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:

1.1 Define important statistical terms, including population, sample, and parameter, as they relate to descriptive and inferential statistics.

1.2 Explain the difference between variables, measurement, and data, and compare the four different levels of data: nominal, ordinal, interval, and ratio.

1.3 Explain the differences between the four dimensions of big data.

1.4 Compare and contrast the three categories of business analytics.

1.5 Describe the data mining and data visualization processes.

Decision Dilemma

Statistics Describe the State of Business in India's Countryside

Kailash Kumar/Getty Images

India is the second-most populous country in the world, with more than 1.4 billion people. Nearly 70% of the people live in rural areas, scattered about the countryside in 600,000 villages. In fact, it may be said that more than one in every ten people in the world live in rural India. While it has a per capita income of US$1.50 per day, rural India, which has been described in the past as poor and semi-literate, now contributes about one-half of the country's gross national product (GNP). However, rural India still has the most households in the world without electricity.

Despite its poverty and economic disadvantages, there are compelling reasons for companies to market their goods and services to rural India. This market has been steadily growing. 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 of their lifestyles and of opportunities for a better life.

Agriculture is no longer the main source of income in rural India, with only 23% coming from farming. Nearly three-quarters of rural India has a mobile connection. Rural consumers are saving, on average, 25% of their income, thereby presenting a significant opportunity for an expansion in the use of and demand for banking and investment products. In addition, there has been an increased consumer priority on health and hygiene products. The rural Fast-Moving Consumer Goods market in India is expected to grow from US$23.63 billion in fiscal year 2018 to US$220 billion by 2025.

Because of such factors, many global and Indian firms, such as Microsoft, General Electric, Nestlé, Kellogg's, Colgate-Palmolive, Hindustan-Unilever, Godrej, Nirma, Novartis, Dabur, Tata, Hero, Bajaj, and Vodafone India, among many others, have entered the rural Indian market with enthusiasm. Marketing to rural customers often involves persuading them to try 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 the 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 statistics to make better decisions about entering this marketplace?**
6. **What big data might be available on rural India?**

Sources: Adapted from "India—Population 2021," statisticstimes.com, December 9, 2021; "India's Rural-Urban Divide: Village Worker Earns Less Than Half of City Peer," financialexpress.com, December 12, 2019; "Living in the Dark: 240 Million Indians Have No Electricity," bloomberg.com, January 24, 2017; "Only 23% of Rural Income from Farming, Reveals NABARD 2016-2017 Survey," indianexpress.com, November 1, 2021; "Agriculture No Longer Main Rural Income Source," business-standard.com, November 16, 2017; "Tring Tring! Nearly Three-Fourths of Rural India Has a Mobile Connection Now," thebetterindia.com, January 24, 2019; "Online Now the Second Largest Consumed Media after TV in Rural India: Kantar-Dialogue Factory," mediabrief.com, September 8, 2021; "Fast Moving Consumer Goods (FMCG)," readkong.com, January 2021.

Introduction

Every minute of the working day, decisions are made by businesses around the world that determine whether companies will be profitable and grow or 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 by 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.

There is a wide variety of uses and applications of statistics in business. Several examples follow.

- A survey of 3,893 Canadians aged 18 or older conducted by Bank of Canada indicated that 80% of Canadians have no plans to become cashless in the next five years, whereas 12% of Canadians are already cashless.[1]
- According to an Ipsos survey of 1,001 Canadians, 44% intend to ask for a raise later this year in response to high rates of inflation and rising interest rates.[2]
- According to Statistics Canada, 84.4% of Canadians over 15 years of age have a smartphone for personal use. For 53.2% of Canadians, checking their smartphone is the first thing they do upon waking up in the morning. This habit is more apparent in younger age brackets, with 74.0% and 77.3% of males and females, respectively, between 15 and 24 waking up to their smartphones, compared to 24.8% and 20.3% of males and females, respectively aged 65 and over.[3]
- A Deloitte Retail "Green" survey of 1,080 adults revealed that 54% agreed that plastic, non-compostable shopping bags should be banned.
- A J.D. Power survey of 6,699 Canadians found that satisfaction with credit card customer service had fallen from 811 to 794 (out of 1,000) over a one-year period, with 22% of respondents indicating that they had switched their primary card to avoid paying annual fees.[4]

[1]Heng Chen, Walter Engert, Marie-Hélène Felt, Kim Huynh, Gradon Nicholls, Daneal O'Habib, and Julia Zhu, "Cash and COVID-19: The Impact of the Second Wave in Canada," Bank of Canada, July 2021, https://www.bankofcanada.ca/2021/07/staff-discussion-paper-2021-12/.

[2]Darrell Bricker, "85% of Canadians Are Concerned That Inflation Will Make Things Less Affordable," Ipsos, June 22, 2022, https://www.ipsos.com/en-ca/news-polls/85-percent-of-canadians-concerned-inflation-less-affordable.

[3]Statistics Canada, "Smartphone Personal Use and Selected Smartphone Habits by Gender and Age Group," Table 22-10-0143-01, June 22, 2021, https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2210014301.

[4]J.D. Power, "Credit Card Customers in Canada Spending Less but Have Higher Satisfaction, J.D. Power Finds," September 16, 2021, https://www.jdpower.com/business/press-releases/2021-canada-credit-card-satisfaction-study.

- In Deloitte's Food Consumer Survey of approximately 1,000 respondents, 23% were satisfied with their online grocery shopping experience. The 18- to 34-year-old age group was more satisfied (28%) than the national average, but still purchased 84% of their monthly groceries in bricks-and-mortar stores.[5]

Note that, in most of these examples, business analysts have conducted a study and provided rich and interesting information that can be used in business decision-making.

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 in a less expensive and more timely manner. We will explore various ways to forecast future values and we will 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's begin!

1.1 Basic Statistical Concepts

LEARNING OBJECTIVE 1.1

Define important statistical terms, including population, sample, and parameter, as they relate to descriptive and inferential statistics.

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*. This word 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** captures the key elements of business statistics.

statistics A science dealing with the collection, analysis, interpretation, and presentation of numerical data.

population A collection of persons, objects, or items of interest.

census A process of gathering data from the whole population for a given measurement of interest.

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 Escape crossover vehicles produced from Year 1 to Year 2." A population can be a group of people, such as "all workers employed by Microsoft," or it can be a set of objects, such as "all Toyota RAV4s produced in February of Year 1 by Toyota Canada at the Woodstock, Ontario, plant." The analyst defines the population to be whatever he or she is studying. When analysts *gather data from the whole population for a given measurement of interest*, they call it a **census**. Most people are familiar with the Canadian Census. Every five years, the government attempts to measure all persons living in this country. As another example, if an analyst is interested in ascertaining the grade point average for

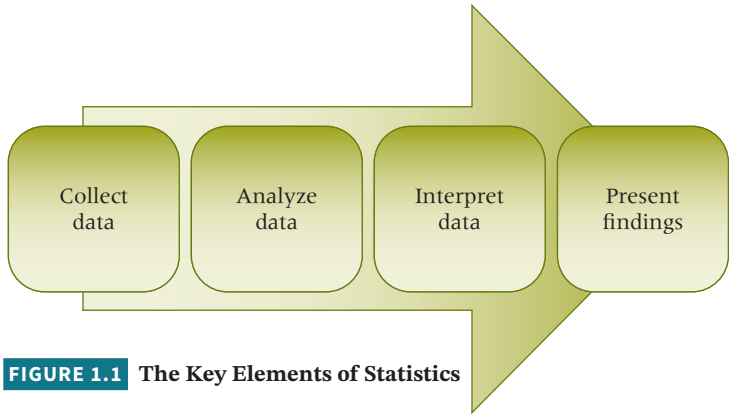

FIGURE 1.1 **The Key Elements of Statistics**

[5]Deloitte, "The Conflicted Consumer—2021 Food Consumer Survey," 2021, https://www2.deloitte.com/ca/en/pages/consumer-business/articles/future-of-food-a-canadian-perspective.html.

all students at the University of Toronto, one way to do so is to conduct a census of all students currently enrolled there.

sample A portion of the whole.

A **sample** is *a portion of the whole* and, if properly taken, is representative of the whole. For various reasons (explained in Chapter 7), analysts 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 smartphone batteries, a smartphone battery manufacturer might randomly sample only 75 batteries 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.

descriptive statistics Statistics that have been gathered on a group to describe or reach conclusions about that same group.

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 results and uses those statistics to reach conclusions about that class only, the statistics are descriptive. The instructor can use these statistics to discuss class average, talk about the range of class scores, or present any other data measurements for the class based on the test.

Most athletic statistics, such as batting averages, save percentages, and first downs, are descriptive statistics because they are used to describe an individual or team effort. Many of the statistical data generated by businesses are descriptive. They might include number of employees on vacation during June, average salary at the Edmonton office, corporate sales for the current fiscal year, average managerial satisfaction score on a company-wide census of employee attitudes, and average return on investment for lululemon for the first ten years of operations.

inferential statistics Statistics that have been gathered from a sample and used to reach conclusions about the population from which the sample was taken.

Another type of statistics is called **inferential statistics**. If an analyst *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, analysts can design experiments with small, randomly selected samples of patients and attempt to reach conclusions and make inferences about the population.

Market analysts 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 analysts want to measure the impact of the new advertisement on various age groups. The analyst 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 analyst to effectively study a wide range of phenomena without having to conduct a census. Most of the topics discussed in this text pertain to inferential statistics.

parameter A descriptive measure of the population.

statistic A descriptive measure of a sample.

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 analyst 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).

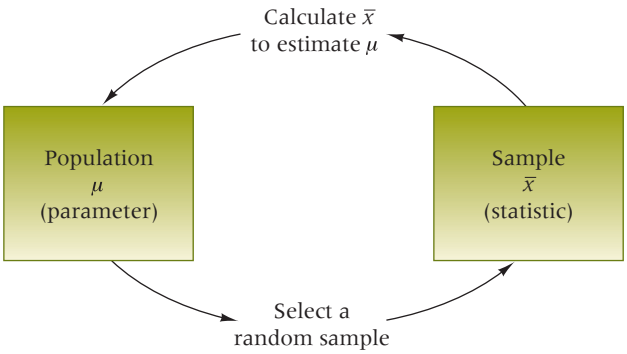

FIGURE 1.2 **Process of Applying Inferential Statistics to Estimate a Population Mean (μ)**

Concept Check

Fill in the blanks.

1. Descriptive statistics can be used to _____ the data to describe a data sample either numerically or graphically.
2. Statistical inference is inference about a _____ from a random data _____ drawn from it.

1.2 Variables, Data, and Data Measurement

LEARNING OBJECTIVE 1.2

Explain the difference between variables, measurement, and data, and compare the four different levels of data: nominal, ordinal, interval, and ratio.

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 might include return on investment, advertising dollars, labour productivity, stock price, historic cost, total sales, market share, age of worker, earnings per share, kilometres 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** occurs *when a standard process is used to assign numbers to particular attributes or characteristics of a variable*. Many measurements are obvious, such as the time a customer spends shopping in a store, the age of the worker, or the number of kilometres driven to work. However, some measurements, such as labour 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 them 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

variable A characteristic of any entity being studied that is capable of taking on different values.

measurement What occurs when a standard process is used to assign numbers to particular attributes or characteristics of a variable.

data Recorded measurements.

Thinking Critically About Statistics in Business Today 1.1

Cellular Phone Use in Japan

The Communications and Information Network Association of Japan conducts an annual study of cellular phone use in Japan. A recent survey was taken as part of this study using a sample of 1,200 cellphone users split evenly between men and women and almost equally distributed over six age brackets. The survey was administered in the greater Tokyo and Osaka metropolitan areas. The study produced several interesting findings.

Of the respondents, 98.3% said that their main-use terminal was a smartphone, while 1.6% 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. Slightly more than 32% of respondents used multiple devices. Of those that did, almost 83% used tablets. Smartphone users in the survey reported that the two decisive factors in purchasing a new smartphone were purchase price (81.0%) and monthly payment cost (87.6%). Over half of all respondents used their device for mobile cashless payment services.

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: Adapted from "CIAJ Releases Report on the Study of Mobile Phone Use," Communications and Information Network Association of Japan (CIAJ), December 9, 2021, www.ciaj.or.jp/en/news/news2020/1071.html.

important that the business analyst pay thoughtful attention to the creation of meaningful, valid data before embarking on analysis and reaching conclusions (see **Thinking Critically About Statistics in Business Today 1.1**).

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, masses of shipments, and rankings of employees at yearly reviews. Not all such data should be analyzed in 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 masses of two objects being shipped, the ratings received on a consumer test by two different products, or the hockey jersey numbers of a centre and a winger. Although 80 kg is twice as much as 40 kg, the winger is probably not twice as big as the centre! Averaging the two masses seems reasonable but averaging the hockey 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 are:

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

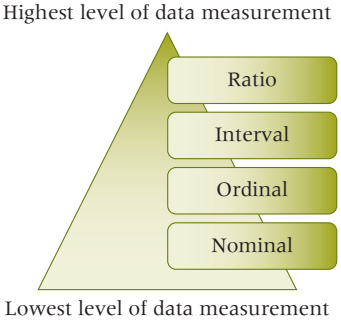

FIGURE 1.3 Hierarchy of Levels of Data

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

Nominal Level

nominal-level data The lowest level of data measurement; used only to classify or categorize.

The *lowest level of data measurement* is the nominal level. Numbers representing **nominal-level data** (the word *level* is often 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 a question that would result in nominal data:

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

a. Educator
b. Construction worker
c. Manufacturing worker
d. Lawyer
e. Doctor
f. Other

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) or any other occupation.

Other types of variables that often produce nominal-level data are marital status, religion, language, geographic location, and employment status. Social insurance numbers, telephone numbers, and employee ID numbers 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 having the nominal-level capabilities, ordinal-level measurement can be used to rank or order 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, the supervisor could not say that the differences in the amount of productivity between the workers ranked 1, 2, and 3 are necessarily the same. With ordinal data, the distances between consecutive numbers are not always equal.

ordinal-level data Next-higher level of data from nominal-level data; can be used to order or rank items, objects, or people.

Some Likert-type scales on questionnaires are considered by many analysts to be ordinal in level. The following is an example of such a scale:

This computer tutorial is

_____	_____	_____	_____	_____
not helpful	somewhat helpful	moderately helpful	very helpful	extremely 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 descriptions. Virtually everyone would agree that a 5 is higher than a 4 on this scale and that it is possible to rank responses. 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 are sometimes rated in terms of investment risk by using measures of default risk, currency risk, and interest rate risk. These three measures are applied to investments by rating them as 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 50 best employers in Canada in *Report on Business* 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 such data.

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*.

nonmetric data Nominal- and ordinal-level data. Also called *qualitative data.*

Interval Level

interval-level data Next to highest level of data. These data have all the properties of ordinal-level data, but in addition, intervals between consecutive numbers have meaning.

ratio-level data Highest level of data measurement; contains the same properties as interval-level data, with the additional property that zero has meaning and represents the absence of the phenomenon being measured.

metric data Interval- and ratio-level data. Also called *quantitative data*.

parametric statistics A class of statistical techniques that contains assumptions about the population and that is used only with interval- and ratio-level data.

nonparametric statistics A class of statistical techniques that makes few assumptions about the population and is particularly applicable to nominal- and ordinal-level data.

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 Celsius temperature. With Celsius temperature numbers, the temperatures can be ranked, and the increases in 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 Celsius 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 share 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, mass, time, volume, and Kelvin temperature. With ratio data, an analyst can state that 180 kg of mass is twice as much as 90 kg or, in other words, make a ratio of 180:90. Many of the data gathered by machines 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 distance, number of trucks sold, complaints per 10,000 flyers, and number of employees. With ratio-level data, no b factor is required in converting units from one measurement to another, that is, $y = ax$. As an example, in converting height from metres to feet, 1 m = 3.28 ft.

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

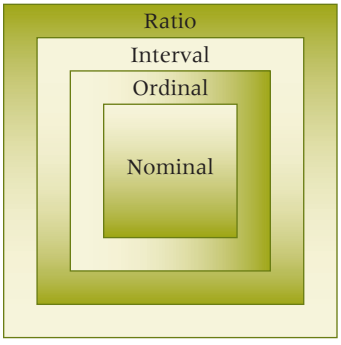

FIGURE 1.4 Usage Potential of Various Levels of Data

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 analyst to perform any analysis that can be done with nominal data and some additional ones. With ratio data, an analyst can make ratio comparisons and appropriately do any analysis that can be performed on nominal, ordinal, or interval data. Some statistical techniques require ratio data and cannot be used to analyze other levels of data.

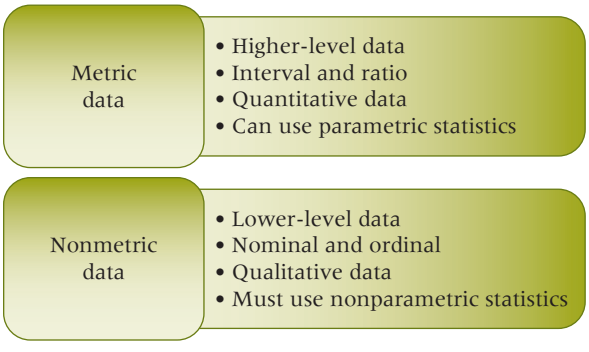

FIGURE 1.5 Metric vs. Nonmetric Data

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 present nonparametric techniques. Thus, much of the material in this text requires interval or ratio data. **Figure 1.5** contains a summary of metric data and nonmetric data.

Concept Check

Match the level with the correct measurement.

1. Nominal	**a.** Measures with a fixed zero that means "no quantity"; a constant interval (distance on the scale)
2. Ordinal	**b.** Measures require a fixed distance, but the zero point is arbitrary
3. Interval	**c.** Measures classified by shared attributes (characteristics)
4. Ratio	**d.** Measures by orderings (ranks)

Demonstration Problem 1.1

The health-care system is constantly changing. Because of the growing pressure to deliver a high level of service at increasingly low costs and the need to determine how providers can better serve their patients, hospital administrators sometimes administer a quality satisfaction survey to their patients after their release from hospital. The following types of questions are sometimes asked on such a survey. What level of data measurement will these questions result in?

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 unit
___Intensive care unit
___Maternity care unit
___Medical unit
___Pediatric/children's unit
___Surgical unit

3. How serious was your condition when you were first admitted to the hospital?

___Critical ___Serious ___Moderate ___Minor

4. Rate the skill of your doctor:

___Excellent ___Very Good ___Good ___Fair ___Poor

5. Rate the nursing care on the following scale from 1 to 7:

Poor 1 2 3 4 5 6 7 Excellent

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 4 are likely to result in ordinal-level data. Suppose a number is assigned to each descriptor in each of these two questions. For question 3, "critical" might be assigned a 4, "serious" a 3, "moderate" a 2, and "minor" a 1. Certainly, the higher the number, the more serious the patient's condition is. 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 numerical values assigned in question 4.

Question 5 displays seven numerical choices with equal distances between the numbers shown on the scale and no adjective descriptors assigned to the numbers. Many analysts would declare this to be interval-level measurement because of the equal distance between numbers and the absence of a true zero on this scale. Other analysts might argue that, because of the imprecision of the scale and the vagueness of selecting values between "poor" and "excellent," the measurement is only ordinal in level.

1.3 Big Data

LEARNING OBJECTIVE 1.3

Explain the differences between the four dimensions of big data.

In the current world of business, the data available to decision-makers to help them produce better business outcomes are growing exponentially. Growing sources of data are available from the internet, social media, governments, transportation systems, health care, environmental organizations, and a plethora of business data sources, among others. Business data include, but are not limited to, consumer information, labour statistics, financials, product and resource tracking information, supply chain information, operations information, and human resource information. According to vCloud, 2.5 quintillion bytes of data are created every day. 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, which are then added to its petabytes of historical data.[6]

big data A large amount of either organized or unorganized data from different sources that is difficult to process using traditional data management and processing applications, and is analyzed to make an informed decision or evaluation.

The advent of such growth in the amount and types of data available to analysts, data scientists, and business decision-makers has resulted in a new term, "Big Data." **Big data** has been defined as *a collection of large and complex data sets from different sources that are difficult to process using traditional data management and processing applications*.[7] In addition, big data can be seen as *a large amount of either organized or unorganized data that are analyzed to make an informed decision or evaluation*.[8] All data are not created in the same way, nor do they represent the same things. Thus, analysts recognize that there are at least four characteristics or dimensions associated with big data.[9] These are:

1. Variety
2. Velocity
3. Veracity
4. Volume

variety Refers to the many different forms of data based on data sources.

velocity Refers to the speed at which the data are available and at which they can be processed.

veracity Has to do with the quality, correctness, and accuracy of data.

volume Has to do with the ever-increasing size of data and databases.

Variety refers to *the many different forms of data based on data sources*. A wide variety of data are 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 at which the data is available and can be processed*. The velocity characteristic of data is important in ensuring that data is current and updated in real time.[10] **Veracity** of data *has to do with data quality, correctness, and accuracy*.[11] 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 v*alue*. Analysis of data that does not generate value makes no contribution to an organization.[12]

[6]"How Walmart Makes Data Work for Its Customers," SAS.com, 2016, at www.sas.com/en_us/insights/articles/analytics/how-walmart-makes-data-work-for-its-customers.html.

[7]Shih-Chia Huang, Suzanne McIntosh, Stanislav Sobolevsky, and Patrick C. K. Hung, "Big Data Analytics and Business Intelligence in Industry," *Information Systems Frontiers* 19 (2017): 1229–32.

[8]Ibid.

[9]Hans W. Ittmann, "The Impact of Big Data and Business Analytics on Supply Chain Management," *Journal of Transport and Supply Chain Management* 9, no. 1 (2015): a165.

[10]Huang et al., "Big Data Analytics and Business Intelligence in Industry."

[11]Ittmann, "The Impact of Big Data and Business Analytics on Supply Chain Management."

[12]Roger H. L. Chiang, Varun Grover, Ting-Peng Lian, and Zhang Dongsong, "Special Issue: Strategic Value of Big Data and Business Analytics," *Journal of Management Information Systems* 35, no. 2: 383–87.

Along with this unparalleled explosion of data, major advances have been made in computing performance, network functioning, data handling, device mobility, and other areas. When businesses capitalize on these developments, new opportunities become available for discovering greater and deeper insights that could produce significant improvements in the way businesses operate and function. So how do businesses go about capitalizing on this potential?

1.4 Business Analytics

LEARNING OBJECTIVE 1.4

Compare and contrast the three categories of business analytics.

To benefit from the challenges, opportunities, and potentialities presented to business decision-makers by big data, the new field of "business analytics" has emerged. There are different definitions of business analytics in the literature. However, one that most 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*.[13] Because big data sources are so large and complex, new questions have arisen that cannot always be effectively answered with traditional methods of analysis.[14] 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."[15] Other names sometimes used to refer to business analytics are 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.[16] For example, when applied to the business environment, data analytics becomes synonymous with business analytics.[17]

business analytics The application of processes and techniques that transform raw data into meaningful information to improve decision-making.

Business analytics provides added value to data, resulting in deeper and broader business insights that can be used to improve decision-makers' understanding 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. Today's executives 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.[18]

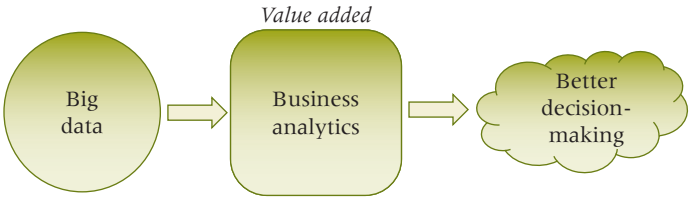

FIGURE 1.6 Business Analytics Add Value to Data

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. There is a plethora of such techniques available, drawing from such areas as statistics, operations research, mathematical modelling, data

[13]Coleen R. Wilder and Ceyhun O. Ozgur, "Business Analytics Curriculum for Undergraduate Majors," *Informs* 15, no. 2 (January 2015): 180–87, doi.org/10.1287/ited.2014.0134.

[14]Dursun Delen and Hamed M. Zolbanin, "The Analytics Paradigm in Business Research," *Journal of Business Research* 90 (2018): 186–95.

[15]M. J. Mortenson, N. F. Doherty, and S. Robinson, "Operational Research from Taylorism to Terabytes: A Research Agenda for the Analytic Sage," *European Journal of Operational Research* 241, no. 3 (2015): 583–95.

[16]R. Sharda, D. Delen, and E. Turban, *Business Intelligence Analytics and Data Science: A Managerial Perspective* (Upper Saddle River, NJ: Pearson, 2017).

[17]C. Aasheim, S. Williams, P. Rutner, and A. Gardner, "Data Analytics vs. Data Science: A Study of Similarities and Differences in Undergraduate Programs Based on Course Descriptions," *Journal of Information Systems Education* 26, no. 2 (2015): 103–15.

[18]Dan LeClair, "Integrating Business Analytics in the Marketing Curriculum: Eight Recommendations," *Marketing Education Review* 28, no. 1 (2018): 6–13.

mining, and artificial intelligence, 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 Often the first step in the analytics process, it takes traditional data and describes what has happened or is happening in a business. It can be used to condense big data into smaller, more useful, data. Sometimes referred to as *reporting 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 analytics process, **descriptive analytics** *takes traditional data and describes what has happened or is happening in a business*. It can be used to condense big data into smaller, more useful data.[19] Also referred to as reporting analytics, descriptive analytics can be used to discover hidden relationships in the data and identify undiscovered patterns.[20] Descriptive analytics drills down in the data to uncover useful and important details, and mines data for trends.[21] At this step, visualization can be a key technique for 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.[22] We could probably add correlation and various clustering techniques to the mix, along with data mining and data visualization techniques.

predictive analytics The second step in the analytics process, which finds relationships in data that are not readily apparent with descriptive analytics.

Predictive Analytics The next step in data reduction is **predictive analytics**, *which finds relationships in the data that are not readily apparent with descriptive analytics*.[23] With predictive analytics, patterns or relationships are extrapolated forward in time, and the past is used to make predictions about the future.[24] Predictive analytics also provides answers that move beyond using the historical data as the principal basis for decisions.[25] It builds and assesses algorithmic models that are intended to make empirical rather than theoretical predictions, and are designed to predict future observations.[26] Predictive analytics can help managers develop likely scenarios.[27]

Topics in predictive analytics can include regression, time-series, forecasting, simulation, data mining (see Section 1.5), statistical modelling, machine-learning techniques, and others. They can also include classifying techniques, such as decision-tree models and neural networks.[28]

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 made.

Prescriptive Analytics The final stage of business analytics is **prescriptive analytics**, which is still in its early stages of development.[29] Prescriptive analytics follows descriptive and predictive analytics in an attempt to find the best course of action under certain circumstances.[30] The goal is to examine current trends and likely forecasts and use that information to make better decisions.[31] 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*.[32] It does this by exploring a set of possible actions

[19]Jeff Bertolucci, "Big Data Analytics: Descriptive vs. Predictive vs. Prescriptive," *Information Week*, December 31, 2018, www.informationweek.com/big-data/big-data.

[20]C. K. Praseeda and B. L. Shivakumar, "A Review of Trends and Technologies in Business Analytics," *International Journal of Advanced Research in Computer Science* 5, no. 8 (November–December 2014): 225–29.

[21]Watson ioT, "Descriptive, Predictive, Prescriptive: Transforming Asset and Facilities Management with Analytics," IBM paper, 2017, www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=SA&subtype=WH&htmlfid=TIW14162USEN.

[22]Wilder and Ozgur, "Business Analytics Curriculum for Undergraduate Majors."

[23]B. Daniel, "Big Data and Analytics in Higher Education: Opportunities and Challenges," *British Journal of Educational Technology* 46, no. 5 (2015): 904–20, onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12230.

[24]Praseeda and Shivakumar, "A Review of Trends and Technologies in Business Analytics."

[25]Ibid.

[26]Delen and Zolbanin, "The Analytics Paradigm in Business Research."

[27]Watson ioT, "Descriptive, Predictive, Prescriptive."

[28]Sharda, Delen, and Turban, *Business Intelligence Analytics and Data Science.*

[29]Sharda, Delen, and Turban, *Business Intelligence Analytics and Data Science*.

[30]Delen and Zolbanin, "The Analytics Paradigm in Business Research."

[31]Sharda, Delen, and Turban, *Business Intelligence Analytics and Data Science.*

[32]Praseeda and Shivakumar, "A Review of Trends and Technologies in Business Analytics."

based on descriptive and predictive analyses of complex data, and then suggesting courses of action.[33] Prescriptive analytics evaluates data and determines new ways to operate while balancing all constraints,[34] at the same time continually and automatically processing new data to improve recommendations and provide better decision options.[35] It not only foresees what will happen and when, but also suggests why it will happen and recommends how to act in order to take advantage of the predictions.[36] Predictive analytics uses a set of mathematical techniques that computationally determine an optimal action or decision given a complex set of objectives, requirements, and constraints.[37]

Topics in prescriptive analytics come from the fields of management science or operations research and are generally aimed at optimizing the performance of a system, using tools such as mathematical programming, simulation, network analysis, and others.[38]

1.5 Data Mining and Data Visualization

LEARNING OBJECTIVE 1.5

Describe the data mining and data visualization processes.

The dawning of the big data era has given rise to new and promising prospects for improving business decisions and outcomes. Two main components in the process of transforming the mountains of data now available into useful business information are data mining and data visualization.

data mining 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.

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 meaningful information that can potentially lead to some business advantage. Data mining allows business people to discover and interpret useful information that will help them 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. Data scientists often refer to the first three steps of this process as ETL (extract, transform, and load). Extracting involves locating the data by asking the question "Where can it be found?" Countless pieces of data are unearthed the world over in a great variety of forms and from multiple disparate sources. Because of this, extracting data can be the most time-consuming step.[39] 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 that determines 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.[40] Often the data are sorted into columns and rows to improve usability

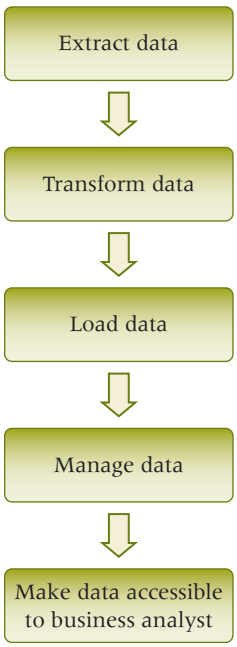

FIGURE 1.7 Process of Data Mining

[33]Watson ioT, "Descriptive, Predictive, Prescriptive."

[34]Daniel, "Big Data and Analytics in Higher Education."

[35]A. Basu, "Five Pillars of Prescriptive Analytics Success," *Analytics* (March/April 2013): 8–12, analytics-magazine.org/executive-edge-five-pillars-of-prescriptive-analytics-success/.

[36]Praseeda and Shivakumar, "A Review of Trends and Technologies in Business Analytics."

[37]I. Lusting, B. Dietrich, C. Johnson, and C. Dziekan, "The Analytics Journey," *Analytics Magazine* 3, no. 6 (2010): 11–3.

[38]Sharda, Delen, and Turban, *Business Intelligence Analytics and Data Science*.

[39]Shirley Zhao, "What Is ETL? (Extract, Transform, Load)," paper from Experian Data Quality, October 20, 2017, www.edq.com/blog/what-is-etl-extract-transform-load/.

[40]S. Wu, "A Review on Coarse Warranty Data and Analysis," *Reliability Engineering and System Safety* 114 (2013): 1–11, doi.org/10.1016/j.ress.2012.12.021.

and searchability.[41] 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, which is a software system that enables users to define, create, maintain, and control access to the database.[42] 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, which has been described as perhaps the most useful component of business analytics, and the element that makes it truly unique.[43] So what is data visualization? Generally, 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*.

data visualization The study of the visual representation of data 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 it holds and to making the most informed decisions.[44] Using visual techniques to convey information hidden in data allows a broader audience with a wider range of backgrounds to view and understand its meaning. Data visualization tools help make 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.[45]

Visualization Example

As an example of data visualization, consider the top five organizations in the manufacturing industry to receive Canadian government funding and their respective amounts funded (in dollars) in a recent year, displayed in **Table 1.1**.[46]

One of the leading companies to develop data visualization software is Tableau©. **Figure 1.8** is a bubble chart of the Table 1.1 data, developed using Tableau©. One of the advantages of using visualization to display data is the variety of types of graphs or charts that can be produced. **Figure 1.9** contains a Tableau©-produced bar chart of the same information to give the user a different perspective.

TABLE 1.1 Top Five Manufacturing Firms to Receive Canadian Government Funding

Company Name	Dollars Funded
FCA Canada Inc. (Chrysler)	$85,800,000
Bombardier Inc.	$54,150,000
Produits Kruger	$39,500,000
Sonaca Montréal Inc.	$23,250,000
Hanwha L&C Canada Inc.	$15,000,000

[41]Zhao, "What Is ETL?"

[42]Thomas M. Connolly and Carolyn E. Begg, *Database Systems: A Practical Approach to Design Implementation and Management*, 6th ed. (Harlow, UK: Pearson, 2014), 64.

[43]James R. Evans. *Business Analytics: Methods, Models, and Decisions*, 2nd ed. (Boston: Pearson, 2016), 7.

[44]R. Roberts, R. Laramee, P. Brookes, G.A. Smith, T. D'Cruze, and M.J. Roach, "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*, vol. 3 (Setúbal, Portugal: SciTePress, 2018), 319–26.

[45]Stephen Few, "Eenie, Meenie, Minie, Moe: Selecting the Right Graph for Your Message," paper from Perceptual Edge, 2004, www.perceptualedge.com/articles/ie/the_right_graph.pdf.

[46]"Top 50 Companies That Received Canadian Government Funding," *The Globe and Mail*, July 19, 2017, www.theglobeandmail.com/report-on-business/top-50-report-on-business-the-funding-portal/article19192109/.

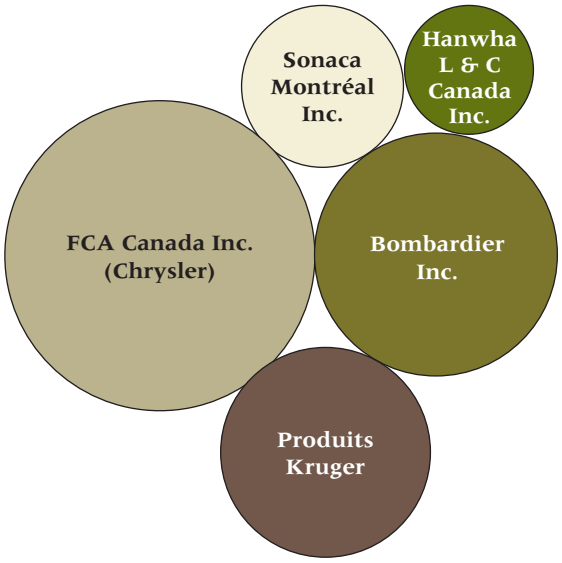

FIGURE 1.8 **Bubble Chart of the Top Five Manufacturing Firms to Receive Canadian Government Funding**

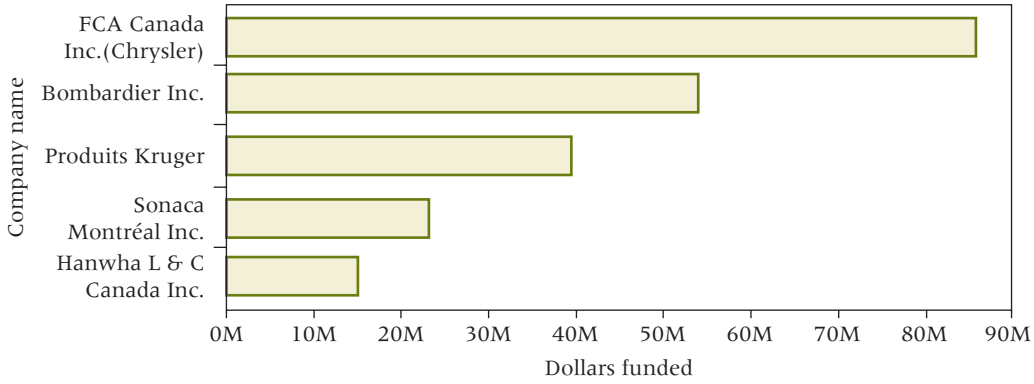

FIGURE 1.9 **Bar Chart of the Top Five Manufacturing Firms to Receive Canadian Government Funding**

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 India 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 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 are then used to estimate population parameters, such as percentage of households with a mobile connection, and so forth.

In conducting research on a topic like consumer consumption in rural India, a wide variety of statistics can be gathered that represent several levels of data. For example, ratio-level measurements of items such as income, number of children, age of household heads, number of heads of livestock, and grams of toothpaste consumed per year might be obtained. On the other hand, if analysts use a Likert-type scale (1-to-5 measurements) to gather responses about rural Indian consumers' interests, likes, and preferences, an ordinal-level measurement would be obtained, as with the ranking of products or brands in market research studies. Other variables, such as geographic location, 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 and operations capacity, scheduling issues, transportation challenges, financial commitments, managerial growth or reassignment issues, accounting issues (accounting for rural India may differ from techniques used in urban markets), information systems, and other related areas. With so much on the line, company decision-makers need as much relevant information 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 approximately 900 million people living in rural India certainly make it a market segment worth studying further.

Key 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 behaviour to use statistics out of context. Unethical business people 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 favour.

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 businesses 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.

Why Statistics Is Relevant

In the contemporary business world, good decisions are driven by data. In all areas of business, amazing amounts of diverse data are available for interpretation and quantitative insight. Business managers and professionals are increasingly required to justify decisions on the basis of data analysis. Thus, the ability to extract useful information from data is one of the most important, and marketable, skills business managers and professionals can acquire. This text is an introduction to the theory and, more importantly, the methods used to intelligently collect, analyze, and interpret data relevant to business decision-making.

Summary of Learning Objectives

LEARNING OBJECTIVE 1.1 Define important statistical terms, including population, sample, and parameter, as they relate to descriptive and inferential statistics.

Statistics is an important decision-making tool in business and is used in virtually every area of business. In this text, the word *statistics* is defined as the science of gathering, analyzing, interpreting, and presenting 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 from the process of gathering sample data from a group, body, or population and reaching conclusions about the larger group from which the sample was drawn.

LEARNING OBJECTIVE 1.2 Explain the difference between variables, measurement, and data, and compare the four different levels of data: nominal, ordinal, interval, and ratio.

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* is produced when a standard process is used to assign numbers to particular attributes or characteristics of a variable. Measurements of 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 the classification of only data, such as geographic location, language, or social insurance 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 highest level of data measurement, in which the distances represented by consecutive numbers are equal. The highest level of data measurement is ratio. This has all the qualities of interval measurement, but ratio data contain an absolute zero, and ratios between numbers are meaningful. Interval and ratio data are sometimes called *metric* or *quantitative* data. Nominal and ordinal data are sometimes 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.

LEARNING OBJECTIVE 1.3 Explain the differences between the four dimensions of big data.

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 data sets 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 at 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.

LEARNING OBJECTIVE 1.4 Compare and contrast the three categories of business analytics.

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.

LEARNING OBJECTIVE 1.5 Describe the data mining and data visualization processes.

Two main components in the process of transforming the mountains of data now available are data mining and data visualization. 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 10
business analytics 11
census 3
data 5
data mining 13
data visualization 14
descriptive analytics 12
descriptive statistics 4
inferential statistics 4
interval-level data 8
measurement 5
metric data 8
nominal-level data 6
nonmetric data 7
nonparametric statistics 8
ordinal-level data 7
parameter 4
parametric statistics 8
population 3
predictive analytics 12
prescriptive analytics 12
ratio-level data 8
sample 4
statistic 4
statistics 3
variable 5
variety 10
velocity 10
veracity 10
volume 10

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, operations and supply chain management, information systems, production, and management. An example in the marketing area might be "number of sales per month by each salesperson."

1.2. Provide 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 health care. 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, and so on. 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. For each variable, briefly describe a numerical observation that might be the result of measuring the variable.

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

- **a.** The time required to produce each tire on an assembly line
- **b.** The number of litres of milk a family drinks in a month
- **c.** The ranking of four machine operators in your plant according to experience
- **d.** The telephone area codes of clients in Canada
- **e.** The age of each of your employees
- **f.** The dollar sales at the local pizza house each month

g. An employee's identification number

h. The response time of an emergency unit

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

a. The ranking of a company by *Report on Business* magazine's Top 1000

b. The number of tickets sold at a movie theatre on any given night

c. The identification number on a questionnaire

d. Per capita income

e. The trade balance in dollars

f. Profit/loss in dollars

g. A company's tax identification number

h. 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. Video The Mapletech Manufacturing Company makes electric wiring, which it sells to contractors in the construction industry. Approximately 900 electrical contractors purchase wire from Mapletech annually. Mapletech's director of marketing wants to determine electrical contractors' satisfaction with Mapletech's wire. She develops a questionnaire that yields a satisfaction score of 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.

a. What is the population for this study?

b. What is the sample for this study?

c. What is the statistic for this study?

d. What would be a parameter for this study?

Exploring the Databases with Business Analytics

See the databases in Wiley's online resources

Six major databases constructed for this text can be used to apply the business analytics and statistical techniques presented in this course. These databases are located in Wiley's online course, and each one is available in a few electronic formats for your convenience. These six databases represent a wide variety of business areas: the stock market, international labour, finance, energy, and agri-business. The data are gathered from such reliable sources as Statistics Canada, the Toronto Stock Exchange, *Yahoo Finance*, the Fraser Institute, and the Global Environment Outlook (GEO) Data Portal. Four of the six databases contain time-series data that can be especially useful in forecasting and regression analysis. Here is a description of each database, along with information that may help you to interpret outcomes.

Canadian Stock Market Database

The stock market database contains seven variables on the Toronto Stock Exchange. Weekly observations for a period of five years yield a total of 257 observations per variable. The variables are Composite Index, Energy Index, Financial Index, Health Index, Utility Index, I.T. Index, and Gold Index. These data were obtained directly from TMX Inc.

Canadian RRSP Contribution Database

The registered retirement savings plan database contains five variables: Number of Tax Filers, Total RRSP Contributors, Average Age of RRSP Contributors, Total RRSP Contributions ($ × 1,000), and Median RRSP Contributions ($). There are 156 entries for each variable in this database, representing 10 years of data for each of 10 provinces and 3 territories in Canada. The data in this database were obtained from Statistics Canada.

International Labour Database

This time-series database contains the civilian unemployment rates in percent from seven countries (G7) presented yearly over a 30-year period. The data are published by the Bureau of Labor Statistics of the U.S. Department of Labor. The countries are Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States.

Financial Database

The financial database contains observations on 12 variables for 100 companies. The variables are Industry Group, Type of Industry, Total Revenues, Price, Average Yield, Dividend Growth, Average Price/Earnings (P/E) Ratio, Dividends per Share, Total Debt/Total Equity, Price/Cash Flow, Price/Book, and One-Year Total Return. The data were gathered from the Toronto Stock Exchange. The companies represent seven different types of industries. The variable "Type" displays a company's industry type as follows:

1 = Real Estate

2 = Financial Institutions

3 = Chemicals

4 = Mining, Electric, Oil & Gas, Pipelines

5 = Telecommunications, Retail, Commercial Services, Auto Parts & Equipment, Transportation

6 = Insurance

7 = Food, Pharmaceuticals, Electronics, Media, Aerospace/Defence, Hand/Machine Tools, Agriculture, Iron/Steel, Holding Companies, Engineering and Construction, Machinery—Diversified, Forest Products and Paper

Energy Resource Database

The energy resource database consists of data for North America and Europe on nine variables related to energy supply and emission of carbon dioxide over a period of 35 years. The database is adapted from the Global Environment Outlook (GEO) Data Portal. The nine variables are the seven total primary energy supplies of Solar, Wind, Tide, and Wave; Nuclear; Natural Gas; Coal and Coal Products; Crude Oil; Hydro; and Petroleum Products, plus the two sources of emissions of CO_2: Manufacturing Industries and Construction, and Transportation.

Agri-Business Canada Database

The agri-business time-series database contains the monthly weight (in thousands of pounds) of seven different grains over an eight-year period. Each of the seven variables represents 94 months of data. The seven grains are Wheat; Wheat, Excluding Durum; Durum Wheat; Oats; Barley; Flaxseed; and Canola. The data are published by Statistics Canada under Agri-business.

Use the databases to answer the following question.

1. In the Financial Database, what is the level of data for each of the following variables?
   - **a.** Type of Industry
   - **b.** Average Price
   - **c.** Price/Cash Flow

Case

Canadian Farmers Dealing with Stress

Farming is an extremely demanding job whose success depends largely on the dedication of the farmers. In order to tackle the rigorous tasks of the trade, farmers must be in good physical and mental condition. The University of Guelph conducted a study in 2017–18 which showed that Canadian farmers experienced higher levels of anxiety and depression and displayed lower levels of help-seeking behaviour than the general Canadian populace. The study indicated that Canadian farmers were at a heightened risk of suicide compared to the rest of the population. The Canadian Agricultural Safety Association (CASA) also recently sponsored research on the stress level of Canadian farmers.

Western Opinion Research Inc. conducted the research study, which was completed by 1,100 farmers across Canada. The survey asked farmers to rate their stress level, ranging from "stressed" to "somewhat stressed" to "very stressed." The results varied and revealed the following: two-thirds of farmers indicated feeling "stressed," 45% indicated feeling "somewhat stressed," and one in five indicated feeling "very stressed." The results of the survey also revealed that the major causes of stress among farmers were (a) financial concerns related to prices of commodities, (b) diseases affecting livestock, and (c) finances regarding general farm expenses. The results also revealed that a large percentage of farmers (35%) were interested in having access to more stress-related resources in order to help alleviate their stress level.

This study allowed CASA to realize that stress within the farming industry is a major issue, making it imperative to take action and offer stress counselling resources to farmers. These resources include (a) confidential meeting with a health care professional, (b) over-the-phone consultation with a health care professional, and (c) attending workshops such as retreats that focus on relaxation techniques and role playing with other farmers.

Over the last few years, CASA has made great progress to help reduce the stress level of Canadian farmers. For example, successful counselling services that deal with farmers and their problems were established using phone, email, and online chat help lines. These include the Manitoba Farm, Rural & Northern Support Services and the Saskatchewan Farm Stress Line. These services are well recognized and are having great success in helping farmers, mainly because they are staffed by paid professional counsellors who all have farming backgrounds, making the farmers feel more comfortable because they are connecting with someone who understands their work-related issues.

Discussion

Think of the market research that was conducted by CASA.

1. What are some of the populations that CASA might have been interested in measuring for these studies? Did CASA attempt to contact entire populations? What samples were taken? In light of these two questions, how was the inferential process used by CASA in its market research? Can you think of any descriptive statistics that might have been used by CASA in its decision-making process?
2. In the various market research efforts made by CASA to determine stress experienced by farmers, some of the possible measurements appear in the following list. Categorize these by level of data. Think of some other measurements that CASA analysts might have taken to help them in this research effort and categorize them by level of data.
   - **a.** Ranking of the level of stress on a stress test
   - **b.** Number of farmers who ask for professional help
   - **c.** Number of farmers who are aware of professional help resources
   - **d.** Number of farmers who try to manage stress on their own
   - **e.** Number of farmers who are interested in having access to more stress-related resources
   - **f.** Number of farmers who are close to being out of business
   - **g.** Number of farmers who would prefer dealing with stress on their own
   - **h.** Number of farmers who would prefer dealing with stress with a professional over the telephone
   - **i.** Number of farmers who would prefer dealing with stress with a professional in person
   - **j.** Age of survey respondent
   - **k.** Gender of survey respondent
   - **l.** Geographical region of survey respondent
   - **m.** Amount of time farmers spend dealing with their stress
   - **n.** Rating of the most stress-related factors on a scale from 1 to 10, where 1 is the least stress-related factor and 10 is the most stress-related factor
   - **o.** Rating of the reasons why farmers do not seek more help for stress on a scale from 1 to 10, where 1 is the least important reason and 10 is the most important reason

Big Data Case

Virtually every chapter in this text will end with a big data case. Unlike the chapter cases, which feature a wide variety of companies and business scenarios, the big data case will be based on data from a single industry, U.S. hospitals. Drawing from the American Hospital Association database of over 2,000 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.

Using Software for Data Analysis

Statistical Analysis Using Excel

The advent of the modern computer opened up 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 analysts before computers were developed. In this textbook, the end-of-chapter "Using the Computer" feature will show you how you can use software to apply the statistical techniques you learn in each chapter.

- Business statisticians use many popular statistical software packages, including MINITAB, SAS, and SPSS. Many computer spreadsheet software packages can also analyze data statistically. In this text, when it is appropriate to use a computer, we will use Microsoft Excel for data analysis.
- Excel is by far the most commonly used spreadsheet for PCs, making it the obvious choice for basic statistical analysis. Excel can perform a variety of calculations and includes a large collection of statistical functions. The Data Analysis ToolPak provides a further suite of statistical macro-functions.
- We note, however, that although Excel is a fine spreadsheet, it is not a professional statistical data analysis package: there are some important limitations. Key limitations to keep in mind include the following: missing values in Excel are difficult to handle when performing data analysis; data organization differs according to analysis, sometimes forcing you to reorganize your data; some output and charts produced by Excel are of poor quality from a statistical point of view and are sometimes inadequately labelled; and there is no record of how an analysis was done if the Data Analysis ToolPak is used.
- Despite these limitations, Excel is the most commonly used package in the business environment and is the package you are most likely to use in your professional life.
- It is important to remember that a statistical software package is not a replacement for a thorough understanding of correct statistical methods. The business analyst is responsible for determining the most appropriate statistical methods for a given business problem. Simply relying on convenient software tools that may be at hand without thinking through the most appropriate approach can lead to errors, oversights, and poor decisions. One of the goals of this text is to show you when and how to use statistical methods to provide information that can be used in business decisions.

Answers

Concept Check 1.1

1. Descriptive statistics can be used to summarize the data to describe a data sample either numerically or graphically. **2.** Statistical inference is inference about a population from a random data sample drawn from it.

Concept Check 1.2

1. c. **2.** d. **3.** b. **4.** a.

Odd-Numbered Supplementary Problems

1.7. a. ratio **b.** ratio **c.** ordinal **d.** nominal **e.** ordinal **f.** ratio **g.** nominal **h.** ratio

1.9. a. 900 electric contractors **b.** 35 contractors **c.** average score for 35 participants **d.** average score for all 900 electric contractors