PART

1

Presenting and describing information

Real People, Real Stats

David McCourt BDO

Which company are you currently working for and what are some of your responsibilities?

I work at BDO, Chartered Accountants and Advisors, in the corporate finance team. My primary responsibilities include the preparation of financial models and valuation reports.

List five words that best describe your personality.

Affable, level-headed, perceptive, analytical, assured (according to my colleagues).

What are some things that motivate you?

Success, working with a team, client satisfaction.

When did you first become interested in statistics?

I never really understood statistics at school and it was a minor part of my university degree. However, statistics play a significant role in many of our valuations, including discounted cash flow valuations and share option valuations.

Complete the following sentence. A world without statistics …

… is not worth thinking about.

LET'S TALK STATS

What do you enjoy most about working in statistics?

We use data services and statistical tools that have been created by third parties. I can use, and talk reasonably knowledgeably about, statistical data without being an expert.

a quick q&a

Describe your first statistics-related job or work experience. Was this a positive or a negative experience?

The first time I can recall using statistics was for a share option valuation. We had to determine the share price volatility based on historical share price data. There are about half a dozen methods that can be used, all with various advantages and disadvantages. I did and still find this analysis interesting.

What do you feel is the most common misconception about your work held by students who are studying statistics? Please explain.

Statistics provides information to support our analysis and decisions. However, the information is never perfect, and subjectivity and commercial common sense play a large part in our work.

Do you need to be good at maths to understand and use statistics successfully?

I think you need to have a logical and well-structured approach to problems. These skills would probably make you good at both maths and statistics.

Is there a high demand for statisticians in your industry (or in other industries)? Please explain.

The finance industry is heavily reliant on statistics. I expect there is high demand for statisticians from the various data providers, and in a number of specialist areas (e.g. insurance).

PRESENTING AND DESCRIBING INFORMATION

Does data collection play an important role in the decisions you make for your business/work? Please explain.

Accurate data collection is essential to our valuation projects. Although our work involves a degree of commercial acumen, it is essential that the data supports and justifies these decisions. We also aggregate data for internal business use to measure staff productivity, business performance and forecasting budgets.

Describe a project that you have worked on recently that might have involved data collection. Please be specific.

We recently valued an infrastructure asset using the discounted cash flow model. The model requires two essential inputs: the forecast of future cash flows of the asset, and the discount rate that reflects the riskiness of those cash flows. To arrive at an appropriate discount rate we generally analyse comparable companies for an indication of the level of risk that should be attributed to the asset to be valued. In this exercise there are several instances of data collection. We collect five-year historical stock data for numerous comparable companies as an initial indication of risk. We then collect data on key financial indicators to assess the degree of comparability between the stock and the asset to be valued. To determine the risk-free rate and the market-risk premium, 10-year government bond rate data is collected.

How are these data usually summarised? What are some positives and negatives of these summary techniques?

We generally organise the collected data into Microsoft Excel workbooks. The main advantage of using this software is the ease of data analysis. Some powerful data analysis tools include data tables, What-If Analysis, Solver, charting and common statistical functions. Some shortcomings we have encountered using Excel is that data sometimes need to be rearranged depending on the analysis, [there can be] problems with inconsistent or missing data, and output can sometimes be incomplete. These factors increase the likelihood of errors in data analysis; however, for the purposes of corporate finance, Excel is generally sufficient as a means of summarising and analysing the data collected.

In your experience, what is the most commonly referred to measure of central tendency? What benefits does this measure offer over others?

In valuations, we generally prefer to use the median as a measure of central tendency rather than mean or mode. We find that the mean has one main disadvantage: it is particularly susceptible to outliers. When looking at comparable companies there are often outliers caused by one-off business issues that are irrelevant for the purposes of comparing our business. We very rarely use mode given that it only really coincides with the central tendency of data where the distribution is centre-heavy and there are generally few recurring figures in the data set.

Why is it important to be aware of the spread/variation of data points in a sample? What are the consequences of not knowing this type of information about your sample?

Without an understanding of the spread and variation of a data set there is no context to the measure of central tendency applied. A measure of central tendency summarises the data into a single value while the spread and variation of data gives an indication of how reliable an average or median summary of collected data is. For example, if the spread of values in the data set is relatively large it suggests the mean is not as representative, and a smoothing of data is required, when compared to a data set with a smaller range. Adopting a mean without reference to the spread can taint our analysis and results in a lack of validity to our decisions that are based on the data.

CHAPTER

1

Defining and Collecting data

THE HONG KONG AIRPORT SURVEY

You are departing Hong Kong International Airport on the next leg of your trip and have cleared Immigration. You are approached by a researcher holding a tablet computer who asks if you can answer a few questions. The first question determines if you are a visitor to Hong Kong or a resident. After establishing that you are a visitor the questions go on to determine the purpose of your visit, the name of your hotel, the activities you have undertaken and much additional information about your visit.

This information is useful for a tourism authority that has the task of marketing Hong Kong as a travel destination and monitoring the quality of visitors' experiences in the city. It may also inform the authority's government and commercial stakeholders, who provide transport, accommodation, and food and shopping for visitors, and be used for forward planning.

LEARNING OBJECTIVES

After studying this chapter you should be able to:

1. identify the types of data used in business
2. identify how statistics is used in business
3. recognise the sources of data used in business
4. distinguish between different survey sampling methods
5. evaluate the quality of surveys

Not so long ago, business students were unfamiliar with the word *data* and had little experience handling data. Today, every time you visit a search engine website or 'ask' your mobile device a question, you are handling data. And if you 'check in' to a location or indicate that you 'like' something, you are *creating* data as well.

You accept as almost true the premises of stories in which characters collect 'a lot of data' to uncover conspiracies, foretell disasters or catch a criminal.

You hear concerns about how the government or business might be able to 'spy' on you in some way or how large social media companies 'mine' your personal data for profit.

You hear the word *data* everywhere and may even have a 'data plan' for your smartphone. You know, in a general way, that data are facts about the world and that most data seem to be, ultimately, a set of numbers – that 34% of students recently polled prefer using a certain Internet browser, or that 50% of citizens believe the country is headed in the right direction, or that unemployment is down 3%, or that your best friend's social media account has 835 friends and 202 recent posts.

You cannot escape from data in this digital world. What, then, should you do? You could try to ignore data and conduct business by relying on hunches or your 'gut instincts'. However, if you want to use only gut instincts, then you probably shouldn't be reading this book or taking business courses in the first place.

You could note that there is so much data in the world – or just in your own little part of the world – that you couldn't possibly get a handle on it.

You could accept other people's data summaries and their conclusions without first reviewing the data yourself. That, of course, would expose yourself to fraudulent practices.

Or you could do things the proper way and realise the benefits of learning the methods of statistics, the subject of this book. You can learn, though, the procedures and methods that will help you make better decisions based on solid evidence. When you begin focusing on the procedures and methods involved in collecting, presenting and summarising a set of data, or forming conclusions about those data, you have discovered statistics.

In the Hong Kong Airport survey scenario it is important that research team members focus on the information that is needed by many different stakeholders when planning for future business and tourist visitors. If the research team fails to collect important information, or misrepresents the opinions of current visitors, stakeholders may make poor decisions about advertising, pricing, facilities and other factors relevant to attracting visitors and hosting them in Hong Kong. Failure to offer suitable facilities and experiences could affect the profitability of businesses in Hong Kong. In deciding how to collect the facts that are needed, it will help if you know something about the basic concepts of statistics.

1.1 BASIC CONCEPTS OF DATA AND STATISTICS

The Meaning of 'Data'

What do we mean by the word *data*? Its common use is somewhat different from its use in statistics. It could be described in a general way as meaning 'facts about the world'. However, statisticians distinguish between the traits or properties that relate to people or things and the actual values that these take.

variables
Characteristics or attributes that can be expected to differ from one individual to another.

data
The observed values of variables.

VARIABLES

Variables are characteristics of items or individuals.

DATA

Data are the observed values of variables.

For a group of people, we could examine the traits of age, country of birth or weight. For a group of cars, we could note the colour, current value or kilometres driven. These characteristics are called **variables**.

Data are the values associated with these traits or properties. As an example, in Table 1.1 we find a set of data collected from six people which represents observations on three different variables.

Table 1.1

Variable	Data
Age in years	24, 18, 53, 16, 22, 31
Country of birth	Australia, China, Australia, Malaysia, India, Australia
Weight in kilograms	50.2, 74.6, 96.3, 45.2, 56.1, 87.3

In this book, the word *data* is always plural to remind you that data are a collection or set of values. While we could say that a single value, such as 'Australia' is a *datum*, the terms *data point, observation, response* or *single data value* are more typically encountered.

operational definition
Defines how a variable is to be measured.

All variables should have an **operational definition** – a universally accepted meaning that is clear to all associated with an analysis. Without operational definitions, confusion can occur. An example of a situation where operational definitions are needed is for the process of data gathering by the Australian Bureau of Statistics (ABS). The ABS needs to collect information about the country of birth of a person and also the countries in which their father and mother were born. While this might seem straightforward, definitional problems arise in the case of people who were adopted or have step- or foster parents or other guardians. So the operational definition used is:

- 'Country of birth of person', which is the country identified as being the one in which the person was born
- 'Country of birth of father', which is the country in which the person's birth father was born, and
- 'Country of birth of mother', which is the country in which the person's birth mother was born (Australian Bureau of Statistics, *Country of Birth Standard,* Cat. No. 1200.0.55.004, 2016).

The Meaning of 'Statistics'

statistics
A branch of mathematics concerned with the collection and analysis of data.

Statistics is the branch of mathematics that examines ways to process and analyse data. It provides procedures to collect and transform data in ways that are useful to business decision makers.

Statistics allows you to determine whether your data represent information that could be used in making better decisions. Therefore, it helps you determine whether differences in the

numbers are meaningful in a significant way or are due to chance. To illustrate, consider the following reports:

- In 'News use across social media platforms 2016' the Pew Research Center reported in May 2016, that 67% of the adult US population had a Facebook account and 66% of users get news from the site (<http://assets.pewresearch.org/wpcontent/uploads/sites/13/2016/05/PJ_2016.05.26_social-media-and-news_FINAL-1.pdf>, accessed 12 June 2017).
- In a blog titled 'The top 10 benefits of newspaper advertising', the 360 Degree Marketing Group says that a study showed newspaper advertising was considered a more trusted paid medium for information (58%) compared with television (54%), radio (49%) or online (27%) (<www.360degreemarketing.com.au/Blog/bid/407663/The-Top-10-Benefits-of-Newspaper-Advertising>, accessed 12 June 2017).

Without statistics, you cannot determine whether the 'numbers' in these stories represent useful information. Without statistics, you cannot validate claims such as the statement that advertising in newspapers or on television is more trusted than online advertising. And without statistics, you cannot see patterns that large amounts of data sometimes reveal.

Statistics is a way of thinking that can help you make better decisions. It helps you solve problems that involve decisions based on data that have been collected. You may have had some statistics instruction in the past. If you ever created a chart to summarise data or calculated values such as averages to summarise data, you have used statistics. But there's even more to statistics than these commonly taught techniques, as the detailed table of contents shows.

Statistics is undergoing important changes today. There are new ways of visualising data that did not exist, were not practicable or were not widely known until recently. And, increasingly, statistics today is being used to 'listen' to what the data might be telling you rather than just being a way to use data to prove something you want to say.

If you associate statistics with doing a lot of mathematical calculations, you will quickly learn that business statistics uses software to perform the calculations for you (and, generally, the software calculates with more precision and efficiency than you could do manually). But while you do not need to be a good manual calculator to apply statistics, because statistics is a way of thinking, you do need to follow a framework or plan to minimise possible errors of thinking and analysis.

One such framework consists of the following tasks to help apply statistics to business decision making:

1. **Define** the data that you want to study in order to solve a problem or meet an objective.
2. **Collect** the data from appropriate sources.
3. **Organise** the data collected by developing tables.
4. **Visualise** the data collected by developing charts.
5. **Analyse** the data collected to reach conclusions and present those results.

Typically, you do the tasks in the order listed. You must always do the first two tasks to have meaningful outcomes, but, in practice, the order of the other three can change or appear inseparable. Certain ways of visualising data will help you to organise your data while performing preliminary analysis as well. In any case, when you apply statistics to decision making, you should be able to identify all five tasks, and you should verify that you have done the first two tasks before the other three.

Using this framework helps you to apply statistics to these four broad categories of business activities:

1. Summarise and visualise business data.
2. Reach conclusions from those data.
3. Make reliable forecasts about business activities.
4. Improve business processes.

Throughout this book, and especially in the scenarios that begin the chapters, you will discover specific examples of how we can apply statistics to business situations.

Statistics is itself divided into two branches, both of which are applicable to managing a business. **Descriptive statistics** focuses on collecting, summarising and presenting a set of data. **Inferential statistics** uses sample data to draw conclusions about a population.

descriptive statistics
The field that focuses on summarising or characterising a set of data.

inferential statistics
Uses information from a sample to draw conclusions about a population.

Descriptive statistics has its roots in the record-keeping needs of large political and social organisations. Refining the methods of descriptive statistics is an ongoing task for government statistical agencies such as the Australian Bureau of Statistics and Statistics New Zealand as they prepare for each Census. In Australia, a Census is scheduled to be carried out every five years (e.g. 2011 and 2016) to count the entire population and to collect data about education, occupation, languages spoken and many other characteristics of the citizens. A large amount of planning and training is necessary to ensure that the data collected represent an accurate record of the population's characteristics at the Census date. However, despite the best planning, such an immense data collection task can be affected by external factors. The Australian Census held in 2016 was badly affected by a computer shutdown on Census night, 9 August. It was blamed on the need to protect the system from denial of service cyber attacks and added approximately $30 million to the cost of the Census (<www.abc.net.au/ news/2016-10-25/turning-router-off-and-on-could-have-prevented-census-outage/7963916>, accessed 13 July 2017).

The foundation of inferential statistics is based on the mathematics of probability theory. Inferential methods use sample data to calculate statistics that provide estimates of the characteristics of the entire population.

Today, applications of statistical methods can be found in different areas of business. Accounting uses statistical methods to select samples for auditing purposes and to understand the cost drivers in cost accounting. Finance uses statistical methods to choose between alternative portfolio investments and to track trends in financial measures over time. Management uses statistical methods to improve the quality of the products manufactured or the services delivered by an organisation. Marketing uses statistical methods to estimate the proportion of customers who prefer one product over another and to draw conclusions about what advertising strategy might be most useful in increasing sales of a product.

Other Important Definitions

Now that the terms *variables, data* and *statistics* have been defined, you need to understand the meaning of the terms *population*, *sample* and *parameter*.

POPULATION

A **population** consists of all the members of a group about which you want to draw a conclusion.

SAMPLE

A **sample** is the portion of the population selected for analysis.

PARAMETER

A **parameter** is a numerical measure that describes a characteristic of a population.

STATISTIC

A **statistic** is a numerical measure that describes a characteristic of a sample.

population
A collection of all members of a group being investigated.

sample
The portion of the population selected for analysis.

parameter
A numerical measure of some population characteristic.

statistic
A numerical measure that describes a characteristic of a sample.

Examples of populations are all the full-time students at a university, all the registered voters in New Zealand and all the people who were customers of the local shopping centre last weekend. The term *population* is not limited to groups of people. We could refer to a

population of all motor vehicles registered in Victoria. Two factors need to be specified when defining a population:

1. the entity (e.g. people or motor vehicles)
2. the boundary (e.g. registered to vote in New Zealand or registered in Victoria for road use).

Samples could be selected from each of the populations mentioned above. Examples include 10 full-time students selected for a focus group; 500 registered voters in New Zealand who were contacted by telephone for a political poll; 30 customers at the shopping centre who were asked to complete a market research survey; and all the vehicles registered in Victoria that are more than 10 years old. In each case, the people or the vehicles in the sample represent a portion, or subset, of the people or vehicles comprising the population.

The average amount spent by all the customers at the local shopping centre last weekend is an example of a *parameter*. Information from all the shoppers in the entire population is needed to calculate this parameter.

The average amount spent by the 30 customers completing the market research survey is an example of a *statistic*. Information from a sample of only 30 of the shopping centre's customers is used in calculating the statistic.

1.2 TYPES OF VARIABLES

As illustrated in Figure 1.1, there are two types of variables – categorical and numerical, sometimes referred to as qualitative and quantitative variables respectively.

The Hong Kong airport survey

Travellers in the departure lounge of the busy Hong Kong International Airport are asked to complete a survey with questions about various aspects of their visit to the city and future travel plans. The interviewer first asks if the traveller is a resident or a visitor. If the traveller is a visitor, the survey proceeds. The survey includes these questions:

- How many visits have you made to Hong Kong prior to this one? _____
- How long is it since your visit here? _____
- How satisfied were you with your accommodation?

  Very satisfied ☐ Satisfied ☐ Undecided ☐ Dissatisfied ☐ Very dissatisfied ☐
- How many times during this visit did you travel by ferry? _____
- Shopping in Hong Kong stores gives good value for money

  Almost always ☐ Very infrequently ☐
  Sometimes ☐ Never ☐
- Was the purpose of your visit business? Yes ☐ No ☐
- Are you likely to return to Hong Kong in the next 12 months? Yes ☐ No ☐

You have been asked to review the survey. What type of data does the survey seek to collect?

What type of information can be generated from the data of the completed survey? How can the research company's clients use this information when planning for future visitors? What other questions would you suggest for the survey?

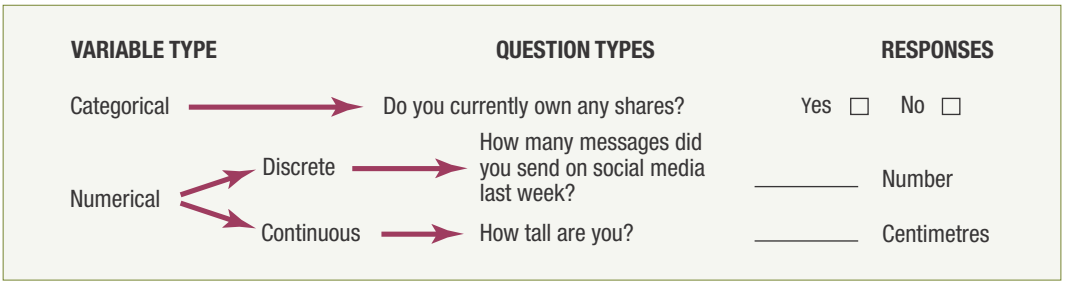

Figure 1.1
Types of variables

categorical variables
Take values that fall into one or more categories.

numerical variables
Take numbers as their observed responses.

LEARNING OBJECTIVE 1
Identify the types of data used in business

discrete variables
Can only take a finite or countable number of values.

continuous variables
Can take any value between specified limits.

Categorical variables yield categorical responses, such as *yes* or *no* or *male* or *female* answers. An example is the response to the question 'Do you currently own any shares?' because it is limited to a simple *yes* or *no* answer. Another example is the response to the question in the Hong Kong Airport survey (presented on page 9), 'Are you likely to return to Hong Kong in the next 12 months?' Categorical variables can also yield more than one possible response; for example, 'On which days of the week are you most likely to use public transport?'

Numerical variables yield numerical responses, such as your height in centimetres. Other examples are 'How many times during this visit did you travel by ferry?' (from the Hong Kong Airport survey) or the response to the question, 'How many messages did you send on social media last week?'

There are two types of numerical variables: discrete and continuous. **Discrete variables** produce numerical responses that arise from a counting process. 'The number of social media messages sent' is an example of a discrete numerical variable because the response is one of a finite number of integers. You send zero, one, two, …, 50 and so on messages.

Continuous variables produce numerical responses that arise from a measuring process. Your height is an example of a continuous numerical variable because the response takes on any value within a continuum or interval, depending on the precision of the measuring instrument. For example, your height may be 158 cm, 158.3 cm or 158.2945 cm, depending on the precision of the available instruments.

No two people are exactly the same height, and the more precise the measuring device used, the greater the likelihood of detecting differences in their heights. However, most measuring devices are not sophisticated enough to detect small differences. Hence, *tied observations* are often found in experimental or survey data even though the variable is truly continuous and, theoretically, all values of a continuous variable are different.

Levels of Measurement and Types of Measurement Scales

Data are also described in terms of their level of measurement. There are four widely recognised levels of measurement: nominal, ordinal, interval and ratio scales.

Nominal and ordinal scales

nominal scale
A classification of categorical data that implies no ranking.

Data from a categorical variable are measured on a nominal scale or on an ordinal scale. A **nominal scale** (Figure 1.2) classifies data into various distinct categories in which no ranking is implied. In the Hong Kong Airport survey, the answer to the question 'Are you likely to return to

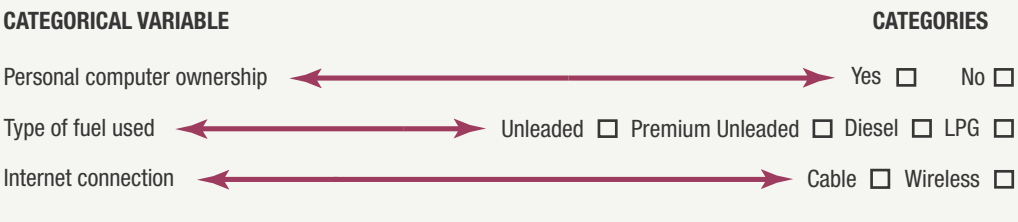

Figure 1.2 Examples of nominal scaling

Hong Kong in the next 12 months?' is an example of a nominally scaled variable, as is your favourite soft drink, your political party affiliation and your gender. Nominal scaling is the weakest form of measurement because you cannot specify any ranking across the various categories.

An **ordinal scale** classifies data into distinct categories in which ranking is implied. In the Hong Kong Airport survey, the answers to the question 'Shopping in Hong Kong stores gives good value for money' represent an ordinal scaled variable because the responses 'almost always, sometimes, very infrequently and never' are ranked in order of frequency. Figure 1.3 lists other examples of ordinal scaled variables.

ordinal scale
Scale of measurement where values are assigned by ranking.

Figure 1.3 Examples of ordinal scaling

Ordinal scaling is a stronger form of measurement than nominal scaling because an observed value classified into one category possesses more or less of a property than does an observed value classified into another category. However, ordinal scaling is still a relatively weak form of measurement because the scale does not account for the amount of the differences *between* the categories. The ordering implies only *which* category is 'greater', 'better' or 'more preferred' – not by *how much*.

Interval and ratio scales

Data from a numerical variable are measured on an interval or ratio scale. An **interval scale** (Figure 1.4) is an ordered scale in which the difference between measurements is a meaningful quantity but does not involve a true zero point. For example, sports shoes for adults are often sold in Australia marked with sizes based on the US or UK system. Neither system has a true zero size. The size below an adult size 1 is a child's size 13. However, in each system the intervals between sizes are equal.

interval scale
A ranking of numerical data where differences are meaningful but there is no true zero point.

Figure 1.4
Examples of interval and ratio scales

A **ratio scale** is an ordered scale in which the difference between the measurements involves a true zero point, as in length, weight, age or salary measurements, and the ratio of two values is meaningful. In the Hong Kong Airport survey, the number of times a visitor travelled by ferry is an example of a ratio scaled variable, as six trips is three times as many as two trips. As another example, a carton that weighs 40 kg is twice as heavy as one that weighs 20 kg.

ratio scale
A ranking where the differences between measurements involve a true zero point.

Data measured on an interval scale or on a ratio scale constitute the highest levels of measurement. They are stronger forms of measurement than an ordinal scale, because you can determine not only which observed value is the largest but also by how much. Interval and ratio scales may apply for either discrete or continuous data.

Telephone polling

think about this

Companies such as Newspoll regularly undertake market research and political polling conducted by phone interviews. A phone poll conducted by Newspoll in Sydney in November 2014 asked questions about a number of topics. Some were demographic questions about the number of people who lived in the household and the age, income, occupation and marital status of the participant. What would be the purpose of asking such questions?

The other questions could be divided into three sections. The first section related to voting intentions for the next state election and the level of satisfaction with the premier and the opposition leader. The second section asked the participant's opinion on the renewal of the federal government's ban on super trawlers. The third section asked a number of questions about domestic and international air travel undertaken in the past year. These questions covered areas such as the purpose of travel, the airlines used and level of satisfaction.

Who would use the data collected in this poll? If you were designing a similar poll, how would you construct questions to collect data for the variables referred to above?

More recently, political and business functions of Newspoll have been separated. To see how results of the latest political polls are published in the Australian, go to <www.theaustralian.com.au/national-affairs/newspoll>. To see some public opinion poll reports, go to <www.omnipoll.com.au>.

Problems for Section 1.2

LEARNING THE BASICS

1.1 Three different types of drinks are sold at a fast-food restaurant – soft drinks, fruit juices and coffee.
- **a.** Explain why the type of drinks sold is an example of a categorical variable.
- **b.** Explain why the type of drinks sold is an example of a nominally scaled variable.

1.2 Coffee is sold in three sizes in takeaway cardboard cups – small, medium and large. Explain why the size of the coffee cup is an example of an ordinal scaled variable.

1.3 Suppose that you measure the time it takes to download an MP3 file from the Internet.
- **a.** Explain why the download time is a numerical variable.
- **b.** Explain why the download time is a ratio scaled variable.

APPLYING THE CONCEPTS

1.4 For each of the following variables, determine whether the variable is categorical or numerical. If the variable is numerical, determine whether the variable is discrete or continuous. In addition, determine the level of measurement.
- **a.** Number of mobile phones per household
- **b.** Length (in minutes) of the longest mobile call made per month
- **c.** Whether all mobile phones in the household use the same telecommunications provider
- **d.** Whether there is a landline telephone in the household

1.5 The following information is collected from students as they leave the campus bookshop during the first week of classes:
- **a.** Amount of time spent shopping in the bookshop
- **b.** Number of textbooks purchased
- **c.** Name of degree
- **d.** Gender

Classify each of these variables as categorical or numerical. If the variable is numerical, determine whether the variable is discrete or continuous. In addition, determine the level of measurement.

1.6 For each of the following variables, determine whether the variable is categorical or numerical. If the variable is numerical, determine whether the variable is discrete or continuous. In addition, determine the level of measurement.
- **a.** Name of Internet provider
- **b.** Amount of time spent surfing the Internet per week
- **c.** Number of emails received per week
- **d.** Number of online purchases made per month

1.7 Suppose the following information is collected from Andrew and Fiona Chen on their application for a home loan mortgage at Metro Home Loans:
- **a.** Monthly expenses: $2,056
- **b.** Number of dependants being supported by applicant(s): 2
- **c.** Annual family salary income: $105,000
- **d.** Marital status: Married

Classify each of the responses by type of data and level of measurement.

1.8 One of the variables most often included in surveys is income. Sometimes the question is phrased, 'What is your income (in thousands of dollars)?' In other surveys, the respondent is asked to 'Place an X in the circle corresponding to your income group' and given a number of ranges to choose from.
- **a.** In the first format, explain why income might be considered either discrete or continuous.
- **b.** Which of these two formats would you prefer to use if you were conducting a survey? Why?
- **c.** Which of these two formats would probably bring you a greater rate of response? Why?

1.9 The director of research at the e-business section of a major department store wants to conduct a survey throughout a Australia to determine the amount of time working women spend shopping online for clothing in a typical month.
- **a.** Describe the population and the sample of interest, and indicate the type of data the director might wish to collect.
- **b.** Develop a first draft of the questionnaire needed in (a) by writing a series of three categorical questions and three numerical questions that you feel would be appropriate for this survey.

1.10 A university researcher designs an experiment to see how generous participants will be in giving to charity. Discuss the types of variables the experiment might give compared with a survey of the same subjects about donations to charity.

1.11 Before a company undertakes an online marketing campaign it needs to consider information about its own current sales and the sales made by its competitors. What categorical data might it use?

1.3 COLLECTING DATA

LEARNING OBJECTIVE 2
Identify how statistics is used in business

In the Hong Kong Airport scenario, identifying the data that need to be collected is an important step in the process of marketing the city and operational planning. Some of the data will come from consumers through market research. It is important that the correct inferences are drawn from the research and that appropriate statistical methods assist planners and designers to make the right decisions.

Managing a business effectively requires collecting the appropriate data. In most cases, the data are measurements acquired from items in a sample. The samples are chosen from populations in such a manner that the sample is as representative of the population as possible. The most common technique to ensure proper representation is to use a random sample. (See section 1.4 for a detailed discussion of sampling techniques.)

Many different types of circumstances require the collection of data:
- A marketing research analyst needs to assess the effectiveness of a new television advertisement.
- A pharmaceutical manufacturer needs to determine whether a new drug is more effective than those currently in use.
- An operations manager wants to monitor a manufacturing process to find out whether the quality of output being produced is conforming to company standards.
- An auditor wants to review the financial transactions of a company to determine whether or not the company is in compliance with generally accepted accounting principles.
- A potential investor wants to determine which firms within which industries are likely to have accelerated growth in a period of economic recovery.

Identifying Sources of Data

LEARNING OBJECTIVE 3
Recognise the sources of data used in business

Identifying the most appropriate source of data is a critical aspect of statistical analysis. If biases, ambiguities or other types of errors flaw the data being collected, even the most sophisticated statistical methods will not produce accurate information. Five important sources of data are:
- data distributed by an organisation or an individual
- a designed experiment
- a survey
- an observational study
- data collected by ongoing business activities.

primary sources
Provide information collected by the data analyser.

secondary sources
Provide data collected by another person or organisation.

Data sources are classified as either **primary sources** or **secondary sources**. When the data collector is the one using the data for analysis, the source is primary. When another organisation or

individual has collected the data that are used for analysis by an organisation or individual, the source is secondary.

Organisations and individuals that collect and publish data typically use this information as a primary source and then let others use the data as a secondary source. For example, the Australian federal government collects and distributes data in this way for both public and private purposes. The Australian Bureau of Statistics oversees a variety of ongoing data collection in areas such as population, the labour force, energy, and the environment and health care, and publishes statistical reports. The Reserve Bank of Australia collects and publishes data on exchange rates, interest rates and ATM and credit card transactions.

Market research firms and trade associations also distribute data pertaining to specific industries or markets. Investment services such as Morningstar provide financial data on a company-by-company basis. Syndicated services such as Nielsen provide clients with data enabling the comparison of client products with those of their competitors. Daily newspapers in print and online formats are filled with numerical information about share prices, weather conditions and sports statistics.

As listed above, conducting an experiment is another important data-collection source. For example, to test the effectiveness of laundry detergent, an experimenter determines which brands in the study are more effective in cleaning soiled clothes by actually washing dirty laundry instead of asking customers which brand they believe to be more effective. Proper experimental designs are usually the subject matter of more advanced texts, because they often involve sophisticated statistical procedures. However, some fundamental experimental design concepts are considered in Chapter 11.

Conducting a survey is a third important data source. Here, the people being surveyed are asked questions about their beliefs, attitudes, behaviours and other characteristics. Responses are then edited, coded and tabulated for analysis.

Conducting an observational study is the fourth important data source. In such a study, a researcher observes the behaviour directly, usually in its natural setting. Observational studies take many forms in business. One example is the **focus group**, a market research tool that is used to elicit unstructured responses to open-ended questions. In a focus group, a moderator leads the discussion and all the participants respond to the questions asked. Other, more structured types of studies involve group dynamics and consensus building and use various organisational-behaviour tools such as brainstorming, the Delphi technique and the nominal-group method. Observational study techniques are also used in situations in which enhancing teamwork or improving the quality of products and services are management goals.

focus group
A group of people who are asked about attitudes and opinions for qualitative research.

Data collected through ongoing business activities are a fifth data source. Such data can be collected from operational and transactional systems that exist in both physical 'bricks-and-mortar' and online settings but can also be gathered from secondary sources such as third-party social media networks and online apps and website services that collect tracking and usage data. For example, a bank might analyse a decade's worth of financial transaction data to identify patterns of fraud, and a marketer might use tracking data to determine the effectiveness of a website.

'Big Data'

Relatively recent advances in information technology allow businesses to collect, process, and analyse very large volumes of data. Because the operational definition of 'very large' can be partially dependent on the context of a business – what might be 'very large' for a sole proprietorship might be commonplace and small for a multinational corporation – many use the term *big data*.

big data
Large data sets characterised by their volume, velocity and variety.

Big data is more of a fuzzy concept than a term with a precise operational definition, but it implies data that are being collected in huge volumes and at very fast rates (typically in real time) and data that arrive in a variety of forms, both organised and unorganised. These attributes of 'volume, velocity, and variety', first identified in 2001 (see reference 1), make big data different from any of the data sets used in this book.

Big data increases the use of business analytics because the sheer size of these very large data sets makes preliminary exploration of the data using older techniques impracticable. This effect is explored in Chapter 20.

Big data tends to draw on a mix of primary and secondary sources. For example, a retailer interested in increasing sales might mine Facebook and Twitter accounts to identify sentiment about certain products or to pinpoint top influencers and then match those data to its own data collected during customer transactions.

Data Formatting

The data you collect may be formatted in more than one way. For example, suppose that you wanted to collect electronic financial data about a sample of companies. The data you seek to collect could be formatted in any number of ways, including:

- tables of data
- contents of standard forms
- a continuous data stream
- messages delivered from social media websites and networks.

These examples illustrate that data can exist in either a *structured* or an *unstructured* form. **Structured data** are data that follow some organising principle or plan, typically a repeating pattern. For example, a simple ASX share price search record is structured because each entry would have the name of a company, the last sale, change in price, bid price, volume traded, and so on. Due to their inherent organisation, tables and forms are also structured. In a table, each row contains a set of values for the same columns (i.e. variables), and in a set of forms, each form contains the same set of entries. For example, once we identify that the second column of a table or the second entry on a form contains the family name of an individual, then we know that all entries in the second column of the table or all of the second entries in all copies of the form contain the family name of an individual.

structured data
Data that follow an organised pattern.

In contrast, **unstructured data** follows no repeating pattern. For example, if five different people sent you an email message concerning the share trades of a specific company, that data could be anywhere in the message. You could not reliably count on the name of the company being the first words of each message (as in the ASX search), and the pricing, volume and percentage of change data could appear in any order. Earlier in this section, *big data* was defined, in part, as data that arrive in a variety of forms, both organised and unorganised. You can restate that definition as '*big data* exists as both structured and unstructured data'.

unstructured data
Data that have no repeated pattern.

The ability to handle unstructured data represents an advance in information technology. Chapter 20 discusses business analytics methods that can analyse structured data as well as unstructured data or *semi-structured* data. (Think of an application form that contains structured form-fills but also contains an unstructured free-response portion.)

With the exception of some of the methods discussed in Chapter 20, the methods taught and the software techniques used in this book involve structured data. Your beginning point will always be tabular data, and for many problems and examples you can begin with that data in the form of a Microsoft Excel worksheet that you can download and use (see companion website).

Electronic formats and **encoding** need to be considered. Data can exist in more than one electronic format. This affects data formatting, as some electronic formats are more immediately usable than others. For example, which data would you like to use: data in an electronic worksheet file or data in a scanned image file that contains one of the worksheet illustrations in this book? Unless you like to do extra work, you would choose the first format because the second would require you to employ a translation process – perhaps a character-scanning program that can recognise numbers in an image.

electronic formats
Data in a form that can be read by a computer.

encoding
Representing data by numbers or symbols to convert the data into a usable form.

Data can also be *encoded* in more than one way, as you may have learned in an information systems course. Different encodings can affect the precision of values for numerical variables, and that can make some data not fully compatible with other data you have collected.

Data Cleaning

No matter how you choose to collect data, you may find irregularities in the values you collect, such as undefined or impossible values. For a categorical variable, an undefined value would be

a value that does not represent one of the categories defined for the variable. For a numerical variable, an impossible value would be a value that falls outside a defined range of possible values for the variable. For a numerical variable without a defined range of possible values, you might also find **outliers**, values that seem excessively different from most of the rest of the values. Such values may or may not be errors, but they demand a second review.

outliers
Values that appear to be excessively large or small compared with most values observed.

missing values
Refers to when no data value is stored for one or more variables in an observation.

Missing values are another type of irregularity. They are values that were not able to be collected (and therefore are not available for analysis). For example, you would record a nonresponse to a survey question as a missing value. You can represent missing values in some computer programs and such values will be properly excluded from analysis. The more limited Excel has no special values that represent a missing value. When using Excel, you must find and then exclude missing values manually.

When you spot an irregularity, you may have to 'clean' the data you have collected. A full discussion of data cleaning is beyond the scope of this book. (See reference 2 for more information.)

Recoding Variables

After you have collected data, you may discover that you need to reconsider the categories that you have defined for a categorical variable, or that you need to transform a numerical variable into a categorical variable by assigning the individual numeric data values to one of several groups. In either case, you can define a **recoded variable** that supplements or replaces the original variable in your analysis. For example, when defining households by their location, the suburb or town recorded might be replaced by a new variable of the postcode.

recoded variable
A variable that has been assigned new values that replace the original ones.

When recoding variables, be sure that the category definitions cause each data value to be placed in one and only one category, a property known as being **mutually exclusive**. Also ensure that the set of categories you create for the new, recoded variables include all the data values being recoded, a property known as being **collectively exhaustive**. If you are recoding a categorical variable, you can preserve one or more of the original categories, as long as your recoded values are both mutually exclusive and collectively exhaustive.

mutually exclusive
Two events that cannot occur simultaneously.

collectively exhaustive
Set of events such that one of the events must occur.

When recoding numerical variables, pay particular attention to the operational definitions of the categories you create for the recoded variable, especially if the categories are not self-defining ranges. For example, while the recoded categories 'Under 12', '12–20', '21–34', '35–59' and '60 and over' are self-defining for age, the categories 'Child', 'Youth', 'Young adult', 'Middle aged' and 'Senior' need their own operational definitions.

Problems for Section 1.3

APPLYING THE CONCEPTS

1.12 The Data and Story Library (DASL) is an online library of data files and stories that illustrate the use of basic statistical methods. Visit <http://.lib.stat.cmu.edu/DASL>, click Power search, and explore a datafile of interest to you. Which of the five sources of data best describes the sources of the datafile you selected?

1.13 Visit the website of Ipsos Australia at <www.ipsos.com.au>. Read about a recent poll or news story. What type of data source is this based on?

1.14 Visit the website of the Pew Research Center at <www.pewresearch.org>. Read one of today's top stories. What type of data source is the story based on?

1.15 Transportation engineers and planners want to address the dynamic properties of travel behaviour by describing in detail the driving characteristics of drivers over the course of a month. What type of data collection source do you think the transportation engineers and planners should use?

1.16 Visit the homepage of the Statistics Portal 'Statista' at <www.statista.com>. Go to Statistics>Popular Statistics, then choose one item to examine. What type of data source is the information presented here based on?

1.4 TYPES OF SURVEY SAMPLING METHODS

LEARNING OBJECTIVE 4

Distinguish between different survey sampling methods

In Section 1.1 a sample was defined as the portion of the population that has been selected for analysis. You collect your data from either a *population* or a *sample* depending on whether all items or people about whom you wish to reach conclusions are included. Rather than taking a complete census of the whole population, statistical sampling procedures focus on collecting a small representative group of the larger population. The resulting sample results are used to estimate characteristics of the entire population. The three main reasons for drawing a sample are:

1. A sample is less time-consuming than a census.
2. A sample is less costly to administer than a census.
3. A sample is less cumbersome and more practical to administer than a census.

The sampling process begins by defining the frame. The **frame** is a listing of items that make up the population. Frames are data sources such as population lists, directories or maps. Samples are drawn from these frames. Inaccurate or biased results can occur if the frame excludes certain groups of the population. Using different frames to generate data can lead to opposite conclusions.

frame
A list of the items in the population of interest.

Once you select a frame, you draw a sample from the frame. As illustrated in Figure 1.5, there are two kinds of samples: the non-probability sample and the probability sample.

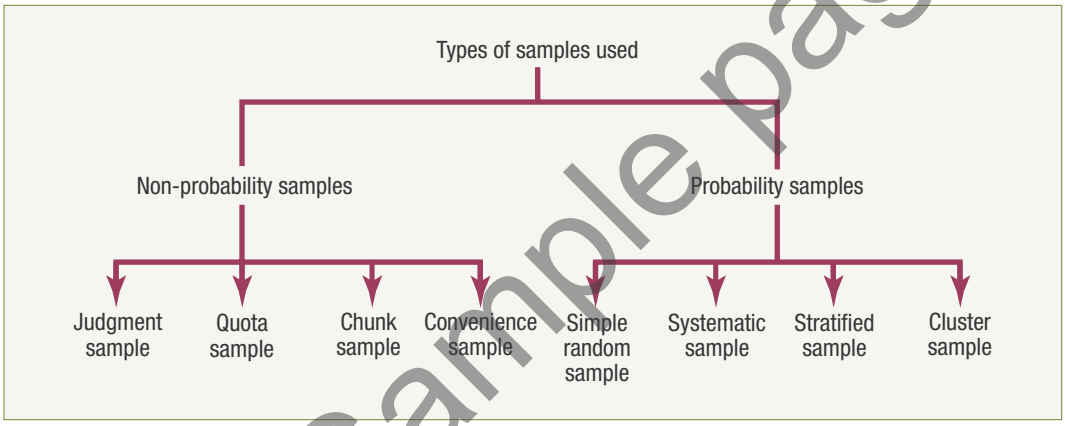

Figure 1.5
Types of samples

In a **non-probability sample**, you select the items or individuals without knowing their probabilities of selection. Thus, the theory that has been developed for probability sampling cannot be applied to non-probability samples. A common type of non-probability sampling is convenience sampling. In **convenience sampling**, items are selected based only on the fact that they are easy, inexpensive or convenient to sample. In some cases, participants are self-selected. For example, many companies conduct surveys by giving visitors to their website the opportunity to complete survey forms and submit them electronically. The response to these surveys can provide large amounts of data quickly, but the sample consists of self-selected web users. For many studies, only a non-probability sample such as a judgment sample is available. In a **judgment sample**, you get the opinions of preselected experts in the subject matter as to who should be included in the survey. Some other common procedures of non-probability sampling are quota sampling and chunk sampling. These are discussed in detail in specialised books on sampling methods (see references 3 and 4).

non-probability sample
One where selection is not based on known probabilities.

convenience sampling
Selection using a method that is easy or inexpensive.

judgment sample
Gives the opinions of preselected experts.

Non-probability samples can have certain advantages such as convenience, speed and lower cost. However, their lack of accuracy due to selection bias and their poorer capacity to provide generalised results more than offset these advantages. Therefore, you should restrict the use of non-probability sampling methods to situations in which you want to get rough

approximations at low cost to satisfy your curiosity about a particular subject, or to small-scale studies that precede more rigorous investigations.

probability sample
One where selection is based on known probabilities.

In a **probability sample**, you select the items based on known probabilities. Whenever possible, you should use probability sampling methods. The samples based on these methods allow you to make unbiased inferences about the population of interest. In practice, it is often difficult or impossible to take a probability sample. However, you should work towards achieving a probability sample and acknowledge any potential biases that might exist. The four types of probability samples most commonly used are simple random, systematic, stratified and cluster. These sampling methods vary in their cost, accuracy and complexity.

Simple Random Sample

simple random sample
One where each item in the frame has an equal chance of being selected.

In a **simple random sample**, every item from a frame has the same chance of selection as every other item. In addition, every sample of a fixed size has the same chance of selection as every other sample of that size. Simple random sampling is the most elementary random sampling technique. It forms the basis for the other random sampling techniques.

With simple random sampling, you use n to represent the sample size and N to represent the frame size. You number every item in the frame from 1 to N. The chance that you will select any particular member of the frame on the first draw is $1/N$.

sampling with replacement
An item in the frame can be selected more than once.

You select samples with replacement or without replacement. **Sampling with replacement** means that after you select an item you return it to the frame, where it has the same probability of being selected again. Imagine you have a barrel which contains the shopping dockets of N shoppers at a major retail centre who are entering a competition. First assume that each shopper can have only one entry but can win more than one prize. The barrel is rolled, opened and the entry of Jason O'Brien is selected. His docket is replaced, the barrel is rolled again and a second docket is chosen. Jason's docket has the same probability of being selected again, $1/N$. You repeat this process until you have selected the desired sample size n. However, it is usually more desirable to have a sample of different items than to permit a repetition of measurements on the same item.

sampling without replacement
Each item in the frame can be selected only once.

Sampling without replacement means that once you select an item it cannot be selected again. The chance that you will select any particular item in the frame, say the shopping docket of Jason O'Brien on the first draw is $1/N$. The chance that you will select any shopping docket not previously selected on the second draw is now 1 out of $N - 1$. This process continues until you have selected the desired sample of size n.

Regardless of whether you have sampled with or without replacement, barrel draw methods have a major drawback for sample selection. In a crowded barrel, it is difficult to mix the entries thoroughly and ensure that the sample is selected randomly. As barrel draw methods are not very useful, you need to use less cumbersome and more scientific methods of selection.

table of random numbers
Shows a list of numbers generated in a random sequence.

One such method uses a **table of random numbers** (see Table E.1 in Appendix E of this book) for selecting the sample. A table of random numbers consists of a series of digits listed in a randomly generated sequence (see reference 5). Because the numeric system uses 10 digits (0, 1, 2, …, 9), the chance that you will randomly generate any particular digit is equal to the probability of generating any other digit. This probability is 1 out of 10. Hence, if a sequence of 800 digits is generated, you would expect about 80 of them to be the digit 0, 80 to be the digit 1, and so on. In fact, those who use tables of random numbers usually test the generated digits for randomness prior to using them. Table E.1 has met all such criteria for randomness. Because every digit or sequence of digits in the table is random, the table can be read either horizontally or vertically. The margins of the table designate row numbers and

column numbers. The digits themselves are grouped into sequences of five in order to make reading the table easier.

To use such a table instead of a barrel for selecting the sample, you first need to assign code numbers to the individual members of the frame. Then you get the random sample by reading the table of random numbers and selecting those individuals from the frame whose assigned code numbers match the digits found in the table. Example 1.1 demonstrates the process of sample selection.

EXAMPLE 1.1

SELECTING A SIMPLE RANDOM SAMPLE USING A TABLE OF RANDOM NUMBERS

A company wants to select a sample of 32 full-time workers from a population of 800 full-time employees in order to collect information on expenditures concerning a company-sponsored dental plan. How do you select a simple random sample?

SOLUTION

The company can contact all employees by email but assumes that not everyone will respond to the survey, so you need to distribute more than 32 surveys to get the desired 32 responses. Assuming that 8 out of 10 full-time workers will respond to such a survey (i.e. a response rate of 80%), you decide to email 40 surveys.

The frame consists of a listing of the names and email addresses of all $N = 800$ full-time employees taken from the company personnel files. Thus, the frame is an accurate and complete listing of the population. To select the random sample of 40 employees from this frame, you use a table of random numbers, as shown in Table 1.2 on page 20. Because the population size (800) is a three-digit number, each assigned code number must also be three digits so that every full-time worker has an equal chance of selection. You give a code of 001 to the first full-time employee in the population listing, a code of 002 to the second full-time employee in the population listing, and so on, until a code of 800 is given to the Nth full-time worker in the listing. Because $N = 800$ is the largest possible coded value, you discard all three-digit code sequences greater than N (i.e. 801 to 999 and 000).

To select the simple random sample, you choose an arbitrary starting point from the table of random numbers. One method you can use is to close your eyes and strike the table of random numbers with a pencil. Suppose you use this procedure and select row 06, column 05, of Table 1.2 (which is extracted from Table E.1) as the starting point. Although you can go in any direction, in this example you will read the table from left to right in sequences of three digits without skipping.

The individual with code number 003 is the first full-time employee in the sample (row 06 and columns 05–07), the second individual has code number 364 (row 06 and columns 08–10) and the third individual has code number 884. Because the highest code for any employee is 800, you discard this number. Individuals with code numbers 720, 433, 463, 363, 109, 592, 470 and 705 are selected third to tenth, respectively.

You continue the selection process until you get the needed sample size of 40 full-time employees. During the selection process, if any three-digit coded sequence is repeated, you include the employee corresponding to that coded sequence again as part of the sample, if sampling with replacement. You discard the repeating coded sequence if sampling without replacement.

Table 1.2
Using a table of random numbers

Source: Data from the Rand Corporation, from *A Million Random Digits with 100,000 Normal Deviates* (Glencoe, IL: The Free Press, 1955) (displayed in Table E.1 in Appendix E of this book).

		Column							
		00000	00001	11111	11112	22222	22223	33333	33334
	Row	12345	67890	12345	67890	12345	67890	12345	67890
	01	49280	88924	35779	00283	81163	07275	89863	02348
	02	61870	41657	07468	08612	98083	97349	20775	45091
	03	43898	65923	25078	86129	78496	97653	91550	08078
	04	62993	93912	30454	84598	56095	20664	12872	64647
	05	33850	58555	51438	85507	71865	79488	76783	31708
Begin	06	97340	03364	88472	04334	63919	36394	11095	92470
selection	07	70543	29776	10087	10072	55980	64688	68239	20461
(row 06,	08	89382	93809	00796	95945	34101	81277	66090	88872
column 5)	09	37818	72142	67140	50785	22380	16703	53362	44940
	10	60430	22834	14130	96593	23298	56203	92671	15925
	11	82975	66158	84731	19436	55790	69229	28661	13675
	12	39087	71938	40355	54324	08401	26299	49420	59208
	13	55700	24586	93247	32596	11865	63397	44251	43189
	14	14756	23997	78643	75912	83832	32768	18928	57070
	15	32166	53251	70654	92827	63491	04233	33825	69662
	16	23236	73751	31888	81718	06546	83246	47651	04877
	17	45794	26926	15130	82455	78305	55058	52551	47182
	18	09893	20505	14225	68514	46427	56788	96297	78822
	19	54382	74598	91499	14523	68479	27686	46162	83554
	20	94750	89923	37089	20048	80336	94598	26940	36858
	21	70297	34135	53140	33340	42050	82341	44104	82949
	22	85157	47954	32979	26575	57600	40881	12250	73742
	23	11100	02340	12860	74697	96644	89439	28707	25815
	24	36871	50775	30592	57143	17381	68856	25853	35041
	25	23913	48357	63308	16090	51690	54607	72407	55538

Systematic Sample

systematic sample
A method that involves selecting the first element randomly then choosing every *k*th element thereafter.

In a **systematic sample**, you partition the N items in the frame into n groups of k items where:

$$k = \frac{N}{n}$$

You round k to the nearest integer. To select a systematic sample, you choose the first item to be selected at random from the first k items in the frame. Then you select the remaining $n - 1$ items by taking every kth item thereafter from the entire frame.

If the frame consists of a listing of prenumbered cheques, sales receipts or invoices, a systematic sample is faster and easier to take than a simple random sample. A systematic sample is also a convenient mechanism for collecting data from telephone directories, class rosters and consecutive items coming off an assembly line.

To take a systematic sample of $n = 40$ from the population of $N = 800$ employees, you partition the frame of 800 into 40 groups, each of which contains 20 employees. You then select a random number from the first 20 individuals, and include every 20th individual after the first selection in the sample. For example, if the first number you select is 008, your subsequent selections are 028, 048, 068, 088, 108, … , 768 and 788.

Although they are simpler to use, simple random sampling and systematic sampling are generally less efficient than other, more sophisticated probability sampling methods. Even greater possibilities for selection bias and lack of representation of the population characteristics occur from systematic samples than from simple random samples. If there is a pattern in the

frame, you could have severe selection biases. To overcome the potential problem of disproportionate representation of specific groups in a sample, you can use either stratified sampling methods or cluster sampling methods.

Stratified Sample

In a **stratified sample**, you first subdivide the N items in the frame into separate subpopulations, or **strata**. A stratum is defined by some common characteristic. You select a simple random sample, in proportion to the size of the strata, and combine the results from the separate simple random samples. This method is more efficient than either simple random sampling or systematic sampling because you are assured of the representation of items across the entire population. The homogeneity of items within each stratum provides greater precision in the estimates of underlying population parameters.

stratified sample
Items randomly selected from each of several populations or strata.

strata
Subpopulations composed of items with similar characteristics in a stratified sampling design.

SELECTING A STRATIFIED SAMPLE

EXAMPLE 1.2

A company wants to select a sample of 32 full-time workers from a population of 800 full-time employees in order to estimate expenditures from a company-sponsored dental plan. Of the full-time employees, 25% are managerial and 75% are non-managerial workers. How do you select the stratified sample so that the sample will represent the correct proportion of managerial workers?

SOLUTION

If you assume an 80% response rate, you need to distribute 40 surveys to get the desired 32 responses. The frame consists of a listing of the names and company email addresses of all $N = 800$ full-time employees included in the company personnel files. Since 25% of the full-time employees are managerial, you first separate the population frame into two strata: a subpopulation listing of all 200 managerial-level personnel and a separate subpopulation listing of all 600 full-time non-managerial workers. Since the first stratum consists of a listing of 200 managers, you assign three-digit code numbers from 001 to 200. Since the second stratum contains a listing of 600 non-managerial-level workers, you assign three-digit code numbers from 001 to 600.

To collect a stratified sample proportional to the sizes of the strata, you select 25% of the overall sample from the first stratum and 75% of the overall sample from the second stratum. You take two separate simple random samples, each of which is based on a distinct random starting point from a table of random numbers (Table E.1). In the first sample you select 10 managers from the listing of 200 in the first stratum, and in the second sample you select 30 non-managerial workers from the listing of 600 in the second stratum. You then combine the results to reflect the composition of the entire company.

Cluster Sample

In a **cluster sample**, you divide the N items in the frame into several clusters so that each cluster is representative of the entire population. You then take a random sample of clusters and study all items in each selected cluster. **Clusters** are naturally occurring designations, such as postcode areas, electorates, city blocks, households or sales territories.

Cluster sampling is often more cost-effective than simple random sampling, particularly if the population is spread over a wide geographical region. However, cluster sampling often requires a larger sample size to produce results as precise as those from simple random sampling or stratified sampling. A detailed discussion of systematic sampling, stratified sampling and cluster sampling procedures can be found in references 3, 4 and 6.

cluster sample
The frame is divided into representative groups (or clusters), then all items in randomly selected clusters are chosen.

cluster
A naturally occurring grouping, such as a geographical area.

Problems for Section 1.4

LEARNING THE BASICS

1.17 For a population containing $N = 902$ individuals, what code number would you assign for:
- **a.** the first person on the list?
- **b.** the fortieth person on the list?
- **c.** the last person on the list?

1.18 For a population of $N = 902$, verify that, by starting in row 05 of the table of random numbers (Table E.1), you need only six rows to select a sample of $n = 60$ *without* replacement.

1.19 Given a population of $N = 93$, starting in row 29 of the table of random numbers (Table E.1) and reading across the row, select a sample of $n = 15$:
- **a.** *without* replacement
- **b.** *with* replacement

APPLYING THE CONCEPTS

1.20 For a study that consists of personal interviews with participants (rather than mail or phone surveys), explain why a simple random sample might be less practical than some other methods.

1.21 You want to select a random sample of $n = 1$ from a population of three items (called *A*, *B* and *C*). The rule for selecting the sample is: flip a coin; if it is heads, pick item *A*; if it is tails, flip the coin again; this time, if it is heads, choose *B*; if it is tails, choose *C*. Explain why this is a random sample but not a simple random sample.

1.22 A population has four members (call them *A*, *B*, *C* and *D*). You would like to draw a random sample of $n = 2$, which you decide to do in the following way: flip a coin; if it is heads, the sample will be items *A* and *B*; if it is tails, the sample will be items *C* and *D*. Although this is a random sample, it is not a simple random sample. Explain why. (If you did problem 1.21, compare the procedure described there with the procedure described in this problem.)

1.23 The town planning department of a Sydney council with a population of $N = 40{,}000$ registered voters is asked by the mayor to conduct a survey to measure community attitudes to urban consolidation. The table following contains a breakdown of the 40,000 registered voters by gender and ward of residence.

	Ward of residence				
Gender	**North**	**South**	**East**	**West**	**Total**
Female	7,000	5,200	5,000	4,800	22,000
Male	5,600	4,600	4,000	3,800	18,000
Total	12,600	9,800	9,000	8,600	40,000

The planning department intends to take a probability sample of $n = 2{,}000$ voters and project the results from the sample to the entire population of voters.
- **a.** If the frame available from the council files is an alphabetical listing of the names of all $N = 40{,}000$ registered voters, what type of sample could you take? Discuss.
- **b.** What is the advantage of selecting a simple random sample in (a)?
- **c.** What is the advantage of selecting a systematic sample in (a)?
- **d.** If the frame available from the council's files is a listing of the names and addresses of all $N = 40{,}000$ registered voters, compiled from eight separate alphabetical lists based on the gender and address breakdowns shown in the ward-of-residence table, what type of sample should you take? Discuss.
- **e.** At present East Ward has many high-rise apartments, West Ward and South Ward have single dwellings only and North Ward has a mixture of low- and medium-density housing. What would be the danger in randomly choosing 40 street names and systematically sampling 50 of the residents of those streets?

1.24 Suppose that 5,000 sales invoices are separated into four strata. Stratum 1 contains 50 electrical invoices, stratum 2 contains 500 paint invoices, stratum 3 contains 1,000 plumbing supplies invoices and stratum 4 contains 3,450 hardware invoices. A sample of 500 sales invoices is needed.
- **a.** What type of sampling method should you use? Why?
- **b.** Explain how you would carry out the sampling according to the method stated in (a).
- **c.** Why is the sampling in (a) not simple random sampling?

LEARNING OBJECTIVE 5

Evaluate the quality of surveys

1.5 EVALUATING SURVEY WORTHINESS

Nearly every day you read or hear about survey or opinion poll results in newspapers, on the Internet or on radio or television. To identify surveys that lack objectivity or credibility, you must critically evaluate what you read and hear by examining the worthiness of the survey. First, you must evaluate the purpose of the survey, why it was conducted and for whom it was conducted. An opinion poll or survey conducted to satisfy curiosity is mainly for entertainment. Its result is an end in itself rather than a means to an end. You should be sceptical of such a survey because the result should not be put to further use.

The second step in evaluating the worthiness of a survey is for you to determine whether it was based on a probability or a non-probability sample (as discussed in Section 1.4). You need to remember that the only way to make correct statistical inferences from a sample to a population is through the use of a probability sample. Surveys that use non-probability sampling methods are subject to serious, perhaps unintentional, *bias* that may render the results meaningless.

Survey Errors

Even when surveys use random probability sampling methods, they are subject to potential errors. Four types of survey errors are:

- coverage error
- non-response error
- sampling error
- measurement error.

Good survey research design attempts to reduce or minimise these various survey errors, often at considerable cost.

Coverage error

The key to proper sample selection is an adequate frame. Remember, a frame is an up-to-date list of all the items from which you will select the sample. **Coverage error** occurs if certain groups of items are excluded from this frame so that they have no chance of being selected in the sample. Coverage error results in a *selection bias*. If the frame is inadequate because certain groups of items in the population were not properly included, any random probability sample selected will provide an estimate of the characteristics of the frame, not the *actual* population. Computer-based surveys are useful for certain studies where the subjects all have Internet access. Coverage error could result if the unemployed, the elderly or indigenous communities are not selected in the frame due to their lack of Internet or email access.

coverage error
Occurs when all items in a frame do not have an equal chance of being selected. This causes selection bias.

Non-response error

Not everyone is willing to respond to a survey. In fact, research has shown that individuals in the upper and lower socioeconomic classes tend to respond less frequently to surveys than people in the middle class. **Non-response error** arises from the failure to collect data on all items in the sample and results in a *non-response bias*. Because you cannot generally assume that people who do not respond to surveys are similar to those who do, you need to follow up on the non-responses after a specified period of time. You should make several attempts to persuade these individuals to complete the survey. The follow-up responses are then compared with the initial responses in order to make valid inferences from the survey (references 3, 4 and 6).

The mode of response you use affects the rate of response. The personal interview and the telephone interview usually produce a higher response rate than a mail survey – but at a higher cost.

non-response error
Occurs due to the failure to collect information on all items chosen for the sample; this causes non-response bias.

Sampling error

There are three main reasons for selecting a sample rather than taking a complete census. It is more expedient, less costly and more efficient. However, chance dictates which individuals or items will or will not be included in the sample. **Sampling error** reflects the heterogeneity, or 'chance differences', from sample to sample, based on the probability of certain individuals or items being selected in particular samples.

When you read about the results of surveys or polls in newspapers or magazines, there is often a statement regarding margin of error or precision; for example, 'the results of this poll are expected to be within ±4 percentage points of the actual value'. This margin of error is the sampling error. You can reduce sampling error by taking larger sample sizes, although this also increases the cost of conducting the survey.

sampling error
The difference in results for different samples of the same size.

think about this

The problem of online survey rigging

As the use of online methods for collecting information grows more prevalent we need to be aware that individuals will not all act honestly, especially when they have something to gain. There are many methods being used to contravene the rules of online competitions, such as paying companies to vote, setting up multiple email addresses or Facebook accounts, and using methods to mask the true IP address of the computer being used. Even if a small incentive is offered for completing a survey, similar problems can arise.

At an Australian university, students were recently asked to complete a survey about a peer-assisted learning program and were offered the chance to win movie tickets as an incentive to give feedback. The survey was carried out anonymously in order to elicit frank responses, but on completion students were automatically sent to a second site where they could register their student ID in order to enter a draw to win movie tickets. One student registered 105 times in order to increase the chance of winning the movie tickets. It is not clear how many times this person completed the survey itself.

How could this type of behaviour potentially affect survey results? What could you do to minimise this type of survey error if you were designing an online survey?

Measurement error

In the practice of good survey research, you design a questionnaire with the intention of gathering meaningful information. But you have a dilemma here – getting meaningful measurements is easier said than done. Consider the following proverb:

> *A man with one watch always knows what time it is.*
> *A man with two watches always searches to identify the correct one.*
> *A man with ten watches is always reminded of the difficulty in measuring time.*

Unfortunately, the process of getting a measurement is often governed by what is convenient, not what is needed. The measurements are often only a proxy for the ones you really desire. Much attention has been given to measurement error that occurs because of a weakness in question wording (reference 6). A question should be clear, not ambiguous. And, to avoid *leading questions*, you need to present them in a neutral manner.

measurement error
The difference between survey results and the true value of what is being measured.

There are three sources of **measurement error**: ambiguous wording of questions, the halo effect and respondent error. The Australian Bureau of Statistics is very conscious of minimising error caused by questionnaire design and survey operations. For the National Health Survey in 2010–11 it used Computer Assisted Interview techniques to collect information. It states:

> the CAI instrument allows:
> - data to be captured electronically at the point of interview, which obviates the cost, logistical, timing and quality issues associated with transport, storage and security of paper forms, and transcription/data entry of information from forms into electronic format
> - the ability to use complex sequencing to define specific populations for questions, and ensure word substitutes used in the questions were appropriate to each respondent's characteristics and prior responses
> - the ability, through data validation (edits), to check responses entered against previous responses, reduce data entry errors by interviewers, and enable seemingly inconsistent responses to be clarified with respondents at the time of interview. The audit trail recorded in the instrument also provides valuable information about the operation of particular questions, and associated data quality issues. (Australian Bureau of Statistics, *Australian Health Survey: Users' Guide, 2011–2013*, electronic publication, Cat. No. 4363.0.55.001, 2013)

The *halo effect* occurs when the respondent feels obligated to please the interviewer. Proper interviewer training can minimise the halo effect. *Respondent error* occurs as a result of overzealous or underzealous effort by the respondent. You can minimise this error in two ways: (1) by carefully scrutinising the data and calling back those individuals whose responses seem unusual, and (2) by establishing a program of random call-backs to determine the reliability of the responses.

Other sources of error besides measurement error can result from clerical or recording errors. See references 7, 8 and 9 for a more detailed discussion of measurement error and the difficulties of avoiding it.

Ethical Issues

Ethical considerations arise with respect to the four types of potential errors that can occur when designing surveys that use probability samples: coverage error, non-response error, sampling error and measurement error. Coverage error can result in selection bias and becomes an ethical issue if particular groups or individuals are *purposely* excluded from the frame so that the survey results are skewed, indicating a position more favourable to the survey's sponsor. Non-response error can lead to non-response bias and becomes an ethical issue if the sponsor knowingly designs the survey in such a manner that particular groups or individuals are less likely to respond. Sampling error becomes an ethical issue if the findings are purposely presented without reference to sample size and margin of error, so that the sponsor can promote a viewpoint that might otherwise be truly insignificant. Measurement error becomes an ethical issue in one of three ways: (1) a survey sponsor chooses leading questions that guide the responses in a particular direction; (2) an interviewer, through mannerisms and tone, purposely creates a halo effect or otherwise guides the responses in a particular direction; (3) a respondent, having a disdain for the survey process, wilfully provides false information.

Ethical issues also arise when the results of non-probability samples are used to form conclusions about the entire population. When you use a non-probability sampling method, you need to explain the sampling procedures and state that the results cannot be generalised beyond the sample.

Problems for Section 1.5

APPLYING THE CONCEPTS

1.25 'A survey indicates that the vast majority of university students own their own personal computer.' What information would you want to know before you accepted the results of this survey?

1.26 A simple random sample of $n = 300$ full-time employees is selected from a company list containing the names of all $N = 5{,}000$ full-time employees in order to evaluate job satisfaction.

a. Give an example of possible coverage error.
b. Give an example of possible non-response error.
c. Give an example of possible sampling error.
d. Give an example of possible measurement error.

1.27 According to a recent cyber security report, 'millennials remain the most common victims of cybercrime, with 40 percent having experienced cybercrime in the past year'. Reasons given for this include slack online security habits and password sharing (*2016 Norton Cyber Security Insights Report*, <www.symantec.com/content/dam/symantec/docs/reports/2016-norton-cyber-security-insights-report.pdf>, accessed 16 June 2017). What information would you want to know before you accepted the results of the survey?

1.28 Kiribati is a small, poor Pacific nation under threat from global warming. According to the CIA *World Factbook*, Kiribati comprises a group of 33 coral atolls in the Pacific Ocean straddling the equator, with elevations varying from 0 to 81 metres above sea level. The low level of some of the islands makes them sensitive to changes in sea level (Central Intelligence Agency, *The World Factbook*, <www.cia.gov/library/publications/the-world-factbook/geos/kr.html> accessed 16 June 2017). Suppose that an environmental economist has seen results from a survey which claims that 30% of inhabitants of Kiribati are already affected by roads having been permanently cut by rising seawater. What information would she want to know before accepting the results of the survey?

1.29 Reality TV shows have incorporated surveys of audience opinion into their formats. In Australia several shows have allowed the audience to vote on whether contestants should remain on the show or be excluded. Consider a show where voting is by SMS, premium rate phone call, Facebook or another online site, and viewers are limited to 10 votes using each method. Compare this type of survey with a random poll of viewers without replacement conducted by phone for the TV show.

a. How might the results differ?

b. What are the costs and benefits for the owners of the show for each voting method?

1.30 The online restaurant search site Dimmi <www.dimmi.com.au> encourages diners to rate restaurants they have been to by giving them reward points which can be accumulated until a meal discount is available. A restaurant at The Rocks in Sydney has been rated as follows: Recommended 8.7; Food 8.5; Service 8.7; Value for money 7.8; Atmosphere 8.4. What differences could arise from this type of survey compared with ratings derived from a random sample of diners?

1.6 THE GROWTH OF STATISTICS AND INFORMATION TECHNOLOGY

During the past century, statistics has played an important role in spurring the use of information technology and, in turn, such technology has spurred the wider use of statistics. At the beginning of the twentieth century, the expanding data-handling requirements associated with the United States Federal Census led directly to the development of tabulating machines that were the forerunners of today's business computer systems. Statisticians such as Pearson, Fisher, Gosset, Neyman, Wald and Tukey established the techniques of modern inferential statistics as an alternative to analysing large sets of population data that had become increasingly costly, time-consuming and cumbersome to collect. The development of early computer systems permitted others to develop computer programs to ease the calculation and data-processing burdens imposed by those techniques. Over time, greater use of statistical methods by business decision makers and advances in computer capacity have led to the development of even more sophisticated statistical methods.

Today, when you hear of retailers investing in a 'customer-relationship management system', or CRM, or a packaged goods producer engaging in 'data mining' to uncover consumer preferences, you should realise that statistical techniques form the foundations of such cutting-edge applications of information technology. As global information storage increases dramatically, businesses are rapidly coming to terms with how to analyse big data – data sets so large and varied that conventional software cannot readily handle them. (Think of the huge volume of data produced each day by people using Visa, Facebook, eBay and Twitter.) Even though cutting-edge applications might require custom programming, for many years businesses have had access to **statistical packages** such as Minitab, SPSS/PASW Statistics, SAS and Stata – standardised sets of programs that help managers use a wide range of statistical techniques by automating the data processing and calculations these techniques require.

statistical packages
Computer programs designed to perform statistical analysis.

The leasing and training costs associated with statistical packages have led many to consider using some of the graphical and statistical functions of Microsoft Excel. However, you need to be aware that many statisticians have concerns about the accuracy and completeness of the statistical results produced by early versions of Excel. Invalid results could be produced, especially when the data sets were very large or had unusual statistical properties (see reference 10). Microsoft Excel 2010 and subsequent versions made some significant improvements in statistical functions (see references 11 and 12) but it would still be wise to be careful about the data and the analysis you are undertaking.

Assess your progress 1

Summary

In this chapter you have studied data collection and the various types of data used in business. In the Hong Kong International Airport scenario you were asked to review the visitor survey which will be used to provide information to the tourism authority planning staff (see page 9). Three of the questions shown will produce numerical data and four will produce categorical data. The responses to the first question (number of previous visits to Hong Kong) are discrete, and the responses to the second question (length of time since last visit) are continuous. After the data have been collected, they must be organised and prepared in order to make various analyses. You have also learned about commonly used sampling methods and ways to prepare data for analysis such as encoding, cleaning and recoding. The next two chapters develop tables and charts and a variety of descriptive numerical measures that are useful for data analysis.

Key terms

References

1. Laney, D., *3D Data Management: Controlling Data Volume, Velocity, and Variety* (Stamford, CT: META Group. February 6, 2001).
2. Osbourne, J. *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data* (Thousand Oaks, CA: Sage Publications, 2013).
3. Cochran, W. G., *Sampling Techniques*, 3rd edn (New York: Wiley, 1977).
4. Lohr, S. L., *Sampling Design and Analysis*, 2nd edn (Boston, MA: Brooks/Cole Cengage Learning, 2010).
5. Rand Corporation, *A Million Random Digits with 100,000 Normal Deviates* (Glencoe, IL: The Free Press, 1955).
6. Groves R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer & R. Tourangeau, *Survey Methodology*, 2nd edn (New York: John Wiley, 2009).
7. Sudman, S., N. M. Bradburn & N. Schwarz. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology* (San Francisco, CA: Jossey-Bass, 1996).
8. Biemer, P. B., R. M. Graves, L. E. Lyberg, A. Mathiowetz & S. Sudman, *Measurement Errors in Survey* (New York: Wiley Interscience, 2004).
9. Fowler, F. J., *Improving Survey Questions: Design and Evaluation*, *Applied Special Research Methods Series*, Vol. 38 (Thousand Oaks, CA: Sage Publications, 1995).

10. McCullough, B. D. & B. Wilson, 'On the accuracy of statistical procedures in Microsoft Excel 97', *Computational Statistics and Data Analysis*, 31 (1999): 27–37.

11. Microsoft Corporation at <http://office.microsoft.com/en-au/excel-help/what-s-new-changes-made-to-excel-functions-HA010355760.aspx>, accessed June 2017.

12. Microsoft Corporation at <http://office.microsoft.com/en-001/excel-help/new-functions-in-excel-2013-HA103980604.aspx>, accessed June 2017.

Chapter review problems

CHECKING YOUR UNDERSTANDING

1.31 What is the difference between a sample and a population?

1.32 What is the difference between a statistic and a parameter?

1.33 What is the difference between descriptive and inferential statistics?

1.34 What is the difference between a categorical and a numerical variable?

1.35 What is the difference between a discrete and a continuous variable?

1.36 What is an operational definition and why is it so important?

1.37 What are the four types of measurement scales?

1.38 What are some potential problems with using 'barrel draw' methods to select a simple random sample?

1.39 What is the difference between sampling *with* replacement and sampling *without* replacement?

1.40 What is the difference between a simple random sample and a systematic sample?

1.41 What is the difference between a simple random sample and a stratified sample?

1.42 What is the difference between a stratified sample and a cluster sample?

APPLYING THE CONCEPTS

1.43 The Australasian Data and Story Library OZDASL <www.statsci.org/data> is an online library of data files and stories that illustrate the use of basic statistical methods. The stories are classified by method and by topic. Go to this site and click on 'First Course in Statistics'. Pick a story and summarise how statistics were used in the story.

1.44 Make a list of six ways you have used or encountered statistics in the past week. Think about what you read or heard in a news report or saw on a commercial website. Also think whether you made a bet or participated in a survey.

1.45 The Australian Bureau of Statistics <www.abs.gov.au> site contains survey information on people, business, geography and other topics. Go to the site and find the latest version of *Labour Force, Australia* (Cat. No. 6202.0).

- **a.** Briefly describe the Labour Force survey.
- **b.** Give an example of a categorical variable found in this survey.
- **c.** Give an example of a numerical variable found in this survey.
- **d.** Is the variable you selected in (c) discrete or continuous?

1.46 The Australian Bureau of Statistics website allows users to access a large amount of Census data online. Go to <www.abs.gov.au/census> and in the Data by Products section click on the latest Census year, enter a location and search for QuickStats.

- **a.** Give an example of a categorical variable found in this summary of survey results.
- **b.** Give an example of a numerical variable found in this summary of survey results.
- **c.** Is the variable you selected in (b) discrete or continuous?

1.47 Detailed information on airport and airline on-time performance can be found at <www.flightstats.com>. Explore the departures performance data for different airports and regions.

- **a.** Which of the five types of data sources listed in Section 1.3 do you think were used here?
- **b.** Name a categorical variable for which observations were collected.
- **c.** Name a numerical variable for which observations were collected.
- **d.** What type of recoding has been used here and why?

1.48 Late in 2016 the National Roads and Motorists' Association (NRMA), a major Australian motoring organisation, released results of a survey that sought to check members' attitudes to traffic congestion and a motorway extension (see <www.mynrma.com.au/about/media/local-support-for-SouthConnex-strengthens-nrma-survey.htm>).

- **a.** Describe the population(s) for this survey.
- **b.** Describe the sample(s) for this survey.
- **c.** Can you identify potential difficulties in comparing these results with results from a similar 2005 survey?

1.49 A manufacturer of flavoured milk is planning to survey households in Tasmania to determine the purchasing habits of consumers. Among the questions to be included are those that relate to:

1. where flavoured milk is primarily purchased
2. what flavour of milk is purchased most often
3. how many people living in the household drink flavoured milk
4. the total number of millilitres of flavoured milk drunk in the past week by members of the household

- **a.** Describe the population.
- **b.** For each of the four items listed, indicate whether the variable is categorical or numerical. If numerical, is it discrete or continuous?
- **c.** Develop five categorical questions for the survey.
- **d.** Develop five numerical questions for the survey.

1.50 A new bus network is proposed for a north-eastern Sydney region. A survey is sent out to residents asking questions which relate to:

1. the resident's age
2. frequency of bus use
3. usual ticket type purchased
4. main purpose of using the bus

- **a.** Describe the population.
- **b.** Indicate whether each of the questions above is categorical or numerical.

c. Develop two more numerical questions and state whether the variables are discrete or continuous.
d. Develop two more categorical questions.

1.51 Political polling has traditionally used telephone interviews. Researchers at a polling organisation argue that Internet polling is less expensive and faster, and offers higher response rates than telephone surveys. Critics are concerned about the scientific reliability of this approach. Even amid this strong criticism, Internet polling is becoming more and more common. What concerns, if any, do you have about Internet polling?

1.52 Statistics New Zealand mentions a number of possible sources of non-sampling error in economic surveys in *A Guide to Good Survey Design*, 3rd edition, which can be downloaded from <www.stats.govt.nz>.
a. Which of the four types of survey error from Section 1.5 are identified on this site as a non-sampling error?
b. Discuss which errors would be more difficult to eliminate.

1.53 Researchers at a university wish to conduct a survey of past students to ascertain how frequently they are using statistical techniques in the workforce. The researchers have permission from the ethics committee to use the last recorded email and postal addresses to contact ex-students, but these may be out of date, particularly as many students have returned to homes overseas without updating their records. The emails and letters are sent out simultaneously. The response to the survey is low.
a. What type of errors or biases should the researchers be especially concerned with?
b. What step(s) should the researchers take to try to overcome the problems noted in (a)?
c. What could have been done differently to improve the survey's worthiness?

1.54 According to a survey conducted by the Australian Interactive Media Industry Association, 77% of mobile phone users surveyed pay by a monthly phone bill compared to 21% who are on pre-paid plans. The percentage of respondents that have data included in their payment plans is 84% (M. M. Mackay, *Australian Mobile Phone Lifestyle Index*, 9th edn, October 2013, <www.aimia.com.au/ampli>, accessed 24 January 2014).
a. What other information would you want to know before you accepted the results of this survey?
b. Suppose that you wished to conduct a similar survey for the geographic region you live in. Describe the population for your survey.
c. Explain how you could minimise the chance of a coverage error in this type of survey.
d. Explain how you could minimise the chance of a nonresponse error in this type of survey.
e. Explain how you could minimise the chance of a sampling error in this type of survey.
f. Explain how you could minimise the chance of a measurement error in this type of survey.

Continuing cases

Tasman University

Tasman University's Tasman Business School (TBU) regularly surveys business students on a number of issues. In particular, students within the school are asked to complete a student survey when they receive their grades each semester. The results of Bachelor of Business (BBus) students who responded to the latest undergraduate (UG) survey are stored in **< TASMAN_UNIVERSITY_BBUS_STUDENT_SURVEY >**.

a For each question asked in the survey, determine whether the variable is categorical or numerical. If you determine that the variable is numerical, identify whether it is discrete or continuous.

b A separate survey has been carried out for Master of Business Administration (MBA) students. Results for these postgraduate (PG) students are in the file **< TASMAN_UNIVERSITY_MBA_STUDENT_SURVEY >**. Repeat the analysis you carried out in (a) for the postgraduate survey results.

As Safe as Houses

To analyse the real estate market in non-capital cities and towns in states A and B, Safe-As-Houses Real Estate, a large national real estate company, has collected samples of recent residential sales from a sample of non-capital cities and towns in these states. The data are stored in **< REAL_ESTATE >**.

a Identify data sources and discuss the type of sampling that was most likely used to collect these data.

b Suggest any additional variables that could be collected in order to explain property prices, and determine if they are numerical or categorical, discrete or continuous.

Chapter 1 Excel Guide

EG1.1 GETTING STARTED WITH MICROSOFT EXCEL

Microsoft Excel is the electronic worksheet program of Microsoft Office. Although not a specialised statistical program, Excel contains basic statistical functions, and the Excel 2016 PC and Mac versions include Data Analysis Toolpak procedures that you can use to perform selected advanced statistical methods. To use the Data Analysis Toolpak you must select it as an Excel add-in. You can also install the PHStat add-in (available for separate purchase or with some textbooks) to extend and enhance the Data Analysis Toolpak that Microsoft Excel contains. (You do not need to use PHStat in order to use Microsoft Excel with this text, although using PHStat will simplify using Excel for statistical analysis.)

In Microsoft Excel, you create or open and save files that are called **workbooks**. Workbooks are collections of worksheets and related items, such as charts, that contain the original data as well as the calculations and results associated with one or more analyses. Because of its widespread distribution, Microsoft Excel is a convenient program to use, but some statisticians have expressed concern about its lack of fully reliable and accurate results for some statistical procedures. Although Microsoft has recently improved many statistical functions, especially from Excel 2010 onwards, you should be somewhat cautious about using Microsoft Excel to perform analyses on data other than the data used in this text. (If you plan to install PHStat, make sure you first read Appendix F and any PHStat read-me file.)

You can use Excel to learn and apply the statistical methods discussed in this book and as an aid in solving end-of-section and end-of-chapter problems. For many topics, you may choose to use the 'Excel How-to' instructions. These instructions use pre-constructed worksheets as models or **templates** for a statistical solution. You learn how to adapt these worksheets to construct your own solutions. Many of these sections feature a specific Excel Guide workbook that contains worksheets that are *identical* to the worksheets that PHStat creates. Because both of these methods create the same results and the same worksheets, you can use a combination of them as you read through this book.

The 'Excel How-to' instructions and the Excel Guide workbooks work best with the latest Versions of Microsoft Excel, including Excel 2016 and Excel 2013 (Microsoft Windows), Excel 2016 for Mac, and Office 365. (Excel Guides also contain instructions for using the Analysis ToolPak add-in that is included with most of the latest Microsoft Excel versions.) (*Microsoft Excel 2016*, Microsoft Corporation, 2015)

You will want to master the basic skills listed in Table EG1.1 before you begin using Microsoft Excel to understand statistical concepts and solve problems. If you plan to use the 'Excel How-to' instructions, you will also need to master the skills listed in the lower part of

Table EG1.1
Basic skills for using Microsoft Excel

Excel skill	Specifics
Excel data entry	• Organising worksheet data in columns • Entering numerical and categorical data
File operations	• Open • Save • Print
Worksheet operations	• Create • Copy and paste
Formula skills	• Concept of a formula • Cell references • Absolute and relative cell references • How to enter a formula • How to enter an array formula
Workbook presentation	• How to apply format changes that affect the display of worksheet cell contents
Chart formatting correction	• How to correct the formatting of charts that Excel improperly creates
Discrete histogram creation	• How to create a properly formatted histogram for a discrete probability distribution

Operation	Examples	Notes
Keyboard keys	• **Enter** • **Ctrl** • **Shift**	Names of keys are always the object of the verb *press*, as in 'press Enter'.
Keystroke combinations	• **Ctrl+C** • **Ctrl+Shift+Enter** • **Command+Enter**	Keyboarding actions that require you to press more than one key at the same time. ***Ctrl+C*** means press **C** while holding down **Ctrl**. **Ctrl+Shift+Enter** means press **Enter** while holding down both **Ctrl** and **Shift**.
Click or select operations	• **Click** OK • **Select** the first 2-D Bar gallery item	Mouse pointer actions that require you to single click an onscreen object. This book uses the verb *select* when the object is either a worksheet cell or an item in a gallery, menu, list or Ribbon tab.
Menu or ribbon selection	• **File → New** • **Layout → Legend → None**	A sequence of Ribbon or menu selections. **File → New** means first select the **File** tab and then select **New** from the list that appears.
Placeholder object	• ***variable 1 cell range*** • ***bins cell range***	An italicised bold-faced phrase is a placeholder for an object reference. In making entries, you enter the reference (e.g. **A1:A10**) and not the placeholder.

Table EG1.2
Excel typographic conventions

the table. While you do not necessarily need these skills if you plan to use PHStat, knowing them will be useful if you expect to customise the Excel worksheets that PHStat creates or expect to be using Excel beyond the course that uses this book.

The list of skills in Table EG1.1 begins with the more basic skills and progresses towards slightly more advanced skills that you will need to use less frequently.

Table EG1.2 presents the typographic conventions that the Excel Guides in this book use to present computer operations.

EG1.2 OPENING AND SAVING WORKBOOKS

Once you open the Excel program a new workbook will be displayed where you can begin entering data in rows and columns. Figure EG1.1 shows a newly opened workbook in Excel 2016. It contains the elements that are common with most Microsoft Windows programs.

If you wish to use a workbook created previously you will need to use the following commands.

If you are using Microsoft Excel 2016, select **File → Open**.

In the Backstage view you will be given a choice of selecting from Recent Workbooks, OneDrive or the Computer. You can browse, select the file to be opened and then click on the **OK** button. If you cannot find your file, you may need to do one or more of the following:

- Use the scroll bars or the slider, if present, to scroll through the entire list of files.
- Select the correct folder from the drop-down list at the left-hand side of the dialog box.
- To search every file in the folder, leave **All Files** showing at the bottom of the dialog box. If you want a specific type of file such as text files, use the arrow to open a drop-down menu and then select **Text Files**.

In Excel 2016, select **File → Save As**, and in the Backstage view choose the location. In the dialog box enter (or edit) the name of the file in the **File name** box and click on the **OK** button. If applicable, you can also do the following:

- Change to another folder by selecting that folder from the **Save in** drop-down list.
- Change the **Save as type** value to something other than the default choice, **Microsoft Excel Workbook**. **Text (Tab delimited)** or **CSV (Comma delimited)** are two file types sometimes used to share Excel data with other programs.

After saving your work, you should consider saving your file a second time, using a different name, to create a backup copy of your work. Read-only files cannot be saved to their original folders unless the name is changed.

EG1.3 ENTERING DATA

The main worksheet area is composed of rows and columns that you use for data entry. You enter data into the rows and columns of a worksheet. By convention, and the style used

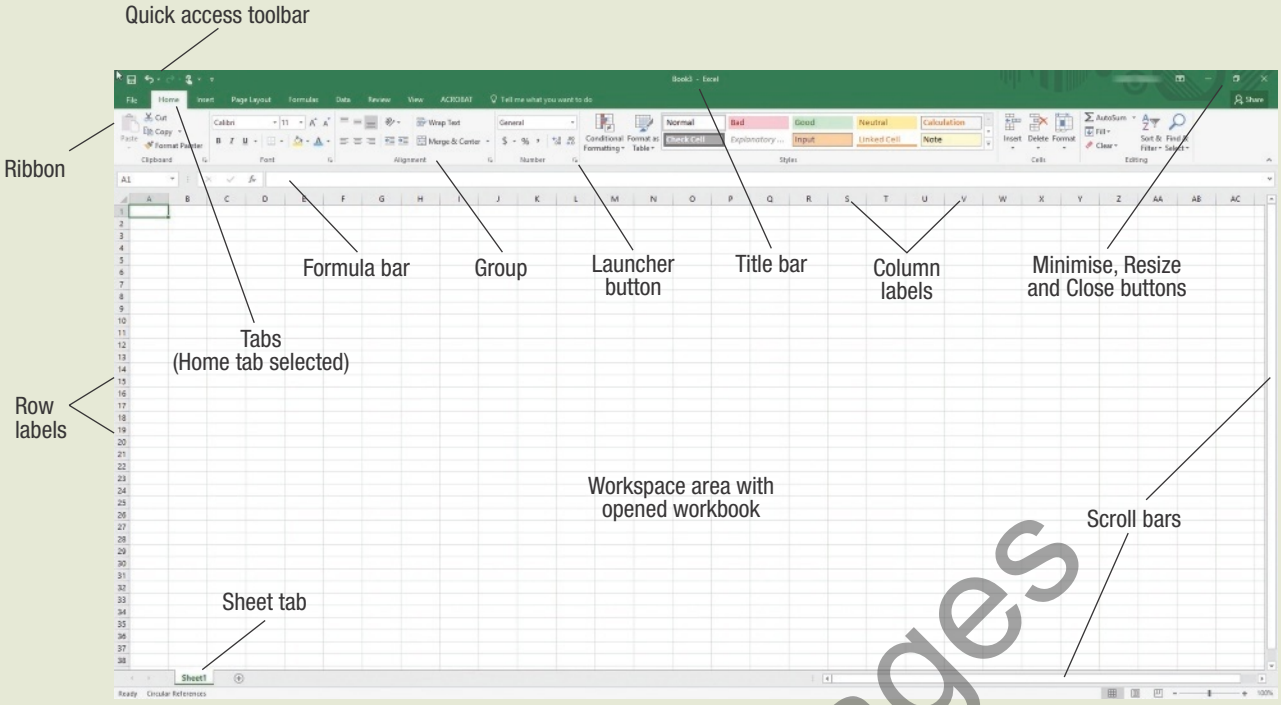

Figure EG1.1
The Excel 2016 window

in this book, when you enter data for a set of variables you enter the name of each variable into the cells of the first row, beginning with column A. Then you enter the data for the variable in the subsequent rows to create a DATA worksheet similar to the one shown in Figure EG1.2, which contains data from an auction sale. Note that the formula used in the active cell F6 can be seen on the formula bar.

To enter data in a specific cell, either use the cursor keys to move the cell pointer to the cell or use your mouse to select the cell directly. As you type, what you type appears in the formula bar. Complete your data entry by pressing **Tab** or **Enter** or by clicking the checkmark button in the formula bar.

When you enter data, never skip any rows in a column and, as a general rule, avoid skipping any columns. Also try to avoid using numbers as row 1 variable headings; if you cannot avoid their use, precede such headings with apostrophes. Pay attention to any special instructions that occur throughout the book for the order of the entry of your data. For some statistical methods, entering your data in an order that Excel does not expect will lead to incorrect results.

To refer to a specific entry, or cell, you use a *Sheetname!ColumnRow* notation. For example, Data!A2 refers to the cell in column A and row 2 in the Data worksheet. To refer to a specific group or **range** of cells, you use a *Sheetname!Upperleftcell:Lowerrightcell* notation. For example, Data!A2:B11 refers to the 20 cells that are in rows 2 to 11 in columns A and B of the Data worksheet. An absolute address for the cell A6 is shown as A6. Even if a formula using this address is copied to another row or column it will still refer to this cell. However, if the formula is written with the relative address A6, moving the formula will change the

F6 =E6*0.25

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Catalogue No.	Price estimate	Auction bids	Sold	Sale price	Buyer's Premium	Total						
2	BG001	$2,300	5	Yes	$2,600	$650.00	$3,250.00						
3	BG002	$800	10	Yes	$1,000	$250.00	$1,250.00						
4	BG003	$4,300	3	Passed in	n.a.	n.a	n.a						
5	BG004	$1,500	4	Yes	$1,380	$345.00	$1,725.00						
6	BG005	$725	6	Yes	$930	$232.50	$1,162.50						
7													
8													

Figure EG1.2
An example of a DATA worksheet

reference cell. Both absolute and relative addresses may be necessary in one sheet depending on the operations intended. Also note that $A6 freezes the column but not the row and A$6 freezes the row but allows the column to change.

Each Microsoft Excel worksheet has its own name. Automatically, Microsoft Excel names worksheets in the form of **Sheet1**, **Sheet2** and so on. You should rename your worksheets, giving them more self-descriptive names, by double-clicking on the sheet tabs that appear at the bottom of each sheet, typing a new name and pressing the Enter key.

EG1.4 USING FORMULAS IN EXCEL WORKSHEETS

Formulas are worksheet cell entries that perform a calculation or some other task. You enter formulas by typing the equals sign symbol (=) followed by some combination of mathematical or other data-processing operations.

For simple formulas, you use the symbols +, −, *, / and ^ for the operations addition, subtraction, multiplication, division and exponentiation (a number raised to a power), respectively. For example, the formula

=Data!B2 + Data!B3 + Data!B4 + Data!B5

adds the contents of the cells B2, B3, B4 and B5 of the Data worksheet and displays the sum as the value in the cell containing the formula. You can also use Microsoft Excel *functions* in formulas to simplify formulas. To find lists of the functions that can be selected in Excel, click on the f_x Function Wizard symbol on the Formula bar. For example, the formula =SUM(Data!B2:B5), using the Excel SUM() function, is a shorter equivalent to the formula above.

You can also use cell or cell range references that do not contain the *Sheetname!* part, such as B2 or B2:B5. Such references *always* refer to the worksheet in which the formula has been entered.

Formulas allow you to create generalised solutions and give Excel its distinctive ability to recalculate results automatically when you change the values of the supporting data. Typically, when you use a worksheet, you see only the results of any formulas entered, not the formulas themselves. However, for your reference, many illustrations of Microsoft Excel worksheets in this text also show the underlying formulas adjacent to the results they produce. When using Excel 2016, select **Formulas ➔ Formula Auditing ➔ Show Formulas** to see onscreen the formulas themselves and not their results. To restore the original view, click on **Show Formulas** again.

EG1.5 CREATING CHARTS

The method of creating charts can vary according to the version of Excel you are using. Both these methods are available in Excel 2016.

- **Method 1** A feature in Excel 2016 allows you to create charts easily using the **Quick Analysis** tool. Simply highlight an area of the spreadsheet containing some data you wish to graph by clicking on the top left-hand cell, then dragging the mouse. The range may contain labels. Click on the small box that appears in the bottom right-hand corner to open **Quick Analysis**. Select **Charts**, then, by hovering the mouse over the different chart types, you can see previews of recommended charts for the selected data. You can also choose **More**, which will open a dialog box with a more extensive range of options. Once a chart is selected there are several ways you can modify it by clicking on the icons that appear on its right-hand side. These are **Chart Elements** (+), **Chart Styles** (paintbrush) and **Chart Filters** (filter). You will also now see that multiple design options are shown on the ribbon and that options to change colours or chart type are shown there. By right-clicking on the background area of the chart you can also activate a drop-down menu. If you choose **Format Chart Area** a menu will open on the right-hand side of the spreadsheet that allows you to change the format of the chart and text in many ways. If instead you choose **Move Chart** you can choose a new location on another sheet. To reposition the chart on the existing sheet, simply click on it and drag. To resize it, drag using one of the circles on its border.
- **Method 2** Highlight the area of the spreadsheet with your data as described above. If you wish to select areas that are not adjacent, hold down the Ctrl key while selecting. The area selected must be rectangular. Click on the **Insert** tab, then from the Charts area click on the **Recommended Charts** and select a particular format from the drop-down gallery. Alternatively, you can select a chart type from the icons shown. Once the chart is created it can be formatted or enhanced by clicking on it and following the instructions given for Method 1.

Figure EG1.3 shows an example of a chart created in Excel 2016 with the Format Axis panel open.

EG1.6 PRINTING WORKBOOKS

Before printing you may select a print area if you do not want the whole sheet printed. To print Excel 2016 worksheets, select **File ➔ Print**. A print preview is automatically created, as can be seen in Figure EG1.4. Various print settings are available in the drop-down list boxes. Clicking on **Page Setup** will give access to more choices such as changing from **Portrait** to **Landscape** orientation, as would suit the worksheet shown. When you are satisfied with the settings and look of the preview, click on the Print button.

Note that if you want only a part of the worksheet to be printed it is easier to set this using **Page Layout** tab then **Page setup ➔ Print area**.

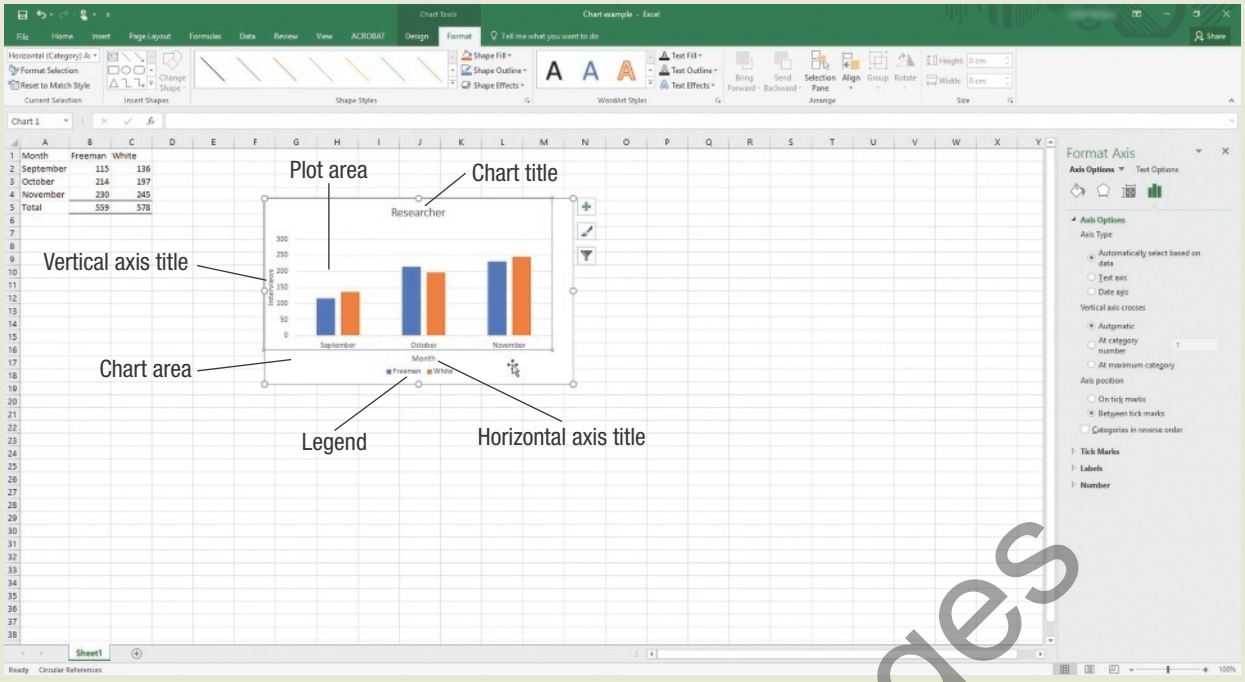

Figure EG1.3
An example of a chart created in Excel 2016 with the Format Axis panel open

Page Setup allows you to customise printing to change the print orientation, add gridlines and so on before printing. Once you are satisfied with the results, click on the **Print** button in the print preview window, then **OK** in the **Print** dialog box.

The Print Backstage view (see Figure EG1.4) contains settings to select the printer to be used, what parts of the workbook to print (the active worksheet is the default) and the number of copies to produce (1 is the default). If you need to change these settings, change them before clicking on the **OK** button.

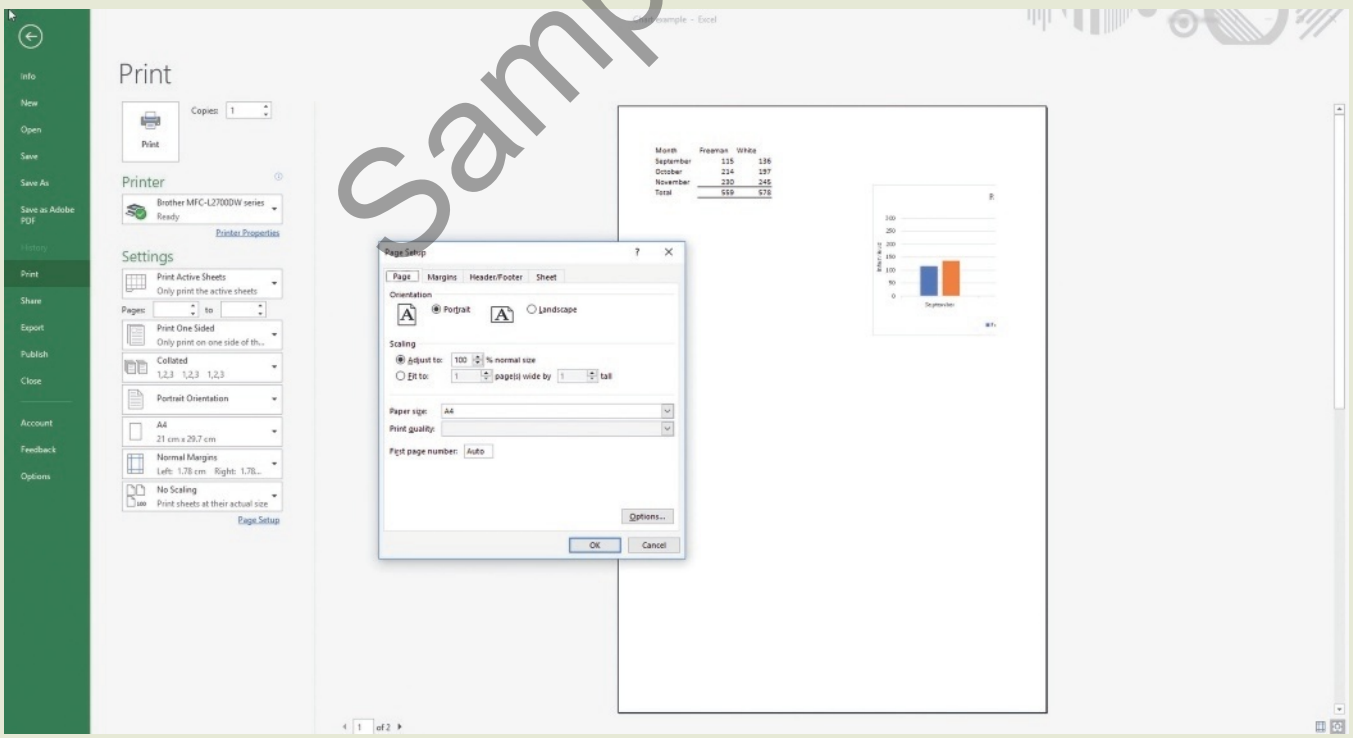

Figure EG1.4
The Excel 2016 Backstage view with Print and Page Setup selected

After printing, you should verify the contents of your printout. Most printing failures will trigger the display of an error message that you can use to work out the source of the failure.

EG1.7 HOW USING EXCEL FOR MAC DIFFERS

Excel 2016 for Mac comes with the add-ins for Analysis Toolpack but earlier versions did not. If you don't have a current version, it is possible to download software made by third-party companies to perform some of the same statistical analysis tasks. The free program StatPlus®:mac LE, for instance, will allow you to run a regression, calculate descriptive statistics and run analysis of variance tests. Further capability is available in the Pro edition at a cost.

In Excel 2016 for Mac you can open a new workbook when the program opens by using New ➔ Blank Workbook ➔ Create.

The easiest way to save a new workbook is to click on the quick access toolbar file icon to **Save**. A **Save As** dialog box will allow you to choose a file name, a location for the file and the file format. You can also choose File ➔ Save to begin this process.

To create a chart in Excel 2016 for Mac, use Method 2 described in section EG1.5. With the chart selected click on the **Chart Design** tab. You will find that extra tabs such as **Add Chart Element, Quick Layout** and **Switch Row/Column** open on the ribbon to allow more formatting.

To print a worksheet or selection use **File ➔ Print** then on the **Printer** select the printer you wish to use. The default is that all active worksheets will be printed but to modify that select **Show Details**. Then choose the option preferred from the drop-down menu, and finally select **Print**.

EG1.8 DEFINING DATA

Establishing the Variable Type

Microsoft Excel infers the variable type from the data you enter into a column. If Excel discovers a column that contains numbers, for example, it treats the column as a numerical variable. If Excel discovers a column that contains words or alphanumeric entries, it treats the column as a non-numerical (categorical) variable.

This imperfect method works most of the time, especially if you make sure that the categories for your categorical variables are words or phrases such as 'yes' and 'no'. However, because you cannot explicitly define the variable type, Excel can mistakenly offer or allow you to do nonsensical things such as using a statistical method that is designed for numerical variables on categorical variables. If you must use coded values such as 1, 2 or 3, enter them preceded by an apostrophe, as Excel treats all values that begin with an apostrophe as non-numerical data. (You can check whether a cell entry includes a leading apostrophe by selecting a cell and viewing the contents of the cell in the formula bar.)

EG1.9 COLLECTING DATA

Recoding Variables

Key technique

To recode a categorical variable, you first copy the original variable's column of data and then use the find-and-replace function on the copied data. To recode a numerical variable, or a categorical variable with only two values, enter a formula that returns a recoded value in a new column.

Example

Imagine that we have collected data at an airport using a survey such as shown on page 9. The **Recode workbook** shows how the original variables of 'Accommodation satisfaction' and 'Business visit' have been recoded.

Excel how-to

Two recoded variables were created by first opening the **Airport Survey worksheet** in the **Recode workbook** and then following these steps:

1. Right-click column **B** (right-click over the shaded 'B' at the top of column B) and click **Copy** in the shortcut menu.
2. Right-click column **C** and click the **first choice** in the **Paste Options** gallery.
3. Enter **Accommodation code** in cell **C1**.
4. Select column **C**. With column C selected, click **Home ➔ Find & Select ➔ Replace**.

In the Replace tab of the Find and Replace dialog box:

5. Enter **Very satisfied as Find what, 1** as **Replace with**, and then click **Replace All**.
6. Click **OK** to close the dialog box that reports the results of the replacement command.
7. Still in the Find and Replace dialog box, enter **Very dissatisfied as Find what** (replacing **Very satisfied**), and **5** as **Replace with**, then click **Replace All**.
8. Click **OK** to close the dialog box that reports the results of the replacement command.
9. Continue to replace the words **Dissatisfied, Satisfied** and **Undecided** with the numbers **4, 2** and **3** respectively using this method. (This creates the recoded variable **Accommodation code** in column C.)
10. **Enter Business visit code** in cell **H1**.
11. Enter the formula **=IF(F2 = "No", 0,1)** in cell **H2**.
12. Copy this formula down the column to the last row that contains Visitor data (row 31). (This creates the recoded variable **Business visit code** in column H.) The **Recode workbook** uses the **IF** function to recode the two categories as numbers. Numerical variables can also be recoded into multiple categories by using a more advanced technique using the **VLOOKUP** function.

EG1.10 TYPES OF SAMPLING METHODS

Simple Random Sample

Key technique

Use the **RANDBETWEEN(*smallest integer*, *largest integer*)** function to generate a random integer that can then be used to select an item from a frame.

Example

Create a simple random sample *with* replacement of size 40 from a population of 800 items.

Excel how-to

Enter a formula that uses this function and then copy the formula down a column for as many rows as is necessary. For example, to create a simple random sample with replacement of size 40 from a population of 800 items, open to a new worksheet. Enter **Sample** in cell **A1** and enter the formula **=RANDBETWEEN(1, 800)** in cell **A2**. Then copy the formula down the column to cell **A41**.

Excel contains no functions to select a random sample *without* replacement. Such samples are most easily created using an add-in such as PHStat or the Analysis ToolPak, as described in the following paragraphs.

Analysis ToolPak

Use **Sampling** to create a random sample *with* replacement.

For the example, assume you have a worksheet that contains the population of 800 items in column A and that contains a column heading in cell A1. Select **Data → Data Analysis**. In the Data Analysis dialog box, select **Sampling** from the **Analysis Tools** list and then click **OK**. In the procedure's dialog box:

1. Enter **A1:A801** as the **Input Range** and check **Labels**.
2. Click **Random** and enter **40** as the **Number of Samples**.
3. Click **New Worksheet Ply** and then click **OK**.

Example

Create a simple random sample *without* replacement of size 40 from a population of 800 items.

PHStat

Use **Random Sample Generation**.

For the example, select **PHStat → Sampling → Random Sample Generation**. In the procedure's dialog box:

1. Enter **40** as the **Sample Size**.
2. Click **Generate list of random numbers** and enter **800** as the **Population Size**.
3. Enter a **Title** and click **OK**.

Unlike most other PHStat results worksheets, the worksheet created contains no formulas.

Excel how-to

Use the **COMPUTE worksheet** of the **Random workbook** as a template. The worksheet already contains 40 copies of the formula **=RANDBETWEEN(1, 800)** in column B. Because the **RANDBETWEEN** function samples *with* replacement as discussed at the start of this section, you may need to add additional copies of the formula in new column B rows until you have 40 unique values.

If your intended sample size is large, you may find it difficult to spot duplicates. See the **ADVANCED worksheet** in the **Random workbook** for more information about an advanced technique that uses formulas to detect duplicate values.