monthly weight (in 1000 lbs.) of cold storage holdings for six different vegetables and for total frozen vegetables over a 14-year period. Each of the seven variables represents 168 months of data. The six vegetables are green beans, broccoli, carrots, sweet corn, onions, and green peas. The data are published by the National Agricultural Statistics Service of the U.S. Department of Agriculture.

AHA Hospital Database

This database contains observations for twelve variables on over 2000 U.S. hospitals. These variables include State, Region, Control, Service, Number of Beds, Number of Admissions, Census, Outpatient Visits, Number of Births, Total Expense, Payroll Expense, and Personnel.

The state variable is coded from 1 to 50. The region variable is coded from 1 to 9, and the numbers represent the following regions:

- 1 = New England: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont

- 2 = Mid Atlantic: New Jersey, New York, Pennsylvania
- 3 = East North Central: Illinois, Indiana, Michigan, Ohio, Wisconsin
- 4 = West North Central: Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota
- 5 = South Atlantic: Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, District of Columbia, West Virginia
- 6 = East South Central: Alabama, Kentucky, Mississippi, Tennessee
- 7 = West South Central: Arkansas, Louisiana, Oklahoma, Texas
- 8 = Mountain: Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming
- 9 = Pacific: Alaska, California, Hawaii, Oregon, Washington

Control is a type of ownership. Six categories of control are included in the database:

- 1 = government nonfederal
- 2 = government federal
- 3 = osteopathic
- 4 = nongovernment nonprofit
- 5 = investor-owned for-profit
- 6 = other nonprofit

Service is the type of hospital. The 15 types of hospitals used in this database are:

- 1 = general medical and surgical
- 2 = rehabilitation
- 3 = orthopedic
- 4 = children's
- 5 = hospital unit of an institution (prison, college, etc.)
- 6 = surgical
- 7 = psychiatric
- 8 = alcoholism and chemical dependency
- 9 = intellectual disabilities
- 10 = long-term acute care
- 11 = heart
- 12 = cancer
- 13 = women's obstetrics & gynecology
- 14 = other specialty
- 15 = chronic disease

This database also contains a subdatabase, Small Hospital database, for hospitals with 40 beds or fewer, using the same variables.

World's Largest Companies

This database contains observations for five variables on 1700 companies worldwide. The variables include Country plus Sales, Profits, Assets, and Market Value in billions of dollars.

AMA Top 50

This database contains observations for the top 50 market research firms in the United States. There are values given on each company for three variables: (1) U.S. Revenue, (2) Non U.S. Revenue, and (3) Number of Employees.

Top 100 Retailers in U.S.

This database contains observations on the top 100 retailers in the United States for the variables of State, Retail Sales, and Number of Stores.

Fifty Largest Banks in the World

This database contains information about the 50 largest banks in the world. The variables include Country, Total Assets, and Market Cap

Inc.5000 2014

This database contains observations about the 5000 U.S. companies experiencing the highest percentage of growth from the previous year. Variables include Number of Workers, State, Percentage Increase of Revenue, Revenue, Industry, and Years on the List.

EMSI United Kingdom Job Change

This database contains information about the number of jobs in particular industries in the United Kingdom and the percentage change over a two-year period. Variables include City, SIC Code, Industry, Number of Jobs in 2011, Number of Jobs in 2014, Raw Change in Number of Jobs, and Percentage Change in Number of Jobs.

EPA Data Average Annual Emissions

This large database is actually seven subdatabases in one. Each of the subdatabases contains emissions data along with data on possible sources. Emissions data is given for carbon monoxide, nitrogen oxide, particle matter 10 micrometers and smaller, particle matter 2.5 micrometers and smaller, sulfur dioxide, volatile organic compounds, and ammonia. All of the subdatabases are time-series databases.

Furniture and Home Furnishings Stores

This time-series database contains monthly data on the number of stores in the United States in the furniture and home furnishings category over a five-year period.

Personal Savings Rate

This database, created by the U.S. Federal Reserve Bank of St. Louis, contains time-series monthly data of U.S. personal savings rates over a 33-year period.

Student Survey Database

This database contains responses from over 900 commuter university students on 7 continuous variables and 3 discrete variables (9,000 observations) gathered by the author. The continuous variables are age, miles to class, minutes to class, credit hours taken this semester, work hours per week, number of meals eaten out per week, and number of non-class-related books read per year. The discrete variables are the student's sex, favorite food (selected from Mexican, Chinese, Italian, Indian), and if he/she is traveling out-of-state during the semester.

Multi-Family Metrics Database

This multi-family database based on Marcus & Millichap's U.S. Multifamily Investment Forecast, ranks 46 apartment markets based on 12-month forward-looking economic indicators along with supply and demand variables. The database contains data on 5 variables for 46 U.S. cities along with their ranks over a two-year period.

Database Mining

1. When analyzing a database, Tableau—visualizing software—typically divides variables into two categories based on the level of data measurement: dimensions and measures. Tableau considers any column or row of data containing qualitative data to be a *dimension* and any column or row of data containing quantitative information

to be a *measure*. The type of graph that can be constructed by Tableau depends on how many dimensions and/or measures a person is attempting to visualize. As shown in Figure 1.5, interval and ratio data are quantitative data and hence would be treated by Tableau as a measure. Nominal- and ordinal-level data are qualitative data and are treated by Tableau as a dimension.

Examine the 20 databases associated with this text. Using what you have learned about levels of data measurement and about qualitative

and quantitative data, find a database that has both a dimension and a measure, thereby making it a good candidate for several Tableau graphs.

2. Search through the 20 databases associated with this text to find one variable for each of the four levels of data. You may have to draw variables from different databases. List the four variables, tell which databases they come from, and write at least one sentence each explaining why the data in that variable are a particular level of data.

Chapter Case

DiGiorno Pizza: Introducing a Frozen Pizza to Compete with Carry-Out

Kraft Foods successfully introduced DiGiorno Pizza into the marketplace in 1996, with first-year sales of \$120 million, followed by \$200 million in sales in 1997. It was neither luck nor coincidence that DiGiorno Pizza was an instant success. Kraft conducted extensive research about the product and the marketplace before introducing this product to the public. Many questions had to be answered before Kraft began production. For example, why do people eat pizza? When do they eat pizza? Do consumers believe that carry-out pizza is always more tasty?

SMI-Alcott conducted a research study for Kraft in which they sent out 1000 surveys to pizza lovers. The results indicated that people ate pizza during fun social occasions or at home when no one wanted to cook. People used frozen pizza mostly for convenience but selected carry-out pizza for a variety of other reasons, including quality and the avoidance of cooking. The Loran Marketing Group conducted focus groups for Kraft with women aged 25 to 54. Their findings showed that consumers used frozen pizza for convenience but wanted carry-out pizza taste. Kraft researchers realized that if they were to launch a successful frozen pizza that could compete with carry-out pizza, they had to develop a frozen pizza that (a) had restaurant takeout quality, (b) possessed flavor variety, (c) was fast and easy to prepare, and (d) had the convenience of freezer storage. To satisfy these seemingly divergent goals, Kraft developed DiGiorno Pizza, which rises in the oven as it cooks. This impressed focus group members; and in a series of blind taste tests conducted by Product Dynamics, DiGiorno Pizza beat out all frozen pizzas and finished second overall behind one carry-out brand.

DiGiorno Pizza has continued to grow in sales and market share over the years. By 2005, sales had topped the \$600 million mark, and DiGiorno Pizza held nearly a quarter of the market share of frozen pizza sales. In each of the last two quarters of 2009, DiGiorno sales increased 20%. On January 6, 2010, Kraft agreed to sell its North American frozen pizza business, including its DiGiorno products, to Nestlé for \$3.7 billion. According to data reported by Statista, DiGiorno was by far the top frozen pizza brand in the United States in 2017 with over \$1 billion in sales when compared to the next brand which had sales of \$572 million.

Discussion

Think about the market research that was conducted by Kraft and the fact that it used several companies.

1. What are some of the populations that Kraft might have been interested in measuring for these studies? Did Kraft actually attempt to contact entire populations? What samples were taken? In light of these two questions, how was the inferential process used by Kraft in their market research? Can you think of any descriptive statistics that might have been used by Kraft in their decision-making process?
2. In the various market research efforts made by Kraft for DiGiorno, some of the possible measurements appear in the following list. Categorize these by level of data. Think of some other measurements that Kraft researchers might have made to help them in this research effort, and categorize them by level of data.

- a. Number of pizzas consumed per week per household
- b. Age of pizza purchaser
- c. Zip code of the survey respondent
- d. Dollars spent per month on pizza per person
- e. Time in between purchases of pizza
- f. Rating of taste of a given pizza brand on a scale from 1 to 10, where 1 is very poor taste and 10 is excellent taste
- g. Ranking of the taste of four pizza brands on a taste test
- h. Number representing the geographic location of the survey respondent
- i. Quality rating of a pizza brand as excellent, good, average, below average, or poor
- j. Number representing the pizza brand being evaluated
- k. Sex of survey respondent

Source: Adapted from “Upper Crust,” *American Demographics*, March 1999, p. 58; “Kraft Trading Pizza for Chocolate,” *MarketWatch*, October 25, 2010, at: <http://www.marketwatch.com/story/kraft-trading-pizza-for-chocolate-2010-01-05>; “Sales of the Leading Frozen Pizza brands of the United States from 2011 to 2017,” at: <https://www.statista.com/statistics/189650/top-frozen-pizza-brands-in-the-united-states/>

Big Data Case

Every chapter in this text will end with a big data case. Unlike the chapter cases, which feature a wide variety of companies and busi-

ness scenarios, the big data case will be based mainly on data from a single industry, U.S. hospitals. Drawing from the American Hospital

Association database of over 2000 hospitals, we will use business analytics to mine data on these hospitals in different ways for each chapter based on analytics presented in that chapter.

The hospital database features data on 12 variables, thereby offering over 24,000 observations to analyze. Hospital data for each of

the 50 states are contained in the database with qualitative data representing state, region of the country, type of ownership, and type of hospital. Quantitative measures include Number of Beds, Number of Admissions, Census, Number of Outpatients, Number of Births, Total Expenditures, Payroll Expenditures, and Personnel.

References

1. “How Walmart Makes Data Work for its Customers,” SAS.com (2016), at: https://www.sas.com/en_us/insights/articles/analytics/how-walmart-makes-data-work-for-its-customers.html
2. Huang, Shih-Chia, Suzanne McIntosh, Stanislav Sobolevsky, & Patrick C. K. Hung (2017). “Big Data Analytics and Business Intelligence in Industry,” *Information Systems Frontiers*, 19:1229–1232.
3. Ibid.
4. Ittmann, Hans W. (2015). “The Impact of Big Data and Business Analytics on Supply Chain Management,” *Journal of Transport and Supply Chain Management*, doi: 10.4102/jtscm.v9i1.165
5. Huang.
6. Ittmann.
7. Chiang, Roger H. L., V. Grover, T-P. Lian, & Z. Dongsong (2018). “Special Issue: Strategic Value of Big Data and Business Analytics,” *Journal of Management Information Systems*, 35(2):383–387.
8. Wilder, C. R., & C. O. Ozgur (January 2015). “Business Analytics Curriculum for Undergraduate Majors,” *Infirms*, 15(2):180–187.
9. Delen, D., & H. M. Zolbanin (2018). “The Analytics Paradigm in Business Research,” *Journal of Business Research*, 90:186–195.
10. Mortenson, M. J., N. F. Doherty, & S. Robinson (2015). “Operational Research from Taylorism to Terabytes: A Research Agenda for the Analytic Sage,” *European Journal of Operational Research*, 241(3):583–595.
11. Sharda, R., D. Delen, & E. Turban (2017). *Business Intelligence Analytics and Data Science: A Managerial Perspective*. Upper Saddle River, NJ: Pearson.
12. Aasheim, C., S. Williams, P. Rutner, & A. Gardner (2015). “Data Analytics vs. Data Science: A Study of Similarities and Differences in Undergraduate Programs Based on Course Descriptions,” *Journal of Information Systems Education*, 26(2):103–115.
13. Finch, G., C. Reese, R. Shockley, & R. Balbon (2013). “Analytics: A Blueprint for Value. Converting Big Data and Analytics Insights into Results,” IBM Institute for Business Value. Retrieved from <http://www-935.ibm.com/services/us/gbs/thoughtleadership/ninelevers/>
14. LeClair, D. (2018). “Integrating Business Analytics in the Marketing Curriculum: Eight Recommendations,” *Marketing Education Review*, 28(1):6–13.
15. Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, & A. Byers (2011). “Big Data: The Next Frontier for Innovation, Competition, and Productivity,” A report by the McKinsey Global Institute. Available at: https://www.mckinsey.com/~media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_exec_summary.ashx
16. Bertolucci, J. (December 31, 2013). “Big Data Analytics: Descriptive vs. Predictive vs. Prescriptive,” *Information Week*, available at: <https://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-descriptive-vs-predictive-vs-prescriptive/d/d-id/1113279>
17. Praseeda, C. K., & B. L. Shivakumar (November–December 2014). “A Review of Trends and Technologies in Business Analytics,” *International Journal of Advanced Research in Computer Science*, 5(8):225–229.
18. Watson IoT (2017). “Descriptive, Predictive, Prescriptive: Transforming Asset and Facilities Management with Analytics,” IBM paper, at: <http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype5SA&subtype5WH&htmlfid5TIW14162USEN>
19. Wilder, C., & C. Ozgur (2015). “Business Analytics Curriculum for Undergraduate Majors,” *INFORMS Transactions on Education*, 15(2):180–187.
20. Praseeda & Shivakumar.
21. Praseeda & Shivakumar.
22. Delen & Zolbanin.
23. Watson IoT.
24. Sharda, Delen, & Turban.
25. Praseeda & Shivakumar.
26. Delen & Zolbanin.
27. Sharda, Delen, & Turban.
28. Praseeda & Shivakumar.
29. Watson IoT.
30. Gorman.
31. Rouse, M. (2012). “Prescriptive Analytics.” Retrieved from: <https://searchcio.techtarget.com/definition/Prescriptive-analytics>
32. Praseeda & Shivakumar.
33. Lusting, I., B. Dietrich, C. Johnson, & C. Dziekan (2010). “The Analytics Journey,” *Analytics Magazine*, 3(6):11–13.
34. Sharda, Delen, & Tuurban.
35. Zhao, S. (October 20, 2017). “What Is ETL? (Extract, Transform, Load),” *Experian Data Quality*, at: <https://www.edq.com/blog/what-is-etl-extract-transform-load/>
36. Wu, S. (2013). “A Review on Coarse Warranty Data and Analysis,” *Reliability Engineering and System Safety*, 114:1–11.
37. Zhao.
38. Connolly, T. M., & C. E. Begg (2014). *Database Systems: A Practical Approach to Design Implementation and Management* (6th ed.). Upper Saddle River, NJ: Pearson, p. 64.
39. Evans, J.R. (2016). *Business Analytics: Methods, Models, and Decisions* (2nd ed.). Boston: Pearson Education, Inc., p. 7.
40. Roberts, R., R. Laramee, P. Brookes, G. A. Smith, T. D. Cruze, & M. J. Roach (2018). “A Tale of Two Visions: Exploring the Dichotomy of Interest Between Academia and Industry in Visualisation.” In: *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications—Volume 3: IVAPP, Funchal, Portugal, 27–29 January 2018*. SciTePress: Setúbal, Portugal, pp. 319–326.
41. Stephen, F. (September 18, 2004) “Eenie, Meenie, Minie, Moe: Selecting the Right Graph for Your Message” [PDF], *Perceptual Edge*, at: https://www.perceptualedge.com/articles/ie/the_right_graph.pdf