1 Data and Statistics

CHAPTER CONTENTS

Statistics in Practice The Economist

LEARNING OBJECTIVES

After reading this chapter and doing the exercises, you should be able to:

1. Appreciate the breadth of statistical applications in business and economics.
2. Understand the meaning of the terms elements, variables and observations, as they are used in statistics.
3. Understand the difference between qualitative, quantitative, cross-sectional and time series data.
4. Find out about data sources available for statistical analysis both internal and external to the firm.
5. Appreciate how errors can arise in data.
6. Understand the meaning of descriptive statistics and statistical inference.
7. Distinguish between a population and a sample.
8. Understand the role a sample plays in making statistical inferences about the population.

Frequently, we see the following kinds of statements in newspaper and magazine articles:

- The Ifo World Economic Climate Index fell again substantially in January 2009. The climate indicator stands at 50.1 (1995 = 100); its historically lowest level since introduction in the early 1980s (CESifo, April 2009).
- The IMF projected the global economy would shrink 1.3 per cent in 2009 (*Fin24*, 23 April 2009).
- The Footsie finished the week on a winning streak despite shock figures that showed the economy has contracted by almost 2 per cent already in 2009 (*This is Money*, 25 April 2009).
- China's growth rate fell to 6.1 per cent in the year to the first quarter (*The Economist*, 16 April 2009).

- GM receives further $2bn in loans (*BBC News*, 24 April 2009).
- Handset shipments to drop by 20 per cent (*In-Stat*, 2009).

The numerical facts in the preceding statements (50.1, 1.3 per cent, 2 per cent, 6.1 per cent, $2bn, 20 per cent) are called statistics. Thus, in everyday usage, the term *statistics* refers to numerical facts. However, the field, or subject, of statistics involves much more than numerical facts. In a broad sense, **statistics** is the art and science of collecting, analyzing, presenting and interpreting data. Particularly in business and economics, the information provided by collecting, analyzing, presenting and interpreting data gives managers and decision-makers a better understanding of the business and economic environment and thus enables them to make more informed and better decisions. In this text, we emphasize the use of statistics for business and economic decision-making.

Chapter 1 begins with some illustrations of the applications of statistics in business and economics. In Section 1.2 we define the term *data* and introduce the concept of a data set. This section also introduces key terms such as *variables* and *observations*, discusses the difference between quantitative and categorical data, and illustrates the uses of cross-sectional and time series data. Section 1.3 discusses how data can be obtained from existing sources or through survey and experimental studies designed to obtain new data. The important role that the Internet now plays in obtaining data is also highlighted. The use of data in developing descriptive statistics and in making statistical inferences is described in Sections 1.4 and 1.5. The last two sections of Chapter 1 outline respectively the role of computers in statistical analysis and introduce the relatively new field of data mining.

STATISTICS IN PRACTICE
The Economist

Founded in 1843, *The Economist* is an international weekly news and business magazine written for top-level business executives and political decision-makers. The publication aims to provide readers with in-depth analyses of international politics, business news and trends, global economics and culture.

The Economist is published by the Economist Group – an international company employing nearly 1000 staff worldwide – with offices in London, Frankfurt, Paris and Vienna; in New York, Boston and Washington, DC; and in Hong Kong, mainland China, Singapore and Tokyo.

Between 1998 and 2008 the magazine's worldwide circulation grew by 100 per cent – recently exceeding 180 000 in the UK, 230 000 in continental Europe, 780 000 plus copies in North America and nearly 130 000 in the Asia-Pacific region. It is read in more than 200 countries and with a readership of four million, is one of the world's most influential business publications. Along with the *Financial Times*, it is arguably one of the two most successful print publications to be introduced in the US market during the past decade.

Complementing *The Economist* brand within the Economist Brand family, the Economist Intelligence Unit provides access to a comprehensive database of worldwide indicators and forecasts covering more than 200 countries, 45 regions and eight key industries. The Economist Intelligence Unit aims to help executives make informed business decisions through dependable intelligence delivered online, in print, in customized research as well as through conferences and peer interchange.

Alongside the Economist Brand family, the Group manages and runs the CFO and Government brand families for the benefit of senior finance executives and government decision-makers (in Brussels and Washington respectively).

1.1 APPLICATIONS IN BUSINESS AND ECONOMICS

In today's global business and economic environment, anyone can access vast amounts of statistical information. The most successful managers and decision-makers understand the information and know how to use it effectively. In this section, we provide examples that illustrate some of the uses of statistics in business and economics.

Accounting

Public accounting firms use statistical sampling procedures when conducting audits for their clients. For instance, suppose an accounting firm wants to determine whether the amount of accounts receivable shown on a client's balance sheet fairly represents the actual amount of accounts receivable. Usually the large number of individual accounts receivable makes reviewing and validating every account too time-consuming and expensive. As common practice in such situations, the audit staff selects a subset of the accounts called a sample. After reviewing the accuracy of the sampled accounts, the auditors draw a conclusion as to whether the accounts receivable amount shown on the client's balance sheet is acceptable.

Finance

Financial analysts use a variety of statistical information to guide their investment recommendations. In the case of stocks, the analysts review a variety of financial data including price/earnings ratios and dividend yields. By comparing the information for an individual stock with information about the stock market averages, a financial analyst can begin to draw a conclusion as to whether an individual stock is over- or under-priced. Similarly, historical trends in stock prices can provide a helpful indication on when investors might consider entering (or re-entering) the market. For example, *Money Week* (3 April 2009) reported a Goldman Sachs analysis that indicated, because stocks were unusually cheap at the time, real average returns of up to 6 per cent in the US and 7 per cent in Britain might be possible over the next decade – based on long-term cyclically adjusted price/earnings ratios.

Marketing

Electronic scanners at retail checkout counters collect data for a variety of marketing research applications. For example, data suppliers such as ACNielsen purchase point-of-sale scanner data from grocery stores, process the data and then sell statistical summaries of the data to manufacturers. Manufacturers spend vast amounts per product category to obtain this type of scanner data. Manufacturers also purchase data and statistical summaries on promotional activities such as special pricing and the use of in-store displays. Brand managers can review the scanner statistics and the promotional activity statistics to gain a better understanding of the relationship between promotional activities and sales. Such analyses often prove helpful in establishing future marketing strategies for the various products.

Production

Today's emphasis on quality makes quality control an important application of statistics in production. A variety of statistical quality control charts are used to monitor the output of a production process. In particular, an x-bar chart can be used to monitor the average output. Suppose, for example, that a machine fills containers with 330g of a soft drink. Periodically, a production worker selects a sample of containers and computes the average number of grams in the sample. This average, or x-bar value, is plotted on an x-bar chart. A plotted value above the chart's upper control limit indicates overfilling, and a plotted value below the chart's lower control limit indicates underfilling. The process is termed 'in control' and allowed to continue as long as the plotted x-bar values fall between the chart's upper and lower control limits. Properly interpreted, an x-bar chart can help determine when adjustments are necessary to correct a production process.

Economics

Economists frequently provide forecasts about the future of the economy or some aspect of it. They use a variety of statistical information in making such forecasts. For instance, in forecasting inflation rates, economists use statistical information on such indicators as the Producer Price Index, the unemployment rate and manufacturing capacity utilization. Often these statistical indicators are entered into computerized forecasting models that predict inflation rates.

Applications of statistics such as those described in this section are an integral part of this text. Such examples provide an overview of the breadth of statistical applications. To supplement these examples, chapter-opening Statistics in Practice articles obtained from a variety of topical sources are used to introduce the material covered in each chapter. These articles show the importance of statistics in a wide variety of business and economic situations.

1.2 DATA

Data are the facts and figures collected, analyzed and summarized for presentation and interpretation. All the data collected in a particular study are referred to as the **data set** for the study. Table 1.1 shows a data set summarizing information for equity (share) trading at the 22 European Stock Exchanges in March 2009.

TABLE 1.1 European stock exchange monthly statistics domestic equity trading (electronic order book transactions) March 2009

	Total	
Exchange	**Trades**	**Turnover**
Athens	599 192	2 009.8
Borsa Italiana	5 921 099	44 385.9
Bratislava	111	0.1
Bucharest	79 921	45.3
Budapest	298 871	1 089.6
Bulgarian	14 040	64.4
Cyprus	31 167	76.1
Deutsche Börse	7 642 241	86 994.5
Euronext	15 282 996	116 488
Irish	79 973	549.8
Ljublijana	11 172	35.6
London	16 539 588	114 283.6
Luxembourg	1 152	125
Malta	638	1.9
NASDAQ OMX Nordic	4 550 073	40 927.4
Oslo Bars	981 362	9 755.1
Prague	65 153	1 034.8
SIX Swiss	440 578	2 667.1
Spanish (BME)	2 799 329	60 387.6
SWX Europe	n/a	n/a
Warsaw	1 155 379	2 468.6
Wiener Borse	433 545	2 744
TOTAL	**56 927 580**	**486 021.7**

Source: European Stock Exchange monthly statistics (www.fese.be/en/?inc=art&id=3)

EXCHANGES 2009

Elements, variables and observations

Elements are the entities on which data are collected. For the data set in Table 1.1, each individual European exchange is an element; the element names appear in the first column. With 22 exchanges, the data set contains 22 elements.

A **variable** is a characteristic of interest for the elements. The data set in Table 1.1 includes the following three variables:

- *Exchange*: at which the equities were traded.
- *Trades*: number of trades during the month.
- *Turnover*: value of trades (€m) during the month.

Measurements collected on each variable for every element in a study provide the data. The set of measurements obtained for a particular element is called an **observation**. Referring to Table 1.1, we see that the set of measurements for the first observation (Athens Exchange) is 599 192 and 2009.8. The set of measurements for the second observation (Borsa Italiana) is 5 921 099 and 44 385.9; and so on. A data set with 22 elements contains 22 observations.

Scales of measurement

Data collection requires one of the following scales of measurement: nominal, ordinal, interval or ratio. The scale of measurement determines the amount of information contained in the data and indicates the most appropriate data summarization and statistical analyses.

When the data for a variable consist of labels or names used to identify an attribute of the element, the scale of measurement is considered a **nominal scale**. For example, referring to the data in Table 1.1, we see that the scale of measurement for the exchange variable is nominal because Athens Exchange, Borsa Italiana … Wiener Börse are labels used to identify where the equities are traded. In cases where the scale of measurement is nominal, a numeric code as well as non-numeric labels may be used. For example, to facilitate data collection and to prepare the data for entry into a computer database, we might use a numeric code by letting 1, denote the Athens Exchange, 2, the Borsa Italiana … and 22, Wiener Börse. In this case the numeric values 1, 2, … 22 provide the labels used to identify where the stock is traded. The scale of measurement is nominal even though the data appear as numeric values.

The scale of measurement for a variable is called an **ordinal scale** if the data exhibit the properties of nominal data and the order or rank of the data is meaningful. For example, Eastside Automotive sends customers a questionnaire designed to obtain data on the quality of its automotive repair service. Each customer provides a repair service rating of excellent, good or poor. Because the data obtained are the labels – excellent, good or poor – the data have the properties of nominal data. In addition, the data can be ranked, or ordered, with respect to the service quality. Data recorded as excellent indicate the best service, followed by good and then poor. Thus, the scale of measurement is ordinal. Note that the ordinal data can also be recorded using a numeric code. For example, we could use 1 for excellent, 2 for good and 3 for poor to maintain the properties of ordinal data. Thus, data for an ordinal scale may be either non-numeric or numeric.

The scale of measurement for a variable becomes an **interval scale** if the data show the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Interval data are always numeric. Graduate Management Admission Test (GMAT) scores are an example of interval-scaled data. For example, three students with GMAT scores of 620 550 and 470 can be ranked or ordered in terms of best performance to poorest performance. In addition, the differences between the scores are meaningful. For instance, student one scored 620 – 550 = 70 points more than student two, while student two scored 550 – 470 = 80 points more than student three.

The scale of measurement for a variable is a **ratio scale** if the data have all the properties of interval data and the ratio of two values is meaningful. Variables such as distance, height, weight and time use the ratio scale of measurement. This scale requires that a zero value be included to indicate that nothing exists for the variable at the zero point. For example, consider the cost of a car. A zero value for the cost would

indicate that the car has no cost and is free. In addition, if we compare the cost of €30 000 for one car to the cost of €15 000 for a second car, the ratio property shows that the first car is €30 000/€15 000 = two times, or twice, the cost of the second car.

Categorical and quantitative data

Data can be further classified as either categorical or quantitative. **Categorical data** include labels or names used to identify an attribute of each element. Categorical data use either the nominal or ordinal scale of measurement and may be non-numeric or numeric. **Quantitative data** require numeric values that indicate how much or how many. Quantitative data are obtained using either the interval or ratio scale of measurement.

A **categorical variable** is a variable with categorical data, and a **quantitative variable** is a variable with quantitative data. The statistical analysis appropriate for a particular variable depends upon whether the variable is categorical or quantitative. If the variable is categorical, the statistical analysis is rather limited. We can summarize categorical data by counting the number of observations in each category or by computing the proportion of the observations in each category. However, even when the categorical data use a numeric code, arithmetic operations such as addition, subtraction, multiplication and division do not provide meaningful results. Section 2.1 discusses ways for summarizing categorical data.

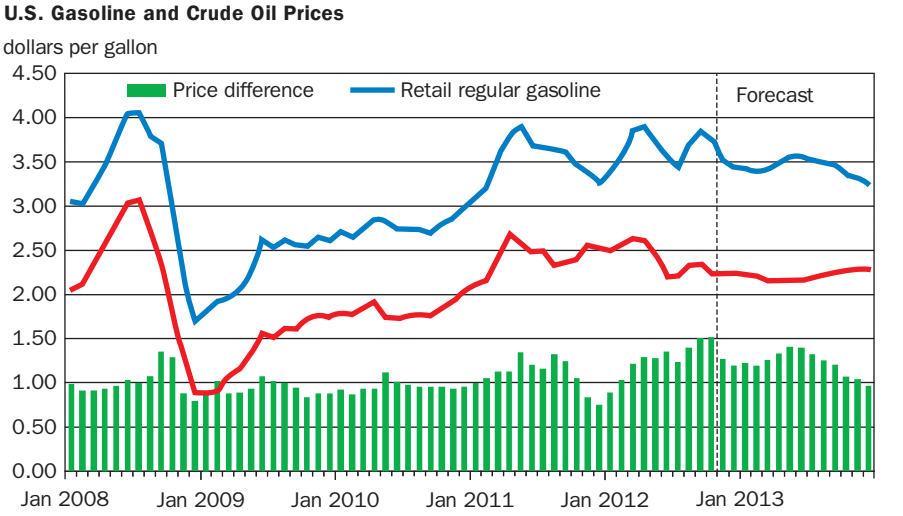
On the other hand, arithmetic operations often provide meaningful results for a quantitative variable. For example, for a quantitative variable, the data may be added and then divided by the number of observations to compute the average value. This average is usually meaningful and easily interpreted. In general, more alternatives for statistical analysis are possible when the data are quantitative. Section 2.2 and Chapter 3 provide ways of summarizing quantitative data.

Cross-sectional and time series data

For purposes of statistical analysis, distinguishing between cross-sectional data and time series data is important. **Cross-sectional data** are data collected at the same or approximately the same point in time. The data in Table 1.1 are cross-sectional because they describe the two variables for the 22 exchanges at the same point in time. **Time series data** are data collected over several time periods. For example, Figure 1.1 provides a graph of the wholesale price (US$) of crude oil per gallon for the period January 2008 and January 2012. It shows that starting around July 2008 the average price dipped sharply to less than $2 per gallon. However, by November 2011 it had recovered to $3 per gallon since when it has mostly hovered between $3.50 and $4 per gallon. Most of the statistical methods presented in this text apply to cross-sectional rather than time series data.

Quantitative data that measure how many are discrete. Quantitative data that measure how much are continuous because no separation occurs between the possible data values.

FIGURE 1.1
Wholesale price of crude oil per gallon (US$) 2008–2012
EIA (www.eia.doe.gov/)



Crude oil price is composite refiner acquisition cost. Retail prices include state and federal

Source: Short-Term Energy Outlook, November 2012

1.3 DATA SOURCES

Data can be obtained from existing sources or from surveys and experimental studies designed to collect new data.

Existing sources

In some cases, data needed for a particular application already exist. Companies maintain a variety of databases about their employees, customers and business operations. Data on employee salaries, ages and years of experience can usually be obtained from internal personnel records. Other internal records contain data on sales, advertising expenditures, distribution costs, inventory levels and production quantities. Most companies also maintain detailed data about their customers. Table 1.2 shows some of the data commonly available from internal company records.

Organizations that specialize in collecting and maintaining data make available substantial amounts of business and economic data. Companies access these external data sources through leasing arrangements or by purchase. Dun & Bradstreet, Bloomberg and the Economist Intelligence Unit are three sources that provide extensive business database services to clients. ACNielsen built successful businesses collecting and processing data that they sell to advertisers and product manufacturers.

Data are also available from a variety of industry associations and special interest organizations. The European Tour Operators, Association and European Travel Commission provide information on tourist trends and travel expenditures by visitors to and from countries in Europe. Such data would be of interest to firms and individuals in the travel industry. The Graduate Management Admission Council maintains data on test scores, student characteristics and graduate management education programmes. Most of the data from these types of sources are available to qualified users at a modest cost.

The Internet continues to grow as an important source of data and statistical information. Almost all companies maintain websites that provide general information about the company as well as data on sales, number of employees, number of products, product prices and product specifications. In addition, a number of companies now specialize in making information available over the Internet. As a result, one can obtain access to stock quotes, meal prices at restaurants, salary data and an almost infinite variety of information. Government agencies are another important source of existing data. For instance, Eurostat maintains considerable data on employment rates, wage rates, size of the labour force and union membership. Table 1.3 lists selected governmental agencies and some of the data they provide. Most government agencies that collect and process data also make the results available through a website. For instance, the Eurostat has a wealth of data at its website, http://ec.europa.eu/eurostat. Figure 1.2 shows the homepage for the Eurostat.

TABLE 1.2 Examples of data available from internal company records

Source	Some of the data typically available
Employee records	Name, address, social security number, salary, number of vacation days, number of sick days and bonus
Production records	Part or product number, quantity produced, direct labour cost and materials cost
Inventory records	Part or product number, number of units on hand, reorder level, economic order quantity and discount schedule
Sales records	Product number, sales volume, sales volume by region and sales volume by customer type
Credit records	Customer name, address, phone number, credit limit and accounts receivable balance
Customer profile	Age, gender, income level, household size, address and preferences

TABLE 1.3 Examples of data available from selected European sources

Source	Some of the data available
Europa rates (http://europa.eu)	Travel, VAT (value added tax), euro exchange employment, population and social conditions
Eurostat (http://epp.eurostat.ec.europa.eu/)	Education and training, labour market, living conditions and welfare
European Central Bank (www.ecb.int/)	Monetary, financial markets, interest rate and balance of payments statistics, unit labour costs, compensation per employee, labour productivity, consumer prices, construction prices

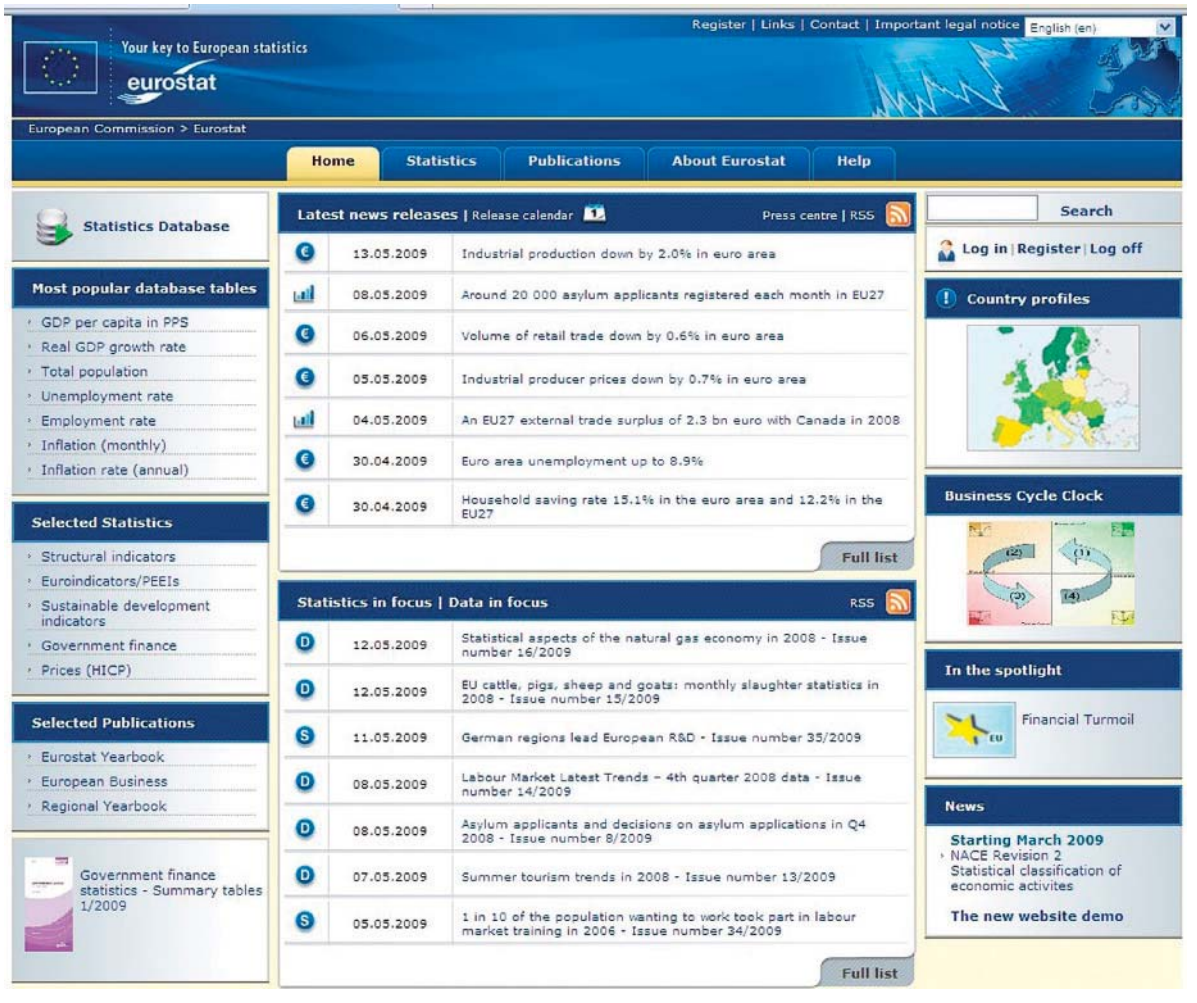


FIGURE 1.2
Eurostat homepage

Statistical studies

Sometimes the data needed for a particular application are not available through existing sources. In such cases, the data can often be obtained by conducting a statistical study. Statistical studies can be classified as either *experimental* or *observational*.

In an experimental study, a variable of interest is first identified. Then one or more other variables are identified and controlled so that data can be obtained about how they influence the variable of interest. For example, a pharmaceutical firm might be interested in conducting an experiment to learn about how a new drug affects blood pressure. Blood pressure is the variable of interest in the study. The dosage level of the new drug is another variable that is hoped to have a causal effect on blood pressure. To obtain data about the effect of the new drug, researchers select a sample of individuals. The dosage level of the new drug is controlled, as different groups of individuals are given different dosage levels. Before and after data on blood pressure are collected for each group. Statistical analysis of the experimental data can help determine how the new drug affects blood pressure.

Non-experimental, or observational, statistical studies make no attempt to control the variables of interest. A survey is perhaps the most common type of observational study. For instance, in a personal interview survey, research questions are first identified. Then a questionnaire is designed and administered to a sample of individuals. Some restaurants use observational studies to obtain data about their customers' opinions of the quality of food, service, atmosphere and so on. A questionnaire used by the Lobster Pot Restaurant in Limerick City, Ireland, is shown in Figure 1.3. Note that the customers completing the questionnaire are asked to provide ratings for five variables: food quality, friendliness of service, promptness of service, cleanliness and management. The response categories of excellent, good, satisfactory and unsatisfactory provide ordinal data that enable Lobster Pot's managers to assess the quality of the restaurant's operation.

Managers wanting to use data and statistical analyses as an aid to decision-making must be aware of the time and cost required to obtain the data. The use of existing data sources is desirable when data must be obtained in a relatively short period of time.

The LOBSTER Pot
RESTAURANT

We are happy you stopped by the Lobster Pot Restaurant and want to make sure you will come back. So, if you have a little time, we will really appreciate it if you will fill out this card. Your comments and suggestions are extremely important to us. Thank you!

Server's Name _____

	Excellent	Good	Satisfactory	Unsatisfactory
Food Quality	❑	❑	❑	❑
Friendly Service	❑	❑	❑	❑
Prompt Service	❑	❑	❑	❑
Cleanliness	❑	❑	❑	❑
Management	❑	❑	❑	❑

Comments _____

What prompted your vist to us? _____

Please drop in suggestion box at entrance. Thank you.

FIGURE 1.3
Customer opinion questionnaire used by the Lobster Pot Restaurant, Limerick City, Ireland

If important data are not readily available from an existing source, the additional time and cost involved in obtaining the data must be taken into account. In all cases, the decision-maker should consider the contribution of the statistical analysis to the decision-making process. The cost of data acquisition and the subsequent statistical analysis should not exceed the savings generated by using the information to make a better decision.

Data acquisition errors

Managers should always be aware of the possibility of data errors in statistical studies. Using erroneous data can be worse than not using any data at all. An error in data acquisition occurs whenever the data value obtained is not equal to the true or actual value that would be obtained with a correct procedure. Such errors can occur in a number of ways. For example, an interviewer might make a recording error, such as a transposition in writing the age of a 24-year-old person as 42, or the person answering an interview question might misinterpret the question and provide an incorrect response.

Experienced data analysts take great care in collecting and recording data to ensure that errors are not made. Special procedures can be used to check for internal consistency of the data. For instance, such procedures would indicate that the analyst should review the accuracy of data for a respondent shown to be 22 years of age but reporting 20 years of work experience. Data analysts also review data with unusually large and small values, called outliers, which are candidates for possible data errors. In Chapter 3 we present some of the methods statisticians use to identify outliers.

Errors often occur during data acquisition. Blindly using any data that happen to be available or using data that were acquired with little care can result in misleading information and bad decisions. Thus, taking steps to acquire accurate data can help ensure reliable and valuable decision-making information.

1.4 DESCRIPTIVE STATISTICS

Most of the statistical information in newspapers, magazines company reports and other publications consists of data that are summarized and presented in a form that is easy for the reader to understand. Such summaries of data, which may be tabular, graphical or numerical, are referred to as **descriptive statistics**.

Refer again to the data set in Table 1.1 showing data on 22 European stock exchanges. Methods of descriptive statistics can be used to provide summaries of the information in this data set. For example, a tabular summary of the data for the six busiest exchanges by trade for the categorical variable exchange is shown in Table 1.4. A graphical summary of the same data, called a bar graph, is shown in Figure 1.4. These types of tabular and graphical summaries generally make the data easier to interpret. Referring to Table 1.4 and Figure 1.4, we can see easily that the majority of trades are for the London exchange (covering trading in Paris, Brussels, Amsterdam and Lisbon). On a percentage basis, 29.1 per cent of all trades for the 22 European stock exchanges occur through London. Similarly 26.8 per cent occur for Euronext and 13.4 per cent for Deutsche Börse. Note from Table 1.4 that 93 per cent of all trades take place in just six of the 22 European exchanges.

TABLE 1.4 Per cent frequencies for six busiest exchanges by trades

Exchange	% of Trades
London	29.1
Euronext	26.8
Deutsche Börse	13.4
Borsa Italiana	10.4
NASDAQ OMX Nordic	8.0
Spanish (BME)	4.9
TOTAL	**92.6**

FIGURE 1.4
Bar graph for the exchange variable

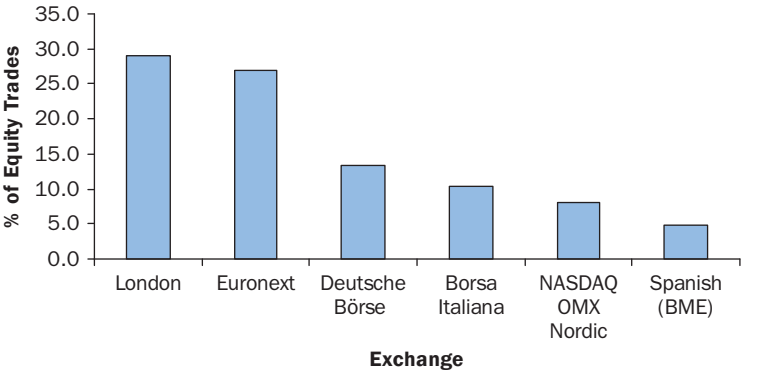
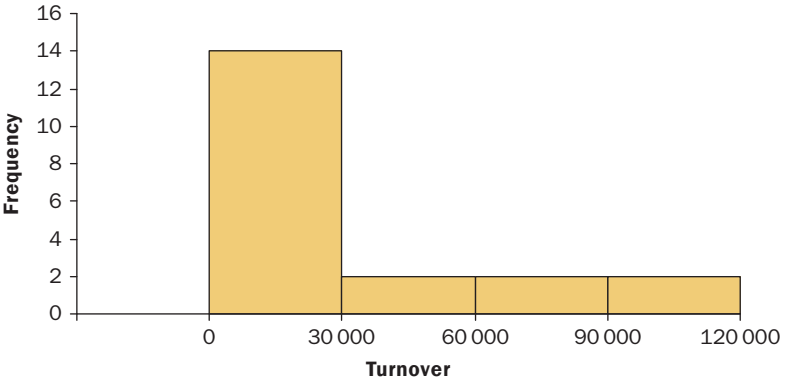


FIGURE 1.5
Histogram of turnover (€m)



A graphical summary of the data for the quantitative variable turnover for the exchanges, called a histogram, is provided in Figure 1.5. The histogram makes it easy to see that the turnover ranges from €0.0 to €120 000m, with the highest concentrations between €0 and €30 000m.

In addition to tabular and graphical displays, numerical descriptive statistics are used to summarize data. The most common numerical descriptive statistic is the average, or mean. Using the data on the variable turnover for the exchanges in Table 1.1, we can compute the average turnover by adding the turnover for the 21 exchanges where turnover has been declared and dividing the sum by 21. Doing so provides an average turnover of €23 144 million. This average demonstrates a measure of the central tendency, or central location, of the data for that variable.

In a number of fields, interest continues to grow in statistical methods that can be used for developing and presenting descriptive statistics. Chapters 1 and 3 devote attention to the tabular, graphical and numerical methods of descriptive statistics.

1.5 STATISTICAL INFERENCE

Many situations require data for a large group of elements (individuals, companies, voters, households, products, customers and so on). Because of time, cost and other considerations, data can be collected from only a small portion of the group. The larger group of elements in a particular study is called the **population**, and the smaller group is called the **sample**. Formally, we use the following definitions.

Population

A *population* is the set of all elements of interest in a particular study.

Sample

A *sample* is a subset of the population.

The process of conducting a survey to collect data for the entire population is called a **census**. The process of conducting a survey to collect data for a sample is called a **sample survey**. As one of its major contributions, statistics uses data from a sample to make estimates and test hypotheses about the characteristics of a population through a process referred to as **statistical inference**.

As an example of statistical inference, let us consider the study conducted by Electronica Nieves. Nieves manufactures a high-intensity light bulb used in a variety of electrical products. In an attempt to increase the useful life of the light bulb, the product design group developed a new light bulb filament. In this case, the population is defined as all light bulbs that could be produced with the new filament. To evaluate the advantages of the new filament, 200 bulbs with the new filament were manufactured and tested. Data collected from this sample showed the number of hours each light bulb operated before the filament burned out or the bulb failed. See Table 1.5.

Suppose Nieves wants to use the sample data to make an inference about the average hours of useful life for the population of all light bulbs that could be produced with the new filament. Adding the 200 values in Table 1.5 and dividing the total by 200 provides the sample average lifetime for the light bulbs: 76 hours. We can use this sample result to estimate that the average lifetime for the light bulbs in the population is 76 hours. Figure 1.6 provides a graphical summary of the statistical inference process for Electronica Nieves.

TABLE 1.5 Hours until failure for a sample of 200 light bulbs for the Electronica Nieves example

107	73	68	97	76	79	94	59	98	57
54	65	71	70	84	88	62	61	79	98
66	62	79	86	68	74	61	82	65	98
62	116	65	88	64	79	78	79	77	86
74	85	73	80	68	78	89	72	58	69
92	78	88	77	103	88	63	68	88	81
75	90	62	89	71	71	74	70	74	70
65	81	75	62	94	71	85	84	83	63
81	62	79	83	93	61	65	62	92	65
83	70	70	81	77	72	84	67	59	58
78	66	66	94	77	63	66	75	68	76
90	78	71	101	78	43	59	67	61	71
96	75	64	76	72	77	74	65	82	86
66	86	96	89	81	71	85	99	59	92
68	72	77	60	87	84	75	77	51	45
85	67	87	80	84	93	69	76	89	75
83	68	72	67	92	89	82	96	77	102
74	91	76	83	66	68	61	73	72	76
73	77	79	94	63	59	62	71	81	65
73	63	63	89	82	64	85	92	64	73

NIEVES

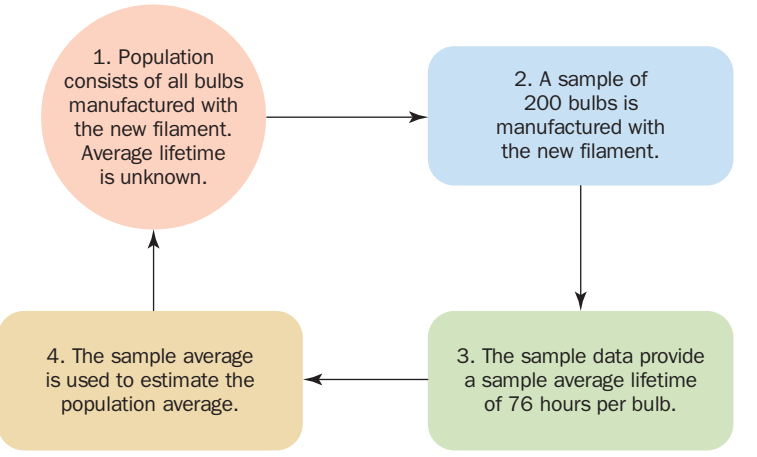


FIGURE 1.6
The process of statistical inference for the Electronica Nieves example

Whenever statisticians use a sample to estimate a population characteristic of interest, they usually provide a statement of the quality, or precision, associated with the estimate. For the Nieves example, the statistician might state that the point estimate of the average lifetime for the population of new light bulbs is 76 hours with a margin of error of ± four hours. Thus, an interval estimate of the average lifetime for all light bulbs produced with the new filament is 72 hours to 80 hours. The statistician can also state how confident he or she is that the interval from 72 hours to 80 hours contains the population average.

1.6 COMPUTERS AND STATISTICAL ANALYSIS

Because statistical analysis typically involves large amounts of data, analysts frequently use computer software for this work. For instance, computing the average lifetime for the 200 light bulbs in the Electronica Nieves example (see Table 1.5) would be quite tedious without a computer. To facilitate computer usage, the larger data sets in this book are available on the website that accompanies the text. A logo in the left margin of the text (e.g. Nieves) identifies each of these data sets. The data files are available in MINITAB, SPSS and EXCEL formats. In addition, we provide instructions on the website for carrying out many of the statistical procedures using MINITAB, SPSS and EXCEL.

1.7 DATA MINING

With the aid of magnetic card readers, bar code scanners, and point-of-sale terminals, most organizations obtain large amounts of data on a daily basis. And, even for a small local restaurant that uses touch screen monitors to enter orders and handle billing, the amount of data collected can be significant. For large retail companies, the sheer volume of data collected is hard to conceptualize, and determining how to effectively use these data to improve profitability is a challenge. For example, mass retailers such as Wal-Mart capture data on 20 to 30 million transactions every day, telecommunication companies such as Vodafone generated in 2011 an average of a billion call records per day, and Visa processes 6800 payment transactions per second or approximately 600 million transactions per day. Storing and managing the transaction data is a significant undertaking.

The term data warehousing is used to refer to the process of capturing, storing and maintaining the data. Computing power and data collection tools have reached the point where it is now feasible to store and retrieve extremely large quantities of data in seconds. Analysis of the data in the warehouse may result in decisions that will lead to new strategies and higher profits for the organization.

The subject of **data mining** deals with methods for developing useful decision-making information from large data bases. Using a combination of procedures from statistics, mathematics and computer science, analysts 'mine the data' in the warehouse to convert it into useful information, hence the name

data mining. Data mining systems that are the most effective use automated procedures to extract information from the data using only the most general or even vague queries by the user. And data mining software automates the process of uncovering hidden predictive information that in the past required hands-on analysis.

The major applications of data mining have been made by companies with a strong consumer focus, such as retail businesses, financial organizations and communication companies. Data mining has been successfully used to help retailers such as Amazon and Barnes & Noble determine one or more related products that customers who have already purchased a specific product are also likely to purchase. Then, when a customer logs on to the company's website and purchases a product, the website uses pop-ups to alert the customer about additional products that the customer is likely to purchase. In another application, data mining may be used to identify customers who are likely to spend more than €20 on a particular shopping trip. These customers may then be identified as the ones to receive special email or regular mail discount offers to encourage them to make their next shopping trip before the discount termination date.

Data mining is a technology that relies heavily on methodology such as statistics, clustering, decision trees and rule induction. But it takes a creative integration of all these methods and computer science technologies involving artificial intelligence and machine learning to make data mining effective. A significant investment in time and money is required to implement commercial data mining software packages developed by firms such as IBM SPSS and SAS. The statistical concepts introduced in this text will be helpful in understanding the statistical methodology used by data mining software packages and enable you to better understand the statistical information that is developed.

Because statistical models play an important role in developing predictive models in data mining, many of the concerns that statisticians deal with in developing statistical models are also applicable. For instance, a concern in any statistical study involves the issue of model reliability. Finding a statistical model that works well for a particular sample of data does not necessarily mean that it can be reliably applied to other data. One of the common statistical approaches to evaluating model reliability is to divide the sample data set into two parts: a training data set and a test data set. If the model developed using the training data is able to accurately predict values in the test data, we say that the model is reliable. One advantage that data mining has over classical statistics is that the enormous amount of data available allows the data mining software to partition the data set so that a model developed for the training data set may be tested for reliability on other data. In this sense, the partitioning of the data set allows data mining to develop models and relationships and then quickly observe if they are repeatable and valid with new and different data. On the other hand, a warning for data mining applications is that with so much data available, there is a danger of over-fitting the model to the point that misleading associations and cause/effect conclusions appear to exist. Careful interpretation of data mining results and additional testing will help avoid this pitfall.

Although statistical methods play an important role in data mining, both in terms of discovering relationships in the data and predicting future outcomes, a thorough coverage of the topic is outside the scope of this text.

EXERCISES

COMPLETE SOLUTIONS

1. Discuss the differences between statistics as numerical facts and statistics as a discipline or field of study.

2. Every year *Condé Nast Traveler* conducts an annual survey of subscribers to determine the best new places to stay throughout the world. Table 1.6 shows the ten hotels that were most highly ranked in their 2006 'hot list' survey. Note that (daily) rates quoted are for double rooms and are variously expressed in US dollars, British pounds or euros.

   a. How many elements are in this data set?
   b. How many variables are in this data set?

c. Which variables are categorical and which variables are quantitative?
d. What type of measurement scale is used for each of the variables?

3. Refer to Table 1.6:

- What is the average number of rooms for the ten hotels?
- If €1 = US$1.3149 = £0.8986 compute the average room rate in euros.

TABLE 1.6 The ten best new hotels to stay in, in the world

Hot list ranking	Name of property	Country	Room rate	Number of rooms
1	Amangalla, Galle	Sri Lanka	US$574	30
2	Amanwella, Tangalle	Sri Lanka	US$275	30
3	Bairro Alto Hotel, Lisbon	Portugal	€180	55
4	Basico, Playa Del Carmen	Mexico	US$166	15
5	Beit Al Mamlouka	Syria	£75	8
6	Brown's Hotel, London	England	£347	117
7	Byblos Art Hotel Villa Amista, Verona	Italy	€270	60
8	Cavas Wine Lodge, Mendoza	Argentina	US$375	14
9	Convento Do Espinheiro Heritage Hotel & Spa, Evora	Portugal	€213	59
10	Cosmopolitan, Toronto	Canada	£150	97

Source: *Condé Nast Traveler*, May 2006 (www.cntraveller.com/magazine/the-hot-list-2006)

HOTELS

- What is the percentage of hotels located in Portugal?
- What is the percentage of hotels with 20 rooms or fewer?

4. Audio systems are typically made up of an MP3 player, a mini disk player, a cassette player, a CD player and separate speakers. The data in Table 1.7 show the product rating and retail price range for a popular selection of systems. Note that the code Y is used to confirm when a player is included in the system, N when it is not. Output power (watts) details are also provided (Kelkoo Electronics 2006).

a. How many elements does this data set contain?
b. What is the population?
c. Compute the average output power for the sample.

COMPLETE SOLUTIONS

5. Consider the data set for the sample of eight audio systems in Table 1.7.

a. How many variables are in the data set?
b. Which of the variables are quantitative and which are categorical?
c. What percentage of the audio systems has a four star rating or higher?
d. What percentage of the audio systems includes an MP3 player?

TABLE 1.7 A sample of eight audio systems

Brand and model	Product rating (# of stars)	Price (£)	MP3 player	Mini disk player	Cassette player	CD (watts) player	Output
Technics SCEH790	1	320–400	Y	N	Y	Y	360
Yamaha M170	3	162–290	N	N	N	Y	50
Panasonic SCPM29	5	188	Y	N	Y	Y	70
Pure Digital DMX50	3	180–230	N	N	N	Y	80
Sony CMTNEZ3	5	60–100	Y	N	Y	Y	30
Philips FWM589	4	143–200	Y	N	N	Y	400
Philips MCM9	5	93–110	Y	N	Y	Y	100
Samsung MM-C6	5	100–130	Y	N	N	Y	40

Source: Kelkoo (http://audiovisual.kelkoo.co.uk)

AUDIO-SYSTEMS

6. State whether each of the following variables is categorical or quantitative and indicate its measurement scale.

a. Annual sales.
b. Soft drink size (small, medium, large).
c. Occupational classification (SOC 2000).
d. Earnings per share.
e. Method of payment (cash, cheque, credit card).

COMPLETE SOLUTIONS

7. The Health & Wellbeing Survey ran over a three-week period (ending 19 October 2007) and 389 respondents took part. The survey asked the respondents to respond to the statement, 'How would you describe your own physical health at this time?' (http://inform. glam.ac.uk/news/2007/10/24/health-wellbeing-staff-survey-results/). Response categories were strongly agree, agree, neither agree or disagree, disagree and strongly disagree.

a. What was the sample size for this survey?
b. Are the data categorical or quantitative?
c. Would it make more sense to use averages or percentages as a summary of the data for this question?
d. Of the respondents, 57 per cent agreed with the statement. How many individuals provided this response?

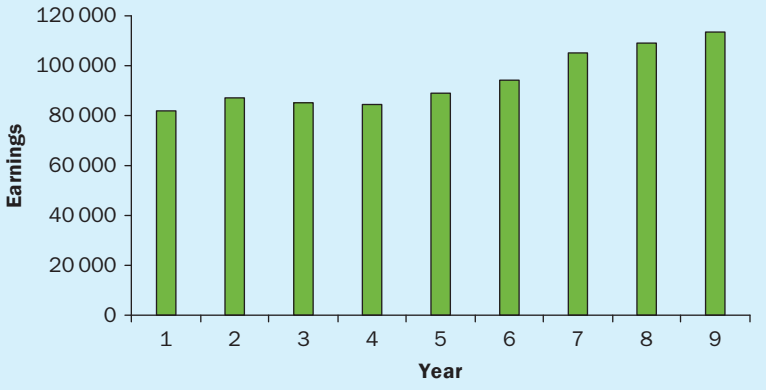
8. State whether each of the following variables is categorical or quantitative and indicate its measurement scale.

a. Age.
b. Gender.
c. Class rank.
d. Make of car.
e. Number of people favouring closer European integration.

9. Figure 1.7 provides a bar chart summarizing the actual earnings for Volkswagen for the years 2000 to 2008 (Source: *Volkswagen AG Annual Reports 2001–2008*).

a. Are the data categorical or quantitative?
b. Are the data times series or cross-sectional?
c. What is the variable of interest?
d. Comment on the trend in Volkswagen's earnings over time. Would you expect to see an increase or decrease in 2009?

FIGURE 1.7
Volkswagen's earnings (€m) 1998–2009



10. The Hawaii Visitors' Bureau collects data on visitors to Hawaii. The following questions were among 16 asked in a questionnaire handed out to passengers during incoming airline flights.

- This trip to Hawaii is my: 1st, 2nd, 3rd, 4th, etc.
- The primary reason for this trip is: (ten categories including vacation, convention, honeymoon).
- Where I plan to stay: (11 categories including hotel, apartment, relatives, camping).
- Total days in Hawaii.

a. What is the population being studied?
b. Is the use of a questionnaire a good way to reach the population of passengers on incoming airline flights?
c. Comment on each of the four questions in terms of whether it will provide categorical or quantitative data.

11. A manager of a large corporation recommends a $10 000 raise be given to keep a valued subordinate from moving to another company. What internal and external sources of data might be used to decide whether such a salary increase is appropriate?

12. In a recent study of causes of death in men 60 years of age and older, a sample of 120 men indicated that 48 died as a result of some form of heart disease.

a. Develop a descriptive statistic that can be used as an estimate of the percentage of men 60 years of age or older who die from some form of heart disease.
b. Are the data on cause of death categorical or quantitative?
c. Discuss the role of statistical inference in this type of medical research.

13. In 2007, 75.4 per cent of *Economist* readers had stayed in a hotel on business in the previous 12 months with 32.4 per cent of readers using first business class for travel.

a. What is the population of interest in this study?
b. Is class of travel a categorical or quantitative variable?
c. If a reader had stayed in a hotel on business in the previous 12 months, would this be classed as a categorical or quantitative variable?
d. Does this study involve cross-sectional or time series data?
e. Describe any statistical inferences *The Economist* might make on the basis of the survey.

ONLINE RESOURCES

For the data files and additional online resources for Chapter 1, go to the accompanying online platform. (See the 'About the Digital Resources' page in the front of the book for more information on access.)

SUMMARY

Statistics is the art and science of collecting, analyzing, presenting and interpreting data. Nearly every college student majoring in business or economics is required to take a course in statistics. We began the chapter by describing typical statistical applications for business and economics.

Data consist of the facts and figures that are collected and analyzed. A set of measurements obtained for a particular element is an observation, Four scales of measurement used to obtain data on a particular variable include nominal, ordinal, interval and ratio. The scale of measurement for a variable is nominal when the data use labels or names to identify an attribute of an element. The scale is ordinal if the data demonstrate the properties of nominal data and the order or rank of the data is meaningful. The scale is interval if the data demonstrate the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Finally, the scale of measurement is ratio if the data show all the properties of interval data and the ratio of two values is meaningful.

For purposes of statistical analysis, data can be classified as categorical or quantitative.

Categorical data use labels or names to identify an attribute of each element. Categorical data use either the nominal or ordinal scale of measurement and may be non-numeric or numeric. Quantitative data are numeric values that indicate how much or how many. Quantitative data use either the interval or ratio scale of measurement. Ordinary arithmetic operations are meaningful only if the data are quantitative. Therefore, statistical computations used for quantitative data are not always appropriate for categorical data.

In Sections 1.4 and 1.5 we introduced the topics of descriptive statistics and statistical inference. Definitions of the population and sample were provided and different types of descriptive statistics – tabular, graphical and numerical – used to summarize data. The process of statistical inference uses data obtained from a sample to make estimates or test hypotheses about the characteristics of a population.

The last two sections of the chapter provide information on the role of computers in statistical analysis and a brief overview of the relative new field of data mining.

KEY TERMS

Categorical data
Categorical variable
Census
Cross-sectional data
Data
Data mining
Data set
Descriptive statistics
Elements
Interval scale
Nominal scale
Observation
Ordinal scale
Population
Quantitative data
Quantitative variable
Ratio scale
Sample
Sample survey
Statistical inference
Statistics
Time series data
Variable