CHAPTER 1

What Is Statistics?

GOALS

When you have completed this chapter, you will be able to:

- Explain what is meant by statistics
- Identify the role of statistics in the development of knowledge and everyday life
- Explain what is meant by descriptive statistics and inferential statistics
- Distinguish between a qualitative variable and a quantitative variable
- Distinguish between a discrete variable and a continuous variable
- Collect data from published and unpublished sources
- Distinguish among the nominal, ordinal, interval, and ratio levels of measurement
- Identify abuses of statistics
- Gain an overview of the art and science of statistics. We recommend that you read this chapter at least twice, once at the beginning and once at the end of your course!

1498
1548
1598
1648
1698
1748
1898
1948
2000

"God does not play dice with Nature", said Einstein, but what about His creation?

What do games, gods, and gambling have to do with statistics? Interestingly enough, human pursuits in games of chance, divinity, and gambling have served as the sources of inspiration for the development of modern statistics.

Games of chance date back to antiquity. Archaeologists have found dice and dice-like bones in many early civilizations dating back from 1000 to 3000 B.C., including regions in the Indus valley, Babylonia, Mesopotamia, Greece, and Rome. The great Vedic epic *The Mahabharata*, written about 1000 B.C., describes how gambling among the princes resulted in the Great War and an eventual fall of the great empire. Augustus (63 B.C.–A.D. 14), the first Roman emperor, once wrote to his daughter: "I send you 250 *denarii*, the sum I gave to each of my guests in case they wished to play at dice or at odd and even during dinner"![1] Claudius (10 B.C.–A.D. 54) is known to have published a book on the art of dicing, and is known "to play while driving, having the board game fitted to his carriage in such a way as to prevent his game from being disturbed."[2]

Chance outcomes have been used as an indication of the divine will in many ancient religions. The goddess Fortuna (Roman and Greek: see the picture above) and the goddess Laxmi (Hindu) were (and still are, by many) often prayed to for a favourable lot at games of chance. In the Old Testament, "the lot is cast into the lap; but the whole disposing thereof is of the Lord" (Proverb 16:33). In ancient Assyria, the king gave his own name to the first year of his reign; the names of subsequent years were determined by a lot. On a die found in Assyria (from the year 833 B.C.), it is inscribed: "O great lord, Assur! O great lord, Adad! This is the lot of Jahali,... king of Assyria,... make prosper the harvest of Assyria... . May his lot come up!"[3] The Chinese used *I Ching* (one of the five books written by Confucius) together with coins or milfoil to create an oracle that they believed to be the result of a partnership between god and humans.

The search for formal methods for experimental knowledge did not begin until the 15th century. Often inspired by games of chance and problems in gambling, efforts in this direction were pioneered by Leonardo da Vinci and Cardano from Italy, and Fermat and Pascal from France. Some of the reasons speculated for such a late development of the experimental methods include the prevalence of an Aristotelian (deterministic) mindset and an opposition of the Church to all types of games of chance and casting of lots.

INTRODUCTION

More than 100 years ago, H.G. Wells, an English author and historian, noted that statistical thinking will one day be as necessary for efficient citizenship as the ability to read. He made no mention of business, because the Industrial Revolution was just beginning. Were he to comment on statistical thinking today, he would probably say that "statistical thinking is necessary not only for effective citizenship but also for effective decision making in various facets of business".

The late W. Edwards Deming, a noted statistician and quality control expert, insisted that statistics education should begin before high school. He liked to tell the story of an 11-year-old who devised a quality control chart to track the on-time performance of his school bus. Deming commented, "He's got a good start in life." We hope that this book will give you a solid foundation in statistics for your future studies and work in economics and business related fields.

Almost daily we apply statistical concepts to our lives. For example, to start the day you turn on the shower and let it run for a few moments. Then you put your hand in the shower to sample the temperature and decide to add more hot water or more cold water. You then conclude that the temperature is just right and enter the shower. As a second example, suppose you are at the grocery store looking to buy a frozen pizza. One of the pizza makers has a sample stand, and they offer a small wedge of their pizza. After sampling the pizza, you decide whether to purchase the pizza or not. In both the shower and pizza examples, you make a decision and select a course of action based on a sample.

Businesses are faced with similar problems. The Kellogg Company must ensure that the average amount of Raisin Bran in the 25.5 g package meets the label specifications. To do so, they select periodic random samples from the production area and weigh the contents. In politics, a candidate for the position of Member of Parliament wants to know what percentage of the voters in his riding will support him in the upcoming election. There are several ways he could go about answering this question. He could have his staff call all those people in his riding who plan to vote in the upcoming election and ask for whom they plan to vote. He could go out on a street in his riding, stop 10 people who look to be of voting age, and ask them for whom they plan to vote. He could select a random cross-section of about 1000 voters from his riding, contact these voters and, based on this cross-section, make an estimate of the percentage who will vote for him in the upcoming election. In this text we will show you why the third choice is the best course of action.

1.1 WHAT IS MEANT BY STATISTICS?

The word *statistics* owes its origin to the German word *statistik*, used to describe numerical information on economic, social, political, cultural, or other characteristics of a nation state. In popular usage, the word is still used to express some numerical information not necessarily related to the characteristics of a state. Examples include the number of Lotto 6/49 tickets sold before the next draw, the number of students enrolled at Concordia University this year, the value of donuts sold at Tim Hortons in Calgary last week, the number of tourists visiting PEI last summer, or the change in the value of the Toronto Stock Exchange Index last Friday. In these examples, statistics is a single number.

STATISTICS IN ACTION

Improvement in the Level of Well-Being

The level of well-being of an individual depends on, not only income, but also on several other factors such as the stock of wealth (including human capital and natural resources); the environment; security from unemployment; illness, and poverty; and income equality. The Centre for the Study of Living Standards (www.csls.ca) computes a well-being index based on a number of such factors. A recent calculation of the well-being index by CSLS for Canada and the provinces is shown in Table 1-1. The chart shows how Canada and each of the provinces have enhanced the quality of life during the period 1971 to 1997. To explain the progress in the form of an index, the CSLS sets the value of well-being in 1971 equal to 100. Thus the number for PEI (125.3) indicates that well-being in PEI improved by 25.3% during the 1971–97 period.

Likewise, the word *statistics* is also used to describe a large amount of information through summary measures, called *descriptive statistics*. Examples include the average income of all households in Canada, the average driving speed of cars on Highway 401, and the variation (say, minimum and maximum values) in Bell Canada stock prices during the last week. More examples are given below.

- In 1998, 45 percent of Canadian households owned a computer and 25 percent were connected to the Internet.
- In 2000, 358 870 adults and 113 598 youths were charged in criminal incidents in Canada.
- In 1998, there were 54 763 police officers and 58 198 teachers in universities and community colleges in Canada.
- On average, Canadians spend 1.3 hours per day commuting, and 1.5 hours per day with their children.
- In July 2000, average weekly earnings (all employees) of Canadians were $665.41.
- In 2000, John Roth, president and CEO of Nortel Networks Corp., was the best paid ($71 million) Canadian company executive.
- On average, Canadians listen to the radio for nearly 21 hours per week.

Larger amounts of information can also be organized in tabular form or as graphical displays (see Table 1-1 and Chart 1-1) to describe patterns in the data. The graphical display of a large quantity of information allows easier interpretation than the tabular form. We study these methods in the next chapter.

While the descriptive or numerical methods have played a significant role in the early development of statistics, it is the interplay of the numerical methods and the theories underlying probability and probability distributions that constitute the core of modern statistics. This enables us to draw inferences about the *whole* from information on only a *part* of that whole. This is called *statistical inference*. For example, we can draw inferences about average income and/or variations in the incomes of about 12 million Canadian households based on the information on only, say, 1200 or 12 000 households. Consistent with the inferential aspects of statistics, statisticians define the word *statistic* as an estimator of a certain attribute (average, variation, etc.) of the entire

Index of Provincial Well-Being and Improvement in the Level of Well-Being

TABLE 1-1

Province	Well-Being
CAN	112.53
NF	106.98
PEI	125.26
NS	107.12
NB	114.09
QC	106.89
ON	113.78
MB	120.74
SK	119.53
AB	120.05
BC	113.67

CHART 1-1: Improvement in the Level of Well-Being (1971–97, 1971=100)

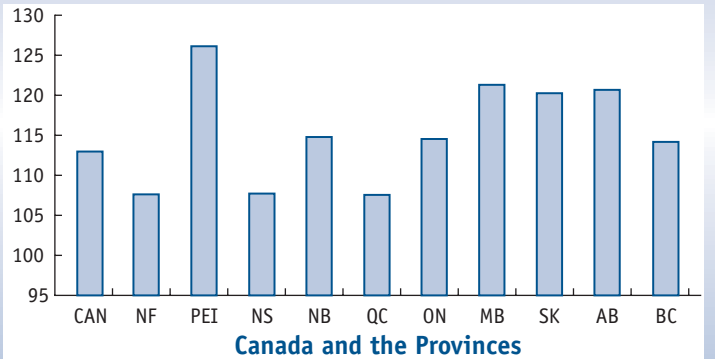
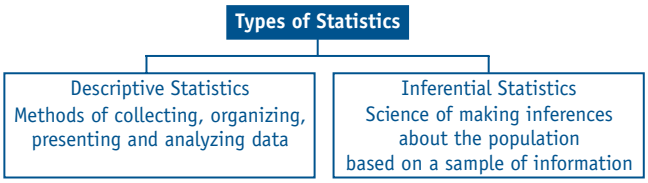


CHART 1-2: Types of Statistics



data (called a "population") based on only a small part of that data (called a "sample"). In other words, statistics is a method of inquiry that enables us to make scientific generalizations about the world from the knowledge of only a small part of that world. We shall discuss this process in some detail in the next section. Thus, the field of statistics includes both descriptive statistics and inferential statistics. Further, the statistical methods can be used for prediction and policy analysis. For a visual representation of the types of statistics, see Chart 1-2 above.

Statistics The art and science of collecting, organizing and presenting data, drawing inferences from a sample of information about the entire population, as well as prediction and policy analysis.

Statistics is both an art and a science. It is a science inasmuch as it depends on scientific theories underlying inference. It is an art in the sense that it requires a good understanding of how to bring various statistical techniques together to understand real-world phenomena and use them for prediction and policy analysis.

Note the words "population" and "sample" in the definition of inferential statistics. We often make reference to the population of 31 million people in Canada, or more than 1 billion people in China. However, in statistics the word *population* has a broader meaning. A population may consist of *individuals*, such as all the students enrolled at the University of Toronto, all the students in Accounting 2001, or all the inmates at the Kingston prison. A population may also consist of *objects*, such as copies of this book printed by McGraw-Hill Ryerson this year, or all the $100 bills in circulation in Canada. A population may also consist of a group of *measurements*, such as all the weights of the defensive linemen on the University of British Columbia's football team or all the heights of the basketball players at Memorial University of Newfoundland. Thus, a population in the statistical sense of the word does not necessarily refer to people.

POPULATION AND PARAMETER

Population is defined as a collection or as a set of all elements (individuals, objects, or measurements) of interest in an investigation. A *parameter* is a summary measure of some characteristic such as an average, a proportion, or a variation in all elements of the population.

To infer something about a population, we usually take a sample from the population.

SAMPLE AND STATISTIC

Sample is a portion, or subset, of all elements in the population. A statistic is an estimator of the parameter of interest in the population.

REASONS FOR SAMPLING

Why take a sample instead of studying every member (element) of the population?

- ***Costs of surveying the entire population may be too large or prohibitive:*** A sample of registered voters is necessary because of the prohibitive cost of contacting millions of voters before an election.
- ***Destruction of elements during investigation:*** If Sylvania tried to estimate the life of all its 60-watt light bulbs by actually using them until they burned out, there would be nothing left to sell. Testing wheat for moisture content destroys the wheat, thus making sampling imperative.
- ***Accuracy of results:*** Assuming proper techniques of sampling are used, results based on samples may be close in accuracy to the results based on population. Further, as discussed in Chapter 8, we can identify the extent of error due to sampling in most cases.

As noted, taking a sample to learn something about a population is done extensively in business, economics, agriculture, politics, and government. Example 1-1 illustrates the use of a sample in relation to population (compare the relative sizes of the sample and the population) by Statistics Canada.

Example 1-1

Internet Shopping

The HIUS (Household Internet Use Survey) is administered by Statistics Canada to a sub-sample of dwellings included in the Labour Force Survey (LFS), and therefore its sample design is closely tied to the LFS. The LFS is a monthly household survey whose sample is representative of the households with civilian, non-institutionalized population, 15 years of age or older, in Canada's 10 provinces.

In total, 43 034 households were eligible for the HIUS survey. Interviews were completed for 36 241 of these households for a response rate of 84.2 percent. Results were weighted to the entire count of households in Canada. The annual estimate for the number of households in Canada is projected from the census of population. The HIUS used a population projection based on the 1996 Census of population (11.632 million households).

Based on the information collected on 36 241 households, the authors of the report conclude:

> In 1999, 1.8 million households indicated that at least one member of their household had engaged in some aspect of Internet shopping from home, either using the Internet as part of their buying process by researching characteristics and prices of goods and services (window shopping) or placing orders for purchases on line… . There were 806 000 households that took the extra step and actually engaged in e-commerce (*those Internet shoppers that did place at least one order over the Internet from home*).[4]

Some additional examples of the use of a sample in relation to population are provided below.

- Television networks constantly monitor the popularity of their programs by hiring Nielsen and other organizations to sample the preferences of TV viewers. These program ratings are used to set advertising rates and to cancel programs.
- A public accounting firm selects a random sample of 100 invoices and checks each invoice for accuracy. There were errors on five of the invoices; hence, the accounting firm estimates that 5 percent of the entire population of invoices contains an error.
- A random sample of 260 accounting graduates from community colleges showed that the mean starting salary was $32 694. We therefore conclude that the mean starting salary for all accounting graduates from community colleges is $32 694.

- In the week after the terrorist attacks on New York and Washington on September 11, 2001, Prime Minister Jean Chrétien's personal approval rating for his political performance shot up from 57 to 65 percent. The results are based on polls conducted jointly by *The Globe and Mail*, CTV, and Ipsos-Reid. The calculated margin of error for these polls is 3.1 percent.

In the above examples, the mean starting salary of $32 694 and Jean Chrétien's approval ratings of 57 and 65 percent are estimates of mean starting salary and approval rating (parameters) in the respective populations. Actual values of the parameters will not be known unless we calculate the mean starting salary or approval rating based on all elements in each population. Since it is too costly or time-consuming to collect information on all elements in the population, we use statistics to draw inferences about the population parameters. Statistical techniques enable us estimate parameters and draw inferences with a level of confidence such as a maximum of 3.1 percent margin of error (in the pollsters' estimate of the parameter) mentioned in the approval rating case.

1.2 THE ROLE OF STATISTICS IN THE DEVELOPMENT OF KNOWLEDGE

The development of knowledge, since antiquity, has proceeded through a number of methods such as intuition, revelation, abstraction, and experimentation. However, abstraction and experimentation have been the most popular methods of advancing the frontiers of knowledge in the material world.

THE METHOD OF ABSTRACTION

The method of abstraction involves *deducing* a hypothesis about a phenomenon in the real world from a set of definitions and assumptions and following the rules of logic. For example, suppose you are interested in understanding why people buy more apples at lower prices. A theorist may explain (deduce) this phenomenon as follows:

DEFINITIONS AND ASSUMPTIONS

1. People are rational. That is, they have well-defined preferences and always prefer more of a commodity to less of that commodity.
2. Except for the price of apples, all other factors that may influence purchases of apples (such as people's incomes, prices of substitutes for apples, people's tastes for apples) remain unchanged.
3. Each additional apple they buy/consume gives them a lower level of satisfaction. Thus, if we could define satisfaction in terms of some imaginary units such as *utils* (short for utility), then we might assume that the first apple gives them 50 *utils*, the second apple 35 *utils*, and the third apple 22 *utils* of satisfaction, and so on.
4. Parting with money yields dissatisfaction. Suppose that each penny spent gives them one *util* of dissatisfaction. Thus, if they spend 50 cents it gives them 50 *utils* of dissatisfaction (negative 50 *utils*).

REASONING

Since people derive lower satisfaction from each additional apple, and since they lose one unit of satisfaction from each penny spent, the maximum they would be willing to

pay for the first apple is 50 cents; for the second apple 35 cents; for the third apple 22 cents, and so on.

IMPLICATIONS/HYPOTHESES

People will buy more apples at lower prices. This is also called the *law of demand*. You can even make a prediction that a tax on apples that increases the price of apples would result in lower sales/purchases of apples.

Thus, in this method, we start from our observation of a general pattern in some real-world phenomenon and draw a particular conclusion about that phenomenon. Our journey in this method is therefore *from general to particular*. Since in this method we deduce a particular conclusion/hypothesis based on a general pattern, we call it the *deductive* method.

THE METHOD OF EXPERIMENTATION

The method of experimentation, on the other hand, proceeds in the opposite direction. For example, in the case of demand for apples we would need to decide how the relevant information on the quantity and price of apples would be collected (the design of the experiment), collect data for the experiment, use statistical methods to analyze the data, and then draw inferences about the buying behaviour of apple consumers in the real world.

In this method, we induce a relation based on the information from the real world. In this method, in general, we proceed as follows:

DEFINE THE EXPERIMENTAL GOAL OR A WORKING HYPOTHESIS

The goal or the working hypothesis may be based on some theory or experience. For example, we may have seen people buying a lower quantity of apples at a higher price, or we may use the law of demand derived by an economist.

DESIGN AN EXPERIMENT

We develop a method of collection of data for the relevant information on the variables of interest in a way that is consistent with both the theoretical hypothesis and the statistical techniques. For example, to verify the law of demand we would need to collect data in a way that, except for the price of apples, all other forces including income, tastes, prices of substitutes for apples, and so on remain unchanged. We could also use an experimental design that would enable us to isolate the effects of changes in price on quantity demanded. In general, the nature of experimental design is determined by the objective(s) of investigation. We discuss this aspect in Chapters 8 and 12.

COLLECT DATA

Given the experimental design, we collect data and check for its adequacy and accuracy. Application of proper survey methods is an essential part of data collection. Data may be collected from published sources, if available and reliable.

ESTIMATE THE VALUES/RELATIONS

We use a suitable statistical technique to estimate the value of an attribute, such as average, for a single variable or a relationship for multiple variables. By using the technique of two-variable regression analysis (Chapter 13), we can estimate the relationship between quantity and price of apples. Techniques for estimating certain attributes of a single variable, such as an average or variation, are discussed in Chapters 3 and 4.

STATISTICS IN ACTION

Productivity and Standard of Living

According to a theory, productivity (measured in terms of output per worker) plays a significant role in improving the standard of living (measured in terms of wages per worker). Chart 1-4 establishes an empirical relationship between productivity and standard of living based on a cross-section of observations of a few countries. Countries with higher productivity in general are seen with a higher standard of living. The scatter of observations (and the line going through these observations) shows a strong relationship between productivity and standard of living (see Chart 1-4). To improve the Canadian standard of living, Canadian policymakers can use this relationship as a guide to devise policies that would result in improvements in productivity.

DRAW INFERENCES

In most cases we use only part of all possible observations to estimate an attribute or a relationship underlying all observations. Accuracy of our estimates as a representation for all possible observations is therefore subject to error. Chapters 8 and 9 discuss the accuracy and precision aspects of an inference based on the foundations of probability and probability distributions discussed in Chapters 5 to 7. Using techniques developed in this book, we can draw a conclusion about the relationship between quantity demanded and price of apples. A negative relationship between the two variables will imply that people would buy fewer apples at higher prices.

PREDICTION AND POLICY ANALYSIS

Given an estimate of consumers' response in terms of change in quantity demanded to changes in prices, we can use the information to predict consumer purchases of apples to a policy-induced change in price. For example, the government may realize the importance of "an apple a day keeps the doctor away" and may therefore wish to subsidize the apple growers in an effort to reduce the price of apples (thus increasing the consumption of apples) and thereby reducing health-care costs!

Thus, in this method, we start from particular observations from the real world and draw conclusions about the general patterns in the real world. Our journey in this method is therefore from the *particular to the general*. Since in this method we induce a conclusion/hypothesis based on a particular set of observations, we call it the **inductive method**. *In practice, the two methods are often complementary in advancing the frontiers of knowledge*. Apples falling from the tree (rather than rising from the ground) can give an idea to a person like Newton to develop a theory of gravitation, and the hypotheses on gravitational forces can further be confirmed/corroborated through experimental methods. These ideas on the construction of knowledge based on two methods, as outlined above, are shown through a simplified diagram in Chart 1-3.

However, we should caution you against jumping to the conclusion of proving or disproving a theory based on statistical evidence. While theories are based on definitions, assumptions, and tight rules of logic, statistical evidence is based on accuracy of experimental design, a particular set of observations produced by the experiment, and continuously evolving methods of estimation and the science of measurement of uncertainty.

CHART 1-3: Characteristics of Abstraction and Experimentation

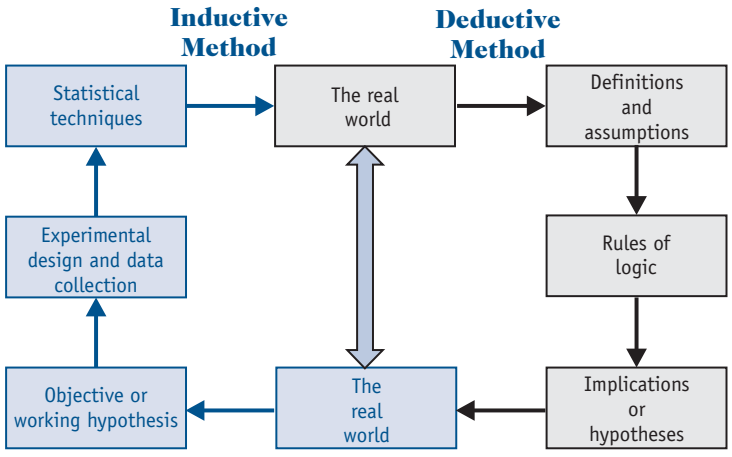


CHART 1-4: High Productivity Is Key to a High Standard of Living

Productivity is the most important determinant of a country's standard of living.

By international standards, Canada is doing relatively well, displaying both high productivity levels and a high standard of living.

—But, Canada can and should do better.

Productivity growth is the key to maintaining and improving living standards—it gives us the biggest "bang for the buck".
Conference Board
Performance & Potential, 1997

*In manufacturing
Source: International Yearbook of Industrial Statistics, Rochik (1998)

Source: Improving Productivity: The Key to Higher Living Standards: Industry Canada, Special Report, 2000, p. k-1

1.3 THE ROLE OF STATISTICS IN EVERYDAY LIFE

If you look through your college/university catalogue, you will find that statistics is a required course for many programs. Why is this so? What are the differences in the statistics courses taught in engineering, psychology, sociology, or business? The biggest difference is the examples used; the course content is basically the same. In engineering we may be interested in how many units are manufactured on a particular machine; in psychology or sociology we are interested in test scores; and in business we are interested in such things as profits, hours worked, and wages. However, all three are interested in what is a typical value and how much variation there is in the data. There may also be a difference in the level of mathematics required. An engineering statistics course usually requires calculus. Statistics courses in colleges of business and education usually teach the course at a more applied level. You should be able to handle the mathematics in this text if you have completed high-school algebra.

So why is statistics required in so many programs? The *first reason* is that numerical information is everywhere; look in newspapers (*The Globe and Mail* or your local newspaper), newsmagazines (*TIME*, *Newsweek*), business magazines (*The Economist*, *BusinessWeek*, or *Forbes*), general-interest magazines (*MACLEAN'S*), women's magazines (*Chatelaine, Canadian Living*), or sports magazines (*Sports Illustrated*), and you will be bombarded by numerical information. Here are some examples.

- Alberta's oil sands, by most estimates, contain more oil in the so-called "tar sands" than there is in all of Saudi Arabia, or about 300 billion barrels that are recoverable

using existing technology. That's enough to supply the United States for more than 40 years—plus there are another 1.5 to 2 trillion barrels on top of that, which would be harder to extract. That's 10 times what Saudi Arabia has.

- The Canadian Radio-Television Commission's (CRTC) first annual report on competition in Canadian telecommunications markets, released September 28, 2001, said this country's telecom-services industry was worth $28.7 billion in 2000 and has grown at an average of 9 percent a year since 1996.
- Golfer Tiger Woods turned pro in mid-1996 and racked up $2.7 million in tournament winnings in his first year on the Tour. At that rate, he was a bargain: according to one estimate, Woods generated $650 million in new revenues for television networks, equipment manufacturers, and other businesses in that first year. He now earns $20 million a year from his endorsement of Nike.
- Graduates of McGill University's Master's of Business Administration Program had a mean starting salary of $54 000 and 91 percent were employed within three months of graduation.
- On July 20, 2001, DaimlerChrysler reported a second-quarter loss of $125 million ($US), much better than the expected loss of $700 million ($US) and a big improvement from the first-quarter loss of $1.2 billion ($US).

How are we to determine if the conclusions reported are reasonable? Were the samples large enough? How were the sampled units selected? To be an educated consumer of this information, we need to be able to read the charts and graphs and understand the discussion of the numerical information. An understanding of the concepts of basic statistics will be a big help.

The *second reason* for taking a statistics course is that statistical techniques are used to make decisions that affect our daily lives. That is, they affect our personal welfare. Here are a few examples.

- Insurance companies use statistical analysis to set rates for home, automobile, life, and health insurance. Tables are available that summarize the probability of an auto accident by a 20-year-old female and a 20-year-old male. The differences in the probabilities are revealed in the differences in their insurance premiums.
- Air Canada slashed 5000 jobs at the end of September 2001 and cautioned that passengers will face less frequent service as it implements plans to ground 84 planes and cut one fifth of its capacity.
- As a result of the technology "meltdown" in 2001, Nortel slashed 30 000 jobs by September 2001 and planned to cut 20 000 more jobs in the remainder of the year.
- About 15 000 forestry workers in B.C. have been laid off as a result of the 19.3-percent US duty imposed on Canadian softwood lumber.
- Medical researchers study the cure rates for diseases, based on the use of different drugs and different forms of treatment. For example, what is the effect of treating a certain type of knee injury surgically or with physical therapy? If you take an Aspirin each day, does that reduce your risk of a heart attack?

A *third reason* for taking a statistics course is that the knowledge of statistical methods will help you understand how decisions are made and give you a better understanding of how they affect you. No matter what line of work you select, you will find yourself faced with decisions where an understanding of data analysis is helpful.

To make informed decisions, you will need to be able to:

- Define the objective (or hypothesis) of your inquiry/investigation.
- Determine the method and framework (design of experiment) for collection of the required information.
- Collect the required data from published and/or unpublished sources as necessary.
- Determine the adequacy and accuracy of the collected data and make changes as necessary.
- Estimate the required characteristics of the population as identified in the objective of your inquiry.
- Analyze the results.
- Draw inferences while assessing the risk of an incorrect conclusion.

The statistical methods presented in the text will provide you with a framework for the decision-making process.

In summary, there are at least three reasons for studying statistics: (1) data are everywhere, (2) statistical techniques are used to make many decisions that affect our lives, and (3) no matter what your future line of work, you will make decisions that involve data. An understanding of statistical methods will help you make these decisions more effectively.

SELF-REVIEW 1-1

Halifax-based Market Facts asked a sample of 1960 consumers to try a newly developed frozen fish dinner by Morton Foods called Fish Delight. Of the 1960 sampled, 1176 said they would purchase the dinner if it was marketed.

(a) What would Market Facts report to Morton Foods regarding acceptance of Fish Delight in the population?

(b) Is this an example of descriptive statistics or inferential statistics? Explain.

1.4 TYPES OF VARIABLES

QUALITATIVE VARIABLES

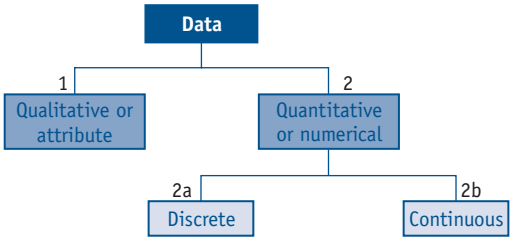
When the characteristic or variable being studied is non-numeric, it is called a *qualitative variable* or an *attribute*. Examples of qualitative variables are gender, religious affiliation, type of automobile owned, province of birth, and eye colour. When the data being studied are qualitative, we are usually interested in how many or what proportion falls into each category. For example, what percentage of the population has blue eyes? How many Catholics and how many Protestants are there in Canada? What percentage of the total number of cars sold last month were Buicks? Qualitative data are often summarized in charts and bar graphs (see Chapter 2).

QUANTITATIVE VARIABLES

When the variable studied can be reported numerically, the variable is called a *quantitative variable*. Examples of quantitative variables are the balance in your chequing account, the ages of company presidents, the life of a battery, the speeds of automobiles travelling along a highway, and the number of children in a family.

CHART 1-5: The Types of Variables

Examples:
Colours of pens in a drawer (1)
Distance between Winnipeg and Bangkok (2b)
Gender (1)
Kilometres driven between oil changes (2b)
Number of children (2a)
Number of employees (2a)
Number of TV sets sold last year (2a)
Type of car owned (1)
Weight of a shipment (2b)



Quantitative variables are either discrete or continuous. **Discrete variables** can assume only certain values, and there are usually "gaps" between the values. Examples of discrete variables are the number of bedrooms in a house (1, 2, 3, 4, etc.), the number of cars arriving at a Tim Hortons drive-through in downtown Fredericton during an hour (16, 19, 30, etc.), and the number of students in each section of a statistics course (25 in section A, 42 in section B, and 18 in section C). Notice that a home can have 3 or 4 bedrooms, but it cannot have 3.56 bedrooms. Thus, there is a "gap" between possible values. Typically, discrete variables result from counting. We count, for example, the number of cars arriving at the Tim Hortons, and we count the number of students in each section in a statistics course.

Observations on a **continuous variable** can assume any value within a specific range. Examples of continuous variables are the air pressure in a tire and the weight of a shipment of grain (which, depending on the accuracy of the scales, could be 15.0 tonnes, 15.01 tonnes, 15.03 tonnes, etc.). The amount of Raisin Bran in a box and the time it takes to fly from Fredericton to Toronto are other variables of a continuous nature. The Fredericton to Toronto flight could take 1 hour and 50 minutes; or 1 hour, 45 minutes, and 45 seconds; or 1 hour, 55 minutes, and 45.1 seconds, depending on the accuracy of the timing device. Typically, continuous variables result from measuring something.

For analytical purposes, it is often useful to remember a distinction between flow variables and stock variables. **Stock variables** refer to variables measured at a point in time. **Flow variables**, on the other hand, are variables measured over a specific period of time (a function of time). For example, the amount of water entering or leaving a reservoir per period (per minute, per day, per week, etc.) is a flow variable, whereas the amount of water in the reservoir at any particular point in time (December 31 at noon) is a stock variable. Thus, the flow variable represents a rate of change in stock over a period of time. Variables such as capital goods, wealth, or government debt are measured at a point in time (such as October 24, 2001), and are therefore called *stock variables*. Investment, savings, or government budget deficit are measured over a period of time (such as per month, per quarter, or per year) and are therefore called *flow variables*.

The types of variables are summarized in Chart 1-5 above.

1.5. LEVELS OF MEASUREMENT

Data can be classified according to levels of measurement. The level of measurement of the data often dictates the calculations that can be done to summarize and present the data. It will also determine the appropriate statistical tests to be performed.

For example, there are six colours of candies in a bag of M&M's. Suppose we assign the brown candy a value of 1, yellow 2, blue 3, orange 4, green 5, and red 6. From a bag of M&M's, we add the assigned colour values and divide by the number of candies and report that the mean colour is 3.56. Does this mean that the average colour is blue or orange? As a second example, in a high-school track meet there are 8 competitors in the 400-metre run. We report the order of finish and that the mean finish is 4.5. What does the mean finish tell us? In both of these instances, we have not properly used the level of measurement.

There are actually four levels of measurement: nominal, ordinal, interval, and ratio. The "lowest" or the weakest measurement is the nominal level. The highest, or the level that gives us the most information about the observation, is the ratio level of measurement.

NOMINAL LEVEL DATA

In the nominal level of measurement, the observations can be only classified or counted. There is no particular order to the labels. The classification of the six colours of M&M's is an example of the nominal level of measurement. We simply classify the candies by colour. There is no natural order. That is, we could report the brown candies first, the orange first, or any of the colours first. Gender is another example of the nominal level of measurement. Suppose we count the number of students entering a football game with a student ID and report how many are male and how many are female. Or, we could report how many are single or married or divorce or widowed. There is no order implied in the presentation. For the nominal level of measurement there is no measurement involved, only counts. Table 1-2 shows the breakdown of Canadians (15 years and over) by their marital status. This is the nominal level of measurement because we placed the number of people according to the category in which they belonged.

Note that the variable in this case is status, and not the numbers belonging to each status. Numbers are just counts of people falling under one of the status categories. These categories are **mutually exclusive**, meaning that a person cannot belong to more than one category. The categories in Table 1-2 are also **exhaustive**, meaning that every member of the population, or sample, must appear in one of the categories.

Mutually exclusive An individual, object, or measurement is included in only one category.

Exhaustive Each individual, object, or measurement must appear in one of the categories.

TABLE 1-2: Marital Status in Canada, 1999

(Population 15 years and over)

Status	Number	%
Single (never married)	7 114 681	29.0
Married*	14 535 881	59.2
Divorced	1 417 136	5.8
Widowed	1 506 231	6.1
Total	24 573 929	100.0

*Includes persons legally married and separated and persons living in common-law unions

Source: Statistics Canada: Canada at a Glance, 2nd edition

To process data on telephone usage, gender, and employment by industry, and so forth, the categories are often coded 1, 2, 3, and so on, with 1 representing single, 2 representing married, and so on. This facilitates counting by the computer. However, because we have assigned numbers to the various categories, this does not give us licence to manipulate the numbers. For example, 1 + 2 does not equal 3; that is, single plus married does not equal divorced. To summarize, the nominal level data have the following properties:

1. Data categories are mutually exclusive, so an object belongs to only one category.
2. Data categories are exhaustive, so that all observations are included in one of the categories.
3. Data categories have no logical order, nor can be they compared with each other.

In brief, nominal level data does not imply any order.

ORDINAL LEVEL DATA

Ordinal level data implies order. Each data point can be expressed in terms of the arithmetic operation > or <. In other words, each data point can be compared to other data points. Table 1-3 lists the student ratings of Professor James Brunner in an Introduction to Finance course. Each student in the class answered the question, "Overall, how did you rate the instructor in this class?" This illustrates the use of the ordinal scale of measurement. One category is "higher" or "better" than the next one. That is, "superior" is better than "good," "good" is better than "average," and so on. However, we are not able to distinguish anything about the magnitude of the differences between groups. Is the difference between superior and good the same as the difference between good and average? We cannot tell. If we substitute a 5 for superior and a 4 for good, we can conclude that the rating of superior is better than the rating of good, but we cannot add or subtract a ranking of superior and a ranking of good with the result being meaningful. Further, we cannot conclude that a rating of good (rating is 4) is necessarily twice as good as a "poor" (rating is 2). We can only conclude that a rating of good is better than a rating of poor. We cannot conclude how much better the rating is. Note that the variable is rating and not frequency (the number of students providing the rating).

In summary, the properties of ordinal level data are:

1. The data categories are mutually exclusive and exhaustive.
2. Data categories are ranked or ordered according to the particular trait they possess.
3. Only the rating values are comparable, but not the differences between the rating values.
4. We cannot do any arithmetic operation on the data other than set up inequalities. Arithmetic operations of subtraction and addition are not meaningful. However, it is not unusual to see such operations being done in practice! You are familiar with

TABLE 1-3: Rating of a Finance Professor

Rating	Frequency
5. Superior	6
4. Good	28
3. Average	25
2. Poor	12
1. Inferior	3

your grades being averaged. All it means is that users of information have decided to use the data as interval-level data.

INTERVAL-LEVEL DATA

The interval level of measurement is the next highest level. In addition to the setting up of inequalities, interval-level data allows arithmetic operations of subtraction and addition. It includes all the characteristics of the ordinal level, and in addition, the differences between values are now meaningful. Numerical differences of equal size between any two pairs of values represent equal changes in the attribute being measured. Examples of the interval level of measurement are temperature (Fahrenheit and Celsius), calendar time, and potential energy. Suppose temperatures on three consecutive winter days in Toronto are –5°C, –2°C, and 1°C. These temperatures can be easily ranked, but we can also determine the difference between temperatures. This is possible because 1°C represents a constant unit of measurement. Equal differences between two temperatures mean the same amount of change in the temperature (the attribute), regardless of their position on the scale. That is, the difference between 10°C and 15°C represents the same amount of change in the temperature as the difference between 20°C and 25°C. However, we cannot say that 20°C indicates twice the amount of heat compared to 10°C. This happens because both the Celsius scale and the Fahrenheit scale of measurement have artificial origins (0). You can convince yourself by converting these values to Fahrenheit (F). Since $F = 32 + 1.8C$; $10C = 50^F$; and $20^C = 68^F$. Obviously, 20/10 on the Celsius scale is not equal to 68/50 on the Fahrenheit scale. For another example of interval scale, see Table 1-1 on the cap Index of Provincial Well Being.

The properties of the interval scale are:

1. Data categories are mutually exclusive and exhaustive.
2. Equal differences in the characteristics are represented by equal differences in the numbers assigned to the categories.
3. Arithmetic operations of addition and subtraction are meaningful, but ratios of two values are *not* meaningful.

RATIO-LEVEL DATA

The ratio level is the "highest" level of measurement. The ratio level of measurement has all the characteristics of the interval level, but in addition, the 0 (zero) point is meaningful and the ratio between two numbers is meaningful. Examples of the ratio scale of measurement include wages, units of production, weight, height, area, pressure, density, and so on. Money is a good illustration. If you have zero dollars, then you have no money. Weight is another example. If the dial on the scale is at zero, then there is a complete absence of weight. The ratio of two numbers is now meaningful. If Jim earns $30 000 per year selling insurance and Rob earns $60 000 per year selling cars, then Rob earns twice as much as Jim.

The properties of the ratio level are that:

1. Data categories are mutually exclusive and exhaustive.
2. Data categories are scaled according to the amount of the characteristic they possess.
3. Equal differences in the characteristic are represented by equal differences in the numbers assigned to the categories.
4. The point 0 reflects the absence of the characteristic. All arithmetic operations are possible.

TABLE 1-4: The World's 10 Most Valuable Athletes (2000)

Rank	Name	Money ($US millions)	Performance Last Season	Sport
1	Tiger Woods	63.1	A+	Golf
2	Michael Schumacher	59	A+	Motor Sport
3	Shaquille O'Neal	24	A+	Basketball
4	Alex Rodriguez	35.2	B+	Baseball
5	Mike Tyson	48	B+	Boxing
6	Allen Iverson	14.3	A–	Basketball
7	Marion Jones	2.7	A	Track
8	Vince Carter	4.2	B	Basketball
9	David Beckham	10.6	B	Soccer
10	Ken Griffey Jr.	11.3	B	Baseball

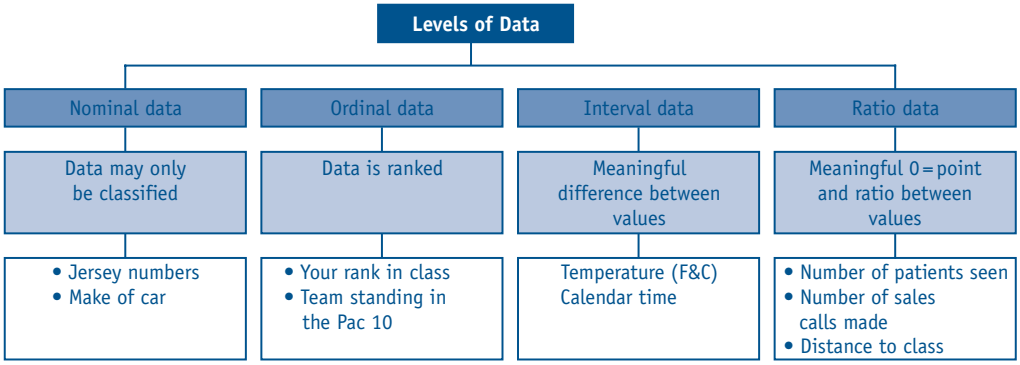
(Source: www.robmagazine.com, October 7, 2001.)

In Table 1-4, money earned by each athlete illustrates the use of the ratio scale of measurement. Unlike the interval scale, the ratio level allows us to carry out multiplication and division operations as well. Thus, we can say that Mike Tyson earned twice as much money as Shaquille O'Neal.

Chart 1-6 summarizes the characteristics of various levels of measurement through an organizational chart.

In brief, nominal scale is good for displaying data consisting of names, ordinal scale for order, interval scale for meaningful intervals, and ratio scale for meaningful ratios. A higher level of measurement scale possesses properties of all lower-level scales.

CHART 1-6: Characteristics of Levels of Measurement



SELF-REVIEW 1-2

Data relating to five students (out of 25) in Stats 2163 are given below. Identify the level of measurement in each case.

(a) Students' ID numbers are:
911992 912345 913465 915429 913978

(b) Students' ranks in their class are:
3 8 15 6 11

(c) Students' annual GPAs (for all courses taken that year) are:
3.9 3.6 2.5 3.5 3.0
(d) Amounts of student loans (in $) owed by the students are:
10 500 5450 12 200 0 8300

EXERCISES 1-1 TO 1-5

1-1. In Table 1-4, what is the level of measurement for
(a) Ranks for the athletes?
(b) Sports?
(c) Performance?

1-2. What is the level of measurement for each of the following variables?
(a) Student IQ ratings.
(b) Distance students travel from home (or residence) to school.
(c) Student scores on the first statistics test.
(d) A classification of students by province of birth.
(e) A ranking of students by first-year, second-year, third-year, and fourth-year.
(f) Number of hours students study per week.

1-3. What is the level of measurement for these items related to the newspaper business (*The Globe and Mail*)?
(a) The number of Saturday copies of *The Globe and Mail* sold last week.
(b) The number of employees in each department (editorial, advertising, sports, etc.).
(c) A summary of the number of papers sold by county.
(d) The number of years employed by the paper for each employee.

1-4. Look in the latest edition of *The Globe and Mail* or your local newspaper and find examples of each level of measurement. Write a brief memo summarizing your findings.

1-5. For each of the following, determine whether the group is a sample or a population.
(a) The participants in a study of a new diabetes drug.
(b) All the drivers who received a speeding ticket on Highway 401 last month.
(c) All families below the income of $20 000 per year in York-Sunbury county (Fredericton).
(d) Ten of the top 50 athletes in the world (by earnings per year).

1.6 SOURCES OF STATISTICAL DATA

Statistical data are available from two types of sources. Data collected by the investigator that are not available in the published form are called *primary data*. The data available in published form are called *secondary data*.

PUBLISHED DATA

Conducting research on problems involving such topics as crime, health, imports and exports, production, and hourly wages generally requires published data. We may want

information on the total number of housing starts in Canada in 2000 (151 700), the total value of retail trade in 2000 ($277 billion), the total number of persons in the Canadian labour force (15 and over) in a particular month, such as April 2001 (16 271 700), the unemployment rate in each of the Canadian provinces, the inflation rate this month, employment and average weekly earnings by industry, and so on.

Almost every country has a statistical agency responsible for collecting and publishing data on social, economic, business, and other aspects of life in the country. In Canada, our statistical agency is called Statistics Canada. Statistics Canada collects both micro-level (individual businesses and households, communities) data and macro-level (economy as a whole) data. Published data in hard copy are available in libraries. However, increasing amounts of data are now available in the form of electronic files on Statistics Canada's Web site (www.statcan.ca). Federal government and provincial government departments also collect data relevant to their objectives. These data can be accessed from their respective Web sites. Often these data are also available from Statistics Canada; the Statistics Canada site has links to most governmental Web sites as well as many international Web sites. In addition to Statistics Canada, the most important governmental sources for Canadian business, industry, trade, and financial data are:

- Industry Canada: www.strategis.gc.ca (for data on Canadian business, industry, and trade).
- Bank of Canada: www.bankofcanada.ca (for data on Canadian monetary conditions).
- Links to all provincial Web sites are available from www.gc.ca and www.statcan.ca
- Links for several international statistical Web sites are also available from www.statcan.ca

For US data, see www.census.gov. Summary data on other countries are available from:

- United Nations: www.un.org and its sub-sites.
- Organization for Economic Co-operation and Development: www.oecd.org
- International Monetary Fund: www.imf.org
- World Bank: www.worldbank.org

In addition to regularly published data, daily/weekly publications such as financial data of a daily, weekly, or monthly nature are available on www.globeandmail.com and its sub-sites, such as www.globefund.com, www.globeinvestor.com, and www.robmagazine.com, all available by links from www.globeandmail.com. Further, most chartered banks also have a wealth of financial data on their own Web sites. Other business-related Web sites are www.fortune.com, www.forbes.com, and www.economist.com. Sports-related data can be obtained from www.sportserver.com and www.canadiansport.com and from links available from these Web sites.

LEARNING RESOURCES AT STATISTICS CANADA (Σ-STAT)

Statistics Canada maintains a Web site called Σ-Stat that is freely available to all educational institutions in Canada (see Chart 1-7). There are more than 700 000 series of data on socio-economic conditions in Canada. We refer you to this Web site several times in this book. If your educational institution is not a registered user of the site, it can register and make the services available to all students. Alternatively, McGraw-Hill Ryerson, the publisher of this textbook, would provide an ID and PIN to access this Web site for all students who are using this textbook in their course.

CHART 1-7: The Statistics Canada Web Site

To collect data from this Web site, first go to the Statistics Canada Web site at www.statcan.ca, then to Learning Resources, then to Σ-Stat. At this point, you will be asked to accept a licence agreement. Say yes. You will then be presented with a small screen requesting your ID and password. Enter the ID and password provided to you by your educational institution or instructor. Select Enter. Now you have access to the wealth of information that Statistics Canada has to offer! You can access data either by series number, if you know it, or by finding data on the topic of your interest through a search engine. There is a user guide and an animated tutorial on the Web site. We strongly advise you to at least go through the animated tutorial to avoid frustrations.

UNPUBLISHED DATA

Like the sources for published data, there are sources for unpublished data collected and analyzed by the researchers working at universities and research institutes. Often such data are contained in working papers, discussion papers, and dissertations. You can access all Canadian universities and their business, economics, and statistics departments through the Association of Universities and Colleges of Canada Web site at www.aucc.ca. You can access a wide variety of resources on statistics including data, texts, animated lectures/tutorials on selected topics, universities around the world, and even jokes (http://noppa5.pc.helsinki.fi/links.html).

Published data are not always available on every subject of interest. Individuals may be contacted in a shopping mall, at their homes, over the telephone, or by mail. The respondents' answers are usually tabulated either by hand or using a computer. You have probably seen and completed many questionnaires. You can often see a poll conducted on *The Globe and Mail* Web site on a current topic of public interest. Perhaps you will be presented with such a questionnaire at the end of this course. Here are the results from a survey conducted by *Working Mother* magazine.

Working Mother commissioned Gallup to study how satisfied working mothers are with their dual role. Gallup polled 1000 working mothers nationwide. Some of the findings are listed below.

1. Seven out of 10 women said they work to feel good about themselves, regardless of the job they do or the amount of money they make.

2. Eight out of 10 working mothers were "extremely satisfied" or "very satisfied" with the job they are doing as mothers.
3. Ninety percent said their children are happy.
4. Three-quarters said they "like" or "love" their jobs.
5. Four percent said they "hate" their work.

1.7 USES AND ABUSES OF STATISTICS

LIES, DAMN LIES, AND STATISTICS

"It ain't so much the things we don't know that get us into trouble. It's the things we know that just ain't so."—Artemus Ward

You have probably heard the old saying that there are three kinds of lies: lies, damn lies, and statistics. This saying is attributable to Benjamin Disraeli, a British Prime Minister, nearly a century ago. It has also been said that "Charts don't lie: liars chart." Both of these statements refer to the abuses of statistics in which data are presented in ways that are misleading. Many abusers of statistics are simply ignorant or careless, while others have an objective to mislead the reader by emphasizing data that support their position while leaving out data that may be detrimental to their position. *One of our major goals in this text is to make you a critical consumer of information.* When you see charts or data in a newspaper, in a magazine, or on TV, always ask yourself these questions: What is the person trying to tell me? Does that person have an agenda? The following are several examples of the abuses of statistical analysis.

AN AVERAGE MAY NOT BE REPRESENTATIVE OF ALL THE DATA

The term *average* refers to several different measures of central tendency that we discuss in Chapter 3. To most people, an average is found by adding the values in the data set and then dividing the total by the number of values in the data. So if a real estate developer tells a client that the average home in a particular subdivision sold for $150 000, we assume that $150 000 is a representative selling price for all the homes. But suppose there are only five homes in the subdivision and they sold for $50 000, $50 000, $60 000, $90 000, and $500 000. We can correctly claim that the average selling price is $150 000, but does $150 000 really seem like a "typical" selling price? Would you like to also know that the same number of homes sold for more than $150 000 as sold for less than $150 000? Or that $150 000 is the selling price that occurred most frequently? So what selling price really is the most "typical"? This example illustrates how a reported average can be misleading. We will discuss averages, or measures of central tendency, in Chapter 3.

GRAPHS CAN BE MISLEADING

Pictographs are often used as a visual aid for an easy interpretation. However, if they are not drawn carefully, they can lead to misinterpretation of information. Suppose the cost of heating a typical home in Toronto increased from $100 a month to $200 a month over the past 20 years; that is, the heating cost per month doubled. To show this change, the dollar sign on the right in Chart 1-8a is twice as tall as the one on the left. It is also twice as wide, making the area of the dollar sign on the right four times (not twice) that of the one on the left. When we double the dimensions of a two-dimensional

CHART 1-8a: Cost of Heating in Toronto

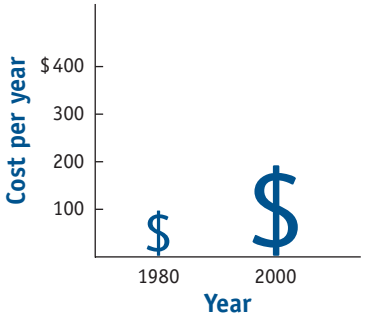
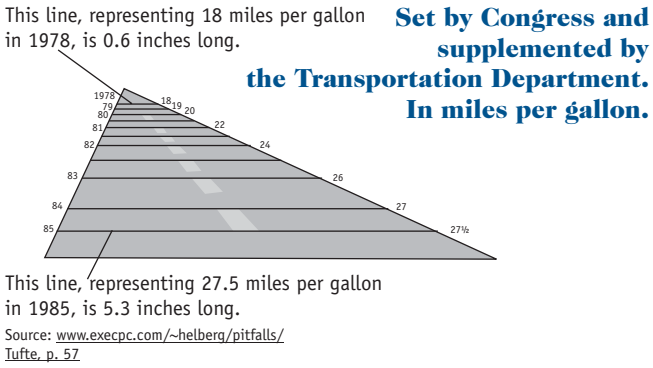


CHART 1-8b: Full Economy Standards for Autos



Source: www.execpc.com/~helberg/pitfalls/ Tufte, p. 57

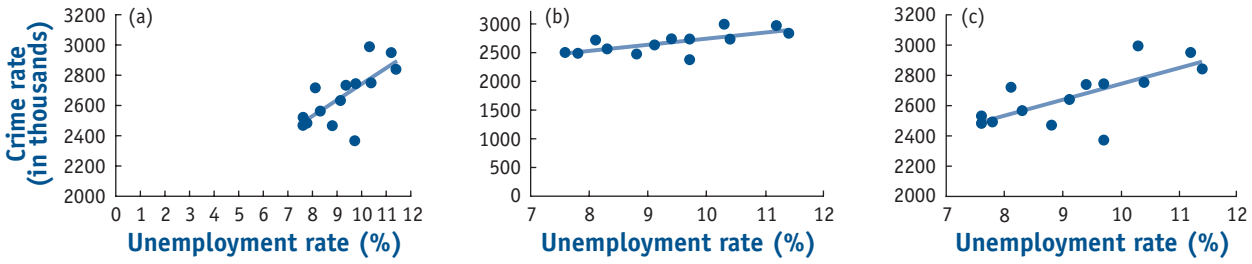
object, we increase the area by a factor of four. The chart is misleading because visually the increase seems much larger than it really is. In Chart 1-8b, the line representing 27.5 miles per gallon in 1985 was 5.3 inches long and the line for 18 miles per gallon only 0.6 inches long in the original presentation.

Edward R. Tufte, in his book *The Visual Display of Qualitative Information* (Cheshire, CT: Graphics Press, 1983) has given many examples of how to recognize misleading graphs and how to construct a good graph. In his book, Tufte introduces a concept called the "**lie factor**". It can be defined as the percentage change in the graphic elements divided by the percentage change in the actual quantities represented by those graphic elements. By this definition, the value of the lie factor should equal 1 for an accurate and informative graph. In Chart 1.8b, the lie factor can be calculated as $\frac{(5.3-0.6)/0.6}{(27.5-18)/18} = 14.8$. Thus, the lie factor is 14.8!

Graphs and charts of data, such as histograms, line charts, and bar charts, can also be misleading if they are not drawn appropriately. We cover these graphs and charts in detail in the next chapter. A misleading visual interpretation in the context of charts arises often due to a presentation of only part of the data, or using the horizontal and/or vertical axis inappropriately.

Charts 1-9a and 1-9b are designed to show a relationship between unemployment rate (in percent) and crime rate (in thousands, per year) in Canada based on the same data for the years 1986 to 1999. In Chart 1-9, we have broken the vertical axis at 2000, and thus show a strong relation between unemployment rate and crime. In Chart 1-9b, we have broken the horizontal axis at a 7-percent rate of unemployment. In this

CHART 1-9: Unemployment Rate and Crime Rate in Canada, 1986–99



graph, we get an impression of a weaker relation between unemployment rate and crime. A more accurate depiction of the relationship can be obtained by using values near the minimum values of the variables as starting points on each axis. Thus, a break on the vertical axis at 2000 and on the horizontal axis at 7 percent will give you a more accurate picture of the relationship. See Chart 1-9c.

There are many graphing techniques, but there are no hard and fast rules about drawing a graph. It is therefore both a science and an art. Your aim should always be a truthful representation of the data. The objectives and the assumptions underlying the data must be kept in mind and mentioned briefly along with graphs. The visual impressions conveyed by the graphs must correspond to the underlying data. The graphs should reveal as much information as possible with the greatest precision and accuracy. *Graphical excellence is achieved when a viewer can get the most accurate and comprehensive picture of the situation underlying the data set in the shortest possible time.* In brief, a graph should act like a mirror between the numerical data and the viewer. According to a popular saying, "*Numbers speak for themselves.*" This is true for small data sets. For large data sets, it may be difficult to discern any patterns by looking at numbers alone. We therefore *need accurate portrayal of data through graphs that can speak for numbers*, and can give a quick overview of the data. We discuss graphic techniques in detail in the next chapter.

STUDIES BASED ON INADEQUATE SURVEYS CAN BE MISLEADING

Several years ago, a series of TV advertisements reported that "2 out of 3 dentists surveyed indicated they would recommend Brand X toothpaste to their patients." The implication is that 67 percent of all dentists would recommend the product to their patients. What if they surveyed only three dentists? It would certainly not be a true representation of the real situation. The trick is that the manufacturer of the toothpaste could take *many* surveys of three dentists and report *only* the surveys of three dentists in which two dentists indicated they would recommend Brand X. This is concealing the information to mislead the public. Further, a survey of more than three dentists is needed, and it must be unbiased and representative of the population of all dentists. We discuss sampling methods in Chapter 8.

Example 1-2

Professor Best's Encounter with Bad Statistics[5]

A graduate student started his dissertation with a quote (possibly to impress his dissertation committee): "Every year since 1950, the number of American children gunned down has doubled." When Professor Best, a member of the student's dissertation committee, read the quotation, he did not believe it. He went to the library and looked up the article the student had cited. In the journal's 1995 volume, he found exactly the same sentence.

"What makes this statistic so bad?" asks Professor Best. "Just for the sake of argument, let's assume that the number of American children gunned down in 1950 was one. If the number doubled each year, there must have been two children gunned down in 1951, four in 1952, eight in 1953, and so on. By 1970, the number would have passed one million; by 1980, one billion (more than four times the total US population in that year). By 1995, when the article was published, the annual number of victims would have been over 35 trillion—a really big number…." Professor Best asked the article's author about the source of this statistic. The author's response was that he had seen the statistic in a document published by the Children's Defense Fund, a well-known advocacy group for children. The CDF's *The State of America's Children Yearbook, 1994* does state that: "The number of American children killed each year by guns has

doubled since 1950." Note the difference in wording—the CDF claimed there were twice as many deaths in 1994 as in 1950; the article's author reworded that claim and created a very different meaning.

It is worth examining the history of this statistic. It began with the CDF noting that child gunshot deaths had doubled from 1950 to 1994. This is not quite as dramatic an increase as it might seem. Remember that the US population also rose throughout this period; in fact, it grew about 73 percent, or nearly double. Therefore, we might expect all sorts of things—including the number of child gunshot deaths to increase—to nearly double, just because the population grew. Before we can decide whether twice as many deaths indicate that things are getting worse, we'd have to know more. The CDF statistic raises other issues as well: Where did the statistics come from? Who counts child gunshot deaths, and how? What is meant by a "child"? (Some CDF statistics about violence include everyone under age 25.) What is meant by "killed by guns"? (Gunshot-death statistics often include suicides and accidents, as well as homicides.) But people rarely ask questions of this sort when they encounter statistics. Most of the time, most people simply accept statistics without question.

Certainly, the article's author didn't ask many probing, critical questions about the CDF's claim. Impressed by the statistic, the author repeated it—well, meant to repeat it. Instead, by rewording the CDF's claim, the author created a mutant statistic, one garbled almost beyond recognition.

ASSOCIATION DOES NOT NECESSARILY IMPLY CAUSATION

Another area where there can be a misrepresentation of data is the association between variables. In statistical analysis often we find there is a strong *association* between variables. We find there is a strong negative association between outside work and grade point average. The more outside work a student is engaged in, the lower will be his or her grade point average. Does it mean that more outside work causes a lower grade point average? Not necessarily. It is also possible that the lower grade point average does not make the student eligible for a scholarship and therefore the student is required to engage in outside work to finance his or her education. Alternatively, both outside work and lower GPA could be a result of the social circumstances of the student. Unless we have used an experimental design that has successfully controlled the influence of all other factors on grade point average except the outside work, or vice versa, we are not justified in establishing any causation between variables based on statistical evidence alone. In general, *association based on observational (non-experimental) data is neutral with regard to causation*. We study the association between variables in Chapters 13 and 14.

BECOME A BETTER CONSUMER AND A BETTER PRODUCER OF INFORMATION

There are many other ways that statistical information can be deceiving. It may be because (1) The data are not representative of the population; (2) Appropriate statistics have not been used; (3) The data do not satisfy the assumptions required for inferences; (4) The prediction is too far out from the range of observed data; (5) Policy analysis does not meet the requirements of either data or theory or both; (6) Ignorance and/or carelessness on the part of the investigator; (7) A deliberate attempt to introduce bias has been made to mislead the consumer of information. Entire books have been written about the subject. The most famous of these is *How to Lie with Statistics* by Darrell Huff. Understanding the art and science of statistics will make you both a better consumer of information as well as a better producer of information (statistician). This is our aim in writing this book.

1.8 COMPUTER APPLICATIONS

Computers are now available for student use at most colleges and universities. Spreadsheets, such as Microsoft Excel, which have many statistical functions, are also available in most computer labs and on most home computers. We have selected Excel and Minitab for most of the statistical applications in the text. We also use an Excel add-in called MegaStat, software bundled free on the CD-ROM included with this book. This add-in gives Excel the capability to produce additional statistical results. With statistical software such as Excel and Minitab, we can get most of the descriptive and inferential statistics at the click of a few buttons! The statistical software can save an enormous amount of time required for computations, and thereby enable us to devote more of our valuable time to analytical aspects.

CHART 1-10: Excel at Work

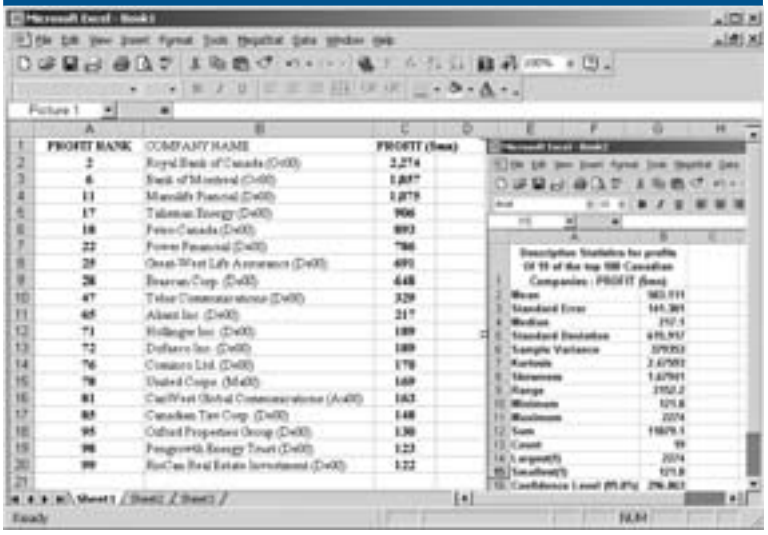
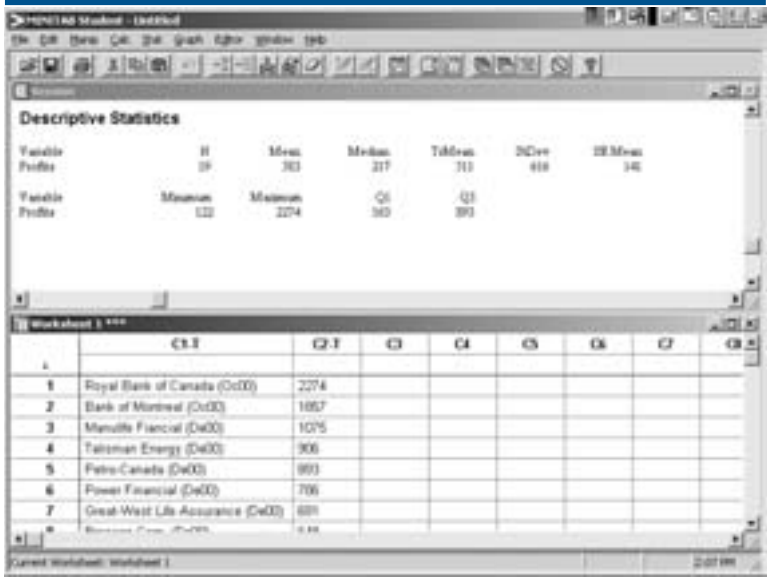


CHART 1-11: Minitab at Work



The following example shows the wide application of computers in statistical analysis. In Chapters 2, 3, and 4 we illustrate methods for summarizing and describing data. Chart 1-10 shows a sample of 19 of the top 100 Canadian companies computing most of the descriptive statistics for profits (in 2000). The Excel output reveals, among other things, that the mean (average) profit was $583 million, and ranged between $122 million and $2274 million. Many other descriptive statistics of data are also shown.

The Minitab output in Chart 1-11 contains much of the same information, although it is arranged somewhat differently. Minitab is user-friendly statistics software. It is versatile software for most of the needs of statistical analysis. If your institution does not have this software on its network, you may purchase or rent the software or download free for a month (on a trial basis) from Minitab's Web site (www.minitab.com). An introductory guide to Minitab, called *Meet Minitab*, is also available (free) as a PDF file that you can download and print. This guide, together with the Minitab Help Menu, will help you in using the software.

CHAPTER OUTLINE

I. Statistics is the art and science of collecting, organizing, analyzing, making inferences from the part about the whole, and prediction and policy analysis.

II. There are two types of statistics:

A. *Descriptive statistics* are procedures used to organize and summarize data.

B. *Inferential statistics* involve taking a sample from a population and making inferences about a population based on the sample results.

C. Statistical methods can also be used for prediction and policy analysis.

1. A *population* is the total collection of individuals or objects. A summary measure that describes some characteristic of a population is called a *parameter*.

2. A *sample* is a part of the population. A summary measure used to estimate a given characteristic of a population from a sample is called a *statistic*.

III. Statistics plays a significant role in the development of knowledge as well as in everyday life.

IV. There are two types of variables:

A. A *qualitative variable* is non-numeric.

1. Usually we are interested in the number or percentage of the observations in each category.

2. Qualitative data are usually summarized in graphs and bar charts.

B. A *quantitative variable* is numeric. There are two types of quantitative variables and they are usually reported numerically.

1. Discrete variables can assume only certain values, and there are usually gaps between values.

2. A continuous variable can assume any value within a specified range.

V. There are four levels of measurement:

A. With the *nominal* level, the data are sorted into categories with no particular order to the categories.

B. The *ordinal* level of measurement presumes that one category is ranked higher than another.

C. The *interval* level of measurement has the ranking characteristic of the ordinal level of measurement plus the characteristic that the distance between values is meaningful.

D. The *ratio* level of measurement has all the characteristics of the interval level as well as a zero point. The ratio of two values is meaningful.

VI. Statistical data can be collected either through surveys or from published sources. Most published data are available in electronic form.

VII. Statistical methods, if not used appropriately, can result in grossly misleading information.

CHAPTER EXERCISES 1-6 TO 1-18

1-6. Explain the difference between qualitative and quantitative data. Give an example of qualitative and quantitative data.

1-7. List the four levels of measurement and give an example (different from those used in the book) of each level of measurement.

1-8. Explain the difference between a sample and a population, and a statistic and a parameter.

1-9. (a) Define the lie factor in graphical presentations. Find the lie factor in Chart 1-8a. Assume the smaller dollar figure is 1 cm wide and 2 cm tall, and the larger dollar figure is 2 cm wide and 4 cm tall.
(b) Explain the possible reasons for misuses of statistics.

1-10. Using data from Statistics Canada and such publications as *The Economist*, *Newsweek*, *The Globe and Mail*, or your local newspaper, give examples of the nominal, ordinal, interval, and ratio level of measurement.

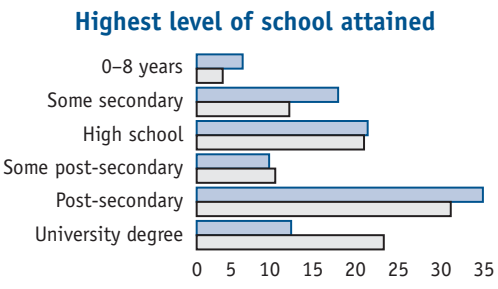
1-11. A random sample of 300 executives out of 2500 employed by a large firm showed that 270 would move to another location if it meant a substantial promotion. Based on these findings, write a brief note to management regarding all executives in the firm.

1-12. A random sample of 500 customers was asked to test a new toothpaste. Of the 500 customers, 400 said it was excellent, 32 thought it was fair, and the remaining customers had no opinion. Based on these sample findings, make an inference about the reaction of all customers to the new toothpaste.

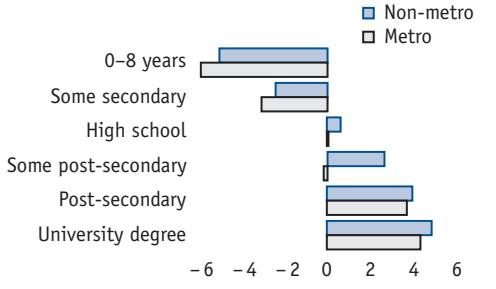
1-13. What is the measurement scale in the following graph? Are scales the same or different in the right-hand and left-hand graphs? Write a brief report on the information contained in the graph.

Both metro and non-metro employment is becoming more knowledge-intensive ...

Highest educational attainment of employed population (15 years+), 1999



Employment growth by educational attainment, 1990–1999



Source: IC calculations based on Statistics Canada data (LFS)

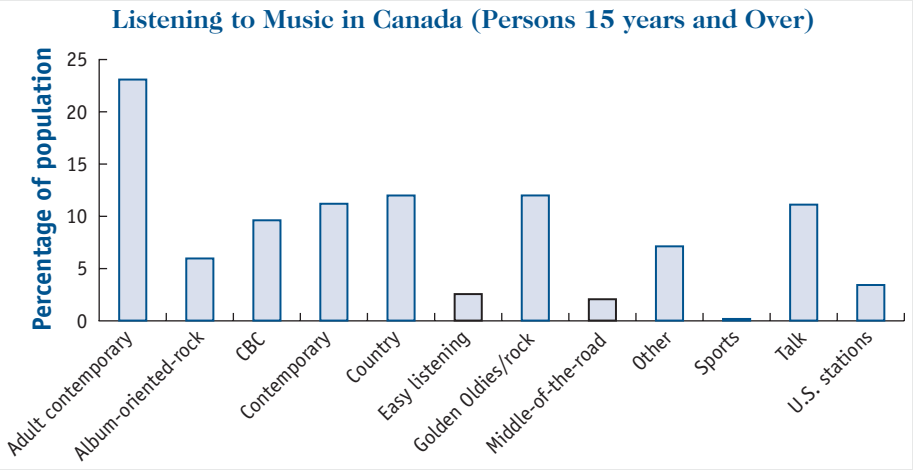
Employment growth among the highly educated has been somewhat faster in non-metro areas than in metro areas.

Still, non-metro Canada has a much smaller share of workers with a university degree and a larger share of workers with less than high school graduation.

– Job growth trends over the 1990–99 period demonstrate the influence of education on employment. Employment growth was faster for those with university and post-secondary education.

Source: MEPA, Industry Canada

1-14. What is the level of measurement in the following graph? Write a brief note on the information contained in the graph. (Source: Statistics Canada, Adapted.)



COMPUTER DATA EXERCISES 1-15 TO 1-18

1-15. Refer to the data set on Real Estate on the CD-ROM and the text Web site, which reports information on homes sold in Victoria, B.C., last year.
(a) Which of the variables are qualitative and which are quantitative?
(b) Determine the level of measurement for each of the variables.

1-16. Refer to the data set on the Major League Baseball Teams on the CD-ROM and the text Web site. Consider the following variables: team salary, attendance, and number of errors.
(a) Which of the variables are qualitative and which are quantitative?
(b) Determine the level of measurement for each of the variables.

1-17. Refer to the data set on the Top 1000 Canadian Corporations on the CD-ROM and the text Web site. Answer the following questions:
(a) Which of the variables are qualitative and which are quantitative?
(b) Determine the level of measurement for each of the variables.

1-18. Refer to the data set on Youth Unemployment in Canada on the CD-ROM and the text Web site.
(a) Which of the variables are qualitative and which are quantitative?
(b) Determine the level of measurement for each of the variables.

INTERNET EXERCISES 1-19 TO 1-21

In these exercises, you are required to go on the Web sites indicated, collect data as required, and write a brief report including the measurement scale of variables, type of statistic, and so on.

1-19. Go to www.statcan.ca and click on Community Profiles (located on the top of the screen). Collect data in your community on the following variables and write a brief report: population (male and female), immigrant population, persons who have completed university degrees, average income of males and females, proportion of married and common-law families, average value of owner-occupied dwellings, and ratio of male to female birth rate and death rate.

1-20. On the same Web site as in Exercise 1-19, go to the Daily and read about the labour force characteristics in your province. Write a brief report.

1-21. Go to the Web site www.globefund.com. Choose the stocks of five companies and look at the variation in prices over the last 30 days. Write a brief report.

CHAPTER 1 ANSWERS TO SELF-REVIEW

1-1. (a) Based on the sample of 1960 consumers, we estimate that, if it is marketed, 60 percent of all consumers will purchase Fish Delight (1176/1960) × 100 = 60 percent.

(b) Inferential statistics, because a sample was used to make an inference about how all consumers in the population would react if Fish Delight were marketed.

1-2. (a) Nominal

(b) Ordinal

(c) Ordinal but often used as interval

(d) Ratio