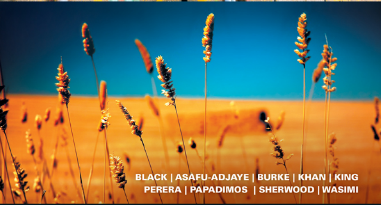
AUSTRALASIAN BUSINESS STATISTICS

4TH EDITION



BLACK | ASAFU-ADJAYE | BURKE | KHAN | KING
PERERA | PAPADIMOS | SHERWOOD | WASIMI

WILEY

AUSTRALASIAN BUSINESS STATISTICS

4TH EDITION

Ken BLACK
John ASAFU-ADJAYE
Paul BURKE
Nazim KHAN
Gerard KING
Nelson PERERA
Andrew PAPADIMOS
Carl SHERWOOD
Saleh WASIMI
Reetu VERMA

WILEY

Fourth edition published 2016 by
John Wiley & Sons Australia, Ltd
42 McDougall Street, Milton Qld 4064

First edition published 2007
Second edition published 2010
Third edition published 2013

Typeset in Minion Pro 10/12

Authorised adaptation of *Business statistics for contemporary decision making*, 4th edition (ISBN 0-471-42983-X), published by John Wiley & Sons, Inc., Hoboken, United States of America under exclusive licence from Leyh Publishing, LLC. Copyright © 2004 in the USA by Leyh Publishing, LLC. All rights reserved.

National Library of Australia
Cataloguing-in-Publication entry

Creator:	Black, Ken (Kenneth Urban), author.
Title:	Australasian business statistics / Ken Black, John Asafu-Adjaye, Paul Burke, Nazim Khan, Gerard King, Nelson Perera, Andrew Papadimos, Carl Sherwood, Saleh Wasimi.
Edition:	4th edition.
ISBN:	9780730312932 (paperback) 9780730324836 (e-text)
Notes:	Includes index.
Subjects:	Commercial statistics — Australasia — Textbooks.
Other Creators/ Contributors:	Asafu-Adjaye, John, author. Burke, Paul, author. Khan, Nazim, author. King, Gerard, author. Perera, Nelson, author. Papadimos, Andrew, author. Sherwood, Carl, author. Wasimi, Saleh A. (Saleh Ahmed), author.
Dewey Number:	519.5

Cover and internal design images: © Daniel Fung / Shutterstock.com (*top*);
Hypervision Creative / Shutterstock (*bottom*)

About the authors

Ken Black is Professor of Decision Sciences in the School of Business and Public Administration at the University of Houston–Clear Lake. He earned a Bachelor of Arts in mathematics from Graceland College; a Master of Arts in mathematics education from the University of Texas at El Paso; a Doctor of Philosophy in business administration in management science; and a Doctor of Philosophy in educational research from the University of North Texas.

Ken has taught all levels of statistics courses: forecasting, management science, market research and production/operations management. He has published 20 journal articles, over 20 professional papers and two textbooks: *Business statistics: an introductory course* and *Business statistics: for contemporary decision making*. Ken has consulted for many different companies, including Aetna, the City of Houston, NYLCare, AT&T, Johnson Space Centre, Southwest Information Resources, Connect Corporation and Eagle Engineering.

John Asafu-Adjaye is an Associate Professor in the School of Economics at the University of Queensland (UQ). He obtained a Bachelor of Science (Honours) in agricultural economics from the University of Ghana and then earned a Master of Science in operations research from the Aston Business School, UK. He completed a Doctor of Philosophy in natural resource economics at the University of Alberta, Edmonton, Canada.

At UQ John teaches business and economic statistics at both the undergraduate and postgraduate levels. His research activities include policy analysis of economic and environmental issues in Africa and the Asia–Pacific region. John is the author or co-author of over 80 research-based publications, including 7 books and monographs, 5 book chapters, 63 peer-reviewed journal articles and 11 commissioned reports.

Paul Burke is a Research Fellow in the School of Marketing and Centre for the Study of Choice (CenSoC) at the University of Technology Sydney (UTS). He obtained a Bachelor of Economics (First Class Honours in Marketing) from the University of Sydney. He holds a Doctor of Philosophy and Graduate Certificate in Higher Education Teaching & Learning from UTS. Paul has won teaching awards for his work in business statistics and large class teaching from UTS as well as national recognition with citations from the Carrick Institute and the Australian Learning Teaching Council. He has published in many international journals including *Research Policy*, *Educational Researcher*, *International Journal of Research in Marketing*, *Journal of Operations Management* and *Journal of Product Innovation Management*. His research interests are in choice modelling, experimental design and consumer behaviour applied in the fields of education, ethical consumerism and innovation. He has been chief investigator on many large-scale grants including Discovery and Linkage grants from the Australian Research Council (ARC), working with many international companies and organisations.

Nazim Khan is a Lecturer and Consultant in the School of Mathematics and Statistics at the University of Western Australia. He earned a Bachelor of Engineering in electrical engineering from the University of Western Australia, a Technical Teachers Certificate from the Fiji Institute of Teaching, and a Bachelor of Science (Honours) in mathematics and a Doctor of Philosophy from the University of Western Australia.

Nazim has taught decision theory at the MBA level, financial mathematics, forecasting and statistics. Nazim is an active researcher in statistics and applications. He has also presented several papers and published several articles in mathematics and statistics education. Nazim has consulted for various companies and research groups in his capacity as Consultant with the UWA Statistical Consulting Group.

Andrew Papadimos is a Lecturer in international business, statistics and economics on the Brisbane campus of Australian Catholic University. His main research interests are the Chinese economy and International Business in the Asia–Pacific region. Apart from a PhD in International Relations and Economics, Andrew also has a Masters in Applied Law from the University

Key features

Opening vignettes are concise case studies showing students the relevance of statistics and how data are used in business and the world they live in.

Chapter cases are brief business-world issues that introduce students to scenarios that use the techniques covered in the chapter to make a business decision. These are based on all-new data sets with a greater focus on cross-sectional data. At the end of each chapter, the **chapter case revisited** uses the techniques and concepts from the chapter to help make the business decision and reinforce the information presented in the text.

Excel-based data analysis is integrated throughout each chapter. Most businesses have access to Microsoft Excel and, accordingly, this text focuses on analysing data using Excel with the techniques learned in each chapter.

Misuse of statistics helps students avoid the pitfalls of using statistics incorrectly in business scenarios by highlighting their potential misuse in easy-to-understand terms.

Problems are included at the end of every section of the text. They usually follow demonstration problems and reinforce the concept learned in that section.

Going further with KaddStat is an online guide with stepped instructions to perform the textbook demonstration problems using enhanced KaddStat Excel functionality. Going further with KaddStat can be downloaded for free from the student website, www.johnwiley.com.au/highered/black4e/kaddstat.

Real-world issues at a glance

CHAPTER	OPENING VIGNETTE	CHAPTER CASE
1 Introduction to statistics	The search for information	
2 Charts and graphs	Red tape	Electronic games
3 Descriptive summary measures	Are you being followed?	Location, location, location!
4 Probability	A conditional workout	Too many leaders
5 Discrete distributions	Binge drinking	Mental health and young people
6 The normal distribution and other continuous distributions	Healthy body temperature	Prawn farm continues to grow
7 Sampling and sampling distributions	Detecting accounting fraud	Prawn farm success tied to strict quality control
8 Statistical inference: estimation for single populations	Rural obesity in Queensland on the rise	Prawn farm up for sale
9 Statistical inference: hypothesis testing for single populations	Australian childcare — enough to make you cry?	Prawn farm expects a bright future
10 Statistical inferences about two populations	Saving for retirement	Life insurance premiums
11 Analysis of variance and design of experiments	Australian teens and luxury brands	Cyberbullying amongst Australian adolescents
12 Chi-square tests	Social media is now crucial for business	Job security at Combaro Ltd
13 Simple regression analysis	Teenage smoking in pregnancy and birth weight	Predicting the selling price of houses in the city of Baycoast
14 Multiple regression analysis	Video gaming and gambling in Australian adolescents	Predicting the prices of houses in Baycoast: using additional variables
15 Time-series forecasting and index numbers	The power of tourism	Forecasting at Combaro Ltd

Acknowledgements

The authors and publisher would like to thank the following copyright holders, organisations and individuals for their permission to reproduce copyright material in *Australasian Business Statistics*, 4th edition.

Images

• Google: **2** © Google. • Shutterstock: **14** © filmfoto; **15 (top)** © auremar; **48** © Twin Design; **49** © bikeriderlondon; **94, 436** © Syda Productions; **96, 228 (top)** © Monkey Business Images; **136** © Piotr Marcinski; **137** © luxorphoto; **170** © AVAVA; **194** © Andrey_Popov; **195** © Dream79; **262 (top)** © Goran Bogicevic; **262 (bottom)** © Mati Nitibhon; **312** © Jerry Sliwowski; **314** © Rawpixel; **368** © william casey; **369** © SpeedKingz; **410** © Ttatty; **411** © Kzenon; **437** © Ekaterina Kamenetsky; **480** © Photosani; **481** © Yulia von Eisenstein; **514** © zhu difeng; **516** © gemphoto. • Interactive Games & Entertainment Association: **15 (bottom), 42** © iGEA 2014. • Alamy: **171** © PhotoStock-Israel. • Picture Media: **228 (bottom)** © REUTERS/Andrea Hopkins/Picture Media.

Text

• Oxford University Press: **601–4** © Durbin, J & Watson, GS 1951, ‘Testing for serial correlation in least squares regression II’, *Biometrika*, vol. 38, pp. 159–78.

Every effort has been made to trace the ownership of copyright material. Information that will enable the publisher to rectify any error or omission in subsequent editions will be welcome. In such cases, please contact the Permissions Section of John Wiley & Sons Australia, Ltd.

John Wiley & Sons, Australia: Terry Burkitt (Publishing Manager), Kylie Challenor (Managing Content Editor), Beth Klan (Project Editor), Tara Seeto (Senior Publishing Assistant), Delia Sala (Graphic Designer), Jo Hawthorne (Senior Production Controller), Rebecca Cam (Digital Content Editor).

Fundamental symbols and abbreviations

Samples, populations and probability

CV	coefficient of variation
E_i	event of interest
$\cap$	intersection, elements common to both sets
λ	mean number of occurrences in the interval in a Poisson distribution; say *lambda*
μ_{x}—	mean of the sample means; say *mu x bar*
A'	not A; not in A; say *A complement*
n_i	number of outcomes in which the event of interest could occur
z-score	number of standard deviations that the variable x is above or below the mean (when th
x_i	number of times the event of interest has occurred
r	Pearson correlation coefficient
s_k	Pearsonian coefficient of skewness
μ	population mean; say *mu*
p	population proportion
N	population size
σ	population standard deviation; say *sigma*
σ^2	population variance
$f(x)$	probability density function
q	probability of failure in a binomial distribution
p	probability of success in a binomial distribution
$P(X\|Y)$	probability of X given Y
$P(E_i)$	probability that an event of interest occurs
x—	sample mean; say *x bar*
pˆ	sample proportion; say *p hat*
n	sample size; number of observations or total number of items
s	sample standard deviation
s^2	sample variance
SE$_{x}$— or σ_{x}—	standard error of the mean or standard deviation of the sample means
Σx	summation of all the numbers in a grouping; say *sigma x*
N_E	total number of outcomes

$\cup$	union, combined elements of both sets

Inference and hypothesis testing

$\mathrm{H_a}$	alternative hypothesis
χ^2	chi-square ratio or chi-square distribution; say *ki square*
z_{crit}	critical value of the test statistic
df	degrees of freedom
F	*F* value; ratio of two sample variances
α	level of significance; probability of Type I error; say *alpha*
ME	margin of error of the confidence interval
μ_D	mean population difference between related samples
d—	mean sample difference
$\mathrm{H_0}$	null hypothesis
$1 - \beta$	power of the test
β	probability of Type II error; say *beta*
s_d	standard deviation of sample difference
$\mathrm{SE}_{\hat{p}}$	standard error of the proportion

Analysis of variance

ANOVA	analysis of variance
MSR	mean block sum of squares (randomised block design)
MSC	mean square of columns
MSE	mean square of errors
MST	mean square of totals
SSE	sum of squares of error
SSR	sum of squares of rows (randomised block design)
SST	sum of squares of totals
SSC	sum of squares of treatments
HSD	Tukey’s honestly significant difference test

Decision making

d_i	decision alternative *i*
EMV	expected monetary value
$P_{i,\, j}$	payoff for decision *i* under state *j*

s_j	state of nature

Regression and forecasting

x_t	actual value for current time period (t)
α	alpha, the exponential smoothing constant, which is between 0 and 1; say *alpha*
r^2	coefficient of determination
R^2	coefficient of multiple determination
SS_{xy}	covariance between x and y
C	cyclical value
df	degrees of freedom
b_k	estimate of regression coefficient k
$E(y_x)$	expected value of y
F_t	forecast value for current time period (t)
F_{t+1}	forecast for the next time period ($t + 1$)
I_i	index number for the year of interest
β_0	intercept of the population line with the Y axis
b_0	intercept of the sample line with the y axis
I	irregular or random value
Y_i	ith value of the dependent variable
X_i	ith value of the independent variable
I_L	Laspeyres price index
d_L	lower critical value of Durbin–Watson statistic
MAD	mean absolute deviation
MSE	mean square error
$\mathrm{MS}_{\mathrm{reg}}$	mean square of the regression
$\mathrm{MS}_{\mathrm{err}}$	mean square of the residual
MA	moving average
k	number of independent variables (not including the constant term)
n	number of observations
D	observed value of Durbin–Watson statistic
I_P	Paasche price index
β_k	partial regression coefficient for independent variable k
ε	population error term
$\hat{Y}$	predicted value of Y (for population data)
$\hat{y}$	predicted value of y (for sample data)

P_i	price in a given year (*i*)
P_0	price in base year (0)
e	sample error term
S	seasonal value
β_1	slope of the population regression line
b_1	slope of the sample regression line
s_e	standard error of the estimate
SSE	sum of squares of error
$\mathrm{SS}_{\mathrm{reg}}$	sum of squares of regression
$\mathrm{SS}_{\mathrm{err}}$	sum of squares of residual
SS_{xx}	sum of squares of *x*
SS_{yy}	sum of squares of *y*
y_t	time-series data value at time *t*
T	trend value
d_U	upper critical value of Durbin–Watson statistic
VIF	variance inflation factor

Nonparametric statistics

d	differences in ranks of each pair (in a Spearman’s rank correlation analysis)
D	Kolmogorov–Smirnov test statistic
K	Kruskal–Wallis test statistic
U	Mann–Whitney *U* test statistic
M_d	median (in a Wilcoxon matched-pairs signed rank test)
n_1	number of items in sample with characteristic 1 (in a runs test)
n_2	number of items in sample with characteristic 2 (in a runs test)
S	number of plus signs in *n* matched pairs with non-zero differences (in a sign test)
R	number of runs
T	smallest sum of ranks (in a Wilcoxon matched-pairs signed rank test)
r	Spearman’s rank correlation coefficient
W_1	sum of ranks for values from group 1 (in a Mann–Whitney *U* test)
W_2	sum of ranks for values from group 2 (in a Mann–Whitney *U* test)

Quality control

p—	average of sample proportions; say *p bar*

$\bar{\bar{x}}$	average sample mean for all samples; say *x double bar*
$\bar{R}$	average sample range for all samples; say *R bar*
$\bar{s}$	average sample standard deviation for all samples; say *s bar*
LCL	lower control limit
$\bar{x}$	sample mean; say *x bar*
TQM	total quality management
UCL	upper control limit

Contents

CHAPTER 1 Introduction to statistics

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

1. define some basic statistical concepts
2. classify data by type and explain why doing so is important
3. describe some common sources of data used in business statistics
4. outline the appropriate use of computers in statistical analysis
5. discuss some examples of the potential consequences of incorrect data analysis.

OPENING VIGNETTE

The search for information

Every day hundreds of millions of people use Google to search for information on the internet. The number of searches per year has been growing exponentially since Google was founded in 1998, and reached a total of 2.2 trillion searches in 2013, or almost 6 billion searches per day. In 1999 it took Google a month to build an index of 50 million pages. Now this task takes less than a minute. Currently Google has 68% of total web search volume. Its nearest rival is Baidu with 19.1%. Google's revenue has grown from $0.5 billion in the first quarter of 2008 to almost $16 billion in the second quarter of 2014.

- What algorithms are used to provide fast searches and data retrieval?
- How are algorithms measured for search quality?
- How does Google optimise advertisement quality?

Answering all of these questions depends on statistical analysis of data. Such analysis is essential for Google to maintain its market dominance and revenue.

It will come as no surprise to learn that Google employs many statisticians. Some are specialists, but many have dual qualifications. Most of them work in the advertising division ('Ads') or the search engine division ('Search'). In the advertising division, quantitative analysts design tools and processes to measure the effectiveness of Google's advertising service and in turn to improve those services. In the search division, statisticians analyse the quality of the results returned by Google's search engine. That information then feeds back to the software engineers to improve the search engine.

With perhaps the biggest collection of data in the world, Google requires its statisticians to work with multi-disciplinary teams to solve a wide range of business problems.

Introduction

Every minute of the working day, businesses around the world make decisions that determine whether they will profit and grow or whether they will stagnate and die. Most of these decisions are made with the assistance of information about the marketplace, economic and financial environments, workforce, competition and other factors. Such information usually comes in the form of data. Business statistics provides the tools through which data are collected, analysed, summarised, interpreted and presented to facilitate the decision-making process. Thus, business statistics plays an important role in decision making within the dynamic world of business.

In this text, we first introduce basic statistical concepts. We then discuss how to organise and present data so they are meaningful and useful to decision makers. We will learn techniques for sampling (from a population) that allow studies of the business world to be conducted promptly at lower cost. We will explore various ways to forecast future values and examine techniques for determining trends. This text also includes many statistical tools for testing hypotheses and for estimating population parameters. These and many other useful statistical techniques await us on this journey through business statistics. Let us begin.

1.1 BASIC STATISTICAL CONCEPTS

In this section some basic concepts will be discussed so that statistical problems can be put into context. These concepts will be covered in detail in later chapters.

Two important concepts in statistics are population and sample. A **population** is a collection of objects (often called units or subjects) of interest. Examples of populations include:

1. all small businesses
2. all workers currently employed by BHP Billiton
3. all dishwashers produced by Fisher & Paykel in Auckland in the past 12 months.

A population (and unit) can be very widely defined, such as ‘all cars’, or narrowly defined such as ‘all red Toyota Corolla hatchbacks produced in 2015’.

Collection of data on a whole population is called a **census**. A **sample** is a subset of the units in a population. If selected using the principles of sampling, a sample can be expected to be representative of the whole population. Sampling has several advantages over a census. In particular, sampling is simpler and cheaper. Further, some forms of data collection are destructive. For example, crash test statistics for a particular model of car are obtained by destroying the car. This makes it impossible to collect crash data on all cars, so sampling is the only option.

There are two steps in analysing data from a sample: exploratory data analysis and statistical inference. These are related and both should be performed for any given data. **Exploratory data analysis**, or EDA, is the first step, in which numerical, tabular and graphical summaries (such as frequency tables, means, standard deviations and histograms) of data are produced to summarise and highlight the key aspects or any special features of the data. Often, such analysis is sufficient for the purpose of the study. However, more often this is a precursor to more formal and extensive analysis of the data.

Statistical inference uses sample data to reach conclusions about the population from which the sample was drawn. This is usually the main aim of any statistical exercise and involves more formal data analysis techniques. An inference is a conclusion that patterns observed in the data (sample) are present in the wider population from which the data were collected. A **statistical inference** is an inference based on a probability model linking the data to the population. Clearly such conclusions assume that the sample data are representative of the population; appropriate data collection is vital for such assumptions to hold true.

As an example, in pharmaceutical research, tests must be limited to a small sample of patients since new drugs are expensive to produce. Researchers design experiments with small, representative samples of patients and draw conclusions about the whole population using techniques of statistical inference.

Note that no inference is required for census data, since a census collects data on the whole population. In this case, EDA is all that is possible. Any inference will be based on simple comparisons of numerical and graphical summaries with a previous census.

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

Distinction between the terms *parameter* and *statistic* is important. A business researcher often wants to estimate the value of a parameter or draw inferences about the parameter. However, the calculation of parameters is usually either impossible or infeasible because of the amount of time and money required to conduct a census. In such cases, the business researcher can take a representative sample of the population and use the corresponding sample statistic to estimate the population parameter. Thus, the sample mean, x¯ $\bar{x}$, is used to estimate the population mean, μ. The basis for inferential statistics, then, is the ability to make decisions about parameters without having to complete a census of the population.

For example, Fisher & Paykel may want to determine the average number of loads that its 8 kg LCD washing machines can wash before needing repairs. The population here is *all* the 8 kg LCD washing machines, and the parameter is the population mean: that is, the average number of washes per machine before repair. A company statistician takes a representative sample of these machines, conducts trials on this sample, recording the number of washes before repair for each machine, and then computes the sample average number of washes before repair. The (population) mean number of washes for this type of washing machine is then estimated from this sample mean.

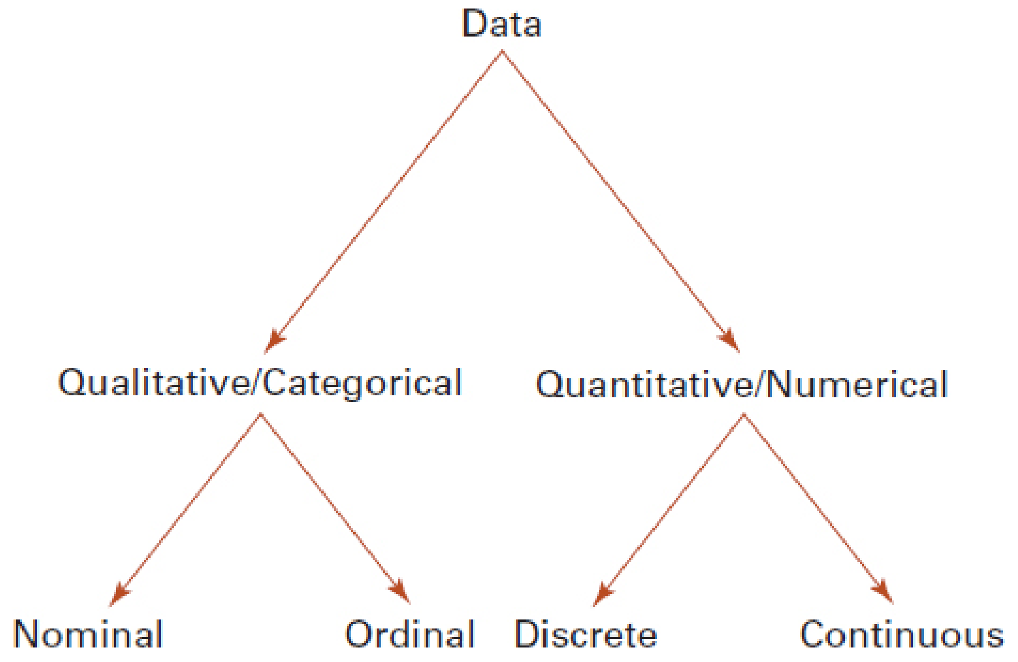
Inferences about parameters are made under uncertainty. Unless parameters are computed directly from a census, the statistician never knows with certainty whether the estimates or inferences made from samples are true. In an effort to estimate the level of confidence in the result of the process, statisticians use probability statements. Therefore, part of this text is devoted to probability.

1.2 TYPES OF DATA

Most available data are numerical. Before we analyse data we need to know what the numbers represent. For example, the data could be the dollar cost of items produced, the geographical location of retail outlets, weights of shipments or rankings of sales staff. These data are of different types and cannot be analysed the same way.

Which exploratory techniques and which inferential methods we use are largely determined by the type of data. Data can be broadly classified as qualitative (also known as categorical) or quantitative (also known as numerical). Categorical data can be further subclassified as nominal or ordinal, and numerical data can be subclassified as discrete or continuous. Figure 1.1 shows this pictorially.

FIGURE 1.1 Types of data



Categorical data

A data type that is simply an identifier or label and has no numerical meaning is **categorical data**.Indeed, such data are often not numbers. For example, the employment of a person (teacher, doctor, lawyer, engineer, business executive, other) is a categorical data type. As another example, the grade in a test (A, B, C, D, E, F) is again simply a label and is a categorical data type. Notice that the two examples are slightly different, in that employment of a person cannot be ranked in any meaningful way, but the test grades have a natural ordering. Thus, the first example is a *nominal* data type, while the second is an *ordinal* data type.

Numerical data

Numerical data have a natural order and the numbers represent some quantity. Two examples are the number of heads in ten tosses of a coin and the weights of rugby players. Note that in the first example we know in advance exactly which values the data may take, namely 0, 1, ..., 10, whereas in the second example all we can give is perhaps a range (say, 80–140 kg). The first example is that of a *discrete* data type, where we can list the possible values. The second example is that of a *continuous* data type, where we can give only a range of possible values for the data. Discrete data often arise from counting processes, while continuous data arise from measurements.

Some data that may be considered to be discrete are often taken as continuous for the purposes of analysis. For example, a person's salary is discrete (that is, in dollars and cents), but because the range of the data is large and often the number of observations is also large, such data are considered to be continuous.

DEMONSTRATION PROBLEM 1.1

PROBLEM: Shoppers in a city are surveyed by the chamber of commerce. Some of the questions in the survey are listed below. What type of data will result from each of the following questions?

1. What is your age (in years)? ________
2. Which mode of transport did you use to travel to the city today?
   □ Public □ Private
3. How far did you travel to the city today (in kilometres)? _________
4. How much did you spend in the city today? ________
5. What did you spend most of your money on today? (choose one)
   □ Clothes □ Shoes □ Food □ Electronic goods □ Services □ Other
6. How satisfied are you with your shopping experience in the city? (circle one)
   Very satisfied Satisfied Neutral Unsatisfied Very unsatisfied

SOLUTION: Question 1 is age in years, so it is a discrete variable. However, for the purpose of analysis age, like salary, is often regarded as continuous.

In question 2, the shopper is asked to categorise the type of transport they used. The responses to this question cannot be ranked or ordered in any meaningful way. Therefore the mode of transport data are categorical, nominal.

Questions 3 and 4 involve measurement and so provide continuous data.

Question 5 results in categorical, nominal data. The data cannot be ranked or ordered.

Question 6 provides categorical ordinal data, as the responses can be ranked or ordered in a sensible and natural way.

Cross-sectional and time-series data

Data that are collected at a fixed point in time are called **cross-sectional data**. Such data give a snapshot of the measured variables at that point in time. For example, Roy Morgan Research conducts and publishes monthly surveys of consumer confidence. The monthly survey provides information on consumer confidence for the given month.

Often data are collected over time. Such data are called **time-series data**. For example, data that consist of consumer confidence over several months or years are time-series data. Note that, unlike cross-sectional data, time-series data are time dependent. Such dependence needs to be appropriately modelled and accounted for in the data analysis.

1.3 OBTAINING DATA

Decisions such as how many units of a product to produce, which processes should be changed to improve quality and which export market to target, can all be better informed through statistics. Statistical analysis begins with data. Deciding what data are needed and how to obtain them are essential for any person or organisation seeking to make an informed business decision. Assessing the quality of data is also important, because it is impossible to produce high-quality statistical analyses — and thus make good decisions — from poor-quality data.

Data that were collected for some other purpose and are already available are known as **secondary data**. Secondary data are available from external and internal sources. External sources of secondary data include government departments, industry associations, academic institutions and commercial research organisations. Internal sources might include sales figures, production records or customer evaluations. In business, the decision maker should always look for secondary data before conducting a data-collecting exercise. Where the required information already exists, further data gathering will likely be a waste of the organisation’s time and money. However, secondary data may not adequately fit the purpose, may be out of date or may have been obtained or processed in some way that is not statistically valid.

Data collected to address a specific need are known as **primary data**. Primary data might be collected using a survey, experiment or some other study.

Obtaining secondary data

Government and government-related organisations are the largest source of secondary data. Most Asia–Pacific nations have a national statistics agency. The types of data typically available from these agencies include key economic, social and demographic indicators. In Australia the agency is the Australian Bureau of Statistics, and in New Zealand it is Statistics New Zealand. Other agencies in the region include the Statistics Bureau of Japan, Korea National Statistical Office, Statistics Indonesia, National Statistical Office of Thailand, Department of Statistics Malaysia and Statistics Singapore. Key economic data are also provided by the Reserve Bank of Australia and equivalent bodies in other countries, such as the Monetary Authority of Singapore, Bank Negara Malaysia, Bank of Indonesia and Reserve Bank of New Zealand. In addition to these national agencies, local government authorities often have significant data holdings in relation to their areas of responsibility.

A number of peak multinational organisations publish secondary data. The United Nations, Organisation for Economic Co-operation and Development (OECD), World Bank, and Association of Southeast Asian Nations (ASEAN) are good sources of data.

Industry associations such as the Australian Recording Industry Association and Horticulture New Zealand are also often excellent sources of secondary data specific to their industries. In the commercial domain, market reports on subjects such as market trends and financial indicators are available for purchase or on a subscription basis from organisations such as Dun and Bradstreet or ANZ Bank. Universities are also great sources of secondary data from their research programs.

When evaluating secondary data, it is extremely important to consider the reliability of the data collection agency (some countries, for example, have better collection capabilities than others), the data’s relevance to your situation, the comparability of the data and the currency of the data.

Obtaining primary data

If a decision has been made to obtain primary data, the necessary research can be done in-house or by an external supplier of research services. Large businesses might have a dedicated research department; smaller ones might have someone with adequate research skills; others might have no in-house research expertise. Even if there is research expertise within a business, it would still be common for an outside service to be engaged, subject to the availability of funds. External research suppliers range from small consultancy operations to large multinationals such as Nielsen and BIS Shrapnel.

Another common way to obtain data is by customer surveys. Surveys need particular care as the results are unreliable unless the sample is taken at random and represents the target population. For example, *voluntary response surveys*, where the subject chooses to be in the survey, suffer from what is called *self-selection bias*; that is, individuals select themselves into the sample, thus producing a biased sample and consequently biasing any statistics based on the sample. Such samples do not represent the target population, and so the results are very unreliable. In particular, ‘phone-in’ surveys conducted by radio and television stations are extremely unreliable, even if the results seem very clear from what are relatively large samples. See section 1.5 for an example of a voluntary response survey.

1.4 STATISTICAL ANALYSIS USING EXCEL

Computers offer many opportunities for statistical analyses. Calculations for advanced statistical techniques are tedious and cumbersome to perform. In the business environment, decision makers will almost always be dealing with very large data sets and will commonly use a computer to analyse the data.

Business statisticians use statistical software packages including SPSS, MINITAB and SAS. Some of these specialist packages are quite expensive and require substantial user training. Fortunately, many spreadsheet software packages can analyse data statistically. In this book, when it is appropriate to use a computer, we will use Microsoft Excel for data analysis; it is the most commonly used package in the business environment and is the package you are most likely to use in your professional life. It must be noted, however, that Excel was not specifically designed for statistical analysis and there are some important limitations. In particular, Excel cannot perform every type of statistical analysis, and some charts produced by Excel are of poor quality from a statistical point of view.

We will use Excel's statistical add-in component and the KaddStat add-in that is provided with this text to give extra functionality. It is important to remember, however, that a statistical package is not a replacement for a thorough understanding of correct statistical methods. The statistician must analyse each business or statistical problem to determine the most appropriate statistical methods. Simply relying on convenient software tools that may be at hand without thinking through the best approach can lead to errors, oversights and poor decisions. This book aims to teach you when and how to use statistical methods to provide information that can be used in business decisions.

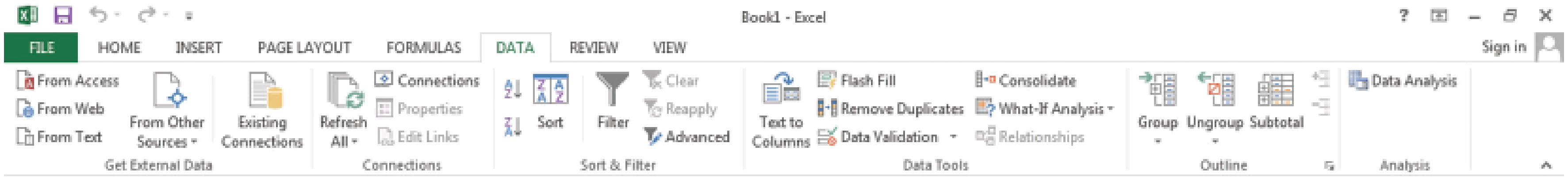
Getting started with Excel and KaddStat

Excel is part of the Microsoft Office suite. We assume you have this software installed. We give concise instructions for each Excel statistics function as you encounter it for the first time. We do, however, assume a basic familiarity with Excel. If you are not familiar with Excel or the Windows operating system, please refer to Microsoft's online documentation. The instructions in this book are for Microsoft Office 2013.

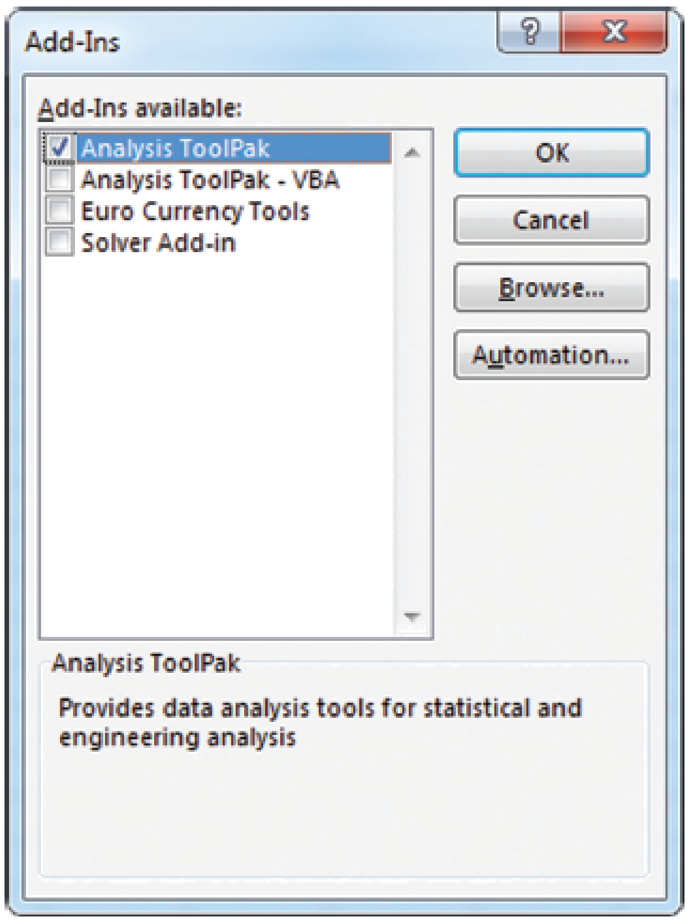
Excel's Analysis ToolPak add-in

Excel comes with a number of add-ins to provide extra functionality. To use Excel for statistical analysis you will need to load the Analysis ToolPak add-in.

1. To determine whether the Analysis ToolPak add-in is already loaded, click the Data tab. If Data Analysis is listed on the Analysis group, the add-in is already loaded and you can skip to the next section. If Data Analysis does not appear, you will need to follow the steps below to load the add-in.



2. Click the File tab at the top left of the screen and select Options.
3. Select Add-ins from the left-hand menu.
4. Choose Excel Add-ins from the Manage drop-down menu and click Go …
5. Check the Analysis ToolPak box and click OK.



☑ Analysis ToolPak
☐ Analysis ToolPak - VBA
☐ Euro Currency Tools
☐ Solver Add-in

Analysis ToolPak

Provides data analysis tools for statistical and engineering analysis

6. If Analysis ToolPak is not listed in the Add-Ins available box, click Browse in the Add-ins dialogue box, and then locate the add-in. If the add-in is not currently installed on your computer, click Yes to install it and follow the instructions for the setup program.

7. Data Analysis will now be listed on the Data tab.

KaddStat

Even with the Analysis ToolPak add-in, Excel's statistical functions are limited. This book comes with the KaddStat statistical analysis add-in to provide further functionality. To perform some of the statistical procedures in this book, you will need to install KaddStat.

1. Download the add-in from the student website, www.johnwiley.com.au/highered/black4e/kaddstat.
2. Unzip the file.
3. Within Excel, click the File tab at the top left of the screen and then click Open.
4. Click Computer and then the Browse icon.
5. Navigate to the downloaded Kadd.xla file. Select it and click Open. If you get a security dialogue box concerning macros, click Enable Macros.
6. If you get an Install KADDstat Option dialogue box, select Install KADDstat as Addin and click Update.
7. Kadd will appear under the Add-Ins tab.

Alternatively, you can copy the Kadd.xla file to a folder on your computer and install it using the same procedure used for installing the Analysis ToolPak add-in.

Using Excel with this book

Most of the demonstration problems in this textbook feature detailed instructions for both a manual solution and one that uses Excel functions, the Analysis ToolPak add-in or the KaddStat add-in. As your knowledge of statistics builds, you may wish to explore KaddStat more fully.

You may note that some of the Excel-generated graphs vary slightly in appearance from your own outputs. These are only cosmetic differences due to individual user settings and can be adjusted via the options on the Home or Page Layout tabs. Should you encounter any cells in scientific notation (i.e. those displaying E+ or E−), they can be reformatted by right-clicking the cell and selecting Format Cells ... You can then select Number from the Category list, adjust the number of decimal places in the Decimal places field and then click OK.

The role of the software is to facilitate calculations. This is very important, as it allows a statistician to explore more complex models and to perform extended analyses of the data. The statistician's role is then to select the best model, interpret it and discuss its implications in the context of the question to be answered.

1.5 WHEN THINGS GO WRONG

Analysing data correctly is important. Incorrect analysis can have disastrous consequences. The following three examples of incorrect statistics and the ensuing consequences are presented for you to consider.

Space shuttle *Challenger*

The *Challenger* space shuttle launch on 28 January 1986, from the Kennedy Space Center in Florida, went disastrously wrong and the shuttle exploded 73 seconds into its flight. The launch was on a particularly cold morning, with subzero temperatures. Some engineers argued that low temperatures could cause the failure of the rubber O-rings that sealed the solid rocket boosters. Data on the performance of these O-rings at different temperatures were available, but some flights were omitted from the analysis. This led to an underestimation of the probability of failure of the O-rings at low temperatures. The launch went ahead with disastrous consequences, costing the lives of all seven crew members and billions of dollars.

The Sally Clark case

Sally Clark was the victim of a miscarriage of justice when incorrect statistics were used to convict her of the murder of her two children. Clark's first son died a few weeks after his birth. Two years later, when her second son also died at just a few weeks of age, Clark was arrested and charged with the murder of both children. At her trial, a paediatrician appeared as an expert medical witness and testified that there was only a 1 in 73 million chance that two children from such a family could die of sudden infant death syndrome (SIDS). This statistic was calculated by taking the probability of one such death (the paediatrician quoted 1 in 8543) and squaring it (to give 1 in 73 million). The figure of 1 in 8543 was for a particular set of circumstances that were not entirely applicable to the case. Further, the technique of squaring the probability assumes that the SIDS deaths would be totally independent, meaning there was no family or environmental link involved in SIDS — this is not a valid assumption.

Clark was convicted and sentenced to prison. She was released after three years when evidence came to light that suggested one of the children had died of natural causes. A couple of years after Clark's conviction, the Royal Statistical Society released a statement pointing out the invalidity of the paediatrician's testimony and calling for the courts to rely only on expert statisticians in such cases. Hundreds of other cases were reviewed after Clark was released from prison and two other women convicted of murdering their children were also released.

The 1936 US presidential election

In the 1932 US presidential election, Democrat candidate Franklin D Roosevelt won a sweeping election victory over the sitting Republican president, Herbert Hoover. During Roosevelt's reelection campaign in 1936, the *Literary Digest* conducted a survey by sending out 10 million ballots to voters selected from telephone books and motor vehicle registrations. More than 2.3 million ballots were returned, which is comparatively a very large sample. The results were published and indicated a landslide victory for the Republican challenger, Alf Landon. However, the election was actually won by Roosevelt with a large margin.

The magazine had run a poll for every election since 1920 and correctly predicted the election outcome in each case. The magazine tried to determine what had gone wrong. Polls of this nature suffer from the problem of *voluntary response bias*; that is, people choose whether they will respond, and this can lead to a very biased sample. For example, the people who respond to online surveys or phone-in surveys conducted by television and radio stations are predominantly those who feel strongly about the issue. The results of such surveys are not reliable, even if the sample size is large.

These examples show that it is important to use statistics correctly, as otherwise the results are at best doubtful and at worse simply incorrect. There can be serious consequences of incorrect statistical analysis, from both business and human cost points of view.

While this first course in statistics will prepare you for conducting many types of statistical analyses, you must be aware of the limitations of your knowledge. It is advisable to seek expert advice for more advanced data analysis and interpretation.

SUMMARY

1. Statistics is a mathematical science concerned with the collection, presentation, analysis and interpretation or explanation of data. Two important concepts in statistics are *population* and *sample*. A population is a set of units of interest and a sample is a subset of the population. A sample should be selected in such a way that it is representative of the population. A census is the process of collecting data on the whole population at a given point in time.

   The two steps in analysing data are exploratory data analysis (EDA) and inferential statistics. EDA aims to summarise and describe data whereas inferential statistics uses sample data to reach conclusions about the population from which the sample was drawn.

2. Data are broadly classified as qualitative (also called categorical) and quantitative (also called numerical). Qualitative data can be further classified into nominal or ordinal, while quantitative data are further classified into discrete or continuous. The type of data and how they were collected determine which EDA and inference techniques should be used.

   Data that are collected at a fixed point in time are called cross-sectional data, while data that are collected over time are called time-series data.

3. Various sources of data are available for answering business questions. Data that are already available are called secondary data. Various agencies routinely collect data and these are useful sources of data for statistical analysis. Data that are collected for a specific purpose are known as primary data.
4. Computers enable fast and accurate statistical analyses of very large amounts of data. In fact, new statistical techniques have become possible due to the increase in computing power. Business statisticians use software packages including SPSS, MINITAB and SAS. Many spreadsheet software packages can also analyse data statistically. Microsoft Excel is the most widely used package for business statistics due to its ease of use, but it is limited in its range of techniques. Remember that software cannot replace a thorough understanding of correct statistical methods.
5. Incorrect data analysis can have disastrous consequences, with large human and business costs. It is important to seek expert help for more complicated statistical analyses.

KEY TERMS

categorical data

census

cross-sectional data

exploratory data analysis (EDA)

numerical data

parameter

population

primary data

sample

secondary data

statistic

statistical inference

time-series data

REVIEW PROBLEMS

Testing your understanding

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

1.2 For each of the following companies, give examples of data that could be gathered and what purpose the data would serve: Bluescope Steel, AAMI, Jetstar, IKEA, Telstra, ANZ Bank, Sydney City Council, and Black and White Taxis.

1.3 Give an example of *descriptive* statistics in the recording industry. Give an example of how *inferential* statistics could be used in the recording industry.

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

1.5 Classify each of the following as nominal, ordinal, discrete or continuous data.

a. The RBA interest rate
b. The return from government bonds
c. The customer satisfaction ranking in a survey of a telecommunications company
d. The ASX 200 index
e. The number of tourists arriving in Australia each month
f. The airline a tourist flies by into Australia
g. The time to serve a customer in a bank queue

1.6 Classify each of the following as nominal, ordinal, discrete or continuous data.

a. The ranking of a company on *BRW*'s top 1000 list
b. The number of tickets sold at a cinema on any given night
c. The identification number on a questionnaire
d. Per capita income
e. The trade balance in dollars
f. Socioeconomic class (low, middle, upper)
g. Profit/loss in dollars
h. A company's ABN
i. Standard & Poor's credit ratings of countries based on the following scale

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

1.7 Powerkontrol Australia designs and manufactures power distribution switchboards and control centres for hospitals, bridges, airports, tunnels, highways and water treatment plants. Powerkontrol's director of marketing wants to determine client satisfaction with its products and services. He developed a questionnaire that yields a satisfaction score between 10 and 50 for participant responses. A random sample of 35 of the company's 900 clients is asked to complete a satisfaction survey. The satisfaction scores for the 35 participants are averaged to produce a mean satisfaction score.

a. What is the population for this study?
b. What is the sample for this study?
c. What is the statistic for this study?
d. What would be a parameter for this study?

1.8 Cricket Australia wants to run a marketing campaign to increase attendance at test matches. You have been hired as a consultant to conduct a survey and prepare a report on your findings.

a. What variables do you consider affect a person's interest in cricket test matches?
b. Design a questionnaire of 10 to 15 questions that will enable you to decide which section of the population the marketing campaign should target.

CHAPTER 2 Charts and graphs

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

1. distinguish between grouped and ungrouped data
2. produce graphical summaries of univariate data — histograms, frequency polygons, ogives, pie charts, stem and leaf plots and Pareto charts
3. produce graphical summaries of two-variable continuous data — scatter plots.

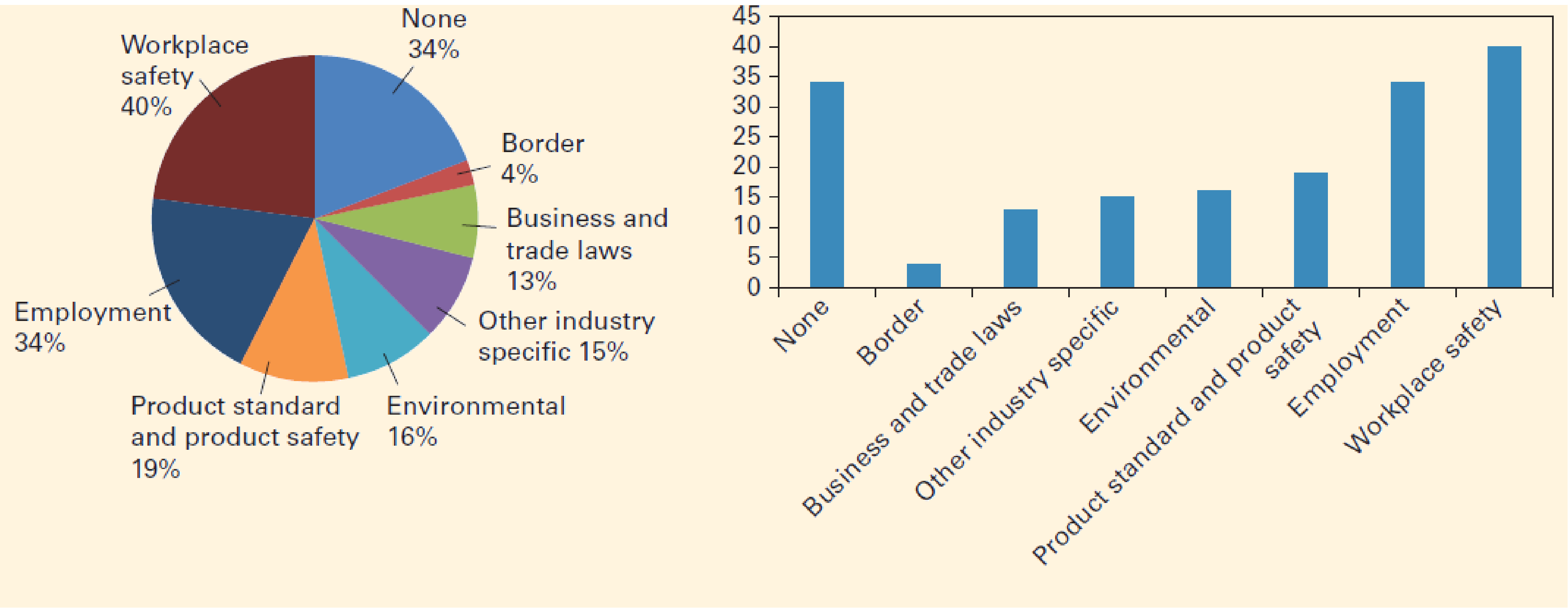
OPENING VIGNETTE

Red tape

Government regulations, sometimes known as 'red tape', can be daunting for businesses. Regulations reflect society's expectations of how a business will operate, and non-compliance can lead to heavy penalties. Many businesses spend considerable time and resources complying with regulations. The table below shows Statistics New Zealand data on businesses' perceptions of their compliance burdens.

TYPE OF REGULATION	PERCENTAGE
None	34.0
Border	4.0
Business and trade laws	13.0
Other industry specific	15.0
Environmental	16.0
Product standard and product safety	19.0
Employment	34.0
Workplace safety	40.0

These data can be displayed as a graph in several ways. Shown here are a pie chart and a vertical bar chart of the data. It may be easier to visually compare categories that are similar in size using a bar chart or a histogram rather than a pie chart. Equally, a Pareto chart could be produced.



Introduction

The first step in any analysis is summarising and presenting the data in such a way that key features of the data can be identified, such as central location, spread, symmetry, distribution and groupings in the data. Two useful techniques for exploratory data analysis (EDA) are graphical presentation and numerical summaries of data. The methods used to display data and the numerical summaries depend on the type of data.

In this chapter we will focus on presenting data as graphs. Continuous data are usually presented as histograms and categorical data as bar charts or pie charts. In each case, often some grouping of the data is required first.

Note that often data that are not continuous are treated as if they were when analysed. For example, salary is a discrete variable, but due to its wide range of values it is treated as continuous. Most data dealing with money are discrete but are treated as continuous.

CHAPTER CASE

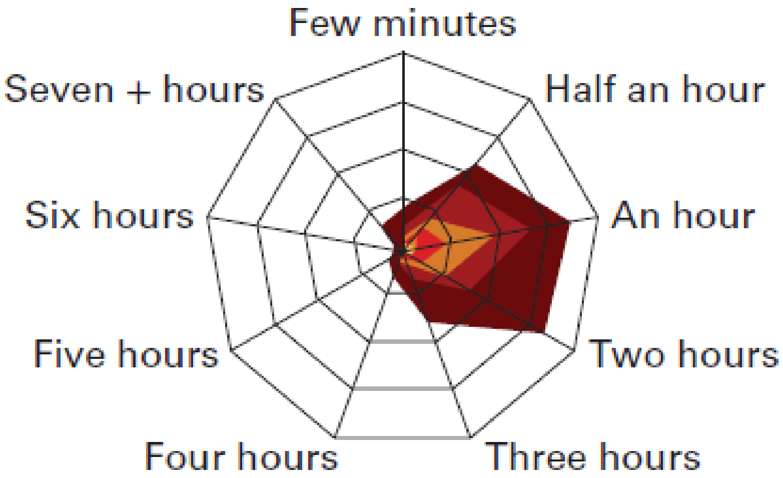
Electronic games

The first interactive video game, ‘Tennis for Two’, was invented by physicist William Higinbotham in 1958, to entertain visitors to an open day at his research lab. Since then computer and electronic games have come a long way — today they are a part of human culture around the globe.

Computer games have evolved over the decades. A growing market in recent years has been games for mobile phones and other handheld devices. This has led to many opportunities in the ‘app’ market for mobile devices and some app developers have become very wealthy from successful games.

As in any type of business, marketing is an essential part of raising the profile of a game. Important considerations for marketing are the profile of gamers, price and platform. A survey by Bond University shows that 75% of gamers are older than 18 years, with the average age being 32. Some 47% of gamers are female. More data from the report are presented in the table and chart below.

Frequency and duration of play



AGE (YEARS)	PERCENTAGE OF TOTAL GAMERS
1 to 5	5
6 to 10	5
11 to 15	10