



Derek L. Waller

# Statistics *for* **BUSINESS**



# Statistics for Business

*This page intentionally left blank*

# Statistics for Business

Derek L Waller



AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD  
PARIS • SAN DIEGO • SAN FRANCISCO • SYDNEY • TOKYO

Butterworth-Heinemann is an imprint of Elsevier



Butterworth-Heinemann is an imprint of Elsevier  
Linacre House, Jordan Hill, Oxford OX2 8DP, UK  
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA

First edition 2008

Copyright © 2008, Derek L Waller  
Published by Elsevier Inc. All rights reserved

The right of Derek L Waller to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: [permissions@elsevier.com](mailto:permissions@elsevier.com). Alternatively you can submit your request online by visiting the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting *Obtaining permission to use Elsevier material*

#### Notice

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein

#### Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

#### British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN: 978-0-7506-8660-0

For information on all Butterworth-Heinemann publications  
visit our web site at [books.elsevier.com](http://books.elsevier.com)

Typeset by Charon Tec Ltd (A Macmillan Company), Chennai, India

Printed and bound in Great Britain

08 09 10 10 9 8 7 6 5 4 3 2 1

Working together to grow  
libraries in developing countries

[www.elsevier.com](http://www.elsevier.com) | [www.bookaid.org](http://www.bookaid.org) | [www.sabre.org](http://www.sabre.org)

ELSEVIER

BOOK AID  
International

Sabre Foundation

*This textbook is dedicated to my family, Christine, Delphine, and Guillaume.  
To the many students who have taken a course in business statistics with me ...  
You might find that your name crops up somewhere in this text!*

*This page intentionally left blank*

# Contents

About this book	ix	Using a Normal Distribution to Approximate a Binomial Distribution	169
<b>1 Presenting and organizing data</b>	<b>1</b>	Chapter Summary	172
Numerical Data	3	Exercise Problems	174
Categorical Data	15	<b>6 Theory and methods of statistical sampling</b>	<b>185</b>
Chapter Summary	23	Statistical Relationships in Sampling for the Mean	187
Exercise Problems	25	Sampling for the Means from an Infinite Population	196
<b>2 Characterizing and defining data</b>	<b>45</b>	Sampling for the Means from a Finite Population	199
Central Tendency of Data	47	Sampling Distribution of the Proportion	203
Dispersion of Data	53	Sampling Methods	206
Quartiles	57	Chapter Summary	211
Percentiles	60	Exercise Problems	213
Chapter Summary	63	<b>7 Estimating population characteristics</b>	<b>229</b>
Exercise Problems	65	Estimating the Mean Value	231
<b>3 Basic probability and counting rules</b>	<b>79</b>	Estimating the Mean Using the Student- <i>t</i> Distribution	237
Basic Probability Rules	81	Estimating and Auditing	243
System Reliability and Probability	93	Estimating the Proportion	245
Counting Rules	99	Margin of Error and Levels of Confidence	248
Chapter Summary	103	Chapter Summary	251
Exercise Problems	105	Exercise Problems	253
<b>4 Probability analysis for discrete data</b>	<b>119</b>	<b>8 Hypothesis testing of a single population</b>	<b>263</b>
Distribution for Discrete Random Variables	120	Concept of Hypothesis Testing	264
Binomial Distribution	127	Hypothesis Testing for the Mean Value	265
Poisson Distribution	130	Hypothesis Testing for Proportions	272
Chapter Summary	134		
Exercise Problems	136		
<b>5 Probability analysis in the normal distribution</b>	<b>149</b>		
Describing the Normal Distribution	150		
Demonstrating That Data Follow a Normal Distribution	161		



The Probability Value in Testing Hypothesis	274	Forecasting Using Non-linear Regression	351
Risks in Hypothesis Testing	276	Seasonal Patterns in Forecasting	353
Chapter Summary	279	Considerations in Statistical Forecasting	360
Exercise Problems	281	Chapter Summary	364
<b>9 Hypothesis testing for different populations</b>	<b>301</b>	Exercise Problems	366
Difference Between the Mean of Two Independent Populations	302	<b>11 Indexing as a method for data analysis</b>	<b>383</b>
Differences of the Means Between Dependent or Paired Populations	309	Relative Time-Based Indexes	385
Difference Between the Proportions of Two Populations with Large Samples	311	Relative Regional Indexes	391
Chi-Square Test for Dependency	313	Weighting the Index Number	392
Chapter Summary	319	Chapter Summary	397
Exercise Problems	321	Exercise Problems	398
<b>10 Forecasting and estimating from correlated data</b>	<b>333</b>	Appendix I: Key Terminology and Formula in Statistics	413
A Time Series and Correlation	335	Appendix II: Guide for Using Microsoft Excel 2003 in This Textbook	429
Linear Regression in a Time Series Data	339	Appendix III: Mathematical Relationships	437
Linear Regression and Causal Forecasting	345	Appendix IV: Answers to End of Chapter Exercises	449
Forecasting Using Multiple Regression	347	Bibliography	509
		Index	511

# About this book

This textbook, *Statistics for Business*, explains clearly in a readable, step-by-step approach the fundamentals of statistical analysis particularly oriented towards business situations. Much of the information can be covered in an intensive semester course or alternatively, some of the material can be eliminated when a programme is on a quarterly basis. The following paragraphs outline the objectives and approach of this book.

## The subject of statistics

Statistics includes the collecting, organizing, and analysing of data for describing situations and often for the purposes of decision-making. Usually the data collected are quantitative, or numerical, but information can also be categorical or qualitative. However, any qualitative data can subsequently be made quantitative by using a numerically scaled questionnaire where subjective responses correspond to an established number scale.

Statistical analysis is fundamental in the business environment as logical decisions are based on quantitative data. Quite simply, if you cannot express what you know, your current situation, or the future outlook, in the form of numbers, you really do not know much about it. And, if you do not know much about it, you cannot manage it. Without numbers, you are just another person with an opinion! This is where statistics plays a role and why it is important to study the subject. For example, by simply displaying statistical data in a visual form you can convince your manager or your client. By using probability analysis you can test your company's

strategy and importantly evaluate expected financial risk. Market surveys are useful to evaluate the probable success of new products or innovative processes. Operations managers in services and manufacturing, use statistical process control for monitoring and controlling performance. In all companies, historical data are used to develop sales forecasts, budgets, capacity requirements, or personnel needs. In finance, managers analyse company stocks, financial performance, or the economic outlook for investment purposes. For firms like General Electric, Motorola, Caterpillar, Gillette (now a subsidiary of Procter & Gamble), or AXA (Insurance), six-sigma quality, which is founded on statistics, is part of the company management culture!

## Chapter organization

There are 11 chapters and each one presents a subject area – organization of information, characteristics of data, probability basics, discrete data, the normal distribution, sampling, estimating, hypothesis testing for single, and multiple populations, forecasting and correlation, and data indexing. Each chapter begins with a box opener illustrating a situation where the particular subject area might be encountered. Following the box opener are the learning objectives, which highlight the principal themes that you will study in the chapter indicating also the subtopics of each theme. These subtopics underscore the elements that you will cover. Finally, at the end of each chapter is a summary organized according to the principal themes. Thus, the box opener, the learning objectives, the chapter itself, and the

chapter summary are logically and conveniently linked that will facilitate navigation and retention of each chapter subject area.

## Glossary

Like many business subjects, statistics contains many definitions, jargon, and equations that are highlighted in bold letters throughout the text. These definitions and equations, over 300, are all compiled in an alphabetic glossary in [Appendix I](#).

## Microsoft excel

This text is entirely based on Microsoft Excel with its interactive spreadsheets, graphical capabilities, and built-in macro-functions. These functions contain all the mathematical and statistical relationships such as the normal, binomial, Poisson, and Student-t distributions. For this reason, this textbook does not include any of the classic statistical tables such as the standardized normal distribution, Student-t, or chi-square values as all of these are contained in the Microsoft Excel package. As you work through the chapters in this book, you will find reference to all the appropriate statistical functions employed. A guide of how to use these Excel functions is contained in [Appendix II](#), in the paragraph “Using the Excel Functions”. The related Table E-2 then gives a listing and the purpose of all the functions used in this text.

The 11 chapters in this book contain numerous tables, line graphs, histograms and pie charts. All these have been developed from data using an Excel spreadsheet and this data has then been converted into the desired graph. What I have done with these Excel screen graphs (or screen dumps as they are sometimes disparagingly called) is to tidy them up by removing the tool bar, the footers, and the numerical column and the alphabetic line headings to give an uncluttered graph. These Excel graphs in PowerPoint format are available on the Web.

A guide of how to make these Excel graphs is given also in [Appendix II](#) in the paragraph, “Generating Excel Graphs”. Associated with this paragraph are several Excel screens giving the stepwise procedure to develop graphs from a particular set of data.

I have chosen Excel as the cornerstone of this book, rather than other statistical packages, as in my experience Excel is a major working tool in business. Thus, when you have completed this book you will have gained a double competence – understanding business statistics and versatility in using Excel!

## Basic mathematics

You may feel a little rusty about your basic mathematics that you did in secondary school. In this case, in [Appendix III](#) is a section that covers all the arithmetical terms and equations that provide all the basics (and more) for statistical analysis.

## Worked examples and end-of-chapter exercises

In every chapter there are worked examples to aid comprehension of concepts. Further there are numerous multipart end-of-chapter exercises and a case. All of these examples and exercises are based on Microsoft Excel. The emphasis of this textbook, as underscored by these chapter exercises, is on practical business applications. The answers for the exercises are given in [Appendix IV](#) and the databases for these exercises and the worked examples are contained on the enclosed CD. (Note, in the text you may find that if you perform the application examples and test exercises on a calculator you may find slightly different answers than those presented in the textbook. This is because all the examples and exercises have been calculated using Excel, which carries up to 14 figures after the decimal point. A calculator will round numbers.)

## International

The business environment is global. This textbook recognizes this by using box openers, examples, exercises, and cases from various countries where the \$US, Euro, and Pound Sterling are employed.

## Learning statistics

Often students become afraid when they realize that they have to take a course in statistics as part of their college or university curriculum. I often hear remarks like:

*“I will never pass this course.” “I am no good at maths and so I am sure I will fail the exam.” “I don’t need a course in statistics as I am going to be in marketing.” “What good is statistics to me, I plan to take a job in human resources?”*

All these remarks are unwarranted and the knowledge of statistics is vital in all areas of business. The subject is made easier, and more fun, by using Microsoft Excel. To aid comprehension,

the textbook begins with fundamental ideas and then moves into more complex areas.

## The author

I have been in industry for over 20 years using statistics and then teaching the subject for the last 21 with considerable success using the subject material, and the approach given in this text. You will find the book shorter than many of the texts on the market but I have only presented those subject areas that in my experience give a solid foundation of statistical analysis for business, and that can be covered in a reasonable time frame. This text avoids working through tedious mathematical computations, often found in other statistical texts that I find which confuse students. You should not have any qualms about studying statistics – it really is not a difficult subject to grasp. If you need any further information, or have questions to ask, please do not hesitate to get in touch through the Elsevier website or at my e-mail address: [derek.waller@wanadoo.fr](mailto:derek.waller@wanadoo.fr).

*This page intentionally left blank*

# Presenting and organizing data

## How not to present data

*Steve was an undergraduate business student and currently performing a 6-month internship with Telephone Co. Today he was feeling nervous as he was about to present the results of a marketing study that he had performed on the sales of mobile telephones that his firm produced. There were 10 people in the meeting including Roger, Susan, and Helen three of the regional sales directors, Valerie Jones, Steve's manager, the Head of Marketing, and representatives from production and product development. Steve showed his first slide as illustrated in Table 1.1 with the comment that "This is the 200 pieces of raw sales data that I have collected". At first there was silence and then there were several very pointed comments. "What does all that mean?" "I just don't understand the significance of those figures?" "Sir, would you kindly interpret that data". After the meeting Valerie took Steve aside and said, "I am sorry Steve but you just have to remember that all of our people are busy and need to be presented information that gives them a clear and concise picture of the situation. The way that you presented the information is not at all what we expect".*

Table 1.1 Raw sales data (\$).

35,378	170,569	104,985	134,859	120,958	107,865	127,895	106,825	130,564	108,654
109,785	184,957	96,598	121,985	63,258	164,295	97,568	165,298	113,985	124,965
108,695	91,864	120,598	47,865	162,985	83,964	103,985	61,298	104,987	184,562
89,597	160,259	55,492	152,698	92,875	56,879	151,895	88,479	165,698	89,486
85,479	64,578	103,985	81,980	137,859	126,987	102,987	116,985	45,189	131,958
73,598	161,895	132,689	120,654	67,895	87,653	58,975	103,958	124,598	168,592
95,896	52,754	114,985	62,598	145,985	99,654	76,589	113,590	80,459	107,865
109,856	101,894	80,157	78,598	86,785	97,562	136,984	89,856	96,215	163,985
83,695	75,894	98,759	133,958	74,895	37,856	90,689	64,189	107,865	123,958
105,987	93,832	58,975	102,986	102,987	144,985	101,498	101,298	103,958	71,589
59,326	121,459	82,198	60,128	86,597	91,786	56,897	112,854	54,128	152,654
99,999	78,562	110,489	86,957	99,486	132,569	134,987	76,589	135,698	118,654
90,598	156,982	87,694	117,895	85,632	104,598	77,654	105,987	78,456	149,562
68,976	50,128	106,598	63,598	123,564	47,895	100,295	60,128	141,298	84,598
100,296	77,498	77,856	134,890	79,432	100,659	95,489	122,958	111,897	129,564
71,458	88,796	110,259	72,598	140,598	125,489	69,584	89,651	70,598	93,876
112,987	123,895	65,847	128,695	66,897	82,459	133,984	98,459	153,298	87,265
72,312	81,456	124,856	101,487	73,569	138,695	74,583	136,958	115,897	142,985
119,654	96,592	66,598	81,490	139,584	82,456	150,298	106,859	68,945	122,654
70,489	94,587	85,975	138,597	97,498	143,985	92,489	146,289	84,592	69,874

## Learning objectives

After you have studied this chapter you will be able to logically **organize** and **present** statistical data in a **visual** form so that you can **convince** your audience and **objectively** get your point across. You will learn how to develop the following support tools for both **numerical** and **categorical** data accordingly as follows.

- ✓ **Numerical data** • Types of numerical data • Frequency distribution • Absolute frequency histogram • Relative frequency histogram • Frequency polygon • Ogive • Stem-and-leaf display • Line graph
- ✓ **Categorical data** • Questionnaires • Pie chart • Vertical histogram • Parallel histogram • Horizontal bar chart • Parallel bar chart • Pareto diagram • Cross-classification or contingency table • Stacked histogram • Pictograms

As the box opener illustrates, in the business environment, it is vital to show data in a clear and precise manner so that everyone concerned understands the ideas and arguments being presented. Management people are busy and often do not have the time to make an in depth analysis of information. Thus a simple and coherent presentation is vital in order to get your message across.

### Numerical Data

Numerical data provide information in a quantitative form. For example, the house has 250 m<sup>2</sup> of living space. My gross salary last year was £70,000 and this year it has increased to £76,000. He ran the Santa Monica marathon in 3 hours and 4 minutes. The firm's net income last year was \$14,500,400. All these give information in a numerical form and clearly state a particular condition or situation. When data is collected it might be **raw data**, which is collected information that has not been organized. The next step after you have raw data is to organize this information and present it in a meaningful form. This section gives useful ways to present numerical data.

### Types of numerical data

Numerical data are most often either univariate or bivariate. **Univariate data** are composed of individual values that represent just one random variable,  $x$ . The information presented in Table 1.1 is univariate data. **Bivariate data** involves two variables,  $x$  and  $y$ , and any data that is subsequently put into graphical form would be bivariate since a value on the  $x$ -axis has a corresponding value on the  $y$ -axis.

### Frequency distribution

One way of organizing univariate data, to make it easier to understand, is to put it into a **frequency distribution**. A frequency distribution is a table, that can be converted into a graph, where the data are arranged into unique **groups**, **categories**, or **classes** according to the frequency, or how often, data values appear in a given class. By grouping data into classes, the data are more manageable than raw data and we can demonstrate clearly patterns in the information. Usually the greater the quantity of data then there should be more classes to clearly show the profile. A guide is to have at least 5 classes but no more than 15 although it really depends on the amount of data



available and what we are trying to demonstrate. In the frequency distribution, the class range or width should be the same such that there is coherency in data analysis. The **class range or class width** is given by the following relationship:

Class range or class width

$$= \frac{\text{Desired range of the complete frequency distribution}}{\text{Number of groups selected}} \quad 1(i)$$

The **range** is the difference between the highest and the lowest value of any set of data. Let us consider the sales data given in Table 1.1. If we use the **[function MAX]** in Excel, we obtain \$184,957 as the highest value of this data. If we use the **[function MIN]** in Excel it gives the lowest value of \$35,378. When we develop a frequency distribution we want to be sure that all of the data is contained within the boundaries that we establish. Thus, to develop a frequency distribution for these sales data, a logical

maximum value for presenting this data is \$185,000 (the nearest value in '000s above \$184,957) and a minimum value is \$35,000 (the nearest value in '000s below \$35,378). By using these upper and lower boundary limits we have included all of the 200 data items. If we want 15 classes then the class range or class width is given as follows using equation 1(i):

Class range or class width

$$= \frac{\$185,000 - \$35,000}{15} = \$10,000$$

The tabulated frequency distribution for the sales data using 15 classes is shown in Table 1.2. The 1st column gives the number of the class range, the 2nd gives the limits of the class range, and the 3rd column gives the amount of data in each range. The lower limit of the distribution is \$35,000 and each class increase by intervals of \$10,000 to the upper limit of \$185,000. In selecting a lower value of \$35,000 and an upper

Table 1.2 Frequency distribution of sales data.

Class no.	Class range (\$)	Amount of data in class	Percentage of data	Midpoint of class range
	>25,000 to ≤35,000	0	0.00	30,000
1	>35,000 to ≤45,000	2	1.00	40,000
2	>45,000 to ≤55,000	6	3.00	50,000
3	>55,000 to ≤65,000	14	7.00	60,000
4	>65,000 to ≤75,000	18	9.00	70,000
5	>75,000 to ≤85,000	22	11.00	80,000
6	>85,000 to ≤95,000	24	12.00	90,000
7	>95,000 to ≤105,000	30	15.00	100,000
8	>105,000 to ≤115,000	20	10.00	110,000
9	>115,000 to ≤125,000	18	9.00	120,000
10	>125,000 to ≤135,000	14	7.00	130,000
11	>135,000 to ≤145,000	12	6.00	140,000
12	>145,000 to ≤155,000	8	4.00	150,000
13	>155,000 to ≤165,000	6	3.00	160,000
14	>165,000 to ≤175,000	4	2.00	170,000
15	>175,000 to ≤185,000	2	1.00	180,000
	>185,000 to ≤195,000	0	0.00	190,000
Total		200	100.00	

value of \$185,000 we have included all the sales data values, and so the frequency distribution is called a **closed-ended frequency distribution** as all data is contained within the limits. (Note that in Table 1.2 we have included a line below \$35,000 of a class range  $>25,000$  to  $\leq 35,000$  and a line above \$185,000 of a class range  $>185,000$  to  $\leq 195,000$ . The reason for this will be explained in the later section entitled, “Frequency polygon”).

In order to develop the frequency distribution using Excel, you first make a single column of the class limits either in the same tab as the dataset or if you prefer in a separate tab. In this case the class limits are \$35,000 to \$185,000 in increments of \$10,000. You then highlight a virgin column, immediately adjacent to the class limits, of exactly the same height and with exactly the corresponding lines as the class limits. Then select **[function FREQUENCY]** in Excel and enter the dataset, that is the information in Table 1.1, and the class limits you developed that are demanded by the Excel screen. When these have been selected, you press the three keys, control-shift-enter [Ctrl - ↑ - ↵] simultaneously and this will give a frequency distribution of the amount of the data as shown in the 3rd column of Table 1.2. Note in the frequency distribution the cut-off points for the class limits. The value of \$45,000 falls in the class range,  $>\$35,000$  and  $\leq \$45,000$ , whereas \$45,001 is in the class range  $>\$45,000$  to  $\leq \$55,000$ . The percentage, or proportion of data, as shown in the 4th column of Table 1.2, is obtained by dividing the amount of data in a particular class by the total amount of data. For example, in the class width  $>\$45,000$  to  $\leq \$55,000$ , there are six pieces of data and  $6/200$  is 3.00%. This is a **relative frequency distribution** meaning that the percentage value is relative to the total amount of data available. Note that once you have created a frequency table or graph you are now making a presentation in bivariate form as all the  $x$  values have a corresponding  $y$  value.

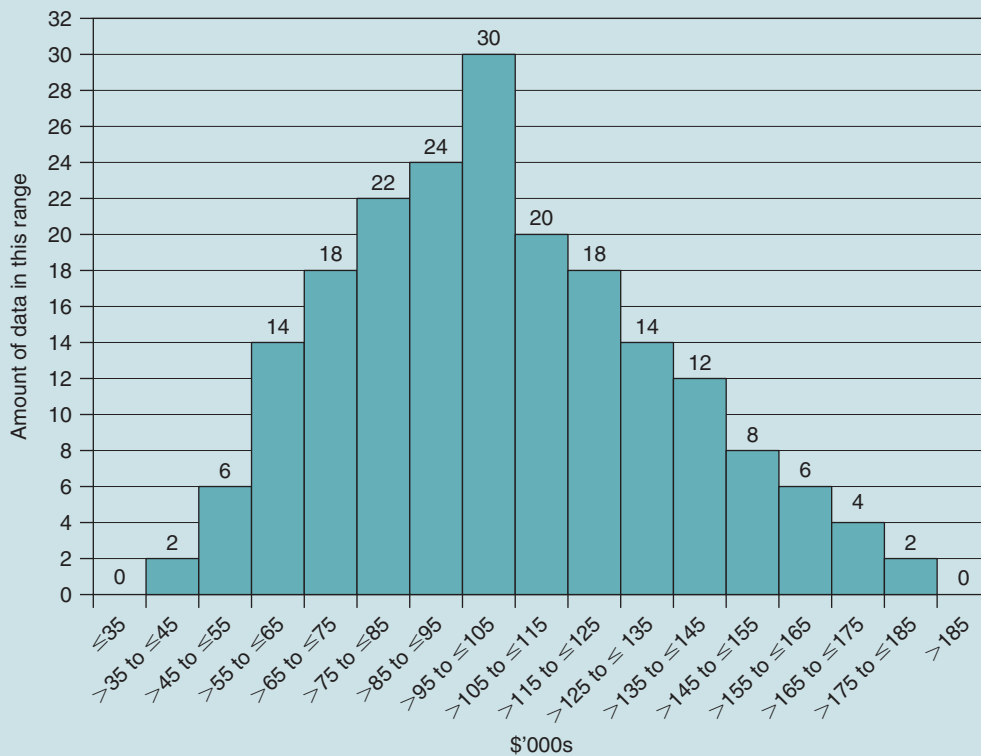
Note that in this example, when we calculated the class range or class width using the maximum and the minimum values for 15 classes we obtained a whole number of \$10,000. Whole numbers such as this make for clear presentations. However, if we wanted 16 classes then the class range would be \$9,375  $[(185,000 - 35,000)/16]$  which is not as convenient. In this case we can modify our maximum and minimum values to say 190,000 and 30,000 which brings us back to a class range of \$10,000  $[(190,000 - 30,000)/16]$ . Alternatively, we can keep the minimum value at \$35,000 and make the maximum value \$195,000 which again gives a class range of \$10,000  $[(195,000 - 35,000)/16]$ . In either case we still maintain a closed-limit frequency distribution.

## Absolute frequency histogram

Once a frequency distribution table has been developed we can convert this into a histogram, which is a visual presentation of the information, using the graphics capabilities in Excel. An **absolute frequency histogram** is a vertical bar chart drawn on an  $x$ - and  $y$ -axes. The horizontal, or  $x$ -axis, is a numerical scale of the desired class width where each class is of equal size. The vertical bars, defined by the  $y$ -axis, have a length proportional to the actual quantity of data, or to the frequency of the amount of data that occurs in a given class range. That is to say, the lengths of the vertical bars are dependent on, or a function of, the range selected by our class width.

Figure 1.1 gives an absolute frequency histogram for the sales data using the 3rd column from Table 1.2. Here we have 15 vertical bars whose lengths are proportional to the amount of contained data. The first bar contains data in the range  $>\$35,000$  to  $\leq \$45,000$ , the second bar has data in the range  $>\$45,000$  to  $\leq \$55,000$ , the third in the range  $>\$55,000$  to  $\leq \$65,000$ , etc. Above each bar is indicated the amount of

Figure 1.1 Absolute frequency distribution of sales data.



data that is included in each class range. There is no space shown between each bar since the class ranges move from one limit to another though each limit has a definite cut-off point. In presenting this information to say, the sales department, we can clearly see the pattern of the data and specifically observe that the amount of sales in each class range increases and then decreases beyond \$105,000. We can see that the greatest amount of sales of the sample of 200, 30 to be exact, lies in the range >\$95,000 to ≤\$ 105,000.

which is an alternative to the absolute frequency histogram where now the vertical bar, represented by the  $y$ -axis, is the percentage or proportion of the total data rather than the absolute amount. The relative frequency histogram of the sales data is given in Figure 1.2 where we have used the percent of data from the 4th column of Table 1.2. The shape of this histogram is identical to the histogram in Figure 1.1. We now see that for revenues in the range >\$95,000 to ≤\$105,000 the proportion of the total sales data is 15%.

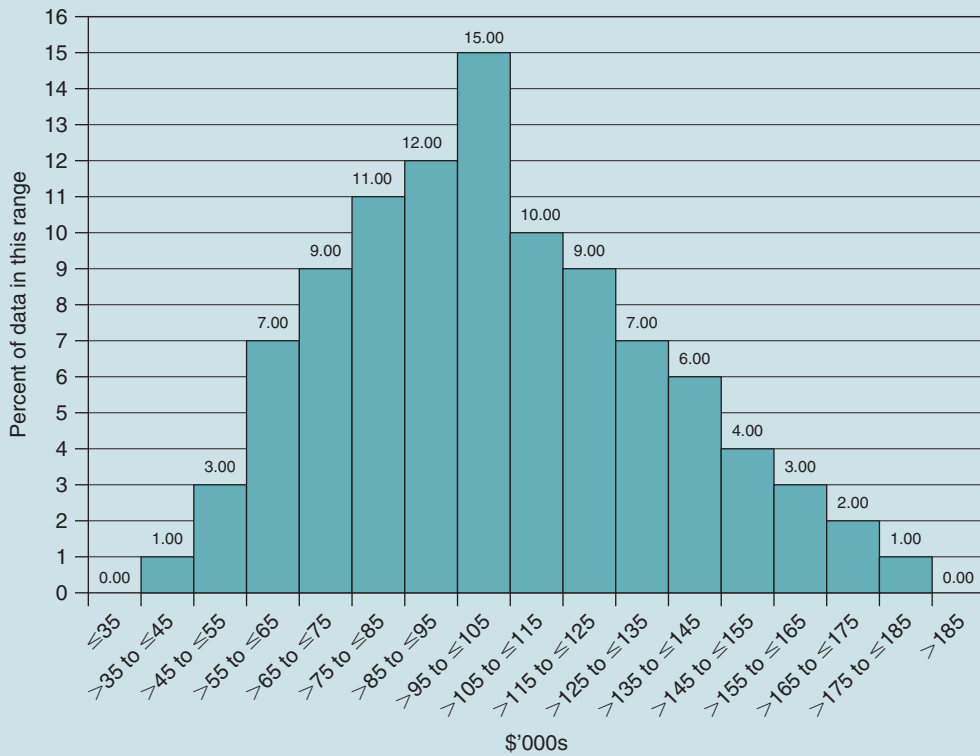
### Relative frequency histogram

Again using the graphics capabilities in Excel we can develop a **relative frequency histogram**,

### Frequency polygon

The absolute frequency histogram, or the relative frequency histogram, can be converted into

Figure 1.2 Relative frequency distribution of sales data.



a line graph or **frequency polygon**. The frequency polygon is developed by determining the midpoint of the class widths in the respective histogram. The **midpoint** of a class range is,

$$\frac{(\text{maximum value} + \text{minimum value})}{2}$$

For example, the midpoint of the class range, >\$95,000 to ≤\$105,000 is,

$$\frac{(95,000 + 105,000)}{2} = \frac{200,000}{2} = 100,000$$

The midpoints of all the class ranges are given in the 5th column of Table 1.2. Note that we

have given an entry, >\$25,000 to ≤\$35,000 and an entry of >\$185,000 to ≤\$195,000 where here the amount of data in these class ranges is zero since in these ranges we are beyond the limits of the closed-ended frequency distribution. In doing this we are able to construct a frequency polygon, which cuts the  $x$ -axis for a  $y$ -value of zero. Figure 1.3 gives the absolute frequency polygon and the relative frequency polygon is shown in Figure 1.4. These polygons are developed using the graphics capabilities in Excel where the  $x$ -axis is the midpoint of the class width and the  $y$ -axis is the frequency of occurrence. Note that the relative frequency polygon has an identical form as the absolute frequency polygon of Figure 1.3 but the

Figure 1.3 Absolute frequency polygon of sales data.

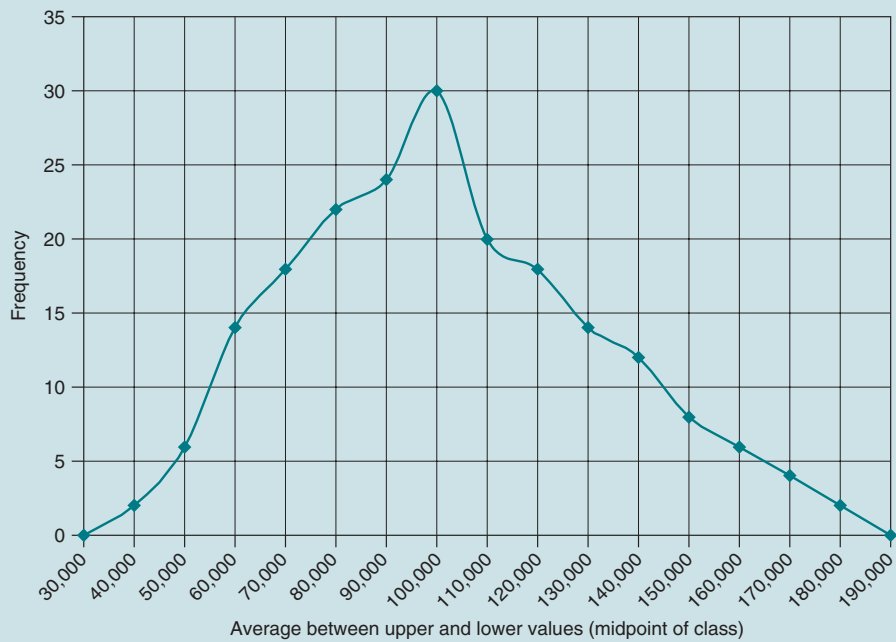
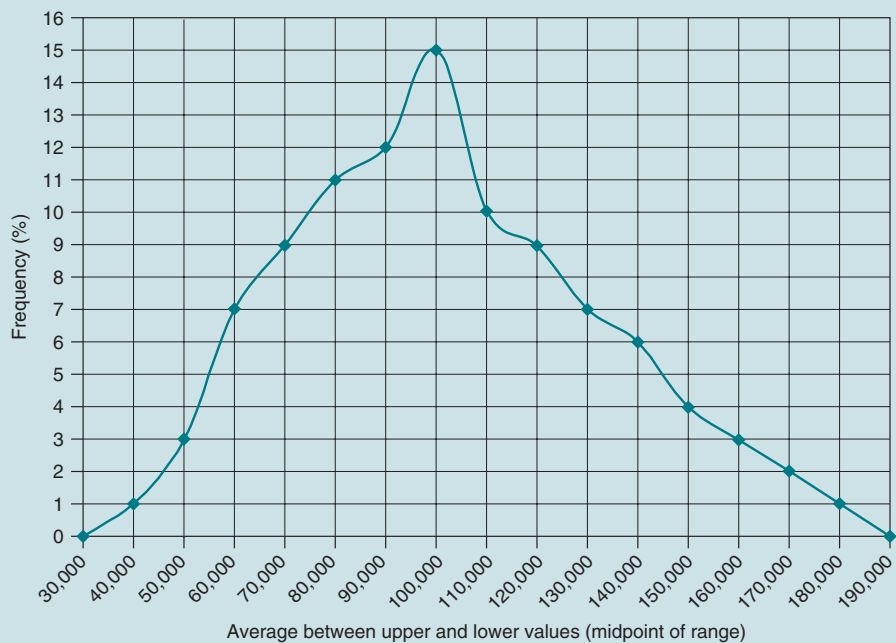


Figure 1.4 Relative frequency polygon of sales data.



$y$ -axis is a percentage, rather than an absolute scale. The difference between presenting the data as a frequency polygon rather than a histogram is that you can see the continuous flow of the data.

## Ogive

An **ogive** is an adaptation of a frequency distribution, where the data values are progressively totalled, or cumulated, such that the resulting table indicates how many, or the proportion of, observations that lie above or below certain limits. There is a **less than ogive**, which indicates the amount of data below certain limits. This ogive, in graphical form, has a positive slope

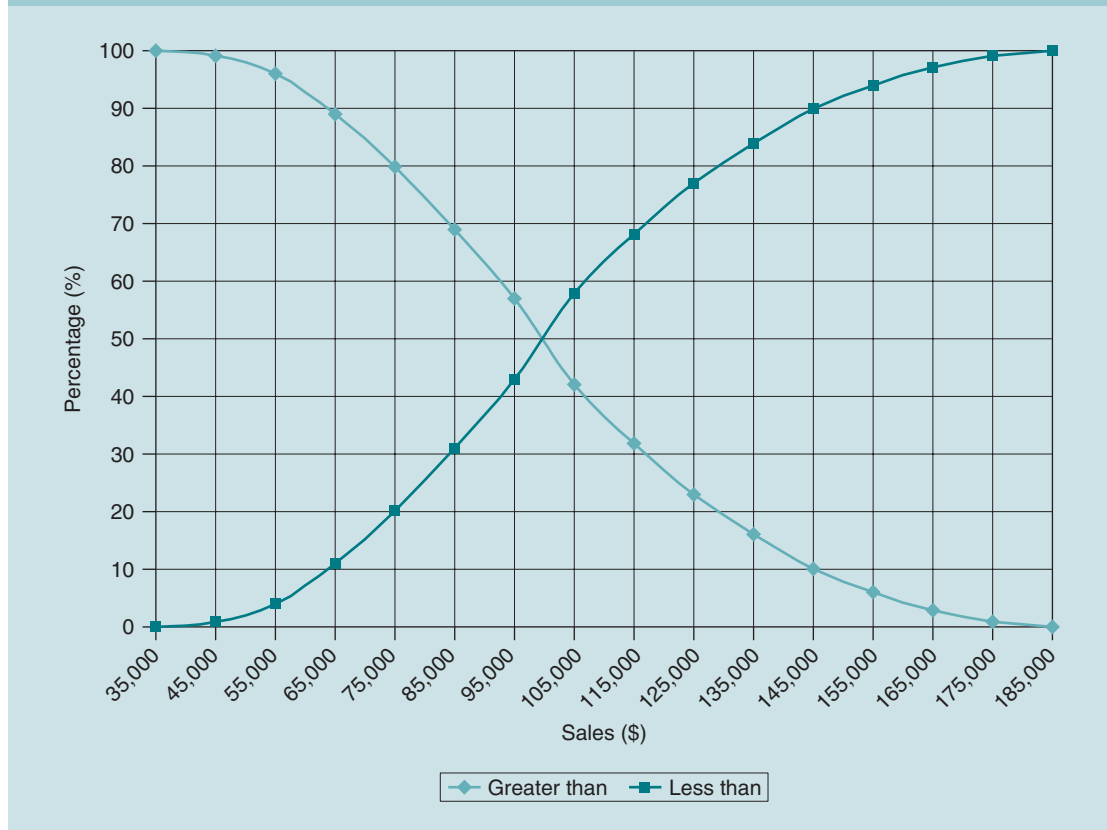
such that the  $y$  values increase from left to right. The other is a **greater than ogive** that illustrates data above certain values. It has a negative slope, where the  $y$  values decrease from left to right.

The frequency distribution data from Table 1.2 has been converted into an ogive format and this is given in Table 1.3, which shows the cumulated data in an absolute form and a relative form. The relative frequency ogives, developed from this data, are given in Figure 1.5. The usefulness of these graphs is that interpretations can be easily made. For example, from the greater than ogive we can see that 80.00% of the sales revenues are at least \$75,000. Alternatively, from the less than ogive, we can

Table 1.3 Ogives of sales data.

Class limit, $n$	Range of class limits ('000s)	Ogive using absolute data			Ogive using relative data		
		No. $\leq n$ but $> (n - 1)$	No. $>$ class limit, $n$	Number $<$ limit	Percentage age $\leq n$ but $> (n - 1)$	Percentage age $>$ class limit, $n$	Percentage $\leq$ limit
25,000							
35,000	$\leq 35$	0	200	0	0.00	100.00	0.00
45,000	$> 35$ to $\leq 45$	2	198	2	1.00	99.00	1.00
55,000	$> 45$ to $\leq 55$	6	192	8	3.00	96.00	4.00
65,000	$> 55$ to $\leq 65$	14	178	22	7.00	89.00	11.00
75,000	$> 65$ to $\leq 75$	18	160	40	9.00	80.00	20.00
85,000	$> 75$ to $\leq 85$	22	138	62	11.00	69.00	31.00
95,000	$> 85$ to $\leq 95$	24	114	86	12.00	57.00	43.00
105,000	$> 95$ to $\leq 105$	30	84	116	15.00	42.00	58.00
115,000	$> 105$ to $\leq 115$	20	64	136	10.00	32.00	68.00
125,000	$> 115$ to $\leq 125$	18	46	154	9.00	23.00	77.00
135,000	$> 125$ to $\leq 135$	14	32	168	7.00	16.00	84.00
145,000	$> 135$ to $\leq 145$	12	20	180	6.00	10.00	90.00
155,000	$> 145$ to $\leq 155$	8	12	188	4.00	6.00	94.00
165,000	$> 155$ to $\leq 165$	6	6	194	3.00	3.00	97.00
175,000	$> 165$ to $\leq 175$	4	2	198	2.00	1.00	99.00
185,000	$> 175$ to $\leq 185$	2	0	200	1.00	0.00	100.00
195,000	$> 185$	0			0.00	0.00	
Total		200			100.00		

Figure 1.5 Relative frequency ogives of sales data.



see that 90.00% of the sales are no more than \$145,000. The ogives can also be presented as an absolute frequency ogive by indicating on the  $y$ -axis the number of data entries which lie above or below given values. This is shown for the sales data in Figure 1.6. Here we see, for example, that 60 of the 200 data points are sales data that are less than \$85,000. The relative frequency ogive is probably more useful than the absolute frequency ogive as proportions or percentages are more meaningful and easily understood than absolute values. In the latter case, we would need to know to what base we are referring. In this case a **sample** of 200 pieces of data.

## Stem-and-leaf display

Another way of presenting data according to the frequency of occurrence is a **stem-and-leaf display**. This organizes data showing how values are distributed and cluster around the range of observations in the dataset. The display separates data entries into leading digits, or **stems** and trailing digits, or **leaves**. A stem-and-leaf display shows all individual data entries whereas a frequency distribution groups data into class ranges.

Let us consider the raw data that is given in Table 1.4, which is the sales receipts, in £'000s for one particular month for 60 branches of a supermarket in the United Kingdom. First the

Figure 1.6 Absolute frequency ogives of sales data.

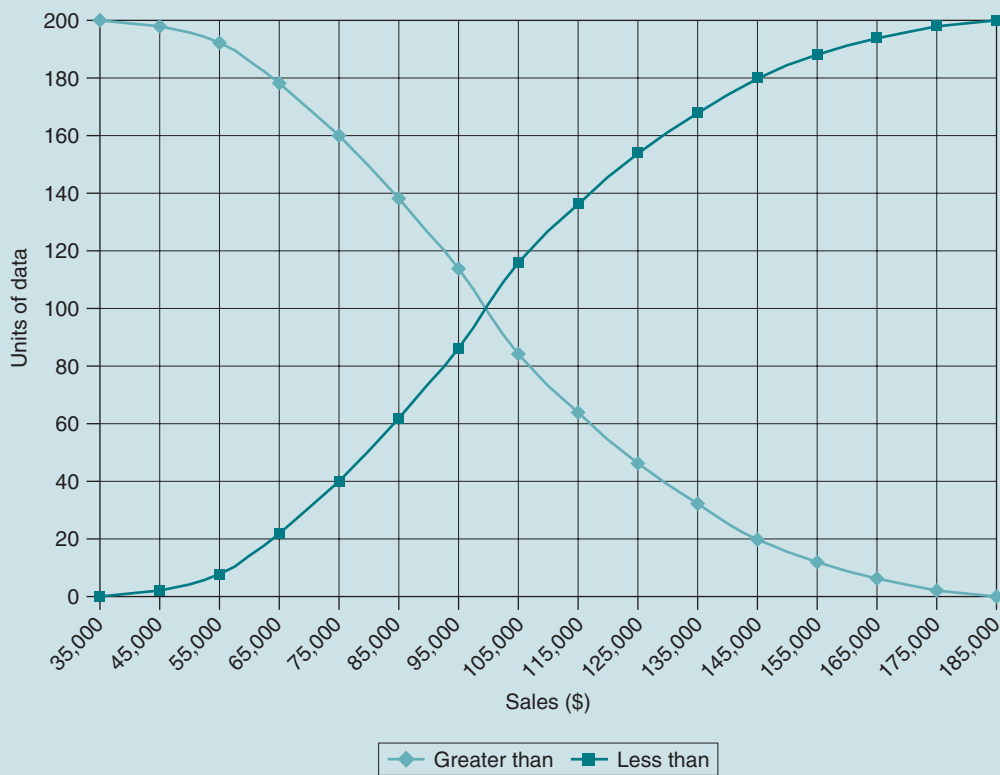


Table 1.4 Raw data of sales revenue from a supermarket (£'000s).

15.5	7.8	12.7	15.6	14.8	8.5	11.5	13.5	8.8	9.8
10.7	16.0	9.0	9.1	13.6	14.5	8.9	11.7	11.5	14.9
15.4	16.0	16.1	13.8	9.2	13.1	15.8	13.2	12.6	10.9
12.9	9.6	12.1	15.2	11.9	10.4	10.6	13.7	14.4	13.8
9.6	12.0	11.0	10.5	12.4	11.5	11.7	14.1	11.2	12.2
12.5	10.8	10.0	11.1	10.2	11.2	14.2	11.0	12.1	12.5

data is sorted from lowest to the highest value using the Excel command [SORT] from the menu bar **Data**. This gives an **ordered dataset** as shown in Table 1.5. Here we see that the lowest values are in the seven thousands while the highest are in the sixteen thousands. For the stem and leaf

we have selected the thousands as the stem, or those values to the left of the decimal point, and the leaf as the hundreds, or those values to the right of the decimal point. The stem-and-leaf display appears in Figure 1.7. The stem that has a value of 11 indicates the data that occurs most



Table 1.5 Ordered data of sales revenue from a supermarket (£'000s).

7.8	8.5	8.8	8.9	9.0	9.1	9.2	9.6	9.6	9.8
10.0	10.2	10.4	10.5	10.6	10.7	10.8	10.9	11.0	11.0
11.1	11.2	11.2	11.5	11.5	11.5	11.7	11.7	11.9	12.0
12.1	12.1	12.2	12.4	12.5	12.5	12.6	12.7	12.9	13.1
13.2	13.5	13.6	13.7	13.8	13.8	14.1	14.2	14.4	14.5
14.8	14.9	15.2	15.4	15.5	15.6	15.8	16.0	16.0	16.1

Figure 1.7 Stem-and-leaf display for the sales revenue of a supermarket (£'000s).

Stem	Leaf											No. of items
7	8											1
8	5	8	9									3
9	0	1	2	6	6	8						6
10	0	2	4	5	6	7	8	9				8
11	0	0	1	2	2	5	5	5	7	7	9	11
12	0	1	1	2	4	5	5	6	7	9		10
13	1	2	5	6	7	8	8					7
14	1	2	4	5	8	9						6
15	2	4	5	6	8							5
16	0	0	1									3
Total												60

frequently or in this case, those sales from £11,000 to less than £12,000.

The frequency distribution for the same data is shown in Figure 1.8. The pattern is similar to the stem-and-leaf display but the individual values are not shown. Note that in the frequency distribution, the  $x$ -axis has the range greater than the lower thousand value while the stem-and-leaf display includes this value. For example, in the stem-and-leaf display, 11.0 appears in the stem 11 to less than 12. In the frequency distribution, 11.0 appears in the class range  $>10$  to  $\leq 11$ . Alternatively, in the stem that has a value of 16 there are three values (16.0; 16.0; 16.1), whereas in the frequency distribution for the class  $>16$  to  $\leq 17$  there is only one value (16.1) as 16.0 is not greater than 16. These differences are simply because this is the way that the

frequency function operates in Microsoft Excel.

If you have no add-on stem-and-leaf display in Excel (a separate package) then the following is a way to develop the display using the basic Excel program:

- Arrange all the raw data in a horizontal line.
- Sort the data in ascending order by line. (Use the Excel function **SORT** in the menu bar **Data**.)
- Select the stem values and place in a column.
- Transpose the ordered data into their appropriate stem giving just the leaf value. For example, if there is a value 9.75 then the stem is 9, and the leaf value is 75.

Another approach to develop a stem-and-leaf display is not to sort the data but to keep it in its raw form and then to indicate the leaf values in chronological order for each stem. This has a disadvantage in that you do not see immediately which values are being repeated. A stem-and-leaf display is one of the techniques in **exploratory data analysis** (EDA), which are those methods that give a sense or initial feel about data being studied. A box and whisker plot discussed in Chapter 2 is also another technique in EDA.

## Line graph

A **line graph**, or usually referred to just as a graph, gives bivariate data on the  $x$ - and  $y$ -axes. It illustrates the relationship between the variable

Figure 1.8 Frequency distribution of the sales revenue of a supermarket (£).

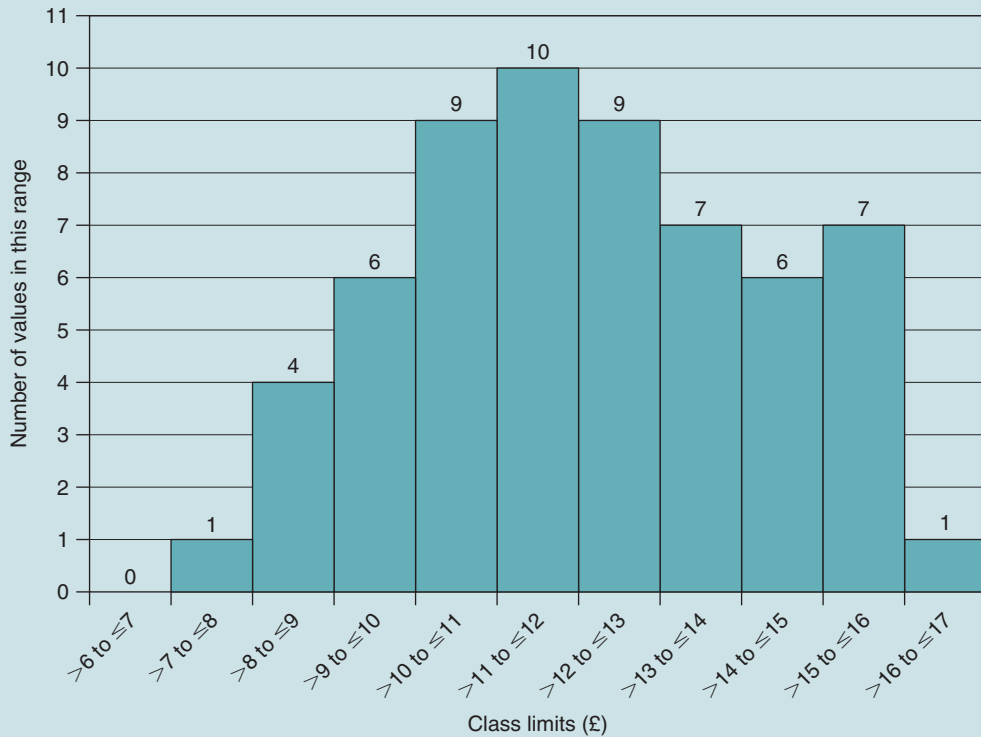


Table 1.6 Sales data for the last 12 years.

Period	Year	Sales (\$'000s)
1	1992	1,775
2	1993	2,000
3	1994	2,105
4	1995	2,213
5	1996	2,389
6	1997	2,415
7	1998	2,480
8	1999	2,500
9	2000	2,665
10	2001	2,810
11	2002	2,940
12	2003	3,070

on the  $x$ -axis and the corresponding value on the  $y$ -axis. If time represents part of the data this is always shown in the  $x$ -axis. A line graph is not necessarily a straight line but can be curvilinear. Attention has to be paid to the scales on the axes as the appearance of the graph can change and decision-making can be distorted. Consider for example, the sales revenues given in Table 1.6 for the 12-year period from 1992 to 2003.

Figure 1.9 gives the graph for this sales data where the  $y$ -axis begins at zero and the increase on the axis is in increments of \$500,000. Here the slope of the graph, illustrating the increase in sales each year, is moderate. Figure 1.10 now shows the same information except that the  $y$ -axis starts at the value of \$1,700,000 and the

Figure 1.9 Sales data for the last 12 years for "Company A".

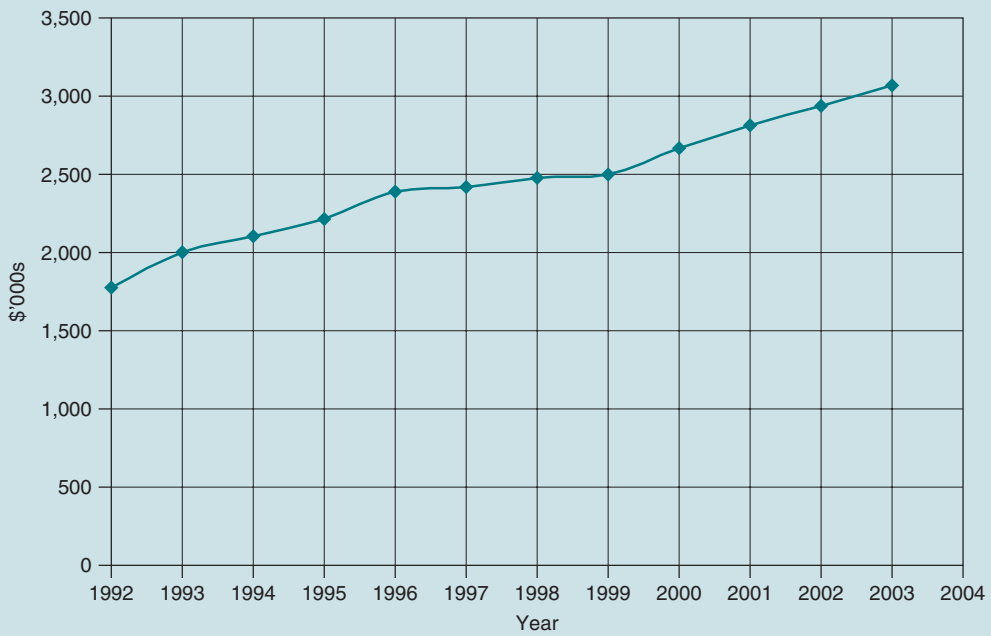
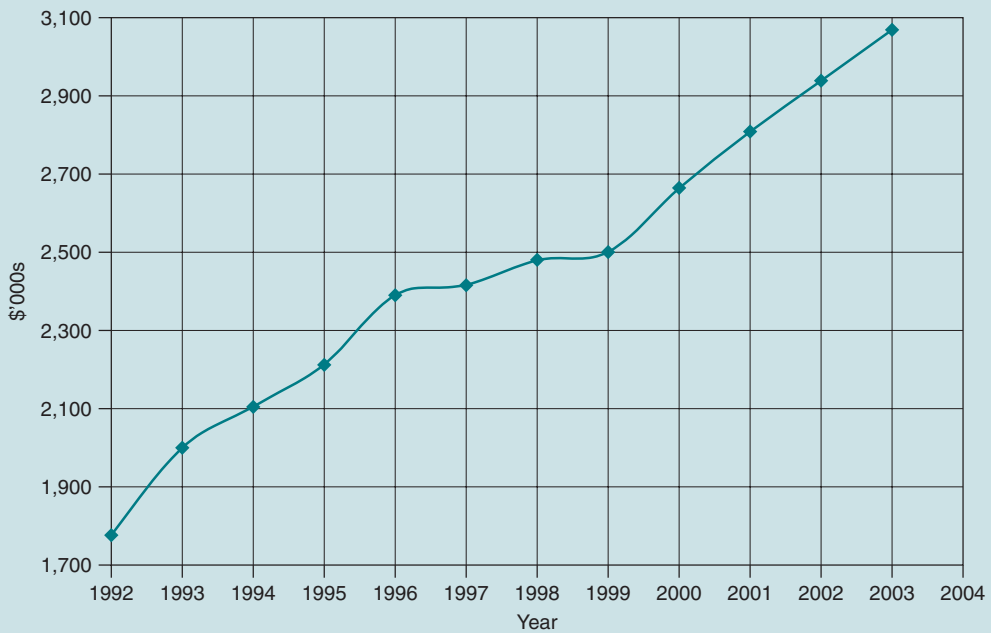


Figure 1.10 Sales data for the last 12 years for "Company B".



incremental increase is \$200,000 or 2.5 times smaller than in Figure 1.9. This gives the impression that the sales growth is very rapid, which is why the two figures are labelled “Company A” and “Company B”. They are of course the same company. Line graphs are treated further in Chapter 10.

## Categorical Data

Information that includes a qualitative response is **categorical data** and for this information there may be no quantitative data. For example, the house is the largest on the street. My salary increased this year. He ran the Santa Monica marathon in a fast time. Here the categories are large, increased, and fast. The responses, “Yes” or “No”, to a survey are also categorical data. Alternatively categorical data may be developed from numerical data, which is then organized and given a label, a category, or a name. For example, a firm’s sales revenues, which are quantitative data, may be presented according to geographic region, product type, sales agent, business unit, etc. A presentation of this type can be important to show the strength of the firm.

## Questionnaires

Very often we use **questionnaires** in order to evaluate customers’ perception of service level, students’ appreciation of a university course, or subscribers’ opinion of a publication. We do this

because we want to know if we are “doing it right” and if not what changes should we make. A questionnaire may take the form as given in Table 1.7. The first line is the **category** of the response. This is obviously subjective information. For example with a university course, Student A may have a very different opinion of the same programme as Student B. We can give the categorical response a **score**, or a quantitative value for the subjective response, as shown in the second line. Then, if the number of responses is sufficiently large, we can analyse this data in order to obtain a reasonable opinion of say the university course. The analysis of this type of questionnaire is illustrated in Chapter 2, and there is additional information in Chapter 6.

## Pie chart

If we have numerical data, this can be converted into a **pie chart** according to desired categories. A pie chart is a circle representing the data and divided into segments like portions of a pie. Each segment of the pie is proportional to the total amount of data it represents and can be labelled accordingly. The complete pie represents 100% of the data and the usefulness of the pie chart is that we can see clearly the pattern of the data. As an illustration, the sales data of Table 1.1 has now been organized by country and this tabular information is given in Table 1.8 together with the percentage amount of data for each country. This information, as a pie chart, is shown in Figure 1.11. We can clearly see now what the data represents and the contribution from each geographical territory. Here for example, the United Kingdom has the greatest contribution to sales revenues, and Austria the least. When you develop a pie chart for data, if you have a category called “other” be sure that this proportion is small relative to all the other categories in the pie chart; otherwise, your audience will question what is included in this mysterious “other” slice. When you develop a pie chart you can

Table 1.7 A scaled questionnaire.

Category	Very poor	Poor	Satisfactory	Good	Very good
Score	1	2	3	4	5

**Table 1.8** Raw sales data according to country (\$).

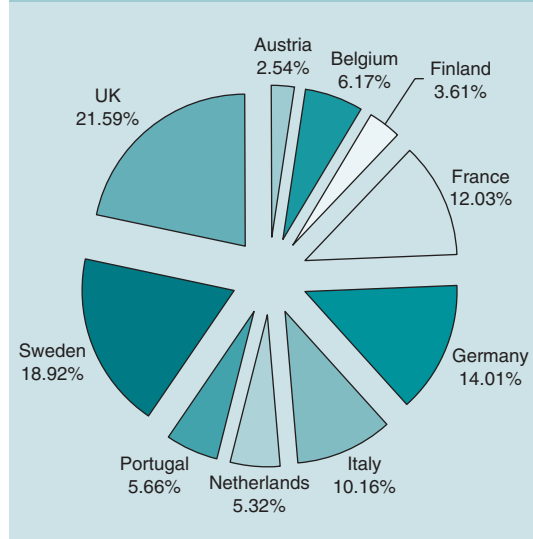
Group	Country	Sales revenues (\$)	Percentage
1	Austria	522,065	2.54
2	Belgium	1,266,054	6.17
3	Finland	741,639	3.61
4	France	2,470,257	12.03
5	Germany	2,876,431	14.01
6	Italy	2,086,829	10.16
7	Netherlands	1,091,779	5.32
8	Portugal	1,161,479	5.66
9	Sweden	3,884,566	18.92
10	United Kingdom	4,432,234	21.59
Total		20,533,333	100.00

only have two columns, or two rows of data. One column, or row, is the category, and the adjacent column, or row, is the numerical data. Note that in developing a pie chart in Excel you do not have to determine the percentage amount in the table. The graphics capability in Excel does this automatically.

## Vertical histogram

An alternative to a pie chart is to illustrate the data by a **vertical histogram** where the vertical bars on the *y*-axis show the percentage of data, and the *x*-axis the categories. Figure 1.12 gives an absolute histogram of the above pie chart sales information where the vertical bars show the absolute total sales and the *x*-axis has now been given a category according to geographic region. Figure 1.13 gives the relative frequency histogram for this same information where the *y*-axis is now a percentage scale. Note, in these histograms, the bars are separated, as one category does not directly flow to another, as is the case of a histogram of a complete numerically based frequency distribution.

**Figure 1.11** Pie chart for sales.



## Parallel histogram

A **parallel or side-by-side histogram** is useful to compare categorical data often of different time periods as illustrated in Figure 1.14. The figure shows the unemployment rate by country for two different years. From this graph we can compare the change from one period to another.<sup>1</sup>

## Horizontal bar chart

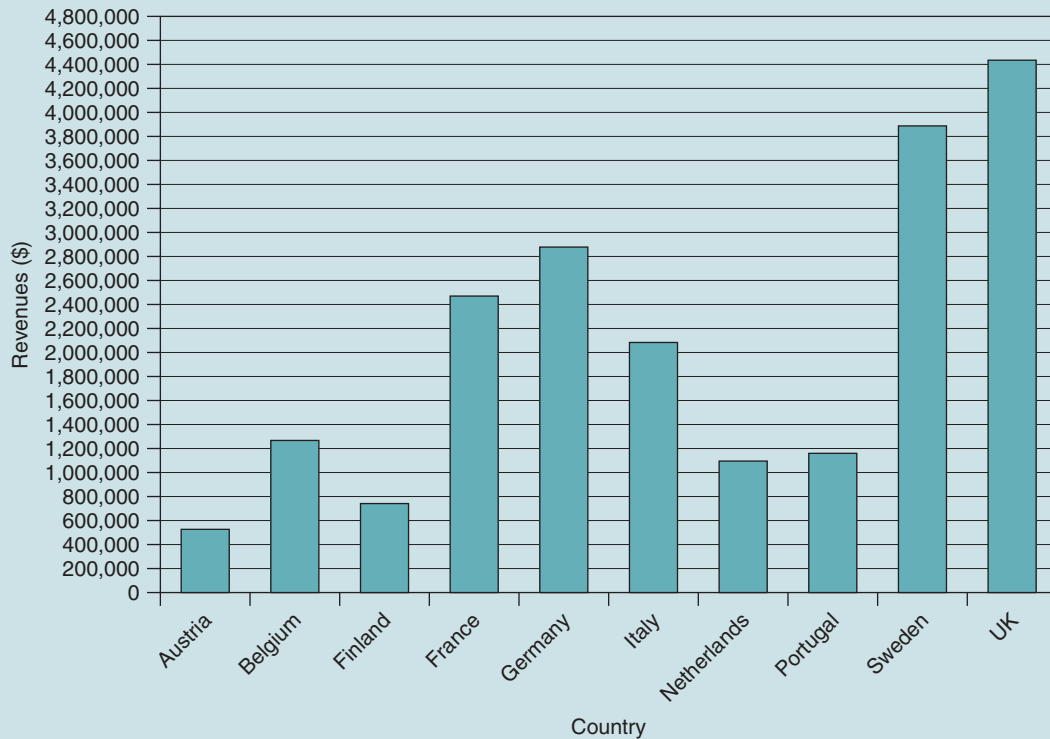
A **horizontal bar chart** is a type of histogram where the *x*- and *y*-axes are reversed such that the data are presented in a horizontal, rather than a vertical format. Figure 1.15 gives a bar chart for the sales data. Horizontal bar charts are sometimes referred to as Gantt charts after the American engineer Henry L. Gantt (1861–1919).

## Parallel bar chart

Again like the histogram, a **parallel or side-by-side bar chart** can be developed. Figure 1.16 shows a

<sup>1</sup> Economic and financial indicators, *The Economist*, 15 February 2003, p. 98.

Figure 1.12 Histogram of sales – absolute revenues.



side-by-side bar chart for the unemployment data of Figure 1.14.

## Pareto diagram

Another way of presenting data is to combine a line graph with a categorical histogram as shown in Figure 1.17. This illustrates the problems, according to categories, that occur in the distribution by truck of a chemical product. The  $x$ -axis gives the categories and the left-hand  $y$ -axis is the percent frequency of occurrence according to each of these categories with the vertical bars indicating their magnitude. The line graph that is shown now uses the right-hand  $y$ -axis and the same  $x$ -axis. This is now the cumulative frequency

of occurrence of each category. If we assume that the categories shown are exhaustive, meaning that all possible problems are included, then the line graph increases to 100% as shown. Usually the presentation is illustrated so that the bars are in descending order from the most important on the left to the least important on the right so that we have an organized picture of our situation. This type of presentation is known as a **Pareto diagram**, (named after the Italian economist, Vilfredo Pareto (1848–1923), who is also known for coining the 80/20 rule often used in business). The Pareto diagram is a visual chart used often in quality management and operations auditing as it shows those categorical areas that are the most critical and perhaps should be dealt with first.

Figure 1.13 Histogram of sales as a percentage.

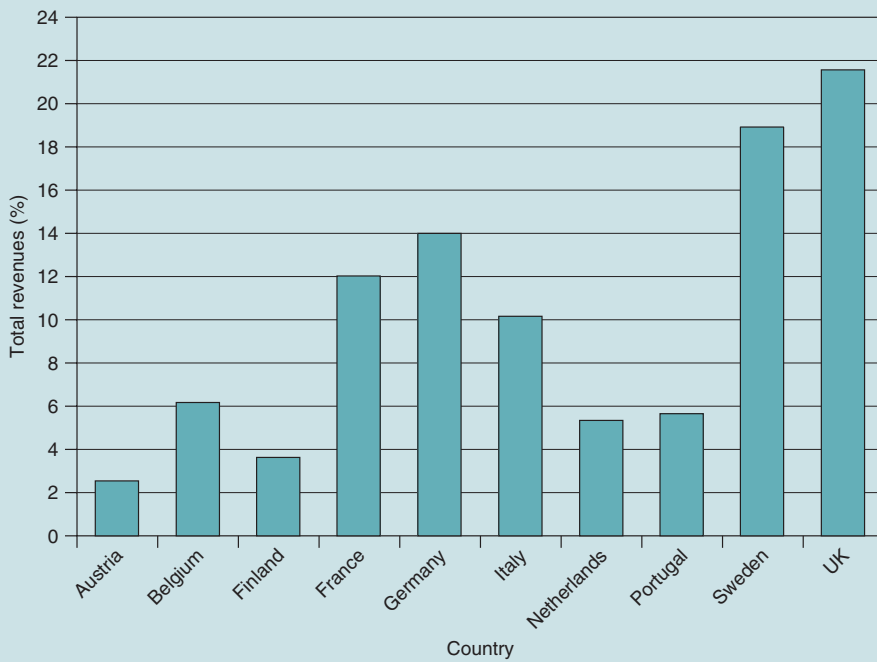


Figure 1.14 Unemployment rate.

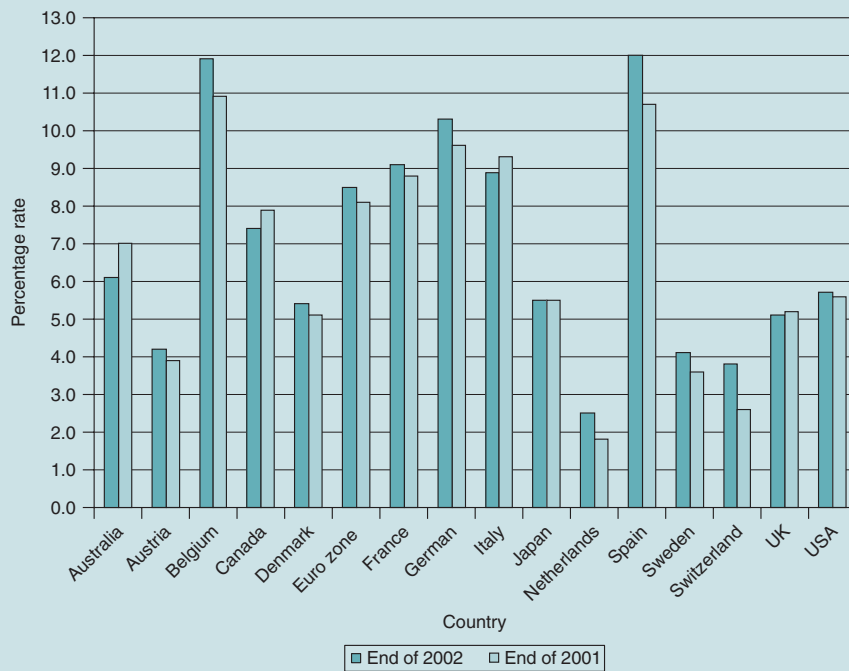


Figure 1.15 Bar chart for sales revenues.

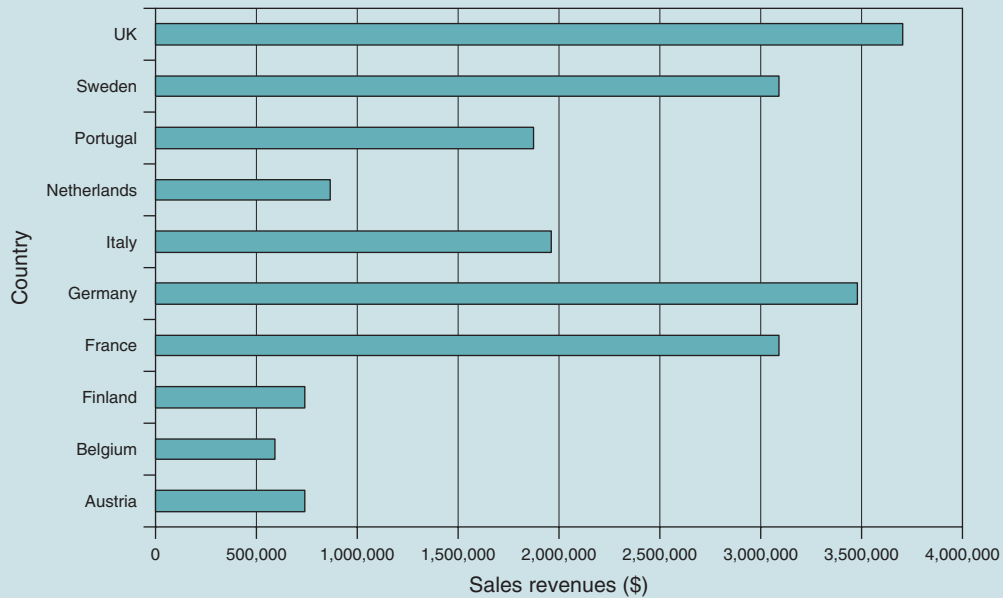


Figure 1.16 Unemployment rate.

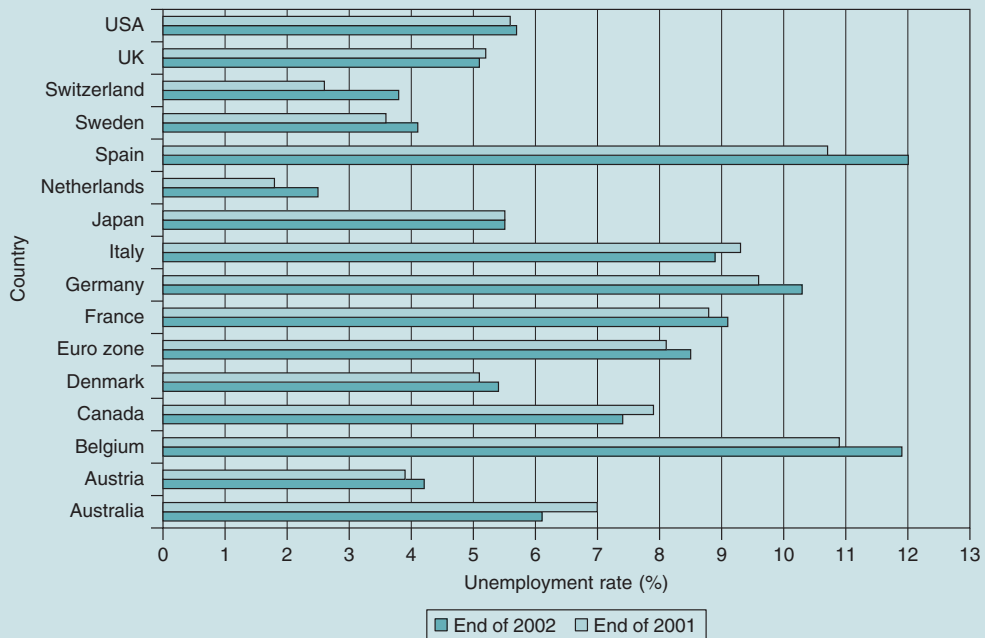
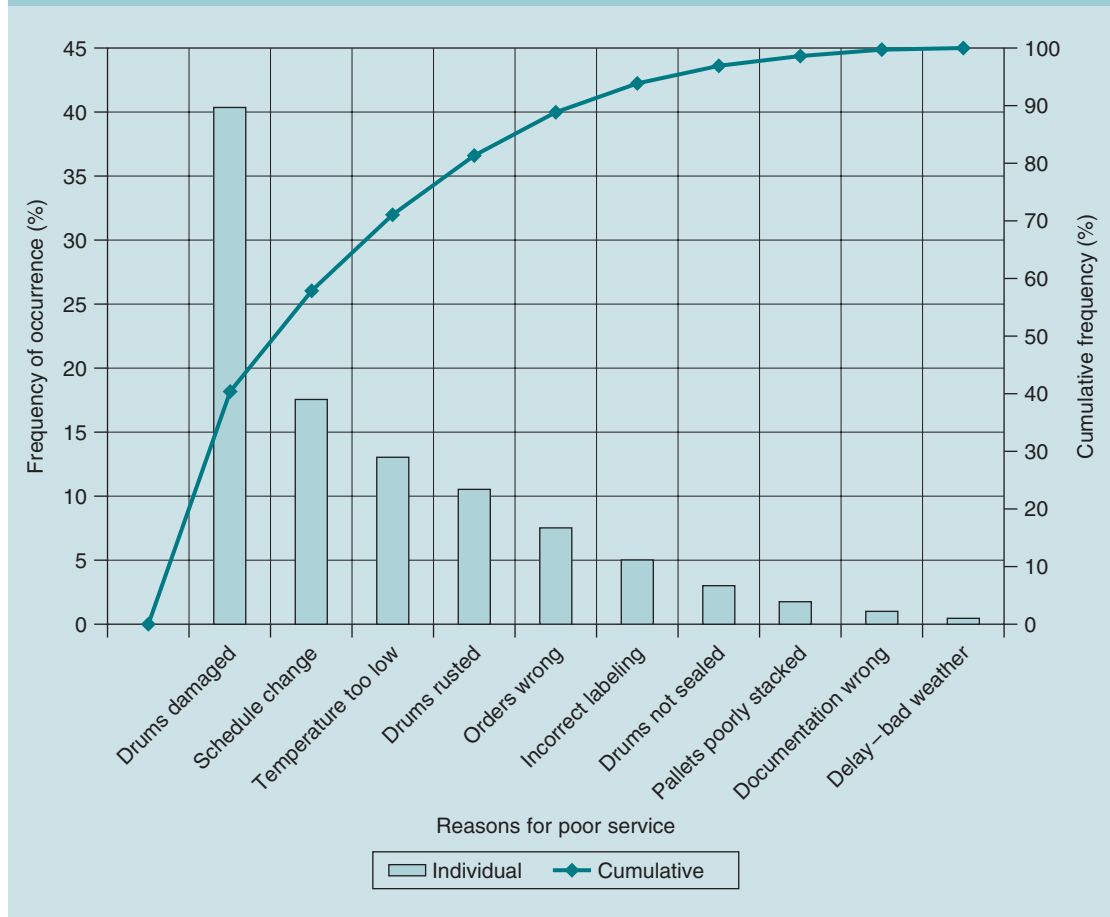




Figure 1.17 Pareto analysis for the distribution of chemicals.



## Cross-classification or contingency table

A **cross-classification** or **contingency table** is a way to present data when there are several variables and we are trying to indicate the relationship between one variable and another. As an illustration, Table 1.9 gives a cross-classification table for a sample of 1,550 people in the United States and their professions according to certain states. From this table we can say, for example, that 51 of the teachers are **contingent** of residing in Vermont. Alternatively, we can say that 24 of the residents of South Dakota are contingent of working for the government.

(Contingent means that values are dependent or conditioned on something else.)

## Stacked histogram

Once you have developed the cross-classification table you can present this visually by developing a **stacked histogram**. Figure 1.18 gives a stacked histogram for the cross-classification in Table 1.9 according to the state of employment. Portions of the histogram indicate the profession. Alternatively, Figure 1.19 gives a stacked histogram for the same table but now according to profession. Portions of the histogram now give the state of residence.

Table 1.9 Cross-classification or contingency table for professions in the United States.

State	Engineering	Teaching	Banking	Government	Agriculture	Total
California	20	19	12	23	23	97
Texas	34	62	15	51	65	227
Colorado	42	32	23	42	26	165
New York	43	40	23	35	54	195
Vermont	12	51	37	25	46	171
Michigan	24	16	15	16	35	106
South Dakota	34	35	12	24	25	130
Utah	61	25	19	29	61	195
Nevada	12	32	18	31	23	116
North Carolina	6	62	14	41	25	148
Total	288	374	188	317	383	1,550

Figure 1.18 Stacked histogram by state in the United States.

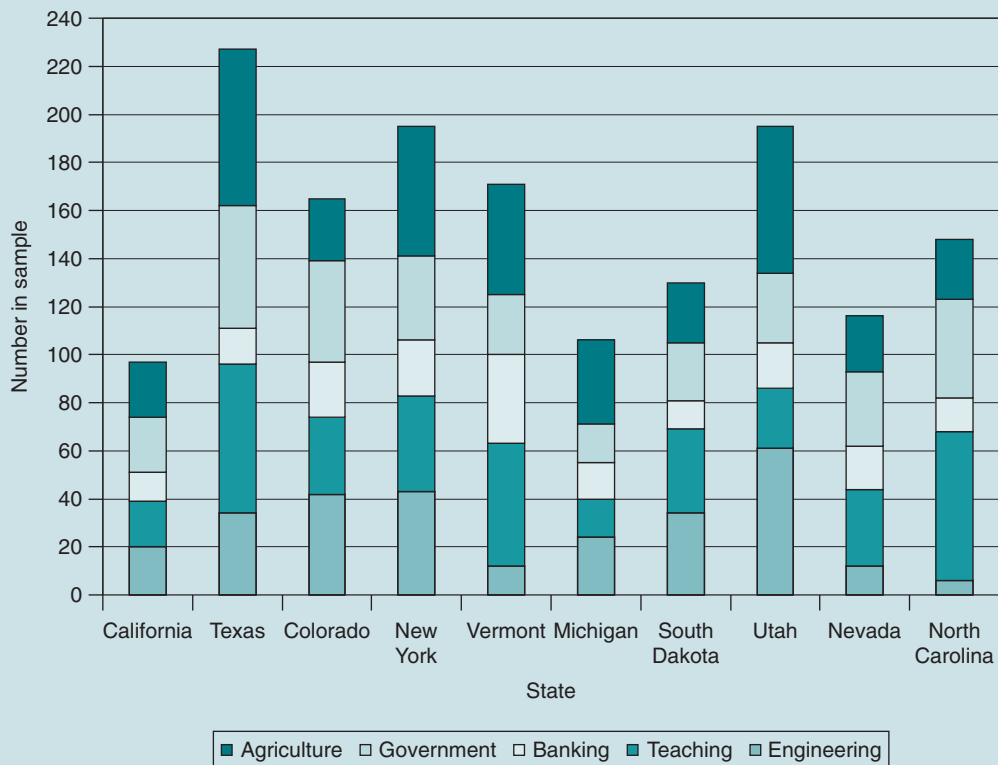


Figure 1.19 Stacked histogram by profession in the United States.

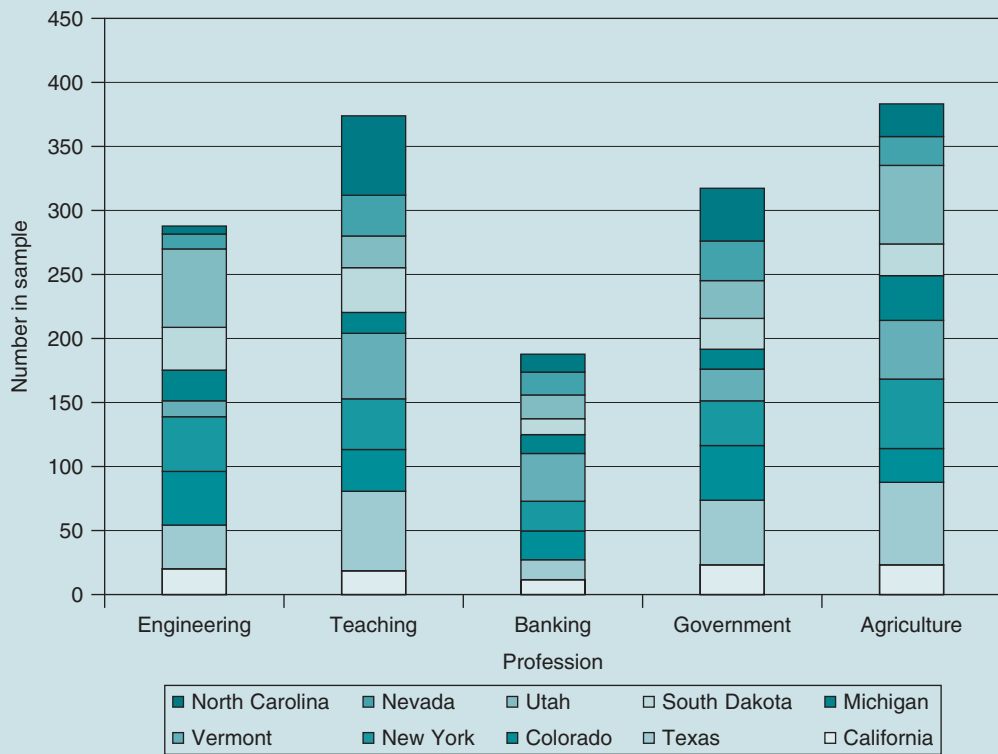


Figure 1.20 A pictogram to illustrate inflation.



The value of your money today



The value of your money tomorrow

## Pictograms

A **pictogram** is a picture, icon, or sketch that represents quantitative data but in a categorical, qualitative, or comparative manner. For example, a coin might be shown divided into sections indicating that portion of sales revenues that go to taxes, operating costs, profits, and capital expenditures. Magazines such as Business Week, Time, or Newsweek make heavy use of pictograms.

**Pictograph** is another term often employed for pictogram. Figure 1.20 gives an example of how inflation might be represented by showing a large sack of money for today, and a smaller sack for tomorrow. Attention must be made when using pictograms as they can easily distort the real facts of the data. For example in the figure given, has our money been reduced by a factor of 50%, 100%, or 200%? We cannot say clearly. Pictograms are not covered further in this textbook.

This chapter has presented several tools useful for presenting data in a concise manner with the objective of clearly getting your message across to an audience. The chapter is divided into discussing numerical and categorical data.

### Numerical data

Numerical data is most often univariate, or data with a single variable, or bivariate which is information that has two related variables. Univariate data can be converted into a frequency distribution that groups the data into classes according to the frequency of occurrence of values within a given class. A frequency distribution can be simply in tabular form, or alternatively, it can be presented graphically as an absolute, or relative frequency, histogram. The advantage of a graphical display is that you see clearly the quantity, or proportion of information, that appear in defined classes. This can illustrate key information such as the level of your best, or worst, revenues, costs, or profits. A histogram can be converted into a frequency polygon which links the midrange of each of the classes. The polygon, either in absolute or relative form, gives the pattern of the data in a continuous form showing where major frequencies occur. An extension of the frequency distribution is the less than, or greater than ogive. The usefulness of ogive presentations is that it is visually apparent the amount, or percentage, that is more or less than certain values and may be indicators of performance. A stem-and-leaf display, a tool in EDA, is a frequency distribution where all data values are displayed according to stems, or leading values, and leaves, or trailing values of the data. The commonly used line graph is a graphical presentation of bivariate data correlating the  $x$  variable with its  $y$  variable. Although we use the term line graph, the display does not have to be a straight line but it can be curvilinear or simply a line that is not straight!

### Categorical data

Categorical data is information that includes qualitative or non-quantitative groupings. Numerical data can be represented in a categorical form where parts of the numerical values are put into a category such as product type or geographic location. In statistical analysis a common tool using categorical responses is the questionnaire, where respondents are asked

opinions about a subject. If we give the categorical response a numerical score, a questionnaire can be easily analysed. A pie chart is a common visual representation of categorical data. The pie chart is a circle where portions of the “pie” are usually named categories and a percentage of the complete data. The whole circle is 100% of the data. A vertical histogram can also be used to illustrate categorical data where the  $x$  scale has a name, or label, and the  $y$ -axis is the amount or proportion of the data within that label. The vertical histogram can also be shown as a parallel or side-by-side histogram where now each label contains data say for two or more periods. In this way a comparison of changes can be made within named categories. The vertical histogram can be shown as a horizontal bar chart where it is now the  $y$ -axis that has the name, or label, and the  $x$ -axis the amount or proportion of data within that label. Similarly the horizontal bar chart can be shown as a parallel bar chart where now each label contains data say for two or more periods. Whether to use a vertical histogram or a horizontal bar chart is really a matter of personal preference. A visual tool often used in auditing or quality control is the Pareto diagram. This is a combination of vertical bars showing the frequency of occurrence of data according to given categories and a line graph indicating the accumulation of the data to 100%. When data falls into several categories the information can be represented in a cross-classification or contingency table. This table indicates the amount of data within defined categories. The cross-classification table can be converted into a stacked histogram according to the desired categories, which is a useful graphical presentation of the various classifications. Finally, this chapter mentions pictograms, which are pictorial representations of information. These are often used in newspapers and magazines to represent situations but they are difficult to rigorously analyse and can lead to misrepresentation of information. No further discussion of pictograms is given in this textbook.

## EXERCISE PROBLEMS

### 1. Buyout – Part I

#### Situation

Carrefour, France, is considering purchasing the total 50 retail stores belonging to Hardway, a grocery chain in the Greater London area of the United Kingdom. The profits from these 50 stores, for one particular month, in £'000s, are as follows.

8.1	11.8	8.7	10.6	9.5
9.3	11.5	10.7	11.6	7.8
10.5	7.6	10.1	8.9	8.6
11.1	10.2	11.1	9.9	9.8
11.6	15.1	12.5	6.5	7.5
10.3	12.9	9.2	10.7	12.8
12.5	9.3	10.4	12.7	10.5
10.3	11.1	9.6	9.7	14.5
13.7	6.7	11.5	8.4	10.3
13.7	11.2	7.3	5.3	12.5

#### Required

1. Illustrate this information as a closed-ended absolute frequency histogram using class ranges of £1,000 and logical minimum and maximum values for the data rounded to the nearest thousand pounds.
2. Convert the absolute frequency histogram developed in Question 1 into a relative frequency histogram.
3. Convert the relative frequency histogram developed in Question 2 into a relative frequency polygon.
4. Develop a stem-and-leaf display for the data using the thousands for the stem and the hundreds for the leaf. Compare this to the absolute frequency histogram.
5. Illustrate this data as a greater than and a less than ogive using both absolute and relative frequency values.
6. After examining the data presented in the figure from Question No. 1, Carrefour management decides that it will purchase only those stores showing profits greater than £12,500. On this basis, determine from the appropriate ogive how many of the Hardway stores Carrefour would purchase?

### 2. Closure

#### Situation

A major United States consulting company has 60 offices worldwide. The following are the revenues, in million dollars, for each of the offices for the last fiscal year. The average

annual operating cost per office for these, including salaries and all operating expenses, is \$36 million.

49.258	43.660	32.190	39.220	35.150	29.532
34.410	54.257	39.590	42.920	33.658	37.125
38.850	28.120	60.120	37.258	31.820	25.324
41.070	37.740	41.070	54.653	36.260	29.584
42.920	59.250	46.250	24.050	27.750	62.543
38.110	47.730	34.040	39.590	69.352	58.965
46.250	34.410	42.653	46.990	38.850	46.235
38.110	41.070	35.520	35.890	53.650	59.210
50.690	24.790	42.550	31.080	42.365	20.210
50.690	41.440	27.010	20.030	46.250	33.564

As a result of intense competition from other consulting firms and declining markets, management is considering closing those offices whose annual revenues are less than the average operating cost.

### Required

In order to present the data to management, so they can understand the impact of their proposed decision, develop the following information.

1. Present the revenue data as a closed-end absolute frequency distribution using logical lower and upper limits rounded to the nearest multiple \$10 million and a class limit range of \$5 million.
2. What is the average margin per office for the consulting firm before any closure?
3. Present on the appropriate frequency distribution (ogive), the number of offices having less than certain revenues. To construct the distribution use the following criterion:
  - Minimum on the revenue distribution is rounded to the closet multiple of \$10 million.
  - Use a range of \$5 million.
  - Maximum on the revenue distribution is rounded to the closest multiple of \$10 million.
4. From the distribution you have developed in Question 3, how many offices have revenues lower than \$36 million and thus risk being closed?
5. If management makes the decision to close that number of offices determined in Question 3 above, estimate the new average margin per office.

### 3. Swimming pool

#### Situation

A local community has a heated swimming pool, which is open to the public each year from May 17 until September 13. The community is considering building a restaurant facility in the swimming pool area but before a final decision is made, it wants to have assurance that the receipts from the attendance at the swimming pool will help finance the construction and operation of the restaurant. In order to give some justification to its decision the community noted the attendance each day for one particular year and this information is given below.

869	755	729	926	821	709	1,088	785	830	709
678	1,019	825	843	940	826	750	835	956	743
835	630	791	795	903	790	931	869	878	808
845	692	830	794	993	847	901	837	755	810
791	609	878	778	761	763	726	745	874	728
870	798	507	763	764	779	678	690	1,004	792
848	823	769	773	919	682	672	829	915	883
699	650	780	743	861	610	582	748	744	680
930	776	871	759	580	669	716	980	724	880
669	712	732	968	620	852	749	860	811	748
822	651	539	658	796	825	685	707	895	806
609	952	565	869	560	751	790	907	621	619

#### Required

1. Develop an absolute value closed-limit frequency distribution table using a data range of 50 attendances and, to the nearest hundred, a logical lower and upper limit for the data. Convert this data into an absolute value histogram.
2. Convert the absolute frequency histogram into a relative frequency histogram.
3. Plot the relative frequency distribution histogram as a polygon. What are your observations about this polygon?
4. Convert the relative frequency distribution into a greater than and less than ogive and plot these two line graphs on the same axis.
5. What is the proportion of the attendance at the swimming pool that is  $>750$  and  $\leq 800$  people?
6. Develop a stem-and-leaf display for the data using the hundreds for the stem and the tens for the leaves.
7. The community leaders came up with the following three alternatives regarding providing the capital investment for the restaurant. Respond to these using the ogive data.
  - (a) If the probability of more than 900 people coming to the swimming pool was at least 10% or the probability of less than 600 people coming to the swimming



pool was not less than 10%. Under these criteria would the community fund the restaurant? Quantify your answer both in terms of the 10% limits and the attendance values.

- (b) If the probability of more than 900 people coming to the swimming pool was at least 10% **and** the probability of less than 600 people coming to the swimming pool was not less than 10%. Under these criteria would the community fund the restaurant? Quantify your answer both in terms of the 10% limits and the attendance values.
- (c) If the probability of between 600 and 900 people coming to the swimming pool was at least 80%. Quantify your answer.

#### 4. Rhine river

##### Situation

On a certain lock gate on the Rhine river there is a toll charge for all boats over 20 m in length. The charge is €15.00/m for every metre above the minimum value of 20 m. In a certain period the following were the lengths of boats passing through the lock gate.

22.00	21.80	32.00	26.70	23.80	27.90	32.50	21.80	32.00	26.70	23.80
31.00	22.00	32.00	25.00	20.33	25.10	25.70	22.00	32.00	25.00	20.33
23.00	20.20	17.00	28.00	19.33	18.00	24.50	20.20	17.00	28.00	19.33
24.50	25.70	29.80	23.00	30.67	17.20	37.50	25.70	29.80	23.00	30.67
19.00	18.70	18.25	26.50	32.00	16.50	36.50	18.70	18.33	26.50	32.00

##### Required

1. Show this information in a stem-and-leaf display.
2. Draw the ogives for this data using a logical maximum and minimum value for the limits to the nearest even number of metres.
3. From the appropriate ogive approximately what proportion of the boats will not have to pay any toll fee?
4. Approximately what proportion of the time will the canal authorities be collecting at least €105 from boats passing through the canal?

#### 5. Purchasing expenditures

##### Situation

The complete daily purchasing expenditures for a large resort hotel for the last 200 days in Euros are given in the table below. The purchases include all food, beverages, and non-food items for the five restaurants in the complex. It also includes energy, water for the

three swimming pools, laundry, which is a purchased service, gasoline for the courtesy vehicles, gardening and landscaping services.

63,680	307,024	188,973	242,746	217,724	194,157	230,211	192,285	235,015	195,577
197,613	332,923	173,876	219,573	113,864	295,731	175,622	297,536	205,173	224,937
195,651	165,355	217,076	86,157	293,373	151,135	187,173	110,336	188,977	332,212
161,275	288,466	99,886	274,856	167,175	102,382	273,411	159,262	298,256	161,075
153,862	116,240	187,173	147,564	248,146	228,577	185,377	210,573	81,340	237,524
132,476	291,411	238,840	217,177	122,211	157,775	106,155	187,124	224,276	303,466
172,613	94,957	206,973	112,676	262,773	179,377	137,860	204,462	144,826	194,157
197,741	183,409	144,283	141,476	156,213	175,612	246,571	161,741	173,187	295,173
150,651	136,609	177,766	241,124	134,811	68,141	163,240	115,540	194,157	223,124
190,777	168,898	106,155	185,375	185,377	260,973	182,696	182,336	187,124	128,860
106,787	218,626	147,956	108,230	155,875	165,215	102,415	203,137	97,430	274,777
179,998	141,412	198,880	156,523	179,075	238,624	242,977	137,860	244,256	213,577
163,076	282,568	157,849	212,211	154,138	188,276	139,777	190,777	141,221	269,212
124,157	90,230	191,876	114,476	222,415	86,211	180,531	108,230	254,336	152,276
180,533	139,496	140,141	242,802	142,978	181,186	171,880	221,324	201,415	233,215
128,624	159,833	198,466	130,676	253,076	225,880	125,251	161,372	127,076	168,977
203,377	223,011	118,525	231,651	120,415	148,426	241,171	177,226	275,936	157,077
130,162	146,621	224,741	182,677	132,424	249,651	134,249	246,524	208,615	257,373
215,377	173,866	119,876	146,682	251,251	148,421	270,536	192,346	124,101	220,777
126,880	170,257	154,755	249,475	175,496	259,173	166,480	263,320	152,266	125,773

### Required

1. Develop an absolute frequency histogram for this data using the maximum value, rounded up to the nearest €10,000, to give the upper limit of the data, and the minimum value, rounded down to the nearest €10,000, to give the lower limit. Use an interval or class width of €20,000. This histogram will be a closed-limit absolute frequency distribution.
2. From the absolute frequency information develop a relative frequency distribution of sales.
3. What is the percentage of purchasing expenditures in the range €180,000 to €200,000?
4. Develop an absolute frequency polygon of the data. This is a line graph connecting the midpoints of each class in the dataset. What is the quantity of data in the highest frequency?
5. Develop an absolute frequency “more than” and “less than” ogive from the dataset.
6. Develop a relative frequency “more than” and “less than” ogive from the dataset.
7. From these ogives, what is an estimate of the percentage of purchasing expenditures less than €250,000?
8. From these ogives, 70% of the purchasing expenditures are greater than what amount?

## 6. Exchange rates

### Situation

The table on next page gives the exchange rates in currency units per \$US for two periods in 2004 and 2005.<sup>2</sup>

	16 November 2005	16 November 2004
Australia	1.37	1.28
Britain	0.58	0.54
Canada	1.19	1.19
Denmark	6.39	5.71
Euro area	0.86	0.77
Japan	119.00	104.00
Sweden	8.25	6.89
Switzerland	1.33	1.17

### Required

1. Construct a parallel bar chart for this data. (Note in order to obtain a graph which is more equitable, divide the data for Japan by 100 and those for Denmark and Sweden by 10.)
2. What are your conclusions from this bar chart?

## 7. European sales

### Situation

The table below gives the monthly profits in Euros for restaurants of a certain chain in Europe.

Country	Profits (€)
Denmark	985,789
England	1,274,659
Germany	225,481
Ireland	136,598
Netherlands	325,697
Norway	123,657
Poland	429,857
Portugal	256,987
Czech Republic	102,654
Spain	995,796

<sup>2</sup>Economic and financial indicators, *The Economist*, 19 November 2005, p. 101.

### Required

1. Develop a pie chart for this information.
2. Develop a histogram for this information in terms of absolute profits and percentage profits.
3. Develop a bar chart for this information in terms of absolute profits and percentage profits.
4. What are the three best performing countries and what is their total contribution to the total profits given?
5. Which are the three countries that have the lowest contribution to profits and what is their total contribution?

## 8. Nuclear power

### Situation

The table below gives the nuclear reactors in use or in construction according to country.<sup>3</sup>

Country	No. of nuclear reactors	Region
Argentina	3	South America
Armenia	1	Eastern Europe
Belgium	7	Western Europe
Brazil	2	South America
Britain	27	Western Europe
Bulgaria	4	Eastern Europe
Canada	16	North America
China	11	Far East
Czech Republic	6	Eastern Europe
Finland	4	Western Europe
France	59	Western Europe
Germany	18	Western Europe
Hungary	4	Eastern Europe
India	22	ME and South Asia
Iran	2	ME and South Asia
Japan	56	Far East
Lithuania	2	Eastern Europe
Mexico	2	North America
Netherlands	1	Western Europe
North Korea	1	Far East
Pakistan	2	ME and South Asia
Romania	2	Eastern Europe
Russia	33	Eastern Europe
Slovakia	8	Eastern Europe
Slovenia	1	Eastern Europe
South Africa	2	Africa

(Continued)

<sup>3</sup> *International Herald Tribune*, 18 October 2004.

Country	No. of nuclear reactors	Region
South Korea	20	Far East
Spain	9	Western Europe
Sweden	11	Western Europe
Switzerland	5	Western Europe
Ukraine	17	Eastern Europe
United States	104	North America

ME: Middle East.

### Required

1. Develop a bar chart for this information by country sorted by the number of reactors.
2. Develop a pie chart for this information according to the region.
3. Develop a pie chart for this information according to country for the region that has the highest proportion of nuclear reactors.
4. Which three countries have the highest number of nuclear reactors?
5. Which region has the highest proportion of nuclear reactors and dominated by which country?

## 9. Textbook sales

### Situation

The sales of an author's textbook in one particular year were according to the following table.

Country	Sales (units)	Country	Sales (units)
Australia	660	Mexico	10
Austria	4	Northern Ireland	69
Belgium	61	Netherlands	43
Botswana	3	New Zealand	28
Canada	147	Nigeria	3
China	5	Norway	78
Denmark	189	Pakistan	10
Egypt	10	Poland	4
Eire	25	Romania	3
England	1,632	South Africa	62
Finland	11	South Korea	1
France	523	Saudi Arabia	1
Germany	28	Scotland	10
Greece	5	Serbia	1
Hong Kong	2	Singapore	362
India	17	Slovenia	4
Iran	17	Spain	16
Israel	4	Sri Lanka	2
Italy	26	Sweden	162

Country	Sales (units)	Country	Sales (units)
Japan	21	Switzerland	59
Jordan	3	Taiwan	938
Latvia	1	Thailand	2
Lebanon	123	UAE	2
Lithuania	1	Wales	135
Luxemburg	69	Zimbabwe	3
Malaysia	2		

### Required

1. Develop a histogram for this data by country and by units sold, sorting the data from the country in which the units sold were the highest to the lowest. What is your criticism of this visual presentation?
2. Develop a pie chart for book sales by continent. Which continent has the highest percentage of sales? Which continent has the lowest book sales?
3. Develop a histogram for absolute book sales by continent from the highest to the lowest.
4. Develop a pie chart for book sales by countries in the European Union. Which country has the highest book sales as a proportion of total in Europe? Which country has the lowest sales?
5. Develop a histogram for absolute book sales by countries in the European Union from the highest to the lowest.
6. What are your comments about this data?

## 10. Textile wages

### Situation

The table below gives the wage rates by country, converted to \$US, for persons working in textile manufacturing. The wage rate includes all the mandatory charges which have to be paid by the employer for the employees benefit. This includes social charges, medical benefits, vacation, and the like.<sup>4</sup>

Country	Wage rate (\$US/hour)
Bulgaria	1.14
China (mainland)	0.49
Egypt	0.88
France	19.82
Italy	18.63
Slovakia	3.27
Turkey	3.05
United States	15.78

<sup>4</sup> *Wall street Journal Europe*, 27 September 2005, p. 1.

### Required

1. Develop a bar chart for this information. Show the information sorted.
2. Determine the wage rate of a country relative to the wage rate in China.
3. Plot on a combined histogram and line graph the sorted wage rate of the country as a histogram and a line graph for the data that you have calculated in Question 2.
4. What are your conclusions from this data that you have presented?

## 11. Immigration to Britain

### Situation

Nearly a year and a half after the expansion of the European Union, hundreds of East Europeans have moved to Britain to work. Poles, Lithuanians, Latvians and others are arriving at an average rate of 16,000 a month, as a result of Britain's decision to allow unlimited access to the citizens of the eight East Europeans that joined the European Union in 2004. The immigrants work as bus drivers, farmhands, dentists, waitresses, builders, and sales persons. The following table gives the statistics for those new arrivals from Eastern Europe since May 2004.<sup>5</sup>

Nationality of applicant	Registered to work
Czech Republic	14,610
Estonia	3,480
Hungary	6,900
Latvia	16,625
Lithuania	33,755
Poland	131,290
Slovakia	24,470
Slovenia	250
Age range of applicant	Percentage in range
18–24	42.0
25–34	40.0
35–44	11.0
45–54	6.0
55–64	1.0
Employment sector of applicant	No. applied to work (May 2004–June 2005)
Administration, business, and management	62,000
Agriculture	30,400
Construction	9,000
Entertainment and leisure	4,000

<sup>5</sup>Fuller, T., Europe's great migration: Britain absorbing influx from the East, *International Herald Tribune*, 21 October 2005, pp. 1, 4.

Employment sector of applicant	No. applied to work (May 2004–June 2005)
Food processing	11,000
Health care	10,000
Hospitality and catering	53,200
Manufacturing	19,000
Retail	9,500
Transport	7,500
Others	9,500

### Required

1. Develop a bar chart of the nationality of the immigrant and the number who have registered to work.
2. Transpose the information from Question 1 into a pie chart.
3. Develop a pie chart for the age range of the applicant and the percentage in this range.
4. Develop a bar chart for the employment sector of the immigrant and those registered for employment in this sector.
5. What are your conclusions from the charts that you have developed?

## 12. Pill popping

### Situation

The table below gives the number of pills taken per 1,000 people in certain selected countries.<sup>6</sup>

Country	Pills consumed per 1,000 people
Canada	66
France	78
Italy	40
Japan	40
Spain	64
United Kingdom	36
USA	53

### Required

1. Develop a bar chart for the data in the given alphabetical order.
2. Develop a pie chart for the data and show on this the country and the percentage of pill consumption based on the information provided.

<sup>6</sup> *Wall Street Journal Europe*, 25 February 2004.



3. Which country consumes the highest percentage of pills and what is this percentage amount to the nearest whole number?
4. How would you describe the consumption of pills in France compared to that in the United Kingdom?

### 13. Electoral College

#### Situation

In the United States for the presidential elections, people vote for a president in their state of residency. Each state has a certain number of electoral college votes according to the population of the state and it is the tally of these electoral college votes which determines who will be the next United States president. The following gives the electoral college votes for each of the 50 states of the United States plus the District of Columbia.<sup>7</sup> Also included is how the state voted in the 2004 United States presidential elections.<sup>8</sup>

State	Electoral college votes	Voted to
Alabama	9	Bush
Alaska	3	Bush
Arizona	10	Bush
Arkansas	6	Bush
California	55	Kerry
Colorado	9	Bush
Connecticut	7	Kerry
Delaware	3	Kerry
District of Columbia	3	Kerry
Florida	27	Bush
Georgia	15	Bush
Hawaii	4	Kerry
Idaho	4	Bush
Illinois	21	Kerry
Indiana	11	Bush
Iowa	7	Bush
Kansas	6	Bush
Kentucky	8	Bush
Louisiana	9	Bush
Maine	4	Kerry
Maryland	10	Kerry
Massachusetts	12	Kerry
Michigan	17	Kerry
Minnesota	10	Kerry
Mississippi	6	Bush

<sup>7</sup> *Wall Street Journal Europe*, 2 November 2004, p. A12.

<sup>8</sup> *The Economist*, 6 November 2004, p. 23.

State	Electoral college votes	Voted to
Missouri	11	Bush
Montana	3	Bush
Nebraska	5	Bush
Nevada	5	Bush
New Hampshire	4	Kerry
New Jersey	15	Kerry
New Mexico	5	Bush
New York	31	Kerry
North Carolina	15	Bush
North Dakota	3	Bush
Ohio	20	Bush
Oklahoma	7	Bush
Oregon	7	Kerry
Pennsylvania	21	Kerry
Rhode Island	4	Kerry
South Carolina	8	Bush
South Dakota	3	Bush
Tennessee	11	Bush
Texas	34	Bush
Utah	5	Bush
Vermont	3	Kerry
Virginia	13	Bush
Washington	11	Kerry
West Virginia	5	Bush
Wisconsin	10	Kerry
Wyoming	3	Bush

### Required

1. Develop a pie chart of the percentage of electoral college votes for each state.
2. Develop a histogram of the percentage of electoral college votes for each state.
3. How were the electoral college votes divided between Bush and Kerry? Show this on a pie chart.
4. Which state has the highest percentage of electoral votes and what is the percentage of the total electoral college votes?
5. What is the percentage of states including the District of Columbia that voted for Kerry?

## 14. Chemical delivery

### Situation

A chemical company is concerned about the quality of its chemical products that are delivered in drums to its clients. Over a 6-month period it used a student intern to

measure quantitatively the number of problems that occurred in the delivery process. The following table gives the recorded information over the 6-month period. The column “reason” in the table is considered exhaustive.

Reason	No. of occurrences in 6-months
Delay – bad weather	70
Documentation wrong	100
Drums damaged	150
Drums incorrectly sealed	3
Drums rusted	22
Incorrect labelling	7
Orders wrong	11
Pallets poorly stacked	50
Schedule change	35
Temperature too low	18

### Required

1. Construct a Pareto curve for this information.
2. What is the problem that happens most often and what is the percentage of occurrence? This is the problem area that you would probably tackle first.
3. Which are the four problem areas that constitute almost 80% of the quality problems in delivery?

## 15. Fruit distribution

### Situation

A fruit wholesaler was receiving complaints from retail outlets on the quality of fresh fruit that was delivered. In order to monitor the situation the wholesaler employed a student to rigorously take note of the problem areas and to record the number of times these problems occurred over a 3-month period. The following table gives the recorded information over the 3-month period. The column “reasons” in the table is considered exhaustive.

Reason	No. of occurrences in 3 months
Bacteria on some fruit	9
Boxes badly loaded	62
Boxes damaged	17
Client documentation incorrect	23
Fruit not clean	25
Fruit squashed	74
Fruit too ripe	14

Labelling wrong	11
Orders not conforming	6
Route directions poor	30

---

### Required

1. Construct a Pareto curve for this information.
2. What is the problem that happens most often and what is the percentage of occurrence? Is this the problem area that you would tackle first?
3. What are the problem areas that cumulatively constitute about 80% of the quality problems in delivery of the fresh fruit?

## 16. Case: Soccer

### Situation

When an exhausted Chris Powell trudged off the Millennium stadium pitch on the afternoon of 30 May 2005, he could not have been forgiven for feeling pleased with himself. Not only had he helped West Ham claw their way back into the Premiership league for the 2005–2006 season, but the left back had featured in 42 league cup and play off matches since reluctantly leaving Charlton Athletic the previous September. It had been a good season since opposition right-wingers had been vanquished and Powell and Mathew Etherington had formed a formidable left-sided partnership. If you did not know better, you might have suspected the engaging 35-year old was a decade younger.<sup>9</sup>

For many people in England, and in fact, for most of Europe, football or soccer is their passion. Every Saturday many people, the young and the not-so-young, faithfully go and see their home team play. Football in England is a huge business. According to the accountants, Deloitte and Touche, the 20 clubs that make up the Barclay's Bank sponsored English Premiership league, the most watched and profitable league in Europe, had total revenues of almost £2 billion (\$3.6 billion) in the 2003–2004 season. The best players command salaries of £100,000 a week excluding endorsements.<sup>10</sup> In addition, at the end of the season, the clubs themselves are awarded prize money depending on their position on the league tables at the end of the year. These prize amounts are indicated in Table 1 for the 2004–2005 season. The game results are given in Table 2 and the final league results in Table 3 and from these you can determine the amount that was awarded to each club.<sup>11</sup>

<sup>9</sup> Aizlewood, J., Powell back at happy valley, *The Sunday Times*, 28 August 2005, p. 11.

<sup>10</sup> Theobald, T. and Cooper, C., *Business and the Beautiful Game*, Kogan Page, *International Herald Tribune*, 1–2 October 2005, p. 19 (Book review on soccer).

<sup>11</sup> *News of the World Football Annual 2005–2006*, Invincible Press, an imprint of Harper Collins, 2005.

Table 1

Position	Prize money (£)	Position	Prize money (£)
1	9,500,000	11	4,750,000
2	9,020,000	12	4,270,000
3	8,550,000	13	3,800,000
4	8,070,000	14	3,320,000
5	7,600,000	15	2,850,000
6	7,120,000	16	2,370,000
7	6,650,000	17	1,900,000
8	6,170,000	18	1,420,000
9	5,700,000	19	950,000
10	5,220,000	20	475,000

### Required

These three tables give a lot of information on the premier leaguer football results for the 2004–2005 season. How could you put this in a visual form to present the information to a broad audience?

Table 2

Club	Games played	Home					Away				
		Win	Draw	Lost	For	Away	Win	Draw	Lost	For	Away
Arsenal	38	13	5	1	54	19	12	3	4	33	17
Aston Villa	38	8	6	5	26	17	5	5	10	19	35
Birmingham City	38	8	6	5	24	15	4	6	10	16	31
Blackburn Rovers	38	5	8	6	24	22	3	7	8	11	21
Bolton Wanderers	38	9	5	5	25	18	7	5	7	24	26
Charlton Athletic	38	8	4	7	29	29	4	6	9	13	29
Chelsea	38	14	5	0	35	6	15	3	1	37	9
Crystal Palace	38	6	5	8	21	19	1	7	11	20	43
Everton	38	12	2	5	24	15	6	5	8	21	31
Fulham	38	8	4	7	29	26	3	4	11	23	34
Liverpool	38	12	4	3	31	15	5	3	11	21	26
Manchester City	38	8	6	5	24	14	5	7	7	23	25
Manchester United	38	12	6	1	31	12	10	5	4	27	14
Middlesbrough	38	9	6	4	29	19	5	7	7	24	27

Club	Games played	Home					Away				
		Win	Draw	Lost	For	Away	Win	Draw	Lost	For	Away
Newcastle United	38	7	7	5	25	25	4	7	9	22	32
Norwich City	38	7	5	7	29	32	0	7	12	13	45
Portsmouth	38	8	4	7	30	26	4	5	12	13	33
Southampton	38	5	9	5	30	30	1	5	13	15	36
Tottenham	38	9	5	5	36	22	5	5	9	11	19
WBA	38	5	8	6	17	24	2	8	10	19	37

Table 3

	Arsenal	Aston Villa	Birmingham City	Blackburn Rovers	Bolton Wanderers	Charlton Athletic	Chelsea	Crystal Palace	Everton
Arsenal	- - -	3 - 1	3 - 0	3 - 0	2 - 2	3 - 0	2 - 2	5 - 1	7 - 0
Aston Villa	1 - 3	- - -	1 - 2	1 - 0	1 - 1	4 - 1	0 - 0	1 - 1	1 - 3
Birmingham City	2 - 1	2 - 0	- - -	2 - 1	1 - 2	5 - 2	0 - 1	0 - 1	0 - 1
Blackburn Rovers	0 - 1	2 - 2	3 - 3	- - -	0 - 1	6 - 3	0 - 1	1 - 0	0 - 0
Bolton Wanderers	1 - 0	1 - 2	1 - 1	0 - 1	- - -	0 - 1	0 - 2	1 - 0	3 - 2
Charlton Athletic	1 - 3	3 - 0	3 - 1	1 - 0	1 - 2	- - -	0 - 4	2 - 2	2 - 0
Chelsea	0 - 0	1 - 0	1 - 1	4 - 0	2 - 2	4 - 0	- - -	4 - 1	1 - 0
Crystal Palace	1 - 1	2 - 0	2 - 0	0 - 0	0 - 1	0 - 0	0 - 2	- - -	1 - 3
Everton	1 - 4	1 - 1	1 - 1	0 - 1	3 - 2	0 - 1	0 - 1	4 - 0	- - -
Fulham	0 - 3	1 - 1	2 - 3	0 - 2	2 - 0	0 - 2	1 - 4	3 - 1	2 - 0
Liverpool	2 - 1	2 - 1	0 - 1	0 - 0	1 - 0	0 - 0	0 - 1	3 - 2	2 - 1
Manchester City	0 - 1	2 - 0	3 - 0	1 - 1	0 - 1	1 - 1	1 - 0	3 - 1	0 - 1
Manchester United	2 - 0	3 - 1	2 - 0	0 - 0	2 - 0	0 - 0	1 - 3	5 - 2	0 - 0
Middlesbrough	0 - 1	3 - 0	2 - 1	1 - 0	1 - 1	1 - 0	0 - 1	2 - 1	1 - 1
Newcastle United	0 - 1	0 - 3	2 - 1	3 - 0	2 - 1	3 - 0	1 - 1	0 - 0	1 - 1
Norwich City	1 - 4	0 - 0	1 - 0	1 - 1	3 - 2	1 - 1	1 - 3	1 - 1	2 - 3
Portsmouth	0 - 1	1 - 2	1 - 1	0 - 1	1 - 1	0 - 1	0 - 2	3 - 1	0 - 1
Southampton	1 - 1	2 - 3	0 - 0	3 - 2	1 - 2	3 - 2	1 - 3	2 - 2	2 - 2
Tottenham	4 - 5	5 - 1	1 - 0	0 - 0	1 - 2	0 - 0	0 - 2	1 - 1	5 - 2
WBA	0 - 2	1 - 1	2 - 0	1 - 1	2 - 1	1 - 1	1 - 4	2 - 2	1 - 0

Fulham	Liverpool	Manchester City	Manchester United	Middlesbrough	Newcastle United	Norwich City	Portsmouth	Southampton	Tottenham	WBA
2 - 0	3 - 1	1 - 1	2 - 4	5 - 3	1 - 0	4 - 1	3 - 0	2 - 2	1 - 0	1 - 1
2 - 0	1 - 1	1 - 2	0 - 1	2 - 0	4 - 2	3 - 0	3 - 0	2 - 0	1 - 0	1 - 1
1 - 2	2 - 0	1 - 0	0 - 0	2 - 0	2 - 2	1 - 1	0 - 0	2 - 1	1 - 1	4 - 0
1 - 3	2 - 2	0 - 0	1 - 1	0 - 4	2 - 2	3 - 0	1 - 0	3 - 0	0 - 1	1 - 1
3 - 1	1 - 0	0 - 1	2 - 2	0 - 0	2 - 1	1 - 0	0 - 1	1 - 1	3 - 1	1 - 1
2 - 1	1 - 2	2 - 2	0 - 4	1 - 2	1 - 1	4 - 0	2 - 1	0 - 0	2 - 0	1 - 4
2 - 1	1 - 0	0 - 0	1 - 0	2 - 0	4 - 0	4 - 0	3 - 0	2 - 1	0 - 0	1 - 0
2 - 0	1 - 0	1 - 2	0 - 0	0 - 1	0 - 2	3 - 3	0 - 1	2 - 2	3 - 0	3 - 0
1 - 0	1 - 0	2 - 1	1 - 0	1 - 0	2 - 0	1 - 0	2 - 1	1 - 0	0 - 1	2 - 1
- - -	2 - 4	1 - 1	1 - 1	0 - 2	1 - 3	6 - 0	3 - 1	1 - 0	2 - 0	1 - 0
3 - 1	- - -	2 - 1	0 - 1	1 - 1	3 - 1	3 - 0	1 - 1	1 - 0	2 - 2	3 - 0
1 - 1	1 - 0	- - -	0 - 2	1 - 1	1 - 1	1 - 1	2 - 0	2 - 1	0 - 1	1 - 1
1 - 0	2 - 1	0 - 0	- - -	1 - 1	2 - 1	2 - 1	2 - 1	3 - 0	0 - 0	1 - 1
1 - 1	2 - 0	3 - 2	0 - 2	- - -	2 - 2	2 - 0	1 - 1	1 - 3	1 - 0	4 - 0
1 - 4	1 - 0	4 - 3	1 - 3	0 - 0	- - -	2 - 2	1 - 1	2 - 1	0 - 1	3 - 1
0 - 1	1 - 2	2 - 3	2 - 0	4 - 4	2 - 1	- - -	2 - 2	2 - 1	0 - 2	3 - 2
4 - 3	1 - 2	1 - 3	2 - 0	2 - 1	1 - 1	1 - 1	- - -	4 - 1	1 - 0	3 - 2
3 - 3	2 - 0	0 - 0	1 - 2	2 - 2	1 - 2	4 - 3	2 - 1	- - -	1 - 0	2 - 2
2 - 0	1 - 1	2 - 1	0 - 1	2 - 0	1 - 0	0 - 0	3 - 1	5 - 1	- - -	1 - 1
1 - 1	0 - 5	2 - 0	0 - 3	1 - 2	0 - 0	0 - 0	2 - 0	0 - 0	1 - 1	- - -



*This page intentionally left blank*

# Characterizing and defining data

## Fast food and currencies

*How do you compare the cost of living worldwide? An innovative way is to look at the prices of a McDonald's Big Mac in various countries as The Economist has been doing since 1986. Their 2005 data is given in Table 2.1.<sup>1</sup> From this information you might conclude that the Euro is overvalued by 17% against the \$US; that the cost of living in Switzerland is the highest; and that it is cheaper to live in Malaysia.*

*Alternatively you would know that worldwide the average price of a Big Mac is \$2.51; that half of the Big Macs are less than \$2.40 and that half are more than \$2.40; and that the range of the prices of Big Macs is \$3.67. These are some of the characteristics of the prices of data for Big Macs. These are some of the properties of statistical data that are covered in this chapter.*

---

<sup>1</sup> "The Economist's Big Mac index: Fast food and strong currencies", *The Economist*, 11 June 2005.

Table 2.1 Price of the Big Mac worldwide.

Country	Price (\$US)	Country	Price (\$US)
Argentina	1.64	Mexico	2.58
Australia	2.50	New Zealand	3.17
Brazil	2.39	Peru	2.76
Britain	3.44	Philippines	1.47
Canada	2.63	Poland	1.96
Chile	2.53	Russia	1.48
China	2.27	Singapore	2.17
Czech Republic	2.30	South Africa	2.10
Denmark	4.58	South Korea	2.49
Egypt	1.55	Sweden	4.17
Euro zone	3.58	Switzerland	5.05
Hong Kong	1.54	Taiwan	2.41
Hungary	2.60	Thailand	1.48
Indonesia	1.53	Turkey	2.92
Japan	2.34	United States	3.06
Malaysia	1.38	Venezuela	2.13

## Learning objectives

After you have studied this chapter you will be able to **determine the properties** of statistical data, to **describe** clearly their meaning, to **compare** datasets, and to **apply** these properties in decision-making. Specifically, you will learn the following **characteristics**.

- ✓ **Central tendency of data** • Arithmetic mean • Weighted average • Median value • Mode • Midrange • Geometric mean
- ✓ **Dispersion of data** • Range • Variance and standard deviation • Expression for the variance • Expression for the standard deviation • Determining the variance and the standard deviation • Deviations about the mean • Coefficient of variation and the standard deviation
- ✓ **Quartiles** • Boundary limit of quartiles • Properties of quartiles • Box and whisker plot • Drawing the box and whisker plot with Excel
- ✓ **Percentiles** • Development of percentiles • Division of data

It is useful to characterize data as these characteristics or properties of data can be compared or benchmarked with other datasets. In this way decisions can be made about business situations and certain conclusions drawn. The two common general **data characteristics** are central tendency and dispersion.

### Central Tendency of Data

The clustering of data around a central or a middle value is referred to as the **central tendency**. The central tendency that we are most familiar with is average or mean value but there are others. They are all illustrated as follows.

#### Arithmetic mean

The **arithmetic mean** or most often known as the **mean** or **average value**, and written by  $\bar{x}$ , is the most common measure of central tendency. It is

determined by the sum of the all the values of the observations,  $x$ , divided by the number of elements in the observations,  $N$ . The equation is,

$$\bar{x} = \frac{\sum x}{N} \quad 2(i)$$

For example, assume the salaries in Euros of five people working in the same department are as in Table 2.2. The total of these five values is €172,000 and 172,000/5 gives a mean value of €34,400. (On a grander scale, Goldman Sachs, the world's leading investment bank reports that the average pay-package of its 24,000 staff in 2005 was \$520,000 and that included a lot of assistants and secretaries!<sup>2</sup>)

The arithmetic mean is easy to understand, and every dataset has a mean value. The mean value in a dataset can be determined by using **[function AVERAGE]** in Excel.

Table 2.2 Arithmetic mean.

Eric	Susan	John	Helen	Robert
40,000	50,000	35,000	20,000	27,000

<sup>2</sup> "On top of the world – In its taste for risk, the world's leading investment bank epitomises the modern financial system", *The Economist*, 29 April 2006, p. 9.

Table 2.3 Arithmetic mean not necessarily affected by the number values.

Eric	Susan	John	Helen	Robert	Brian	Delphine
40,000	50,000	35,000	20,000	27,000	34,000	34,800

Note that the arithmetic mean can be influenced by extreme values or **outliers**. In the above situation, John has an annual salary of €35,000 and his salary is currently above the average. Now, assume that Susan has her salary increased to €75,000 per year. In recalculating the mean the average salary of the five increases to €39,400 per year. Nothing has happened to John's situation but his salary is now below average. Is John now at a disadvantage? What is the reason that Susan received the increase? Thus, in using average values for analysis, you need to understand if it includes outliers and the circumstance for which the mean value is being used.

The number of values does not necessarily influence the arithmetic mean. In the above example, using the original data, suppose now that Brian and Delphine join the department at respective annual salaries of €34,000 and €34,800 as shown in Table 2.3. The average is still €34,400.

## Weighted average

The **weighted average** is a measure of central tendency and is a mean value that takes into account the importance, or weighting of each value in the overall total. For example in Chapter 1 we introduced a questionnaire as a method of evaluating customer satisfaction. Table 2.4 is the type of questionnaire used for evaluating customer satisfaction. Here the questionnaire has the responses of 15 students regarding satisfaction of

a course programme. (Students are the customers of the professors!) The X in each cell is the response of each student and the total responses for each category are in the last line.

The weighted average of the student response is given by,

Weighted average

$$= \sum \frac{\text{Number of responses} * \text{score}}{\text{Total responses}}$$

From the table we have,

Weighted average

$$= \frac{2 * 1 + 1 * 2 + 1 * 3 + 5 * 4 + 6 * 5}{15} = 3.80$$

Thus using the criterion of the weighted average the central tendency of the evaluation of the university programme is 3.80, which translates into saying the programme is between satisfactory and good and closer to being good. Note in Excel this calculation can be performed by using **[function SUMPRODUCT]**.

Another use of weighted averages is in product costing. Assume that a manufacturing organization uses three types of labour in the manufacture of Product 1 and Product 2 as shown in Table 2.5. In making the finished product the semi-finished components must pass through the activities of drilling, forming, and assembly before it is completed. Note that in these different activities the hourly wage rate is different. Thus to calculate the correct average cost of

Table 2.4 Weighted average.

Category	Very poor	Poor	Satisfactory	Good	Very good	Total responses
Score	1	2	3	4	5	
Student 1					X	1
Student 2				X		1
Student 3		X				1
Student 4				X		1
Student 5	X				1	1
Student 6			X			1
Student 7					X	1
Student 8				X		1
Student 9					X	1
Student 10	X					1
Student 11					X	1
Student 12					X	1
Student 13				X		1
Student 14					X	1
Student 15				X		1
Total responses	2	1	1	5	6	15

Table 2.5 Weighted average.

Labour operation	Hourly wage rate	Labour hours/unit Product A	Labour hours/unit Product B
Drilling	\$10.50	2.50	1.50
Forming	\$12.75	3.00	2.25
Assembly	\$14.25	1.75	2.00
Total		7.25	5.25

labour per finished unit, weighted averages are used as follows:

Product A, labour cost, \$/unit is

$$\begin{aligned} & \$10.50 * 2.50 + 12.75 * 3.00 + 14.25 * 1.75 \\ & = \$89.44 \end{aligned}$$

Product B, labour cost, \$/unit is

$$\begin{aligned} & \$10.50 * 1.50 + 12.75 * 2.25 + 14.25 * 2.00 \\ & = \$72.94 \end{aligned}$$

If simply the average hourly wage rate was used, the hourly labour cost would be:

$$\frac{10.50 + 12.75 + 14.25}{3} = \$12.50/\text{hour}$$

Then if we use this hourly labour cost to determine unit product cost we would have,

$$\text{Product A} \quad 12.50 * 7.25 = \$90.63/\text{unit}$$

$$\text{Product B} \quad \$12.50 * 5.75 = \$71.88/\text{unit}$$

This is an incorrect way to determine the unit cost since we must use the contribution of each activity to determine the correct amount.

## Median value

The **median** is another measure of central tendency that divides information or data into two equal parts. We come across the median when we talk about the median of a road. This is the white line that divides the road into two parts such that there is the same number of lanes on

Table 2.6 Median value – raw data.

9	13	12	7	6	11	12
---	----	----	---	---	----	----

Table 2.7 Median value – ordered data.

6	7	9	11	12	12	13
---	---	---	----	----	----	----

one side than on the other. When we have quantitative data it is the middle value of the **data array** or the ordered set of data. Consider the dataset in Table 2.6.

To determine the median value it must first be rearranged in ascending (or descending) order. In ascending order this is as in Table 2.7. Since there are seven pieces of data, the middle, or the median, value is the 4th number, which in this case is 11. The median value is of interest as indicates that half of the data lies above the median, and half below. For example, if the median, price of a house in a certain region is \$200,000 then this indicates that half of the number of houses is above \$200,000 and the other half is below.

When  $n$ , the number of values in a data array is odd, the median is given by,

$$\frac{n+1}{2} \quad 2(\text{ii})$$

Thus, if there are seven values in the dataset, then the median is  $(7+1)/2$  or the 4th value as in the above example.

When  $n$ , the number of values, is even, the median value is the average of the values determined from the following relationship:

$$\frac{n}{2} \quad \text{and} \quad \frac{(n+2)}{2} \quad 2(\text{iii})$$

When there are 6 values in a set of data, the median is the value of  $6/2$  and  $(6+2)/2$  or the linear average of the 3rd and 4th value.

Table 2.8 Median value – salaries.

Eric	Susan	John	Helen	Robert
40,000	50,000	35,000	20,000	27,000

Table 2.9 Median value – salaries ordered.

Helen	Robert	John	Eric	Susan
20,000	27,000	35,000	40,000	50,000

Table 2.10 Median value – salaries unaffected by extreme values.

Helen	Robert	John	Eric	Susan
20,000	27,000	35,000	40,000	75,000

The value of the median is unaffected by extreme values. Consider again the salary situation of the five people in John's department as in Table 2.8.

Ordering this data gives Table 2.9.

John's salary is at the median value. Again, if Susan's salary is increased to €75,000 then the revised information is as in Table 2.10.

John still has the median salary and so that on this basis, nothing has changed for John. However, when we used the average value as above, there was a change.

The number of values affects the median. Assume Stan joins the department in the example above at the same original salary as Susan. The salary values are thus as Table 2.11.

There is an even number of values in the dataset and now the median is  $(35,000 + 40,000)/2$  or €37,500. John's salary is now below the median. Again, nothing has happened

**Table 2.11** Median value – number of values affects the median.

Helen	Robert	John	Eric	Susan	Stan
20,000	27,000	35,000	40,000	50,000	50,000

to John's salary but on a comparative basis it appears that he is worse off!

The median value in any size dataset can be determined by using the [\[function MEDIAN\]](#) in Excel. We do not have to order the data or even to take into account whether there is an even or odd number of values as Excel automatically takes this into consideration. For example, if we determine the median value of the sales data given in Table 1.1, we call up [\[function MEDIAN\]](#) and enter the dataset. For this dataset the median value is 100,296.

## Mode

The [mode](#) is another measure of central tendency and is that value that occurs most frequently in a dataset. It is of interest because that value that occurs most frequently is probably a response that deserves further investigation. For example, Table 2.12 are the monthly sales in \$millions for the last year.

The mode is 12 since it occurs 3 times. Thus in forecasting future sales we might conclude that there is a higher probability that sales will be \$12 million in any given month. The mode is unaffected by extreme values. For example, if the sales in January were \$100 million instead of \$10 million, the mode is still 12. However, the number of values might affect the mode. For example, if we use the following sales data in Table 2.13 over the last 15 months, the modal value is now \$14 million since it occurs 4 times.

Unlike the mean and median, the mode can be used for qualitative as well as for quantitative

**Table 2.12** Mode – that value that occurs most frequently.

January	10
February	12
March	11
April	14
May	12
June	14
July	12
August	16
September	9
October	19
November	10
December	13

**Table 2.13** Mode – might be affected by the number of values.

January	10
February	12
March	11
April	14
May	12
June	14
July	12
August	16
September	9
October	19
November	10
December	13
January	14
February	10
March	14

data. For example, in a questionnaire, people were asked to give their favourite colour. The responses were according to Table 2.14.

The modal value is blue since this response occurred 3 times. This type of information is useful say in the textile business when a firm is planning the preparation of new fabric or the automobile industry when the company is planning to put



**Table 2.14** Mode can be determined for colours.

Yellow	Green	Purple	Blue
Red	Blue	Brown	Pink
Green	Violet	Rose	Blue

**Table 2.15** Bi-modal.

9	13	17	19	7	3	13	8	22	4	7	9
---	----	----	----	---	---	----	---	----	---	---	---

**Table 2.16** Midrange.

9	13	12	7	6	11	12
---	----	----	---	---	----	----

new cars on the market. The modal value in a dataset for quantitative data can be determined by using the **[function MODE]** in Excel.

A dataset might be multi-modal when there are data values that occur equally frequently. For example, **bi-modal** is when there are two values in a dataset that occur most frequently. The dataset in Table 2.15 is bi-modal as both the values 9 and 13 occur twice. When a dataset is bi-modal that indicates that there are two pieces of data that are of particular interest. Data can be **tri-modal**, **quad-modal**, etc. meaning that there are three, four, or more values that occur most frequently.

## Midrange

The **midrange** is also a measure of central tendency and is the average of the smallest and largest observation in a dataset. In Table 2.16, the midrange is,

$$\frac{13 + 6}{2} = \frac{19}{2} = 9.5$$

**Table 2.17** Midrange.

Helen	Robert	John	Eric	Susan
20,000	27,000	35,000	40,000	50,000

**Table 2.18** Midrange – affected by extreme values.

Helen	Robert	John	Eric	Susan
20,000	27,000	35,000	40,000	75,000

The midrange is of interest to know where data sits compared to the midrange. In the salary information of Table 2.17,

The midrange is  $(50,000 + 20,000)/2$  or 35,000 and so John's salary is exactly at the midrange. Again assume Susan's salary is increased to €75,000 to give the information in Table 2.18.

Then the midrange is  $(20,000 + 75,000)/2$  or €47,500 and John's salary is now below the midrange. Thus, the midrange can be distorted by extreme values.

## Geometric mean

The **geometric mean** is a measure of central tendency used when data is changing over time. Examples might be the growth of investments, the inflation rate, or the change of the gross national product. For example, consider the growth of an initial investment of \$1,000 in a savings account that is deposited for a period of 5 years. The interest rate, which is accumulated annually, is different for each year. Table 2.19 gives the interest and the growth of the investment.

The average growth rate, or geometric mean, is calculated by the relationship:

$${}^n\sqrt{(\text{product of growth rates})} \quad 2(\text{iv})$$

Table 2.19 Geometric mean.

Year	Interest rate (%)	Growth factor	Value year-end
1	6.0	1.060	\$1,060.00
2	7.5	1.075	\$1,139.50
3	8.2	1.082	\$1,232.94
4	7.9	1.079	\$1,330.34
5	5.1	1.051	\$1,398.19

In this case the geometric mean is,

$$\sqrt[5]{1.060 * 1.075 * 1.082 * 1.079 * 1.051} = 1.0693$$

This is an average growth rate of 6.93% per year ( $1.0693 - 1 = 0.0693$  or 6.93%). Thus, the value of the \$1,000 at the end of 5 years will be,

$$\$1,000 * 1.0693^5 = \$1,398.19$$

The same value as calculated in Table 2.19.

If the arithmetic average of the growth rates was used, the mean growth rate would be:

$$\frac{1.060 + 1.075 + 1.082 + 1.079 + 1.051}{5} = 1.0690$$

or a growth rate slightly less of 6.90% per year.

Using this mean interest rate, the value of the initial deposit at the end of 5 years would be,

$$\$1,000 \times 1.0690^5 = \$1,396.01$$

This is less than the amount calculated using the geometric mean. The difference here is small but in cases where interest rates are widely fluctuating, and deposit amounts are large, the difference can be significant. The geometrical mean can be determined by using [\[function GEOMEAN\]](#) in Excel applied to the growth rates.

Table 2.20 Range.

Eric	Susan	John	Helen	Robert
40,000	50,000	35,000	20,000	27,000

## Dispersion of Data

Dispersion is how much data is separated, spread out, or varies from other data values. It is important to know the amount of dispersion, variation, or spread, as data that is more dispersed or separated is less reliable for analytical purposes. Datasets can have different measures of dispersion or variation but may have the same measure of central tendency. In many situations, we may be more interested in the variation, than in the central value, since variation can be a measure of inconsistency. The following are the common measures of the [dispersion](#) of data.

### Range

The [range](#) is the difference between the maximum and the minimum value in a dataset. We have seen the use of the range in Chapter 1 in the development of frequency distributions. Another illustration is represented in Table 2.20 which is the salary data presented earlier in Table 2.8. Here the range is the difference of the salaries for Susan and Helen, or  $\text{€}50,000 - \text{€}20,000 = \text{€}30,000$ .

The range is affected by extreme values. For example, if we include in the dataset the salary of Francis, the manager of the department, who has a salary of  $\text{€}125,000$  we then have Table 2.21. Here the range is  $\text{€}125,000 - \text{€}20,000 = \text{€}105,000$ .

The number of values does not necessarily affect the range. For example let us say that we

**Table 2.21** Range is affected by extreme values.

Eric	Susan	John	Francis	Helen	Robert
40,000	50,000	35,000	125,000	20,000	27,000

**Table 2.22** Range is not necessarily affected by the number of values.

Eric	Susan	Julie	John	Francis	Helen	Robert
40,000	50,000	37,000	35,000	125,000	20,000	27,000

add the salary of Julie at €37,000 to the dataset in Table 2.21 to give the dataset in Table 2.22. Then the range is unchanged at €105,000.

The larger the range in a dataset, then the greater is the dispersion, and thus the uncertainty of the information for analytical purposes. Although we often talk about the range of data, the major drawback in using the range as a measure of dispersion is that it only considers two pieces of information in the dataset. In this case, any extreme, or outlying values, can distort the measure of dispersion as is illustrated by the information in Tables 2.21 and 2.22.

## Variance and standard deviation

The variance and the related measure, standard deviation, overcome the drawback of using the range as a measure of dispersion as in their calculation every value in the dataset is considered. Although both the variance and standard deviation are affected by extreme values, the impact is not as great as using the range since an aggregate of all the values in the dataset are considered. The variance and particularly the standard deviation are the most often used measures of dispersion in statistics. The variance is in

squared units and measures the dispersion of a dataset around the mean value. The standard deviation has the same units of the data under consideration and is the square root of the variation. We use the term “standard” in standard deviation as it represents the typical deviation for that particular dataset.

## Expression for the variance

There is a variance and standard deviation both for a population and a sample. The **population variance**, denoted by  $\sigma_x^2$ , is the sum of the squared difference between each observation,  $x$ , and the mean value,  $\mu$ , divided by the number of data observations,  $N$ , or as follows:

$$\sigma_x^2 = \frac{\sum (x - \mu_x)^2}{N} \quad 2(v)$$

- For each observation of  $x$ , the mean value  $\mu_x$  is subtracted. This indicates how far this observation is from the mean, or the range of this observation from the mean.
- By squaring each of the differences obtained, the negative signs are removed.
- By dividing by  $N$  gives an average value.

The expression for the **sample variance**,  $s^2$ , is analogous to the population variance and is,

$$s^2 = \frac{\sum (x - \bar{x})^2}{(n - 1)} \quad 2(vi)$$

In the sample variance,  $\bar{x}$ , or  $x$ -bar the average of the values of  $x$  replaces  $\mu_x$  of the population variance and  $(n - 1)$  replaces  $N$  the population size. One of the principal uses of statistics is to take a sample from the population and make estimates of the population parameters based only on the sample measurements. By convention when we use the symbol  $n$  it means we have taken a sample of size  $n$  from the population of size  $N$ . Using  $(n - 1)$  in the denominator reflects the fact that we have used  $\bar{x}$  in the formula and so we have lost one degree of freedom in our calculation. For example, consider you have a sum of \$1,000 to

distribute to your six co-workers based on certain criteria. To the first five you have the freedom to give any amount say \$200, \$150, \$75, \$210, \$260. To the sixth co-worker you have no degree of freedom of the amount to give which has to be the amount remaining from the original \$1,000, which in this case is \$105. When we are performing sampling experiments to estimate the population parameter with  $(n - 1)$  in the denominator of the sample variance formula we have an unbiased estimate of the true population variance. If the sample size,  $n$ , is large, then using  $n$  or  $(n - 1)$  will give results that are close.

## Expression for the standard deviation

The standard deviation is the square root of the variance and thus has the same units as the data used in the measurement. It is the most often used measure of dispersion in analytical work. The **population standard deviation**,  $\sigma_x$ , is given by,

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{\frac{\sum (x - \mu_x)^2}{N}} \quad 2(\text{vii})$$

The **sample standard deviation**,  $s$ , is as follows:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x - \bar{x})^2}{(n - 1)}} \quad 2(\text{viii})$$

For any dataset, the closer the value of the standard deviation is to zero, then the smaller is the dispersion which means that the data values are closer to the mean value of the dataset and that the data would be more reliable for subsequent analytical purposes.

Note that the expression  $\sigma$  is sometimes used to denote the population standard deviation rather than  $\sigma_x$ . Similarly  $\mu$  is used to denote the mean value rather than  $\mu_x$ . That is the subscript  $x$  is dropped, the logic being that it is understood that the values are calculated using the random variable  $x$  and so it is not necessary to show them with the mean and standard deviation symbols!

**Table 2.23** Variance and standard deviation.

9	13	12	7	6	11	12
---	----	----	---	---	----	----

## Determining the variance and the standard deviation

Let us consider the dataset given in Table 2.23. If we use equations 2(v) through to 2(viii) then we will obtain the population variance, the population standard deviation, the sample variance, and the sample standard deviation. These values and the calculation steps are shown in Table 2.24. However with Excel it is not necessary to go through these calculations as their values can be simply determined by using the following Excel functions:

- Population variance [**function** VARP]
- Population standard deviation [**function** STDEVP]
- Sample variance [**function** VAR]
- Sample standard deviation [**function** STDEV].

Note that for any given dataset when you calculate the population variance it is always smaller than the sample variance since the denominator,  $N$ , in the population variance is greater than the value  $N - 1$ . Similarly for the same dataset the population standard deviation is always less than the calculated sample standard deviation. Table 2.25, which is a summary of the final results of Table 2.24, illustrates this clearly.

## Deviation about the mean

The **deviation about the mean** of all observations,  $x$ , about the mean value,  $\bar{x}$ , is zero or mathematically,

$$\sum (x - \bar{x}) = 0 \quad 2(\text{ix})$$

Table 2.24 Variance and standard deviation.

	$x$	$(x - \mu)$	$(x - \mu)^2$
	2	-1	1
	13	3	9
	12	2	4
	7	-3	9
	6	-4	16
	11	1	1
	12	2	4
Number of values, $N$	7		
Total of values	70		44
Mean value, $\mu$	10		
Population variance, $\sigma^2$			6.2857
Population standard deviation, $\sigma$			2.5071
$N - 1$	6		
Sample variance, $s^2$			7.3333
Sample standard deviation, $s$			2.7080

Table 2.25 Variance and standard deviation.

Measure of dispersion	Value
Population variance	6.2857
Sample variance	7.3333
Population standard deviation	2.5071
Sample standard deviation	2.7080

Table 2.26 Deviations about the mean value.

9	13	12	7	6	11	12
---	----	----	---	---	----	----

In the dataset of Table 2.26 the mean is 10.

And the deviation of the data around the mean value of 10 is as follows:

$$(9 - 10) + (13 - 10) + (12 - 10) + (7 - 10) \\ + (6 - 10) + (11 - 10) + (12 - 10) = 0$$

This is perhaps a logical conclusion since the mean value is calculated from all the dataset values.

## Coefficient of variation and the standard deviation

The standard deviation as a measure of dispersion on its own is not easy to interpret. In general terms a small value for the standard deviation indicates that the dispersion of the data is low and conversely the dispersion is large for a high value of the standard deviation. However the magnitude of these values depends on what you are analysing. Further, how small is small and what about the units? If you say that the standard deviation of the total travel time, including waiting, to fly from London to Vladivostok is 2 hours, the number 2 is small. However, if you convert that to minutes the value is 120, and a high 7,200, if you use seconds. But in any event, the standard deviation has not changed!

A way to overcome the difficulty in interpreting the standard deviation is to include the value of the mean of the dataset and use the **coefficient of variation**. The coefficient of variation is a relative measure of the standard deviation of a distribution,  $\sigma$ , to its mean,  $\mu$ . The

coefficient of variation can be either expressed as a proportion or a percentage of the mean. It is defined as follows:

$$\text{Coefficient of variation} = \frac{\sigma}{\mu} \quad 2(x)$$

As an illustration, say that a machine is cutting steel rods used in automobile manufacturing where the average length is 1.5 m, and the standard deviation of the length of the rods that are cut is 0.25 cm or 0.0025 m. In this case the coefficient of variation is 0.25/150 (keeping all units in cm), which is 0.0017 or 0.17%. This value is small and perhaps would be acceptable from a quality control point of view. However, say that the standard deviation is 6 cm or 0.06 m. The value 0.06 is a small number but it gives a coefficient of variation of  $0.06/1.50 = 0.04$  or 4%. This value is probably unacceptable for precision engineering in automobile manufacturing.

The coefficient of variation is also a useful measure to compare two sets of data. For example, in a manufacturing operation two operators are working on each of two machines. Operator A produces an average of 45 units/day, with a standard deviation of the number of pieces produced of 8 units. Operator B completes on average 125 units/day with a standard deviation of 14 units. Which operator is the most consistent in the activity? If we just examine the standard deviation, it appears that Operator B has more variability or dispersion than Operator A, and thus might be considered more erratic. However if we compare the coefficient of variations, the value for Operator A is  $8/45$  or 17.78% and for Operator B it is  $14/125$  or 11.20%. On this comparative basis, the variability for Operator B is less than for Operator A because the mean output for Operator B is more. Table 2.27 gives a summary.

The term  $\sigma/\mu$  is strictly for the population distribution. However, in absence of the values for the population, sample values of  $s/\bar{x}$  will give you an estimation of the coefficient of variation.

Table 2.27 Coefficient of variation.

	Mean output $\mu$	Standard deviation $\sigma$	Coefficient of variation $\sigma/\mu$ (%)
Operator A	45	8	17.78
Operator B	125	14	11.20

## Quartiles

In the earlier section on *Central Tendency of Data* we introduced the median or the value that divides ordered data into two equal parts. Another divider of data is the **quartiles** or those values that divide ordered data into four equal parts, or four equal quarters. With this division of data the positioning of information within the quartiles is also a measure of dispersion. Quartiles are useful to indicate where data such as student's grades, a person's weight, or sales' revenues are positioned relative to standardized data.

### Boundary limits of quartiles

The lower limit of the quartiles is the minimum value of the dataset, denoted as  $Q_0$ , and the upper limit is the maximum value  $Q_4$ . Between these two values is contained 100% of the dataset. There are then three quartiles within these outer limits. The 1st quartile is  $Q_1$ , the 2nd quartile  $Q_2$ , and the 3rd quartile  $Q_3$ . We then have the **boundary limits of the quartiles** which are those values that divide the dataset into four equal parts such that within each of these boundaries there is 25% of the data. In summary then there are the following five boundary limits:

$Q_0$	$Q_1$	$Q_2$	$Q_3$	$Q_4$
-------	-------	-------	-------	-------

The quartile values can be determined by using in Excel **[function QUARTILE]**.

Table 2.28 Quartiles for sales revenues.

35,378	170,569	104,985	134,859	120,958	107,865	127,895	106,825	130,564	108,654
109,785	184,957	96,598	121,985	63,258	164,295	97,568	165,298	113,985	124,965
108,695	91,864	120,598	47,865	162,985	83,964	103,985	61,298	104,987	184,562
89,597	160,259	55,492	152,698	92,875	56,879	151,895	88,479	165,698	89,486
85,479	64,578	103,985	81,980	137,859	126,987	102,987	116,985	45,189	131,958
73,598	161,895	132,689	120,654	67,895	87,653	58,975	103,958	124,598	168,592
95,896	52,754	114,985	62,598	145,985	99,654	76,589	113,590	80,459	111,489
109,856	101,894	80,157	78,598	86,785	97,562	136,984	89,856	96,215	163,985
83,695	75,894	98,759	133,958	74,895	37,856	90,689	64,189	107,865	123,958
105,987	93,832	58,975	102,986	102,987	144,985	101,498	101,298	103,958	71,589
59,326	121,459	82,198	60,128	86,597	91,786	56,897	112,854	54,128	152,654
99,999	78,562	110,489	86,957	99,486	132,569	134,987	76,589	135,698	118,654
90,598	156,982	87,694	117,895	85,632	104,598	77,654	105,987	78,456	149,562
68,976	50,128	106,598	63,598	123,564	47,895	100,295	60,128	141,298	84,598
100,296	77,498	77,856	134,890	79,432	100,659	95,489	122,958	111,897	129,564
71,458	88,796	110,259	72,598	140,598	125,489	69,584	89,651	70,598	93,876
112,987	123,895	65,847	128,695	66,897	82,459	133,984	98,459	153,298	87,265
72,312	81,456	124,856	101,487	73,569	138,695	74,583	136,958	115,897	142,985
119,654	96,592	66,598	81,490	139,584	82,456	150,298	106,859	68,945	122,654
70,489	94,587	85,975	138,597	97,498	143,985	92,489	146,289	84,592	69,874

Quartile	Position	Value
$Q_0$	0	35,378
$Q_1$	1	79,976
$Q_2$	2	100,296
$Q_3$	3	123,911
$Q_4$	4	184,957

$Q_3 - Q_1$	Mid-spread	43,935
$(Q_3 - Q_1)/2$	Quartile deviation	21,968
$(Q_3 + Q_1)/2$	Mid-hinge	101,943
Mean		102,667

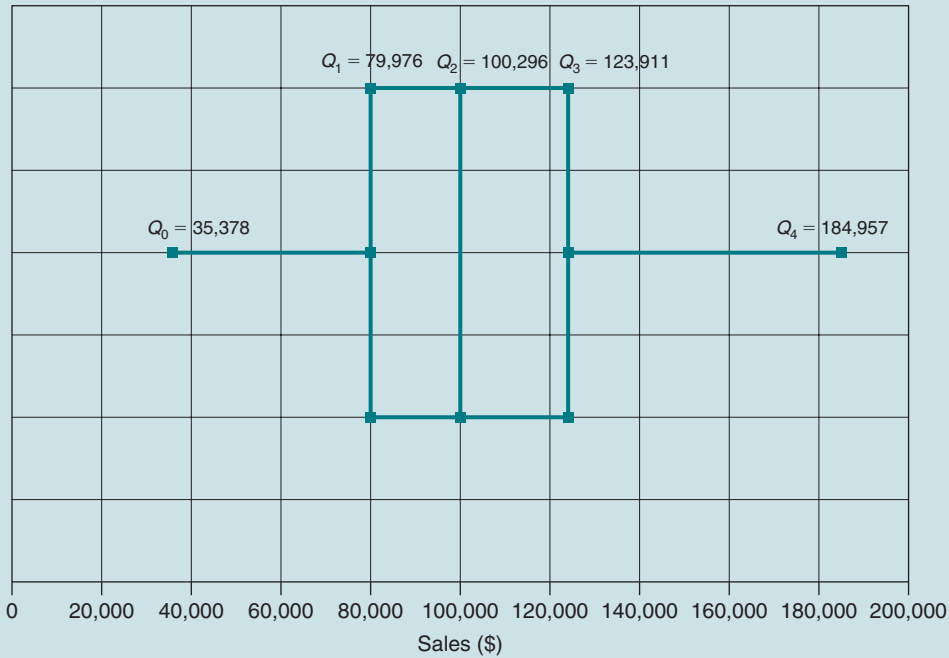
## Properties of quartiles

For the sales data of Chapter 1, we have developed the quartile values using the quartile function in Excel. This information is shown in Table 2.28, which gives the five quartile boundary limits plus additional properties related to the quartiles. Also indicated is the **inter-quartile range**, or **mid-spread**, which is the difference between the 3rd and the 1st quartile in a dataset or  $(Q_3 - Q_1)$ . It measures the range of the middle 50% of the data. One half of the inter-quartile range,

$(Q_3 - Q_1)/2$ , is the **quartile deviation** and this measures the average range of one half of the data. The smaller the quartile deviation, the greater is the concentration of the middle half of the observations in the dataset. The **mid-hinge**, or  $(Q_3 + Q_1)/2$ , is a measure of central tendency and is analogous to the midrange. Although like the range, these additional quartile properties only use two values in their calculation, distortion from extreme values is limited as the quartile values are taken from an ordered set of data.



Figure 2.1 Box and whisker plot for the sales revenues.



## Box and whisker plot

A useful visual presentation of the quartile values is a **box and whisker plot** (from the face of a cat – if you use your imagination!) or sometimes referred to as a **box plot**. The box and whisker plot for the sales data is shown in Figure 2.1. Here, the middle half of the values of the dataset or the 50% of the values that lie in the inter-quartile range are shown as a box. The vertical line making the left-hand side of the box is the 1st quartile, and the vertical line of the right-hand side of the box is the 3rd quartile. The 25% of the values that lie to the left of the box and the 25% of the values to the right of the box, or the other 50% of the dataset, are shown as two horizontal lines, or whiskers. The extreme left part of the first whisker is the minimum value,  $Q_0$ , and the extreme right part of the second whisker is the

maximum value,  $Q_4$ . The larger the width of the box relative to the two whiskers indicates that the data is clustered around the middle 50% of the values.

The box and whisker plot is **symmetrical** if the distances from  $Q_0$  to the median,  $Q_2$ , and the distance from  $Q_2$  to  $Q_4$  are the same. In addition, the distance from  $Q_0$  to  $Q_1$  equals the distance from  $Q_3$  to  $Q_4$  and the distance from  $Q_1$  to  $Q_2$  equals the distance from the  $Q_2$  to  $Q_3$  and further the mean and the median values are equal. The box and whisker plot is **right-skewed** if the distance from  $Q_2$  to  $Q_4$  is greater than the distance from  $Q_0$  to  $Q_2$  and the distance from  $Q_3$  to  $Q_4$  is greater than the distance from  $Q_0$  to  $Q_1$ . Also, the mean value is greater than the median. This means that the data values to the right of the median are more dispersed than those to the



left of the median. Conversely, the box and whisker plot is **left-skewed** if the distance from  $Q_2$  to  $Q_4$  is less than the distance from  $Q_0$  to  $Q_2$  and the distance from  $Q_3$  to  $Q_4$  is less than the distance from  $Q_0$  to  $Q_1$ . Also, the mean value is less than the median. This means that the data values to the left of the median are more dispersed than those to the right. The box and whisker plot in Figure 2.1 is slightly right-skewed. There is further discussion on the skewed properties of data in Chapter 5 in the paragraph entitled *Asymmetrical Data*.

## Drawing the box and whiskerplot with Excel

If you do not have add-on functions with Microsoft Excel one way to draw the box and whisker plot is to develop a horizontal and vertical line graph. The  $x$ -axis is the quartile values and the  $y$ -axis has the arbitrary values 1, 2, and 3. As the box and whisker plot has only three horizontal lines the lower part of the box has the arbitrary  $y$ -value of 1; the whiskers and the centre part of the box have the arbitrary value of 2; and the upper part of the box has the arbitrary value of 3. The procedure for drawing the box and whisker plot is as follows.

Determine the five quartile boundary values  $Q_0$ ,  $Q_1$ ,  $Q_2$ ,  $Q_3$ , and  $Q_4$  using the Excel quartile function. Set the coordinates for the box and whisker plot in two columns using the format in Table 2.29. For the 2nd column you enter the corresponding quartile value.

The reason that there are 13 coordinates is that when Excel creates the graph it connects every coordinate with a horizontal or vertical straight line to arrive at the box plot including going over some coordinates more than once.

Say once we have drawn the box and whisker plot, the sales data from which it is constructed is considered our reference or **benchmark**. We now ask the question, where would we position Region A which has sales of \$60,000, Region

**Table 2.29** Coordinates for a box and whisker plot.

Point No.	X	Y
1	$Q_0$	2
2	$Q_1$	2
3	$Q_1$	3
4	$Q_2$	3
5	$Q_2$	1
6	$Q_1$	1
7	$Q_1$	3
8	$Q_3$	3
9	$Q_3$	1
10	$Q_2$	1
11	$Q_3$	1
12	$Q_3$	2
13	$Q_4$	2

B which has sales of \$90,000, Region C which has sales of \$120,000, and Region D which has sales of \$150,000? From the box and whisker plot of Figure 2.1 an amount of \$60,000 is within the 1st quartile and not a great performance; \$90,000 is within the 2nd quartile or within the box or the middle 50% of sales. Again the performance is not great. An amount of \$120,000 is within the 3rd quartile and within the box or the middle 50% of sales and is a better performance. Finally, an amount of \$150,000 is within the 4th quartile and a superior sales performance. As mentioned in Chapter 1, a box and whisker plot is another technique in exploratory data analysis (EDA) that covers methods to give an initial understanding of the characteristics of data being analysed.

## Percentiles

The **percentiles** divide data into 100 equal parts and thus give a more precise positioning of where information stands compared to the quartiles. For example, paediatricians will measure

Table 2.30 Percentiles for sales revenues.

Percentile (%)	Value (\$)	Percentile (%)	Value (\$)	Percentile (%)	Value (\$)	Percentile (%)	Value (\$)	Percentile (%)	Value (\$)
0	35,378								
1	45,116	21	76,589	41	92,717	61	107,865	81	132,592
2	47,894	22	77,620	42	93,858	62	108,670	82	133,963
3	52,675	23	78,318	43	95,101	63	109,811	83	134,864
4	55,437	24	78,589	44	96,075	64	110,342	84	135,101
5	56,896	25	79,976	45	96,595	65	111,632	85	136,962
6	58,975	26	81,197	46	97,533	66	112,899	86	137,962
7	60,072	27	81,848	47	98,040	67	113,720	87	138,811
8	61,204	28	82,384	48	99,137	68	115,277	88	140,682
9	63,199	29	83,337	49	99,830	69	117,267	89	143,095
10	64,130	30	84,404	50	100,296	70	118,954	90	145,085
11	65,707	31	85,206	51	100,972	71	120,614	91	146,584
12	66,861	32	85,865	52	101,492	72	121,098	92	150,426
13	68,809	33	86,723	53	102,407	73	122,166	93	152,657
14	69,499	34	87,160	54	102,987	74	123,116	94	153,519
15	70,397	35	87,680	55	103,958	75	123,911	95	160,341
16	71,320	36	88,682	56	103,985	76	124,660	96	163,025
17	72,189	37	89,556	57	104,764	77	125,086	97	164,325
18	73,394	38	89,778	58	105,407	78	127,187	98	165,756
19	74,396	39	90,654	59	106,238	79	128,877	99	170,709
20	75,694	40	91,833	60	106,839	80	130,843	100	184,957

the height and weight of small children and indicate how the child compares with others in the same age range using a percentile measurement. For example assume the paediatrician says that for your child's height he is in the 10th percentile. This means that only 10% of all children in the same age range have a height less than your child, and 90% have a height greater than that of your child. This information can be used as an indicator of the growth pattern of the child. Another use of percentiles is in examination grading to determine in what percentile, a student's grade falls.

## Development of percentiles

We can develop the percentiles using `[function PERCENTILE]` in Excel. When you call up this function you are asked to enter the dataset and the

value of the  $k$ th percentile where  $k$  is to indicate the 1st, 2nd, 3rd percentile, etc. When you enter the value of  $k$  it has to be a decimal representation or a percentage of 100. For example the 15th percentile has to be written as 0.15 or 15%. As for the quartiles, you do not have to sort the data – Excel does this for you. Using the same sales revenue information that we used for the quartiles, Table 2.30 gives the percentiles for this information using the percentage to indicate the percentile. For example a percentile of 15% is the 15th percentile or a percentile of 23% is the 23rd percentile. Using this data we have developed Figure 2.2, which shows the percentiles as a histogram.

Say once again as we did for the quartiles, we ask the question, where would we position Region A which has sales of \$60,000, Region B which has sales of \$90,000, Region C which has sales of \$120,000, and Region D which has sales of

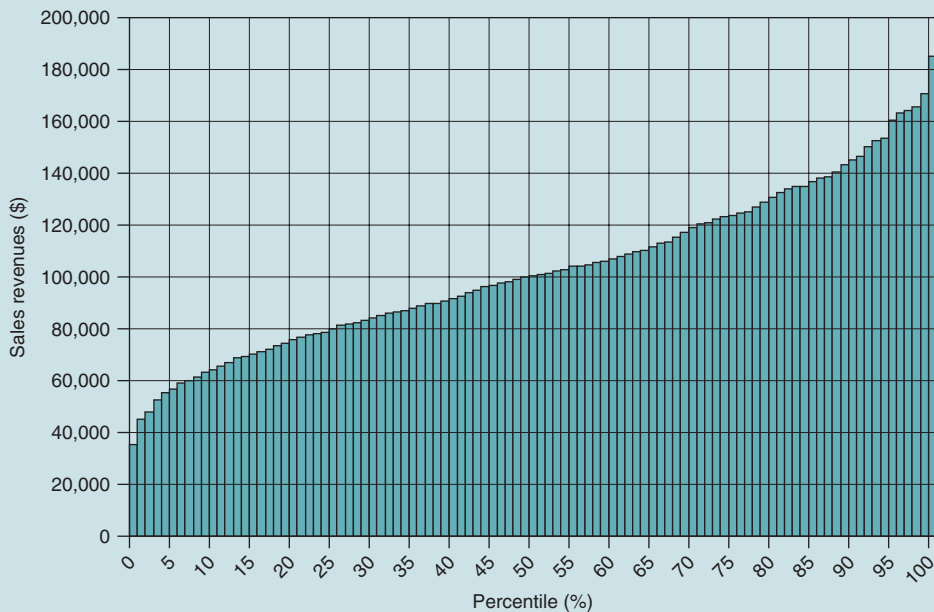
\$150,000? From either Table 2.30 or Figure 2.2 then we can say that for \$60,000 this is at about the 7% percentile, which means that 93% of the sales are greater than this region and 7% are less – a poor performance. For \$90,000 this is roughly the 39% percentile, which means that 61% of the sales are greater than this region and 39% are less – not a good performance. At the \$120,000 level this is about the 71% percentile, which means that 29% of the sales are greater than this region and 71% are less – a reasonable performance. Finally \$150,000 is at roughly the 92% percentile which signifies that only 8% of the sales are greater than this region and 92% are less – a good performance. By describing the data using percentiles rather than using quartiles

we have been able to be more precise as to where the region sales data are positioned.

## Division of data

We can divide up data by using the median – two equal parts, by using the quartiles – four equal parts, or using the percentiles – 100 equal parts. In this case the median value equals the 3rd quartile which also equals the 50th percentile. For the raw sales data given in Table 1.1 the median value is 100,296 (indicated at the end of paragraph median of this chapter), the value of the 2nd quartile,  $Q_2$ , given in Table 2.28, is also 100,296 and the value of the 50th percentile, given in Table 2.30, is also 100,296.

Figure 2.2 Percentiles of sales revenues.



This chapter has detailed the meaning and calculation of properties of statistical data, which we have classified by central tendency, dispersion, quartiles, and percentiles.

### Central tendency of data

Central tendency is the clustering of data around a central or a middle value. If we know the central tendency this gives us a benchmark to situate a dataset and use this central value to compare one dataset with another. The most common measure of central tendency is the mean or average value, which is the sum of the data divided by the number of data points. The mean value can be distorted by extreme value or outliers. We also have the median, or that value that divides data into two halves. The median is not affected by extreme values but may be affected by the number of values. The mode is a measure of central tendency and is that value that occurs most often. The mode can be used for qualitative responses such as the colour that is preferred. There is the midrange, which is the average of the highest and lowest value in the dataset and this is very much dependent on extreme values. We might use the weighted average when certain values are more important than others. If data are changing over time, as for example interest rates each year, then we would use the geometric mean as the measure of central tendency.

### Dispersion of data

The dispersion is the way that data is spread out. If we know how data is dispersed, it gives us an indicator of its reliability for analytical purposes. Data that is highly dispersed is unreliable compared to data that is little dispersed. The range is an often used measure of dispersion but it is not a good property as it is affected by extreme values. The most meaningful measures of dispersion are the variance and the standard deviation, both of which take into consideration every value in the dataset. Mathematically the standard deviation is the square root of the variance and it is more commonly used than the variance since it has the same units of the dataset from which it is derived. The variance has squared units. For a given dataset, the standard deviation of the sample is always more than the standard deviation of the population since it uses the value of the sample size less one in its denominator whereas the population standard deviation uses in the denominator the number of data values. A simple way to compare the relative dispersion of datasets is to use the coefficient of variation, which is the ratio of the standard deviation to its mean value.

### Quartiles

The quartiles are those values that divide ordered data into four equal values. Although, there are really just three quartiles,  $Q_1$  – the first,  $Q_2$  – the second, and  $Q_3$  – the third, we also refer to  $Q_0$ , which is the start value in the quartile framework and also the minimum value. We also have  $Q_4$ , which is the last value in the dataset, or the maximum value. Thus there are five quartile boundary limits. The value of the 2nd quartile,  $Q_2$ , is also the median value as it divides the data into two halves. By developing quartiles we can position information within the quartile framework and this is an indicator of its importance in the dataset. From the quartiles we can

develop a box and whisker plot, which is a visual display of the quartiles. The middle box represents the middle half, or 50% of the data. The left-hand whisker represents the first 25% of the data, and the right-hand whisker represents the last 25%. The box and whisker plot is distorted to the right when the mean value is greater than the median and distorted to the left when the mean is less than the median. Analogous to the range, in quartiles, we have the inter-quartile range, which is the difference between the 3rd and 1st quartile values. Also, analogous to the midrange we have the mid-hinge which is the average of the 3rd and 1st quartile.

## Percentiles

Percentiles are those values that divide ordered data into 100 equal parts. Percentiles are useful in that by positioning where a value occurs in a percentile framework you can compare the importance of this value. For example, in the medical profession an infant's height can be positioned on a standard percentile framework for children's height of the same age group which can then be an estimation of the height range of this child when he/she reaches adulthood. The 50th percentile in a dataset is equal to the 2nd quartile both of which are equal to the median value.

## EXERCISE PROBLEMS

### 1. Billing rate

#### Situation

An engineering firm uses senior engineers, junior engineers, computing services, and assistants on its projects. The billing rate to the customer for these categories is given in the table below together with the hours used on a recent design project.

Category	Senior engineers	Junior engineers	Computing services	Assistants
Billing rate (\$/hour)	85.00	45.00	35.00	22.00
Project hours	23,000	37,000	19,000	9,500

#### Required

1. If this data was used for quoting on future projects, what would be the correct average billing rate used to price a project?
2. If the estimate for performing a future job were 110,000 hours, what would be the billing amount to the customer?
3. What would be the billing rate if the straight arithmetic average were used?

### 2. Delivery

#### Situation

A delivery company prices its services according to the weight of the packages in certain weight ranges. This information together with the number of packages delivered last year is given in the table below.

Weight category	Less than 1 kg	From 1 to 5 kg	From 5 to 10 kg	From 10 to 50 kg	Greater than 50 kg
Price (\$/package)	10.00	8.00	7.00	6.00	5.50
Number of packages	120,000	90,500	82,545	32,500	950

#### Required

1. What is the average price paid per package?
2. If next year it was estimated that 400,000 packages would be delivered, what would be an estimate of revenues?

### 3. Investment

#### Situation

Antoine has \$1,000 to invest. He has been promised two options of investing his money if he leaves it invested over a period of 10 years with interest calculated annually. The interest rates for the following two options are in the tables below.

Option 1		Option 2	
Year	Interest rate (%)	Year	Interest rate (%)
1	6.00	1	8.50
2	7.50	2	3.90
3	8.20	3	9.20
4	7.50	4	3.20
5	4.90	5	4.50
6	3.70	6	7.30
7	4.50	7	4.70
8	6.70	8	3.20
9	9.10	9	6.50
10	7.50	10	9.70

#### Required

1. What is the average annual growth rate, geometric mean for Option 1?
2. What is the average annual growth rate, geometric mean for Option 2?
3. What would be the value of his investment at the end of 10 years if he invested in Option 1?
4. What would be the value of his investment at the end of 10 years if he invested in Option 2?
5. Which is the preferred investment?
6. What would need to be the interest rate in the 10th year for Option 2 in order that the value of his asset at the end of 10 years for Option 2 is the same as for Option 1?

### 4. Production

#### Situation

A custom-made small furniture company has produced the following units of furniture over the past 5 years.

Year	2000	2001	2002	2003	2004
Production (units)	13,250	14,650	15,890	15,950	16,980

**Required**

1. What is the average percentage growth in this period?
2. If this average growth rate is maintained, what would be the production level in 2008?

**5. Euro prices****Situation**

The table below gives the prices in Euros for various items in the European Union.<sup>3</sup>

	Milk (1 l)	Renault Mégane	Big Mac	Stamp for postcard	Compact disc	Can of Coke
Austria	0.86	15,650	2.50	0.51	19.95	0.50
Belgium	0.84	13,100	2.95	0.47	21.99	0.47
Finland	0.71	21,700	2.90	0.60	21.99	1.18
France	1.11	15,700	3.00	0.48	22.71	0.40
Germany	0.56	17,300	2.65	0.51	17.99	0.35
Greece	1.04	16,875	2.11	0.59	15.99	0.51
Ireland	0.83	17,459	2.54	0.38	21.57	0.70
Italy	1.34	14,770	2.50	0.41	14.98	0.77
Luxembourg	0.72	12,450	3.10	0.52	17.50	0.37
The Netherlands	0.79	16,895	2.60	0.54	22.00	0.45
Portugal	0.52	20,780	2.24	0.54	16.93	0.44
Spain	0.69	14,200	2.49	0.45	16.80	0.33

**Required**

1. Determine the maximum, minimum, range, average, midrange, median, sample standard deviation, and the estimated coefficient of variation using the sample values for all of the items indicated.
2. What observations might you draw from these characteristics?

**6. Students****Situation**

A business school has recorded the following student enrolment over the last 5 years

Year	1997	1998	1999	2000	2001
Students	3,275	3,500	3,450	3,600	3,800

**Required**

1. What is the average percentage increase in this period?
2. If this rate of percentage increase is maintained, what would be the student population in 2005?

<sup>3</sup> *International Herald Tribune*, 5/6 January 2002, p. 4.



## 7. Construction

### Situation

A firm purchases certain components for its construction projects. The price of these components over the last 5 years has been as follows.

Year	1996	1997	1998	1999	2000
Price (\$/unit)	105.50	110.80	115.45	122.56	125.75

### Required

1. What is the average percentage price increase in this period?
2. If this rate of price increase is maintained, what would be the price in 2003?

## 8. Net worth

### Situation

A small firm has shown the following changes in net worth over a 5-year period.

Year	2000	2001	2002	2003	2004
Growth (%)	6.25	9.25	8.75	7.15	8.90

### Required

1. What is the average change in net worth over this period?

## 9. Trains

### Situation

A sample of the number of late trains each week, on a privatized rail line in the United Kingdom, was recorded over a period as follows.

25	13	20	32	3
15	42	25	25	38
20	39	15	30	7
17	45	36	25	10
42	35	7	15	25

### Required

1. From this information, what is the average number of trains late?
2. From this information, what is the median value of the number of trains late?
3. From this information, what is the mode value of the number of trains late? How many times does this modal value occur?
4. From this information, what is the range?
5. From this information, what is the midrange?
6. From this information, what is the sample variance?
7. From this information, what is the sample standard deviation?
8. From this sample information, what is an estimate of the coefficient of variation?
9. What can you say about the distribution of the data?

## 10. Summer Olympics 2004

### Situation

The table below gives the final medal count for the Summer Olympics 2004 held in Athens, Greece.<sup>4</sup>

Country	Gold	Silver	Bronze	Country	Gold	Silver	Bronze
Argentina	2	0	4	Japan	16	9	12
Australia	17	16	16	Kazakhstan	1	4	3
Austria	2	4	1	Kenya	1	4	2
Azerbaijan	1	0	4	Latvia	0	4	0
Bahamas	1	0	1	Lithuania	1	2	0
Belarus	2	6	7	Mexico	0	3	1
Belgium	1	0	2	Mongolia	0	0	1
Brazil	4	3	3	Morocco	2	1	0
Britain	9	9	12	The Netherlands	4	9	9
Bulgaria	2	1	9	New Zealand	3	2	0
Cameroon	1	0	0	Nigeria	0	0	2
Canada	3	6	3	North Korea	0	4	1
Chile	2	0	1	Norway	5	0	1
China	32	17	14	Paraguay	0	1	0
Columbia	0	0	1	Poland	3	2	5
Croatia	1	2	2	Portugal	0	2	1
Cuba	9	7	11	Romania	8	5	6
Czech Republic	1	3	4	Russia	27	27	38
Denmark	2	0	6	Serbia-Montenegro	0	2	0
Dominican Republic	1	0	0	Slovakia	2	2	2
Egypt	1	1	3	Slovenia	0	1	3
Eritrea	0	0	1	South Africa	1	3	2
Estonia	0	1	2	South Korea	9	12	9
Ethiopia	2	3	2	Spain	3	11	5
Finland	0	2	0	Sweden	4	1	2
France	11	9	13	Switzerland	1	1	3
Georgia	2	2	0	Syria	0	0	1
Germany	14	16	18	Taiwan	2	2	1
Greece	6	6	4	Thailand	3	1	4
Hong Kong	0	1	0	Trinidad and Tobago	0	0	1
Hungary	8	6	3	Turkey	3	3	4
India	0	1	0	Ukraine	9	5	9
Indonesia	1	1	2	United Arab Emirates	1	0	0
Iran	2	2	2	United States	35	39	29
Ireland	1	0	0	Uzbekistan	2	1	2
Israel	1	0	1	Venezuela	0	0	2
Italy	10	11	11	Zimbabwe	1	1	1
Jamaica	2	1	2				

<sup>4</sup>International Herald Tribune, 31 August 2004, p. 20.

### Required

1. If the total number of medals won is the criterion for rating countries, which countries in order are in the first 10?
2. If the number of gold medals won is the criterion for rating countries, which countries in order are in the first 10?
3. If there are three points for a gold medal, two points for a silver medal, and one point for a bronze medal, which countries in order are in the first 10? Indicate the weighted average for these 10 countries.
4. What is the average medal count per country for those who competed in the Summer Olympics?
5. Develop a histogram for the percentage of gold medals by country for those who won a gold medal. Which three countries have the highest percentage of gold medals out of all the gold medals awarded?

## 11. Printing

### Situation

A small printing firm has the following wage rates and production time in the final section of its printing operation.

Operation	Binding	Trimming	Packing
Wages (\$/hour)	14.00	13.70	15.25
Hours per 100 units	1.50	1.75	1.25

### Required

1. For product costing purposes, what is the correct average rate per hour for 100 units for this part of the printing operation?
2. If we added in printing, where the wages are \$25.00 hour and the production time is 45 minutes per 100 units, then what would be the new correct average wage rate for the operation?

## 12. Big Mac

### Situation

The table below gives the price a Big Mac Hamburger in various countries converted to the \$US.<sup>5</sup> (This is the information presented in the Box Opener.)

<sup>5</sup>See Note 1.

Country	Price (\$US)	Country	Price (\$US)
Argentina	1.64	Mexico	2.58
Australia	2.50	New Zealand	3.17
Brazil	2.39	Peru	2.76
Britain	3.44	Philippines	1.47
Canada	2.63	Poland	1.96
Chile	2.53	Russia	1.48
China	2.27	Singapore	2.17
Czech Republic	2.30	South Africa	2.10
Denmark	4.58	South Korea	2.49
Egypt	1.55	Sweden	4.17
Euro zone	3.58	Switzerland	5.05
Hong Kong	1.54	Taiwan	2.41
Hungary	2.60	Thailand	1.48
Indonesia	1.53	Turkey	2.92
Japan	2.34	United States	3.06
Malaysia	1.38	Venezuela	2.13

### Required

- Determine the following characteristics of this data:
  - Maximum
  - Minimum
  - Average value
  - Median
  - Range
  - Midrange
  - Mode and how many modal values are there?
  - Sample standard deviation
  - Coefficient of variation using the sample standard deviation
- Illustrate the price of a Big Mac on a horizontal bar chart sorted according to price.
- What are the boundary limits of the quartiles?
- What is the inter-quartile range?
- Where in the quartile distribution do the prices of the Big Mac occur in Indonesia, Singapore, Hungary, and Denmark? What initial conclusions could you draw from this information?
- Draw a box and whisker plot for this data.

## 13. Purchasing expenditures – Part II

### Situation

The complete daily purchasing expenditures for a large resort hotel for the last 200 days in Euros are given in the table below. The purchases include all food and non-food items,

and wine for the five restaurants in the complex, energy including water for the three swimming pools, laundry which is a purchased service, gasoline for the courtesy vehicles, gardening and landscaping services.

63,680	307,024	188,973	242,746	217,724	194,157	230,211	192,285	235,015	195,577
197,613	332,923	173,876	219,573	113,864	295,731	175,622	297,536	205,173	224,937
195,651	165,355	217,076	86,157	293,373	151,135	187,173	110,336	188,977	332,212
161,275	288,466	99,886	274,856	167,175	102,382	273,411	159,262	298,256	161,075
153,862	116,240	187,173	147,564	248,146	228,577	185,377	210,573	81,340	237,524
132,476	291,411	238,840	217,177	122,211	157,775	106,155	187,124	224,276	303,466
172,613	94,957	206,973	112,676	262,773	179,377	137,860	204,462	144,826	194,157
197,741	183,409	144,283	141,476	156,213	175,612	246,571	161,741	173,187	295,173
150,651	136,609	177,766	241,124	134,811	68,141	163,240	115,540	194,157	223,124
190,777	168,898	106,155	185,375	185,377	260,973	182,696	182,336	187,124	128,860
106,787	218,626	147,956	108,230	155,875	165,215	102,415	203,137	97,430	274,777
179,998	141,412	198,880	156,523	179,075	238,624	242,977	137,860	244,256	213,577
163,076	282,568	157,849	212,211	154,138	188,276	139,777	190,777	141,221	269,212
124,157	90,230	191,876	114,476	222,415	86,211	180,531	108,230	254,336	152,276
180,533	139,496	140,141	242,802	142,978	181,186	171,880	221,324	201,415	233,215
128,624	159,833	198,466	130,676	253,076	225,880	125,251	161,372	127,076	168,977
203,377	223,011	118,525	231,651	120,415	148,426	241,171	177,226	275,936	157,077
130,162	146,621	224,741	182,677	132,424	249,651	134,249	246,524	208,615	257,373
215,377	173,866	119,876	146,682	251,251	148,421	270,536	192,346	124,101	220,777
126,880	170,257	154,755	249,475	175,496	259,173	166,480	263,320	152,266	125,773

### Required

- Using the raw data determine the following data characteristics:
  - Maximum value (you may have done this in the exercise from the previous chapter)
  - Minimum value (you may have done this in the exercise from the previous chapter)
  - Range
  - Midrange
  - Average value
  - Median value
  - Mode and indicate the number of modal values
  - Sample variance
  - Standard deviation (assuming a sample)
  - Coefficient of variation on the basis of a sample
- Determine the boundary limits for the quartile values for this data.
- Construct a box and whisker plot.
- What can you say about the distribution of this data?
- Determine the percentile values for this data. Plot this information on a histogram with the  $x$ -axis being the percentile value, and the  $y$ -axis the dollar value of the retail sales. Verify that the median value, 2nd quartile, and the 50th quartile are the same.

## 14. Swimming pool – Part II

### Situation

A local community has a heated swimming pool, which is open to the public each year from May 17 until September 13. The community is considering building a restaurant facility in the swimming pool area but before a final decision is made, it wants to have assurance that the receipts from the attendance at the swimming pool will help finance the construction and operation of the restaurant. In order to give some justification to its decision the community noted the attendance for one particular year and this information is given below.

869	755	729	926	821	709	1,088	785	830	709
678	1,019	825	843	940	826	750	835	956	743
835	630	791	795	903	790	931	869	878	808
845	692	830	794	993	847	901	837	755	810
791	609	878	778	761	763	726	745	874	728
870	798	507	763	764	779	678	690	1,004	792
848	823	769	773	919	682	672	829	915	883
699	650	780	743	861	610	582	748	744	680
930	776	871	759	580	669	716	980	724	880
669	712	732	968	620	852	749	860	811	748
822	651	539	658	796	825	685	707	895	806
609	952	565	869	560	751	790	907	621	619

### Required

- From this information determine the following properties of the data:
  - The sample size
  - Maximum value
  - Minimum value
  - Range
  - Midrange
  - Average value
  - Median value
  - Modal value and how many times does this value occur?
  - Standard deviation if the data were considered a sample (which it is)
  - Standard deviation if the data were considered a population
  - Coefficient of variation
  - The quartile values
  - The inter-quartile range
  - The mid-hinge
- Using the quartile values develop a box and whisker plot.
- What are your observations about the box plot?
- Determine the percentiles for this data and plot them on a histogram.

## 15. Buyout – Part II

### Situation

Carrefour, France, is considering purchasing the total 50 retail stores belonging to Hardway, a grocery chain in the Greater London area of the United Kingdom. The profits from these 50 stores, for one particular month, in £'000s, are as follows.

8.1	11.8	8.7	10.6	9.5
9.3	11.5	10.7	11.6	7.8
10.5	7.6	10.1	8.9	8.6
11.1	10.2	11.1	9.9	9.8
11.6	15.1	12.5	6.5	7.5
10.3	12.9	9.2	10.7	12.8
12.5	9.3	10.4	12.7	10.5
10.3	11.1	9.6	9.7	14.5
13.7	6.7	11.5	8.4	10.3
13.7	11.2	7.3	5.3	12.5

### Required

- Using the raw data determine the following data characteristics:
  - Maximum value (this will have been done in the previous chapter)
  - Minimum value (this will have been done in the previous chapter)
  - Range
  - Midrange
  - Average value
  - Median value
  - Modal value and indicate the order of modality (single, bi, tri, etc.)
  - Standard deviation assuming the data was a sample
  - Standard deviation taking the data correctly as the population
- Determine the quartile values for the data and use these to develop a box and whisker plot.
- Determine the percentile values for the data and plot these on a histogram.

## 16. Case: Starting salaries

### Situation

A United States manufacturing company in Chicago has several subsidiaries in the 27 countries of the European Union including Calabas, Spain; Watford, United Kingdom; Bonn, Germany and Louny, Czech Republic. It is planning to hire new engineers to work in these subsidiaries and needs to decide on the starting salary to offer these new hires. These new engineers will be hired from their country of origin to work in their home country. The human resource department of the parent firm in Chicago, who is not too

familiar with the employment practices in Europe, has the option to purchase a database of annual starting salaries for engineers in the European Union from a consulting firm in Paris. This database, with values converted to Euros, is given in the table below. It was compiled from European engineers working in the automobile, aeronautic, chemicals, pharmaceutical, textiles, food, and oil refining sectors. At the present time, the Chicago firm is considering hiring Markus Schroeder, offering a starting salary of €36,700, Xavier Perez offering a salary of €30,500, Joan Smith a salary of €32,700 and Jitka Sikorova a starting salary of €28,900. All these starting salaries include all social benefits and mandatory employer charges which have to be paid for the employee.

### Required

Assume that you work with the human resource department in Chicago. Use the information from this current chapter, and also from Chapter 1 to present in detail the salary database prepared by the Paris consulting firm. Then in using your results, describe the characteristics of the four starting salaries that have been offered and give you comments.

34,756	30,196	29,164	37,022	32,842	28,356	43,504	31,380	31,030	28,366
25,700	40,750	33,012	33,726	37,594	33,038	30,004	33,388	38,224	29,728
33,400	35,450	31,658	35,044	36,104	27,894	37,224	34,754	33,122	32,310
33,800	27,662	33,208	31,752	39,724	33,866	36,032	34,132	30,180	32,380
31,634	24,370	35,136	33,858	30,456	30,538	29,052	29,796	34,978	29,110
34,786	33,936	33,586	30,530	30,568	31,178	24,652	27,580	40,160	33,926
33,928	32,932	30,774	30,914	36,750	27,280	26,886	33,152	36,580	35,324
27,956	26,016	25,802	29,722	34,454	35,964	23,282	29,908	29,772	27,200
37,198	31,056	34,852	30,370	30,828	26,776	28,650	33,958	28,974	35,204
26,752	28,478	29,264	37,898	32,724	34,082	29,948	34,410	32,456	29,928
32,884	25,974	21,566	26,310	31,836	35,898	27,396	28,292	35,784	32,220
24,342	36,302	32,184	34,788	32,098	30,044	31,610	36,282	24,842	24,742
29,514	35,566	27,556	39,886	31,468	33,302	37,980	38,174	35,644	37,370
35,072	27,400	35,838	29,858	24,062	28,606	28,012	34,442	35,018	31,638
26,154	29,706	33,850	28,668	31,870	25,572	26,576	28,758	32,580	24,114
34,878	31,860	31,216	31,668	37,490	42,072	29,242	28,086	25,054	33,248
33,654	25,892	34,902	33,294	31,712	32,312	27,546	31,472	34,020	32,704
40,202	27,252	28,870	36,414	29,586	28,906	35,434	34,332	33,564	34,268
24,246	35,214	33,102	29,274	29,454	34,126	29,412	31,588	30,766	31,052
34,614	31,630	31,024	29,242	30,924	35,032	37,334	26,660	36,616	25,342
30,076	31,902	35,114	32,348	41,184	28,972	34,588	27,312	30,404	31,478
26,422	31,648	33,078	33,640	34,240	25,632	24,528	31,188	30,006	34,650
28,466	27,616	33,994	35,368	33,804	27,050	34,070	36,012	36,410	31,840
39,782	28,378	29,328	36,144	33,010	28,592	32,782	36,774	39,144	36,902
28,662	27,522	29,200	31,992	30,564	30,762	25,860	35,620	25,192	41,490
34,250	25,212	35,678	24,912	35,648	36,622	36,884	29,488	29,060	38,692
29,052	33,884	35,202	38,824	33,376	36,488	31,982	34,902	33,068	34,518
25,146	27,834	38,990	34,944	23,394	30,276	37,124	26,756	32,142	30,388
27,624	28,718	36,828	26,528	29,168	31,612	28,822	29,296	28,374	29,990

(Continued)



30,914	34,812	37,508	30,446	35,390	38,916	33,842	25,442	28,088	28,234
34,652	34,286	26,474	35,394	36,636	34,596	36,196	32,412	31,272	31,822
29,696	33,552	28,900	30,384	32,274	22,320	28,934	35,738	36,010	39,038
22,044	34,022	37,750	27,146	34,570	29,514	31,042	30,672	33,482	34,774
36,518	29,638	28,976	31,146	38,434	27,468	39,570	28,502	31,762	38,600
27,134	32,334	27,928	31,150	31,858	31,544	27,254	27,716	41,482	27,082
32,470	28,128	28,584	33,120	36,764	35,450	32,854	31,848	33,474	26,842
33,396	27,358	33,832	38,088	35,074	29,114	26,380	31,256	37,080	29,622
33,060	25,426	31,616	31,876	35,838	29,376	36,654	30,398	36,030	34,196
29,732	29,744	30,544	31,854	30,884	23,768	31,520	30,336	27,442	29,796
39,302	25,424	28,924	32,072	29,204	34,906	32,434	23,710	31,964	33,328
39,956	33,386	38,184	35,326	34,468	37,616	35,588	37,312	32,484	29,522
31,332	35,136	35,186	32,964	31,962	34,070	41,396	28,170	35,352	31,300
24,190	37,292	33,146	32,972	30,260	31,178	26,772	39,376	31,860	37,080
32,568	27,990	32,378	22,508	32,644	27,158	31,868	33,050	29,624	32,368
28,176	27,664	31,840	26,800	33,252	32,622	35,966	29,264	31,546	26,292
31,116	30,834	30,254	30,690	23,930	31,202	32,166	30,396	33,698	29,704
31,496	33,730	31,714	34,046	29,756	38,372	35,666	31,344	35,976	33,036
27,500	33,882	40,496	32,218	30,110	36,168	31,654	28,880	27,502	29,082
36,524	30,512	33,882	34,350	39,062	24,674	33,384	27,472	21,954	27,934
33,346	33,426	31,722	29,566	31,000	30,522	33,942	32,490	35,134	29,644
38,754	32,214	31,220	32,604	23,588	29,648	32,470	38,824	30,820	34,294
22,856	29,514	28,000	35,398	31,934	27,104	34,994	25,006	31,186	35,164
28,352	26,626	36,052	31,134	34,064	32,186	29,724	36,968	32,558	34,596
34,646	29,832	33,784	36,346	33,692	41,182	29,374	36,574	26,868	37,596
25,132	30,618	23,684	33,918	35,336	26,862	35,756	31,754	28,090	28,236
30,780	32,894	26,608	30,890	33,530	34,210	31,072	36,742	32,982	41,776
33,250	36,836	32,390	29,626	38,642	29,406	27,086	27,902	36,370	30,522
29,572	30,944	33,000	34,314	31,148	35,300	24,016	27,878	38,818	33,910
26,838	34,214	33,470	31,070	32,100	28,982	27,632	28,432	31,854	32,852
32,596	25,810	36,426	33,452	31,704	34,938	30,704	35,736	34,682	36,700
34,238	27,012	34,812	30,624	30,418	34,730	33,134	30,692	32,142	34,450
34,766	31,824	29,126	34,594	31,088	39,328	27,676	34,518	30,296	35,742
22,830	33,332	33,486	31,544	32,932	29,596	28,628	35,662	27,524	35,074
37,378	33,426	30,336	29,462	32,180	35,530	36,288	32,148	27,738	30,110
29,610	29,252	36,378	33,632	28,574	26,076	33,118	28,660	35,970	35,806
30,698	32,620	28,642	35,738	34,744	34,828	29,520	23,676	32,424	32,538
27,782	29,062	27,266	30,916	29,868	29,746	35,976	32,204	30,992	35,100
38,164	30,698	31,002	34,276	30,846	29,952	27,972	35,484	31,812	32,620
31,974	31,932	33,348	27,468	33,736	27,100	31,120	30,492	39,210	28,310
27,216	31,428	33,268	29,196	29,868	35,784	31,938	33,570	27,300	37,214
28,758	30,968	33,402	36,310	37,372	35,490	35,254	23,456	29,628	29,966
32,102	41,046	31,504	26,562	33,400	28,768	32,270	27,726	32,422	30,504
27,662	29,844	30,178	33,942	32,794	27,536	27,354	29,754	31,814	29,426
31,498	39,300	26,742	40,572	36,102	32,950	20,376	35,892	29,254	36,222
30,880	30,040	25,360	36,004	28,592	29,334	26,960	25,978	27,216	32,292
33,090	27,826	31,052	31,774	32,562	32,112	26,386	37,556	28,554	23,048
36,176	27,392	37,216	24,314	24,410	36,304	29,568	33,214	31,284	37,264
25,518	33,618	36,218	31,148	26,620	31,178	31,490	28,338	26,770	31,498

31,404	32,206	34,552	34,842	26,664	24,960	32,798	22,856	33,082	32,514
26,856	29,672	33,786	30,502	31,766	31,854	35,450	29,188	32,692	29,830
28,858	36,308	30,292	30,298	32,124	31,730	33,534	35,440	28,990	29,606
29,554	25,062	28,502	34,388	31,052	34,826	34,024	33,926	32,330	33,460
32,216	32,160	40,642	27,986	33,040	36,398	36,084	25,664	29,852	37,400
29,674	33,100	32,048	30,606	34,902	34,538	32,438	28,844	30,502	29,178
26,656	26,730	26,690	31,236	35,788	29,438	33,088	28,930	27,342	32,070
36,686	30,786	27,364	35,570	39,390	28,258	35,902	33,858	27,742	31,358
39,762	33,386	37,550	30,652	24,938	33,852	30,508	30,422	34,022	29,790
33,316	26,600	29,916	31,562	22,092	32,998	34,746	35,340	30,336	32,256
30,336	31,462	31,918	31,994	25,040	30,986	32,220	26,830	28,882	29,426
33,048	31,510	33,382	32,680	35,802	36,704	29,836	31,160	33,318	25,824
37,688	34,382	34,504	31,868	30,872	36,156	42,592	33,636	38,870	25,470
34,658	30,430	36,060	37,306	39,048	35,334	28,598	32,664	34,958	39,414
42,786	43,258	35,260	35,068	30,454	30,880	34,776	29,942	26,144	26,432
25,936	31,368	26,992	26,452	28,084	28,036	28,780	36,382	35,248	32,926
36,662	27,056	27,762	28,616	34,842	28,582	37,860	31,134	36,704	29,992
37,560	32,108	29,358	27,562	29,490	31,316	35,590	33,520	30,462	28,802
35,772	34,220	34,490	29,224	37,310	30,246	27,920	30,000	35,144	29,814
28,584	32,202	35,650	24,874	36,094	34,774	38,626	30,520	24,750	28,578
39,180	30,170	30,220	29,564	34,306	33,834	34,368	27,344	32,548	32,702
30,792	23,460	31,302	29,472	25,530	29,028	34,350	33,748	35,530	31,732

*This page intentionally left blank*

# Basic probability and counting rules

## The wheel of fortune

*For many, gambling casinos are exciting establishments. The one-arm-bandits are colourful machines with flashing lights, which require no intelligence to operate. When there is a “win” coins drop noisily into aluminium receiving tray and blinking lights indicate to the world the amount that has been won. The gaming rooms for poker, or blackjack, and the roulette wheel have an air of mystery about them. The dealers and servers are beautiful people, smartly dressed, who say very little and give an aura of superiority. Throughout the casinos there are no clocks or windows so you do not see the time passing. Drinks are cheap, or maybe free, so having “a few” encourages you to take risk. The carpet patterns are busy so that you look at where the action is rather than looking at the floor.*

*When you want to go to the toilet you have to pass by rows of slot machines and perhaps on the way you try your luck!*

*Gambling used to be a by-word for racketeering. Now it has cleaned up its act and is more profitable than ever. Today the gambling industry is run by respectable corporations instead of by the Mob and it is confident of winning public acceptance. In 2004 in the United States, some 54.1 million people, or more than one-quarter of all American adults visited a casino, on average 6 times each. Poker is a particular growth area and some 18% of Americans played poker in 2004, which was a 50% increase over 2003. Together, the United States’ 445 commercial casinos, that means excluding those*

*owned by Indian tribes, had revenues in 2004 of nearly \$29 billion. Further, it paid state gaming taxes of \$4.74 billion or almost 10% more than in 2003. A survey of 201 elected officials and civic leaders, not including any from gambling dependent Nevada and New Jersey, found that 79% believed casinos had had a positive impact on their communities. Europe is no different. The company Partouche owns and operates very successful casinos in Belgium, France, Switzerland, Spain, Morocco, and Tunisia. And, let us not forget the famed casino in Monte Carlo. Just about all casinos are associated with hotels and restaurants and many others include resort settings and spas. Las Vegas immediately springs to mind. This makes the whole combination, gambling casinos, hotels, resorts, and spas a significant part of the service industry. This is where statistics plays a role.<sup>1,2</sup>*

---

<sup>1</sup> The gambling industry, *The Economist*, 24 September 2005.

<sup>2</sup> <http://www.partouche.fr>, consulted 27 September 2005.

## Learning objectives

After you have studied this chapter you will understand **basic probability rules**, **risk in system reliability**, and **counting rules**. You will then be able to apply these concepts to practical situations. The following are the specific topics to be covered.

- ✓ **Basic probability rules** • Probability • Risk • An event in probability • Subjective probability • Relative frequency probability • Classical probability • Addition rules in classical probability • Joint probability • Conditional probabilities under statistical dependence • Bayes' Theorem • Venn diagram • Application of a Venn diagram and probability in services: *Hospitality management* • Application of probability rules in manufacturing: *A bottling machine* • Gambling, odds, and probability.
- ✓ **System reliability and probability** • Series or parallel arrangements • Series systems • Parallel or backup systems • Application of series and parallel systems: *Assembly operation*.
- ✓ **Counting rules** • A single type of event: Rule No. 1 • Different types of events: Rule No. 2 • Arrangement of different objects: Rule No. 3 • Permutations of objects: Rule No. 4 • Combinations of objects: Rule No. 5.

In statistical analysis the outcome of certain situations can be reliably estimated, as there are mathematical relationships and rules that govern choices available. This is useful in decision-making since we can use these relationships to make probability estimates of certain outcomes and at the same time reduce risk.

### Basic Probability Rules

A principal objective of statistics is **inferential statistics**, which is to infer or make logical decisions about situations or populations simply by taking and measuring the data from a **sample**. This sample is taken from a **population**, which is the entire group in which we are interested. We use the information from this sample to infer conclusions about the population. For example, we are interested to know how people will vote in a certain election. We sample the opinion of 5,500 of the electorate and we use this result to estimate the opinion of the population of 35 million. Since we are extending our sample results beyond the data that we have measured,

this means that there is no guarantee but only a **probability** of being correct or of making the right decision. The corollary to this is that there is a probability or **risk** of being incorrect.

### Probability

The concept of probability is the chance that something happens or will not happen. In statistics it is denoted by the capital letter **P** and is measured on an inclusive numerical scale of 0 to 1. If we are using percentages, then the scale is from 0% to 100%. If the probability is 0% then there is absolutely no chance that an outcome will occur. Under present law, if you live in the United States, but you were born in Austria, the probability of you becoming president is 0% – in 2006, the current governor of California! At the top end of the probability scale is 100%, which means that it is certain the outcome will occur. The probability is 100% that someday you will die – though hopefully at an age way above the statistical average! Between the two extremes of 0 and 1 something might occur or might not occur. The meteorological office may announce that there is a 30% chance of rain

today, which also means that there is a 70% chance that it will not. The opposite of probability is **deterministic** where the outcome is certain on the assumption that the input data is reliable. For example if revenues are £10,000 and costs are £7,000 then it is sure that the gross profit is £3,000 (£10,000 – £7,000).

With probability something happens or it does not happen, that is the situation is **binomial**, or there are only two possible outcomes. However that does not mean that there is a 50/50 chance of being right or wrong or a 50/50 chance of winning. If you toss a fair-sided coin, one that has not been “fixed”, you have a 50% chance of obtaining heads or 50% chance of throwing tails. If you buy one ticket in a fund raising raffle then you will either win or lose. However, if there are 2,000 tickets that have been sold you have only a 1/2,000 or 0.05% chance of winning and a 1,999/2,000 or a 99.95% chance of losing!

## Risk

An extension of probability, often encountered in business situations, but also in our personal life, is risk. Here, when we extend probability to risk we are putting a value on the outcomes. In business we might invest in new technology and say that there is a 70% probability of increasing market share but this also might mean that there is a risk of losing \$100 million. To insurance companies, the probability of an automobile driver aged between 18 and 25 years having an accident is considered greater than for people in higher age groups. Thus, to the insurance company young people present a high risk and so their premiums are higher than normal. If you drink and drive the probability of you having an accident is high. In this case you risk having an accident, or perhaps the risk of killing yourself. In this case the “value” on the outcome is more than monetary.

## An event in probability

In probability we talk about an **event**. An event is the result of an activity or experiment that

has been carried out. If you obtain heads on the tossing of a coin, then “obtaining heads” would be an event. If you draw the King of Hearts from a pack of cards, then “drawing the King of Hearts” would be an event. If you select a light bulb from a production lot and it is defective then the “selection of a defective light bulb” would be an event. If you obtain an A grade on an examination, then “obtaining an A grade” would be an event. If Susan wins a lottery, “Susan winning the lottery” would be an event. If Jim wins a slalom ski competition, “Jim winning the slalom” would be an event.

## Subjective probability

One type of probability is **subjective probability**, which is qualitative, sometimes emotional, and simply based on the belief or the “gut” feeling of the person making the judgment. For example, you ask Michael, a single 22-year-old student what is the probability of him getting married next year? His response is 0%. You ask his friend John, what he thinks is the probability of Michael getting married next year and his response is 50%. These are qualitative responses. There are no numbers involved, and this particular situation has never occurred before. (Michael has never been married.)

Subjective probability may be a function of a person’s experience with a situation. For example, Salesperson A says that he is 80% certain of making a sale with a certain client, as he knows the client well. However, Salesperson B may give only a 50% probability level of making that sale. Both are basing their arguments on subjective probability. A manager who knows his employees well may be able to give a subjective probability of his department succeeding in a particular project. This probability might differ from that of an outsider assessing the probability of success. Very often, the subjective probability of people who are prepared to take risks, or risk takers, is higher than those persons who are risk averse, or afraid to take risks, since

Table 3.1 Composition of a pack of cards with no jokers.

Suit															Total
Hearts	Ace	1	2	3	4	5	6	7	8	9	10	Jack	Queen	King	13
Clubs	Ace	1	2	3	4	5	6	7	8	9	10	Jack	Queen	King	13
Spades	Ace	1	2	3	4	5	6	7	8	9	10	Jack	Queen	King	13
Diamonds	Ace	1	2	3	4	5	6	7	8	9	10	Jack	Queen	King	13
Total	4	4	4	4	4	4	4	4	4	4	4	4	4	4	52

the former are more optimistic or gung ho individuals.

### Relative frequency probability

A probability based on information or data collected from situations that have occurred previously is **relative frequency probability**. We have already seen this in Chapter 1, when we developed a relative frequency histogram for the sales data given in Figure 1.2. Here, if we assume that future conditions are similar to past events, then from this Figure 1.2 we could say that there is a 15% probability that future sales will lie in the range of £95,000 to £105,000.

Relative frequency probabilities have use in many business situations. For example, data taken from a certain country indicate that in a sample of 3,000 married couples under study, one-third were divorced within 10 years of marriage. Again, on the assumption that future conditions will be similar to past conditions, we can say that in this country, the probability of being divorced before 10 years of marriage is 1/3 or 33.33%. This demographic information can then be extended to estimate needs of such things as legal services, new homes, and child-care. In collecting data for determining relative frequency probabilities, the reliability is higher if the conditions from which the data has been collected are stable and a large amount of data has been measured. Relative frequency probability is also called **empirical probability** as it is

based on previous experimental work. Also, the data collected is sometimes referred to as **historical data** as the information after it has been collected is history.

### Classical probability

A probability measure that is also the basis for gambling or betting, and thus useful if you frequent casinos, is **classical probability**. Classical probability is also known as **simple probability** or **marginal probability** and is defined by the following ratio:

$$\text{Classical probability} = \frac{\text{Number of outcomes where the event occurs}}{\text{Total number of possible outcomes}} \quad 3(i)$$

In order for this expression to be valid, the probability of the outcomes, as defined by the numerator (upper part of the ratio) must be equally likely. For example, let us consider a full pack of 52 playing cards, which is composed of the individual cards according to Table 3.1.

The total number of possible outcomes is 52, the number of cards in the pack. We know in advance that the probability of drawing an Ace of Spades, or in fact any one single card, is 1/52 or 1.92%.

Similarly in the throwing of one die there are six possible outcomes, the numbers 1, 2, 3, 4, 5, or 6. Thus, we know in advance that the probability of throwing a 5 or any other number is



1/6 or 16.67%. In the tossing of a coin there are only two possible outcomes, heads or tails. Thus the probability of obtaining heads or tails is  $\frac{1}{2}$  or 50%. These illustrations of classical probability are also referred to as **a priori probability** since we know the probability of an event in advance without the need to perform any experiments or trials.

## Addition rules in classical probability

In probability situations we might have a **mutually exclusive** event. A mutually exclusive event means that there is no connection between one event and another. They exhibit **statistical independence**. For example, obtaining heads on the tossing of a coin is mutually exclusive from obtaining tails since you can have either heads, or tails, but not both. Further, if you obtain heads on one toss of a coin this event will have no impact of the following event when you toss the coin again. In many chance situations, such as the tossing of coin, each time you make the experiment, everything resets itself back to zero. My Canadian cousins had three girls and they really wanted a boy. They tried again thinking after three girls there must be a higher probability of getting a boy. This time they had twins – two girls! The fact that they had three girls previously had no bearing on the gender of the baby on the 4th trial.

When two events are mutually exclusive then the probability of *A* or *B* occurring can be expressed by the following **addition rule for mutually exclusive events**

$$P(A, \text{ or } B) = P(A) + P(B) \quad 3(\text{ii})$$

For example, in a pack of cards, the probability of drawing the Ace of Spades,  $A_S$ , or the Queen of Hearts,  $Q_H$ , with **replacement** after the first draw, is by equation 3(ii). Replacement means that we draw a card, note its face value, and then put it back into the pack

$$P(A_S \text{ or } Q_S) = \frac{1}{52} + \frac{1}{52} = \frac{1}{26} = 3.85\%$$

If we do not replace the first card that is withdrawn, and this first card is neither the Ace of Spades, or the Queen of Hearts then the probability is given by the expression,

$$\begin{aligned} P(A_S \text{ or } Q_S) &= \frac{1}{52} + \frac{1}{51} \\ &= 0.0192 + 0.0196 = 3.88\% \end{aligned}$$

That is, a slightly higher probability than in the case with replacement.

If two events are **non-mutually exclusive**, this means that it is possible for both events to occur. If we consider for example, the probability of drawing either an Ace or a Spade from a deck of cards, then the event Ace and Spade can occur together since it is possible that the Ace of Spades could be drawn. Thus an Ace and a Spade are not mutually exclusive events. In this case, equation 3(ii) for mutually exclusive events must be adjusted to avoid double accounting, or to reduce the probability of drawing an Ace, or a Spade, by the chance we could draw both of them together, that is, the Ace of Spades. Thus, equation 3(ii) is adjusted to become the following **addition rule for non-mutually exclusive events**

$$P(A, \text{ or } B) = P(A) + P(B) - P(AB) \quad 3(\text{iii})$$

Here  $P(AB)$  is the probability of *A* and *B* happening together. Thus from equation 3(iii) the probability of drawing an Ace or a Spade is,

$$\begin{aligned} P(\text{Ace or Spade}) &= \frac{4}{52} + \frac{13}{52} - \frac{4}{52} * \frac{13}{52} \\ &= \frac{17}{52} - \frac{1}{52} = \frac{16}{52} = 30.77\% \end{aligned}$$

Or we can look at it another way:

$$\begin{aligned} P(\text{Ace}) &= \frac{4}{52} \\ &= 7.69\% \end{aligned}$$

$$\begin{aligned} P(\text{Spade}) &= \frac{13}{52} \\ &= 25.00\% \end{aligned}$$

$$P(\text{Ace of Spades}) = \frac{1}{52} = 1.92\%$$

$P(\text{Ace or a Spade}) = 7.69 + 25.00 - 1.92 = 30.77\%$  to avoid double accounting.

## Joint probability

The probability of two or more independent events occurring together or in succession is **joint probability**. This is calculated by the product of the individual marginal probabilities

$$P(AB) = P(A) * P(B) \quad 3(\text{iv})$$

Here  $P(AB)$  is the joint probability of events  $A$  and  $B$  occurring together or in succession.  $P(A)$  is the marginal probability of  $A$  occurring and  $P(B)$  is the marginal probability of  $B$  occurring. The joint probability is always less than the marginal probability since we are determining the probability of more than one event occurring together in our experiment.

Consider for example again in gambling where we are using one pack of cards. The classical or marginal probability of drawing the Ace of Spades from a pack is  $1/52$  or  $1.92\%$ . The probability of drawing the Ace of Spades both times in two successive draws with replacement is as follows:

$$\frac{1}{52} * \frac{1}{52} = \frac{1}{2,704} = 0.037\%$$

Here the value of  $0.037\%$  for drawing the Ace of Spades twice in two draws is much less than the marginal productivity of  $1.92\%$  of drawing the Ace of Spades once in a single drawing.

Assume in another gambling game, two dice are thrown together, and the total number obtained is counted. In order for the total count to be 7, the various combinations that must come up together on the dice are as given in Table 3.2.

From classical probability we know that the chance of throwing a 1 and a 6 together, the

**Table 3.2** Possible combinations for obtaining 7 on the throw of two dice.

1st die	1	2	3	4	5	6
2nd die	6	5	4	3	2	1
Total throw	7	7	7	7	7	7

combination in the 1st column, is from joint probability

$$\frac{1}{6} * \frac{1}{6} = \frac{1}{36} = 2.78\%$$

The chance of throwing a 2 and a 5 together, the combination in the 2nd column, is from joint probability

$$\frac{1}{6} * \frac{1}{6} = \frac{1}{36} = 2.78\%$$

Similarly, the joint probability for throwing a 3 and 4 together, a 4 and 3, a 5 and 2, and a 6 and 1 together is always  $2.78\%$ . Thus, the probability that all 6 can occur is determined as follows from the addition rule

$$2.78\% + 2.78\% + 2.78\% + 2.78\% + 2.78\% + 2.78\% = 16.67\%$$

This is the same result using the criteria of classical or marginal probability of equation 3(i),

$$\frac{\text{Number of outcomes where the event occurs}}{\text{Total number of possible outcomes}}$$

Here, the number of possible outcomes where the number 7 occurs is six. The total number of possible outcomes are 36 by the joint probability of  $6 * 6$ .

Thus, the probability of obtaining a 7 on the throw of two dice is  $6/36 = 16.67\%$

In order to obtain the number 5, the combinations that must come up together are according to Table 3.3.

**Table 3.3** Possible combinations for obtaining 5 on the throw of two dice.

1st die	1	2	3	4
2nd die	4	3	2	1
Total throw	5	5	5	5

The probability that all four can occur is then from the addition rule,

$$\begin{aligned}
 &2.78\% + 2.78\% + 2.78\% + 2.78\% \\
 &= 11.12\% \text{ (actually } 11.11\% \text{ if we} \\
 &\quad \text{round at the end of the} \\
 &\quad \text{calculation)}
 \end{aligned}$$

Again from marginal probabilities this is  $4/36 = 11.11\%$ .

Thus again this is *a priori* probability since in the throwing of two dice, we know in advance that the probability of obtaining a 5 is  $4/36$  or  $11.11\%$  (see also the following section *counting rules*).

In gambling with slot machines or a **one-arm-bandit**, often the winning situation is obtaining three identical objects on the pull of a lever according to Figure 3.1, where we show three apples. The probability of winning is joint probability and is given by,

$$P(A_1 \cdot A_2 \cdot A_3) = P(A_1) * P(A_2) * P(A_3) \quad 3(v)$$

If there are six different objects on each wheel, but each wheel has the same objects, then the marginal probability of obtaining one object is  $1/6 = 16.67\%$ . Then the joint probability of obtaining all three objects together is thus,

$$0.1667 * 0.1667 * 0.1667 = 0.0046 = 0.46\%$$

If there are 10 objects on each wheel then the marginal probability for each wheel is  $1/10 = 0.10$ . In this case the joint probability is  $0.10 * 0.10 * 0.10 = 0.001 = 0.10\%$  as shown in the Figure 3.1.

**Figure 3.1** Joint probability.

Probability of obtaining the same three fruits on a one-arm-bandit where there are 10 different fruits on each of the three wheels.

$$P(ABC) = P(A) * P(B) * P(C)$$



$$\text{Probability} = 0.10 * 0.10 * 0.10 = 0.0010 = 0.10\%$$

This low value explains why in the long run, most gamblers lose!

## Conditional probabilities under statistical dependence

The concept of **statistical dependence** implies that the probability of a certain event is dependent on the occurrence of another event. Consider the lot of 10 cubes given in Figure 3.2. There are four different formats. One cube is dark green and dotted; two cubes are light green and striped; three cubes are dark green and striped; and four cubes are light green and dotted. As there are 10 cubes, there are 10 possible events and the probability of selecting any one cube at random from the lot is 10%. The possible outcomes are shown in Table 3.4 according to the configuration of each cube.

Alternatively, this information can be presented in a two by two **cross-classification** or **contingency** table as in Table 3.5. This shows that we have one cube that is dark green and dotted, three cubes that are dark green and striped, four cubes that are light green and dotted, and two cubes that are light green and striped. These formats are also shown in Figure 3.3.

Assume that we select a cube at **random** from the lot. Random means that each cube has an equally chance of being chosen.

Figure 3.2 Probabilities under statistical dependence.

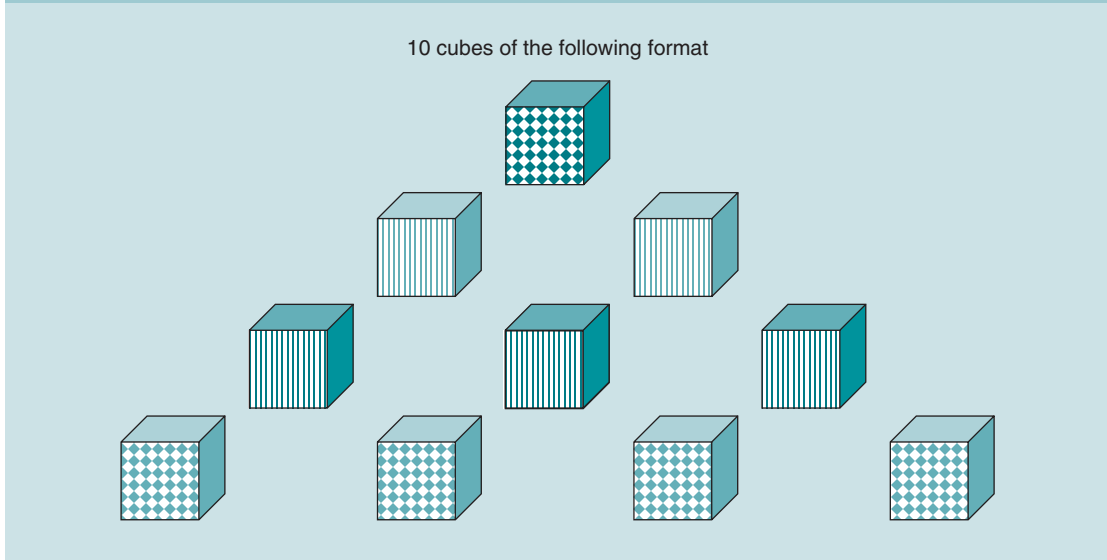


Table 3.4 Possible outcomes of selecting a coloured cube.

Event	Probability (%)	Colour	Design
1	10	Dark green	Dotted
2	10	Dark green	Striped
3	10	Dark green	Striped
4	10	Dark green	Striped
5	10	Light green	Striped
6	10	Light green	Striped
7	10	Light green	Dotted
8	10	Light green	Dotted
9	10	Light green	Dotted
10	10	Light green	Dotted

Figure 3.3 Probabilities under statistical dependence.

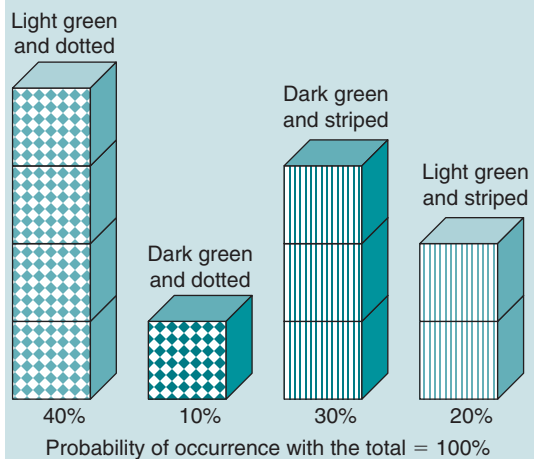


Table 3.5 Cross-classification table for coloured cubes.

	Dark green	Light green	Total
Dotted	1	4	5
Striped	3	2	5
Total	4	6	10

- The probability of the cube being light green is 6/10 or 60%.
- The probability of the cube being dark green is 4/10 or 40%.
- The probability of the cube being striped is 5/10 or 50%.

- The probability of the cube being dotted is 5/10 or 50%.
- The probability of the cube being dark green and striped is 3/10 or 30%.
- The probability of the cube being light green and striped is 2/10 or 20%.
- The probability of the cube being dark green and dotted is 1/10 or 10%
- The probability of the cube being light green and dotted is 4/10 or 40%.

Now, if we select a light green cube from the lot, what is the probability of it being dotted? The condition is that we have selected a light green cube. There are six light green cubes and of these, four are dotted, and so the probability is 4/6 or 66.67%. If we select a striped cube from the lot what is the probability of it being light green? The condition is that we have selected a striped cube. There are five striped cubes and of these two are light green, thus the probability is 2/5 or 40%. This conditional probability under statistical dependence can be written by the relationship,

$$P(B | A) = \frac{P(BA)}{P(A)} \quad 3(vi)$$

This is interpreted as saying that the probability of  $B$  occurring, on the condition that  $A$  has occurred, is equal to the joint probability of  $B$  and  $A$  happening together, or in succession, divided by the marginal probability of  $A$ .

Using the relationship from equation 3(vi) and referring to Table 3.5,

$$\begin{aligned} P(\text{striped, given light green}) &= \frac{P(\text{striped and light green})}{P(\text{light green})} \\ &= \frac{2/10}{6/10} = \frac{1}{3} = 13.33\% \end{aligned}$$

$$\begin{aligned} P(\text{dotted, given light green}) &= \frac{P(\text{dotted and light green})}{P(\text{light green})} \\ &= \frac{4/10}{6/10} = \frac{2}{3} = 33.33\% \end{aligned}$$

$$\begin{aligned} P(\text{light green, given striped}) &= \frac{P(\text{light green and striped})}{P(\text{striped})} \\ &= \frac{2/10}{5/10} = \frac{2}{5} = 40.00\% \end{aligned}$$

$$\begin{aligned} P(\text{dark green, given dotted}) &= \frac{P(\text{dark green and dotted})}{P(\text{dotted})} \\ &= \frac{1/10}{5/10} = \frac{1}{5} = 20.00\% \end{aligned}$$

The relationship,

$$\begin{aligned} &P(\text{striped, given light green}) \\ &+ P(\text{dotted, given light green}) \\ &= \frac{4}{6} + \frac{2}{6} = 1.00 \end{aligned}$$

The relationship,

$$\begin{aligned} &P(\text{striped, given dark green}) \\ &+ P(\text{dotted, given dark green}) \\ &= \frac{3}{4} + \frac{1}{4} = 1.00 \end{aligned}$$

## Bayes' Theorem

The relationship given in equation 3(vi) for conditional probability under statistical dependence is attributed to the Englishman, The Reverend Thomas Bayes (1702–1761) and is also referred to as **Bayesian decision-making**. It illustrates that if you have additional information, or based on the fact that *something has occurred*, certain probabilities may be revised to give *posterior* probabilities (*post* meaning afterwards).

Consider that you are a supporter of Newcastle United Football team. Based on last year's performance you believe that there is a high probability they have a chance of moving to the top of the league this year. However, as the current season moves on Newcastle loses many of the games even on their home turf. In addition, two of their

best players have to withdraw because of injuries. Thus, based on these new events, the probability of Newcastle United moving to the top of the league has to be revised downwards. Take into account another situation where insurance companies have actuary tables for the life expectancy of individuals. Assume that your 18-year-old son is considered for a life insurance. His life expectancy is in the high 70s. However, as time moves on, your son starts smoking heavily. With this new information, your son's life expectancy drops as the risk of contracting life-threatening diseases such as lung cancer increases. Thus, based on this posterior information, the probabilities are again revised downwards.

Thus, if Bayes' rule is correctly used it implies that it maybe unnecessary to collect vast amounts of data over time in order to make the best decisions based on probabilities. Or, another way of looking at Bayes' posterior rule is applying it to the often-used phrase of Hamlet, "*he who hesitates is lost*". The phrase implies that we should quickly make a decision based on the information we have at hand – buy stock in Company A, purchase the house you visited, or take the high-paying job you were offered in Algiers, Algeria.<sup>3</sup> However, if we wait until new information comes along – Company A's financial accounts turns out are inflated, the house you thought about buying turns out is on the path of the construction of a new auto route, or new elections in Algeria make the political situation in the country unstable with a security risk for the population. In these cases, procrastination may be the best approach and, "*he who hesitates comes out ahead*".

## Venn diagram

A **Venn diagram**, named after John Venn an English mathematician (1834–1923), is a useful

way to visually demonstrate the concept of mutually exclusive and non-mutually exclusive events. A surface area such as a circle or rectangle represents an entire sample space, and a particular outcome of an event is represented by part of this surface. If two events, *A* and *B*, are mutually exclusive, their areas will not overlap as shown in Figure 3.4. This is a visual representation for a pack of cards using a rectangle for the surface. Here the number of boxes is 52, which is entire sample space, or 100%. Each card occupies 1 box and when we are considering two cards, the sum of occupied areas is 2 boxes or  $2/52 = 3.85\%$ . If two events, are not mutually exclusive their areas would overlap as shown in Figure 3.5. Here again the number of boxes is 52, which is the entire sample space. Each of the cards, 13 Spades and 4 Aces would normally occupy 1 box or a total of 17 boxes. However, one card is common to both events and so the sum of occupied areas is  $17 - 1$  boxes or  $16/52 = 30.77\%$ .

## Application of a Venn diagram and probability in services: Hospitality management

A business school has in its curriculum a hospitality management programme. This programme covers hotel management, the food industry, tourism, casino operation, and health spa management. The programme includes a specialization in hotel management and tourist management and for these programmes the students spend an additional year of enrolment. In one particular year there are 80 students enrolled in the programme. Of these 80 students, 15 elect to specialize in tourist management, 28 in hotel management, and 5 specializing in both tourist and hotel management. This information is representative of the general profile of the hospitality management programme.

<sup>3</sup>Based on a real situation for the Author in the 1980s.

Figure 3.4 Venn diagram: mutually exclusive events.

			1st Card				2nd Card					

Number of boxes = 52 which is entire sample space = 100%  
 Each card occupies 1 box  
 Sum of occupied areas = 2 boxes or  $2/52 = 3.85\%$

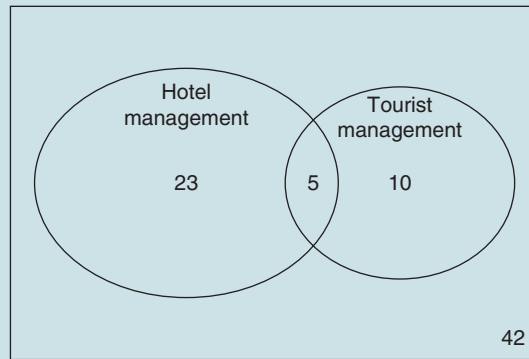
Figure 3.5 Venn diagram: non-mutually exclusive events.

Ace H												
Ace S	2	3	4	5	6	7	8	9	10	J	Q	K
Ace D												
Ace C												

Number of boxes = 52 which is entire sample space  
 Each card would normally occupy 1 box = 17 boxes  
 However, one card is common to both events  
 Sum of occupied areas =  $17 - 1$  boxes or  $16/52 = 30.77\%$



Figure 3.6 Venn diagram for a hospitality management programme.



1. Illustrate this situation on a Venn diagram

The Venn diagram is shown in Figure 3.6. There are  $(23 + 5)$  in hotel management shown in the circle (actually an ellipse) on the left. There are  $(10 + 5)$  in tourist management in the circle on the right. The two circles overlap indicating the 5 students who are specializing in both hotel and tourist management. The rectangle is the total sample space of 80 students, which leaves  $(80 - 23 - 5 - 10) = 42$  students as indicated not specializing.

2. What is the probability that a random selected student is in tourist management?

From the Venn diagram this is the total in tourist management divided by total sample space of 80 students or,

$$P(T) = \frac{5 + 10}{80} = 18.75\%$$

3. What is the probability that a random selected student is in hotel management?

From the Venn diagram this is the total in hotel management divided by total sample space of 80 students or,

$$P(H) = \frac{23 + 5}{80} = 35.00\%$$

4. What is the probability that a random selected student is in hotel or tourist management?

From the Venn diagram this is,

$$P(H \text{ or } T) = \frac{23 + 5 + 10}{80} = 47.50\%$$

This can also be expressed by the counting rule equation 3(iii):

$$\begin{aligned} P(H \text{ or } T) &= P(H) + P(T) - P(HT) \\ &= \frac{28}{80} + \frac{15}{80} - \frac{5}{80} = 47.50\% \end{aligned}$$

5. What is the probability that a random selected student is in hotel and tourist management?

From the Venn diagram this is  $P(\text{both } H \text{ and } T) = 5/80 = 6.25\%$

6. Given a student is specializing in hotel management, what is the probability that a random selected student is also specializing in tourist management?

This is expressed as  $P(T|H)$ , and from the Venn diagram this is  $5/28 = 17.86\%$ .

From equation 3(vi), this is also written as,

$$P(T|H) = \frac{P(TH)}{P(H)} = \frac{5/80}{28/80} = \frac{5}{28} = 17.86\%$$



7. Given a student is specializing in tourist management, what is the probability that a random selected student is also specializing in hotel management?

This is expressed as  $P(H|T)$ , and from the Venn diagram this is  $5/15 = 33.33\%$ .

From equation 3(vi), this is also written as,

$$P(H|T) = \frac{P(HT)}{P(T)} = \frac{5/80}{15/80} = \frac{5}{15} = 33.33\%$$

### Application of probability rules in manufacturing: A bottling machine

On an automatic combined beer bottling and capping machine, two major problems that occur are overfilling and caps not fitting correctly on the bottle top. From past data it is known that 2% of the bottles are overfilled. Further past data shows that if a bottle is overfilled then 25% of the bottles are faulty capped as the pressure differential between the bottle and the capping machine is too low. Even if a bottle is filled correctly, then still 1% of the bottles are not properly capped.

1. What are the four simple events in this situation?

The four simple events are:

- An overfilled bottle
- A normally filled bottle
- An incorrectly capped bottle
- A correctly capped bottle.

2. What are joint events for this situation?

There are four joint events:

- An overfilled bottle and correctly capped
- An overfilled bottle and incorrectly capped
- A normally filled bottle and correctly capped
- A normally filled bottle and incorrectly capped.

3. What is the percentage of bottles that will be faulty capped and thus have to be rejected before final packing?

Here there are two conditions where a bottle is rejected before packing. A bottle overfilled and faulty capped. A bottle normally filled but faulty capped.

- Joint probability of a bottle being overfilled **and** faulty capped is  $0.02 * 0.25 = 0.0050 = 0.5\%$
- Joint probability of a bottle filled normally **and** faulty capped is  $(1 - 0.02) * 0.01 = 0.0098 = 0.98\%$
- By the addition rule, a bottle is faulty capped if it is overfilled and faulty capped **or** normally filled and faulty capped =  $0.0050 + 0.0098 = 0.0148 = 1.48\%$  of the time.

4. If the analysis were made looking at a sample of 10,000 bottles, how would this information appear in a cross-classification table?

The cross-classification table is shown in Table 3.6.

This is developed as follows.

- Sample size is 10,000 bottles
- There are 2% bottles overfilled or  $10,000 * 2\% = 200$
- There are 98% of bottles filled correctly or  $10,000 * 98\% = 9,800$
- Of the bottles overfilled, 25% are faulty capped or  $200 * 25\% = 50$
- Thus bottles overfilled but correctly capped is  $200 - 50 = 150$
- Bottles filled correctly but 1% are faulty capped or  $9,800 * 1\% = 98$
- Thus filled correctly and correctly capped is  $9,800 - 98 = 9,702$
- Thus, bottles correctly capped is  $9,702 + 150 = 9,852$
- Thus, all bottles incorrectly capped is  $10,000 - 9,852 = 148 = 1.48\%$ .

### Gambling, odds, and probability

Up to this point in the chapter you might argue that much of the previous analysis is related to gambling and then you might say, “*but the business*

**Table 3.6** Cross-classification table for bottling machine.

Volume	Capping		Total
	Number that fit	Number that does not fit	
Right amount	9,702	98	9,800
Overfilled	150	50	200
<b>Total</b>	<b>9,852</b>	<b>148</b>	<b>10,000</b>

*world is not just gambling*". That is true but do not put gambling aside. Our capitalistic society is based on risk, and as a corollary, gambling, as is indicated by the Box Opener. We are confronted daily with gambling through government organized lotteries, buying and selling stock, and gambling casinos. This service-related activity represents a non-negligible part of our economy!

In risk, gambling, or betting we refer to the **odds** of wining. Although the odds are related to probability they are a way of looking at risk. The probability is the number of favourable outcomes divided by the total number of possible outcomes. The odds of winning are the ratio of the chances of losing to the chances of winning. Earlier we illustrated that the probability of obtaining the number 7 in the tossing of two dice was 6 out of 36 throws, or 1 out of 6. Thus the probability of not obtaining the number 7 is 30 out of 36 throws or 5 out of 6. Thus the odds of obtaining the number 7 are 5 to 1.

This can be expressed mathematically as,

$$\frac{5/6}{1/6} = \frac{5}{1}$$

The odds of drawing the Ace of Spades from a full pack of cards are 51 to 1. Although the odds depend on probability, it is the odds that matter when you are placing a bet or taking a risk!

## System Reliability and Probability

Probability concepts as we have just discussed can be used to evaluate system **reliability**. A system includes all the interacting components or activities needed for arriving at an end result or product. In the system the reliability is the confidence that we have in a product, process, service, work team, or individual, such that we can operate under prescribed conditions without failure, or stopping, in order to produce the required output. In the supply chain of a firm for example, reliability might be applied to whether the trucks delivering raw materials arrive on time, whether the suppliers produce quality components, whether the operators turn up for work, or whether the packing machines operate without breaking down. Generally, the more components or activities in a product or a process, then the more complex is the system and in this case the greater is the risk of failure, or **unreliability**.

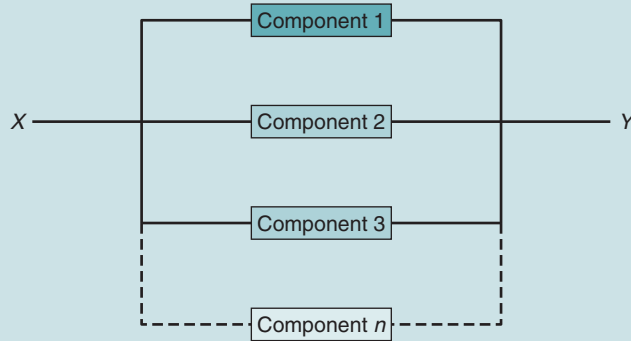
### Series or parallel arrangement

A product or a process might be organized in a **series arrangement** or **parallel arrangement** as illustrated schematically in Figure 3.7. This is a general structure, which contains  $n$  components in the case of a product, or  $n$  activities for processes. The value  $n$  can take on any integer value. The upper scheme shows a purely series arrangement and the lower a parallel arrangement. Alternatively a system may be a combination of both series and parallel arrangements.

### Series systems

In the series arrangement, shown in the upper diagram of Figure 3.7, it means that for a system to operate we have to pass in sequence through Component 1, Component 2, Component 3, and eventually to Component  $n$ .

Figure 3.7 Reliability: Series and parallel systems.

**System connected in series****System connected in parallel (backup)**

For example, when an electric heater is operating the electrical current comes from the main power supply (Component 1), through a cable (Component 2), to a resistor (Component 3), from which heat is generated.

The **reliability of a series system**,  $R_S$ , is the joint probability of the number of interacting components,  $n$ , according to the following relationship:

$$R_S = R_1 * R_2 * R_3 * R_4 * \dots R_n \quad 3(vii)$$

Here  $R_1$ ,  $R_2$ ,  $R_3$ , etc. represent the reliability of the individual components expressed as a fraction or percentage. The relationship in equation 3(vii) assumes that each component is independent of the other and that the reliability of one does not depend on the reliability of the other. In the electric heater example, the main power supply, the electric cable, and the resistor are all independent of each other. However, the complete electric heating system does depend on all the

components functioning, or in the system they are interdependent. If one component fails, then the system fails. For the electric heater, if the power supply fails, or the cable is cut, or the resistor is broken then the heater will not function. The reliability, or the value of  $R$ , will be  $<100\%$  (nothing is perfect) and may have a value of say 99%. This means that a component will perform as specified 99% of the time, or it will fail 1% of the time ( $100 - 99$ ). This is a binomial relationship since the component either works or it does not. Binomial means there are only two possible outcomes such as yes or no, true or false.

Consider the system between point X and Y in the series scheme of Figure 3.7 with three components. Assume that component  $R_1$  has a reliability of 99%,  $R_2$  a reliability of 98%, and  $R_3$  a reliability of 97%. The system reliability is then:

$$\begin{aligned} R_S &= R_1 * R_2 * R_3 = 0.99 * 0.98 * 0.97 \\ &= 0.9411 = 94.11\% \end{aligned}$$

Table 3.7 System reliability for a series arrangement.

Number of components	1	3	5	10	25	50	100	200
System reliability (%)	98.00	94.12	90.39	81.71	60.35	36.42	13.26	1.76

In a situation where the components have the same reliability then the system reliability is given by the following general equation, where  $n$  is the number of components

$$R_S = R^n \quad 3(\text{viii})$$

Note that as already mentioned for joint probability, the system reliability  $R_S$  is always less than the reliability of the individual components. Further, the reliability of the system, in a series arrangement of multiple components, decreases rapidly with the number of components. For example assume that we have a system where the average reliability of each component is 98%, then as shown in Table 3.7 the system reliability drops from 94.12% for three components to 1.76% for 200 components. Further, to give a more complete picture, Figure 3.8 gives a family of curves showing the system reliability, for various values of the individual component reliability from 100% to 95%. These curves illustrate the rapid decline in the system reliability as the number of components increases.

## Parallel or backup systems

The parallel arrangement is illustrated in the lower diagram of Figure 3.6. This illustrates that in order for equipment to operate we can pass through Component 1, Component 2, Component 3, or eventually Component  $n$ .

Assume that we have two components in a parallel system,  $R_1$  the main component and  $R_2$  the backup or auxiliary component. The reliability of a parallel system,  $R_S$ , is then given by the relationship,

$$R_S = \text{Probability of } R_1 \text{ working} + \text{Probability of } R_2 \text{ working} * \text{Probability of needing } R_2$$

The probability of needing  $R_2$  is when  $R_1$  is not working or  $(1 - R_1)$ . Thus,

$$R_S = R_1 + R_2(1 - R_1) \quad 3(\text{ix})$$

Reorganizing equation 3(ix)

$$R_S = R_1 + R_2 - R_2 * R_1$$

$$R_S = 1 + R_1 + R_2 - R_2 * R_1 - 1$$

$$R_S = 1 - (1 - R_1 - R_2 + R_2 * R_1)$$

$$R_S = 1 - (1 - R_1)(1 - R_2) \quad 3(\text{x})$$

If there are  $n$  components in a parallel arrangement then the system reliability becomes

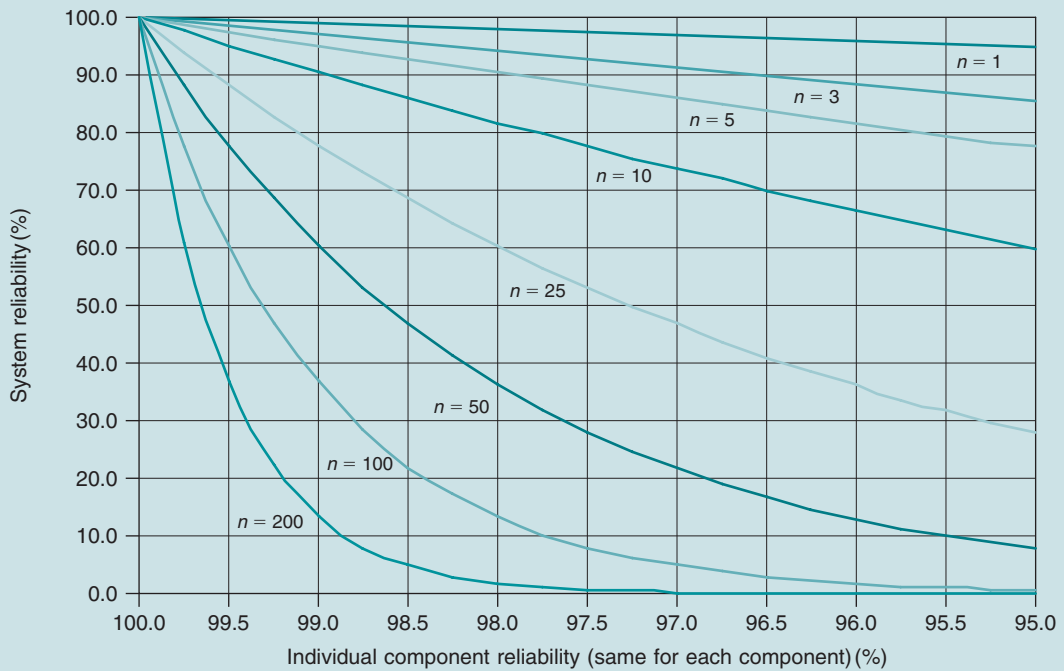
$$R_S = 1 - (1 - R_1)(1 - R_2)(1 - R_3) \dots (1 - R_n) \quad 3(\text{xi})$$

where  $R_1, R_2, \dots, R_n$  represent the reliability of the individual components. The equation can be interpreted as saying that the more the number of backup units, then the greater is the system reliability. However, this increase of reliability comes at an increased cost since we are adding backup which may not be used for any length of time.

When the backup components of quantity,  $n$ , have an equal reliability, then the system reliability is given by the relationship,

$$R_S = 1 - (1 - R)^n \quad 3(\text{xii})$$

Consider the three component system in the lower scheme of Figure 3.7 between point X and Y with the principal component  $R_1$  having a

Figure 3.8 System reliability in series according to number of components,  $n$ .

reliability of 99%,  $R_2$  the first backup component having a reliability of 98%, and  $R_3$  the second backup component having a reliability of 97% (the same values as used in the series arrangement). The system reliability is then from equation 3(xi),

$$R_S = 1 - (1 - R_1)(1 - R_2)(1 - R_3)$$

$$R_S = 1 - (1 - 0.99)(1 - 0.98)(1 - 0.97)$$

$$R_S = 1 - 0.000006 = 0.999994 = 99.994\%$$

That is, a system reliability greater than with using a single generator.

If we only had the first backup unit,  $R_2$  then the system reliability is,

$$R_S = 1 - (1 - R_1)(1 - R_2)$$

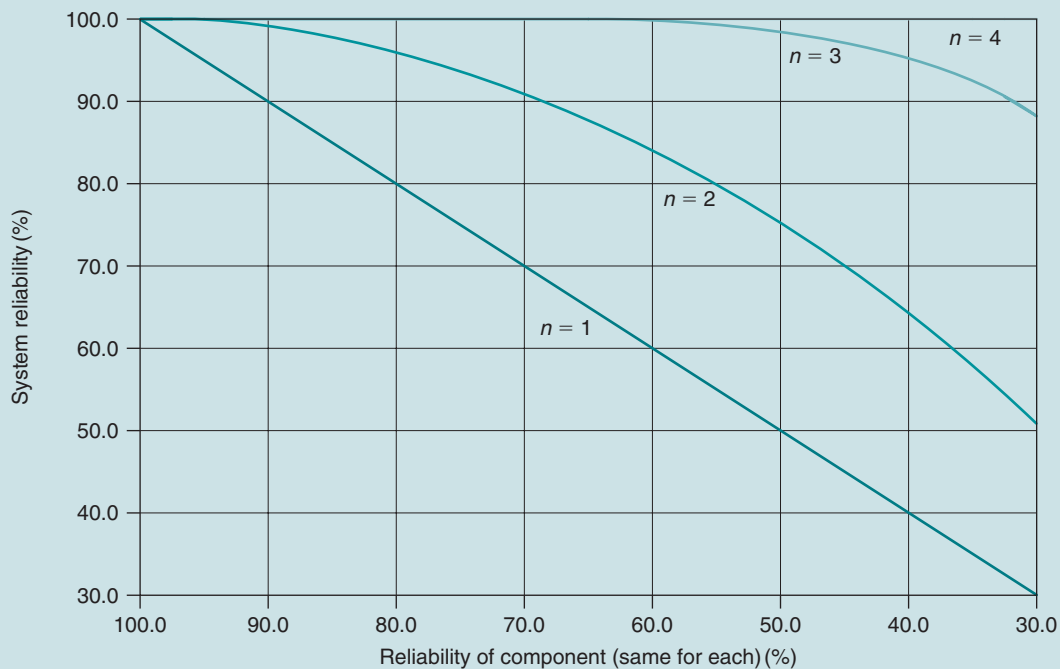
$$R_S = 1 - (1 - 0.99)(1 - 0.98)$$

$$R_S = 1 - 0.01 * 0.02$$

$$R_S = 1 - 0.0002 = 0.9998 = 99.98\%$$

Again, this is a reliability greater than the reliability of the individual components. Since the components are in parallel they are called backup units. The more the number of backup units, then the greater is the system reliability as illustrated in Figure 3.9. Here the curves give the reliability with no backups ( $n = 1$ ) to three backup components ( $n = 4$ ). Of course, ideally, we would always want close to 100% reliability, however, with greater reliability, the greater is the cost.

Hospitals have back up energy systems in case of failure of the principal power supply. Most banks and other firms have backup computer systems containing client data should one system

Figure 3.9 System reliability of a parallel or backup system according to number of components,  $n$ .

fail. The IKEA distribution platform in South-eastern France has a backup computer in case its main computer malfunctions. Without such a system, IKEA would be unable to organize delivery of its products to its retail stores in France, Spain, and Portugal.<sup>4</sup> Aeroplanes have backup units in their design such that in the eventual failure of one component or subsystem there is recourse to a backup. For example a Boeing 747 can fly on one engine, although at a much reduced efficiency. To a certain extent the human body has a backup system as it can function with only one lung though again at a reduced efficiency. In August 2004, my wife

<sup>4</sup>After a visit to the IKEA distribution platform in St. Quentin Falavier, Near Lyon, France, 18 November 2005.

and I were in a motor home in St. Petersburg Florida when hurricane Charlie was about to land. We were told of four possible escape routes to get out of the path. The emergency services had designated several backup exit routes – thankfully! When backup systems are in place this implies redundancy since the backup units are not normally operational.

The following is an application example of mixed series and parallel systems.

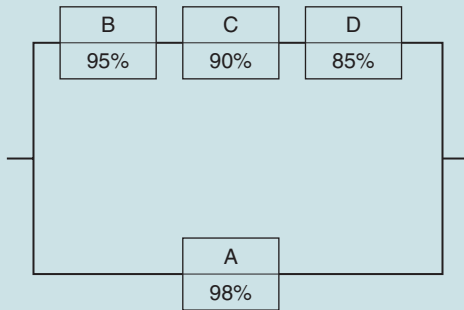
### Application of series and parallel systems: Assembly operation

In an assembly operation of a certain product there are four components A, B, C, and D that have an individual reliability of 98%, 95%, 90%, and 85%, respectively. The possible ways

Figure 3.10 Assembly operation:  
Arrangement No. 1.



Figure 3.11 Assembly operation:  
Arrangement No. 2.



of assembly the four components are given in Figures 3.10–3.13. Determine the system reliability of the four arrangements.

#### Arrangement No. 1

Here this is completely a series arrangement and the system reliability is given by the joint probability of the individual reliabilities:

- Reliability is  $0.98 * 0.95 * 0.90 * 0.85 = 0.7122 = 71.22\%$ .
- Probability of system failure is  $(1 - 0.7122) = 0.2878 = 28.78\%$ .

#### Arrangement No. 2

Here this is a series arrangement in the top row in parallel with an assembly in the bottom row. The system reliability is calculated by first the

Figure 3.12 Assembly operation:  
Arrangement No. 3.

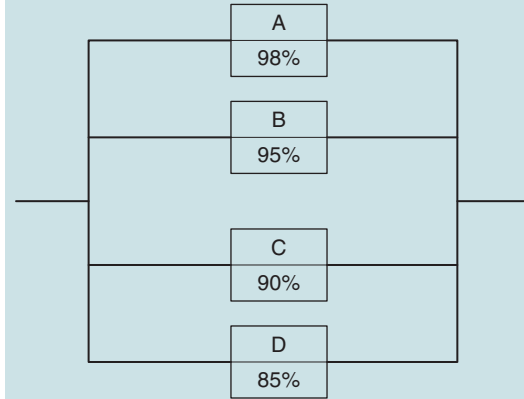
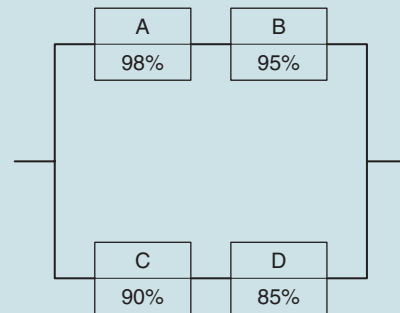


Figure 3.13 Assembly operation:  
Arrangement No. 4.



joint probability of the individual reliabilities in the top row, in parallel with the reliability in the second row.

- Reliability of top row is  $0.95 * 0.90 * 0.85 = 0.7268 = 72.68\%$ .
- Reliability of system is  $1 - (1 - 0.7268) * (1 - 0.9800) = 0.9945 = 99.45\%$ .
- Probability of system failure is  $(1 - 0.9945) = 0.0055 = 0.55\%$ .



Table 3.8 Possible outcomes of the tossing of a coin 8 times.

Outcome	1	2	3	4	5	6	7	8
First toss	Heads	Heads	Heads	Tails	Tails	Tails	Tails	Heads
Second toss	Heads	Heads	Tails	Heads	Tails	Tails	Heads	Tails
Third toss	Heads	Tails	Heads	Heads	Tails	Heads	Tails	Tails

### Arrangement No. 3

Here we have four units in parallel and thus the system reliability is,

- $1 - (1 - 0.9800) * (1 - 0.9500) * (1 - 0.9000) * (1 - 0.8500) = 0.999985 = 99.9985\%$ .
- Probability of system failure is  $(1 - 0.999985) = 0.000015 = 0.0015\%$ .

### Arrangement No. 4

Here we have two units each in series and then the combination in parallel.

- Joint reliability of top row is  $0.98 * 0.95 = 0.9310 = 93.10\%$ .
- Joint reliability of bottom row is  $0.90 * 0.85 = 0.7650 = 76.50\%$ .
- Reliability of system is  $1 - (1 - 0.9310) * (1 - 0.7650) = 0.9983 = 98.38\%$ .
- Probability of system failure is  $(1 - 0.9838) = 0.0162 = 1.62\%$ .

In summary, when systems are connected in parallel, the reliability is the highest and the probability of system failure is the lowest.

## Counting Rules

Counting rules are the mathematical relationships that describe the possible outcomes, or results, of various types of experiments, or trials. The **counting rules** are in a way *a priori* since you have the required information before

you perform the analysis. However, there is no probability involved. The usefulness of counting rules is that they can give you a precise answer to many basic design or analytical situations. The following gives five different counting rules.

### A single type of event: Rule No. 1

If the number of events is  $k$ , and the number of trials, or experiments is  $n$ , then the total possible **outcomes of single types events** are given by  $k^n$ . Suppose for example that a coin is tossed 3 times. Then the number of trials,  $n$ , is 3 and the number of events,  $k$ , is 2 since only heads or tails are the two possible events. The events, obtaining heads or tails are mutually exclusive since you can only have heads or tails in one throw of a coin. The **collectively exhaustive** outcome is  $2^3$ , or 8. In Excel we use **[function POWER]** to calculate the result.

Table 3.8 gives the possible outcomes of the coin toss experiment. For example as shown for throw No. 1 in the three tosses of the coin, heads could be obtained each time. Alternatively as shown for throw No. 6 the first two tosses could be tails, and then the third heads.

In tossing a coin just 3 times it is impossible to say what will be the possible outcomes. However, if there are many tosses say a 1,000 times, we can reasonably estimate that we will obtain approximately 500 heads and 500 tails. That is, the larger the number of trials, or experiments, the closer the result will be to the **characteristic probability**. In this case the characteristic probability,  $P(x)$  is 50% since there is



Table 3.9 Possible outcomes of the tossing of two dice.

Throw No.	1	2	3	4	5	6	7	8	9	10	11	12
1st die	1	2	3	4	5	6	1	2	3	4	5	6
2nd die	1	1	1	1	1	1	2	2	2	2	2	2
Total of both dice	2	3	4	5	6	7	3	4	5	6	7	8
Throw No.	13	14	15	16	17	18	19	20	21	22	23	24
1st die	1	2	3	4	5	6	1	2	3	4	5	6
2nd die	3	3	3	3	3	3	4	4	4	4	4	4
Total of both dice	4	5	5	7	8	9	5	6	7	8	9	10
Throw No.	25	26	27	28	29	30	31	32	33	34	35	36
1st die	1	2	3	4	5	6	1	2	3	4	5	6
2nd die	5	5	5	5	5	5	6	6	6	6	6	6
Total of both dice	6	7	8	9	10	11	7	8	9	10	11	12

an equal chance of obtaining either heads or tails. Thus the outcome is  $n * P(x)$  or  $1,000 * 50\% = 500$ . This idea is further elaborated in the law of averages in Chapter 4.

### Different types of events: Rule No. 2

If there are  $k_1$  possible events on the 1st trial or experiment,  $k_2$  possible events on the 2nd trial,  $k_3$  possible events on the 3rd trial, and  $k_n$  possible events on the  $n$ th trial, then the total possible **outcomes of different events** are calculated by the following relationship:

$$k_1 * k_2 * k_3 \dots k_n \quad 3(\text{xiii})$$

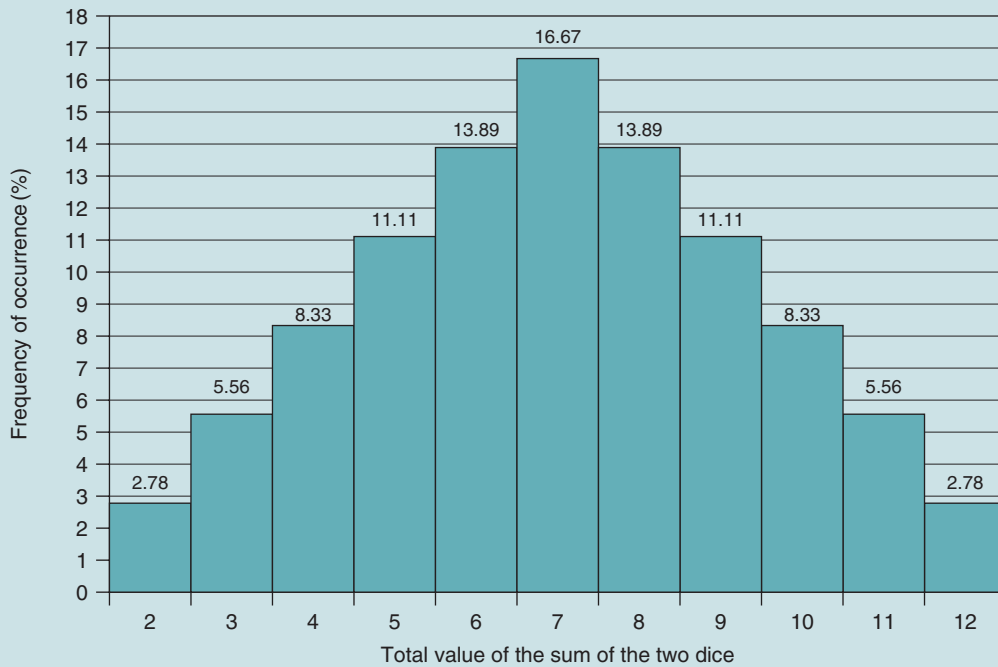
Suppose in gambling, two dice are used. The possible events from throwing the first die are six since we could obtain the number 1, 2, 3, 4, 5, or 6. Similarly, the possible events from throwing the second die are also six. Then the total possible different outcomes are  $6 * 6$  or 36. Table 3.9 gives the 36 possible combinations.

The relative frequency histogram of all the possible outcomes is shown in Figure 3.14.

Note, that the number 7 has the highest possibility of occurring at 6 times or a probability of 16.67% ( $6/36$ ). This is the same value we found in the previous section on joint probabilities.

Consider another example to determine the total different licence plate registrations that a country or community can possibly issue. Assume that the format for a licence plate is 212TPV. (This was the licence plate number of my first car, an Austin A40, in England, that I owned as a student in the 1960s the time of the Beatles – la belle époque!) In this format there are three numbers, followed by three letters. For numbers, there are 10 possible outcomes, the numbers from 0 to 9. For letters, there are 26 possible outcomes, the letters A to Z. Thus the first digit of the licence plate can be the number 0 to 9, the same for the second, and the third. Similarly, the first letter can be any letter from A to Z, the same for the second letter, and the same for the third. Thus the total possible different combinations, or the number of licence plates is 17,566,000 on the assumption that 0 is possible in the first place

Figure 3.14 Frequency histogram of the outcomes of throwing two dice.



$$10 * 10 * 10 * 26 * 26 * 26 = 17,576,000$$

If zero is not permitted in the first place, then the number possible is 15,818,000

$$9 * 10 * 10 * 26 * 26 * 26 = 15,818,000$$

### Arrangement of different objects: Rule No. 3

In order to determine the number of ways that we can arrange  $n$  objects is  $n!$ , or  $n$  factorial, where,

$$n! = n(n-1)(n-2)(n-3) \dots 1 \quad 3(\text{xiv})$$

This is the **factorial rule**. Note, the last term in equation 3(xiv) is really  $(n - n)$  or 0, but in the factorial relationship,  $0! = 1$ .

For example, the number of ways that the three colours, red, yellow, and blue can be arranged is,

Table 3.10 Possible arrangement of three different colours.

1	2	3	4	5	6
Red	Red	Yellow	Yellow	Blue	Blue
Yellow	Blue	Blue	Red	Red	Yellow
Blue	Yellow	Red	Blue	Yellow	Red

$$3! = 3 * 2 * 1 = 6$$

Table 3.10 gives these six possible arrangements. In Excel we use **[function FACT]** to calculate the result.

### Permutations of objects: Rule No. 4

A **permutation** is a combination of data arranged in a particular order. The number

Table 3.11 Permutations in organizing an operating committee.

Choice	1	2	3	4	5	6	7	8	9	10	11	12
President	Dan	Dan	Dan	Sue	Sue	Sue	Jim	Jim	Jim	Ann	Ann	Ann
Secretary	Sue	Jim	Ann	Jim	Ann	Dan	Ann	Dan	Sue	Dan	Sue	Jim

of ways, or permutations, of arranging  $x$  objects selected in order from a total of  $n$  objects is,

$${}^n P_x = \frac{n!}{(n-x)!} \quad 3(xv)$$

Suppose there are four candidates Dan, Sue, Jim, and Ann, who have volunteered to work on an operating committee: the number of ways a president and secretary can be chosen is by equation 3(xv),

$${}^4 P_2 = \frac{4!}{(4-2)!} = 12$$

In Excel we use [\[function PERMUT\]](#) to calculate the result. Table 3.11 gives the various permutations. Here the same two people can be together, providing they have different positions. For example in the 1st choice, Dan is the president and Sue is the secretary. In the 6th choice their positions are reversed. Sue is the president and Dan is the secretary.

## Combinations of objects: Rule No. 5

A **combination** is a selection of distinct items regardless of order. The number of ways, or combinations, of arranging  $x$  objects, regardless of order, from  $n$  objects is given by,

Table 3.12 Combinations for organizing an operating committee.

Choice	1	2	3	4	5	6
President	Dan	Dan	Dan	Sue	Sue	Jim
Vice president	Sue	Jim	Ann	Jim	Ann	Ann

$${}^n C_x = \frac{n!}{x!(n-x)!} \quad 3(xvi)$$

Again, assume that there are four candidates for two positions in an operating committee: Dan, Sue, Jim, and Ann. The number of ways a president and secretary can be chosen now without the same two people working together, regardless of position is by equation 3(xvi)

$${}^4 C_2 = \frac{4!}{2!(4-2)!} = 6$$

Table 3.12 gives the combinations. In Excel we can use [\[function COMBIN\]](#) to directly calculate the result.

Note that the Rule No. 4, permutations, differs from Rule No. 5, combinations, by the value of  $x!$  in the denominator. For a given set of items the number of permutations will always be more than the number of combinations because with permutations the order of the data is important, whereas it is unimportant for combinations.

This chapter has introduced rules governing basic probability and then applied these to reliability of system design. The last part of the chapter has dealt with mathematical counting rules.

### Basic probability rules

Probability is the chance that something happens, or does not happen. An extension of probability is risk, where we can put a monetary value on the outcome of a particular action. In probability we talk about an event, which is the outcome of an experiment that has been undertaken. Probability may be subjective and this is the “gut” feeling or emotional response of the individual making the judgment. Relative frequency probability is derived from collected data and is thus also called empirical probability. A third is classical or marginal probability, which is the ratio of the number of desired outcomes to the total number of possible outcomes. Classical probability is also *a priori* probability because before any action occurs we know in advance all possible outcomes. Gambling involving dice, cards, or roulette wheels are examples of classical probability since before playing we know in advance that there are six faces on a die, 52 cards in a pack. (We do not know in advance the number of slots on the roulette wheel – but the casino does!). Within classical probability, the addition rule gives the chance that two or more events occur, which can be modified to avoid double accounting. To determine the probability of two or more events occurring together, or in succession, we use joint probability. When one event has already occurred then this gives *posterior* probability meaning the new chance based on the condition that another event has already happened. Posterior probability is Bayes’ Theorem. To visually demonstrate relationships in classical probabilities we can use Venn diagrams where a surface area, such as a circle, represents an entire sample space, and a particular outcome of an event is shown by part of this surface. In gambling, particularly in horse racing, we refer to the odds of something happening. Odds are related to probability but odds are the ratio of the chances of losing to the chances of winning.

### System reliability and probability

A system is a combination of components in a product or many of the process activities that makes a business function. We often refer to the system reliability, which is the confidence that we have in the product or process operating under prescribed conditions without failure. If a system is made up of series components then we must rely on all these series components working. If one component fails, then the system fails. To determine the system reliability, or system failure, we use joint probability. When the probability of failure, even though small, can be catastrophic such as for an airplane in flight, the power system in a hospital, or a bank’s computer-based information system, components are connected in parallel. This gives a backup to the system. The probability of failure of parallel systems is always less than the probability of failure for series systems for given individual component probabilities. However, on the downside, the cost is always higher for a parallel arrangement since we have a backup that (we hope) will hardly, or never, be used.

## Counting rules

Counting rules do not involve probabilities. However, they are a sort of *a priori* conditions, as we know in advance, with given criteria, exactly the number of combinations, arrangements, or outcomes that are possible. The first rule is that for a fixed number of possible events,  $k$ , then for an experiment with a sample of size,  $n$ , the possible arrangements is given by  $k^n$ . If we throw a single die 4 times then the possible arrangements are  $6^4$  or 1,296. The second rule is if we have events of different types say  $k_1$ ,  $k_2$ ,  $k_3$  and  $k_4$  then the possible arrangements are  $k_1 * k_2 * k_3 * k_4$ . This rule will indicate, for example, the number of licence plate combinations that are possible when using a mix of numbers and letters. The third rule uses the factorial relationship,  $n!$  for the number of different ways of organizing  $n$  objects. The fourth and fifth rules are permutations and combinations, respectively. Permutations gives the number of possible ways of organizing  $x$  objects from a sample of  $n$  when the order is important. Combinations determine the number of ways of organizing  $x$  objects from a sample of  $n$  when the order is irrelevant. For given values of  $n$  and  $x$  the value using permutations is always higher than for combinations.

## EXERCISE PROBLEMS

### 1. Gardeners' gloves

#### Situation

A landscape gardener employs several students to help him with his work. One morning they come to work and take their gloves from a communal box. This box contains only five left-handed gloves and eight right-handed gloves.

#### Required

1. If two gloves are selected at random from the box, without replacement, what is the probability that both gloves selected will be right handed?
2. If two gloves are selected at random from the box, without replacement, what is the probability that a pair of gloves will be selected? (One glove is right handed and one glove is left handed.)
3. If three gloves are selected at random from the box, with replacement, what is the probability that all three are left handed?
4. If two gloves are selected at random from the box, with replacement, what is the probability that both gloves selected will be right handed?
5. If two gloves are selected at random from the box, with replacement, what is the probability that a correct pair of gloves will be selected?

### 2. Market Survey

#### Situation

A business publication in Europe does a survey of some of its readers and classifies the survey responses according to the person's country of origin and their type of work. This information according to the number of respondents is given in the following contingency table.

Country	Consultancy	Engineering	Investment banking	Product marketing	Architecture
Denmark	852	232	541	452	385
France	254	365	842	865	974
Spain	865	751	695	358	845
Italy	458	759	654	587	698
Germany	598	768	258	698	568

#### Required

1. What is the probability that a survey response taken at random comes from a reader in Italy?

2. What is the probability that a survey response taken at random comes from a reader in Italy and who is working in engineering?
3. What is the probability that a survey response taken at random comes from a reader who works in consultancy?
4. What is the probability that a survey response taken at random comes from a reader who works in consultancy and is from Germany?
5. What is the probability that a survey response taken at random from those who work in investment banking comes from a reader who lives in France?
6. What is the probability that a survey response taken at random from those who live in France is working in investment banking?
7. What is the probability that a survey response taken at random from those who live in France is working in engineering or architecture?

### 3. Getting to work

#### Situation

George is an engineer in a design company. When the weather is nice he walks to work and sometimes he cycles. In bad weather he takes the bus or he drives. Based on past habits there is a 10% probability that George walks, 30% he uses his bike, 20% he drives, and 40% of the time he takes the bus. If George walks, there is a 15% probability of being late to the office, if he cycles there is a 10% chance of being late, a 55% chance of being late if he drives, and a 20% chance of being late if he takes the bus.

1. On any given day, what is the probability of George being late to work?
2. Given that George is late 1 day, what is the probability that he drove?
3. Given that George is on time for work 1 day, what is the probability that he walked?
4. Given that George takes the bus 1 day, what is the probability that he will arrive on time?
5. Given that George walks to work 1 day, what is the probability that he will arrive on time?

### 4. Packing machines

#### Situation

Four packing machines used for putting automobile components in plastics packs operate independently of one another. The utilization of the four machines is given below.

Packing machine A	Packing machine B	Packing machine C	Packing machine D
30.00%	45.00%	80.00%	75.00%

**Required**

1. What is the probability at any instant that both packing machine A and B are not being used?
2. What is the probability at any instant that all machines will be idle?
3. What is the probability at any instant that all machines will be operating?
4. What is the probability at any instant of packing machine A and C being used, and packing machines B and D being idle?

**5. Study Groups****Situation**

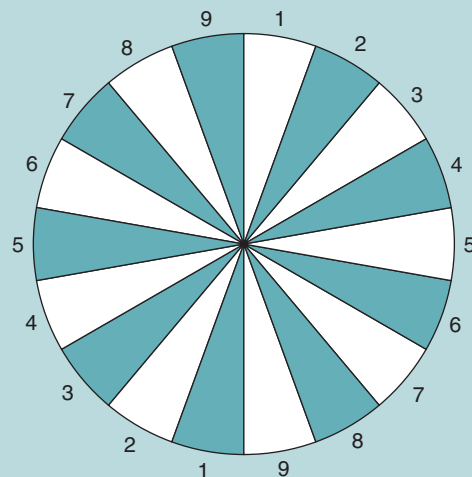
In an MBA programme there are three study groups each of four people. One study group has three ladies and one man. One has two ladies and two men and the third has one lady and three men.

**Required**

1. One person is selected at random from each of the three groups in order to make a presentation in front of the class. What is the probability that this presentation group will be composed of one lady and two men?

**6. Roulette****Situation**

A hotel has in its complex a gambling casino. In the casino the roulette wheel has the following configuration.





There are two games that can be played:

#### Game No. 1

Here a player bets on any single number. If this number turns up then the player gets back 7 times the bet. There is always only one ball in play on the roulette wheel.

#### Game No. 2

Here a player bets on a simple chance such as the colours white or dark green, or an odd or even number. If this chance occurs then the player doubles his/her bet. If the number 5 turns up, then all players lose their bets. There is always only one ball in play on the roulette wheel.

#### Required

1. In Game No. 1 a player places £25 on number 3. What is the probability of the player receiving back £175? What is the probability that the player loses his/her bet?
2. In Game No. 1 a player places £25 on number 3 and £25 on number 4. What is the probability of the player winning? What is the probability that the player loses his/her bet? If the player wins how much money will he/she win?
3. In Game No. 1 if a player places £25 on each of several different numbers, then what is the maximum numbers on which he/she should bet in order to have a chance of winning? What is this probability of winning? In this case, if the player wins how much will he/she win? What is the probability that the player loses his entire bet? How much would be lost?
4. In Game No. 2 a player places £25 on the colour dark green. What is the probability of the player doubling the bet? What is the probability of the player losing his/her bet?
5. In Game No. 2 a player places £25 on obtaining the colour dark green and also £25 on obtaining the colour white. In this case what is the probability a player will win some money? What is the probability of the player losing both bets?
6. In Game No. 2 a player places £25 on an even number. What is the probability of the player doubling the bet? What is the probability of the player losing his/her bet?
7. In Game No. 2 a player places £25 on an odd number. What is the probability of the player doubling the bet? What is the probability of the player losing his/her bet?

## 7. Sourcing agents

#### Situation

A large international retailer has sourcing agents worldwide to search out suppliers of products according the best quality/price ratio for products that it sells in its stores in the United States. The retailer has a total of 131 sourcing agents internationally. Of these 51 specialize in textiles, 32 in footwear, and 17 in both textiles and footwear. The remainder are general sourcing agents with no particular specialization. All the sourcing agents are in a general database with a common E-mail address. When a purchasing manager from any of the retail stores needs information on its sourced products they send an E-mail to the general database address. Anyone of the 131 sourcing agents is able to respond to the E-mail.

1. Illustrate the category of the specialization of the sourcing agents on a Venn diagram.
2. What is the probability that at any time an E-mail is sent it will be received by a sourcing agent specializing in textiles?
3. What is the probability that at any time an E-mail is sent it will be received by a sourcing agent specializing in both textiles and footwear?
4. What is the probability that at any time an E-mail is sent it will be received by a sourcing agent with no specialty?
5. Given that the E-mail is received by a sourcing agent specializing in textiles what is the probability that the agent also has a specialty in footwear?
6. Given that the E-mail is received by a sourcing agent specializing in footwear what is the probability that the agent also has a specialty in textiles?

## 8. Subassemblies

### Situation

A subassembly is made up of three components A, B, and C. A large batch of these units are supplied to the production site and the proportion of defective units in these is 5% of the component A, 10% of the component B, and 4% of the component C.

### Required

1. What proportion of the finished subassemblies will contain no defective components?
2. What proportion of the finished subassemblies will contain exactly one defective component?
3. What proportion of the finished subassemblies will contain at least one defective component?
4. What proportion of the finished subassemblies will contain more than one defective component?
5. What proportion of the finished subassemblies will contain all three defective components?

## 9. Workshop

### Situation

In a workshop there are the four operating posts with their average utilization as given in the following table. Each operating post is independent of the other.

Operating post	Utilization (%)
Drilling	50
Lathe	40
Milling	70
Grinding	80

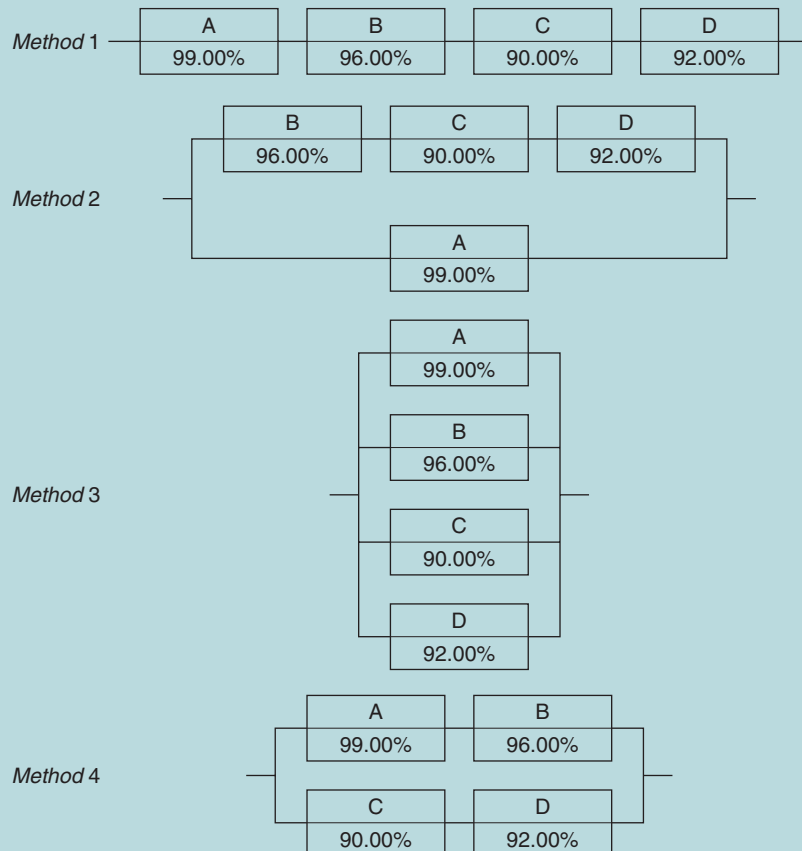
### Required

1. What is the probability of both the drilling and lathe work post not being used at any time?
2. What is the probability of all work posts being idle?
3. What is the probability of all the work posts operating?
4. What is the probability of the drilling and the lathe work post operating and the milling and grinding not operating?

## 10. Assembly

### Situation

In an assembly operation of a certain product there are four components A, B, C, and D which have an individual reliability of 98%, 95%, 90%, and 85%, respectively. The possible ways of assembly the four components making certain adjustments, are as follows.



**Required**

1. Determine the system reliability of each of the four possible ways of assembling the components.
2. Determine the probability of system failure for each of the four schemes.

**11. Bicycle gears****Situation**

The speeds on a bicycle are determined by a combination of the number of sprocket wheels on the pedal sprocket and the rear wheel sprocket. The sprockets are toothed wheels over which the bicycle chain is engaged and the combination is operated by a derailleur system. To change gears you move a lever, or turn a control on the handlebars, which derails the chain onto another sprocket. A bicycle manufacturer assembles customer made bicycles according to the number of speeds desired by clients.

**Required**

1. Using the counting rules, complete the following table regarding the number of sprockets and the number of gears available on certain options of bicycles.

Bicycle model	Pedal sprocket	Rear wheel sprocket	Number of gears
A	1	1	
B		2	2
C	2		6
D	2	4	
E		5	10
F	3		12
G	3	7	
H		7	28
I	4		32
J	4	9	

**12. Film festival****Situation**

The city of Cannes in France is planning its next film festival. The festival will last 5 days and there will be seven films shown each day. The festival committee has selected the 35 films which they plan to show.

**Required**

1. How many different ways can the festival committee organize the films on the first day?

2. If the order of showing is important, how many different ways can the committee organize the showing of their films on the first day? (Often the order of showing films is important as it can have an impact on the voting results.)
3. How many different ways can the festival committee organize the films on (a) the second, (b) the third, (c) the fourth, (d) and the fifth and last day?
4. With the conditions according the Question No. 3, and again the order of showing the films is important, how many different ways are possible on (a) the second, (b) the third, (c) the fourth, (d) and the fifth and last day?

### 13. Flag flying

#### Situation

The Hilton Hotel Corporation has just built two large new hotels, one in London, England and the other in New York, United States. The hotel manager wants to fly appropriate flags in front of the hotel main entrance.

#### Required

1. If the hotel in London wants to fly the flag of every members of the European Union, how many possible ways can the hotel organize the flags?
2. If the hotel in London wants to fly the flag of 10 members of the European Union, how many possible ways can the flags be organized assuming that the hotel will consider all the flags of members of the European Union?
3. If the hotel in London wants to fly the flag of just five members of the European Union, how many possible ways can the flags be organized assuming that the hotel will consider all the flags of members of the European Union?
4. If the hotel in New York wants to fly the flag of all of the states of the United States how many possible ways can the flags be organized?
5. If the hotel in New York wants to fly the flag of all of the states of the United States in alphabetical order by state how many possible ways can the flags be organized?

### 14. Model agency

#### Situation

A dress designer has 21 evening gowns, which he would like to present at a fashion show. However at the fashion show there are only 15 suitable models to present the dresses and the designer is told that the models can only present one dress, as time does not permit the presentation of more than 15 designs.

#### Required

1. How many different ways can the 21 different dress designs be presented by the 15 models?

2. Once the 15 different dress designs have been selected for the available models in how many different orders can the models parade these on the podium if they all walk together in a single file?
3. Assume there was time to present all the 21 dresses. Each time a presentation is made the 15 models come onto the podium in a single file. In this case how many permutations are possible in presenting the dresses?

## 15. Thalassothérapie

### Situation

Thalassothérapie is a type of health spa that uses seawater as a base of the therapy treatment (*thalassa* from the Greek meaning sea). The thalassothérapie centres are located in coastal areas in Morocco, Tunisia, and France, and are always adjacent or physical attached to a hotel such that clients will typically stay say a week at the hotel and be cared for by (usually) female therapists at the thalassothérapie centre. A week stay at a hotel with breakfast and dinner, and the use of the health spa may cost some £6,000 for two people. A particular thalassothérapie centre offers the following eight choices of individual treatments.<sup>5</sup>

1. Bath and seawater massage (*bain hydromassant*). This is a treatment that lasts 20 minutes where the client lies in a bath of seawater at 37°C where mineral salts have been added. In the bath there are multiple water jets that play all along the back and legs, which help to relax the muscles and improve blood circulation.
2. Oscillating shower (*douche oscillante*). In this treatment the client lies face down while a fine warm seawater rain oscillates across the back and legs giving the client a relaxing and sedative water massage (duration 20 minutes).
3. Massage under a water spray (*massage sous affusion*). This treatment is an individual massage by a therapist over the whole body under a fine shower of seawater. Oils are used during the massage to give a tonic rejuvenation to the complete frame (duration 20 minutes).
4. Massage with a water jet (*douche à jet*). Here the client is sprayed with a high-pressure water jet at a distance by a therapist who directs the jet over the soles of the feet, the calf muscles, and the back. This treatment tones up the muscles, has an anti-cramp effect and increases the blood circulation (duration 10 minutes).
5. Envelopment in seaweed (*enveloppement d'algues*). In this treatment the naked body is first covered with a warm seaweed emulsion. The client is then completely wrapped from the neck down in a heavy heated mattress. This treatment causes the client to perspire eliminating toxins and recharges the body with iodine and other trace elements from the seaweed (duration 30 minutes).

<sup>5</sup>Based on the Thalassothérapie centre (Thalazur), avenue du Parc, 33120 Arcachon, France July 2005.

6. Application of seawater mud (*application de boue marine*). This treatment is very similar to the envelopment in seaweed except that mud from the bottom of the sea is used instead of seaweed. Further, attention is made to apply the mud to the joints as this treatment serves to ease the pain from rheumatism and arthritis (duration 30 minutes).
7. Hydro-jet massage (*hydrojet*). In this treatment the client lies on their back on the bare plastic top of a water bed maintained at 37°C. High-pressure water jets within the bed pound the legs and back giving a dry tonic massage (duration 15 minutes).
8. Dry massage (*massage à sec*). This is a massage by a therapist where oils are rubbed slowly into the body toning up the muscles and circulation system (duration 30 minutes).

In addition to the individual treatments, there are also the following four treatments that are available in groups or which can be used at any time:

1. Relaxation (*relaxation*). This is a group therapy where the participants have a gym session consisting of muscle stretching, breathing, and mental reflection (duration 30 minutes).
2. Gymnastic in a seawater swimming pool (*Aquagym*). This is a group therapy where the participants have a gym session of running, walking, and jumping in a swimming pool (duration 30 minutes).
3. Steam bath (*hammam*). The steam bath originated in North Africa and is where clients sit or lie in a marble covered room in which hot steam is pumped. This creates a humid atmosphere where the client perspires to clean the pores of the skin (maximum recommended duration, 15 minutes).
4. Sauna. The sauna originated in Finland and is a room of exotic wood panelling into which hot dry air is circulated. The temperature of a sauna can reach around 100°C and the dryness of the air can be tempered by pouring water over hot stones that add some humidity (maximum recommended duration, 10 minutes).

### Required

1. Considering just the eight individual treatments, how many different ways can these be sequentially organized?
2. Considering just the four non-individual treatments, how many different ways can these be sequentially organized?
3. Considering all the 12 treatments, how many different ways can these be sequentially organized?
4. One of the programmes offered by the thalassothérapie centre is 6 days for five of the individual treatments alternating between the morning and afternoon. The morning session starts at 09:00 hours and finishes at 12:30 hours and the afternoon session starts at 14:00 hours and finishes at 17:00 hours. In this case, how many possible ways can a programme be put together without any treatment appearing twice on the same day? Show a possible weekly schedule.

## 16. Case: *Supply chain management class*

### Situation

A professor at a Business School in Europe teaches a popular programme in supply chain management. In one particular semester there are 80 participants signed up for the class. When the participants register they are asked to complete a questionnaire regarding their sex, age, country of origin, area of experience, marital status, and the number of children. This information helps the professor organize study groups, which are balanced in terms of the participant's background. This information is contained in the table below. The professor teaches the whole group of 80 together and there is always 100% attendance. The professor likes to have an interactive class and he always asks questions during his class.

### Required

When you have a database with this type of information, there are many ways to analyse the information depending on your needs. The following gives some suggestions, but there are several ways of interpretation.

1. What is the probability that if the professor chooses a participant at random then that person will be:
  - (a) From Britain?
  - (b) From Portugal?
  - (c) From the United States?
  - (d) Have experience in Finance?
  - (e) Have experience in Marketing?
  - (f) Be from Italy?
  - (g) Have three children?
  - (h) Be female?
  - (i) Is greater than 30 years in age?
  - (j) Are aged 25 years?
  - (k) Be from Britain, have experience in engineering, and be single?
  - (l) From Europe?
  - (m) Be from the Americas?
  - (n) Be single?
2. Given that a participant is from Britain then, what is the probability that that the person will:
  - (a) Have experience in engineering?
  - (b) Have experience in purchasing?
3. Given that a participant is interested in finance, then what is the probability that person is from an Asian country?
4. Given that a participant has experience in marketing, then what is the probability that person is from Denmark?
5. What is the average number of children per participant?



Number	Sex	Age	Country	Experience	Marital status	Children
1	M	21	United States	Engineering	Married	0
2	F	25	Mexico	Marketing	Single	2
3	F	27	Denmark	Marketing	Married	0
4	F	31	Spain	Engineering	Married	2
5	F	23	France	Production	Married	0
6	M	26	France	Production	Single	3
7	M	25	Germany	Engineering	Single	0
8	F	29	Canada	Production	Single	3
9	M	32	Britain	Engineering	Married	2
10	F	21	Britain	Finance	Single	1
11	M	26	Spain	Engineering	Married	2
12	M	28	United States	Finance	Single	0
13	F	27	China	Engineering	Married	3
14	M	35	Germany	Production	Married	0
15	F	21	France	Engineering	Married	2
16	F	26	Germany	Marketing	Married	3
17	F	25	Britain	Production	Married	3
18	F	31	China	Production	Single	4
19	M	22	Britain	Production	Married	2
20	M	20	Britain	Marketing	Single	3
21	F	26	Germany	Engineering	Married	2
22	M	28	Portugal	Engineering	Single	1
23	M	29	Germany	Engineering	Single	0
24	M	35	Luxembourg	Production	Married	0
25	M	41	Germany	Finance	Married	3
26	F	25	Britain	Marketing	Single	0
27	M	23	Britain	Engineering	Married	3
28	F	23	Denmark	Production	Single	3
29	M	25	Denmark	Marketing	Single	2
30	F	26	Norway	Finance	Married	3
31	F	22	France	Marketing	Single	2
32	F	26	Portugal	Engineering	Married	3
33	F	28	Spain	Engineering	Single	3
34	M	24	Germany	Production	Married	2
35	M	23	Britain	Engineering	Single	1
36	M	25	United States	Production	Married	0
37	M	26	Canada	Engineering	Married	0
38	F	24	Canada	Marketing	Single	2
39	F	25	Denmark	Marketing	Single	0
40	M	28	Norway	Engineering	Married	3
41	F	31	France	Finance	Married	5
42	M	32	Britain	Engineering	Married	2
43	F	26	Britain	Finance	Single	3
44	M	21	Luxembourg	Marketing	Single	2
45	M	25	China	Marketing	Married	5
46	M	24	Japan	Production	Married	2

Number	Sex	Age	Country	Experience	Marital status	Children
47	F	25	France	Marketing	Single	0
48	F	26	Britain	Marketing	Married	3
49	M	24	Germany	Production	Single	2
50	F	21	Taiwan	Engineering	Married	1
51	F	31	China	Engineering	Single	3
52	F	35	Britain	Marketing	Married	0
53	M	38	United States	Marketing	Married	5
54	F	39	China	Engineering	Single	2
55	M	23	Portugal	Purchasing	Married	3
56	F	25	Indonesia	Engineering	Married	2
57	M	26	Portugal	Purchasing	Married	2
58	M	23	Britain	Marketing	Single	0
59	M	25	China	Purchasing	Married	3
60	M	26	Canada	Engineering	Single	0
61	F	24	Mexico	Purchasing	Married	3
62	M	25	China	Engineering	Single	0
63	F	28	France	Production	Married	1
64	M	31	United States	Marketing	Single	2
65	F	32	Britain	Marketing	Married	3
66	F	25	Germany	Engineering	Single	0
67	M	25	Spain	Purchasing	Married	2
68	M	25	Portugal	Engineering	Single	1
69	M	26	Luxembourg	Production	Single	3
70	F	24	Taiwan	Marketing	Single	0
71	M	25	Luxembourg	Production	Married	1
72	F	26	Britain	Engineering	Married	2
73	M	28	United States	Engineering	Single	3
74	F	25	France	Engineering	Married	0
75	M	26	France	Production	Single	0
76	F	31	Germany	Marketing	Single	0
77	M	40	France	Engineering	Married	3
78	F	25	Spain	Marketing	Single	2
79	M	26	Portugal	Purchasing	Married	1
80	M	23	Taiwan	Production	Single	1

*This page intentionally left blank*

# Probability analysis for discrete data

## The shopping mall

*How often do you go to the shopping mall – every day, once a week, or perhaps just once a month? When do you go? Perhaps after work, after dinner, in the morning when you think you can beat the crowds, or on the weekends? Why do you go? It might be that you have nothing else better to do, it is a grey, dreary day and it is always bright and cheerful in the mall, you need a new pair of shoes, you need a new coat, you fancy buying a couple of CDs, you are going to meet some friends, you want to see a film in the evening so you go to the mall a little early and just have a look around. All these variables of when and why people go to the mall represent a complex random pattern of potential customers. How does the retailer manage this randomness? Further, when these potential customers are at the mall they behave in a binomial fashion – either they buy or they do not buy. Perhaps in the shopping mall there is a supermarket. It is Saturday, and the supermarket is full of people buying groceries. How to manage the waiting line or the queue at the cashier desk? This chapter covers some of these concepts.*

## Learning objectives

After you have studied this chapter you will learn the application of **discrete random variables**, and how to use the **binomial** and the **Poisson distributions**. These subjects are treated as follows:

- ✓ **Distribution for discrete random variables** • Characteristics of a random variable • Expected value of rolling two dice • Application of the random variable: *Selling of wine* • Covariance of random variables • Covariance and portfolio risk • Expected value and the law of averages
- ✓ **Binomial distribution** • Conditions for a binomial distribution to be valid • Mathematical expression of the binomial function • Application of the binomial distribution: *Having children* • Deviations from the binomial validity
- ✓ **Poisson distribution** • Mathematical expression for the Poisson distribution • Application of the Poisson distribution: *Coffee shop* • Poisson approximated by the binomial relationship • Application of the Poisson–binomial relationship: *Fenwick's*

**Discrete data** are statistical information composed of **integer values**, or **whole numbers**. They originate from the counting process. For example, we could say that 9 machines are shutdown, 29 bottles have been sold, 8 units are defective, 5 hotel rooms are vacant, or 3 students are absent. It makes little sense to say  $9\frac{1}{2}$  machines are shutdown,  $29\frac{1}{2}$  bottles have been sold,  $8\frac{1}{2}$  units are defective,  $5\frac{1}{2}$  hotel rooms are empty, or  $3\frac{1}{2}$  students are absent. With discrete data there is a clear segregation and the data does not progress from one class to another. It is information that is unconnected.

### Distribution for Discrete Random Variables

If the values of discrete data occur in no special order, and there is no explanation of their configuration or distribution, then they are considered **discrete random variables**. This means that, within the range of the possible values of the data, every value has an equal chance of occurring. In our gambling situation, discussed in Chapter 3, the value obtained by throwing a single die is random and the drawing of a card

from a full pack is random. Besides gambling, there are many situations in the environment that occur randomly and often we need to understand the pattern of randomness in order to make appropriate decisions. For example as illustrated in the Box Opener “The shopping mall”, the number of people arriving at a shopping mall in any particular day is random. If we knew the pattern it would help to better plan staff needs. The number of cars on a particular stretch of road on any given day is random and knowing the pattern would help us to decide on the appropriateness of installing stop signs, or traffic signals for example. The number of people seeking medical help at a hospital emergency centre is random and again understanding the pattern helps in scheduling medical staff and equipment. It is true that in some cases of randomness, factors like the weather, the day of the week, or the hour of the day, do influence the magnitude of the data but often even if we know these factors the data are still random.

### Characteristics of a random variable

Random variables have a mean value and a standard deviation. The **mean value of random data** is the weighted average of all the possible

outcomes of the random variable and is given by the expression:

$$\text{Mean value, } \mu_x = \sum x * P(x) = E(x) \quad 4(i)$$

Here  $x$  is the value of the discrete random variable, and  $P(x)$  is the probability, or the chance of obtaining that value  $x$ . If we assume that this particular pattern of randomness might be repeated we call this mean also the **expected value of the random variable**, or  $E(x)$ .

The **variance of a distribution of a discrete random variable** is given by the expression,

$$\text{Variance, } \sigma^2 = \sum (x - \mu_x)^2 P(x) \quad 4(ii)$$

This is similar to the calculation of the variance of a population given in Chapter 2, except that instead of dividing by the number of data values, which gives a straight average, here we are multiplying by  $P(x)$  to give a weighted average.

The **standard deviation of a random variable** is the square root of the variance or,

$$\text{Standard deviation, } \sigma = \sqrt{\sum (x - \mu_x)^2 P(x)} \quad 4(iii)$$

The following demonstrates the application of analysing the random variable in the throwing of two dice.

## Expected value of rolling two dice

In Chapter 3, we used combined probabilities to determine that the chance of obtaining the Number 7 on the throw of two dice was 16.67%. Let us turn this situation around and ask the question, "What is the expected value obtained in the throwing two dice, A and B?" We can use equation 4(i) to answer this question.

Table 4.1 gives the possible 36 combinations that can be obtained on the throw of two dice. As this table shows of the 36 combinations, there are just 11 different possible total values (2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12) by adding the numbers from the two dice. The number of possible ways that these 11 totals can be achieved is summarized

in Column 2 of Table 4.2 and the probability  $P(x)$  of obtaining these totals is in Column 3 of the same table. Using equation 4(i) we can calculate the expected or mean value of throwing two dice and the calculation and the individual results are in Columns 4 and 5. The total in the last line of Column 4 indicates the probability of obtaining these eleven values as 36/36 or 100%. The expected value of throwing two dice is 7 as shown in the last line of Column 5. The last column of Table 4.2 gives the calculation for the variation of obtaining the Number 7 using equation 4(ii). Finally, from equation 4(iii) the standard deviation is,

$$\sqrt{40.8333} = 6.3901.$$

Another way that we can determine the average value of the number obtained by throwing two dice is by using equation 2(i) for the mean value given in Chapter 2:

$$\bar{x} = \frac{\sum x}{N} \quad 2(i)$$

From Column 1 of Table 4.2 the total value of the possible throws is,

$$\begin{aligned} \sum x &= 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 \\ &\quad + 11 + 12 = 77 \end{aligned}$$

The value  $N$ , or the number of possible throws to give this total value is 11. Thus,

$$\bar{x} = \frac{\sum x}{N} = \frac{77}{11} = 7$$

The following is a business-related application of using the random variable.

## Application of the random variable: Selling of wine

Assume that a distributor sells wine by the case and that each case generates €6.00 in profit. The sale of wine is considered random. Sales data for the last 200 days is given in Table 4.3.

If we consider that this data is representative of future sales, then the frequency of occurrence of sales can be used to estimate the expected or

Table 4.1 Possible outcomes on the throw of two dice.

Throw No.	1	2	3	4	5	6	7	8	9	10	11	12
Die A	1	2	3	4	5	6	1	2	3	4	5	6
Die B	1	1	1	1	1	1	2	2	2	2	2	2
Total	2	3	4	5	6	7	3	4	5	6	7	8
Throw No.	13	14	15	16	17	18	19	20	21	22	23	24
Die A	1	2	3	4	5	6	1	2	3	4	5	6
Die B	3	3	3	3	3	3	4	4	4	4	4	4
Total	4	5	6	7	8	9	5	6	7	8	9	10
Throw No.	25	26	27	28	29	30	31	32	33	34	35	36
Die A	1	2	3	4	5	6	1	2	3	4	5	6
Die B	5	5	5	5	5	5	6	6	6	6	6	6
Total	6	7	8	9	10	11	7	8	9	10	11	12

Table 4.2 Expected value of the outcome of the throwing of two dice.

Value of throw ( $x$ )	Number of possible ways	Probability $P(x)$	$x * P(x)$	Weighted value of $x$	$(x - \mu)$	$(x - \mu)^2$	$(x - \mu)^2 * P(x)$
2	1	1/36	2 * (1/36)	0.0556	-5	25	1.3889
3	2	2/36	3 * (2/36)	0.1667	-4	16	2.6667
4	3	3/36	4 * (3/36)	0.3333	-3	9	3.0000
5	4	4/36	5 * (4/36)	0.5556	-2	4	2.2222
6	5	5/36	6 * (5/36)	0.8333	-1	1	0.8333
7	6	6/36	7 * (6/36)	1.1667	0	0	0.0000
8	5	5/36	8 * (5/36)	1.1111	1	1	1.1111
9	4	4/36	9 * (4/36)	1.0000	2	4	4.0000
10	3	3/36	10 * (3/36)	0.8333	3	9	7.5000
11	2	2/36	11 * (2/36)	0.6111	4	16	9.7778
12	1	1/36	12 * (1/36)	0.3333	5	25	8.3333
Total	36	36/36	$E(x) = 7.0000$				40.8333

average value, of future profits. Here, the values, “days this amount of wine is sold” are used to calculate the probability of future sales using the relationship,

$$\text{Probability of selling amount } x = \frac{\text{days amount of } x \text{ sold}}{\text{total days considered in analysis}} \quad 4(\text{iv})$$

For example, from equation 4(iv)

$$\text{Probability of selling 12 cases is } 80/200 = 40.00\%$$

The complete probability distribution is given in Table 4.4, and the histogram of this frequency distribution of the probability of sale is in Figure 4.1.

Using equation 4(i) to calculate the mean value, we have,

$$\begin{aligned} \mu_x &= 10 * 15\% + 11 * 20\% + 12 * 40\% \\ &\quad + 13 * 25\% = 11.75 \text{ cases} \end{aligned}$$

From this, an estimate of future profits is €6.00 \* 11.75 = €70.50/day.

Using equation 4(ii) to calculate the variance,

$$\begin{aligned} \sigma^2 &= (10 - 11.75)^2 * 15\% + (11 - 11.75)^2 \\ &\quad * 20\% + (12 - 11.75)^2 * 40\% \\ &\quad + (13 - 11.75)^2 * 25\% = 0.9875 \text{ cases}^2 \end{aligned}$$

Using equation 4(iii) to calculate the standard deviation we have,

$$\sigma = \sqrt{0.9875} = 0.9937$$

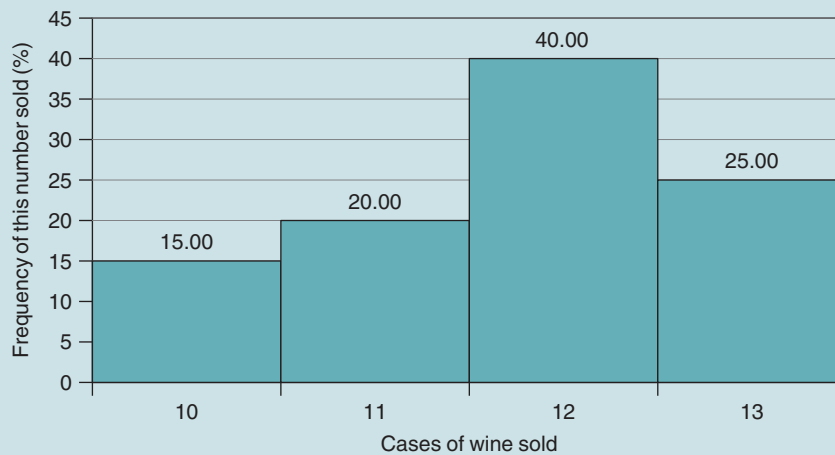
**Table 4.3** Cases of wine sold over the last 200 days.

Cases of wine sold per day	10	11	12	13	Total days
Days this amount of wine is sold	30	40	80	50	200

**Table 4.4** Cases of wine sold over the last 200 days.

Cases sold per day	10	11	12	13	Total
Days this amount of wine is sold	30	40	80	50	200
Probability of selling this amount (%)	15	20	40	25	100

**Figure 4.1** Frequency distribution of the sale of wine.





These calculations give a plausible approach of estimating average long-term future activity on the condition that the past is representative of the future.

## Covariance of random variables

Covariance is an application of the distribution of random variables and is useful to analyse the risk associated with financial investments. If we consider two datasets then the covariance,  $\sigma_{xy}$ , between two discrete random variables  $x$  and  $y$  in each of the datasets is,

$$\sigma_{xy} = \sum (x - \mu_x)(y - \mu_y)P(xy) \quad 4(v)$$

Here  $x$  is a discrete random variable in the first dataset and  $y$  is a discrete random variable in the second dataset. The terms  $\mu_x$  and  $\mu_y$  are the mean or expected values of the corresponding datasets and  $P(xy)$  is the probability of each occurrence.

The expected value of the sum of two random variables is,

$$E(x + y) = E(x) + E(y) = \mu_x + \mu_y \quad 4(vi)$$

The variance of the sum of two random variables is,

$$\text{Variance } (x + y) = \sigma_{(x+y)}^2 = \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy} \quad 4(vii)$$

The standard deviation is the square root of the variance or

$$\text{Standard deviation } (x + y) = \sqrt{\sigma_{(x+y)}^2} \quad 4(viii)$$

## Covariance and portfolio risk

An extension of random variables is **covariance**, which can be used to analyse **portfolio risk**. Assume that you are considering investing in two types of investments. One is a high growth fund, X, and the other is essentially a bond fund, Y. An estimate of future returns, per \$1,000 invested, according to expectations of the future outlook of the macro economy is given in Table 4.5.

Using equation 4(i) to calculate the mean or expected values, we have,

$$\begin{aligned} \mu_x &= 20\% * -\$100 + 35\% * \$125 \\ &\quad + 45\% * \$300 \\ &= \$158.75 \end{aligned}$$

Table 4.5 Covariance and portfolio risk.

Economic change	Contracting	Stable	Expanding
Probability of economic change (%)	20	35	45
High growth fund (X)	−\$100	\$125	\$300
Bond fund (Y)	\$250	\$100	\$10

$$\begin{aligned} \mu_y &= 20\% * -\$250 + 35\% * \$100 \\ &\quad + 45\% * \$10 \\ &= \$89.50 \end{aligned}$$

Using equation 4(ii) to calculate the variance, we have

$$\begin{aligned} \sigma_x^2 &= (-100 - 158.75)^2 * 20\% \\ &\quad + (125 - 158.75)^2 * 35\% \\ &\quad + (300 - 158.75)^2 * 45\% \\ &= \$22,767.19 \end{aligned}$$

$$\begin{aligned} \sigma_y^2 &= (250 - 89.50)^2 * 20\% + (100 - 89.50)^2 \\ &\quad * 35\% + (10 - 89.50)^2 * 45\% \\ &= \$8,034.75 \end{aligned}$$

Using equation 4(iii) to calculate the standard deviation,

$$\sigma_x = \sqrt{22,767.19} = 150.89$$

$$\sigma_y = \sqrt{8,034.75} = 89.64$$

The high growth fund, X, has a higher expected value than the bond fund, Y. However, the standard deviation of the high growth fund is higher and this is an indicator that the investment risk is greater.

Using equation 4(v) to calculate the covariance,

$$\begin{aligned} \sigma_{xy} &= (-100 - 158.75) \\ &\quad * (250 - 89.50) * 20\% + (125 - 158.75) \\ &\quad * (100 - 89.50) * 35\% + (300 - 158.75) \\ &\quad * (10 - 89.50) * 45\% \\ &= -\$13,483.13 \end{aligned}$$

The covariance between the two investments is negative. This implies that the returns on the

investments are moving in the opposite direction, or when the return on one is increasing, the other is decreasing and vice versa.

From equation 4(vi) the expected value of the sum of the two investments is,

$$\mu_x + \mu_y = \$158.75 + \$89.50 = \$248.25$$

From equation 4(vii) the variance of the sum of the two investments is,

$$\sigma_{(x+y)}^2 = 22,767.19 + 8,034.75 + 2 * \$ -13,483.13 = \$3,835.69$$

From equation 4(viii) the standard deviation of the sum of the two investments is,

$$\sqrt{\sigma_{(x+y)}^2} = \sqrt{\$3,835.69} = \$61.93$$

The standard deviation of the two funds is less than standard deviation of the individual funds because there is a negative covariance between the two investments. This implies that there is less risk with the joint investment than just with an individual investment.

If  $\alpha$  is the assigned weighting to the asset X, then since there are only two assets the situation is binomial and thus the weighting for the other asset is  $(1 - \alpha)$ . The portfolio expected return for an investment of two assets,  $E(P)$ , is,

$$E(P) = \mu_p = \alpha\mu_x + (1 - \alpha)\mu_y \quad 4(\text{ix})$$

The risk associated with a portfolio is given by:

$$\sqrt{[\alpha^2\sigma_x^2 + (1 - \alpha)^2\sigma_y^2 + 2\alpha(1 - \alpha)\sigma_{xy}]} \quad 4(\text{x})$$

Assume that we have 40% of our investment in the high-risk fund, which means there is 60% in the bond fund. Then from equation 4(ix) the portfolio expected return is,

$$\mu_p = \alpha\mu_x + (1 - \alpha)\mu_y = 40\% * \$158.75 + 60\% * \$89.50 = \$117.20$$

From equation 4(x) the risk associated with this portfolio is

$$\begin{aligned} & \sqrt{[\alpha^2\sigma_x^2 + (1 - \alpha)^2\sigma_y^2 + 2\alpha(1 - \alpha)\sigma_{xy}]} \\ &= \sqrt{[0.40^2 * \$22,767.19 + 0.60^2 * \$8,034.75 + 2 * 0.40 * 0.60 * \$ -13,483.13]} \\ &= 7.96 \end{aligned}$$

Thus in summary, the portfolio has an expected return of \$117.20, or since this amount is based on an investment of \$1,000, there is a return of 11.72%. Further for every \$1,000 invested there is a risk of \$7.96. Figure 4.2 gives a graph of the expected return according to the associated risk. This shows that the minimum risk is when there is 40% in the high growth fund and 60% in the bond fund. Although there is a higher expected return when the weighting in the high growth fund is more, there is a higher risk.

## Expected values and the law of averages

When we talk about the mean, or expected value in probability situations, this is not the value that will occur next, or even tomorrow. It is the value that is expected to be obtained in the long run. In the short term we really do not know what will happen. In gambling for example, when you play the slot machines, or one-armed bandits, you may win a few games. In fact, quite a lot of the money put into slot machines does flow out as jackpots but about 6% rests with the house.<sup>1</sup> Thus if you continue playing, then in the long run you will lose because the gambling casinos have set their machines so that the casino will be the long-term winner. If, not they would go out of business! With probability, it is the **law of averages** that governs. This law says that the average value obtained in the long term will be close to the expected value, which is the weighted outcome based on each of the probability of occurrence.

The long-term result corresponding to the law of averages can be explained by Figure 4.3. This illustrates the tossing of a coin 1,000 times where we have a 50% probability of obtaining

<sup>1</sup> Henriques, D.B., On bases, problem gamblers battle the odds, *International Herald Tribune*, 20 October 2005, p. 5.

Figure 4.2 Portfolio analysis: expected value and risk.

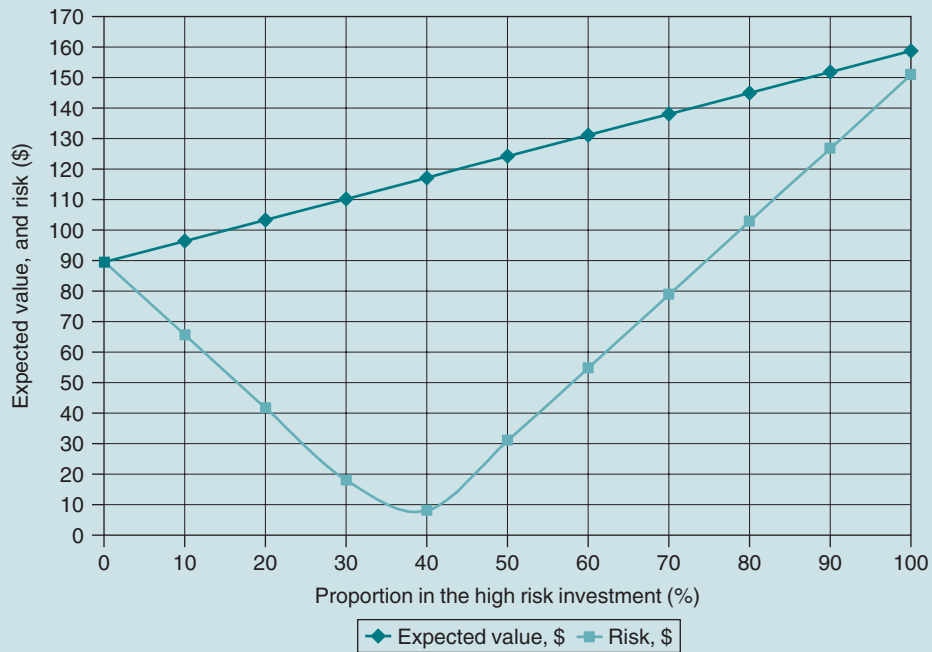
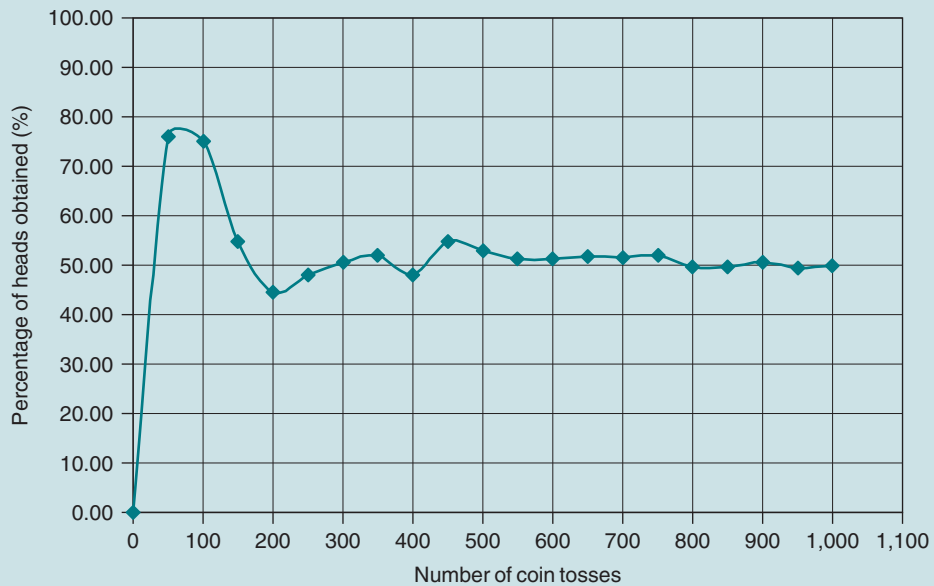


Figure 4.3 Tossing a coin 1,000 times.



heads and a 50% probability of obtaining tails. The  $y$ -axis of the graph is the cumulative frequency of obtaining heads and the  $x$ -axis is the number of times the coin is tossed. In the early throws, as we toss the coin, the cumulative number of heads obtained may be more than the cumulative number of tails as illustrated. However, as we continue tossing the coin, the law of averages comes into play, and the cumulative number of heads obtained approaches the cumulative number of tails obtained. After 1,000 throws we will have approximately 500 heads and 500 tails. This illustration supports the Rule 1 of the counting process given in Chapter 2.

You can perhaps apply the law of averages on a non-quantitative basis to the behaviour in society. We are educated to be honest, respectful, and ethical. This is the norm, or the average, of society's behaviour. There are a few people who might cheat, steal, be corrupt, or be violent. In the short term these people may get away with it. However, often in the long run the law of averages catches up with them. They get caught, lose face, are punished or may be removed from society!

## Binomial Distribution

In statistics, binomial means there are only two possible outcomes from each trial of an experiment. The tossing of a coin is binomial since the only possible outcomes are heads or tails. In quality control for the manufacture of light bulbs the principle test is whether the bulb illuminates or

does not. This is a binomial condition. If in a market survey, a respondent is asked if she likes a product, then the alternative response must be that she does not. Again, this is binomial. If we know beforehand that a situation exhibits a binomial pattern then we can use the knowledge of statistics to better understand probabilities of occurrence and make suitable decisions. We first develop a **binomial distribution**, which is a table or a graph showing all the possible outcomes of performing many times, the binomial-type experiment. The binomial distribution is discrete.

### Conditions for a binomial distribution to be valid

In order for the binomial distribution to be valid we consider that each observation is selected from an infinite population, or one of a very large size usually without replacement. Alternatively, if the population is finite, such as a pack of 52 cards, then the selection has to be with replacement. Since there are only two possible outcomes, if we say that the probability of obtaining one outcome, or "success" is  $p$ , then the probability of obtaining the other, or "failure," is  $q$ . The value of  $q$  must be equal to  $(1 - p)$ . The idea of failure here simply means the opposite of what you are testing or expecting. Table 4.6 gives some various qualitative outcomes using  $p$  and  $q$ .

Other criteria for the binomial distribution are that the probability,  $p$ , of obtaining an outcome must be fixed over time and that the outcome of any result must be independent of a previous result. For example, in the tossing of a coin, the probability of obtaining heads or tails

Table 4.6 Qualitative outcomes for a binomial occurrence.

Probability, $p$	Success	Win	Works	Good	Present	Pass	Open	Odd	Yes
Probability, $q = (1 - p)$	Failure	Lose	Defective	Bad	Absent	Fail	Shut	Even	No

remains always at 50% and obtaining a head on one toss has no effect on what face is obtained on subsequent tosses. In the throwing a die, an odd or even number can be thrown, again with a probability outcome of 50%. For each result one throw has no bearing on another throw. In the drawing of a card from a full pack, the probability of obtaining a black card (spade or clubs) or obtaining a red card (heart or diamond) is again 50%. If a card is replaced after the drawing, and the pack shuffled, the results of subsequent drawings are not influenced by previous drawings. In these three illustrations we have the following relationship:

$$\begin{aligned}\text{Probability, } p &= (1 - p) \\ &= q = 0.5 \text{ or } 50.00\% \quad 4(\text{xi})\end{aligned}$$

## Mathematical expression of the binomial function

The relationship in equation 4(xii) for the binomial distribution was developed by experiments carried out by Jacques Bernoulli (1654–1705) a Swiss/French mathematician and as such the binomial distribution is sometimes referred to as a **Bernoulli process**.

Probability of  $x$  successes, in  $n$  trials

$$= \frac{n!}{x! (n-x)!} \cdot p^x \cdot q^{(n-x)} \quad 4(\text{xii})$$

- $p$  is the characteristic probability, or the probability of *success*,
- $q = (1 - p)$  or the probability of *failure*,
- $x$  = number of successes desired,
- $n$  = number of trials undertaken, or the sample size.

The binomial random variable  $x$  can have any integer value ranging from 0 to  $n$ , the number of trials undertaken. Again, if  $p = 50\%$ , then  $q$  is 50% and the resulting binomial distribution is symmetrical regardless of the sample size,  $n$ . This is the case in the coin toss experiment, obtaining

an even or odd number on throwing a die, or selecting a black and red card from a pack. When  $p$  is not equal to 50% the distribution is skewed.

In the binomial function, the expression,

$$p^x \cdot q^{(n-x)} \quad 4(\text{xiii})$$

is the probability of obtaining exactly  $x$  successes out of  $n$  observations in a particular sequence. The relationship,

$$\frac{n!}{x! (n-x)!} \quad 4(\text{xiv})$$

is how many combinations of the  $x$  successes, out of  $n$  observations are possible. We have already presented this expression in the counting process of Chapter 3.

The **expected value of the binomial distribution**  $E(x)$  or the mean value,  $\mu_x$ , is the product of the number of trials and the **characteristic probability**.

$$\mu_x = E(x) = n \cdot p \quad 4(\text{xv})$$

For example, if we tossed a coin 40 times then the mean or expected value would be,

$$40 \cdot 0.5 = 20$$

The **variance of the binomial distribution** is the product of the number of trials, the characteristic probability of *success*, and the characteristic probability of *failure*.

$$\sigma^2 = n \cdot p \cdot q \quad 4(\text{xvi})$$

The **standard deviation of the binomial distribution** is the square root of the variance,

$$\sigma = \sqrt{\sigma^2} = \sqrt{(n \cdot p \cdot q)} \quad 4(\text{xvii})$$

Again for tossing a coin 40 times,

$$\begin{aligned}\text{Variance} &= \sigma^2 = n \cdot p \cdot q = 40 \cdot 0.5 \cdot 0.5 \\ &= 10.00\end{aligned}$$

Standard deviation,

$$\sigma = \sqrt{\sigma^2} = \sqrt{(n \cdot p \cdot q)} = \sqrt{10} = 3.16$$

## Application of the binomial distribution: *Having children*

Assume that Brad and Delphine are newly married and wish to have seven children. In the genetic makeup of both Brad and Delphine the chance of having a boy and a girl is equally possible and in their family history there is no incidence of twins or other multiple births.

1. What is the probability of Delphine giving birth to exactly two boys?

For this situation,

- $p = q = 50\%$
- $x$ , the random variable can take on the values, 0, 1, 2, 3, 4, 5, 6, and 7.
- $n$ , the sample size is 7

For this particular question,  $x = 2$  and from equation 4(xii),

$$p(x = 2) = \frac{7!}{2! (7-2)!} 0.50^2 0.50^{(7-2)}$$

$$p(x = 2) = \frac{5,040}{2 * 120} 0.25 * 0.0313$$

$$p(x = 2) = 21 * 0.25 * 0.1313 = 16.41\%$$

2. Develop a complete binomial distribution for this situation and interpret its meaning.

We do not need to go through individual calculations as by using in Excel, [function **BINOMDIST**] the complete probability distribution for each of the possible outcomes can be obtained. This is given in Table 4.7 for the individual and cumulative values. The histogram corresponding to this data is shown in Figure 4.4.

We interpret this information as follows:

- Probability of having exactly two boys = 16.41%.
- Probability of having more than two boys (3, 4, 5, 6, or 7 boys) = 77.34%.
- Probability of having at least two boys (2, 3, 4, 5, 6, or 7 boys) = 93.75%.
- Probability of having less than two boys (0 or 1 boy) = 6.25%.

Table 4.7 Probability distribution of giving birth to a boy or a girl.

Sample size (n)	7	
Probability (p)	50.00%	
Random variable (X)	Probability of obtaining exactly this value of x (%)	Probability of obtaining this cumulative value of x (%)
0	0.78	0.78
1	5.47	6.25
2	16.41	22.66
3	27.34	50.00
4	27.34	77.34
5	16.41	93.75
6	5.47	99.22
7	0.78	100.00
Total	100.00	

Mean value is  $n * p = 7 * 0.50 = 3.50$  boys (though not a feasible value)

Standard deviation is

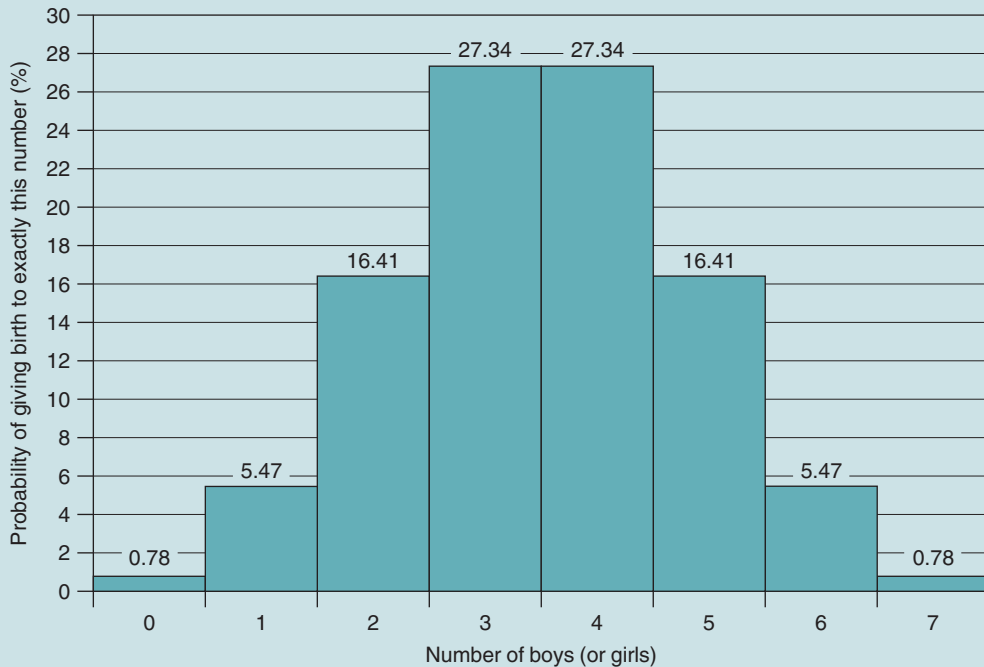
$$\sqrt{(n * p * q)} = \sqrt{(7 * 0.50 * 0.50)}$$

$$= 0.1890$$

## Deviations from the binomial validity

Many business-related situations may appear to follow a binomial situation meaning that the probability outcome is fixed over time, and the result of one outcome has no bearing on another. However, in practice these two conditions might be *violated*. Consider for example, a manager is interviewing in succession 20 candidates for one position in his firm. One of the candidates has to be chosen. Each candidate represents discrete information where their experience and ability are independent of each other. Thus, the interview process is binomial – either a particular

Figure 4.4 Probability histogram of giving birth to a boy (or girl).



candidate is selected or is not. As the manager continues the interviewing process he makes a subliminal comparison of competing candidates, in that if one candidate that is rated positively this results perhaps in a less positive rating of another candidate. Thus, the evaluation is not entirely independent. Further, as the day goes on, if no candidate has been selected, the interviewer gets tired and may be inclined to offer the post to say perhaps one of the last few remaining candidates out of sheer desperation!

In another situation, consider you drive your car to work each morning. When you get into the car, either it starts, or it does not. This is binomial and your expectation is that your car will start every time. The fact that your car started on Tuesday morning should have no effect on whether it starts on Wednesday and should not have been influenced on the fact that it started on Monday morning. However, over

time, mechanical, electrical, and even electronic components wear. Thus, on one day you turn the ignition in your car and it does not start!

## Poisson Distribution

The **Poisson distribution**, named after the Frenchman, Denis Poisson (1781–1840), is another discrete probability distribution to describe events that occur usually during a given time interval. Illustrations might be the number of cars arriving at a tollbooth in an hour, the number of patients arriving at the emergency centre of a hospital in one day, or the number of airplanes waiting in a holding pattern to land at a major airport in a given 4-hour period, or the number of customers waiting in line at the cash checkout



as highlighted in the Box Opener “The shopping mall”.

## Mathematical expression for the Poisson distribution

The equation describing the Poisson probability of occurrence,  $P(x)$  is,

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad 4(\text{xviii})$$

- $\lambda$  (lambda the Greek letter l) is the mean number of occurrences;
- $e$  is the base of the natural logarithm, or 2.71828;
- $x$  is the Poisson random variable;
- $P(x)$  is the probability of exactly  $x$  occurrences.

The **standard deviation of the Poisson distribution** is given by the square root of the mean number of occurrences or,

$$\sigma = \sqrt{\lambda} \quad 4(\text{xix})$$

In applying the Poisson distribution the assumptions are that the mean value can be estimated from past data. Further, if we divide the time period into seconds then the following applies:

- The probability of exactly one occurrence per second is a small number and is constant for every one-second interval.
- The probability of two or more occurrences within a one-second interval is small and can be considered zero.
- The number of occurrences in a given one-second interval is independent of the time at which that one-second interval occurs during the overall prescribe time period.
- The number of occurrences in any one-second interval is independent on the number of occurrences in any other one-second interval.

## Application of the Poisson distribution: Coffee shop

A small coffee shop on a certain stretch of highway knows that on average nine people per

hour come in for service. Sometimes the only waitress on the shop is very busy, and sometimes there are only a few customers.

1. The owner has decided that if there is greater than a 10% chance that there will be at least 13 clients coming into the coffee shop in a given hour, the manager will hire another waitress. Develop the information to help the manager make a decision.

To determine the probability of there being exactly 13 customers coming into the coffee shop in a given hour we can use equation 4(xviii) where in this case  $x$  is 13 and  $\lambda$  is 9.

$$\begin{aligned} P(13) &= \frac{\lambda^{13} e^{-\lambda}}{13!} \\ &= \frac{2,541,865,828,329 * 0.000123}{6,227,020,800} \\ &= 5.04\% \end{aligned}$$

Again as for the binomial distribution, you can simply calculate the distribution using in Excel the **[function POISSON]**. This distribution is shown in Table 4.8. Column 2 gives the probability of obtaining exactly the random number, and Column 3 gives the cumulative values. Figure 4.5 gives the distribution histogram for Column 2, the probability of obtaining the exact random variable. This distribution is interpreted as follows:

- Probability of exactly 13 customers entering in a given hour = 5.04%.
- Probability of more than 13 customers entering in a given hour =  $(100 - 92.61) = 7.39\%$ .
- Probability of at least 13 customers entering in a given hour =  $(100 - 87.58) = 12.42\%$ .
- Probability of less than 13 customers entering in a given hour = 87.58%.

Since the probability of at least 13 customers entering in a given hour is 12.42% or greater than 10% the manager should decide to hire another waitress.



**Table 4.8** Poisson distribution for the coffee shop.

Mean value ( $\lambda$ )		9
Random variable ( $x$ )	Probability of obtaining exactly (%)	Probability of obtaining this cumulative value of $x$ (%)
0	0.01	0.01
1	0.11	0.12
2	0.50	0.62
3	1.50	2.12
4	3.37	5.50
5	6.07	11.57
6	9.11	20.68
7	11.71	32.39
8	13.18	45.57
9	13.18	58.74
10	11.86	70.60
11	9.70	80.30
12	7.28	87.58
13	5.04	92.61
14	3.24	95.85
15	1.94	97.80
16	1.09	98.89
17	0.58	99.47
18	0.29	99.76
19	0.14	99.89
20	0.06	99.96
21	0.03	99.98
22	0.01	99.99
23	0.00	100.00
Total	100.00	

### Poisson approximated by the binomial relationship

When the value of the sample size  $n$  is large, and the characteristic probability of occurrence,  $p$ , is small, we can use the Poisson distribution as a reasonable approximation of the binomial distribution. The criteria most often applied to make this approximation is when  $n$  is greater, or equal to 20, and  $p$  is less than, or equal to 0.05% or 5%.

If this requirement is met then the mean of the binomial distribution, which is given by the product  $n * p$ , can be substituted for the mean of the Poisson distribution,  $\lambda$ . The probability relationship from equation 4(xviii) then becomes,

$$P(x) = \frac{(np)^x e^{-(np)}}{x!} \quad 4(xx)$$

The Poisson random variable,  $x$  in theory ranges from 0 to  $\infty$ . However, when the distribution is used as an approximation of the binomial distribution, the number of successes out of  $n$  observations cannot be greater than the sample size  $n$ . From equation 4(xx) the probability of observing a large number of successes becomes small and tends to zero very quickly when  $n$  is large and  $p$  is small. The following illustrates this approximation.

### Application of the Poisson–binomial approximation: Fenwick's

A distribution centre has a fleet of 25 Fenwick trolleys, which it uses every day for unloading and putting into storage products it receives on pallets from its suppliers. The same Fenwick's are used as needed to take products out of storage and transfer them to the loading area. These 25 Fenwick's are battery driven and at the end of the day they are plugged into the electric supply for recharging. From past data it is known that on a daily basis on average one Fenwick will not be properly recharged and thus not available for use.

1. What is the probability that on any given day, three of the Fenwick's are out of service?

Using the Poisson relationship equation 4(xviii) and generating the distribution in Excel by using [function POISSON] where lambda is 1, we have the Poisson distribution given in Column 2 and Column 5 of Table 4.9. From this table the probability of three Fenwick's being out of service on any given day is 6.1313% or about 6%.

Now if we use the binomial approximation, then the characteristic probability

Figure 4.5 Poisson probability histogram for the coffee shop.

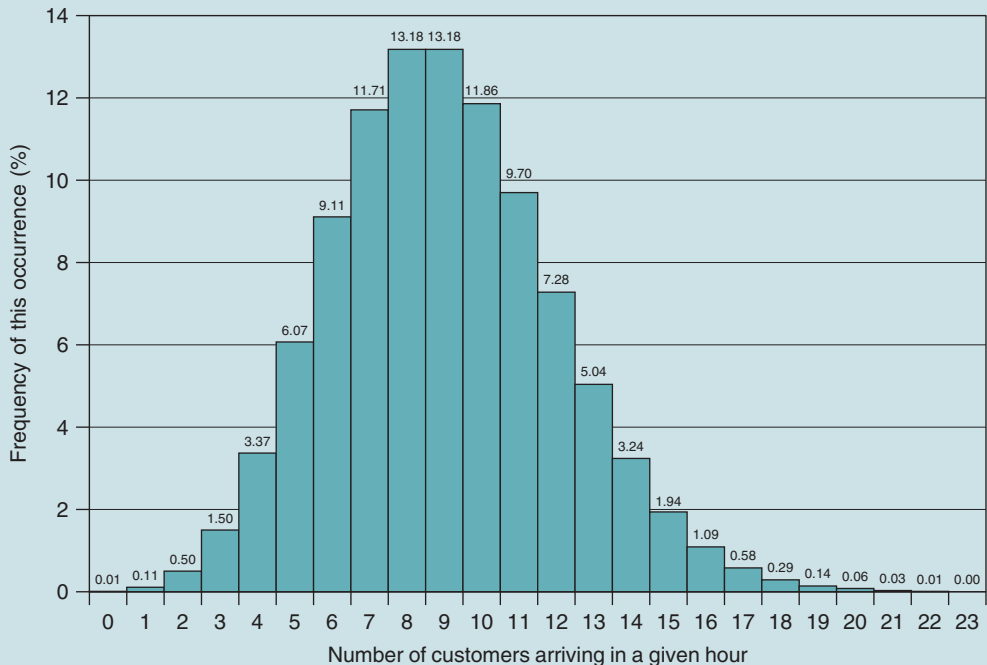


Table 4.9 Poisson and binomial distributions for Fenwick's.

Number of Fenwick's	25				
$\lambda$	1				
$p$	4.00%				
Random variable	Poisson (%)	Binomial (%)	Random variable	Poisson (%)	Binomial (%)
$X$	Exact	Exact	13	0.0000	0.0000
0	36.7879	36.0397	14	0.0000	0.0000
1	36.7879	37.5413	15	0.0000	0.0000
2	18.3940	18.7707	16	0.0000	0.0000
3	6.1313	5.9962	17	0.0000	0.0000
4	1.5328	1.3741	18	0.0000	0.0000
5	0.3066	0.2405	19	0.0000	0.0000
6	0.0511	0.0334	20	0.0000	0.0000
7	0.0073	0.0038	21	0.0000	0.0000
8	0.0009	0.0004	22	0.0000	0.0000
9	0.0001	0.0000	23	0.0000	0.0000
10	0.0000	0.0000	24	0.0000	0.0000
11	0.0000	0.0000	25	0.0000	0.0000
12	0.0000	0.0000	Total	100.00	100.00

is  $1/25$  or 4.00%. The sample size  $n$  is 25, the number of Fenwick's. Then applying the binomial relationship of equation 4(xii) by generating the distribution using [function BINOMDIST] we have the binomial distribution in Column 3 and Column 6 of Table 4.9. This indicates that on any given day, the probability of three Fenwick's being out of service is 5.9962% or again about

6%. This is about the same result as using the Poisson relationship.

Note, in Table 4.9 we have given the probabilities to four decimal places to be able to compare values that are very close. You can also notice that the probability of observing a large number of "successes" tails off very quickly to zero. In this case it is for values of  $x$  beyond the Number 5.

## Chapter Summary

This chapter has dealt with discrete random variables, their corresponding distribution, and the binomial and Poisson distribution.

### Distribution for discrete random variables

When integer or whole number data appear in no special order they are considered discrete random variables. This means that for a given range of values, any number is likely to appear. The number of people in a shopping mall, the number of passengers waiting for the Tube, or the number of cars using the motorway is relatively random. The mean or the expected value of the random variable is the weighted outcome of all the possible outcomes. The variance is calculated by the sum, of the square of the difference between a given random variable and the mean of data multiplied by the probability of occurrence. As always, the standard deviation is the square root of the variance. When we have the expected value and the dispersion or spread of the data, these relationships can be useful in estimating long-term profits, costs, or budget figures. An extension of the random variable is covariance analysis that can be used to estimate portfolio risk. The law of averages in life is underscored by the expected value in random variable analysis. We will never know exactly what will happen tomorrow, or even the day after, however over time or in the long range we can expect the mean value, or the norm, to approach the expected value.

### Binomial distribution

The binomial concept was developed by Jacques Bernoulli a Swiss/French mathematician and as such is sometimes referred to as the Bernoulli process. Binomial means that there are only two possible outcomes, yes or no, right or wrong, works or does not work, etc. For the binomial distribution to be valid the characteristic probability must be fixed over time, and the outcome of an activity must be independent of another. The mean value in a binomial distribution is the product of the sample size and the characteristic probability. The standard deviation is the square root of the product of the sample size, the characteristic probability, and the characteristic failure. If we know that data follows a binomial pattern, and we have the characteristic probability of occurrence, then for a given sample size we can determine, for example, the probability

of a quantity of products being good, the probability of a process operating in a given time period, or the probability outcome of a certain action. Although many activities may at first appear to be binomial in nature, over time the binomial relationship may be violated.

### Poisson distribution

The Poisson distribution, named after the Frenchman, Denis Poisson, is another discrete distribution often used to describe patterns of data that occur during given time intervals in waiting lines or queuing situations. In order to determine the Poisson probabilities you need to know the average number of occurrences,  $\lambda$ , which are considered fixed for the experiment in question. When this is known, the standard deviation of the Poisson function is the square root of the average number of occurrences. In an experiment when the sample size is at least 20, and the characteristic probability is less than 5%, then the Poisson distribution can be approximated using the binomial relationship. When these conditions apply, the probability outcomes using either the Poisson or the binomial distributions are very close.

## EXERCISE PROBLEMS

### 1. HIV virus

#### Situation

The Pasteur Institute in Paris has a clinic that tests men for the HIV virus. The testing is performed anonymously and the clinic has no way of knowing how many patients will arrive each day to be tested. Thus tomorrow's number of patients is a random variable. Past daily records, for the last 200 days, indicate that from 300 to 315 patients per day are tested. Thus the random variable is the number of patients per day – a discrete random variable. This data is given in Table 1.

The Director of the clinic, Professor Michel is preparing his annual budget. The total direct and indirect cost for testing each patient is €50 and the clinic is open 250 days per year.

Table 1

Men tested	Days this level tested
300	2
301	7
302	10
303	12
304	12
305	14
306	18
307	20
308	24
309	22
310	18
311	16
312	12
313	5
314	4
315	4

Table 2

Men tested	Days this level tested
300	1
301	1
302	1
303	1
304	10
305	16
306	30
307	40
308	40
309	30
310	16
311	10
312	1
313	1
314	1
315	1

#### Required

- Using the data in Table 1, what will be a reasonable estimated cost for this particular operation in this budget year? Assume that the records for the past 200 days are representative of the clinic's operation.
- If the historical data for the testing is according to Table 2, what effect would this have on your budget?
- Use the coefficient of variation (ratio of standard deviation to mean value or  $\sigma/\mu$ ) to compare the data.

4. Illustrate the distributions given by the two tables as histograms. Do the shapes of the distributions corroborate the information obtained in Question 3? Which of the data is the most reliable for future analysis, and why?

## 2. Rental cars

### Situation

Roland Ryan operates a car leasing business in Wyoming, United States with 10 outlets in this state. He is developing his budgets for the following year and is proposing to use historical data for estimating his profits for the coming year. For the previous year he has accumulated data from two of his agencies one in Cheyenne, and the other in Laramie. This data shown below gives the number of cars leased, and the number of days at which this level of cars are leased during 250 days per year when the leasing agencies are opened.

Cheyenne		Laramie	
Cars leased	Days at this level	Cars leased	Days at this level
20	2	20	1
21	9	21	1
22	12	22	2
23	14	23	2
24	14	24	12
25	18	25	20
26	24	26	38
27	26	27	49
28	29	28	50
29	27	29	37
30	25	30	19
31	20	31	13
32	15	32	2
33	8	33	2
34	6	34	1
35	1	35	1

### Required

- Using the data from the Cheyenne agency, what is a reasonable estimate of the average number of cars leased per day during the year the analysis was made?
- If each car leased generates \$22 in profit, using the average value from the Cheyenne data, what is a reasonable estimate of annual profit for the coming year for each agency?

3. If the data from Laramie was used, how would this change the response to Question 1 for the average number of cars leased per day during the year the analysis was made?
4. If the data from Laramie was used, how would this change the response to Question 2 of a reasonable estimate of annual profit for the coming year for all 10 agencies?
5. For estimating future activity for the leasing agency, which of the data from Cheyenne or Laramie would be the most reliable? Justify your response visually and quantitatively.

### 3. Road accidents

#### Situation

In a certain city in England, the council was disturbed by the number of road accidents that occurred, and the cost to the city. Some of these accidents were minor just involving damage to the vehicles involved, others involved injury, and in a few cases, death to those persons involved. These costs and injuries were obviously important but also the council wanted to know what were the costs for the services of the police and fire services. When an accident occurred, on average two members of the police force were dispatched together with three members of the fire service. The estimated cost of the police was £35 per hour per person and £47 per hour per person for the fire service. The higher cost for the fire service was because the higher cost of the equipment employed. On average each accident took 3 hours to investigate. This including getting to the scene, doing whatever was necessary at the accident scene, and then writing a report. The council conducted a survey of the number of accidents that occurred and this is in the table below.

No. of accidents ( $x$ )	No. of days occurred
0	7
1	35
2	34
3	46
4	6
5	2
6	31
7	33
8	29
9	31
10	47
11	34
12	30

#### Required

1. Plot a relative frequency probability for this data for the number of accidents that occurred.

2. Using this data, what is a reasonable estimate of the daily number of accidents that occur in this city?
3. What is the standard deviation for this information?
4. Do you think that there is a large variation for this data?
5. What is an estimated cost for the annual services of the police services?
6. What is an estimated cost for the annual services of the fire services?
7. What is an estimated cost for the annual services of the police and fire services?

#### 4. Express delivery

##### Situation

An express delivery company in a certain country in Europe offers a 48-hour delivery service to all regions of the United States for packages weighing less than 1 kg. If the firm is unable to deliver within this time frame it refunds to the client twice the fixed charge of €42.50. The following table gives the number of packages of less than one kilogram, each month, which were not delivered within the promised time-frame over the last three years.

Month	2003	2004	2005
January	6	4	10
February	4	6	7
March	5	2	3
April	3	0	4
May	0	5	4
June	1	6	5
July	10	7	9
August	2	9	3
September	2	10	3
October	2	1	6
November	3	1	4
December	11	3	8

##### Required

1. Plot a relative frequency probability for this data for the number of packages that were not delivered within the 48-hour time period.
2. What is the highest frequency of occurrence for not meeting the promised time delivery?
3. What is a reasonable estimate of the average number of packages that are not delivered within the promised time frame?
4. What is the standard deviation of the number of packages that are not delivered within the promised time frame?
5. If the firm sets an annual target of not paying to the client more than €4,500, based on the above data, would it meet the target?
6. What qualitative comments can you make about this data that might in part explain the frequency of occurrence of not meeting the time delivery?



## 5. Bookcases

### Situation

Jack Sprat produces handmade bookcases in Devon, United Kingdom. Normally he operates all year-round but this year, 2005, because he is unable to get replacement help, he decides to close down his workshop in August and make no further bookcases. However, he will leave the store open for sales of those bookcases in stock. At the end of July 2005, Jack had 19 finished bookcases in his store/workshop. Sales for the previous 2 years were as follows:

Month	2003	2004
January	17	18
February	21	24
March	22	17
April	21	21
May	23	22
June	19	23
July	22	22
August	21	19
September	20	21
October	16	18
November	22	22
December	20	15

### Required

1. Based on the above historical data, what is the expected number of bookcases sold per month?
2. What is the highest probability of selling bookcases, and what is this quantity?
3. If the average sale price of a bookcase were £250.00 using the expected value, what would be the expected financial situation for Jack?
4. What are your comments about the answer to Question 3?

## 6. Investing

### Situation

Sophie, a shrewd investor, wants to analyse her investment in two types of portfolios. One is a high growth fund that invests in blue chip stocks of major companies, plus selected technology companies. The other fund is a bond fund, which is a mixture of United States and European funds backed by the corresponding governments. Using her knowledge of finance and economics Sophie established the following regarding probability and financial returns per \$1,000 of investment.

Economic change	Contracting	Stable	Expanding
Probability of economic change (%)	15	45	40
High growth fund, change (\$/\$1,000)	−50	100	250
Bond fund change (\$/\$1,000)	200	50	10

**Required**

1. Determine the expected values of the high growth fund, and the bond fund.
2. Determine the standard deviation of the high growth fund, and the bond fund.
3. Determine the covariance of the two funds.
4. What is the expected value of the sum of the two investments?
5. What is the expected value of the portfolio?
6. What is the expected percentage return of the portfolio and what is the risk?

**7. Gift store****Situation**

Madame Charban owns a gift shop in La Ciotat. Last year she evaluated that the probability of a customer who says they are just browsing, buys something, is 30%. Suppose that on a particular day this year 15 customers browse in the store each hour.

**Required**

Assuming a binomial distribution, respond to the following questions

1. Develop the individual probability distribution histogram for all the possible outcomes.
2. What is the probability that at least one customer, who says they are browsing, will buy something during a specified hour?
3. What is the probability that at least four customers, who say they are browsing, will buy something during a specified hour?
4. What is the probability that no customers, who say they are browsing, will buy something during a specified hour?
5. What is the probability that no more than four customers, who say they are browsing, will buy something during a specified hour?

**8. European Business School****Situation**

A European business school has a 1-year exchange programme with international universities in Argentina, Australia, China, Japan, Mexico, and the United States. There is a strong demand for this programme and selection is based on language ability for the country in question, motivation, and previous examination scores. Records show that in the 70% of the candidates that apply are accepted. The acceptance for the programme follows a Bernoulli process.

**Required**

1. Develop a table showing all the possible exact probabilities of acceptance if 20 candidates apply for this programme.

2. Develop a table showing all the possible cumulative probabilities of acceptance if 20 candidates apply for this programme.
3. Illustrate, on a histogram, all the possible exact probabilities of acceptance if 20 candidates apply for this programme.
4. If 20 candidates apply, what is the probability that exactly 10 candidates will be accepted?
5. If 20 candidates apply, what is the probability that exactly 15 candidates will be accepted?
6. If 20 candidates apply, what is the probability that at least 15 candidates will be accepted?
7. If 20 candidates apply, what is the probability that no more than 15 candidates will be accepted?
8. If 20 candidates apply, what is the probability that fewer than 15 candidates will be accepted?

## 9. Clocks

### Situation

The Chime Company manufactures circuit boards for use in electric clocks. Much of the soldering work on the circuit boards is performed by hand and there are a proportion of the boards that during the final testing are found to be defective. Historical data indicates that of the defective boards, 40% can be corrected by redoing the soldering. The distribution of defective boards follows a binomial distribution.

### Required

1. Illustrate on a probability distribution histogram all of the possible individual outcomes of the correction possibilities from a batch of eight defective circuit boards.
2. What is the probability that in the batch of eight defective boards, none can be corrected?
3. What is the probability that in the batch of eight defective boards, exactly five can be corrected?
4. What is the probability that in the batch of eight defective boards, at least five can be corrected?
5. What is the probability that in the batch of eight defective boards, no more than five can be corrected?
6. What is the probability that in the batch of eight defective boards, fewer than five can be corrected?

## 10. Computer printer

### Situation

Based on past operating experience, the main printer in a university computer centre, which is connected to the local network, is operating 90% of the time. The head of Information Systems makes a random sample of 10 inspections.

### Required

1. Develop the probability distribution histogram for all the possible outcomes of the operation of the computer printer.
2. In the random sample of 10 inspections, what is the probability that the computer printer is operating in exactly 9 of the inspections?
3. In the random sample of 10 inspections, what is the probability that the computer printer is operating in at least 9 of the inspections?
4. In the random sample of 10 inspections, what is the probability that the computer printer is operating in at most 9 of the inspections?
5. In the random sample of 10 inspections, what is the probability that the computer printer is operating in more than 9 of the inspections?
6. In the random sample of 10 inspections, what is the probability that the computer printer is operating in fewer than 9 of the inspections?
7. In how many inspections can the computer printer be expected to operate?

## 11. Bank credit

### Situation

A branch of BNP-Paribas has an attractive credit programme. Customers meeting certain requirements can obtain a credit card called “BNP Wunder”. Local merchants in surrounding communities accept this card. The advantage is that with this card, goods can be purchased at a 2% discount and further, there is no annual cost for the card.

Past data indicates that 35% of all card applicants are rejected because of unsatisfactory credit. Assuming that credit acceptance, or rejection, is a Bernoulli process, and samples of 15 applicants are made.

### Required

1. Develop a probability histogram for this situation.
2. What is the probability that exactly three applicants will be rejected?
3. What is the probability that at least three applicants will be rejected?
4. What is the probability that more than three applicants will be rejected?
5. What is the probability that exactly seven applicants will be rejected?
6. What is the probability that at least seven applicants will be rejected?
7. What is the probability that more than seven applicants will be rejected?

## 12. Biscuits

### Situation

The Betin Biscuit Company every August offers discount coupons in the Rhône-Alps Region, France for the purchase of their products. Historical data at Betin’s marketing

department indicates that 80% of consumers buying their biscuits do not use the coupons. One day eight customers enter into a store to buy biscuits.

#### Required

1. Develop an individual binomial distribution for the data. Plot this data as a relative frequency distribution.
2. What is the probability that exactly six customers do not use the coupons for the Betin biscuits?
3. What is the probability that exactly seven customers do not use the coupons?
4. What is the probability that more than four customers do not use the coupons for the Betin biscuits?
5. What is the probability that less than eight customers do not use the coupons?
6. What is the probability that no more than three customers do not use the coupons?

### 13. Bottled water

#### Situation

A food company processes sparkling water into 1.5 litre PET bottles. The speed of the bottling line is very high and historical data indicates that after filling, 0.15% of the bottles are ejected. This filling and ejection operation is considered to follow a Poisson distribution.

#### Required

1. For 2,000 bottles, develop a probability histogram from zero to 15 bottles falling from the line.
2. What is the probability that for 2,000 bottles, none are ejected from the line?
3. What is the probability that for 2,000 bottles, exactly four are ejected from the line?
4. What is the probability that for 2,000 bottles, at least four are ejected from the line?
5. What is the probability that for 2,000 bottles, less than four are ejected from the line?
6. What is the probability that for 2,000 bottles, no more than four are ejected from the line?

### 14. Cash for gas

#### Situation

A service station, attached to a hypermarket, has two options for gasoline or diesel purchases. Customers either using a credit card that they insert into the pump, serve themselves with fuel such that payment is automatic. This is the most usual form of purchase. The other option is the cash-for-gas utilization area. Here the customers fill their tank and then drive to the exit and pay cash, to one of two attendants at the exit kiosk. This form of distribution is more costly to the operator principally because of the salaries of the attendants in the kiosk. The owner of this service station wants some assurance that

there is a probability of greater than 90% that 12 or more customers in any hour use the automatic pump. Past data indicates that on average 15 customers per hour use the automatic pump. The Poisson relationship will be used for evaluation.

#### Required

1. Develop a Poisson distribution for the cash-for-gas utilization area.
2. Should the service station owner be satisfied with the cash-for-gas utilization, based on the criteria given?
3. From the information obtained in Question 2 what might you propose for the owner of the service station?

## 15. Cashiers

#### Situation

A supermarket store has 30 cashiers full time for its operation. From past data, the absenteeism due to illness is 4.5%.

#### Required

1. Develop an individual Poisson distribution for the data. Plot this data as a relative frequency distribution?
2. Using the Poisson distribution, what is the probability that on any given day exactly three cashiers do not show up for work?
3. Using the Poisson distribution, what is the probability that less than three cashiers do not show up for work?
4. Using the Poisson distribution, what is the probability that more than three cashiers do not show up for work?
5. Develop an individual binomial distribution for the data. Plot this data as a relative frequency distribution.
6. Using the binomial distribution, what is the probability that on any given day exactly three cashiers do not show up for work?
7. Using the binomial distribution, what is the probability that less than three cashiers do not show up for work?
8. Using the binomial distribution, what is the probability that more than three cashiers do not show up for work?
9. What are your comments about the two frequency distribution that you have developed, and the probability values that you have determined?

## 16. Case: Oil well

#### Situation

In an oil well area of Texas are three automatic pumping units that bring the crude oil from the ground. These pumps are installed to operate continuously, 24 hours per day, 365 days

Required

Describe this situation in probability and financial terms.

[illegible]

Pump No. 1										
1	1	1	1	1	1	0	1	1	1	
1	1	0	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	0	1	1	
1	1	1	1	1	1	1	1	1	1	
1	1	1	1	1	0	1	1	1	1	
1	0	1	1	1	1	1	1	1	1	
1	1	1	1	1						

Pump No. 2										
1	1	1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	1	1	
1	1	1	1	1	1	0	1	1	1	
1	1	1	1	1	1	1	1	1	1	
1	0	1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	0	1	1	
1	1	1	1	1						

Pump No. 3										
1	1	1	1	1	1	1	1	1	1	
1	1	1	0	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	0	1	
1	1	1	1	1	1	1	1	1	1	
1	0	1	1	1	1	0	1	1	1	
0	1	1	0	1						



*This page intentionally left blank*

# Probability analysis in the normal distribution

## Your can of beer or your bar of chocolate

*When you buy a can of beer written 33 cl on the label, you have exactly a volume of 33 cl in the can, right? You are almost certainly wrong as this implies a volume of 33.0000 cl. When you buy a bar of black chocolate it is stamped on the label, net weight 100 g. Again, it is highly unlikely that you have 100.0000 g of chocolate. In operations, where the target, or the machine setting is to obtain a certain value it is just about impossible to always obtain this value. Some values will be higher, and some will be lower, just because of the variation of the filling process for the cans of beer or the moulding operation for the chocolate bars. The volume of the beer in the can, or the weight of the bar of chocolate, should not be consistently high since over time this would cost the producing firm too much money. Conversely, the volume or weight cannot be always too low as the firm will not be respecting the information given on the label and clearly this would be unethical. These measurement anomalies can be explained by the normal distribution.*

## Learning objectives

After you have studied this chapter you will understand and be able to apply the most widely used tool in statistics, *the normal distribution*. The theory and concepts of this distribution are presented as follows:

- ✓ **Describing the normal distribution** • Characteristics • Mathematical expression • Empirical rule for the normal distribution • Effect of different means and/or different standard deviations • Kurtosis in frequency distributions • Transformation of a normal distribution • The standard normal distribution • Determining the value of  $z$  and the Excel function • Application of the normal distribution: *Light bulbs*
- ✓ **Demonstrating that data follow a normal distribution** • Verification of normality • Asymmetrical data • Testing symmetry and asymmetry by a normal probability plot • Percentiles and the number of standard deviations
- ✓ **Using normal distribution to approximate a binomial distribution** • Conditions for approximating the binomial distribution • Application of the normal–binomial approximation: *Ceramic plates* • Continuity correction factor • Sample size to approximate the normal distribution

The normal distribution is developed from **continuous random variables** that unlike discrete random variables, are not whole numbers, but take fractional or decimal values. As we have illustrated in the box opener “Your can of beer or your bar of chocolate”, the nominal volume of beer in a can, or that amount indicated on the label, is 33 cl. However, the actual volume when measured may be in fact 32.8579. The nominal weight of a bar of chocolate is 100 g but the actual weight when measured may be in fact 99.7458 g. We may note that the runner completed the Santa Barbara marathon in 3 hours and 4 minutes and 32 seconds. For all these values of volume, weight, and time there is no distinct cut-off point between the data values and they can overlap into other class ranges.

### Describing the Normal Distribution

A **normal distribution** is the most important probability distribution, or frequency of occurrence,

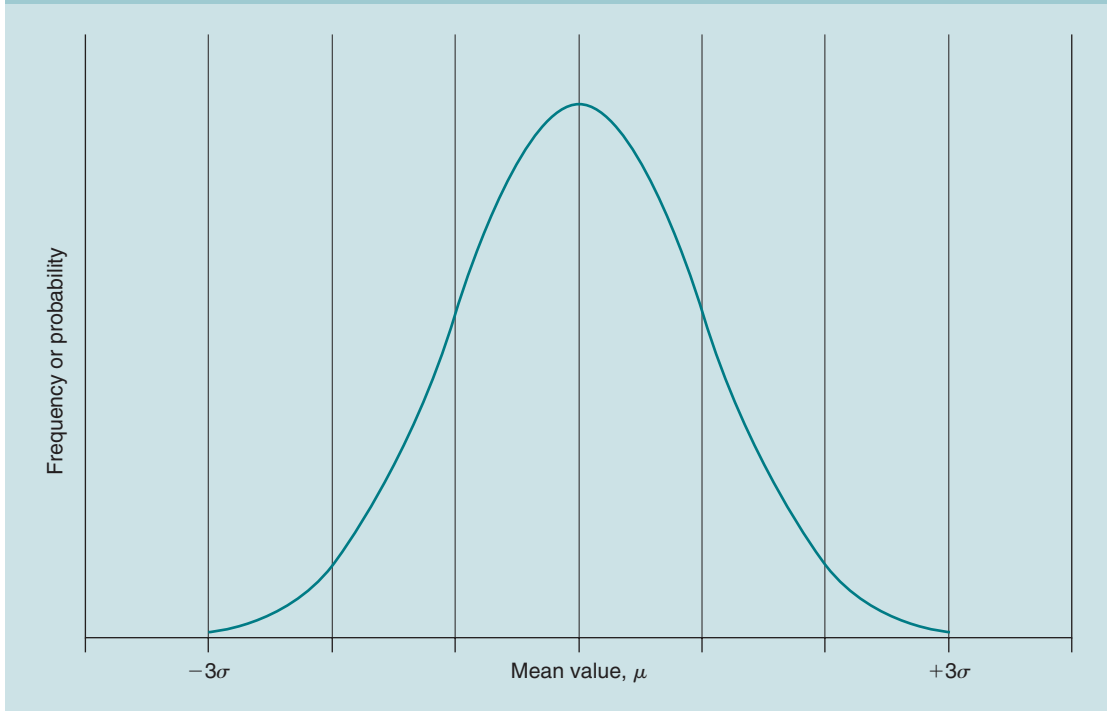
to describe a continuous random variable. It is widely used in statistical analysis. The concept was developed by the German, Karl Friedrich Gauss (1777–1855) and is thus it is also known as the Gaussian distribution. It is valuable to understand the characteristics of the normal distribution as this can provide information about probability outcomes in the business environment and can be a vital aid in decision-making.

### Characteristics

The shape of the normal distribution is illustrated in Figure 5.1. The  $x$ -axis is the value of the random variable, and the  $y$ -axis is the frequency of occurrence of this random variable. As we mentioned in Chapter 3, if the frequency of occurrence can represent future outcomes, then the normal distribution can be used as a measure of probability. The following are the basic characteristics of the distribution:

- It is a continuous distribution.
- It is bell, mound, or humped shaped and it is symmetrical around this hump. When it is

Figure 5.1 Shape of the normal distribution.



symmetrical it means that the left side is a mirror image of the right side.

- The central point, or the hump of the distribution, is at the same time the mean, median, mode, and midrange. They all have the same value.
- The left and right extremities, or the two tails of the normal distribution, may extend far from the central point implying that the associated random variable,  $x$ , has a range,  $-\infty > x < +\infty$ .
- The inter-quartile range is equal to 1.33 standard deviations.

Regarding the tails of the distributions most real-life situations do not extend indefinitely in both directions. In addition, negative values or extremely high positive values would not be possible. However, for these situations

the normal distribution is still a reasonable approximation.

### Mathematical expression

The mathematical expression for the normal distribution, and from which the continuous curve is developed, is given by the **normal distribution density function**,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-(1/2)[(x-\mu_x)/\sigma_x]^2} \quad 5(i)$$

- $f(x)$  is the probability density function.
- $\pi$  is the constant pie equal to 3.14159.
- $\sigma_x$  is the standard deviation.
- $e$  is the base of the natural logarithm equal to 2.71828.
- $x$  is the value of the random variable.
- $\mu_x$  is the mean value of the distribution.

## Empirical rule for the normal distribution

There is an **empirical rule for the normal distribution** that states the following:

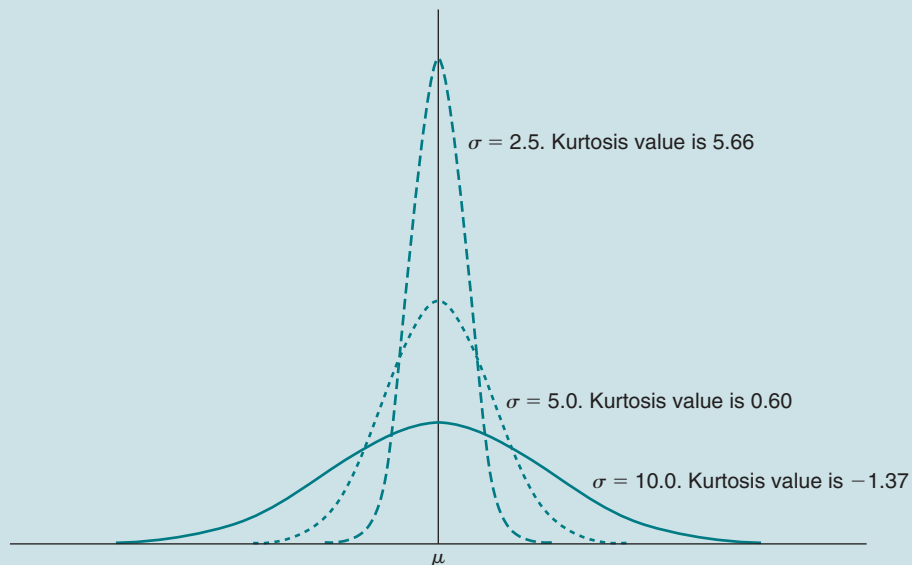
- No matter the values of the mean or the standard deviation, the area under the curve is always unity. This means that the area under the curve represents all or 100% of the data.
- About 68% (the exact value is 68.26%) of all the data falls within  $\pm 1$  standard deviations from the mean. This means that the boundary limits of this 68% of the data are  $\mu \pm \sigma$ .
- About 95% (the exact value is 95.44%) of all the data falls within  $\pm 2$  standard deviations from the mean. This means that the boundary limits of this 95% of the data are  $\mu \pm 2\sigma$ .
- Almost 100% (the exact value is 99.73%) of all the data falls within  $\pm 3$  standard deviations from the mean. This means that the boundary limits of this almost 100% of the data are  $\mu \pm 3\sigma$ .

## Effect of different means and/or different standard deviations

The mean measures the central tendency of the data, and the standard deviation measures its spread or dispersion. Datasets in a normal distribution may have the following configurations:

- The same mean, but different standard deviations as illustrated in Figure 5.2. Here there are three distributions with the same mean but with standard deviations of 2.50, 5.00, and 10.00 respectively. The smaller the standard deviation, here 2.50, the curve is narrower and the data congregates around the mean. The larger the standard deviation, here 10.0, the flatter is the curve and the deviation around the mean is greater.
- Different means but the same standard deviation as illustrated in Figure 5.3. Here the standard deviation is 10.00 for the three curves and their shape is identical. However their means

Figure 5.2 Normal distribution: the same mean but different standard deviations.



are  $-10$ ,  $0$ , and  $20$  so that they have different positions on the  $x$ -axis.

- Different means and also different standard deviations are illustrated in Figure 5.4. Here the flatter curve has a mean of  $-10.00$  and a standard deviation of  $10.00$ . The middle curve has a mean of  $0$  and a standard deviation of  $5.00$ . The sharper curve has a mean of  $20.00$  and a standard deviation of  $2.50$ .

In conclusion, the shape of the normal distribution is determined by its standard deviation, and the mean value establishes its position on the  $x$ -axis. As such, there is an infinite combination of curves according to their respective mean and standard deviation. However, a set of data can be uniquely defined by its mean and standard deviation.

## Kurtosis in frequency distributions

Since continuous distributions may have the same mean, but different standard deviations, the different standard deviations alter the sharpness or hump of the peak of the curve as illustrated by the three normal distributions given in Figure 5.2. This difference in shape is the **kurtosis**, or the characteristic of the peak of a frequency distribution curve.

The curve that has a small standard deviation,  $\sigma = 2.5$  is **leptokurtic** after the Greek word *lepto* meaning slender. The peak is sharp, and as shown in Figure 5.2, the kurtosis value is  $5.66$ . The curve that has a standard deviation,  $\sigma = 10.0$  is **platykurtic** after the Greek word *platy* meaning broad, or flat, and this flatness can be seen also in Figure 5.2. Here the kurtosis value is  $-1.37$ .

Figure 5.3 Normal distribution: the same standard deviation but different means.

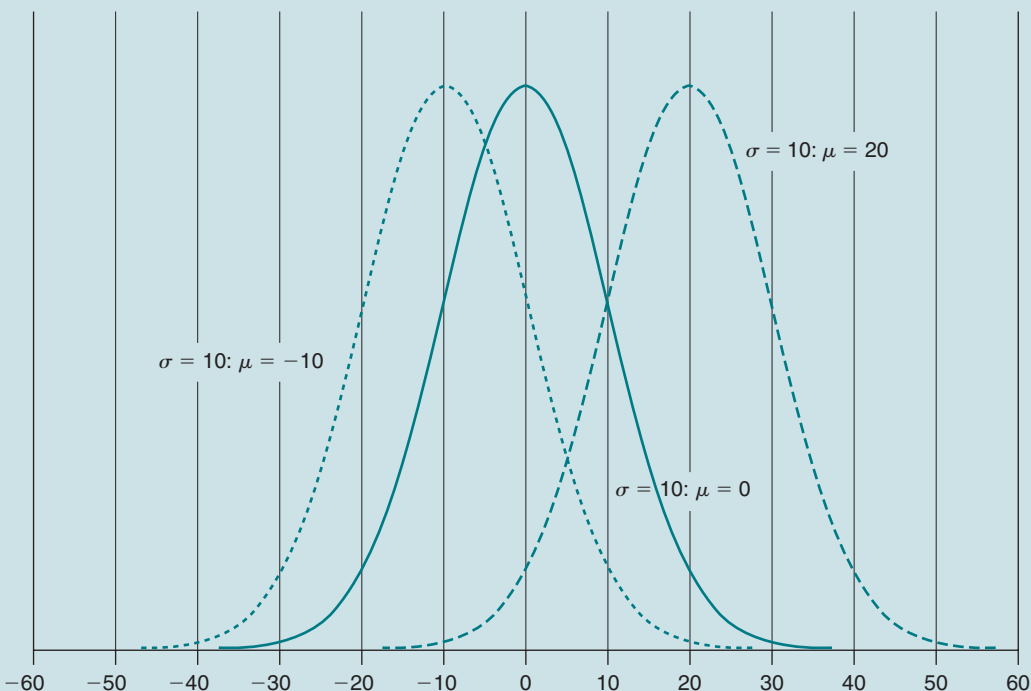
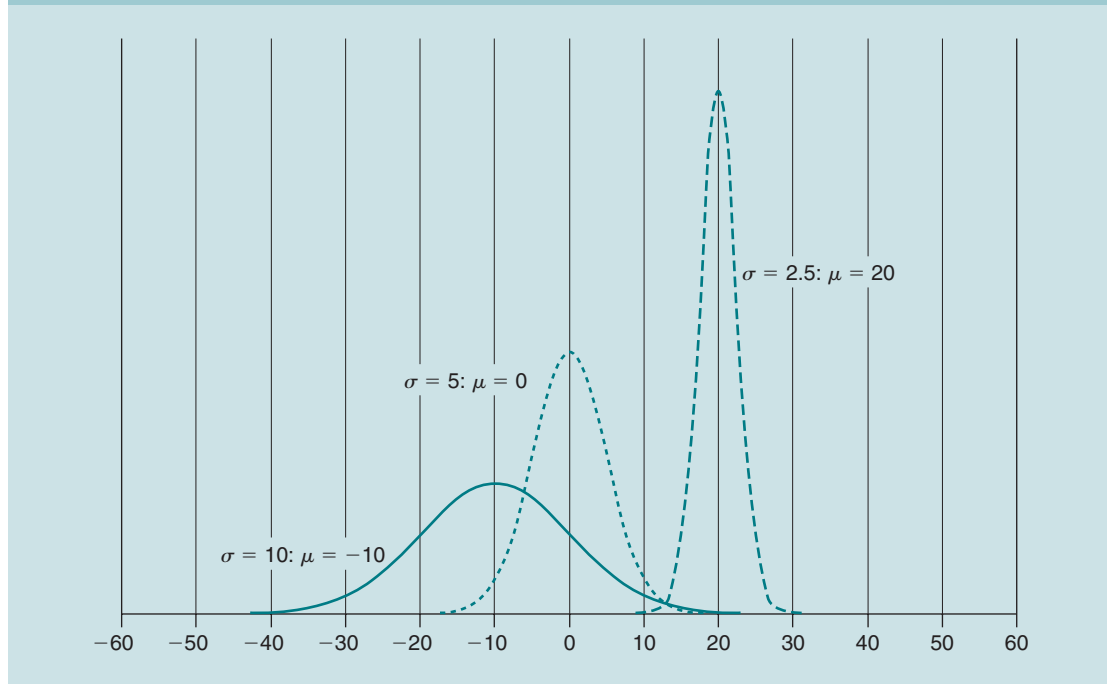


Figure 5.4 Normal distribution: different means and different standard deviations.



The intermediate curve where the standard deviation  $\sigma = 5.0$  is called **mesokurtic** since the peak of the curve is in between the two others. Meso from the Greek means intermediate. Here the kurtosis value is 0.60. In statistics, recording the kurtosis value of data gives a measure of the sharpness of the peak and as a corollary a measure of its dispersion. The kurtosis value of a relatively flat peak is negative, whereas for a sharp peak it is positive and becomes increasingly so with the sharpness. The importance of knowing these shapes is that a curve that is leptokurtic is more reliable for analytical purposes. The kurtosis value can be determined in Excel by using **[function KURT]**.

## Transformation of a normal distribution

Continuous datasets might be for example, the volume of beer in cans, the weight of chocolate bars, or the distance travelled by an automobile

tyre. In the normal distribution the units for these measurements for the mean and the standard deviation are different. There are centilitres for the beer, grams for the chocolate, or kilometres for the tyres. However, all these datasets can be transformed into a standard normal distribution using the following **normal distribution transformation relationship**:

$$z = \frac{x - \mu_x}{\sigma_x} \quad 5(ii)$$

- $x$  is the value of the random variable.
- $\mu_x$  is the mean of the distribution of the random variables.
- $\sigma_x$  is the standard deviation of the distribution.
- $z$  is the number of standard deviations from  $x$  to the mean of this distribution.

Since the numerator and the denominator (top and bottom parts of the equation) have the

same units, there are no units for the value of  $z$ . Further, since the value of  $x$  can be more, or less, than the mean value, then  $z$  can be either plus or minus. For example, for a certain format the mean value of beer in a can is 33 cl and from past data we know that the standard deviation of the bottling process is 0.50 cl. Assume that a single can of beer is taken at random from the bottling line and its volume is 33.75 cl. In this case using equation 5(ii),

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{33.75 - 33.00}{0.50} = \frac{0.75}{0.50} = 1.50$$

Alternatively, the mean value of a certain size chocolate bar is 100 g and from past data we know that the standard deviation of a production lot of these chocolate bars is 0.40 g. Assume one slab of chocolate is taken at random from the production line and its weight is 100.60 g. In this case using equation 5(ii),

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{100.60 - 100.00}{0.40} = \frac{0.60}{0.40} = 1.50$$

Again assume that the mean value of the life of a certain model tyre is 35,000 km and from past data we know that the standard deviation of the life of a tyre is 1,500 km. Then suppose that one tyre is taken at random from the production line and tested on a rolling machine. The tyre lasts 37,250 km. Then using equation 5(ii),

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{37,250 - 35,000}{1,500} = \frac{2,250}{1,500} = 1.50$$

Thus in each case we have the same number of standard deviations,  $z$ . This is as opposed to the value of the standard deviation,  $\sigma$ , in using three different situations each with different units. We have converted the data to a standard normal distribution. This is how the normal frequency distribution can be used to estimate the probability of occurrence of certain situations.

## The standard normal distribution

A **standard normal distribution** has a mean value,  $\mu$ , of zero. The area under the curve to the left of the mean is 50.00% and the area to the right of the mean is also 50.00%. For values of  $z$  ranging from  $-3.00$  to  $+3.00$  the area under the curve represents 99.73% or almost 100% of the data. When the values of  $z$  range from  $-2.00$  to  $+2.00$ , then the area under the curve represents 95.45%, or close to 95%, of the data. And, for values of  $z$  ranging from  $-1.00$  to  $+1.00$  the area under the curve represents 68.27% or about 68% of the data. These relationships are illustrated in Figure 5.5. These areas of the curve are indicated with the appropriate values of  $z$  on the  $x$ -axis. Also, indicated on the  $x$ -axis are the values of the random variable,  $x$ , for the case of a bar of chocolate of a nominal weight of 100.00 g, and a population standard deviation of 0.40 g, as presented earlier. These values of  $x$  are determined as follows.

Reorganizing equation 5(ii) to make  $x$  the subject, we have,

$$x = \mu_x + z\sigma_x \quad 5(\text{iii})$$

Thus, when  $z$  is 2 the value of  $x$  from equation 5(iii) is,

$$x = 100.00 + 2 * 0.4 = 100.80$$

Alternatively, when  $z$  is  $-3$  the value of  $x$  from equation 5(iii) is,

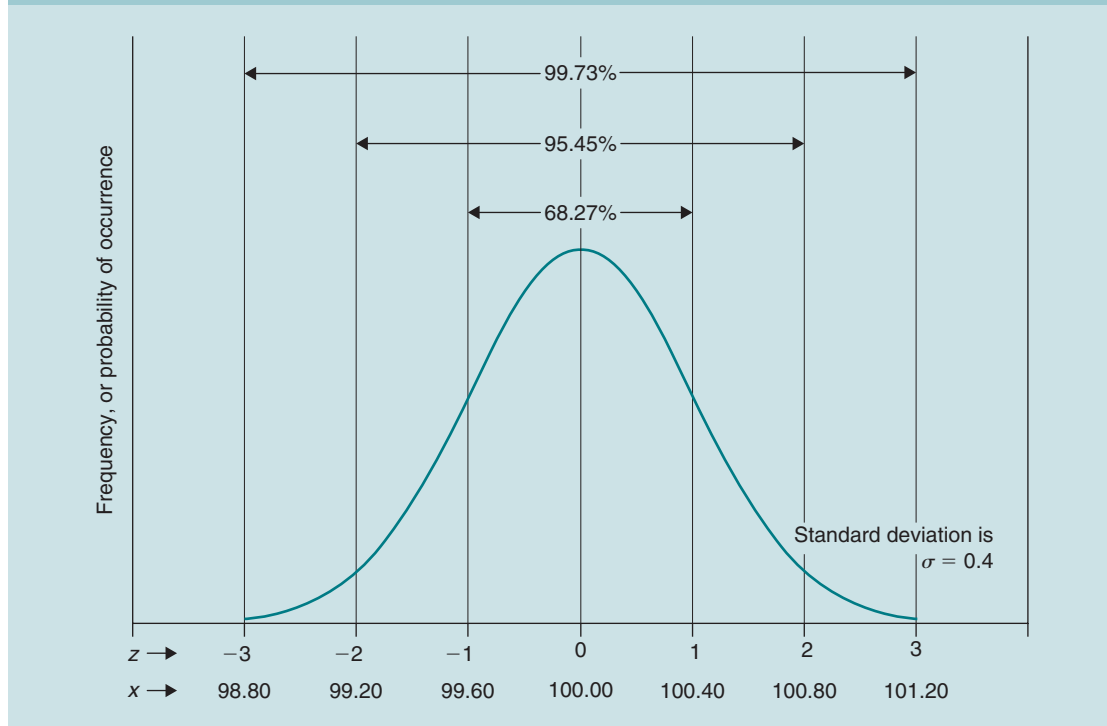
$$x = 100.00 + (-3) * 0.4 = 98.80$$

The other values of  $x$  are calculated in a similar manner.

Note the value of  $z$  is not necessarily a whole number but it can take on any numerical value such as  $-0.45$ ,  $0.78$ , or  $2.35$ , which give areas under the curve from the left-hand tail to the  $z$ -value of 32.64%, 78.23%, and 99.06%, respectively. When  $z$  is negative it means that the area under the curve from the left is less than 50% and when  $z$  is positive it means that the area from the



Figure 5.5 Areas under a standard normal distribution.



left of the curve is greater than 50%. These area values can also be interpreted as probabilities. Thus for any data of any continuous units such as weight, volume, speed, length, etc. all intervals containing the same number of standard deviations,  $z$  from the mean, will contain the same proportion of the total area under the curve for any normal probability distribution.

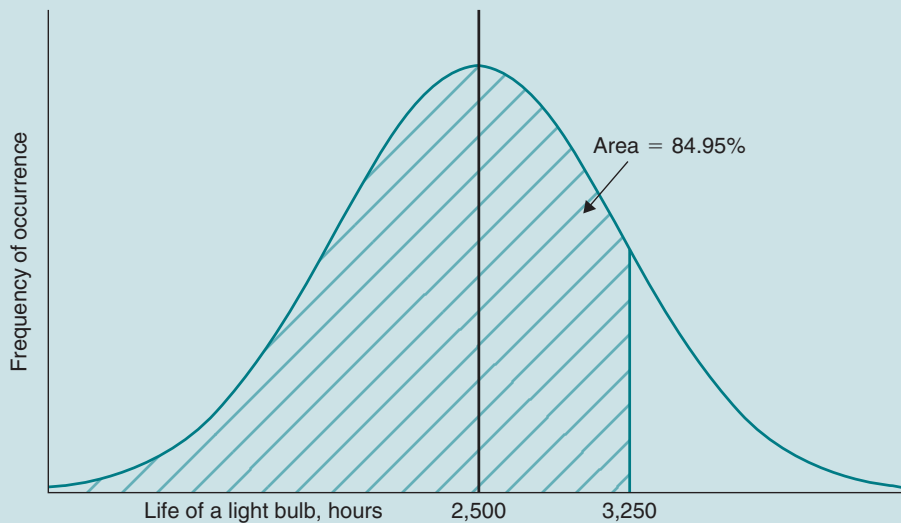
### Determining the value of $z$ and the Excel function

Many books on statistics and quantitative methods publish standard tables for determining  $z$ . These tables give the area of the curve either to the right or the left side of the mean and from these tables probabilities can be estimated. Instead of tables, this book uses the Microsoft Excel function for the normal distribution,

which has a complete database of the  $z$ -values. The logic of the  $z$ -values in Excel is that the area of the curve increases from 0% at the left to 100% as we move to the right of the curve. The following four useful normal distribution functions are found in Excel.

- **[function NORMDIST]** determines the area under the curve, or probability  $P(x)$ , given the value of the random variable  $x$ , the mean value,  $\mu$ , of the dataset, and the standard deviation,  $\sigma$ .
- **[function NORMINV]** determines the value of the random variable,  $x$ , given the area under the curve or the probability,  $P(x)$ , the mean value,  $\mu$ , and the standard deviation,  $\sigma$ .
- **[function NORMSDIST]** the value of the area or probability,  $p$ , given  $z$ .
- **[function NORMSINV]** the value of  $z$  given the area or probability,  $P(x)$ .

Figure 5.6 Probability that the life of a light bulb lasts no more than 3,250 hours.



It is not necessary to learn by heart which function to use because, as for all the Excel functions, when they are selected, it indicates what values to insert to obtain the result. Thus, knowing the information that you have available, tells you what normal function to use. The application of the normal distribution using the Excel normal distribution function is illustrated in the following example.

### Application of the normal distribution: *Light bulbs*

General Electric Company has past data concerning the life of a particular 100-Watt light bulb that shows that on average it will last 2,500 hours before it fails. The standard deviation of this data is 725 hours and the illumination time of a light bulb is considered to follow a normal distribution. Thus for this situation, the mean value,  $\mu$ , is considered a constant at 2,500 hours and the standard deviation,  $\sigma$ , is also a constant with a value of 725 hours.

1. What is the probability that a light bulb of this kind selected at random from the production line will last no more than 3,250 hours?

Using equation 5(ii), where the random variable,  $x$ , is 3,250,

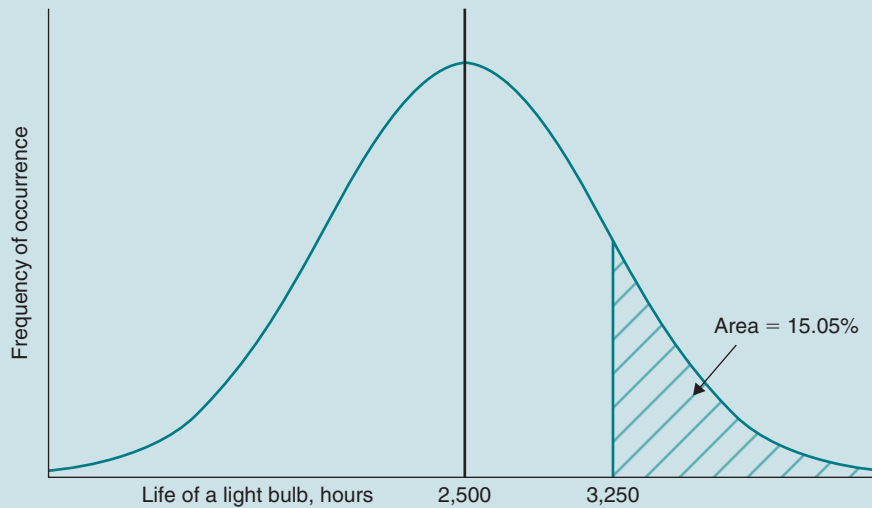
$$z = \frac{3,250 - 2,500}{725} = \frac{750}{725} = 1.0345$$

From [function NORMSDIST] the area under the curve from left to right, for  $z = 1.0345$ , is 84.95%. Thus we can say that the probability of a single light bulb taken from the production line has an 84.95% probability of lasting not more than 3,250 hours. This concept is shown on the normal distribution in Figure 5.6.

2. What is the probability that a light bulb of this kind selected at random from the production line will last at least 3,250 hours?

Here we are interested in the area of the curve on the right where  $x$  is at least 3,250 hours. This area is  $(100\% - 84.95\%)$  or 15.05%. Thus we can say that there is a 15.05%

Figure 5.7 Probability that the life of a light bulb lasts at least 3,250 hours.



probability that a single light bulb taken from the production line has a 15.05% probability of lasting at least 3,250 hours. This is shown on the normal distribution in Figure 5.7.

3. What is the probability that a light bulb of this kind selected at random will last no more than 2,000 hours?

Using equation 5(ii), where the random variable,  $x$ , is now 2,000 hours,

$$z = \frac{2,000 - 2,500}{725} = -\frac{500}{725} = -0.6897$$

The fact that  $z$  has a negative value implies that the random variable lies to the left of the mean; which it does since 2,000 hour is less than 2,500 hours. From [function NORMS-DIST] the area of the curve for  $z = -0.6897$  is 24.52%. Thus, we can say that there is a 24.52% probability that a single light bulb taken from the production line will last no more than 2,000 hours. This is shown on the normal distribution curve in Figure 5.8.

4. What is the probability that a light bulb of this kind selected at random will last between 2,000 and 3,250 hours?

In this case we are interested in the area of the curve between 2,000 hours and 3,250 hours where 2,000 hours is to the left of the mean and 3,250 is greater than the mean. We can determine this probability by several methods.

#### Method 1

- Area of the curve 2,000 hours and below is 24.52% from answer to Question 3.
- Area of the curve 3,250 hours and above is 15.05% from answer to Question 2.

Thus, area between 2,000 and 3,250 hours is  $(100.00\% - 24.52\% - 15.05\%) = 60.43\%$ .

#### Method 2

Since the normal distribution is symmetrical, the area of the curve to the left of the mean is 50.00% and also the area of the curve to the right of the mean is 50.00%. Thus,

- Area of the curve between 2,000 and 2,500 hours is  $(50.00\% - 24.52\%) = 25.48\%$ .
- Area of the curve between 2,500 and 3,250 hours is  $(50.00\% - 15.05\%) = 34.95\%$ .

Figure 5.8 Probability that the life of a light bulb lasts no more than 2,000 hours.

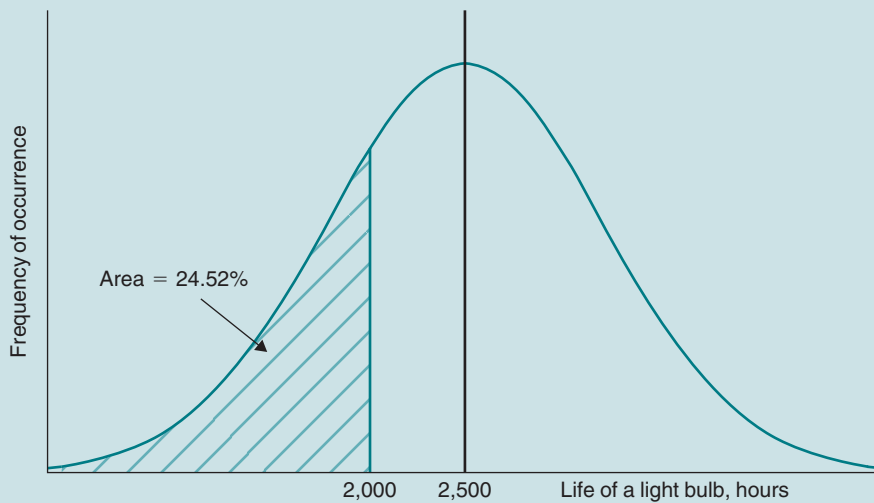
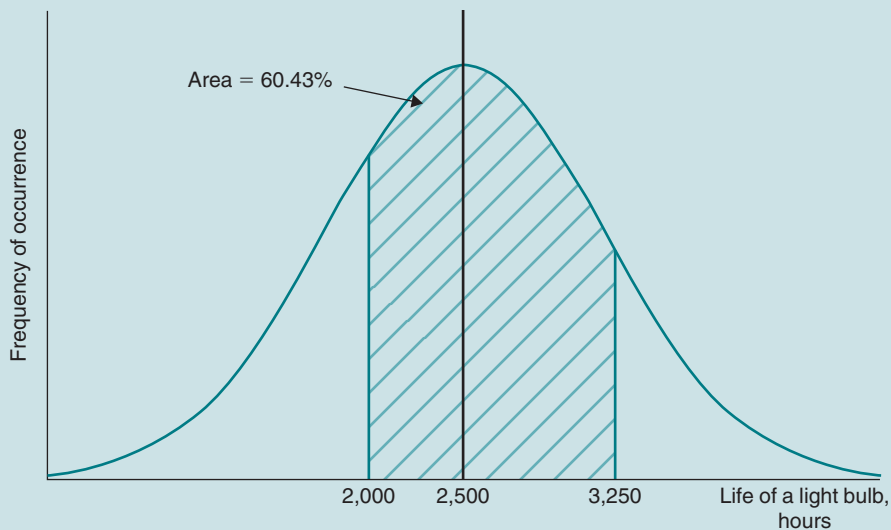


Figure 5.9 Probability that the light bulb lasts between 2,000 and 3,250 hours.



Thus, area of the curve between 2,000 and 3,250 hours is  $(25.48\% + 34.95\%) = 60.43\%$ .

#### Method 3

- Area of the curve at 3,250 hours and below is 84.95%.

- Area of the curve at 2,000 hours and below is 24.52%.

Thus, area of the curve between 2,000 and 3,250 hours is  $(84.95\% - 24.52\%) = 60.43\%$ .

This situation is shown on the normal distribution curve in Figure 5.9.

5. What are the lower and upper limits in hours, symmetrically distributed, at which 75% of the light bulbs will last?

In this case we are interested in 75% of the middle area of the curve. The area of the curve outside this value is  $(100.00\% - 75.00\%) = 25.00\%$ . Since the normal distribution is symmetrical, the area on the left side of the limit, or the left tail, is  $25/2$  or  $12.50\%$ . Similarly, the area on the right of the limit, or the right tail, is also  $12.50\%$  as illustrated in Figure 5.10.

From the normal probability functions in Excel, given the value of  $12.50\%$ , then the numerical value of  $z$  is  $1.1503$ . Again, since the curve is symmetrical the value of  $z$  on the left side is  $-1.1503$  and on the right side, it is  $+1.1503$ .

From equation 5(iii) where  $z$  at the upper limit is  $1.1503$ ,  $\mu_x = 2,500$  and  $\sigma_x$  is  $725$ ,

$$\begin{aligned} x(\text{upper limit}) &= 2,500 + 1.1503 * 725 \\ &= 3,334 \text{ hours} \end{aligned}$$

At the lower limit  $z$  is  $-1.1503$

$$\begin{aligned} x(\text{lower limit}) &= 2,500 - 1.1503 * 725 \\ &= 1,666 \text{ hours} \end{aligned}$$

These values are also shown on the normal distribution curve in Figure 5.10.

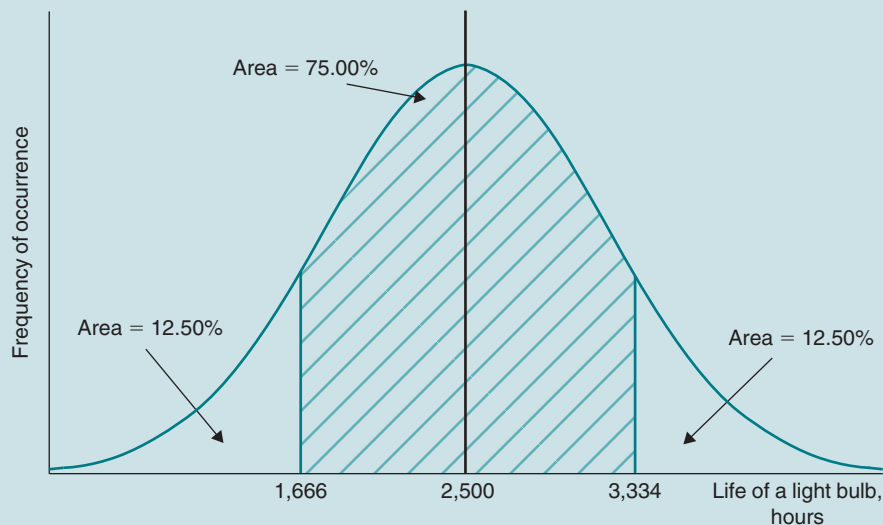
6. If General Electric has 50,000 of this particular light bulb in stock, how many bulbs would be expected to fail at 3,250 hours or less?

In this case we simply multiply the population  $N$ , or 50,000 by the area under the curve by the answer determined in Question No. 1, or  $50,000 * 84.95\% = 42,477.24$  or 42,477 light bulbs rounded to the nearest whole number.

7. If General Electric has 50,000 of this particular light bulb in stock, how many bulbs would be expected to fail between 2,000 and 3,250 hours?

Again, we multiply the population  $N$ , or 50,000, by the area under the curve determined by the answer to Question No. 4, or  $50,000 * 60.43\% = 30,216.96$  or 30,217 light bulbs rounded to the nearest whole number.

Figure 5.10 Symmetrical limits between which 75% of the light bulbs will last.



In all these calculations we have determined the appropriate value by first determining the value of  $z$ . A quicker route in Excel is to use the **[function NORMDIST]** where the mean, standard deviation, and the value of  $x$  are entered. This gives the probability directly. It is a matter of preference which of the functions to use. I like to calculate  $z$ , since with this value it is easy to position the situation on the normal distribution curve.

## Demonstrating That Data Follow a Normal Distribution

A lot of data follows a normal distribution particularly when derived from an operation set to a nominal value. The weight of a nominal 100-g chocolate bar, the volume of liquid in a nominal 33-cl beverage can, or the life of a tyre mentioned earlier follow a normal distribution. Some of the units examined will have values greater than the nominal figure and some less. However, there may be cases when other data may not follow a normal distribution and so if you apply the normal distribution assumptions erroneous conclusions may be made.

## Verification of normality

To verify that data reasonably follows a normal distribution you can make a visual comparison. For small datasets a stem-and-leaf display as presented in Chapter 1, will show if the data appears normal. For larger datasets a **frequency polygon** also developed in Chapter 1 or a box-and-whisker plot, introduced in Chapter 2, can be developed to see if their profiles look normal. As an illustration, Figure 5.11 shows a frequency polygon and the box-and-whisker plot for the sales revenue data presented in Chapters 1 and 2.

Another verification of the normal assumption is to determine the properties of the dataset to see if they correspond to the normal distribution

criteria. If they do then the following relationships should be close.

- The mean is equal to the median value.
- The inter-quartile range is equal to 1.33 times the standard deviation.
- The range of the data is equal to six times the standard deviation.
- About 68% of the data lies between  $\pm 1$  standard deviations of the mean.
- About 95% of the data lies between  $\pm 2$  standard deviations of the mean.
- About 100% of the data lies between  $\pm 3$  standard deviations of the mean.

The information in Table 5.1 gives the properties for the 200 pieces of sales data presented in Chapter 1. The percentage values are calculated by using the equation 5(iii) first to find the limits for a given value of  $z$  using the mean and standard deviation of the data. Then the amount of data between these limits is determined and this

Figure 5.11 Sales revenue: comparison of the frequency polygon and its box-and-whisker plot.

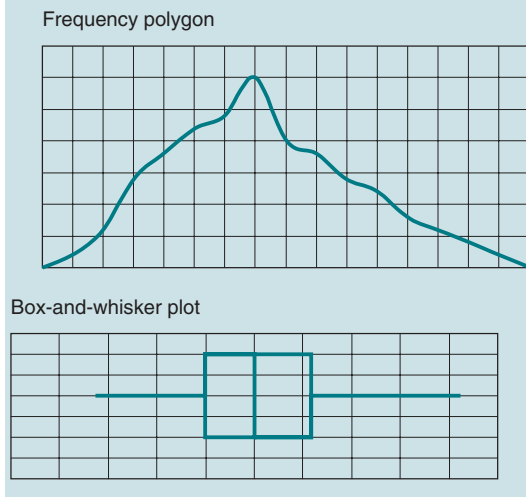


Table 5.1 Sales revenues: properties compared to normal assumptions.

35,378	170,569	104,985	134,859	120,958	107,865	127,895	106,825	130,564	108,654	Property	Value
109,785	184,957	96,598	121,985	63,258	164,295	97,568	165,298	113,985	124,965	Mean	102,666.67
108,695	91,864	120,598	47,865	162,985	83,964	103,985	61,298	104,987	184,562	Median	100,295.50
89,597	160,259	55,492	152,698	92,875	56,879	151,895	88,479	165,698	89,486	Maximum	184,957.00
85,479	64,578	103,985	81,980	137,859	126,987	102,987	116,985	45,189	131,958	Minimum	35,378.00
73,598	161,895	132,689	120,654	67,895	87,653	58,975	103,958	124,598	168,592	Range	149,579.00
95,896	52,754	114,985	62,598	145,985	99,654	76,589	113,590	80,459	111,489	$\sigma$ (population)	30,888.20
109,856	101,894	80,157	78,598	86,785	97,562	136,984	89,856	96,215	163,985	$Q_3$	123,910.75
83,695	75,894	98,759	133,958	74,895	37,856	90,689	64,189	107,865	123,958	$Q_1$	79,975.75
105,987	93,832	58,975	102,986	102,987	144,985	101,498	101,298	103,958	71,589	$Q_3 - Q_1$	43,935.00
59,326	121,459	82,198	60,128	86,597	91,786	56,897	112,854	54,128	152,654	$6\sigma$	185,329.17
99,999	78,562	110,489	86,957	99,486	132,569	134,987	76,589	135,698	118,654	$1.33\sigma$	41,081.30
90,598	156,982	87,694	117,895	85,632	104,598	77,654	105,987	78,456	149,562	Normal plot	Area under curve
68,976	50,128	106,598	63,598	123,564	47,895	100,295	60,128	141,298	84,598	$\pm 1\sigma$	68.27%
100,296	77,498	77,856	134,890	79,432	100,659	95,489	122,958	111,897	129,564	$\pm 2\sigma$	95.45%
71,458	88,796	110,259	72,598	140,598	125,489	69,584	89,651	70,598	93,876	$\pm 3\sigma$	99.73%
112,987	123,895	65,847	128,695	66,897	82,459	133,984	98,459	153,298	87,265	Sales data	Area under curve
72,312	81,456	124,856	101,487	73,569	138,695	74,583	136,958	115,897	142,985	$\pm 1\sigma$	64.50%
119,654	96,592	66,598	81,490	139,584	82,456	150,298	106,859	68,945	122,654	$\pm 2\sigma$	96.00%
70,489	94,587	85,975	138,597	97,498	143,985	92,489	146,289	84,592	69,874	$\pm 3\sigma$	100.00%

is converted to a percentage amount. The following gives an example of the calculation.

$$x \text{ (for } z = -1) = 102,667 - 30,880 = 71,787$$

$$x \text{ (for } z = +1) = 102,667 + 30,880 = 133,547$$

Using Excel, there are 129 pieces of data between these limits and so  $129/200 = 64.50\%$

$$\begin{aligned} x \text{ (for } z = -2) &= 102,667 - 2 * 30,880 \\ &= 40,907 \end{aligned}$$

$$\begin{aligned} x \text{ (for } z = +2) &= 102,667 + 2 * 30,880 \\ &= 164,427 \end{aligned}$$

Using Excel, there are 192 pieces of data between these limits and  $192/200 = 96.00\%$

$$\begin{aligned} x \text{ (for } z = -3) &= 102,667 - 3 * 30,880 \\ &= 10,027 \end{aligned}$$

$$\begin{aligned} x \text{ (for } z = +3) &= 102,667 + 3 * 30,880 \\ &= 19,5307 \end{aligned}$$

Using Excel, there are 200 pieces of data between these limits and  $200/200 = 100.00\%$

Thus from the visual displays, and the properties of the sales data, the normal assumption seems reasonable. As a proof of this, if we go back to Chapter 1 from the ogives for this sales data we showed that,

- From the greater than ogive, 80.00% of the sales revenues are at least \$75,000.
- From the less than ogive, 90.00% of the revenues are no more than \$145,000.

If we assume a normal distribution then at least 80% of the sales revenue will appear in the area of the curve as illustrated in Figure 5.12. The value of  $z$  at the point  $x$  with the Excel normal distribution function is  $-0.8416$ . Using this and the mean, and standard deviation values for the sales data using equation 5(iii) we have,

$$x = 102,667 + (-0.8416) * 30,880 = \$76,678$$

This value is only 2.2% greater than the value of \$75,000 determined from the ogive.

Figure 5.12 Area of the normal distribution containing at least 80% of the data.

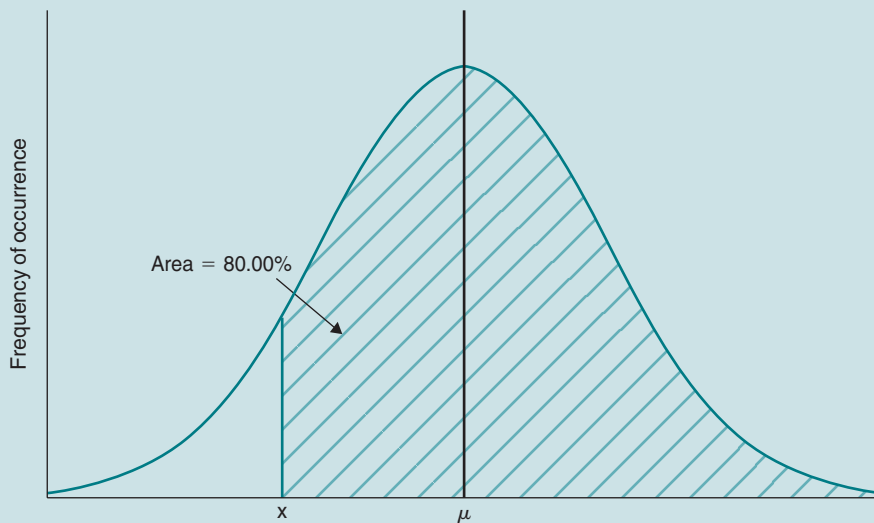
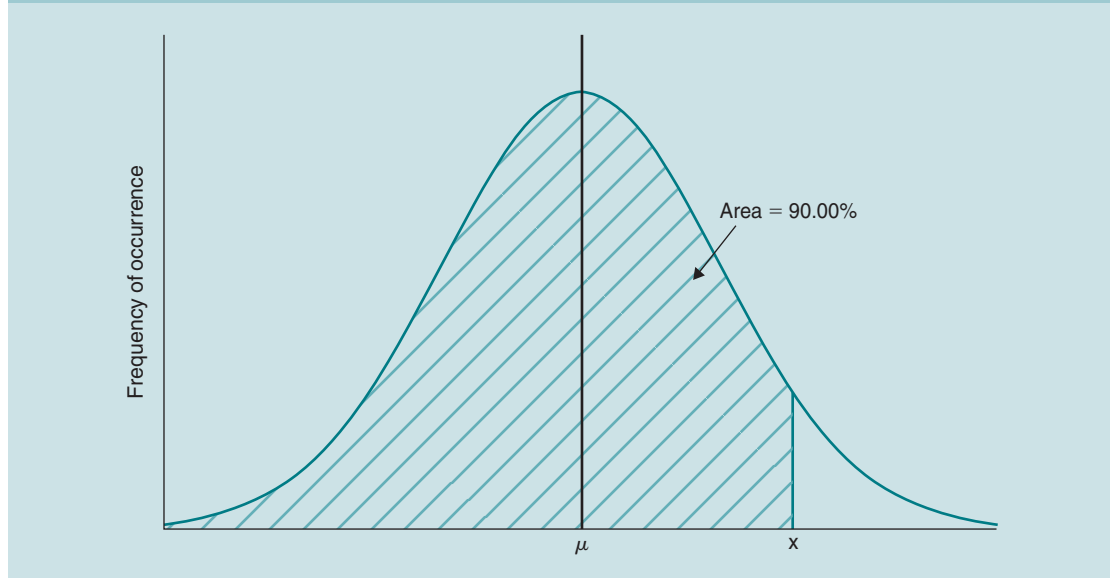




Figure 5.13 Area of the normal distribution giving upper limit of 90% of the data.



Similarly, if we assume a normal distribution then 90% of the sales revenue will appear in the area of the curve as illustrated in Figure 5.13. The value of  $z$  at the point  $x$  with the Excel normal distribution function is +1.2816. Using this and the mean, and standard deviation values for the sales data using equation 5(iii) we have,

$$x = 102,667 + 1.2816 * 30,880 = \$142,243$$

This value is only 1.9% less than the value of \$145,000 determined from the ogive.

## Asymmetrical data

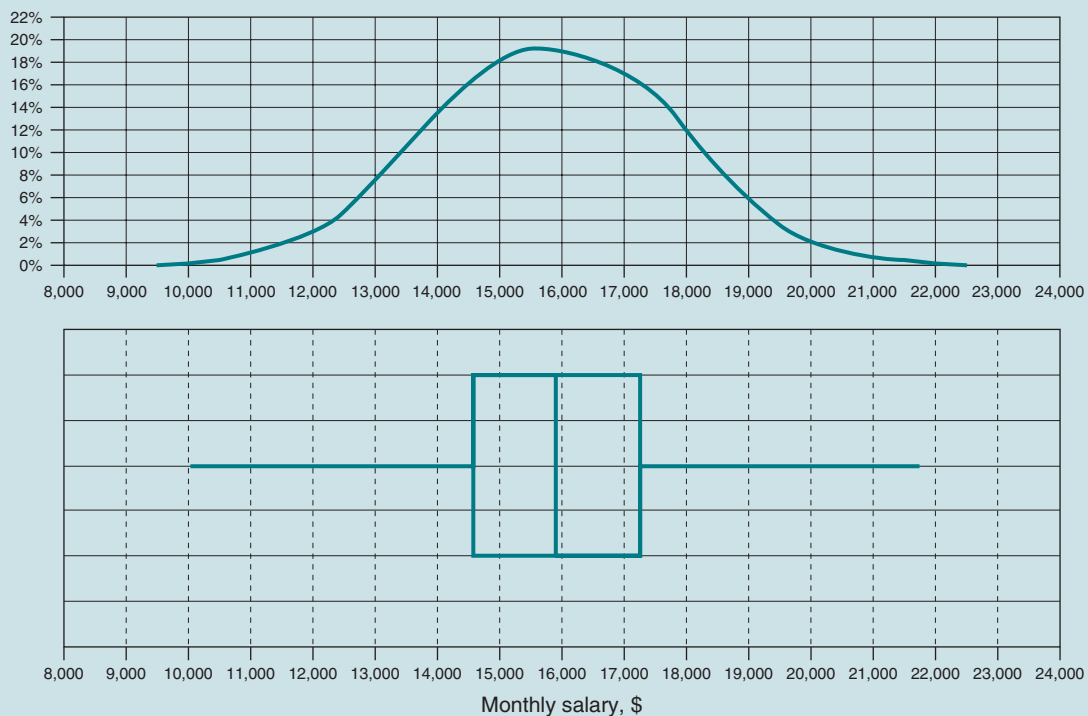
In a dataset when the mean and median are significantly different then the probability distribution is not normal but is asymmetrical or **skewed**. A distribution is skewed because values in the frequency plot are concentrated at either the low (left side) or the high end (right side) of the  $x$ -axis. When the mean value of the dataset is greater than the median value then the distribution of the data is positively or **right-skewed** where the curve tails off to the right. This is because it is the mean that is the most affected by extreme values and is pulled over to the right.

Here the distribution of the data has its mode, the hump, or the highest frequency of occurrence, at the left end of the  $x$ -axis where there is a higher proportion of relatively low values and a lower proportion of high values. The median is the middle value and lies between the mode and the mean.

If the mean value is less than the median, then the data is negatively or **left-skewed** such that the curve tails off to the left. This is because it is the mean that is the most affected by extreme values and is pulled back to the left. Here the distribution of the data has its mode, the hump, or the highest frequency of occurrence, at the right end of the  $x$ -axis where there is a higher proportion of large values and lower proportion of relatively small values. Again, the median is the middle value and lies between the mode and the mean.

This concept of symmetry and asymmetry is illustrated by the following three situations. For a certain consulting Firm A, the monthly salaries of 1,000 of its worldwide staff are shown by the frequency polygon and its associated box-and-whisker plot in Figure 5.14. Here

Figure 5.14 Frequency polygon and its box-and-whisker plot for symmetrical data.



the data is essentially symmetrically distributed. The mean value is \$15,893 and the median value is \$15,907 or the mean is just 0.08% less than the median. The maximum salary is \$21,752 and the minimum is \$10,036. Thus, 500, or 50% of the staff have a monthly salary between \$10,036 and \$15,907 and 500, or the other 50%, have a salary between \$15,907 and \$21,752. From the graph the mode is about \$15,800 with the frequency at about 19.2% or essentially the mean, mode, and median are approximately the same.

Figure 5.15 is for consulting Firm B. Here the frequency polygon and the box-and-whisker plot are right-skewed. The mean value is now \$12,964 and the median value is \$12,179 or the mean is 6.45% greater than the median. The maximum salary is still \$21,752 and the minimum \$10,036. Now, 500, or 50%, of the staff

have a monthly salary between \$10,036 and \$12,179 and 500, or the other 50%, have a salary between \$12,179 and \$21,752 or a larger range of smaller values than in the case of the symmetrical distribution, which explains the lower average value. From the graph the mode is about \$11,500 with the frequency at about 24.0%. Thus in ascending order, we have the mode (\$11,500), median (\$12,179), and mean (\$12,964).

Figure 5.16 is for consulting Firm C. Here the frequency polygon and the box-and-whisker plot are left-skewed. The mean value is now \$18,207 and the median value is \$19,001 or the mean is 4.18% less than the median. The maximum salary is still \$21,752 and the minimum \$10,036. Now, 500, or 50%, of the staff have a monthly salary between \$10,036 and \$19,001 and 500, or the other 50%, have a salary between

Figure 5.15 Frequency polygon and its box-and-whisker plot for right-skewed data.

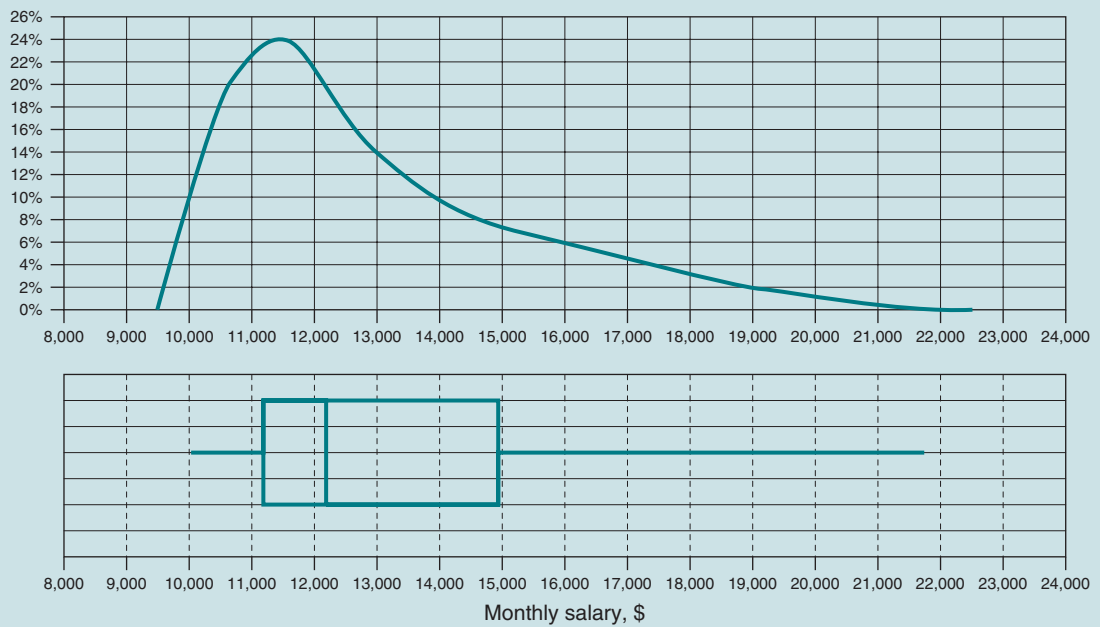
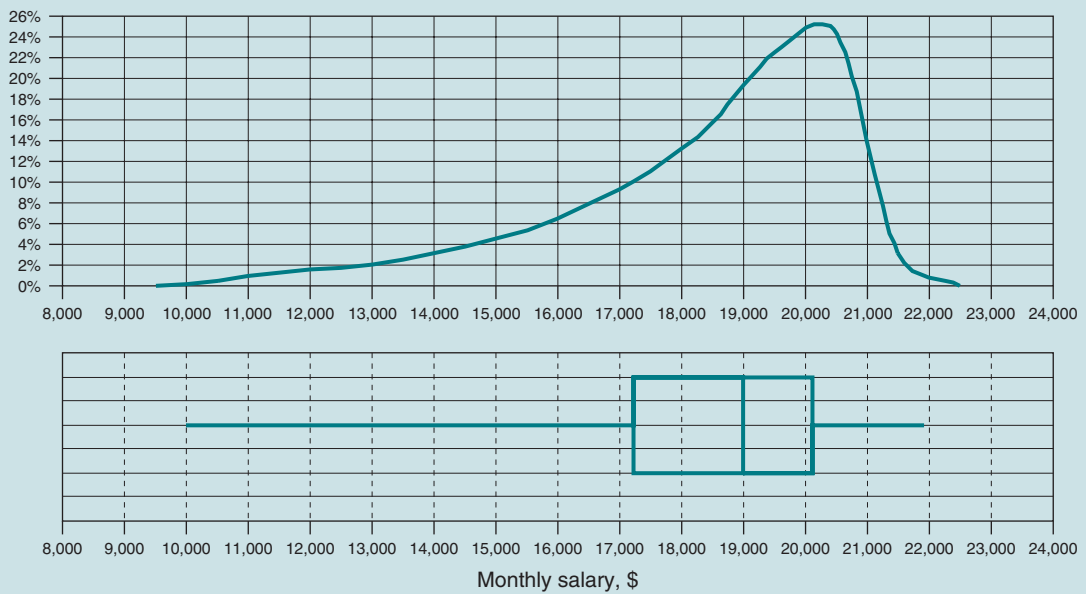


Figure 5.16 Frequency polygon and its box-and-whisker plot for left-skewed data.



\$19,001 and \$21,752 or a smaller range of upper values compared to the symmetrical distribution, which explains the higher mean value. From the graph the mode is about \$20,500 with the frequency at about 24.30%. Thus in ascending order we have the mean (\$18,207), median (\$19,001), and the mode (\$20,500).

## Testing symmetry by a normal probability plot

Another way to establish the symmetry of data is to construct a normal probability plot. This procedure is as follows:

- Organize the data into an ordered data array.
- For each of the data points determine the area under the curve on the assumption that the data follows a normal distribution. For example, if there are 19 data points in the array then the curve has 20 portions. (To divide a segment into  $n$  portions you need  $(n - 1)$  limits.)
- Determine the number of standard deviations,  $z$ , for each area using that normal distribution function in Excel, which gives  $z$  for a given probability. For example, for 19 data values Table 5.2 gives the area under the curve and the corresponding value of  $z$ . Note that the value of  $z$  has the same numerical values moving from left to right and at the median,  $z$  is 0 since this is a standardized normal distribution.
- Plot the data values on the  $y$ -axis against the  $z$ -values on the  $x$ -axis.
- Observe the profile of the graph. If the graph is essentially a straight line with a positive slope then the data follows a normal distribution. If the graph is non-linear of a concave format then the data is right-skewed. If the graph has a convex format then the data is left-skewed.

The three normal probability plots that show clearly the profiles for the normal, right-, and left-skewed datasets for the consulting data of Figures 5.14–5.16 are shown in Figure 5.17.

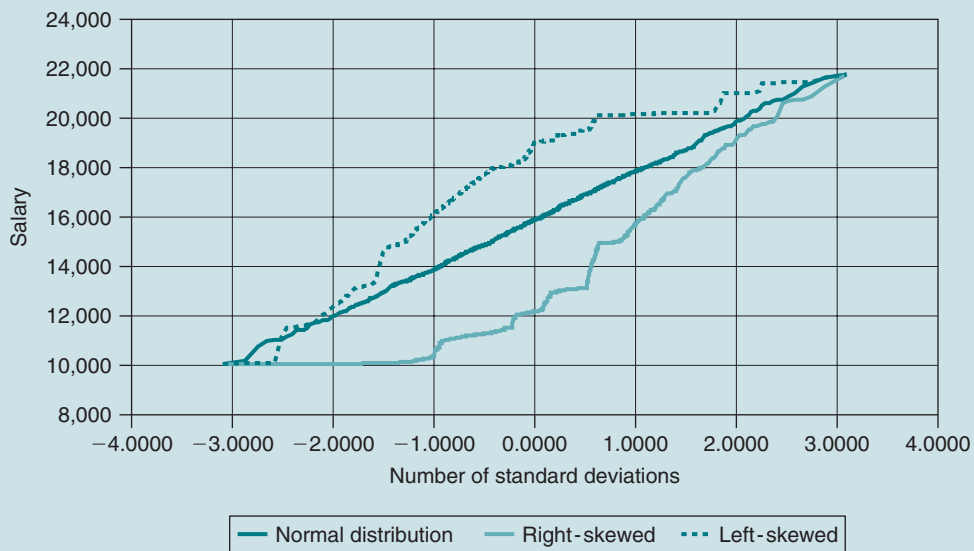
**Table 5.2** Symmetry by a normal probability plot.

Data point	Area to left of data point (%)	No. of standard deviations at data point
1	5.00	−1.6449
2	10.00	−1.2816
3	15.00	−1.0364
4	20.00	−0.8416
5	25.00	−0.6745
6	30.00	−0.5244
7	35.00	−0.3853
8	40.00	−0.2533
9	45.00	−0.1257
10	50.00	0.0000
11	55.00	0.1257
12	60.00	0.2533
13	65.00	0.3853
14	70.00	0.5244
15	75.00	0.6745
16	80.00	0.8416
17	85.00	1.0364
18	90.00	1.2816
19	95.00	1.6449

## Percentiles and the number of standard deviations

In Chapter 2, we used percentiles to divide up the raw sales data originally presented in Figure 1.1 and then to position regional sales information according to its percentile value. Using the concept from the immediate previous paragraph, “Testing symmetry by a normal probability plot”, we can relate the percentile value and the number of standard deviations. In Table 5.3, in the column “ $z$ ”, we show the number of standard deviations going from  $-3.4$  to  $+3.4$  standard deviations. The next column, “percentile” gives the area to the left of this number of standard deviations, which is also the percentile value on the basis the data follows a normal distribution, which we have demonstrated in the paragraph “Demonstrating that data follow a normal distribution” in this chapter. The third

Figure 5.17 Normal probability plot for salaries.

Table 5.3 Positioning of sales data, according to  $z$  and the percentile.

$z$	Percentile (%)	Value (\$)	$z$	Percentile (%)	Value (\$)	$z$	Percentile (%)	Value (\$)
−3.40	0.0337	35,544	−1.10	13.5666	68,976	1.20	88.4930	141,469
−3.30	0.0483	35,616	−1.00	15.8655	71,090	1.30	90.3200	145,722
−3.20	0.0687	35,717	−0.90	18.4060	73,587	1.40	91.9243	150,246
−3.10	0.0968	35,855	−0.80	21.1855	76,734	1.50	93.3193	152,685
−3.00	0.1350	36,044	−0.70	24.1964	78,724	1.60	94.5201	157,293
−2.90	0.1866	36,298	−0.60	27.4253	82,106	1.70	95.5435	162,038
−2.80	0.2555	36,638	−0.50	30.8538	84,949	1.80	96.4070	163,835
−2.70	0.3467	37,088	−0.40	34.4578	87,487	1.90	97.1283	164,581
−2.60	0.4661	37,677	−0.30	38.2089	89,882	2.00	97.7250	165,487
−2.50	0.6210	39,585	−0.20	42.0740	93,864	2.10	98.2136	166,986
−2.40	0.8198	42,485	−0.10	46.0172	97,535	2.20	98.6097	169,053
−2.30	1.0724	45,548	0.00	50.0000	100,296	2.30	98.9276	170,304
−2.20	1.3903	47,241	0.10	53.9828	102,987	2.40	99.1802	175,728
−2.10	1.7864	47,882	0.20	57.9260	105,260	2.50	99.3790	181,264
−2.00	2.2750	49,072	0.30	61.7911	108,626	2.60	99.5339	184,591
−1.90	2.8717	52,005	0.40	65.5422	112,307	2.70	99.6533	184,684
−1.80	3.5930	54,333	0.50	69.1462	117,532	2.80	99.7445	184,756
−1.70	4.4565	56,697	0.60	72.5747	121,682	2.90	99.8134	184,810
−1.60	5.4799	58,778	0.70	75.8036	124,502	3.00	99.8650	184,851
−1.50	6.6807	59,562	0.80	78.8145	128,568	3.10	99.9032	184,881
−1.40	8.0757	61,390	0.90	81.5940	133,161	3.20	99.9313	184,903
−1.30	9.6800	63,754	1.00	84.1345	135,291	3.30	99.9517	184,919
−1.20	11.5070	66,522	1.10	86.4334	138,597	3.40	99.9663	184,931

column, “Value, \$” is the sales amount corresponding to the number of standard deviations and also the percentile.

What does all these mean? From Table 5.1 the standard deviation,  $z = 1$ , for this sales data is \$30,888.20 (let's say \$31 thousand) and the mean 102,666.67 (let's say \$103 thousand). Thus if sales are +1 standard deviations from the mean they would be approximately  $103 + 31 = \$134$  thousand. From Table 5.3 the value is \$135 thousand (rounding), or a negligible difference. Similarly a value of  $z = -1$  puts the sales at  $103 - 31 = \$72$  thousand. From Table 5.3 the value is \$71 thousand which again is close. Thus, using the standard  $z$ -values we have a measure of the dispersion of the data. This is another way of looking at the spread of information.

### Using a Normal Distribution to Approximate a Binomial Distribution

In Chapter 4, we presented the binomial distribution. Under certain conditions, the discrete binomial distribution can be approximated by the continuous normal distribution, enabling us to perform sampling experiments for discrete data but using the more convenient normal distribution for analysis. This is particularly useful for example in statistical process control (SPC).

### Conditions for approximating the binomial distribution

The conditions for approximating the binomial distribution are that the product of the sample size,  $n$ , and the probability of success,  $p$ , is greater, or equal to five and at the same time the product of the sample size and the probability of failure is also greater than or equal to five. That is,

$$np \geq 5 \quad 5(\text{iv})$$

$$n(1 - p) \geq 5 \quad 5(\text{v})$$

From Chapter 4, equation 4(xv), the mean or expected value of the binomial distribution is,

$$\mu_x = E(x) = np$$

And from equation 4(xvii) the standard deviation of the binomial distribution is given by,

$$\sigma = \sqrt{\sigma^2} = \sqrt{np(1 - p)} = \sqrt{npq}$$

When the two normal approximation conditions apply, then from using equation 5(ii) substituting for the mean and standard deviation we have the following **normal-binomial approximation**:

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{x - np}{\sqrt{npq}} = \frac{x - np}{\sqrt{np(1 - p)}} \quad 5(\text{vi})$$

The following illustrates this application.

### Application of the normal-binomial approximation: Ceramic plates

A firm has a continuous production operation to mould, glaze, and fire ceramic plates. It knows from historical data that in the operation 3% of the plates are defective and have to be sold at a marked down price. The quality control manager takes a random sample of 500 of these plates and inspects them.

1. Can we use the normal distribution to approximate the normal distribution?

The sample size  $n$  is 500, and the probability  $p$  is 3%. Using equations 5(iv) and 5(v)

$$np = 500 * 0.03 = 15 \text{ or a value } > 5$$

$$n(1 - p) = 500 * (1 - 0.03) = 500 * 0.97$$

$$= 485 \text{ and again a value } > 5$$

Thus both conditions are satisfied and so we can correctly use the normal distribution as an approximation of the binomial distribution.

2. Using the binomial distribution, what is the probability that 20 of the plates are defective?

Here we use in Excel, [function BINOMDIST] where  $x$  is 20, the characteristic probability  $p$  is 3%, the sample size,  $n$  is 500, and the cumulative value is 0. This gives the probability of exactly 20 plates being defective of 4.16%.

3. Using the normal–binomial approximation what is the probability that 20 of the plates are defective?

From equation 4(xv), the mean value of the binomial distribution is,

$$\mu_x = np = 500 * 0.003 = 15$$

From equation 4(xvii) the standard deviation of the binomial distribution is,

$$\begin{aligned}\sigma &= \sqrt{npq} = \sqrt{(500 * 0.003 * 0.997)} \\ &= 3.8144\end{aligned}$$

Here we use in Excel, [function NORMDIST] where  $x$  is 20, the mean value is 15, the standard deviation is 3.8144, and the cumulative value is 0. This gives the probability of exactly 20 plates being defective of 4.43%. This is a value not much different from 4.16% obtained in Question 2. (Note if we had used a cumulative value = 1 this would give the area from the left of the normal distribution curve to the value of  $x$ .)

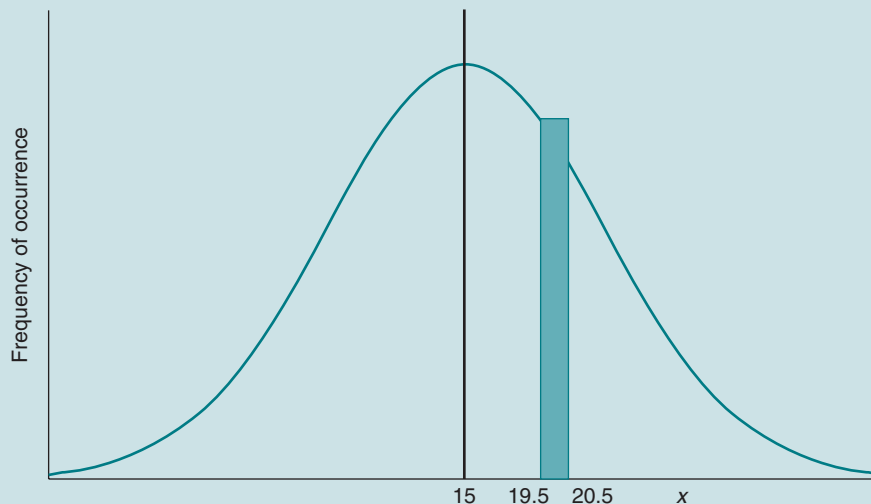
## Continuity correction factor

Now, the normal distribution is continuous, and is shown by a line graph, whereas the binomial distribution is discrete illustrated by a histogram. Another way to make the normal–binomial approximation is to apply a **continuity correction factor** so that we encompass the range of the discrete value recognizing that we are superimposing a histogram to a continuous curve. In the previous ceramic plate example, if we apply a correction factor of 0.5–20, the random variable  $x$  then on the lower side we have  $x_1 = 19.5$  ( $20 - 0.5$ ) and  $x_2 = 20.5$  ( $20 + 0.5$ ) on the upper side. The concept is illustrated in Figure 5.18.

Using equation 5(vi) for these two values of  $x$  gives

$$\begin{aligned}z_1 &= \frac{x_1 - np}{\sqrt{np(1-p)}} = \frac{19.5 - 500 * 0.03}{\sqrt{500 * 0.03(1-0.03)}} \\ &= \frac{19.5 - 15}{\sqrt{14.55}} = \frac{4.5}{3.8144} = 1.1797\end{aligned}$$

Figure 5.18 Continuity correction factor.



$$z_2 = \frac{x_2 - np}{\sqrt{np(1-p)}} = \frac{20.5 - 500 * 0.03}{\sqrt{500 * 0.03(1-0.03)}}$$

$$= \frac{20.5 - 15}{\sqrt{14.55}} = \frac{5.5}{3.8144} = 1.4419$$

Using in Excel, [function NORMSDIST] for a z-value of 1.1797 gives the area under the curve from the left to  $x = 19.5$  of 88.09%. For a value of  $x$  of 20.5 gives the area under the curve of 92.53%. The difference between these two areas is 4.44% (92.53% – 88.09%). This value is again close to those values obtained in the worked example for the ceramic plates.

### Sample size to approximate the normal distribution

The conditions that equations 5(iv) and 5(v) are met depend on the values of  $n$  and  $p$ . When  $p$  is

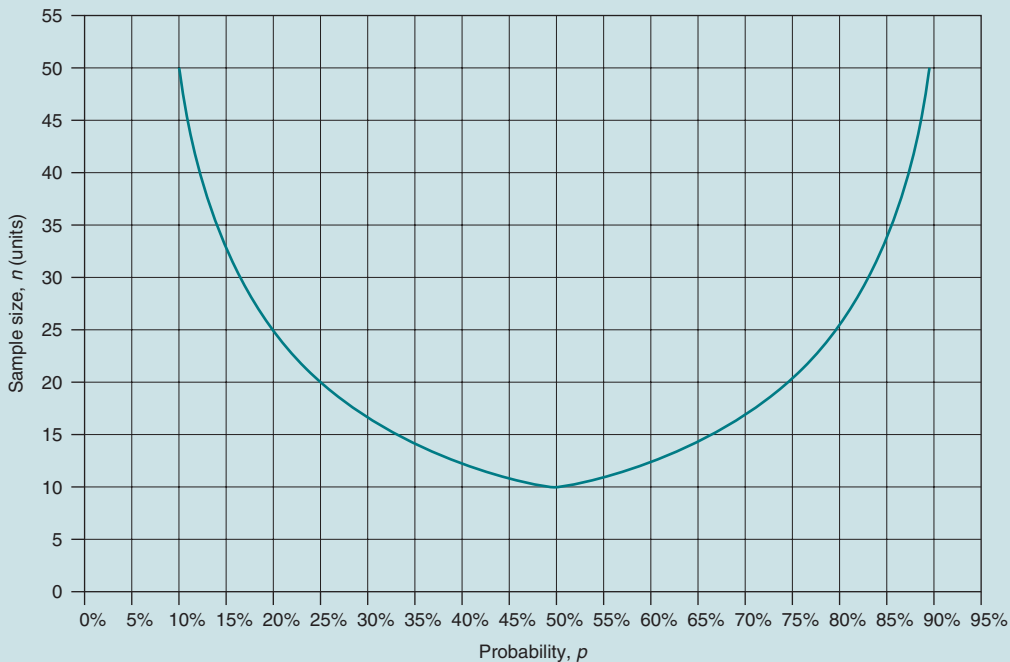
large then for a given value of  $n$  the product  $np$  is large; conversely  $n(1 - p)$  is small. The minimum sample size possible to apply the normal–binomial approximation is 10. In this case the probability,  $p$ , must be equal to 50% as for example in the coin toss experiment. As the probability  $p$  increases in value,  $(1 - p)$  decreases and so for the two conditions to be valid the sample size  $n$  has to be larger. If for example  $p$  is 99%, then the minimum sample size in order to apply the normal distribution assumption is 500 illustrated as follows:

$$p = 99\% \text{ and thus, } np = 500 * 99\% = 495$$

$$(1 - p) = 1\% \text{ and thus, } n(1 - p) = 500 * 1\% = 5$$

Figure 5.19 gives the relationship of the minimum values of the sample size,  $n$ , for values of  $p$  from 10% to 90% in order to satisfy both equations 5(iv) and 5(v).

Figure 5.19 Minimum sample size in a binomial situation to be able to apply the normal distribution assumption.





This chapter has been entirely devoted to the normal distribution.

## Describing the normal distribution

The normal distribution is the most widely used analytical tool in statistics and presents graphically the profile of a continuous random variable. Situations which might follow a normal distribution are those processes that are set to produce products according to a target or a mean value such as a bottle filling operation, the filling of yogurt pots, or the pouring of liquid chocolate into a mould. Simply because of the nature, or randomness of these operations, we will find volume or weight values below and above the set target value.

Visually, a normal distribution is bell or hump shaped and is symmetrical around this hump such that the left side is a mirror image of the right side. The central point of the hump is at the same time the mean, median, mode, and midrange. The left and right extremities, or the two tails of the normal distribution, may extend far from the central point. No matter the value of the mean or the standard deviation, the area under the curve of the normal distribution is always unity. In addition, 68.26% of all the data falls within  $\pm 1$  standard deviations from the mean, 95.44% of the data falls within  $\pm 2$  standard deviations from the mean, and 99.73% of data is  $\pm 3$  standard deviations from the mean. These empirical relationships allow the normal distribution to be used to determine probability outcomes of many situations.

Data in a normal distribution can be uniquely defined by its mean value and standard deviation and these values define the shape or kurtosis of the distribution. A distribution that has a small standard deviation relative to its mean has a sharp peak or is leptokurtic. A distribution that has a large standard deviation relative to its mean has a flat peak and is platykurtic. A distribution between these two extremes is mesokurtic. The importance of knowing these shapes is that a curve that is leptokurtic is more reliable for analytical purposes. When we know the values of the mean value,  $\mu$ , the standard deviation,  $\sigma$ , and the random variable,  $x$ , of a dataset we can transform the absolute values of the dataset into standard values. This then gives us a standard normal distribution which has a mean value of 0 and plus or minus values of  $z$ , the number of standard deviations from the mean corresponding to the area under the curve.

## Demonstrating that data follow a normal distribution

To verify that data follows a normal distribution there are several tests. We can develop a stem-and-leaf display if the dataset is small. For larger datasets we can draw a box-and-whisker plot, or plot a frequency polygon, and see if these displays are symmetrical. Additionally, we can determine the properties of the data to see if the mean is about equal to the median, that the inter-quartile range is equal to 1.33 times the standard deviation, that the data range is about six times the standard deviation, and that the empirical rules governing the number of standard deviations and the area under the curve are respected. If the mean and median value in a dataset are significantly different then the data is asymmetric or skewed. When the mean is greater than the median the distribution is positively or right-skewed, and when the mean is less than the median the distribution is negatively or left-skewed. A more rigorous test of symmetry involves developing a normal

probability plot which involves organizing the data into an ordered array and determining the values of  $z$  for defined equal portions of the data. If the normal probability plot is essentially linear with a positive slope, then the data is normal. If the plot is non-linear and concave then the data is right-skewed, and if it is convex then the data is left-skewed. Since we have divided data into defined portions, the normal probability plot is related to the data percentiles.

### A normal distribution to approximate a binomial distribution

When both the product of sample size,  $n$ , and probability,  $p$ , of success and the product of sample size and probability of failure ( $1 - p$ ) are greater or equal to five then we can use a normal distribution to approximate a binomial distribution. This condition applies for a minimum sample size of 10 when the probability of success is 50%. For other probability values the sample size will be larger. This normal–binomial approximation has practicality in sampling experiments such as statistical process control.

## EXERCISE PROBLEMS

### 1. Renault trucks

#### Situation

Renault Trucks, a division of Volvo Sweden, is a manufacturer of heavy vehicles. It is interested in the performance of its Magnum trucks that it sells throughout Europe to both large and smaller trucking companies. Based on service data throughout the Renault agencies in Europe it knows that on an annual basis the distance travelled by its trucks, before a major overhaul is necessary, is 150,000 km with a standard deviation of 35,000 km. The data is essentially normally distributed, and there were 62,000 trucks in the analysis.

#### Required

1. What proportion of trucks can be expected to travel between 82,000 and 150,000 km per year?
2. What is the probability that a randomly selected truck travels between 72,000 and 140,000 km per year?
3. What percentage of trucks can be expected to travel no more than 50,000 km per year and at least 190,000 km per year?
4. How many of the trucks in the analysis, are expected to travel between 125,000, and 200,000 km in the year?
5. In order to satisfy its maintenance and quality objectives Renault Trucks desires that at least 75% of its trucks travel at least 125,000 km. Does Renault Trucks reach this objective? Justify your answer by giving the distance at which at least 75% of the trucks travel.
6. What is the distance below which 99.90% of the trucks are expected to travel?
7. For analytical purposes for management, develop a greater than ogive based on the data points developed from Questions 1–6.

### 2. Telephone calls

#### Situation

An analysis of 1,000 long distance telephone calls made from a large business office indicates that the length of these calls is normally distributed, with an average time of 240 seconds, and a standard deviation of 40 seconds.

#### Required

1. What percentage of these calls lasted no more than 180 seconds?
2. What is the probability that a particular call lasted between 180 and 300 seconds?
3. How many calls lasted no more than 180 seconds and at least 300 seconds?
4. What percentage of these calls lasted between 110 and 180 seconds?
5. What is the length of a particular call, such that only 1% of all calls are shorter?

### 3. Training programme

#### Situation

An automobile company has installed an enterprise resource planning (ERP) system to better manage the firm's supply chain. The human resource department has been instructed to develop a training programme for the employees to fully understand how the new system functions. This training programme has a fixed lecture period and at the end of the programme there is a self-paced on-line practical examination that the participants have to pass before they are considered competent with the new ERP system. If they fail the examination they are able to retake it as many times as they wish in order to pass. When the employee passes the examination they are considered competent with the ERP system and they immediately receive a 2% salary increase. During the last several months, average completion of the programme, which includes passing the examination, has been 56 days, with a standard deviation of 14 days. The time taken to pass the examination is considered to follow a normal distribution.

#### Required

1. What is the probability that an employee will successfully complete the programme between 40 and 51 days?
2. What is the probability an employee will successfully complete programme in 35 days or less?
3. What is the combined probability that an employee will successfully complete the programme in no more than 34 days or more than 84 days?
4. What is the probability that an employee will take at least 75 days to complete the training programme?
5. What are the upper and lower limits in days within which 80% of the employees will successfully complete the programme?

### 4. Cashew nuts

#### Situation

Salted cashew nuts sold in a store are indicated on the packaging to have a nominal net weight of 125 g. Tests at the production site indicate that the average weight in a package is 126.75 g with a standard deviation of 1.25 g.

#### Required

1. If you buy a packet of these cashew nuts at a store, what is the probability that your packet will contain more than 127 g?
2. If you buy a packet of these cashew nuts at a store, what is the probability that your packet will contain less than the nominal indicated weight of 125 g?
3. What is the minimum and maximum weight of a packet of cashew nuts in the middle 99% of the cashew nuts?
4. In the packets of cashew nuts, 95% will contain at least how much in weight?

## 5. Publishing

### Situation

Cathy Peck is the publishing manager of a large textbook publishing house in England. Based on passed information she knows that it requires on average, 10.5 months to publish a book from receipt of manuscript from the author to getting the book on the market. She also knows that from past publishing data a normal distribution represents the distribution time for publication, and that the standard deviation for the total process from review, publication, to distribution is 3.24 months. In a certain year she is told that she will receive 19 manuscripts for publication.

### Required

1. From the manuscripts she is promised to receive this year for publication, approximately how many can Cathy expect to publish within the first quarter?
2. From the manuscripts she is promised to receive this year for publication, approximately how many can Cathy expect to publish within the first 6 months?
3. From the manuscripts she is promised to receive this year for publication, approximately how many can Cathy expect to publish within the third quarter?
4. From the manuscripts she is promised to receive this year for publication, approximately how many can Cathy expect to publish within the year?
5. If by the introduction of new technology, the publishing house can reduce the average publishing time and the standard deviation by 30%, how many of the 19 manuscripts could be published within the year?

## 6. Gasoline station

### Situation

A gasoline service sells, on average 5,000 litre of diesel oil per day. The standard deviation of this sale is 105 litre per day. The assumption is that the sale of diesel oil follows a normal distribution.

### Required

1. What is the probability that on a given day, the gas station sells at least 5,180 litre?
2. What is the probability that on a given day, the gas station sells no more than 4,850 litre?
3. What is the probability that on a given day, the gas station sells between 4,700 and 5,200 litre?
4. What is the volume of diesel oil sales at which the sales are 80% more?
5. The gasoline station is open 7 days a week and diesel oil deliveries are made once a week on Monday morning. To what level should diesel oil stocks be replenished if the owner wants to be 95% certain of not running out of diesel oil before the next delivery? Daily demand of diesel oil is considered reasonably steady.

## 7. Ping-pong balls

### Situation

In the production of ping-pong balls the mean diameter is 370 mm and their standard deviation is 0.75 mm. The size distribution of the production of ping-pong balls is considered to follow a normal distribution.

### Required

1. What percentage of ping-pong balls can be expected to have a diameter that is between 369 and 370 mm?
2. What is the probability that the diameter of a randomly selected ping-pong ball is between 372 and 369 mm?
3. What is the combined percentage of ping-pong balls can be expected to have a diameter that is no more than 368 mm or is at least 371 mm?
4. If there are 25,000 ping-pong balls in a production lot how many of them would have a diameter between 368 and 371 mm?
5. What is the diameter of a ping-pong ball above which 75% are greater than this diameter?
6. What are the symmetrical limits of the diameters between which 90% of the ping-pong balls would lie?
7. What can you say about the shape of the normal distribution for the production of ping-pong balls?

## 8. Marmalade

### Situation

The nominal net weight of marmalade indicated on the jars is 340 g. The filling machines are set to the nominal weight and the standard deviation of the filling operation is 3.25 g.

### Required

1. What percentage of jars of marmalade can be expected to have a net weight between 335 and 340 g?
2. What percentage of jars of marmalade can be expected to have a net weight between 335 and 343 g?
3. What is the combined percentage of jars of marmalade that can be expected to have a net weight that is no more than 333 g and at least 343 g?
4. If there are 40,000 jars of marmalade in a production lot how many of them would have a net weight between 338 and 345 g?
5. What is the net weight of jars of marmalade above which 85% are greater than this net weight?
6. What are the symmetrical limits of the net weight between which 99% of the jars of marmalade lie?
7. The jars of marmalade are packed in cases of one dozen jars per case. What proportion of cases will be above 4.1 kg in net weight?

## 9. Restaurant service

### Situation

The profitability of a restaurant depends on how many customers can be served and the price paid for a meal. Thus, a restaurant should service the customers as quickly as possible but at the same time providing them quality service in a relaxed atmosphere. A certain restaurant in New York, in a 3-month study, had the following data regarding the time taken to service clients. It believed that it was reasonable to assume that the time taken to service a customer, from showing to the table and seating, to clearing the table after the client had been serviced, could be approximated by a normal distribution.

Activity	Average time (minutes)	Variance
Showing to table, and seating client	4.24	1.1025
Selecting from menu	10.21	5.0625
Waiting for order	14.45	9.7344
Eating meal	82.14	378.3025
Paying bill	7.54	3.4225
Getting coat and leaving	2.86	0.0625
Clearing table	3.56	0.7744

### Required

1. What is the average time and standard deviation to serve a customer such that the restaurant can then receive another client?
2. What is the probability that a customer can be serviced between 90 and 125 minutes?
3. What is the probability that a customer can be serviced between 70 and 140 minutes?
4. What is the combined probability that a customer can be serviced in 70 minutes or less and at least 140 minutes?
5. If in the next month it is estimated that 1,200 customers will come to the restaurant, to the nearest whole number, what is a reasonable estimate of the number of customers that can be serviced between 70 and 140 minutes?
6. Again, on the basis that 1,200 customers will come to the restaurant in the next month, 85% of the customers will be serviced in a minimum of how many minutes?

## 10. Yoghurt

### Situation

The Candy Corporation has developed a new yogurt and is considering various prices for the product. Marketing developed an initial daily sales estimate of 2,400 cartons, with a standard deviation of 45. Prices for the yogurt were then determined based on that forecast. A later revised estimate from marketing was that average daily sales would be 2,350 cartons.

**Required**

1. According to the revised estimate, what is the probability that a day's sale will still be over 2,400 given that the standard deviation remains the same?
2. According to the revised estimate, what is the probability that a day's sale will be at least 98% of 2,400?

**11. Motors****Situation**

The IBB Company has just received a large order to produce precision electric motors for a French manufacturing company. To fit properly, the drive shaft must have a diameter of  $4.2 \pm 0.05$  cm. The production manager indicates that in inventory there is a large quantity of steel rods with a mean diameter of 4.18 cm, and a standard deviation of 0.06 cm.

**Required**

1. What is the probability of a steel rod from this inventory stock, meeting the drive shaft specifications?

**12. Doors****Situation**

A historic church site wishes to add a door to the crypt. The door opening for the crypt is small and the church officials want to enlarge the opening such that 95% of visitors can pass through without stooping. Statistics indicate that the adult height is normally distributed, with a mean of 1.76 m, and a standard deviation of 12 cm.

**Required**

1. Based on the design criterion, what height should the doors be made to the nearest cm?
2. If after consideration, the officials decided to make the door 2 cm higher than the value obtained in Question 1, what proportion of the visitors would have to stoop when going through the door?

**13. Machine repair****Situation**

The following are the three stages involved in the servicing of a machine.

Activity	Mean time (minutes)	Standard deviation (minutes)
Dismantling	20	4
Testing and adjusting	30	7
Reassembly	15	3



**Required**

1. What is the probability that the dismantling time will take more than 28 minutes?
2. What is the probability that the testing and adjusting activity alone will take less than 27 minutes?
3. What is the probability that the reassembly activity alone will take between 13 and 18 minutes?
4. What is the probability that an allowed time of 75 minutes will be sufficient to complete the servicing of the machine including dismantling, testing and adjusting, and assembly?

**14. Savings****Situation**

A financial institution is interested in the life of its regular savings accounts opened at its branch. This information is of interest as it can be used as an indicator of funds available for automobile loans. An analysis of past data indicates that the life of a regular savings account, maintained at its branch, averages 17 months, with a standard deviation of 171 days. For calculation purposes 30 days/month is used. The distribution of this past data was found to be approximately normal.

**Required**

1. If a depositor opens an account with this savings institution, what is the probability that there will still be money in that account in 20 months?
2. What is the probability that the account will have been closed within 2 years?
3. What is the probability that the account will still be open in 2.5 years?
4. What is the chance an account will be open in 3 years?

**15. Buyout – Part III****Situation**

Carrefour, France, is considering purchasing the total 50 retail stores belonging to Hardway, a grocery chain in the Greater London area of the United Kingdom. The profits from these 50 stores, for one particular month, in £ '000s, are as follows. (This is the same information as provided in Chapters 1 and 2.)

8.1	11.8	8.7	10.6	9.5
9.3	11.5	10.7	11.6	7.8
10.5	7.6	10.1	8.9	8.6
11.1	10.2	11.1	9.9	9.8
11.6	15.1	12.5	6.5	7.5
10.3	12.9	9.2	10.7	12.8
12.5	9.3	10.4	12.7	10.5
10.3	11.1	9.6	9.7	14.5
13.7	6.7	11.5	8.4	10.3
13.7	11.2	7.3	5.3	12.5

### Required

1. Carrefour management decides that it will purchase only those stores showing profits greater than £12,500. On the basis that the data follow a normal distribution, calculate how many of the Hardway stores Carrefour would purchase? (You have already calculated the mean and the standard deviation in the *Exercise Buyout – Part II* in Chapter 2.)
2. How does the answer to Question 1 compare to the answer to Question 6 of buyout in Chapter 1 that you determined from the ogive?
3. What are your conclusions from the answers determined from both methods.

## 16. Case: Cadbury's chocolate

### Situation

One of the production lines of Cadbury Ltd turns out 100-g bars of milk chocolate at a rate of 20,000/hour. The start of this production line is a stainless steel feeding pipe that delivers the molten chocolate, at about 80°C, to a battery of 10 injection nozzles. These nozzles are set to inject a little over 100 g of chocolate into flat trays which pass underneath the nozzles. Afterwards these trays move along a conveyer belt during which the chocolate cools and hardens taking the shape of the mould. In this cooling process some of the water in the chocolate evaporates in order that the net weight of the chocolate comes down to the target value of 100 g. At about the middle of the conveyer line, the moulds are turned upside down through a reverse system on the belt after which the belt vibrates slightly such that the chocolate bars are ejected from the mould. The next production stage is the packing process where the bars are first wrapped in silver foil then wrapped in waxed paper onto which is printed the product type and the net weight. The final part of this production line is where the individual bars of chocolate are packed in cardboard cartons. Immediately upstream of the start of the packing process, the bars of chocolate pass over an automatic weighing machine that measures at random the individual weights. A printout of the weights for a sample of 1,000 bars, from a production run of 115,000 units is given in the table below. The production cost for these 100 g chocolate bars is £0.20/unit. They are sold in retail for £3.50.

### Required

From the statistical sample data presented, how would you describe this operation? What are your opinions and comments?

109.99	95.56	92.29	117.16	103.93	89.73	137.67	99.30	98.20	89.77
81.33	128.96	104.47	106.73	118.97	104.55	94.95	105.66	120.96	94.08
105.70	112.18	100.18	110.90	114.25	88.27	117.80	109.98	104.82	102.25
106.96	87.54	105.09	100.48	125.71	107.17	114.03	108.01	95.51	102.47
100.11	77.12	111.19	107.15	96.38	96.64	91.94	94.29	110.69	92.12
110.08	107.39	106.28	96.61	96.73	98.66	78.01	87.28	127.09	107.36
107.37	104.22	97.39	97.83	116.30	86.33	85.08	104.91	115.76	111.78
88.47	82.33	81.65	94.06	109.03	113.81	73.68	94.65	94.22	86.08

117.72	98.28	110.29	96.11	97.56	84.73	90.66	107.46	91.69	111.41
84.66	90.12	92.61	119.93	103.56	107.85	94.77	108.89	102.71	94.71
104.06	82.20	68.25	83.26	100.75	113.60	86.70	89.53	113.24	101.96
77.03	114.88	101.85	110.09	101.58	95.08	100.03	114.82	78.61	78.30
93.40	112.55	87.20	126.22	99.58	105.39	120.19	120.80	112.80	118.26
110.99	86.71	113.41	94.49	76.15	90.53	88.65	108.99	110.82	100.12
82.77	94.01	107.12	90.72	100.85	80.92	84.10	91.01	103.10	76.31
110.37	100.82	98.78	100.22	118.64	133.14	92.54	88.88	79.28	105.22
106.50	81.94	110.45	105.36	100.35	102.25	87.17	99.59	107.66	103.49
127.22	86.24	91.36	115.23	93.63	91.47	112.13	108.65	106.22	108.44
76.73	111.44	104.75	92.64	93.21	107.99	93.08	99.96	97.36	98.27
109.54	100.09	98.18	92.54	97.86	110.86	118.15	84.37	115.87	80.20
95.18	100.96	111.12	102.37	130.33	91.68	109.46	86.43	96.22	99.61
83.61	100.15	104.68	106.46	108.35	81.11	77.62	98.70	94.96	109.65
90.08	87.39	107.58	111.92	106.97	85.60	107.82	113.96	115.22	100.76
125.89	89.80	92.81	114.38	104.46	90.48	103.74	116.37	123.87	116.78
90.70	87.09	92.41	101.24	96.72	97.35	81.84	112.72	79.72	131.30
108.39	79.78	112.91	78.84	112.81	115.89	116.72	93.32	91.96	122.44
91.94	107.23	111.40	122.86	105.62	115.47	101.21	110.45	104.65	109.23
79.58	88.08	123.39	110.58	74.03	95.81	117.48	84.67	101.72	96.16
87.42	90.88	116.54	83.95	92.30	100.04	91.21	92.71	89.79	94.91
97.83	110.16	118.70	96.35	111.99	123.15	107.09	80.51	88.89	89.35
109.66	108.50	83.78	112.01	115.94	109.48	114.54	102.57	98.96	100.70
93.97	106.18	91.46	96.15	102.13	70.63	91.56	113.09	113.96	123.54
69.76	107.66	119.46	85.91	109.40	93.40	98.23	97.06	105.96	110.04
115.56	93.79	91.70	98.56	121.63	86.92	125.22	90.20	100.51	122.15
85.87	102.32	88.38	98.58	100.82	99.82	86.25	87.71	131.27	85.70
102.75	89.01	90.46	104.81	116.34	112.18	103.97	100.78	105.93	84.94
105.68	86.58	107.06	120.53	110.99	92.13	83.48	98.91	117.34	93.74
104.62	80.46	100.05	100.87	113.41	92.96	115.99	96.20	114.02	108.22
94.09	94.13	96.66	100.80	97.73	75.22	99.75	96.00	86.84	94.29
124.37	80.46	91.53	101.49	92.42	110.46	102.64	75.03	101.15	105.47
126.44	105.65	120.84	111.79	109.08	119.04	112.62	118.08	102.80	93.42
99.15	111.19	111.35	104.32	101.15	107.82	131.00	89.15	111.87	99.05
76.55	118.01	104.89	104.34	95.76	98.66	84.72	124.61	100.82	117.34
103.06	88.58	102.46	71.23	103.30	85.94	100.85	104.59	93.75	102.43
89.16	87.54	100.76	84.81	105.23	103.23	113.82	92.61	99.83	83.20
98.47	97.58	95.74	97.12	75.73	98.74	101.79	96.19	106.64	94.00
99.67	106.74	100.36	107.74	94.16	121.43	112.87	99.19	113.85	104.54
87.03	107.22	128.15	101.96	95.28	114.46	100.17	91.39	87.03	92.03
115.58	96.56	107.22	108.70	123.61	78.08	105.65	86.94	69.47	88.40
105.53	105.78	100.39	93.56	98.10	96.59	107.41	102.82	111.18	93.81
122.64	101.94	98.80	103.18	74.65	93.82	102.75	122.86	97.53	108.53
72.33	93.40	88.61	112.02	101.06	85.77	110.74	79.13	98.69	111.28
89.72	84.26	114.09	98.53	107.80	101.85	94.06	116.99	103.03	109.48
109.64	94.41	106.91	115.02	106.62	130.32	92.96	115.74	85.03	118.97
79.53	96.89	74.95	107.34	111.82	85.01	113.15	100.49	88.89	89.35
97.41	104.09	84.20	97.75	106.11	108.26	98.33	116.27	104.37	132.20
105.22	116.57	102.50	93.75	122.28	93.06	85.72	88.30	115.09	96.59
93.58	97.92	104.43	108.59	98.57	111.71	76.00	88.22	122.84	107.31

84.93	108.27	105.92	98.32	101.58	91.72	87.44	89.97	100.80	103.96
103.15	81.68	115.27	105.86	100.33	110.56	97.16	113.09	109.75	116.14
108.35	85.48	110.16	96.91	96.26	109.91	104.85	97.13	101.72	109.02
110.02	100.71	92.17	109.47	98.38	124.46	87.58	109.23	95.87	113.11
72.25	105.48	105.97	99.82	104.22	93.66	90.59	112.85	87.10	110.99
118.28	105.78	96.00	93.23	101.84	112.44	114.84	101.73	87.78	95.28
93.70	92.57	115.12	106.43	90.42	82.52	104.80	90.70	113.83	113.31
97.15	103.23	90.64	113.09	109.95	110.22	93.42	74.92	102.61	102.97
87.92	91.97	86.28	97.84	94.52	94.13	113.85	101.91	98.08	111.08
120.77	97.15	98.11	108.47	97.61	94.78	88.52	112.29	100.67	103.23
101.18	101.05	105.53	86.92	106.76	85.76	98.48	96.49	124.08	89.59
86.13	99.46	105.28	92.39	94.52	113.24	101.07	106.23	86.39	117.77
91.01	98.00	105.70	114.91	118.27	112.31	111.56	74.23	93.76	94.83
101.59	129.89	99.70	84.06	105.70	91.04	102.12	87.74	102.60	96.53
87.54	94.44	95.50	107.41	103.78	87.14	86.56	94.16	100.68	93.12
99.68	124.37	84.63	128.39	114.25	104.27	64.48	113.58	92.58	114.63
97.72	95.06	80.25	113.94	90.48	92.83	85.32	82.21	86.13	102.19
104.72	88.06	98.27	100.55	103.04	101.62	83.50	118.85	90.36	72.94
114.48	86.68	117.77	76.94	77.25	114.89	93.57	105.11	99.00	117.92
80.75	106.39	114.61	98.57	84.24	98.66	99.65	89.68	84.72	99.68
99.38	101.92	109.34	110.26	84.38	78.99	103.79	72.33	104.69	102.89
84.99	93.90	106.92	96.53	100.53	100.80	112.18	92.37	103.46	94.40
91.32	114.90	95.86	95.88	101.66	100.41	106.12	112.15	91.74	93.69
93.53	79.31	90.20	108.82	98.27	110.21	107.67	107.36	102.31	105.89
101.95	101.77	128.61	88.56	104.56	115.18	114.19	81.22	94.47	118.35
93.91	104.75	101.42	96.85	110.45	109.30	102.65	91.28	96.53	92.34
84.35	84.59	84.46	98.85	113.25	93.16	104.71	91.55	86.53	101.49
116.09	97.42	86.59	112.56	124.65	89.42	113.61	107.15	87.79	99.23
125.83	105.65	118.83	97.00	78.92	107.13	96.54	96.27	107.66	94.27
105.43	84.18	94.67	99.88	69.91	104.42	109.96	111.84	96.00	102.08
96.00	99.56	101.01	101.25	79.24	98.06	101.96	84.91	91.40	93.12
104.58	99.72	105.64	103.42	113.30	116.15	94.42	98.61	105.44	81.72
119.27	108.80	109.19	100.85	97.70	114.42	134.78	106.44	123.01	80.60
109.68	96.30	114.11	118.06	123.57	111.82	90.50	103.37	110.63	124.73
135.40	136.89	111.58	110.97	96.37	97.72	110.05	94.75	82.73	83.65
82.08	99.27	85.42	83.71	88.87	88.72	91.08	115.13	111.54	104.20
116.02	85.62	87.85	90.56	110.26	90.45	119.81	98.53	116.15	94.91
118.86	101.61	92.91	87.22	93.32	99.10	112.63	106.08	96.40	91.15
113.20	108.29	109.15	92.48	118.07	95.72	88.35	94.94	111.22	94.35
90.46	101.91	112.82	78.72	114.22	110.04	122.23	96.58	78.32	90.44
123.99	95.47	95.63	93.56	108.56	107.07	108.76	86.53	103.00	103.49
97.44	74.24	99.06	93.27	80.79	91.86	108.70	106.80	112.44	100.42

*This page intentionally left blank*

# Theory and methods of statistical sampling

## The sampling experiment was badly designed!

*A well-designed sample survey can give pretty accurate predictions of the requirements, desires, or needs of a population. However, the accuracy of the survey lies in the phrase “well-designed”. A classic illustration of sampling gone wrong was in 1948 during the presidential election campaign when the two candidates were Harry Truman, the Democratic incumbent and Governor Dewey of New York, the Republican candidate. The Chicago Tribune was “so sure” of the outcome that the headlines in their morning daily paper of 3 November 1948 as illustrated in Figure 6.1, announced, “Dewey defeats Truman”. In fact Harry Truman won by a narrow but decisive victory of 49.5% of the popular vote to Dewey’s 45% and with an electoral margin of 303 to 189. The Chicago Tribune had egg on their face; something went wrong with the design of their sample experiment!<sup>1,2</sup>*

<sup>1</sup> *Chicago Daily Tribune*, 3 November 1948.

<sup>2</sup> Freidel, F., and Brinkley, A. (1982), *America in the Twentieth Century*, 5th edition, McGraw Hill, New York, pp. 371–372.

Figure 6.1 Harold Truman holding aloft a copy of the November 3rd 1948 morning edition of the Chicago Tribune.



## Learning objectives

After you have studied this chapter you will understand the **theory, application, and practical methods of sampling**, an important application of statistical analysis. The topics are broken down according to the following themes:

- ✓ **Statistical relations in sampling for the mean** • Sample size and population • Central limit theory
  - Sample size and shape of the sampling distribution of the means • Variability and sample size
  - Sample mean and the standard error.
- ✓ **Sampling for the means for an infinite population** • Modifying the normal transformation relationship • Application of sampling from an infinite normal population: *Safety valves*
- ✓ **Sampling for the means from a finite population** • Modification of the standard error
  - Application of sampling from a finite population: *Work week*
- ✓ **Sampling distribution of the proportion** • Measuring the sample proportion • Sampling distribution of the proportion • Binomial concept in sampling for the proportion • Application of sampling for proportions: *Part-time workers*
- ✓ **Sampling methods** • Bias in sampling • Randomness in your sample experiment • Excel and random sampling • Systematic sampling • Stratified sampling • Several strata of interest
  - Cluster sampling • Quota sampling • Consumer surveys • Primary and secondary data

In business, and even in our personal life, we often make decisions based on limited data. What we do is take a **sample** from a **population** and then make an inference about the population characteristics, based entirely on the analysis of this sample. For example, when you order a bottle of wine in a restaurant, the waiter pours a small quantity in your glass to taste. Based on that small quantity of wine you accept or reject the bottle of wine as drinkable. The waiter would hardly let you drink the whole bottle before you decide it is no good! The United States Dow Jones Industrial Average consists of just 30 stocks but this sample average is used as a measure of economic changes when in reality there are hundreds of stocks in the United States market where millions of dollars change hands daily. In political elections, samples of people's voting intentions are made and based on the proportion that prefer a particular candidate, the expected outcome of the nation's election may be presented beforehand. In manufacturing, lots of materials, assemblies, or finished products are

sampled at random to see if pieces conform to appropriate specifications. If they do, the assumption is that the entire population, the production line or the lot from where these samples are taken, meet the desired specifications and so all the units can be put onto the market. And, how many months do we date our future spouse before we decide to spend the rest of our life together!

### Statistical Relationships in Sampling for the Mean

The usual purpose of taking and analysing a sample is to make an estimate of the population parameter. We call this **inferential statistics**. As the sample size is smaller than the population we have no guarantee of the population parameter that we are trying to measure, but from the sample analysis, we draw conclusions. If we really wanted to guarantee our conclusion we would have to analyse the whole population but



in most cases this is impractical, too costly, takes too long, or is clearly impossible. An alternative to inferential statistics is **descriptive statistics** which involves the collection and analysis of the dataset in order to characterize just the sampled dataset.

## Sample size and population

A question that arises in our sampling work to infer information about the population is what should be the size of the sample in order to make a reliable conclusion? Clearly the larger the sample size, the greater is the probability of being close to estimating the correct population parameter, or alternatively, the smaller is the risk of making an inappropriate estimate. To demonstrate the impact of the sample size, consider an experiment where there is a population of seven steel rods, as shown in Figure 6.2. The number of the rod and its length in centimetres is indicated in Table 6.1.

Figure 6.2 Seven steel rods and their length in centimetres.

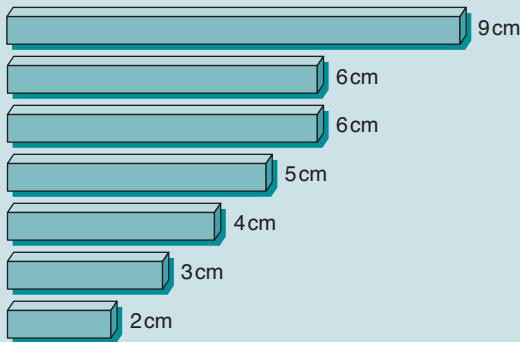


Table 6.1 Size of seven steel rods.

Rod number	1	2	3	4	5	6	7
Rod length (cm)	2.00	3.00	4.00	5.00	6.00	6.00	9.00

The total length of these seven rods is 35 cm ( $2 + 3 + 4 + 5 + 6 + 6 + 9$ ). This translates into a mean value of the length of the rods of 5 cm ( $35/7$ ). If we take samples of these rods from the population, with replacement, then from the counting relations in Chapter 3, the possible combinations of rods that can be taken, the same rod not appearing twice in the sample, is given by the relationship,

$$\text{Combinations} = \frac{n!}{x!(n-x)!} \quad 3(\text{xvi})$$

Here,  $n$  is the size of the population, or in this case 7 and  $x$  is the size of the sample. For example, if we select a sample of size of 3, the number of possible different combinations from equation 3(xvi) is,

$$\begin{aligned} \text{Combinations} &= \frac{7!}{3!(7-3)!} = \frac{7!}{3! \cdot 4!} \\ &= \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 35 \end{aligned}$$

If we increase the sample sizes from one to seven rods, then from equation 3(xvi) the total possible number of different samples is as given in Table 6.2. Thus, we sample from the population first with a sample size of one, then two, three, etc. right through to seven. Each time we select a sample we determine the sample mean value of the length of rods selected. For example, if the sample size is 3 and rods of length 2, 4, and 6 cm are selected, then the mean length,  $\bar{x}$ , of the sample would be,

$$\frac{2 + 4 + 6}{3} = 4.00 \text{ cm}$$

Table 6.2 Number of samples from a population of seven steel rods.

Sample size, $x$	1	2	3	4	5	6	7
No. of possible different samples	7	21	35	35	21	7	7



**Table 6.4** Frequency distribution within sample means for different sample sizes.

Sample mean $\bar{x}$	Sample size						
	1	2	3	4	5	6	7
2.00	1	0	0	0	0	0	0
2.25	0	0	0	0	0	0	0
2.50	0	1	0	0	0	0	0
2.75	0	0	0	0	0	0	0
3.00	1	1	1	0	0	0	0
3.25	0	0	0	0	0	0	0
3.50	0	2	1	1	0	0	0
3.75	0	0	3	2	0	0	0
4.00	1	3	3	2	2	0	0
4.25	0	0	0	3	1	0	0
4.50	0	3	4	4	1	1	0
4.75	0	0	4	3	2	0	0
5.00	1	2	4	4	5	3	1
5.25	0	0	0	4	3	1	0
5.50	0	3	3	4	3	2	0
5.75	0	0	4	3	2	0	0
6.00	2	2	3	3	2	0	0
6.25	0	0	0	1	0	0	0
6.50	0	1	2	1	0	0	0
6.75	0	0	2	0	0	0	0
7.00	0	1	1	0	0	0	0
7.25	0	0	0	0	0	0	0
7.50	0	2	0	0	0	0	0
7.75	0	0	0	0	0	0	0
8.00	0	0	0	0	0	0	0
8.25	0	0	0	0	0	0	0
8.50	0	0	0	0	0	0	0
8.75	0	0	0	0	0	0	0
9.00	1	0	0	0	0	0	0
Total	7	21	35	35	21	7	1

this number divided by the sample number of 35 gives 5. These values are given at the bottom of Table 6.3. What we conclude is that the sample means are always equal to 5 cm, or exactly the same as the population mean.

Next, for each sample size, a frequency distribution of mean length is determined. This data is given in Table 6.4. The left-hand column gives the sample mean and the other columns give the number of occurrences within a class limit

according to the sample size. For example, for a sample size of four there are four sample means greater than 4.25 cm but less than or equal to 4.50 cm. This data is now plotted as a frequency histogram in Figures 6.3 to 6.9 where each of the seven histograms have the same scale on the  $x$ -axis.

From Figures 6.3 to 6.9 we can see that as the sample size increases from one to seven, the dispersion about the mean value of 5 cm becomes smaller or alternatively more sample means lie closer to the population mean. For the sample size of seven, or the whole population, the dispersion is zero. The mean of the sample means,  $\bar{\bar{x}}$ , is always equal to the population mean of 5 or they have the same central tendency. This experiment demonstrates the concept of the **central limit theory** explained in the following section.

## Central limit theory

The foundation of sampling is based on the central limit theory, which is the criterion by which information about a population parameter can be inferred from a sample. The central limit theory states that in sampling, as the size of the sample increases, there becomes a point when the **distribution of the sample means**,  $\bar{x}$ , can be approximated by the normal distribution. This is so even though the distribution of the population itself may not necessarily be normal.

The distribution of the sample means, also called **sampling distribution of the means**, is a probability distribution of all the possible means of samples taken from a population. This concept of sampling and sampling means is illustrated by the information in Table 6.5 for the production of chocolate. Here the production line is producing 500,000 chocolate bars, and this is the population value,  $N$ . The moulding for the chocolate is set such that the weight of each chocolate bar should be 100 g. This is the nominal weight of the chocolate bar and is

Figure 6.3 Samples of size 1 taken from a population of size 7.

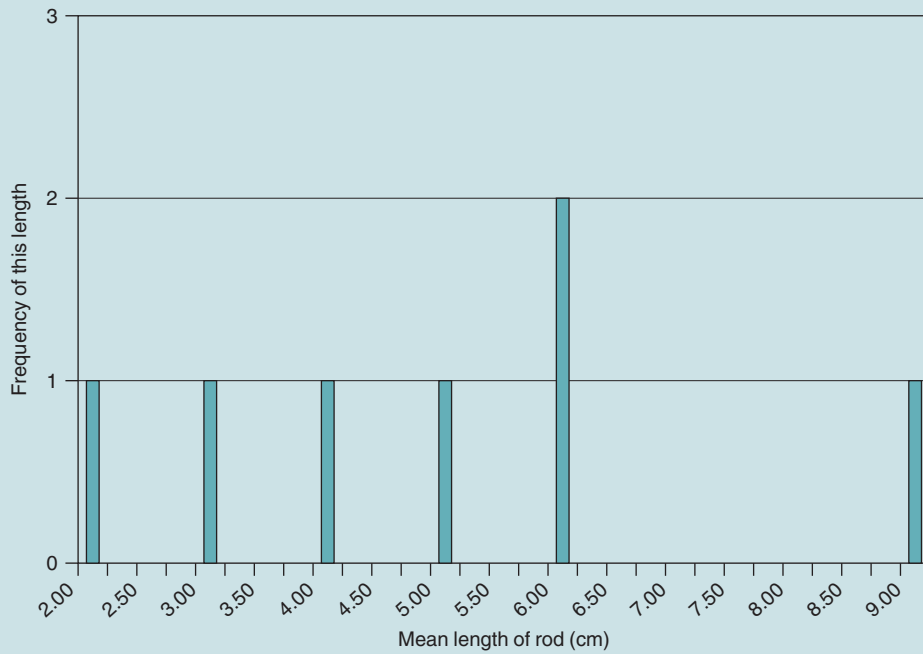


Figure 6.4 Samples of size 2 taken from a population of size 7.

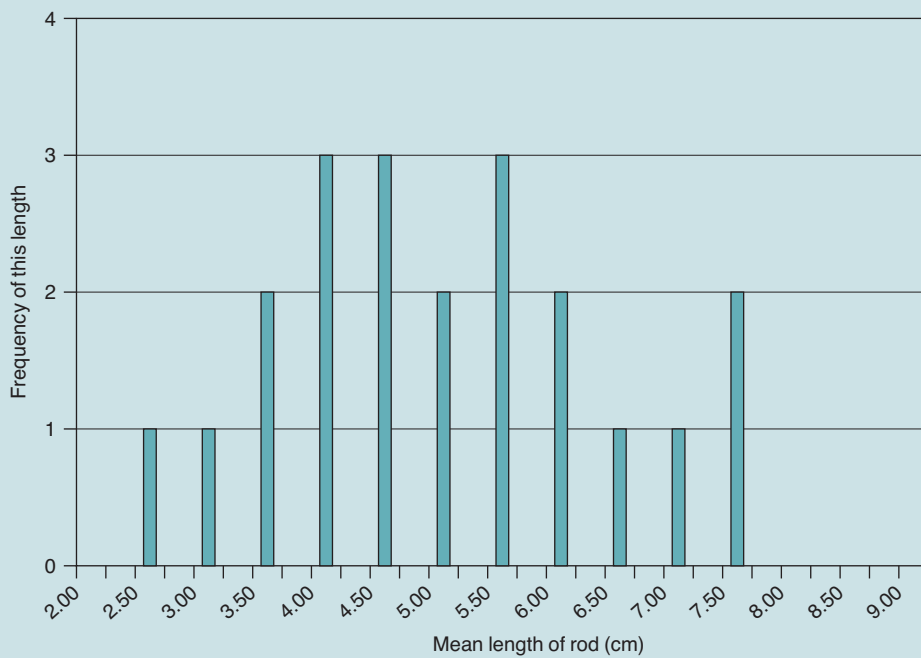


Figure 6.5 Samples of size 3 taken from a population of size 7.

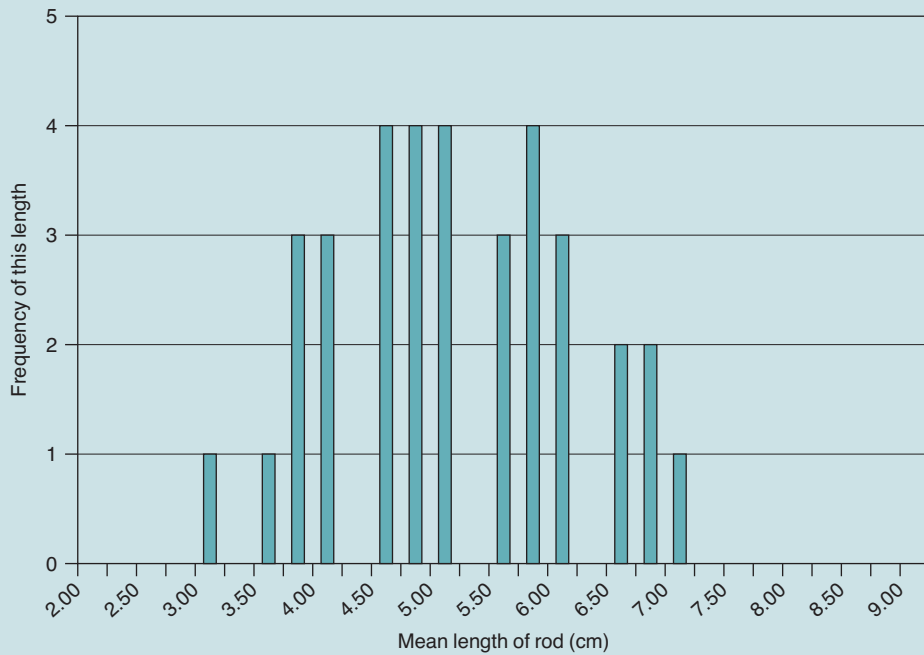


Figure 6.6 Samples of size 4 taken from a population of size 7.

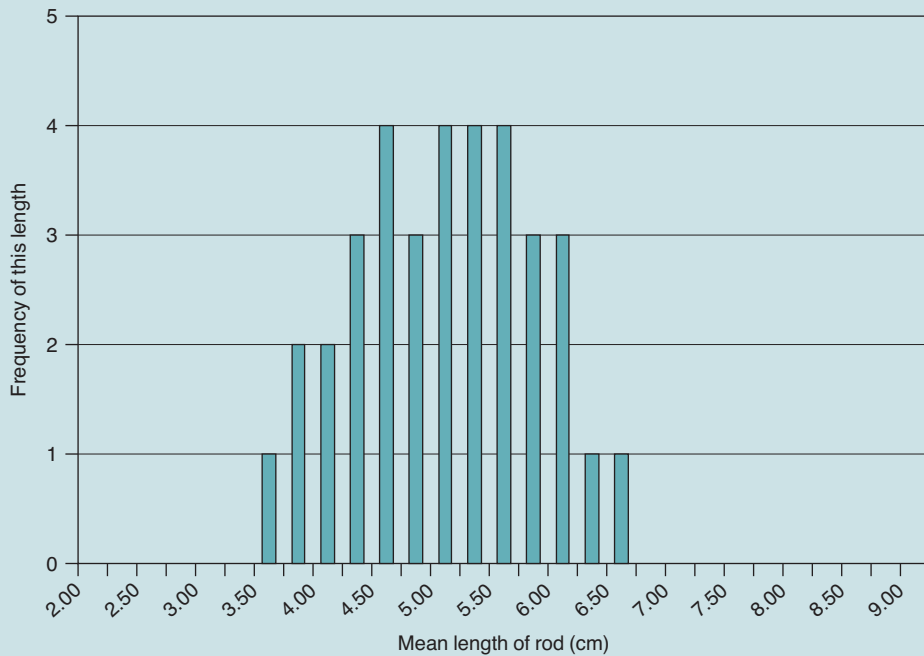


Figure 6.7 Samples of size 5 taken from a population of size 7.

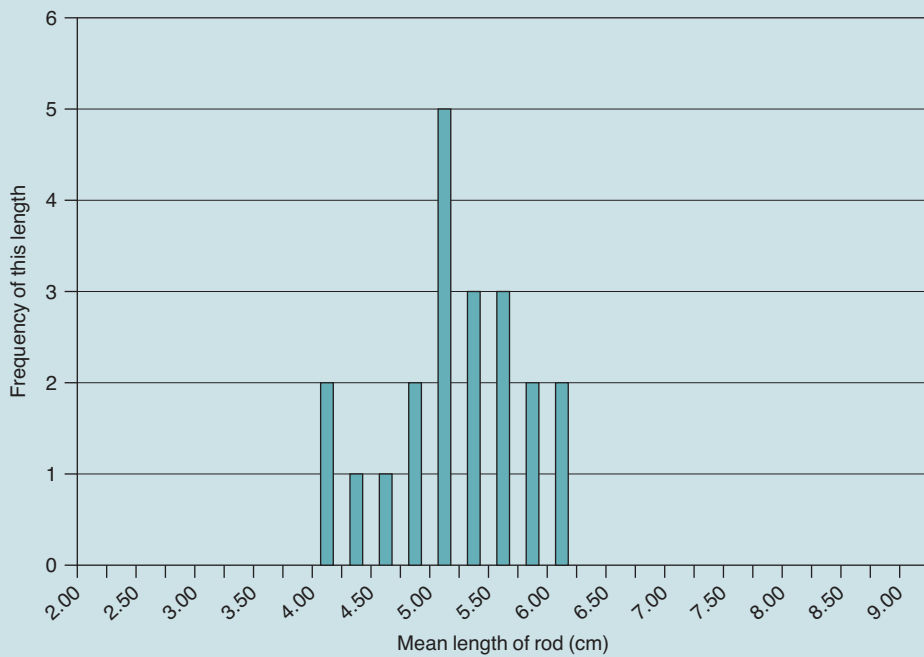


Figure 6.8 Samples of size 6 taken from a population of size 7.

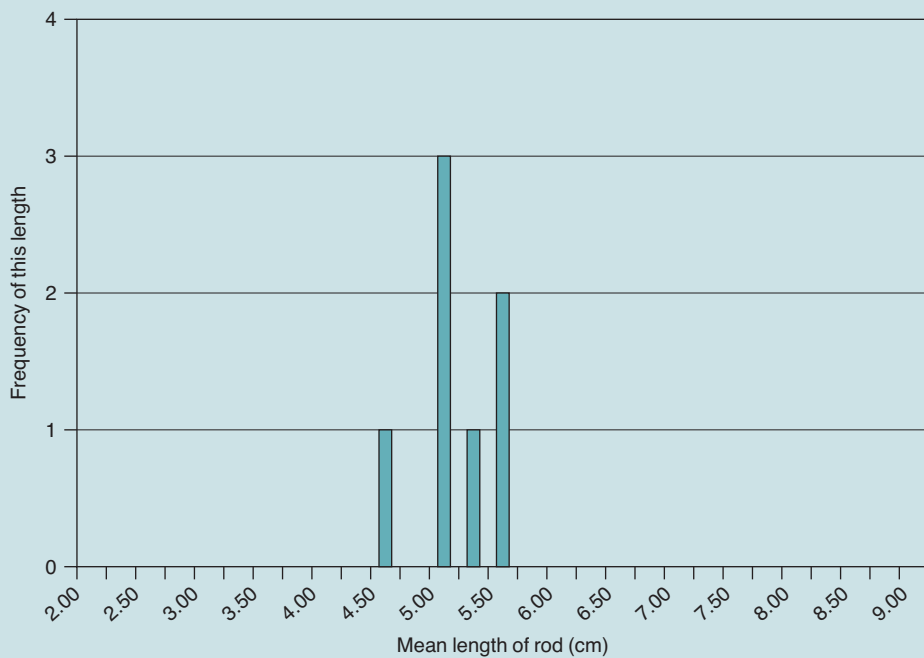
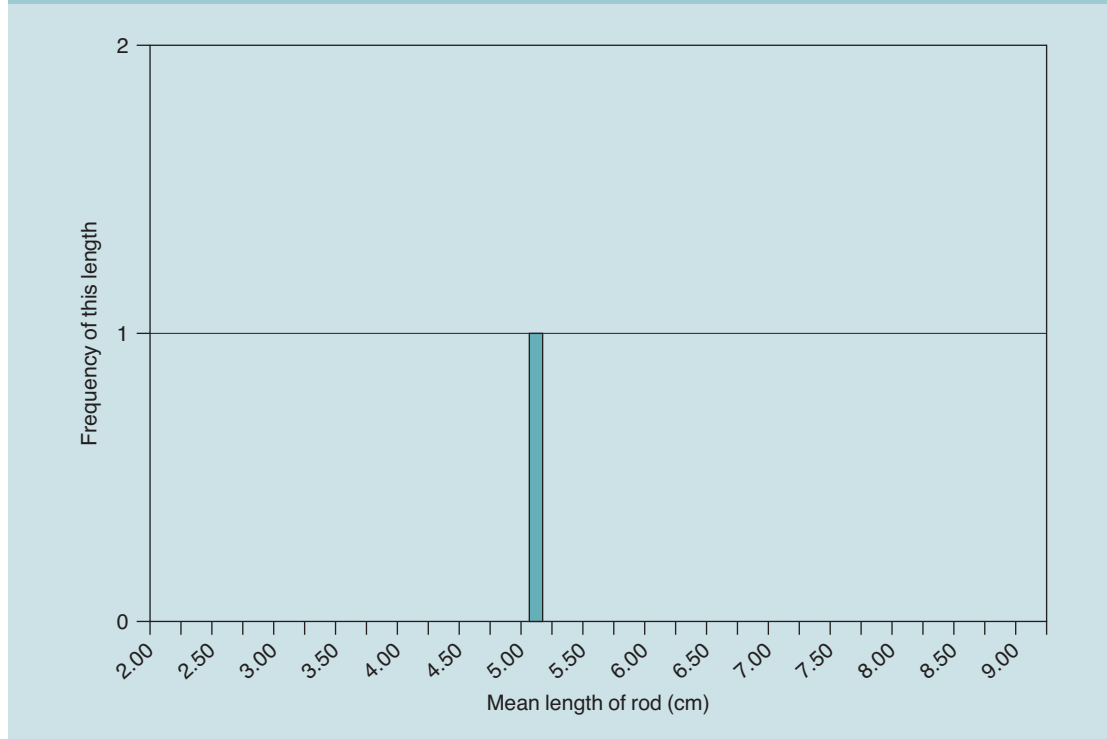


Figure 6.9 Samples of size 7 taken from a population of size 7.



the population mean,  $\mu$ . For quality control purposes an inspector takes 10 random samples from the production line in order to verify that the weight of the chocolate is according to specifications. Each sample contains 15 chocolate bars. Each bar in the sample is weighed and these individual weights, and the mean weight of each sample, are recorded. For example, if we consider sample No. 1, the weight of the 1st bar is 100.16 g, the weight of the 2nd bar is 99.48 g, and the weight of the 15th bar is 98.56 g. The mean weight of this first sample,  $\bar{x}_1$ , is 99.88 g. The mean weight of the 10th sample,  $\bar{x}_{10}$ , is 100.02 g. The mean value of the means of all the 10 samples,  $\bar{\bar{x}}$  is 99.85 g. The values of  $\bar{x}$  plotted in a frequency distribution would give a sampling distribution of the means (though only 10 values are insufficient to show a correct distribution).

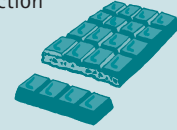
### Sample size and shape of the sampling distribution of the means

We might ask, what is the **shape of the sampling distribution of the means**? From statistical experiments the following has been demonstrated:

- For most population distributions, regardless of their shape, the sampling distribution of the means of samples taken at random from the population will be approximately normally distributed if samples of at least a size of 30 units each are withdrawn.
- If the population distribution is symmetrical, the sampling distribution of the means of samples taken at random from the population will be approximately normal if samples of at least 15 units are withdrawn.
- If the population is normally distributed, the sampling distribution of the means of samples

Table 6.5 Sampling chocolate bars.

- Company is producing a lot (population) of 500,000 chocolate bars
- Nominal weight of each chocolate bar is 100 g
- To verify the weight of the population, an inspector takes 10 random samples from production
- Each sample contains 15 slabs of chocolate
- Mean value of each sample is determined. This is  $\bar{x}$
- Mean value of all the  $\bar{x}$  is  $\bar{\bar{x}}$  or 99.85 g
- A distribution can be plotted with  $\bar{x}$  on the x-axis. The mean will be  $\bar{\bar{x}}$ .



	Sample number										
	1	2	3	4	5	6	7	8	9	10	
1	100.16	100.52	101.2	101.15	98.48	98.31	101.85	101.34	98.56	99.27	
2	99.48	98.3	101.23	101.3	98.75	99.18	99.74	101.38	101.31	101.5	
3	100.66	99.28	98.39	101.61	99.84	100.47	99.72	101.09	101.61	101.62	
4	98.93	98.01	98.06	99.07	98.38	98.3	98.76	98.89	101.26	100.84	
5	98.25	98.42	98.94	99.71	99.42	99.09	100	98.08	98.03	98.94	
6	98.06	99.19	100.53	99.78	99.23	98.23	101.42	101.5	99.74	98.94	
7	100.39	100.15	98.81	98.12	100.98	100.64	98.1	100.44	99.66	99.65	
8	101.16	99.6	99.79	101.58	100.82	98.71	100.49	101.7	98.8	98.82	
9	100.03	98.89	99.07	98.03	101.51	101.23	100.54	100.84	99.04	99.96	
10	101.27	101.94	98.39	100.77	100.17	100.99	101.66	98.4	100.61	100.95	
11	99.18	98.34	99.61	98.6	101.56	99.24	101.68	99.22	99.2	99.86	
12	101.77	100.8	99.66	98.84	100.55	98.13	99.13	99.34	100.52	98.11	
13	99.07	98.79	101.18	100.46	101.59	98.27	98.81	101.23	98.8	100.85	
14	101.17	101.02	99.57	100.3	101.87	98.16	101.73	99.98	99.26	99.17	
15	98.56	98.93	101.27	98.55	99.04	101.35	99.89	98.24	98.87	101.84	
$\bar{x}$	99.88	99.48	99.71	99.86	100.15	99.35	100.23	100.11	99.68	100.02	$\bar{\bar{x}} = 99.85$

taken at random from the population will be normally distributed regardless of the sample size withdrawn.

The practicality of these relationships with the central limit theory is that by sampling, either from non-normal populations or normal populations, inferences can be made about the population parameters without having information about the shape of the population distribution other than the information obtained from the sample.

### Variability and sample size

Consider a large organization such as a government unit that has over 100,000 employees. This is a large enough number so that it can be

considered infinite. Assume that the distribution of the employee salaries is considered normal with an average salary of \$40,000. Sampling of individual salaries is made using random computer selection:

- Assume a random sample of just one salary value is selected that happens to be \$90,000. This value is a long way from the mean value of \$40,000.
- Assume now that random samples of two salaries are taken which happen to be \$60,000 and \$90,000. The average of these is \$75,000  $[(60,000 + 90,000)/2]$ . This is still far from \$40,000 but closer than in the case of a single sample.



- If now random samples of five salaries \$60,000, \$90,000, \$45,000, \$15,000, and \$20,000 come up, the mean value of these is \$46,000 or closer to the population average of \$40,000.

Thus, by taking larger samples there is a higher probability of making an estimate close to the population parameter. Alternatively, increasing the sample size reduces the spread or variability of the average value of the samples taken.

### Sample mean and the standard error

The mean of a sample is  $\bar{x}$  and the mean of all possible samples withdrawn from the population is  $\bar{\bar{x}}$ . From the central limit theory, the mean of the entire sample means taken from the population can be considered equal to the population mean,  $\mu_x$ :

$$\bar{\bar{x}} = \mu_x \quad 6(i)$$

And because of this relationship in equation 6(i), the arithmetic mean of the sample is said to be an unbiased estimator of the population mean.

By the central limit theory, the **standard deviation of the sampling distribution**,  $\sigma_{\bar{x}}$ , is related to the population standard deviation,  $\sigma_x$ , and the sample size,  $n$ , by the following relationship:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \quad 6(ii)$$

This indicates that as the size of the sample increases, the standard deviation of the sampling distribution decreases. The standard deviation of the sampling distribution is more usually referred to as the **standard error of the sample means**, or more simply the **standard error** as it represents the error in our sampling experiment. For example, going back to our illustration of the salaries of the government employees, if we take a series of samples from the employees and measure each time, the  $\bar{x}$  value of salaries, we

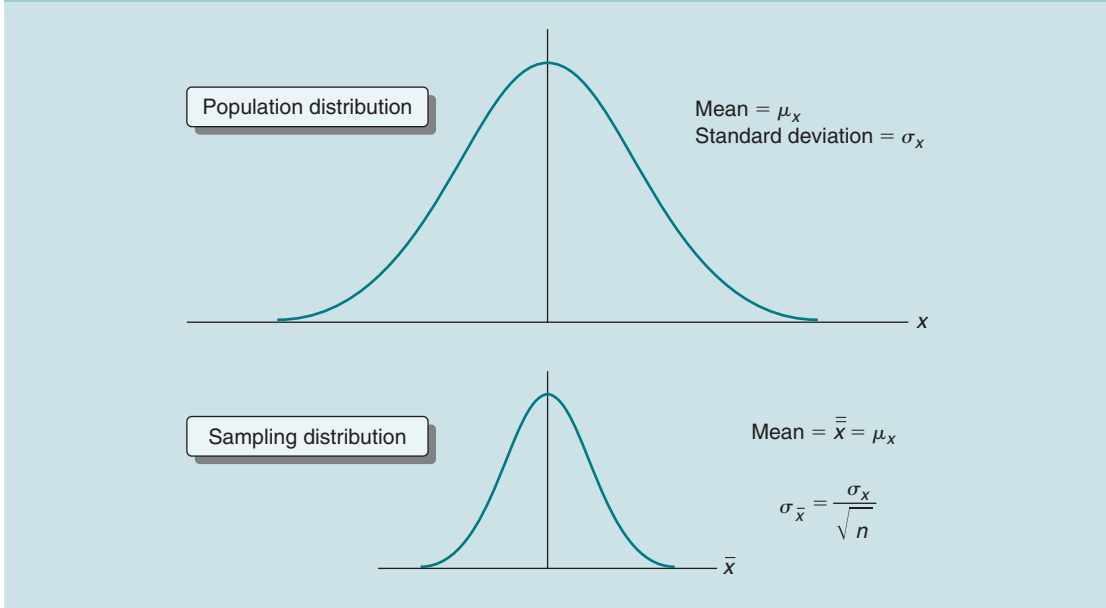
will almost certainly have different values each time simply because the chances are that our salary numbers in our sample will be different. That is, the difference between each sample, among the several samples, and the population causes variability in our analysis. This variability, as measured by the standard error of equation 6(ii), is due to the chance or **sampling error** in our analysis between the samples we took and the population. The standard error indicates the magnitude of the chance error that has been made, and also the accuracy when using a sample statistic to estimate the population parameter. A distribution of sample means that has less variability, or less spread out, as evidenced by a small value of the standard error, is a better estimator of the population parameter than a distribution of sample means that is widely dispersed with a larger standard error.

As a comparison to the standard error, we have a standard deviation of a population. This is not an error but a deviation that is to be expected since by their very nature, populations show variation. There are variations in the age of people, variations in the volumes of liquid in cans of soft drinks, variations in the weights of a nominal chocolate bar, variations in the per capita income of individuals, etc. These comparisons are illustrated in Figure 6.10, which shows the shape of a normal distribution with its standard deviation, and the corresponding profile of the sample distribution of the means with its standard error.

### Sampling for the Means from an Infinite Population

An **infinite population** is a collection of data that has such a large size that **sampling from an infinite population** involving removing or destroying some of the data elements does not significantly impact the population that remains.

Figure 6.10 Population distribution and the sampling distribution.



### Modifying the normal transformation relationship

In Chapter 5, we have shown that the standard relationship between the mean,  $\mu_x$ , the standard deviation,  $\sigma_x$ , and the random variable,  $x$ , in a normal distribution is as follows:

$$z = \frac{x - \mu_x}{\sigma_x} \quad 5(\text{ii})$$

An analogous relationship holds for the sampling distribution as shown in the lower distribution of Figure 6.10 where now:

- the random variable  $x$  is replaced by the sample mean  $\bar{x}$ ;
- the mean value  $\mu_x$  is replaced by the sample mean  $\bar{x}$ ;
- the standard deviation of the normal distribution,  $\sigma_x$ , is replaced by the standard deviation of the sample distribution or the sample error,  $\sigma_{\bar{x}}$ .

The standard equation for the sampling distribution of the means now becomes,

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{\bar{x} - \bar{\bar{x}}}{\sigma_{\bar{x}}} \quad 6(\text{iii})$$

Substituting from equations 6(i) to 6(iii), the standard equation then becomes,

$$z = \frac{\bar{x} - \bar{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_x}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}} \quad 6(\text{iv})$$

This relationship can be used using the four normal Excel functions already presented in Chapter 4, except that now the mean value of the sample mean,  $\bar{x}$ , replaces the random variable,  $x$ , of the population distribution, and the standard error of the sampling distribution  $\sigma_x / \sqrt{n}$  replaces the standard deviation of the population. The following application illustrates the use of this relationship.

## Application of sampling from an infinite normal population: *Safety valves*

A manufacturer produces safety pressure valves that are used on domestic water heaters. In the production process, the valves are automatically preset so that they open and release a flow of water when the upstream pressure in a heater exceeds 7 bars. In the manufacturing process there is a tolerance in the setting of the valves and the release pressure of the valves follows a normal distribution with a standard deviation of 0.30 bars.

1. *What proportion of randomly selected valves has a release pressure between 6.8 and 7.1 bars?* Here we are only considering a single valve, or a sample of size 1, from the population between 6.8 and 7.1 bars on either side of the mean. From equation 5(ii) when  $x = 6.8$  bars,

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{6.8 - 7.0}{0.3} = -\frac{0.2}{0.3} = -0.6667$$

From [function NORMSDIST] in Excel this gives an area from the left end of the curve of 25.25%.

From equation 5(ii) when  $x = 7.1$  bars,

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{7.1 - 7.0}{0.3} = \frac{0.1}{0.3} = 0.3333$$

From [function NORMSDIST] in Excel this gives a value from the left end of the curve of 63.06%. Thus, the probability that a randomly selected valve has a release pressure between 6.8 and 7.1 bars is  $63.06 - 25.25 = 37.81\%$ .

2. *If many random samples of size eight were taken, what proportion of sample means would have a release pressure between 6.8 and 7.1 bars?* Here now we are sampling from the normal population with a sample size of 8. Using equation 6(ii) the standard error is,

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{0.3}{\sqrt{8}} = \frac{0.3}{2.8284} = 0.1061$$

Using this value in equation 6(iv) when  $\bar{x} = 6.8$  bars,

$$\begin{aligned} z &= \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}} = \frac{6.8 - 7.0}{0.1061} = -\frac{0.2}{0.1061} \\ &= -1.8850 \end{aligned}$$

From [function NORMSDIST] in Excel using the standard error in place of the standard deviation, gives the area under the curve from the left of 2.97%.

Again from equation 6(iv) when  $\bar{x} = 7.1$  bars,

$$z = \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}} = \frac{7.1 - 7.0}{0.1061} = \frac{0.1}{0.1061} = 0.9425$$

From [function NORMSDIST] in Excel using the standard error in place of the standard deviation, gives the area under the curve from the left of 82.71%. Thus, the proportion of sample means that would have a release pressure between 6.8 and 7.1 bars is  $82.71 - 2.97 = 79.74\%$ .

3. *If many random samples of size 20 were taken, what proportion of sample means would have a release pressure between 6.8 and 7.1 bars?*

Here now we are sampling from the population with a sample size of 20. Using equation 6(ii) the standard error is,

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{0.3}{\sqrt{20}} = \frac{0.3}{4.4721} = 0.0671$$

Using this value in equation 6(iv) when  $\bar{x} = 6.8$  bars,

$$\begin{aligned} z &= \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}} = \frac{6.8 - 7.0}{0.0671} = -\frac{0.2}{0.0671} \\ &= -2.9814 \end{aligned}$$

From [function NORMSDIST] using the standard error in place of the standard deviation,

gives the area under the curve from the left of 0.14%.

Again from equation 6(iv) when  $\bar{x} = 7.1$  bars,

$$z = \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}} = \frac{7.1 - 7.0}{0.0671} = \frac{0.1}{0.0671} = 1.4903$$

From [function NORMSDIST] using the standard error in place of the standard deviation, gives the area under the curve from the left of 93.20%. Thus, the proportion of sample means that would have a release pressure between 6.8 and 7.1 bars is  $93.20 - 0.14 = 93.06\%$ .

4. If many random samples of size 50 were taken, what proportion of sample means would have a release pressure between 6.8 and 7.1 bars?

Here now we are sampling from the population with a sample size of 50. Using equation 6(ii) the standard error is,

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{0.3}{\sqrt{50}} = \frac{0.3}{7.0711} = 0.0424$$

Using this value in equation 6(iv) when  $\bar{x} = 6.8$  bars,

$$z = \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}} = \frac{6.8 - 7.0}{0.0424} = -\frac{0.2}{0.0424} = -4.714$$

From [function NORMSDIST] using the standard error in place of the standard deviation, gives the area under the curve from the left of 0%.

Again from equation 7(v) when  $\bar{x} = 7.1$  bars,

$$z = \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}} = \frac{7.1 - 7.0}{0.0424} = \frac{0.1}{0.0424} = 2.3585$$

From [function NORMSDIST] using the standard error in place of the standard deviation, gives the area under the curve from the left of 99.08%. Thus, the proportion of sample means that would have a release pressure

Table 6.6 Example, safety valves.

Sample size	1	8	20	50
Standard error, $\sigma_x / \sqrt{n}$	0.3000	0.1061	0.0671	0.0424
Proportion between 6.8 and 7.1 bars (%)	37.81	79.74	93.20	99.08

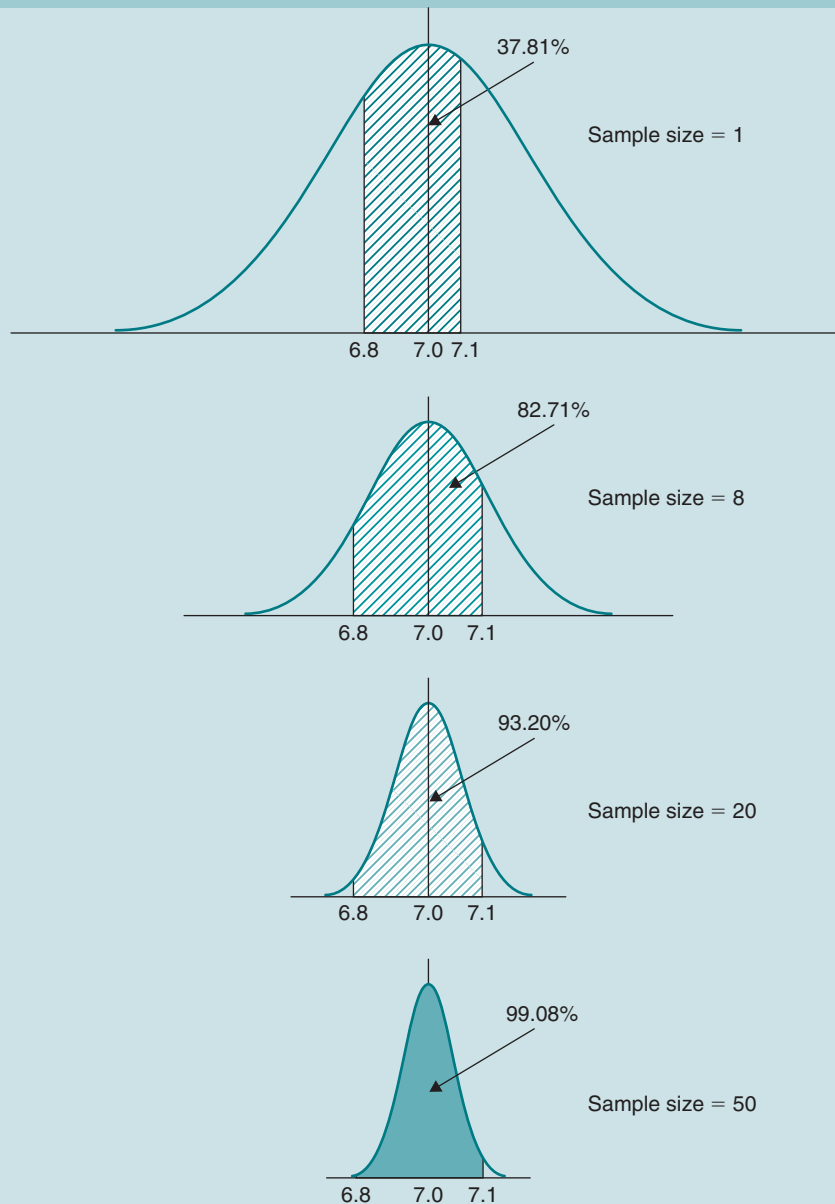
between 6.8 and 7.1 bars is  $99.08 - 0.00 = 99.08\%$ .

To summarize this situation we have the results in Table 6.6 and the concept is illustrated in the distributions of Figure 6.11. What we observe is that not only the standard error decreases as the sample size increases but there is a larger proportion between the values of 6.8 and 7.1 bars. That is a larger cluster around the mean or target value of 7.0 bars. Alternatively, as the sample size increases there is a smaller dispersion of the values. For example, in the case of a sample size of 1 there is 37.81% of the data clustered around the values of 6.8 and 7.1 bars which means that there is 62.19% ( $100\% - 37.81\%$ ) not clustered around the mean. In the case of a sample size of 50 there is 99.08% clustered around the mean and only 0.92% ( $100\% - 99.08\%$ ) not clustered around the mean. Note, in applying these calculations the assumption is that the sampling distributions of the mean follow a normal distribution, and the relation of the central limit theory applies. As in the calculations for the normal distribution, if we wish we can avoid always calculating the value of  $z$  by using the [function NORMSDIST].

## Sampling for the Means from a Finite Population

A **finite population** is a collection of data that has a stated, limited, or small size. It implies that if one piece of the data from the population is

Figure 6.11 Example, safety valves.



destroyed, or removed, there would be a significant impact on the data that remains.

### Modification of the standard error

If the population is considered finite, that is the size is relatively small and there is **sampling with**

**replacement** (after each item is sampled it is put back into the population), then we can use the equation for the standard error already presented,

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \quad 6(ii)$$

However, if we are **sampling without replacement**, the standard error of the mean is modified by the relationship,

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad 6(v)$$

Here the term,

$$\sqrt{\frac{N-n}{N-1}} \quad 6(vi)$$

is the **finite population multiplier**, where  $N$  is the population size, and  $n$  is the size of the sample. This correction is applied when the ratio of  $n/N$  is greater than 5%. In this case, equation 6(iv) now becomes,

$$z = \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}} = \frac{\bar{x} - \mu_x}{\frac{\sigma_x}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}} \quad 6(vii)$$

The application of the finite population multiplier is illustrated in the following application exercise.

### Application of sampling from a finite population: *Work week*

A firm has 290 employees and records that they work an average of 35 hours/week with a standard deviation of 8 hours/week.

1. What is the probability that an employee selected at random will be working between  $\pm 2$  hours/week of the population mean?

In this case, again we have a single unit (an employee) taken from the population where the standard deviation  $\sigma_x$  is 8 hours/week. Thus,  $n = 1$  and  $N = 290$ .  $n/N = 1/290 = 0.34\%$  or less than 5% and so the population multiplier is not needed.

We know that the difference between the random variable and the population,  $(x - \mu_x)$  is equal to  $\pm 2$ . Thus, assuming that the population follows a normal distribution,

then from equation 6(iii) for a value of  $(x - \mu_x) = +2$ ,

$$z = \frac{x - \mu_x}{\sigma_x} = +\frac{2}{8} = 0.2500$$

From **[function NORMSDIST]** in Excel, the area under the curve from the left to a value of  $z$  of 0.2500 is 59.87%.

For a value of  $(x - \mu_x) = -2$  we have again from equation 6(ii),

$$z = \frac{x - \mu_x}{\sigma_x} = -\frac{2}{8} = -0.2500$$

Or we could have simply concluded that  $z$  is  $-0.2500$  since the assumption is that the curve follows a normal distribution, and a normal distribution is, by definition, symmetrical.

From **[function NORMSDIST]** in Excel, the area under the curve from the left to a value of  $z$  of  $-0.2500$  is 40.13%.

Thus, the probability that an employee selected at random will be working between  $\pm 2$  hours/week is,

$$59.87 - 40.13 = 19.74\%$$

2. If a sample size of 19 employees is taken, what is the probability that the sample means lies between  $\pm 2$  hours/week of the population mean?

In this case, again we have a sample,  $n$ , of size 19 taken from a population,  $N$ , of size 290. The ratio  $n/N$  is,

$$\frac{n}{N} = \frac{19}{290} = 0.0655 \text{ or } 6.55\% \text{ of the population}$$

This ratio is greater than 5% and so we use the finite population multiplier in order to calculate the standard error. From equation 6(vi),

$$\begin{aligned}\sqrt{\frac{N-n}{N-1}} &= \sqrt{\frac{(290-19)}{(290-1)}} = \sqrt{\frac{271}{289}} \\ &= \sqrt{0.9377} = 0.9684\end{aligned}$$

From equation 6(v) the corrected standard error of the distribution of the mean is,

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{8}{\sqrt{19}} * 0.9684 \\ &= \frac{8 * 0.9684}{4.3589} = 1.7773\end{aligned}$$

From equation 6(vii) where now  $\bar{x} - \mu_x = \pm 2$ . For  $\bar{x} - \mu_x = +2$  we have,

$$z = \frac{\bar{x} - \mu_x}{\frac{\sigma_x}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}} = \frac{2}{1.7773} = 1.1253$$

From [function NORMSDIST] in Excel, the area under the curve from the left to a value of  $z$  of 1.1253 is 86.98%.

From equation 6(vii) for  $\bar{x} - \mu_x = -2$  we have,

$$z = \frac{\bar{x} - \mu_x}{\frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}} = \frac{2}{1.7773} = -1.1253$$

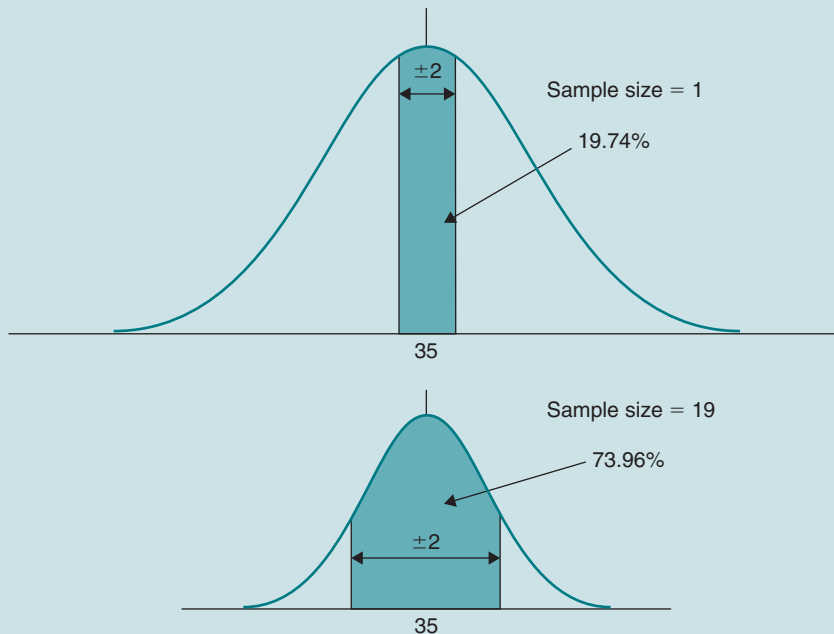
From [function NORMSDIST] in Excel, the area under the curve from the left to a value of  $z$  of  $-1.1253$  is 13.02%.

Thus, the probability that the sample means lie between  $\pm 2$  hours/week is,

$$86.98 - 13.02 = 73.96\%.$$

Note that 73.96% is greater than 19.74%, obtained for a sample of size 1, because as we increase the sample size, the sampling distribution of the means is clustered around the population mean. This concept is illustrated in Figure 6.12.

Figure 6.12 Example, work week.





## Sampling Distribution of the Proportion

In sampling we may not be interested in an absolute value but in a proportion of the population. For example, what proportion of the population will vote conservative in the next United Kingdom elections? What proportion of the population in Paris, France has a salary more than €60,000 per year? What proportion of the houses in Los Angeles country in United States has a market value more than \$500,000? In these cases, we have established a binomial situation. In the United Kingdom elections either a person votes conservative or he or she do not. In Paris either an individual earns a salary more than €60,000/year, or they do not. In Los Angeles country, either the houses have a market value greater than \$500,000 or they do not. In these types of situations we use sampling for proportions.

### Measuring the sample proportion

When we are interested in the proportion of the population, the procedure is to sample from the population and then again use inferential statistics to draw conclusions about the population proportion. The sample proportion,  $\bar{p}$ , is the ratio of that quantity,  $x$ , taken from the sample having the desired characteristic divided by the sample size,  $n$ , or,

$$\bar{p} = \frac{x}{n} \quad 6(\text{viii})$$

For example, assume we are interested in people's opinion of gun control. We sample 2,000 people from the State of California and 1,450 say they are for gun control. The proportion in the sample that says they are for gun control is thus 72.50% (1,450/2,000). We might extend this sample experiment further and say that 72.50% of the population of California is for gun control or even go further and conclude that 72.50% of

the United States population is for gun control. However, these would be very uncertain conclusions since the 2,000-sample size may be neither representative of California, and probably not of the United States. This experiment is binomial because either a person is for gun control or is not. Thus, the proportion in the sample that is against gun control is 27.50% (100% – 72.50%).

### Sampling distribution of the proportion

In our sampling process for the proportion, assume that we take a random sample and measure the proportion having the desired characteristic and this is  $\bar{p}_1$ . We then take another sample from the population and we have a new value  $\bar{p}_2$ . If we repeat this process then we possibly will have different values of  $\bar{p}$ . The probability distribution of all possible values of the sample proportion,  $\bar{p}$ , is the **sampling distribution of the proportion**. This is analogous to the sampling distribution of the means,  $\bar{x}$ , discussed in the previous section.

### Binomial concept in sampling for the proportion

If there are only two possibilities in an outcome then this is binomial. In the binomial distribution the mean number of successes,  $\mu$ , for a sample size,  $n$ , with a characteristic probability of success,  $p$ , is given by the relationship presented in Chapter 4:

$$\mu = np \quad 4(\text{xv})$$

Dividing both sides of this equation by the sample size,  $n$ , we have,

$$\frac{\mu}{n} = \frac{np}{n} = p \quad 6(\text{ix})$$

The ratio  $\mu/n$  is now the **mean proportion of successes** written as  $\mu_{\bar{p}}$ . Thus,

$$\mu_{\bar{p}} = p \quad 6(\text{x})$$



Again from Chapter 4, the standard deviation of binomial distribution is given by the relationship,

$$\sigma = \sqrt{npq} = \sqrt{np(1-p)} \quad 4(\text{xvii})$$

where the value  $q = 1 - p$

And again dividing by  $n$ ,

$$\frac{\sigma}{n} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{pq}{n^2}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}} \quad 6(\text{xi})$$

where the ratio  $\sigma/n$  is the **standard error of the proportion**,  $\sigma_{\bar{p}}$ , and thus,

$$\sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}} \quad 6(\text{xii})$$

From equation 6(iv) we have the relationship,

$$z = \frac{\bar{x} - \bar{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_x}{\sigma_{\bar{x}}} \quad 6(\text{iv})$$

From Chapter 5, we can use the normal distribution to approximate the binomial distribution when the following two conditions apply:

$$np \geq 5 \quad 5(\text{iv})$$

$$n(1-p) \geq 5 \quad 5(\text{v})$$

That is, the products  $np$  and  $n(1-p)$  are both greater or equal to 5. Thus, if these criteria apply then by substituting in equation 6(iv) as follows,

$\bar{x}$ , the sample mean by the average sample proportion,  $\bar{p}$

$\mu_x$ , the population mean by the population proportion,  $p$

$\sigma_{\bar{x}}$ , the standard error of the sample means by  $\sigma_{\bar{p}}$ , the standard error of the proportion

and using the relationship developed in equation 6(iii), we have,

$$z = \frac{\bar{x} - \bar{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_x}{\sigma_{\bar{x}}} = \frac{\bar{p} - p}{\sigma_{\bar{p}}} \quad 6(\text{xiii})$$

Since from equation 6(xii),

$$\sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}} \quad 6(\text{xiv})$$

Then,

$$z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \quad 6(\text{xv})$$

Alternatively, we can say that the difference between the sample proportion  $\bar{p}$  and the population proportion  $p$  is,

$$\bar{p} - p = z \sqrt{\frac{p(1-p)}{n}} \quad 6(\text{xvi})$$

The application of this relationship is illustrated as follows.

### Application of sampling for proportions: *Part-time workers*

The incidence of part-time working varies widely across Organization for Cooperation and Development (OECD) countries. The clear leader is the Netherlands where part-time employment accounts for 33% of all jobs.<sup>3</sup>

1. If a sample of 100 people of the work force were taken in the Netherlands, what proportion between 25% and 35%, in the sample, would be part-time workers?

Now, the sample size is 100 and so we need to test again whether we can use the normal probability assumption by using equations 5(iv) and 5(v). Here  $p$  is still 33%, or 0.33 and  $n$  is 100, thus from equation 5(iv),

$$np = 100 * 0.33 = 33 \text{ or greater than } 5$$

From equation 5(v),

$$n(1-p) = 100(1-0.33) = 67 \text{ or again greater than } 5$$

<sup>3</sup>Economic and financial indicators, *The Economist*, 20 July 2002, p. 88.

Thus, we can apply the normal probability assumption.

The population proportion  $p$  is 33%, or 0.33, and thus from equation 6(xiv) the standard error of the proportion is,

$$\begin{aligned}\sigma_{\bar{p}} &= \sqrt{\frac{0.33(1-0.33)}{100}} = \sqrt{\frac{0.33 * 0.67}{100}} \\ &= \sqrt{0.0022} = 0.0469\end{aligned}$$

The lower sample proportion,  $\bar{p}$ , is 25%, or 0.25 and thus from equation 6(xiii),

$$\begin{aligned}z &= \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0.25 - 0.33}{0.0469} = -\frac{0.0800}{0.0469} \\ &= -1.7058\end{aligned}$$

From [function NORMSDIST] in Excel, the area under the curve from the left to a value of  $z$  of  $-1.7058$  is 4.44%.

The upper sample proportion,  $\bar{p}$ , is 35%, or 0.35 and thus from equation 6(xiii),

$$\begin{aligned}z &= \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0.35 - 0.33}{0.0469} = -\frac{0.02}{0.0469} \\ &= 0.4264\end{aligned}$$

From [function NORMSDIST] in Excel, the area under the curve from the left to a value of  $z$  of 0.4264 is 66.47%.

Thus, the proportion between 25% and 35%, in the sample, that would be part-time workers is,

$$66.47 - 4.44 = 62.03\% \text{ or } 0.6203$$

2. If a sample of 200 people of the work force were taken in the Netherlands, what proportion between 25% and 35%, in the sample, would be part-time workers?

First, we need to test whether we can use the normal probability assumption by using equations 5(iv) and 5(v). Here  $p$  is 33%, or 0.33 and  $n$  is 200, thus from equation 5(iv),

$$np = 200 * 0.33 = 66 \text{ or greater than } 5$$

From equation 5(v)

$$n(1 - p) = 200(1 - 0.33) = 134 \text{ or again greater than } 5$$

Thus, we can apply the normal probability assumption.

The population proportion  $p$  is 33%, or 0.33, and thus from equation 6(xiv) the standard error of the proportion is,

$$\begin{aligned}\sigma_{\bar{p}} &= \sqrt{\frac{0.33(1-0.33)}{200}} = \sqrt{\frac{0.33 * 0.67}{200}} \\ &= \sqrt{0.0011} = 0.0332\end{aligned}$$

The lower sample proportion,  $\bar{p}$ , is 25%, or 0.25 and thus from equation 6(xiii),

$$\begin{aligned}z &= \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0.25 - 0.33}{0.0332} = -\frac{0.0800}{0.0332} \\ &= -2.4061\end{aligned}$$

From [function NORMSDIST] in Excel the area under the curve from the left to a value of  $z$  of  $-2.4061$  is 0.81%.

The upper sample proportion,  $\bar{p}$ , is 35%, or 0.35 and thus from equation 6(xiii),

$$\begin{aligned}z &= \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0.35 - 0.33}{0.0332} = -\frac{0.02}{0.0332} \\ &= 0.6015\end{aligned}$$

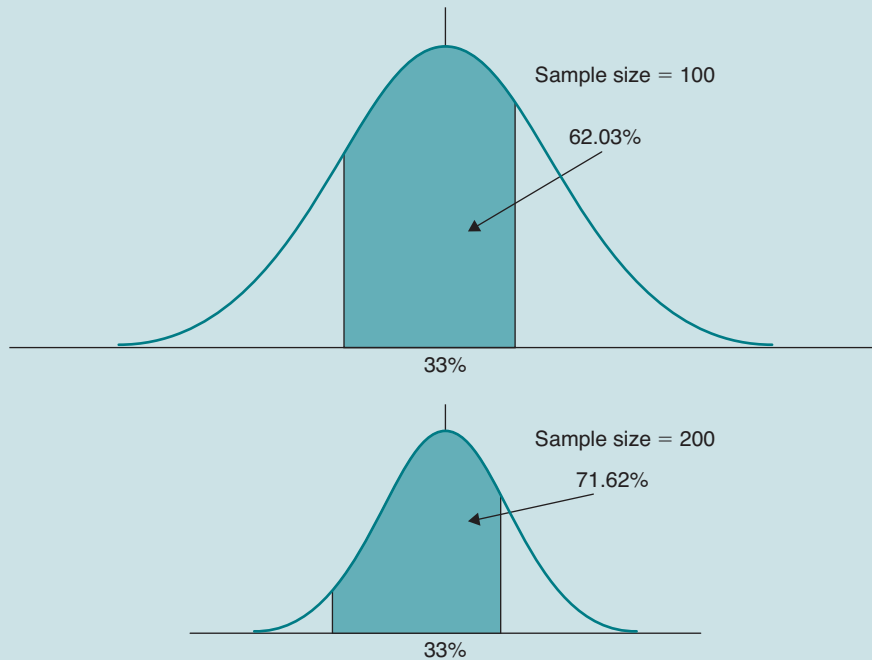
From [function NORMSDIST] in Excel, the area under the curve from the left to a value of  $z$  of 0.6015 is 72.63%.

Thus, the proportion between 25% and 35% in a sample size of 200 that would be part-time workers is,

$$72.63 - 0.81 = 71.82\% \text{ or } 0.7182$$

Note that again this value is larger than in the first situation since the sample size was 100 rather than 200. As for the mean, as the sample size increases the values will cluster around the mean value of the population. Here the mean value of the proportion for the population is 33% and the

Figure 6.13 Example, part-time workers.



sample proportions tested were 25% and 35% or both on the opposite side of the mean value of the proportion. This concept is illustrated in Figure 6.13.

## Sampling Methods

The purpose of sampling is to make reliable estimates about a population. It is usually impossible, and too expensive, to sample the whole population so that when a sampling experiment is developed it should as closely as possible parallel the population conditions. As the box opener “The sampling experiment was badly designed!” indicates, the sampling experiment to determine voter intentions was obviously badly designed. This section gives considerations when undertaking sampling experiments.

## Bias in sampling

When you sample to make estimates of a population you must avoid **bias** in the sampling experiment. Bias is favouritism, purposely or unknowingly, present in the sample data that gives lopsided, misleading, or unrepresentative results. For example, you wish to obtain the voting intentions of the people in the United Kingdom and you sample people who live in the West End of London. This would be biased as the West End is pretty affluent and the voters sampled are more likely to vote Tory (Conservative). To measure the average intelligence quotient (IQ) of all the 18-year-old students in a country you take a sample of students from a private school. This would be biased because private school students often come from high-income families and their education level is higher. To measure the average income of residents of Los Angeles, California you take

a sample of people who live in Santa Monica. This would be biased as people who live in Santa Monica are wealthy.

## Randomness in your sample experiment

A **random sample** is one where each item in the population has an equal chance of being selected. Assume a farmer wishes to determine the average weight of his 200 pigs. He samples the first 12 who come when he calls. They are probably the fittest – thus thinner than the rest! Or, a hotel manager wishes to determine the quality of the maid service in his 90-room hotel. The manager samples the first 15. If the maid works in order, then the first 15 probably were more thoroughly cleaned than the rest – the maid was less tired! These sampling experiments are not random and probably they are not representative of the population.

In order to perform random sampling, you need a framework for your sampling experiment. For example, as an auditor you might wish to analyse 10% of the financial accounts of the firm to see if they conform to acceptable accounting practices. A business might want to sample 15% of its clients to obtain the level of customer satisfaction. A hotel might want to sample 12% of the condition of its hotel rooms to obtain a quality level of its operation.

## Excel and random sampling

In Excel there are two functions for generating random numbers, **[function RAND]** that generates a random number between 0 and 1, and **[function RANDBETWEEN]** that generates a random number between the lowest and highest number that you specify. You first create a random number in a cell and copy this to other cells. Each time you press the function key F9 the random number will change.

Suppose that as an auditor you have 630 accounts in your population and you wish to examine 10% of these accounts or 63. You

**Table 6.7** Table of 63 random numbers between 1 and 630.

389	386	309	75	174	350	314	70	219
380	473	249	56	323	270	147	605	426
440	285	353	339	173	583	620	624	331
84	219	78	560	272	347	171	476	589
396	285	306	557	300	183	406	114	485
105	161	528	438	510	288	437	374	368
512	49	368	25	75	36	415	251	308

**Table 6.8** Table of 12 random numbers between 1 and 200.

142	26	178	146	72	7
156	95	176	144	113	194

number the accounts from 1 to 630. You then generate 63 random numbers between 1 and 630 and you examine those accounts whose numbers correspond to the numbers generated by the random number function. For example, the matrix in Table 6.7 shows 63 random numbers within the range 1 to 630. Thus, you would examine those accounts corresponding to those numbers.

The same procedure would apply to the farmer and his pigs. Each pig would have identification, either a tag, tattoo, or embedded chip giving a numerical indication from 1 to 200. The farmer would generate a list of 12 random numbers between 1 and 200 as indicated in Table 6.8, and weigh those 12 pigs that correspond to those numbers.

## Systematic sampling

When a population is relatively homogeneous and you have a listing of the items of interest such as invoices, a fleet of company cars, physical units such as products coming off a production

line, inventory going into storage, a stretch of road, or a row of houses, then **systematic sample** may be appropriate. You first decide at what frequency you need to take a sample. For example, if you want a 4% sample you analyse every 25th unit – 4% of 100 is 25. If you want a 5% sample you analyse every 20th unit – 5% of 100 is 20. If you want a 0.5% sample you analyse every 200 units – 0.5% of 100 is 200, etc. Care must be taken in using systematic sampling that no bias occurs where the interval you choose corresponds to a pattern in the operation. For example, you use systematic sampling to examine the filling operation of soft drink machine. You sample every 25th can of drink. It so happens that there are 25 filling nozzles on the machine. In this case, you will be sampling a can that has been filled from the same nozzle. The United States population census, undertaken every 10 years, is a form of systematic sample where although every household receives a survey datasheet to complete, every 10th household receives a more detailed survey form to complete.

### Stratified sampling

The technique of **stratified sampling** is useful when the population can be divided into relatively homogeneous groups, or strata, and random sampling is made only on the strata of interest. For example, the strata may be students, people of a certain age range, male or female, married or single households, socio-economic levels, affiliated with the labour or conservative party, etc. Stratified sampling is used because it more

accurately reflects the characteristics of the target population. Single people of a certain socio-economic class are more likely to buy a sports car; people in the 20–25 have a different preference of music and different needs of portable phones than say those in the 50–55-age range. Stratified sampling is used when there is a small variation within each group, but a wide variation among groups. For example, teenagers in the age range 13 to 19 and their parents in the age range 40 to 50 differ very much in their tastes and ideas!

### Several strata of interest

In a given population you may have several well-defined strata and perhaps you wish to take a representative sample from this population. Consider for example, the 1st row of Table 6.9 which gives the number of employees by function in a manufacturing company. Each function is a stratum since it defines a specific activity. Suppose we wish to obtain the employees' preference of changing from the current 8 hours/day, 5 days a week to a proposed 10 hours/day, 4 days/week. In order to limit the cost and the time of the sampling experiment we decide to only survey 60 of the employees. There are a total of 1,200 employees in the firm and so 60 represents 5% of the total workforce (60/1,200). Thus, we would take a random sample of 5% of the employees from each of the departments or strata such that the sampling experiment parallels the population. The number that we would survey is given in the 2nd row of Table 6.9.

Table 6.9 Stratified sampling.

Department	Administration	Operations	Design	R&D	Sales	Accounting	Information Systems	Total
Employees	160	300	200	80	260	60	140	1,200
Sample size	8	15	10	4	13	3	7	60

### Cluster sampling

In **cluster sample** the population is divided into groups, or clusters, and each cluster is then sampled at random. For example, assume Birmingham is targeted for preference of a certain consumer product. The city is divided into clusters using a city map and an appropriate number of clusters are selected for analysis. Cluster sampling is used when there is considerable variation in each group or cluster, but groups are essentially similar. Cluster sampling, if properly designed, can provide more accurate results than simple random sampling from the population.

### Quota sampling

In market research, or market surveys, interviewers carrying out the experiment may use **quota sampling** where they have a specific target quantity to review. In this type of sampling often the population is stratified according to some criteria so that the interviewer's quota is based within these strata. For example, the interviewer may be interested to obtain information regarding a ladies fashion magazine. The interviewer conducts her survey in a busy shopping area such as London's Oxford Street. Using quota sampling, in her survey she would only interview females, perhaps less than 40, and who are elegantly dressed. This *stratification* should give a reasonable probability that the selected candidates have some interest, and thus an opinion, regarding the fashion magazine in question. If you are in an area where surveys are being carried out, it could be that you do not fit the strata desired by the interviewer. For example, you are male and the interviewer is targeting females, you appear to be over 50 and the interviewer is targeting the age group under 40, you are white and the interviewer is targeting other ethnic groups, etc.

### Consumer surveys

If your sampling experiment involves opinions say concerning a product, a concept, or a situation, then you might use a **consumer survey**, where responses are solicited from individuals who are targeted according to a well-defined sampling plan. The sampling plan would use one, or a combination of the methods above – simple random sampling, systematic, stratified, cluster, or quota sampling. The survey information is prepared on **questionnaires**, which might be sent through the mail, completed by telephone, sent by electronic mail, or requested in person. In the latter case this may be either going door-to-door, or soliciting the information in areas frequented by potential consumers such as shopping malls or busy pedestrian areas. The collected survey data, or sample, is then analysed and used to forecast or make estimates for the population from which the survey data was taken. Surveys are often used to obtain ideas about a new product, because required data is unavailable from other sources.

When you develop a consumer survey remember that it is perhaps you who have to analyse it afterwards. Thus, you should structure it so that this task is straightforward with responses that are easy to organize. Avoid open-ended questions. For example, rather than asking the question "How old are you?" give the respondent age categories as for example in Table 6.10. Here these categories are all encompassing. Alternatively, if you want to know the job of the respondent rather than asking, "What is your job?" ask the question, "Which of the following best describes your professional activity?" as for example in Table 6.11.

Table 6.10 Age range for a questionnaire.

Under 25	25–34	35–44	45–54	55–65	Over 65
-------------	-------	-------	-------	-------	------------



**Table 6.11** Which of the following best describes your professional activity?

---

Construction  
 Consulting  
 Design  
 Education  
 Energy  
 Financial services  
 Government  
 Health care  
 Hospitality  
 Insurance  
 Legal  
 Logistics  
 Manufacturing  
 Media communications  
 Research  
 Retail  
 Telecommunications  
 Tourism  
 Other (please describe)

---

This is not all encompassing but there is a category “Other” for activities that may have been overlooked.

Soliciting information from consumers is not easy, “everyone is too busy”. Postal responses have a very low response and their use has declined. Those people who do respond may not be representative in the sample. Telephone surveys give a higher return because voice contact has been obtained. However, again the sample obtained may not be representative as those contacted may be the unemployed, retirees or elderly people, or non-employed individuals who are more likely to be at home when the telephone call is made. The other segment of the population, usually larger, is not available because they are working. Though if you have access to portable phone numbers this may not apply. Electronic mail surveys give a reasonable

response, as it is very quick to send the survey back. However, the questionnaire only reaches those who have E-mail, and then those who care to respond. Person-to-person contact gives a much higher response for consumer surveys since if you are stopped in the street, a relatively large proportion of people will accept to be questioned. Consumer surveys can be expensive. There is the cost of designing the questionnaire such that it is able to solicit the correct response. There is the operating side of collecting the data, and then the subsequent analysis. Often businesses use outside consulting firms specialized in developing consumer surveys.

## Primary and secondary data

In sampling if we are responsible for carrying out the analysis, or at least responsible for designing the consumer surveys, then the data is considered **primary data**. If the sample experiment is well designed then this primary data can provide very useful information. The disadvantage with primary data is the time, and the associated cost, of designing the survey and the subsequent analysis. In some instances it may be possible to use **secondary data** in analytical work. Secondary data is information that has been developed by someone else but is used in your analytical work. Secondary data might be demographic information, economic trends, or consumer patterns, which is often available through the Internet. The advantage with secondary data, provided that it is in the public domain, is that it costs less or at best is free. The disadvantage is that the secondary data may not contain all the information you require, the format may not be ideal, and/or it may be not be up to date. Thus, there may be a trade-off between using less costly, but perhaps less accurate, secondary data, and more expensive but more reliable, primary data.

This chapter has looked at sampling covering specifically, basic relationships, sampling for the mean in infinite and finite populations, sampling for proportions, and sampling methods.

### Statistical relations in sampling for the mean

Inferential statistics is the estimate of population characteristics based on the analysis of a sample. The larger the sample size, the more reliable is our estimate of the population parameter. It is the central limit theory that governs the reliability of sampling. This theory states that as the size of the sample increases, there becomes a point when the distribution of the sample means can be approximated by the normal distribution. In this case, the mean of all sample means withdrawn from the population is equal to the population mean. Further, the standard error of the estimate in a sampling distribution is equal to the population standard deviation divided by the square root of the sample size.

### Sampling for the means for an infinite population

An infinite population is a collection of data of a large size such that by removing or destroying some data elements the population that remains is not significantly affected. Here we can modify the transformation relationship that apply to the normal distribution and determine the number of standard deviations,  $z$ , as the sample mean less the population mean divided by the standard error. When we use this relationship we find that the larger the sample size,  $n$ , the more the sample data clusters around the population mean implying that there is less variability.

### Sampling for the means from a finite population

A finite population in sampling is defined such that the ratio of the sample size to the population size is greater than 5%. This means that the sample size is large relative to the population size. When we have a finite population we modify the standard error by multiplying it by a finite population multiplier, which is the square root of the ratio of the population size minus the sample size to the population size minus one. When we have done this, we can use this modified relationship to infer the characteristics of the population parameter. Again as before, the larger the sample size, the more the data clusters around the population mean and there is less variability in the data.

### Sampling distribution of the proportion

A sample proportion is the ratio of those values that have the desired characteristics divided by the sample size. The binomial relationship governs proportions, since either values in the sample have the desired characteristics or they do not. Using the binomial relationships for the mean and the standard deviation, we can develop the standard error of the proportion. With this standard error of the proportion, and the value of the sample proportion, we can make an estimate of the population proportion in a similar manner to making an estimate of the population mean. Again, the larger the sample size, the closer is our estimate to the population proportion.



## Sampling methods

The key to correct sampling is to avoid bias, that is not taking a sample that gives lopsided results, and to ensure that the sample is random. Microsoft Excel has a function that generates random numbers between given limits. If we have a relatively homogeneous population we can use systematic sampling, which is taking samples at predetermined intervals according to the desired sample size. Stratified sampling can be used when we are interested in a well-defined strata or group. Stratified sampling can be extended when there are several strata of interest within a population. Cluster sampling is another way of making a sampling experiment when the population is divided up into manageable clusters that represent the population, and then sampling an appropriate quantity within a cluster. Quota sampling is when an interviewer has a certain quota, or number of units to analyse that may be according to a defined strata. Consumer surveys are part of sampling where respondents complete questionnaires that are sent through the post, by E-mail, completed over the phone, or face to face contact. When you construct a questionnaire for a consumer surveys, avoid having open-ended questions as these are more difficult to analyse. In sampling there is primary data, or that collected by the researcher, and secondary data that maybe in the public domain. Primary data is normally the most useful but is usually more costly to develop.

## EXERCISE PROBLEMS

### 1. Credit card

#### Situation

From past data, a large bank knows that the average monthly credit card account balance is £225 with a standard deviation of £98.

#### Required

1. What is the probability that in an account chosen at random, the average monthly balance will lie between £180 and £250?
2. What is the probability that in 10 accounts chosen at random, the sample average monthly balance will lie between £180 and £250?
3. What is the probability that in 25 accounts chosen at random, the sample average monthly balance will lie between £180 and £250?
4. Explain the differences.
5. What assumptions are made in determining these estimations?

### 2. Food bags

#### Situation

A paper company in Finland manufactures treated double-strength bags used for holding up to 20 kg of dry dog or cat food. These bags have a nominal breaking strength of  $8 \text{ kg/cm}^2$  with a production standard deviation of  $0.70 \text{ kg/cm}^2$ . The manufacturing process of these food bags follows a normal distribution.

#### Required

1. What percentage of the bags produced has a breaking strength between  $8.0$  and  $8.5 \text{ kg/cm}^2$ ?
2. What percentage of the bags produced has a breaking strength between  $6.5$  and  $7.5 \text{ kg/cm}^2$ ?
3. What proportion of the sample means of size 10 will have breaking strength between  $8.0$  and  $8.5 \text{ kg/cm}^2$ ?
4. What proportion of the sample means of size 10 will have breaking strength between  $6.5$  and  $7.5 \text{ kg/cm}^2$ ?
5. Compare the answers of Questions 1 and 3, and 2 and 4.
6. What distribution would the sample means follow for samples of size 10?

### 3. Telephone calls

#### Situation

It is known that long distance telephone calls are normally distributed with the mean time of 8 minutes, and the standard deviation of 2 minutes.

### Required

1. What is the probability that a call taken at random will last between 7.8 and 8.2 minutes?
2. What is the probability that a call taken at random will last between 7.5 and 8.0 minutes?
3. If random samples of 25 calls are selected, what is the probability that a call taken at random will last between 7.8 and 8.2 minutes?
4. If random samples of 25 calls are selected, what is the probability that a call taken at random will last between 7.5 and 8.0 minutes?
5. If random samples of 100 calls are selected, what is the probability that a call taken at random will last between 7.8 and 8.2 minutes?
6. If random samples of 100 calls are selected, what is the probability that a call taken at random will last between 7.5 and 8.0 minutes?
7. Explain the difference in the results.

## 4. Soft drink machine

### Situation

A soft drinks machine is regulated so that the amount dispensed into the drinking cups is on average 33 cl. The filling operation is normally distributed and the standard deviation is 1 cl no matter the setting of the mean value.

### Required

1. What is the volume that is dispensed such that only 5% of cups contain this amount or less?
2. If the machine is regulated such that only 5% of the cups contained 30 cl or less, by how much could the nominal value of the machine setting be reduced? In this case, on average a customer would be receiving what percentage less of beverage?
3. With a nominal machine setting of 33 cl, if samples of 10 cups are taken, what is the volume that will be exceeded by 95% of sample means?
4. There is a maintenance rule such that if the sample average content of 10 cups falls below 32.50 cl, a technician will be called out to check the machine settings. In this case, how often would this happen at a nominal machine setting of 33 cl?
5. What should be the nominal machine setting to ensure that no more than 2% maintenance calls are made? In this case, on average customers will be receiving how much more beverage?

## 5. Baking bread

### Situation

A hypermarket has its own bakery where it prepares and sells bread from 08:00 to 20:00 hours. One extremely popular bread, called “pave supreme”, is made and sold continuously throughout the day. This bread, which is a nominal 500 g loaf, is individually kneaded, left for 3 hours to rise before being baked in the oven. During the kneading and

baking process moisture is lost but from past experience it is known that the standard deviation of the finished bread is 17 g.

### Required

1. If you go to the store and take at random one pave supreme, what is the probability that it will weigh more than 520 g?
2. You are planning a dinner party and so you go to the store and take at random four pave supremes, what is the probability that the average weight of the four weigh more than 520 g?
3. Say that you are planning a larger dinner party and you go to the store and take at random eight pave supremes, what is the probability that the average weight of the eight breads weigh more than 520 g?
4. If you go to the store and take at random one pave supreme, what is the probability that it will weigh between 480 and 520 g?
5. If you go to the store and take at random four pave supremes, what is the probability that the average weight of the loaves will be between 480 and 520 g?
6. If you go to the store and take at random eight pave supremes, what is the probability that the average weight of the loaves will be between 480 and 520 g?
7. Explain the differences between Questions 1 to 3.
8. Explain the differences between Questions 4 to 6. Why is the progression the reverse of what you see for Questions 1 to 3?

## 6. Financial advisor

### Situation

The amount of time a financial advisor spends with each client has a population mean of 35 minutes, and a standard deviation of 11 minutes.

1. If a random client is selected, what is the probability that the time spent with the client will be at least 37 minutes?
2. If a random client is selected, there is a 35% chance that the time the financial advisor spends with the client will be below how many minutes?
3. If random sample of 16 clients are selected, what is the probability that the average time spent per client will be at least 37 minutes?
4. If a random sample of 16 clients is selected, there is a 35% chance that the sample mean will be below how many minutes?
5. If random sample of 25 clients are selected, what is the probability that the average time spent per client will be at least 37 minutes?
6. If a random sample of 25 clients is selected, there is a 35% chance that the sample mean will be below how many minutes?
7. Explain the differences between Questions 1, 3, and 5.
8. What assumptions do you make in responding to these questions?

## 7. Height of adult males

### Situation

In a certain country, the height of adult males is normally distributed, with a mean of 176 cm and a variance of 225 cm<sup>2</sup>.

### Required

1. If one adult male is selected at random, what is the probability that he will be over 2 m?
2. What are the upper and lower limits of height between which 90% will lie for the population of adult males?
3. If samples of four men are taken, what percentage of such samples will have average heights over 2 m?
4. What are the upper and lower limits between which 90% of the sample averages will lie for samples of size four?
5. If samples of nine men are taken, what percentage of such samples will have average heights over 2 m?
6. What are the upper and lower limits between which 90% of the sample averages will lie for samples of size nine?
7. Explain the differences in the results.

## 8. Wal-Mart

### Situation

Wal-Mart of the United States, after buying ASDA in Great Britain, is now looking to move into France. It has targeted 220 supermarket stores in that country and the present owner of these said that profits from these supermarkets follows a normal distribution, have the same mean, with a standard deviation of €37,500. Financial information is on a monthly basis.

### Required

1. If Wal-Mart selects a store at random what is the probability that the profit from this store will lie within €5,400 of the mean?
2. If Wal-Mart management selects 50 stores at random, what is the probability that the sample mean of profits for these 50 stores will lie within €5,400 of the mean?

## 9. Automobile salvage

### Situation

Joe and three colleagues have created a small automobile salvage company. Their work consists of visiting sites that have automobile wrecks and recovering those parts that can be resold. Often from these wrecks they recoup engine parts, computers from the electrical systems, scrap metal, and batteries. From past work, salvaged components on an average generate €198 per car with a standard deviation of €55. Joe and his three colleagues pay themselves €15 each per hour and they work 40 hours/week. Between them

they are able to complete the salvage work on four cars per day. One particular period they carry out salvage work at a site near Hamburg, Germany where there are 72 wrecked cars.

#### Required

1. What is the correct standard error for this situation?
2. What is the probability that after one weeks work the team will have collected enough parts to generate total revenue of €4,200?
3. On the assumption that the probability outcome in Question No. 2 is achieved, what would be the net income to each team member at the end of 1 week?

## 10. Education and demographics

#### Situation

According to a survey in 2000, the population of the United States in the age range 25 to 64 years, 72% were white. Further in this same year, 16% of the total population in the same age range were high school dropouts and 27% had at least a bachelor's degree.<sup>4</sup>

#### Required

1. If random samples of 200 people in the age range 25 to 64 are selected, what proportion of the samples between 69% and 75% will be white?
2. If random samples of 400 people in the age range 25 to 64 are selected, what proportion of the samples between 69% and 75% will be white?
3. If random samples of 200 people in the age range 25 to 64 are selected, what proportion of the samples between 13% and 19% will be high school dropouts?
4. If random samples of 400 people in the age range 25 to 64 are selected, what proportion of the samples between 13% and 19% will be high school dropouts?
5. If random samples of 200 people in the age range 25 to 64 are selected, what proportion of the samples between 24% and 30% will have at least a bachelors degree?
6. If random samples of 400 people in the age range 25 to 64 are selected, what proportion of the samples between 24% and 30% will have at least a bachelors degree?
7. Explain the difference between each paired question of 1 and 2; 3 and 4; and 5 and 6.

## 11. World Trade Organization

#### Situation

The World Trade Organization talks, part of the Doha Round, took place in Hong Kong between 13 and 18 December 2005. According to data, the average percentage tariff imposed on all imported tangible goods and services in certain selected countries is as follows<sup>5</sup>:

<sup>4</sup>Losing ground, *Business Week*, 21 November 2005, p. 90.

<sup>5</sup>US, EU walk fine line at heart of trade impasse, *The Wall Street Journal*, 13 December 2005, p. 1.

United States	India	European Union	Burkina Faso	Brazil
3.7%	29.1%	4.2%	12.0%	12.4%

### Required

1. If a random sample of 200 imported tangible goods or service into the United States were selected, what is the probability that the average proportion of the tariffs for this sample would be between 1% and 4%?
2. If a random sample of 200 imported tangible goods or service into Burkina Faso were selected, what is the probability that the average proportion of the tariffs for this sample would be between 10% and 14%?
3. If a random sample of 200 imported tangible goods or service into India were selected, what is the probability that the average proportion of the tariffs for this sample would be between 25% and 32%?
4. If a random sample of 400 imported tangible goods or service into the United States were selected, what is the probability that the average proportion of the tariffs for this sample would be between 1% and 4%?
5. If a random sample of 400 imported tangible goods or service into Burkina Faso were selected, what is the probability that the average proportion of the tariffs for this sample would be between 10% and 14%?
6. If a random sample of 400 imported tangible goods or service into India were selected, what is the probability that the average proportion of the tariffs for this sample would be between 25% and 32%?
7. Explain the difference between each paired question of 1 and 2; 3 and 4; and 5 and 6.

## 12. Female illiteracy

### Situation

In a survey conducted in three candidate countries for the European Union – Turkey, Romania, and Croatia and three member countries – Greece, Malta, and Slovakia Europe in 2003, the female illiteracy of those over 15 was reported as follows<sup>6</sup>:

Turkey	Greece	Malta	Romania	Croatia	Slovakia
19%	12%	11%	4%	3%	0.5%

### Required

1. If random samples of 250 females over 15 were taken in Turkey in 2003, what proportion between 12% and 22% would be illiterate?
2. If random samples of 500 females over 15 were taken in Turkey in 2003, what proportion between 12% and 22% would be illiterate?

<sup>6</sup>Too soon for Turkish delight, *The Economist*, 1 October 2005, p. 25.

3. If random samples of 250 females over 15 were taken in Malta in 2003, what proportion between 9% and 13% would be illiterate?
4. If random samples of 500 females over 15 were taken in Malta in 2003, what proportion between 9% and 13% would be illiterate?
5. If random samples of 250 females over 15 were taken in Slovakia in 2003, what proportion between 0.1% and 1.0% would be illiterate?
6. If random samples of 500 females over 15 were taken in Slovakia in 2003, what proportion between 0.1% and 1.0% would be illiterate?
7. What is your explanation for the difference between each paired question of 1 and 2; 3 and 4; and 5 and 6?
8. If you took a sample of 200 females over 15 from Istanbul and the proportion of those females illiterate was 0.25%, would you be surprised?

### 13. Unemployment

#### Situation

According to published statistics for 2005, the unemployment rate among people under 25 in France was 21.7% compared to 13.8% for Germany, 12.6% in Britain, and 11.4% in the United States. These numbers in part are considered to be reasons for the riots that occurred in France in 2005.<sup>7</sup>

#### Required

1. If random samples of 100 people under 25 were taken in France in 2005, what proportion between 12% and 15% would be unemployed?
2. If random samples of 200 people under 25 were taken in France in 2005, what proportion between 12% and 15% would be unemployed?
3. If random samples of 100 people under 25 were taken in Germany in 2005, what proportion between 12% and 15% would be unemployed?
4. If random samples of 200 people under 25 were taken in Germany in 2005, what proportion between 12% and 15% would be unemployed?
5. If random samples of 100 people under 25 were taken in Britain in 2005, what proportion between 12% and 15% would be unemployed?
6. If random samples of 200 people under 25 were taken in Britain in 2005, what proportion between 12% and 15% would be unemployed?
7. If random samples of 100 people under 25 were taken in the United States in 2005, what proportion between 12% and 15% would be unemployed?
8. If random samples of 200 people under 25 were taken in the United States in 2005, what proportion between 12% and 15% would be unemployed?

<sup>7</sup> France's young and jobless, *Business Week*, 21 November 2005, p. 23.



9. What is your explanation for the difference between each paired question of 3 and 4; 5 and 6; and 7 and 8?
10. Why do the data for France in Questions 1 and 2 not follow the same trend as for the questions for the other three countries?

## 14. Manufacturing employment

### Situation

According to a recent survey by the OECD in 2005 employment in manufacturing as a percent of total employment, has fallen dramatically since 1970. The following table gives the information for OECD countries<sup>8</sup>:

Country	Germany	Italy	Japan	France	Britain	Canada	United States
1970	40%	28%	27%	28%	35%	23%	25%
2005	23%	22%	18%	16%	14%	14%	10%

### Required

1. If random samples of 200 people of the working population were taken from Germany in 2005, what proportion between 20% and 26% would be in manufacturing?
2. If random samples of 400 people of the working population were taken from Germany in 2005, what proportion between 20% and 26% would be in manufacturing?
3. If random samples of 200 people of the working population were taken from Britain in 2005, what proportion between 13% and 15% would be in manufacturing?
4. If random samples of 400 people of the working population were taken from Britain in 2005, what proportion between 13% and 15% would be in manufacturing?
5. If random samples of 200 people of the working population were taken from the United States in 2005, what proportion between 6% and 10% would be in manufacturing?
6. If random samples of 400 people of the working population were taken from the United States in 2005, what proportion between 6% and 10% would be in manufacturing?
7. What is your explanation for the difference between each paired question of 1 and 2; 3 and 4; and 5 and 6.
8. If a sample of 100 people was taken in Germany in 2005 and the proportion of the people in manufacturing was 32%, what conclusions might you draw?

<sup>8</sup>Industrial metamorphosis, *The Economist*, 1 October 2005, p. 69.

## 15. Homicide

### Situation

In December 2005, Steve Harvey, an internationally known AIDS outreach worker was abducted at gunpoint from his home in Jamaica and murdered.<sup>9</sup> According to the statistics of 2005, Jamaica is one the world's worst country for homicide. How it compares with some other countries according to the number of homicides per 100,000 people is given in the table below<sup>10</sup>:

Britain	United States	Zimbabwe	Argentina	Russia	Brazil	S. Africa	Columbia	Jamaica
2	6	8	14	21	25	44	47	59

### Required

1. If you lived in Jamaica what is the probability that some day you would be a homicide statistic?
2. If you lived in Britain what is the probability that some day you would be a homicide statistic? Compare this probability with the previous question? What is another way of expressing this probability between the two countries?
3. If random samples of 1,000 people were selected in Jamaica what is the proportion between 0.03% and 0.09% that would be homicide victims?
4. If random samples of 2,000 people were selected in Jamaica what is the proportion between 0.03% and 0.09% that would be homicide victims?
5. Explain the difference between Questions 3 and 4.

## 16. Humanitarian agency

### Situation

A subdivision of the humanitarian organization, doctors without borders, based in Paris has 248 personnel in its database according to the table below. This database gives in alphabetical order the name of the staff members, gender, age at last birthday, years with the organization, the country where the staff member is based, and their training in the medical field. You wish to get information about the whole population included in this database including criteria such as job satisfaction, safety concerns in the country of work, human relationships in the country, and other qualitative factors. For budget reasons you are limited to interviewing a total of 40 people and some of these will be done by telephone but others will be personal interviews in the country of operation.

<sup>9</sup> A murder in Jamaica, *International Herald Tribune*, 14 December 2005, p. 8.

<sup>10</sup> Less crime, more fear, *The Economist*, 1 October 2005, p. 42.

### Required

Develop a sampling plan to select the 40 people. Consider total random sampling, cluster sampling, and strata sampling. In all cases use the random number function in Excel to make the sample selection. Of the plans that you select draw conclusions. Which do you believe is the best experiment? Explain your reasoning:

No.	Name	Gender	Age	Years with agency	Country where based	Medical training
1	Abissa, Yasmina Murielle	F	26	2	Chile	Nurse
2	Adekalom, Maily	F	45	17	Brazil	General medicine
3	Adjei, Abena	F	41	16	Brazil	Nurse
4	Ahihounkpe, Ericka	F	29	5	Kenya	Physiotherapy
5	Akintayo, Funmilayo	F	46	12	Brazil	Nurse
6	Alexandre, Gaëlle	F	46	19	Kenya	Nurse
7	Alibizzata, Mylène	F	31	2	Chile	Radiographer
8	Ama, Eric	M	30	2	Brazil	Nurse
9	Angue Assoumou, Mélodie	F	47	18	Chile	Nurse
10	Arfort, Sabrina	F	47	12	Cambodia	Nurse
11	Aubert, Nicolas	M	50	20	Costa Rica	Nurse
12	Aubery, Olivia	F	34	12	Thailand	Nurse
13	Aulombard, Audrey	F	49	18	Brazil	Physiotherapy
14	Awitor, Euloge	M	36	12	Kenya	Nurse
15	Ba, Oumy	F	27	1	Chile	Nurse
16	Bakouan, Aminata	F	24	3	Vietnam	Nurse
17	Banguebe, Sandrine	F	41	1	Costa Rica	Nurse
18	Baque, Nicolas	M	42	14	Kenya	Nurse
19	Batina, Cédric	M	32	2	Kenya	Nurse
20	Batty-Ample, Agnès	F	31	10	Chile	Nurse
21	Baud, Maxime	F	44	18	Costa Rica	Nurse
22	Belkora, Youssef	M	46	2	Brazil	Radiographer
23	Berard, Emmanuelle	F	41	17	Vietnam	Nurse
24	Bernard, Eloise	F	40	3	Vietnam	Surgeon
25	Berton, Alexandra	M	46	22	Ivory Coast	Nurse
26	Besenwald, Laetitia	F	28	2	Brazil	Nurse
27	Beyschlag, Natalie	F	34	8	Brazil	Nurse
28	Black, Kimberley	F	32	8	Kenya	Nurse
29	Blanchon, Paul	M	23	1	Vietnam	Nurse
30	Blondet, Thomas	M	34	1	Kenya	Nurse
31	Bomboh, Patrick	M	31	11	Chile	Nurse
32	Bordenave, Bertrand	M	32	10	Brazil	Physiotherapy
33	Bossekota, Ariane	F	37	9	Kenya	Nurse
34	Boulay, Grégory	M	36	7	Kenya	Nurse
35	Bouziat, Lucas	M	53	26	Kenya	Nurse
36	Briatte, Pierre-Edouard	M	48	28	Kenya	Nurse

No.	Name	Gender	Age	Years with agency	Country where based	Medical training
37	Brunel, Laurence	F	27	3	Ivory Coast	General medicine
38	Bruntsch-Lesba, Natascha	F	55	30	Cambodia	Nurse
39	Buzingo, Patrick	M	46	5	Thailand	Nurse
40	Cablova, Dagmar	F	53	14	Kenya	Nurse
42	Chabanel, Gael	F	31	2	Ivory Coast	Nurse
43	Chabanier, Maud	F	27	1	Brazil	Nurse
44	Chahboun, Zineb	F	53	23	Brazil	Physiotherapy
45	Chahed, Samy	M	46	12	Thailand	Nurse
46	Chappon, Romain	F	40	18	Costa Rica	Nurse
47	Chartier, Henri	M	28	8	Vietnam	Nurse
48	Chaudagne, Stanislas	M	45	10	Chile	Radiographer
49	Coffy, Robin	M	48	24	Ivory Coast	Nurse
50	Coissard, Alexandre	M	36	8	Chile	Nurse
51	Collomb, Fanny	F	54	18	Ivory Coast	Nurse
52	Coradetti, Louise	F	36	11	Cambodia	Nurse
53	Cordier, Yan	M	43	1	Brazil	Surgeon
54	Crombe, Jean-Michel	M	27	7	Brazil	Nurse
55	Croute, Benjamin	M	42	17	Vietnam	Nurse
56	Cusset, Johansson	M	42	12	Cambodia	Nurse
57	Czajkowski, Mathieu	M	51	10	Brazil	Nurse
58	Dadzie, Kelly	M	34	10	Chile	Nurse
59	Dandjouma, Ainaou	F	50	2	Nigeria	Nurse
60	Dansou, Joel	M	54	27	Brazil	Physiotherapy
61	De Messe Zinsou, Thierry	M	38	2	Ivory Coast	Nurse
62	De Zelicourt, Gonzague	M	55	4	Cambodia	Nurse
63	Debaille, Camille	F	50	26	Kenya	Nurse
64	Declippeleir, Olivier	M	31	9	Chile	Nurse
65	Delahaye, Benjamin	M	47	11	Ivory Coast	Nurse
66	Delegue, Héloïse	F	31	6	Brazil	Radiographer
67	Delobel, Delphine	F	23	3	Kenya	Nurse
68	Demange, Aude	F	30	1	Vietnam	Nurse
69	Deplano, Guillaume	M	54	33	Thailand	Nurse
70	Desplanches, Isabelle	F	34	10	Thailand	Nurse
71	Destombes, Hélène	F	31	7	Brazil	Nurse
72	Diallo, Ralou Maimouna	F	50	25	Ivory Coast	General medicine
73	Diehl, Pierre	M	45	11	Brazil	Nurse
74	Diop, Mohamed	M	25	5	Chile	Nurse
75	Dobeli, Nathalie	F	33	10	Chile	Physiotherapy
76	Doe-Bruce, Othalia Ayele E	F	53	19	Cambodia	Nurse
77	Donnat, Mélanie	F	51	16	Thailand	Nurse
78	Douenne, François-Xavier	M	37	15	Ivory Coast	Surgeon
79	Du Mesnil Du Buisson, Edouard	M	52	21	Vietnam	Nurse
80	Dubourg, Jonathan	M	44	3	Cambodia	Nurse
81	Ducret, Camille	F	50	16	Thailand	Nurse

(Continued)

No.	Name	Gender	Age	Years with agency	Country where based	Medical training
82	Dufau, Guillaume	M	45	25	Chile	Nurse
83	Dufaud, Charly	M	28	5	Costa Rica	Radiographer
84	Dujardin, Agathe	F	36	15	Kenya	Nurse
85	Dutel, Sébastien	M	50	11	Cambodia	Nurse
86	Dutraive, Benjamin	M	33	2	Brazil	Nurse
87	Eberhardt, Nadine	F	26	6	Cambodia	Physiotherapy
88	Ebibie N'ze, Yannick	M	28	8	Chile	Nurse
89	Errai, Skander	M	47	5	Thailand	Nurse
90	Erulin, Caroline	F	42	18	Ivory Coast	General medicine
91	Escarboutel, Christel	F	52	3	Ivory Coast	Nurse
92	Etien, Stéphanie	F	54	12	Brazil	Nurse
93	Felio, Sébastien	M	32	4	Kenya	Nurse
94	Fernandes, Claudio	M	29	5	Vietnam	Surgeon
95	Fillioux, Stéphanie	F	32	9	Brazil	Nurse
96	Flandrois, Nicolas	M	31	10	Ivory Coast	Nurse
97	Gaillardet, Marion	F	23	3	Brazil	Nurse
98	Garapon, Sophie	F	31	3	Kenya	Nurse
99	Garnier, Charles	M	27	6	Costa Rica	Nurse
100	Garraud, Charlotte	F	43	11	Brazil	Radiographer
101	Gassier, Vivienne	F	33	13	Brazil	Nurse
102	Gava, Mathilde	F	26	4	Ivory Coast	Nurse
103	Gerard, Vincent	M	50	5	Thailand	Physiotherapy
104	Germany, Julie	F	29	9	Kenya	Nurse
105	Gesrel, Valentin	M	40	11	Nigeria	Nurse
106	Ginet-Kauders, David	M	54	33	Costa Rica	Nurse
107	Gobber, Aurélie	F	32	1	Costa Rica	Nurse
108	Grangeon, Baptiste	M	33	13	Chile	Nurse
109	Gremmel, Antoine	M	31	3	Brazil	Nurse
110	Gueit, Delphine	F	46	2	Thailand	Nurse
111	Guerite, Camille	M	45	8	Cambodia	Nurse
112	Guillot, Nicholas	M	33	5	Brazil	Nurse
113	Hardy, Gilles	M	30	3	Chile	Nurse
114	Hazard, Guillaume	M	38	9	Cambodia	Nurse
115	Honnegger, Dorothee	F	45	2	Vietnam	Nurse
116	Houdin, Julia	F	49	7	Thailand	Physiotherapy
117	Huang, Shan-Shan	F	35	14	Costa Rica	Nurse
118	Jacquel, Hélène	F	47	16	Vietnam	Nurse
119	Jiguet-Jiglairaz, Sébastien	M	55	35	Chile	Nurse
120	Jomard, Sam	M	34	7	Kenya	Surgeon
121	Julien, Loïc	F	35	2	Brazil	Nurse
122	Kacou, Joeata	F	48	3	Vietnam	Nurse
123	Kasalica, Aneta	F	51	13	Brazil	Nurse
124	Kasalica, Darko	M	24	4	Brazil	Nurse
125	Kassab, Philippe	M	29	1	Brazil	Radiographer
126	Kervaon, Nathalie	F	45	24	Brazil	Nurse

No.	Name	Gender	Age	Years with agency	Country where based	Medical training
127	Kimbakala-Koumba, Madeleine	F	44	11	Costa Rica	Nurse
128	Kolow, Alexandre	M	40	7	Chile	Nurse
129	Latini, Stéphane	F	50	23	Chile	Nurse
130	Lauvaure, Julien	M	42	14	Brazil	Nurse
132	Legris, Baptiste	M	38	18	Ivory Coast	Nurse
133	Lehot, Julien	M	37	16	Vietnam	Physiotherapy
134	Lestangt, Aurélie	F	29	8	Cambodia	Nurse
135	Li, Si Si	F	32	5	Chile	Nurse
136	Liubinskas, Ricardas	M	25	4	Cambodia	Nurse
137	Loyer, Julien	M	34	10	Vietnam	Nurse
138	Lu Shan Shan	F	31	8	Chile	Nurse
139	Marchal, Arthur	M	45	12	Nigeria	Nurse
140	Marganne, Richard	M	25	4	Chile	Nurse
141	Marone, Lati	F	33	4	Brazil	Nurse
142	Martin, Cyrielle	F	42	5	Kenya	Physiotherapy
143	Martin, Stéphanie	F	46	16	Brazil	Nurse
144	Martinez, Stéphanie	F	25	5	Thailand	Nurse
145	Maskey, Lilly	F	23	1	Vietnam	Nurse
146	Masson, Cédric	M	29	8	Brazil	Nurse
147	Mathisen, Mélinda	F	48	12	Cambodia	Nurse
148	Mermet, Alexandra	F	25	1	Brazil	Nurse
149	Mermet, Florence	F	27	1	Brazil	Radiographer
150	Michel, Dorothée	F	54	24	Brazil	Nurse
151	Miribel, Julien	M	53	6	Vietnam	Nurse
152	Monnot, Julien	F	40	5	Chile	Nurse
153	Montfort, Laura	F	53	16	Nigeria	Nurse
154	Murgue, François	M	32	2	Kenya	Nurse
155	Nauwelaers, Emmanuel	F	55	24	Thailand	Nurse
156	Nddalla-Ella, Claude	F	35	14	Thailand	Surgeon
157	Ndiaye, Baye Mor	M	50	23	Vietnam	Nurse
158	Neulat, Jean-Philippe	M	37	17	Cambodia	Physiotherapy
159	Neves, Christophe	M	28	2	Brazil	Nurse
160	Nicot, Guillaume	M	29	8	Brazil	Nurse
161	Oculy, Frédéric	M	45	12	Chile	Nurse
162	Okewole, Maxine	M	51	21	Kenya	Nurse
163	Omba, Nguie	M	47	24	Ivory Coast	Nurse
164	Ostler, Emilie	F	28	1	Brazil	Nurse
165	Owiti, Brenda	F	25	1	Kenya	Surgeon
166	Ozkan, Selda	F	43	21	Nigeria	Nurse
167	Paillet, Maïté	F	55	28	Kenya	Nurse
168	Penillard, Cloé	F	38	5	Ivory Coast	Nurse
169	Perera, William	M	43	17	Nigeria	Nurse
170	Perrenot, Christophe	M	30	3	Kenya	Nurse
171	Pesenti, Johan	M	47	3	Costa Rica	Radiographer

(Continued)

No.	Name	Gender	Age	Years with agency	Country where based	Medical training
172	Petit, Dominique	F	48	17	Thailand	Nurse
173	Pfeiffer, Céline	F	39	7	Ivory Coast	Nurse
174	Philetas, Ludovic	M	24	3	Thailand	Nurse
175	Portmann, Kevin	M	45	21	Chile	Nurse
176	Pourrier, Jennifer	F	41	15	Thailand	Physiotherapy
177	Prou, Vincent	M	42	7	Chile	Nurse
178	Raffaele, Grégory	M	55	26	Cambodia	Nurse
179	Ramanoelisoa, Eliane Gorette	F	49	18	Vietnam	Nurse
180	Rambaud, Philippe	M	27	1	Costa Rica	Nurse
181	Ranjatoelina, Andrew	M	43	5	Brazil	Nurse
182	Ravets, Emmanuelle	F	30	8	Nigeria	Nurse
183	Ribieras, Alexandre	M	45	14	Cambodia	Nurse
184	Richard, Damien	M	23	1	Kenya	Nurse
185	Rocourt, Nicolas	M	41	9	Brazil	Nurse
186	Rossi-Ferrari, Sébastien	M	37	2	Thailand	Nurse
187	Rouviere, Grégory	M	51	31	Ivory Coast	Nurse
188	Roux, Alexis	M	23	1	Kenya	Nurse
189	Roy, Marie-Charlotte	F	51	14	Costa Rica	Nurse
190	Rudkin, Steven	M	41	21	Thailand	General medicine
191	Ruget, Joffrey	M	24	4	Brazil	Nurse
192	Rutledge, Diana	F	38	11	Thailand	Nurse
193	Ruzibiza, Hubert	M	35	12	Brazil	Nurse
194	Ruzibiza, Oriane	F	45	10	Brazil	Nurse
195	Sadki, Khalid	M	35	13	Vietnam	Nurse
196	Saint-Quentin, Florent	M	55	22	Brazil	Physiotherapy
197	Salami, Mistoura	F	45	20	Cambodia	Nurse
198	Sambe, Mamadou	F	31	5	Ivory Coast	Nurse
199	Sanvee, Pascale	F	51	23	Vietnam	Radiographer
200	Saphores, Pierre-Jean	M	32	2	Cambodia	Nurse
201	Sassioui, Mohamed	M	48	10	Chile	Nurse
202	Savall, Arnaud	M	47	18	Brazil	Nurse
203	Savinas, Tamara	F	54	34	Ivory Coast	Nurse
204	Schadt, Stéphanie	F	33	13	Brazil	Surgeon
205	Schmuck, Céline	F	54	9	Costa Rica	Nurse
206	Schneider, Aurélie	F	53	3	Costa Rica	Nurse
207	Schulz, Amir	M	39	10	Vietnam	Nurse
208	Schwartz, Olivier	M	46	21	Brazil	Nurse
209	Seimbille, Alexandra	M	47	1	Chile	Nurse
210	Servage, Benjamin	M	47	23	Ivory Coast	Nurse
211	Sib, Brigitte	F	51	13	Brazil	Nurse
212	Sinistaj, Irena	F	36	12	Brazil	Nurse
213	Six, Martin	M	34	3	Costa Rica	Nurse
214	Sok, Steven	M	26	1	Costa Rica	Nurse
215	Souah, Steve	M	50	23	Ivory Coast	Nurse

No.	Name	Gender	Age	Years with agency	Country where based	Medical training
216	Souchko, Edouard	M	38	7	Nigeria	Radiographer
217	Soumare, Anna	F	52	22	Vietnam	Nurse
218	Straub, Elodie	F	25	2	Brazil	Nurse
219	Sun, Wenjie	F	31	4	Kenya	Physiotherapy
220	SuperVielle Brouques, Claire	F	40	7	Kenya	Surgeon
221	Tahraoui, Davina	F	31	1	Thailand	Nurse
222	Tall, Kadiatou	F	33	8	Thailand	Nurse
223	Tarate, Romain	M	49	3	Costa Rica	Nurse
224	Tessaro, Laure	F	39	18	Ivory Coast	Nurse
225	Tillier, Pauline	F	29	8	Kenya	Nurse
226	Trenou, Kémi	M	44	19	Thailand	Nurse
227	Triquere, Cyril	M	23	1	Brazil	Nurse
228	Tshitungi, Mesenga	F	40	2	Ivory Coast	Nurse
229	Vadivelou, Christophe	M	55	17	Chile	Physiotherapy
230	Vande-Vyre, Julien	M	25	5	Brazil	Nurse
231	Villemur, Claire	F	41	17	Costa Rica	Nurse
232	Villet, Diana	F	33	8	Nigeria	Nurse
233	Vincent, Marion	F	36	2	Ivory Coast	General medicine
234	Vorillon, Fabrice	M	32	4	Chile	Nurse
235	Wadagni, Imelda	F	45	10	Vietnam	Nurse
236	Wallays, Anne	F	30	2	Vietnam	Nurse
237	Wang, Jessica	F	38	18	Brazil	Nurse
238	Weigel, Samy	M	34	13	Ivory Coast	Nurse
239	Wernert, Lucile	F	24	2	Kenya	Nurse
240	Willot, Mathieu	M	52	30	Brazil	Nurse
241	Wlodyka, Sébastien	M	40	6	Kenya	Nurse
242	Wurm, Debora	F	46	15	Chile	Nurse
243	Xheko, Eni	F	28	1	Costa Rica	Nurse
244	Xu, Ning	F	48	13	Brazil	Physiotherapy
245	Yuan, Zhiyi	M	39	18	Brazil	Surgeon
246	Zairi, Leila	F	51	26	Vietnam	Radiographer
247	Zeng, Li	F	25	3	Ivory Coast	Nurse
248	Zhao, Lizhu	F	33	9	Thailand	General medicine



*This page intentionally left blank*

# Estimating population characteristics

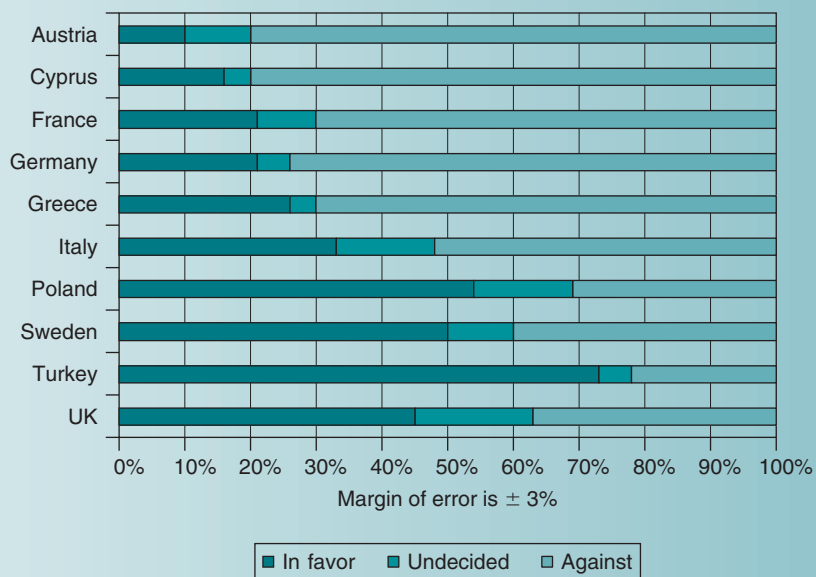
## Turkey and the margin of error

*The European Union, after a very heated debate, agreed in early October 2005 to open membership talks to admit Turkey, a Muslim country of 70 million people. This agreement came only after a tense night-and-day discussion with Austria, one of the 25-member states, who strongly opposed Turkey's membership. Austria has not forgotten fighting back the invading Ottoman armies in the 16th and 17th centuries. Reservations to Turkey's membership is also very strong in other countries as shown in Figure 7.1 where an estimated 70% or more of the population in each of Austria, Cyprus, Germany, France, and Greece are opposed to membership. This estimated information is based on a survey response of a sample of about 1,000 people in each of the 10 indicated countries. The survey was conducted in the period May–June 2005 with an indicated margin of error of  $\pm 3\%$  points. This survey was made to estimate population characteristics, which is the essence of the material in this chapter.<sup>1</sup>*

---

<sup>1</sup> Champion, M., and Karnitschnig, M., "Turkey gains EU approval to begin membership talks", *Wall Street Journal Europe*, 4 October 2005, pp. 1 and 14.

Figure 7.1 Survey response of attitudes to Turkey joining the European Union.



## Learning objectives

After you have studied this chapter you will understand how **sampling can be extended to make estimates of population parameters** such as the **mean** and the **proportion**. To facilitate comprehension the chapter is organized as follows:

- ✓ **Estimating the mean value** • Point estimates • Interval estimates • Confidence level and reliability • Confidence interval of the mean for an infinite population • Application of confidence intervals for an infinite population: *Paper* • Sample size for estimating the mean of an infinite population • Application for determining the sample size: *Coffee* • Confidence interval of the mean for a finite population • Application of the confidence interval for a finite population: *Printing*
- ✓ **Estimating the mean using the Student-*t* distribution** • The Student-*t* distribution • Degrees of freedom in the *t*-distribution • Profile of the Student-*t* distribution • Confidence intervals using a Student-*t* distribution • Excel and the Student-*t* distribution • Application of the Student-*t* distribution: *Kiwi fruit* • Sample size and the Student-*t* distribution • Re-look at the example *kiwi fruit* using the normal distribution
- ✓ **Estimating and auditing** • Estimating the population amount • Application of auditing for an infinite population: *tee-shirts* • Application of auditing for a finite population: *paperback books*
- ✓ **Estimating the proportion** • Interval estimate of the proportion for large samples • Sample size for the proportion for large samples • Application of estimation for proportions: *Circuit boards*
- ✓ **Margin of error and levels of confidence** • Explaining margin of error • Confidence levels

In Chapter 6, we discussed statistical sampling for the purpose of obtaining information about a population. This chapter expands upon this to use sampling to estimate, or infer, population parameters based entirely on the sample data. By its very nature, estimating is **probabilistic** as there is no certainty of the result. However, if the sample experiment is correctly designed then there should be a reasonable confidence about conclusions that are made. Thus from samples we might with confidence estimate the mean weight of airplane passengers for fuel-loading purposes, the proportion of the population expected to vote Republican, or the mean value of inventory in a distribution centre.

### Estimating the Mean Value

The mean or average value of data is the sum of all the data taken divided by the number of

measurements taken. The units of measurement can be financial units, length, volume, weight, etc.

### Point estimates

In estimating, we could use a single value to estimate the true population mean. For example, if the grade point average of a random sample of students is 3.75 then we might estimate that the population average of all students is also 3.75. Or, we might select at random 20 items of inventory from a distribution centre and calculate that their average value is £25.45. In this case we would estimate that the population average of the entire inventory is £25.45. Here we have used the sample mean  $\bar{x}$  as a **point estimate** or an **unbiased estimate** of the true population mean,  $\mu_x$ . The problem with one value or a point estimate is that they are presented as being exact and that unless we have a *super*

*crystal ball*, the probability of them being precisely the right value is low. Point estimates are often inadequate as they are just a single value and thus, they are either right or wrong. In practice it is more meaningful to have an interval estimate and to quantify these intervals by probability levels that give an estimate of the error in the measurement.

## Interval estimates

With an **interval estimate** we might describe situations as follows. The estimate for the project cost is between \$11.8 and \$12.9 million and I am 95% confident of these figures. The estimate for the sales of the new products is between 22,000 and 24,500 units in the first year and I am 90% confidence of these figures. The estimate of the price of a certain stock is between \$75 and \$90 but I am only 50% confident of this information. The estimate of class enrolment for Business Statistics next academic year is between 220 and 260 students though I am not too confident about these figures. Thus the interval estimate is a range within which the population parameter is likely to fall.

## Confidence level and reliability

Suppose a subcontractor A makes refrigerator compressors for client B who assembles the final refrigerators. In order to establish the terms of the final customer warranty, the client needs information about the life of compressors since the compressor is the principal working component of the refrigerator. Assume that a random sample of 144 compressors is tested and that the mean life of the compressors,  $\bar{x}$ , is determined to be 6 years or 72 months. Using the concept of point estimates we could say that the mean life of all the compressors manufactured is 72 months. Here  $\bar{x}$  is the **estimator** of the population mean  $\mu_x$  and 72 months is the **estimate** of the population mean obtained from the sample. However, this

information says nothing about the reliability or confidence that we have in the estimate.

The subcontractor has been making these compressors for a long time and knows from past data that the standard deviation of the working life of compressors is 15 months. Then since our sample size of 144 is large enough, the standard error of the mean can be calculated by using equation 6(ii) from Chapter 6 from the central limit theory:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{15}{\sqrt{144}} = \frac{15}{12} = 1.25 \text{ months}$$

This value of 1.25 months is one standard error of the mean, or it means that  $z = \pm 1.00$ , for the sampling distribution. If we assume that the life of a compressor follows a normal distribution then we know from Chapter 5 that 68.26% of all values in the distribution lie within  $\pm 1$  standard deviations from the mean. From equation 6(iv),

$$\pm z = \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}}$$

or

$$\pm 1 = \frac{\bar{x} - 72}{1.25}$$

When  $z = -1$  then the lower limit of the compressor life is,

$$\bar{x} = 72 - 1.25 = 70.75 \text{ months}$$

When  $z = +1$  then the upper limit is,

$$\bar{x} = 72 + 1.25 = 73.25 \text{ months}$$

Thus we can say that, the mean life of the compressors is about 72 months and there is a 68.26% (about 68%) probability that the mean value will be between 70.75 and 73.25 months.

Two standard errors of the mean, or when  $z = \pm 2$ , is  $2 * 1.25$  or 2.50 months. Again from

Chapter 5, if we assume a normal distribution, 95.44%, of all values in the distribution lie within  $\pm 2$  standard deviations from the mean.

When  $z = -2$  then using equation 6(iv), the lower limit of the compressor life is,

$$\bar{x} = 72 - 2 * 1.25 = 69.50 \text{ months}$$

When  $z = +2$  then the upper limit is,

$$\bar{x} = 72 + 2 * 1.25 = 74.50 \text{ months}$$

Thus we can say that, the mean life of the compressor is about 72 months and there is a 95.44% (about 95%) probability that the mean value will be between 69.50 and 74.50 months.

Finally, three standard errors of the mean is  $3 * 1.25$  or 3.75 months and again from Chapter 5, assuming a normal distribution, 99.73%, of all values in the distribution lie within  $\pm 3$  standard deviations from the mean.

When  $z = -3$  then using equation 6(iv), the lower limit of compressor life is,

$$\bar{x} = 72 - 3 * 1.25 = 68.25 \text{ months}$$

When  $z = +3$  then the upper limit is,

$$\bar{x} = 72 + 3 * 1.25 = 75.75 \text{ months}$$

Thus we can say that the mean life of the compressor is about 72 months and there is almost a 99.73% (about 100%) probability that the mean value will be between 68.25 and 75.75 months.

Thus in summary we say as follows:

1. The best estimate is that the mean compressor life is 72 months and the manufacturer is about 68% confident that the compressor life is in the range 70.75 to 73.25 months. Here the **confidence interval** is between 70.75 and 73.25 months, or a range of 2.50 months.
2. The best estimate is that the mean compressor life is 72 months and the manufacturer is about 95% confident that the compressor life

is in the range 69.50 to 74.50 months. Here the confidence interval is between 69.50 and 74.50 months, or a range of 5.00 months.

3. The best estimate is that the mean compressor life is 72 months and the manufacturer is about 100% confident that the compressor life is in the range 68.25 to 75.75 months. Here the confidence interval is between 68.25 and 75.75 months, or a range of 7.50 months.

It is important to note that as our confidence level increases, going from 68% to 100%, the confidence interval increases, going from a range of 2.50 to 7.50 months. This is to be expected as we become more confident of our estimate, we give a broader range to cover uncertainties.

## Confidence interval of the mean for an infinite population

The confidence interval is the range of the estimate being made. From the above compressor example, considering the  $\pm 2\sigma$  confidence intervals, we have 69.50 and 74.50 months as the respective lower and upper limits. Between these limits this is equivalent to 95.44% of the area under the normal curve, or about 95%. A 95% confidence interval estimate implies that if all possible samples were taken, about 95% of them would include the true population mean,  $\mu$ , somewhere within their interval, whereas, about 5% of them would not. This concept is illustrated in Figure 7.2 for six different samples. The  $\pm 2\sigma$  intervals for sample numbers 1, 2, 4, and 5 contain the population mean  $\mu$ , whereas for samples 3 and 6 do not contain the population mean  $\mu$  within their interval.

The **level of confidence** is  $(1 - \alpha)$ , where  $\alpha$  is the total proportion in the tails of the distribution outside of the confidence interval. Since the distribution is symmetrical, the area in each tail is  $\alpha/2$  as shown in Figure 7.3. As we have shown in the compressor situation, the

Figure 7.2 Confidence interval estimate.

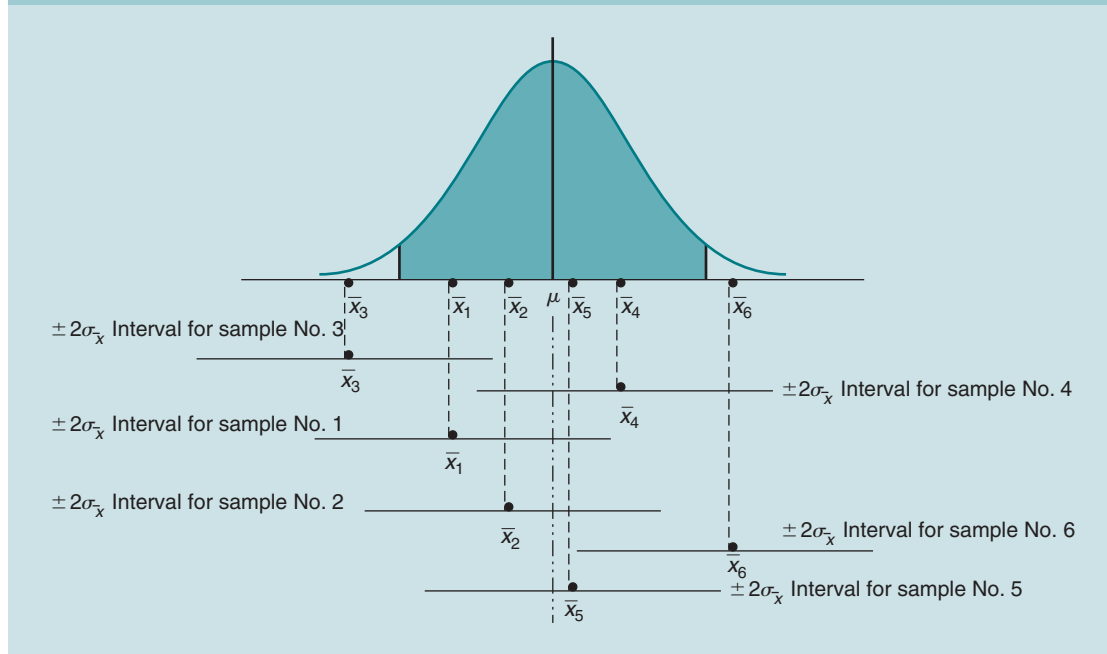
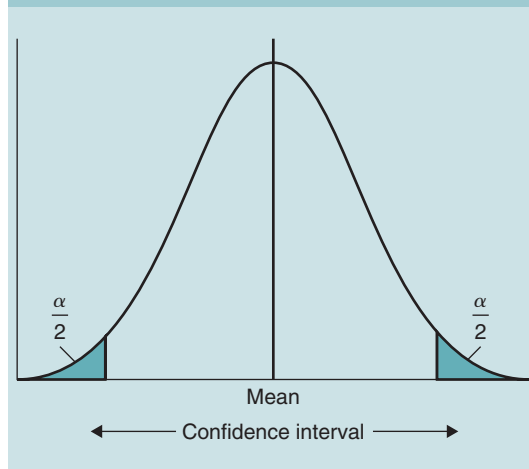


Figure 7.3 Confidence interval and the area in the tails.



confidence intervals for the population estimate for the mean value are thus,

$$\bar{x} \pm z\sigma_{\bar{x}} = \bar{x} \pm z \frac{\sigma_x}{\sqrt{n}} \quad 7(i)$$

This implies that the population mean lies in the range given by the relationship,

$$\bar{x} - z \frac{\sigma_x}{\sqrt{n}} \leq \mu_x \leq \bar{x} + z \frac{\sigma_x}{\sqrt{n}} \quad 7(ii)$$

### Application of confidence intervals for an infinite population: *Paper*

Inacopia, the Portuguese manufacturer of A4 paper commonly used in computer printers wants to be sure that its cutting machine is operating correctly. The width of A4 paper is expected to be 21.00 cm and it is known that the standard deviation of the cutting machine is 0.0100 cm. The quality control inspector pulls a random sample of 60 sheets from the production line and the average width of this sample is 20.9986 cm.

1. Determine the 95% confidence intervals of the mean width of all the A4 paper coming off the production line?

We have the following information:

Sample size,  $n$ , is 60

Sample mean,  $\bar{x}$ , is 20.9986 cm

Population standard deviation,  $\sigma$ , is 0.0100

Standard error of the mean is,

$$\frac{\sigma}{\sqrt{n}} = \frac{0.0100}{\sqrt{60}} = 0.0013$$

The area in the each tail for a 95% confidence limit is 2.5%. Using [function NORM-SINV] in Excel for a value  $P(x)$  of 2.5% gives a lower value of  $z$  of  $-1.9600$ . Since the distribution is symmetrical, the upper value is numerically the same at  $+1.9600$ . (Note: an alternative way of finding the upper value of  $z$  is to enter in [function NORMSINV] the value of 97.50% (2.50% + 95%) which is the area of the curve from the left to the upper value of  $z$ .)

From equation 7(i) the confidence limits are,

$$20.9986 \pm 1.9600 * 0.0013 = 20.9961$$

$$\text{and } 21.0011 \text{ cm}$$

Thus we would say that our best estimate of the width of the computer paper is 20.9986 cm and we are 95% confident that the width is in the range 20.9961–21.0011. Since this interval contains the population expected mean value of 21.0000 cm, we can conclude that there seems to be no problem with the cutting machine.

2. Determine the 99% confidence intervals of the mean width of all the A4 paper coming off the production line.

The area in each tail for a 99% confidence limit is 0.5%. Using [function NORMSINV] in Excel for a value  $P(x)$  of 0.5% gives a lower value of  $z$  of  $-2.5758$ . Since the distribution is symmetrical, the upper value is  $+2.5758$ . (Note: an alternative way of finding the upper value of  $z$  is to enter in [function NORMSINV] the value of 99.50% (0.50% + 99%) which is the area of the curve from the left to the upper value of  $z$ .)

From equation 7(i) the confidence limits are,

$$20.9986 \pm 2.5758 * 0.0013 = 20.9953$$

$$\text{and } 21.0019 \text{ cm}$$

Thus we would say that our best estimate of the width of the computer paper is 20.9986 cm and we are 99% confident that the width is in the range 20.9953–21.0019. Again, since this interval contains the expected mean value of 21.0000 cm, we can conclude that there seems to be no problem with the cutting machine.

Note that the limits in Question 2 are wider than in Question 1 since we have a higher confidence level.

## Sample size for estimating the mean of an infinite population

In sampling it is useful to know the size of the sample to take in order to estimate the population parameter for a given confidence level. We have to accept that unless the whole population is analysed there will always be a sampling error. If the sample size is small, the chances are that the error will be high. If the sample size is large there may be only a marginal gain in reliability in the estimate of our population mean but what is certain is that the analytical experiment will be more expensive. Thus, what is an appropriate sample size,  $n$ , to take for a given confidence level?

The confidence limits are related the sample size,  $n$ , by equation 6(iv) or,

$$\pm z = \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}} \quad 6(\text{iv})$$

The range from the population mean, on the left side of the distribution when  $z$  is negative, is  $-(\bar{x} - \mu_x)$  or  $\mu_x - \bar{x}$  on the left side of the distribution, and  $\bar{x} - \mu_x$  on the right side of the



distribution curve. Reorganizing equation 6(iv) by making the sample size,  $n$ , the subject gives,

$$n = \left( \frac{z\sigma_x}{\bar{x} - \mu_x} \right)^2 \quad 7(\text{iii})$$

The term,  $\bar{x} - \mu_x$ , is the sample error and if we denote this by  $e$ , then the sample size is given by,

$$n = \left( \frac{z\sigma_x}{e} \right)^2 \quad 7(\text{iv})$$

Thus for a given confidence level, which then gives the value of  $z$ , and a given confidence limit the required sample size can be determined. Note in equation 7(iv) since  $n$  is given by squared value it does not matter if we use a negative or positive value for  $z$ . The following worked example illustrates the concept of confidence intervals and sample size for an infinite population.

### Application for determining sample size: *Coffee*

The quality control inspector of the filling machine for coffee wants to estimate the mean weight of coffee in its 200 gram jars to within  $\pm 0.50$  g. It is known that the standard deviation of the coffee filling machine is 2 g.

1. What sample size should the inspector take to be 95% confidence of the estimate?

Using equation 7(iv),

$$n = \left( \frac{z\sigma_x}{e} \right)^2$$

The area in the each tail for a 95% confidence limit is 2.5%. Using [function NORMSINV] in Excel for a value  $P(x)$  of 2.5% gives a lower value of  $z$  of  $-1.9600$ . Since the distribution is symmetrical, the upper value is numerically the same at  $+1.9600$ . (Note: an alternative way of finding the upper value of  $z$  is to enter in [function NORMSINV] the value of 97.50% (2.50% + 95%) which is the area of the curve from the left to the upper value of  $z$ .)

Here,

$z$  is 1.9600 (it does not matter whether we use plus or minus since we square the value)

$\sigma_x$  is 2 g

$e$  is  $\pm 0.50$  g

$$n = \left( \frac{1.9600 * 2.00}{0.50} \right)^2 = 61.463 \\ = 62 \text{ (rounded up)}$$

Thus the quality control inspector should take a sample size of 62 (61 would be just slightly too small).

### Confidence interval of the mean for a finite population

As discussed in Chapter 6 (equation 6(vi)), if the population is considered finite, that is the ratio  $n/N$  is greater than 5%, then the standard error should be modified by the finite population multiplier according to the expression,

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \quad 6(\text{vi})$$

In this case the confidence limits for the population estimation from equation 7(i) are modified as follows:

$$\bar{x} \pm z\sigma_{\bar{x}} = \bar{x} \pm z \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{(N-n)}{(N-1)}} \quad 7(\text{v})$$

### Application of the confidence interval for a finite population: *Printing*

A printing firm runs off the first edition of a textbook of 496 pages. After the book is printed, the quality control inspector looks at 45 random pages selected from the book and finds that the average number of errors in these pages is 2.70. These include printing errors of colour and alignment, but also typing errors which originate from the author and the editor. The

inspector knows that based on past contracts for a first edition of a book the standard deviation of the number of errors per page is 0.5.

1. What is a 95% confidence interval for the mean number of errors in the book?

Sample size,  $n$ , is 45

Population size,  $N$ , is 496

Sample mean,  $\bar{x}$ , errors per page is 2.70

Population standard deviation,  $\sigma$ , is 0.5

Ratio of  $n/N$  is  $45/496 = 9.07\%$

This value is greater than 5%, thus, we must use the finite population multiplier:

$$\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{496-45}{496-1}} = \sqrt{\frac{451}{495}} = 0.9545$$

Uncorrected standard error of the mean is

$$\frac{\sigma_x}{\sqrt{n}} = \frac{0.5}{\sqrt{45}} = 0.0745$$

Corrected standard error of the mean,

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = 0.0745 * 0.9545 \\ &= 0.0711\end{aligned}$$

Confidence level is 95%, thus area in each tail is 2.5%

Using [function NORMSINV] in Excel for a value  $P(x)$  of 2.5% gives a lower value of  $z$  of  $-1.9600$ . Since the distribution is symmetrical, the upper value is numerically the same at  $+1.9600$ .

Thus from equation 7(v) the lower confidence limit is,

$$2.70 - 1.9600 * 0.0711 = 2.56$$

Thus from equation 7(v) the upper confidence limit is,

$$2.70 + 1.9600 * 0.0711 = 2.84$$

Thus we could say that the best estimate of the errors in the book is 2.70 per page and that we are 95% confident that the errors lie between 2.56 and 2.84 errors per page.

## Estimating the Mean Using the Student- $t$ Distribution

There may be situations in estimating when we do not know the population standard deviation and that we have small sample sizes. In this case there is an alternative distribution that we apply called the Student- $t$  distribution, or more simply the  $t$ -distribution.

### The Student- $t$ distribution

In Chapter 6, in the paragraph entitled, “Sample size and shape of the sampling distribution of the means”, we indicated that the sample size taken has an influence on the shape of the sampling distributions of the means. If we sample from population distributions that are normal, such that we know the standard deviation,  $\sigma$ , any sample size will give a sampling distribution of the means that are approximately normal. However, if we sample from populations that are not normal, we are obliged to increase our sampling size to at least 30 units in order that the sampling distribution of the means will be approximately normally distributed. Thus, what do we do when we have small sample sizes that are less than 30 units? To be correct, we should use a **Student- $t$  distribution**.

The Student- $t$  distribution, like the normal distribution, is a continuous distribution for small amounts of data. It was developed by William Gossett of the Guinness Brewery, in Dublin, Ireland in 1908 (presumably when he had time between beer production!) and published under the pseudonym “student” as the Guinness company would not allow him to put his own name to the development. The Student- $t$  distributions are a family of distributions each one having a different shape and characterized by a parameter called the degrees of freedom. The density function, from which the Student- $t$

distribution is drawn, has the following relationship:

$$f(t) = \frac{[(v-1)/2]!}{\sqrt{v\pi}[(v-2)/2]} \left[1 + \frac{t^2}{v}\right]^{-(v+1)/2} \quad 7(\text{vi})$$

Here,  $v$  is the degree of freedom,  $\pi$  is the value of 3.1416, and  $t$  is the value on the  $x$ -axis similar to the  $z$ -value of a normal distribution.

### Degrees of freedom in the Student- $t$ distribution

Literally, the **degrees of freedom** means the choices that you have regarding taking certain actions. For example, what is the degree of freedom that you have in manoeuvring your car into a parking slot? What is the degree of freedom that you have in contract or price negotiations? What is the degree of freedom that you have in negotiating a black run on the ski slopes? In the context of statistics the **degrees of freedom in a Student- $t$  distribution** are given by  $(n-1)$  where  $n$  is the sample size. This then implies that there is a degree of freedom for every sample size. To understand quantitatively the degrees of freedom consider the following.

There are five variables  $v$ ,  $w$ ,  $x$ ,  $y$ , and  $z$  that are related by the following equation:

$$\frac{v + w + x + y + z}{5} = 13 \quad 7(\text{vii})$$

Since there are five variables we have a choice, or the degree of freedom, to select four of the five. After that, the value of the fifth variable is automatically fixed. For example, assume that we give  $v$ ,  $w$ ,  $x$ , and  $y$  the values 14, 16, 12, and 18, respectively. Then from equation 7(vii) we have,

$$\begin{aligned} \frac{14 + 16 + 12 + 18 + z}{5} &= 13 \\ z &= 5 * 13 - (14 + 16 + 12 + 18) \\ &= 65 - 60 = 5 \end{aligned}$$

Thus automatically the fifth variable,  $z$ , is fixed at a value of 5 in order to retain the validity of the equation. Here we had five variables to give a degree of freedom of four. In general terms, for a sample size of  $n$  units, the degree of freedom is the value determined by  $(n-1)$ .

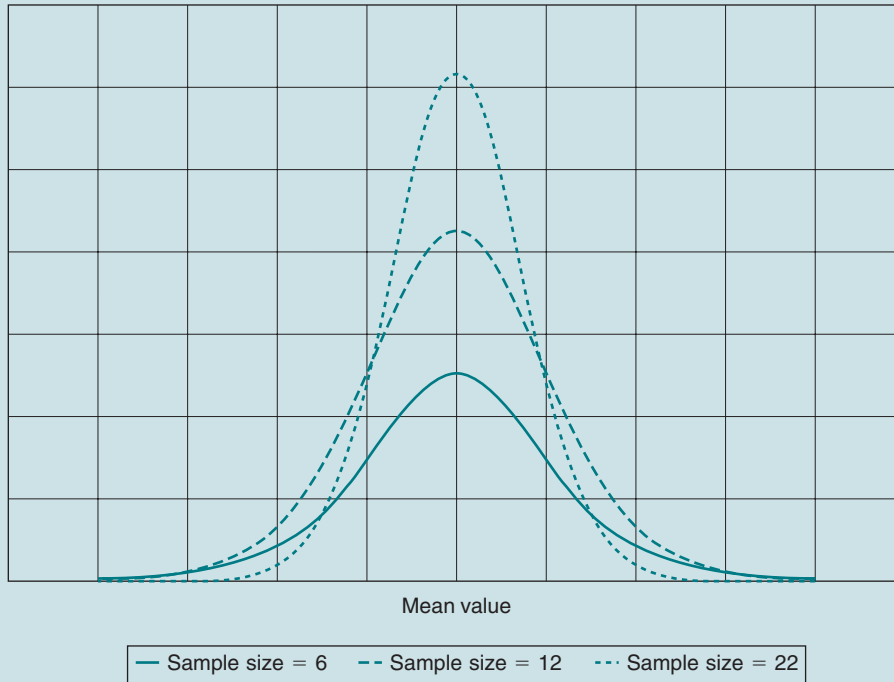
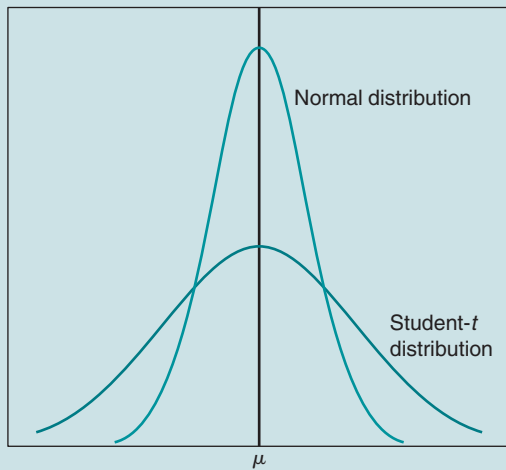
### Profile of the Student- $t$ distribution

Three Student- $t$  distributions, for sample size  $n$  of 6, 12, and 22, or sample sizes less than 30, are illustrated in Figure 7.4. The degrees of freedom for these curves, using  $(n-1)$  are respectively 5, 11, and 21. These three curves have a profile similar to the normal distribution but if we superimposed a normal distribution on a Student- $t$  distribution as shown in Figure 7.5, we see that the normal distribution is higher at the peak and the tails are closer to the  $x$ -axis, compared to the Student- $t$  distribution. The Student- $t$  distribution is flatter and you have to go further out on either side of the mean value before you are close to the  $x$ -axis indicating greater variability in the sample data. This is the penalty you pay for small sample sizes and where the sampling is taken from a non-normal population. As the sample size increases the profile of the Student- $t$  distribution approaches that of the normal distribution and as is illustrated in Figure 7.4 the curve for a sample size of 22 has a smaller variation and is higher at the peak.

### Confidence intervals using a Student- $t$ distribution

When we have a normal distribution the confidence intervals of estimating the mean value of the population are as given in equation 7(i):

$$\bar{x} \pm z \frac{\sigma_x}{\sqrt{n}} \quad 7(\text{i})$$

Figure 7.4 Three Student-*t* distributions for different sample sizes.Figure 7.5 Normal and Student-*t* distributions.

When we are using a Student-*t* distribution, Equation 7(i) is modified to give the following:

$$\bar{x} \pm t \frac{\hat{\sigma}_x}{\sqrt{n}} \quad 7(\text{viii})$$

Here the value of *t* has replaced *z*, and  $\hat{\sigma}$  has replaced  $\sigma$ , the population standard deviation. This new term,  $\hat{\sigma}$ , means an estimate of the population standard deviation. Numerically it is equal to *s*, the sample standard deviation by the relationship,

$$\hat{\sigma} = s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{(n - 1)}} \quad 7(\text{ix})$$

We could avoid writing  $\hat{\sigma}$ , as some texts do, and simply write *s* since they are numerically the same. However, by putting  $\hat{\sigma}$  it is clear that our only alternative to estimate our confidence

limits is to use an estimate of the population standard deviation as measured from the sample.

## Excel and the Student-*t* distribution

There are two functions in Excel for the Student-*t* distribution. One is [function TDIST], which determines the probability or area for a given random variable  $x$ , the degree of freedom, and the number of tails in the distribution. When we use the *t*-distribution in estimating, the number of tails is always two – that is, one on the left and one on the right. (This is not necessarily the case for hypothesis testing that is discussed in Chapter 8.) The other function is [function TINV] and this determines the value of the Student-*t* under the distribution given the total area outside the curve or  $\alpha$ . (Note the difference in the way you enter the variables for the Student-*t* and the normal distribution. For the Student-*t* you enter the area in the tails, whereas for the normal distribution you enter the area of the curve from the extreme left to a value on the  $x$ -axis.)

## Application of the Student-*t* distribution: Kiwi fruit

Sheila Hope, the Agricultural inspector at Los Angeles, California wants to know in milligrams, the level of vitamin C in a boat load of kiwi fruits imported from New Zealand, in order to compare this information with kiwi fruits grown in the Central Valley, California. Sheila took a random sample of 25 kiwis from the ship's hold and measured the vitamin C content. Table 7.1 gives the results in milligrams per kiwi sampled.

1. Estimate the average level of vitamin C in the imported kiwi fruits and give a 95% confidence level of this estimate.

Since we have no information about the population standard deviation, and the sample size of 25 is less than 30, we use a Student-*t* distribution.

Table 7.1 Milligrams of vitamins per kiwi sampled.

109	88	91	136	93
101	89	97	115	92
114	106	94	109	110
97	89	117	105	92
83	79	107	100	93

Using [function AVERAGE], mean value of the sample,  $\bar{x}$ , is 100.24.

Using [function STDEV], standard deviation of the sample,  $s$ , is 12.6731.

Sample size,  $n$ , is 25.

Using [function SQRT], square root of the sample size,  $\sqrt{n}$ , is 5.00.

Estimate of the population standard deviation,  $\hat{\sigma} = s = 12.6731$ .

Standard error of the sample distribution,

$$\frac{\hat{\sigma}_x}{\sqrt{n}} = \frac{12.6731}{5.00} = 2.5346$$

Required confidence level (given) is 95%.

Area outside of confidence interval,  $\alpha$ , (100% – 95%) is 5%.

Degrees of freedom,  $(n - 1)$ , is 24.

Using [function TINV], Student-*t* value is 2.0639.

From equation 7(viii),

$$\begin{aligned} \text{Lower confidence level, } \bar{x} - t \frac{\hat{\sigma}_x}{\sqrt{n}} \\ &= 100.24 - 2.0639 * 2.5346 \\ &= 100.24 - 5.2312 = 95.01 \end{aligned}$$

$$\begin{aligned} \text{Upper confidence level, } \bar{x} + t \frac{\hat{\sigma}_x}{\sqrt{n}} \\ &= 100.24 + 2.0639 * 2.5346 \\ &= 100.24 + 5.2312 = 105.47 \end{aligned}$$

Thus the estimate of the average level of vitamin C in all the imported kiwis is 100.24 mg with a 95% confidence that the lower level of our estimate is 95.01 mg and the upper level

is 105.47 mg. This information is illustrated on the Student- $t$  distribution in Figure 7.6.

## Sample size and the Student- $t$ distribution

We have said that the Student- $t$  distribution should be used when the sample size is less than 30 and the population standard deviation is unknown. Some analysts are more rigid and use a sample size of 120 as the cut-off point. What should we use, a sample size of 30 or a sample size of 120? The movement of the value of  $t$  relative to the value of  $z$  is illustrated by the data in Table 7.2 and the corresponding graph in

Figure 7.6 Confidence intervals for kiwi fruit.

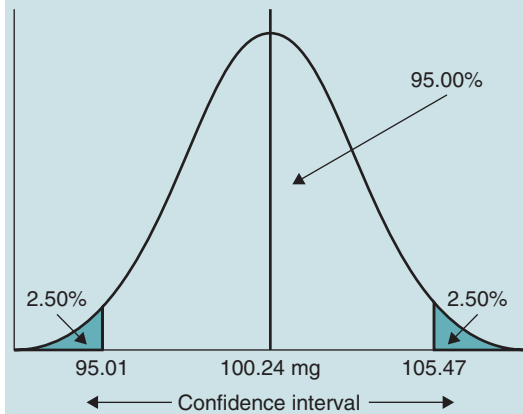


Table 7.2 Values of  $t$  and  $z$  with different sample sizes.

Confidence level	95.00%						
Area outside	5.00%						
Excel (lower)	2.50%						
Excel (upper)	97.50%						
Sample size, $n$	Upper Student- $t$	Upper $z$	$\frac{(t - z)}{z}$	Sample size, $n$	Upper Student- $t$	Upper $z$	$\frac{(t - z)}{z}$
5	2.7765	1.9600	41.66%	105	1.9830	1.9600	1.18%
10	2.2622	1.9600	15.42%	110	1.9820	1.9600	1.12%
15	2.1448	1.9600	9.43%	115	1.9810	1.9600	1.07%
20	2.0930	1.9600	6.79%	120	1.9801	1.9600	1.03%
25	2.0639	1.9600	5.30%	125	1.9793	1.9600	0.99%
30	2.0452	1.9600	4.35%	130	1.9785	1.9600	0.95%
35	2.0322	1.9600	3.69%	135	1.9778	1.9600	0.91%
40	2.0227	1.9600	3.20%	140	1.9772	1.9600	0.88%
45	2.0154	1.9600	2.83%	145	1.9766	1.9600	0.85%
50	2.0096	1.9600	2.53%	150	1.9760	1.9600	0.82%
55	2.0049	1.9600	2.29%	155	1.9755	1.9600	0.79%
60	2.0010	1.9600	2.09%	160	1.9750	1.9600	0.77%
65	1.9977	1.9600	1.93%	165	1.9745	1.9600	0.74%
70	1.9949	1.9600	1.78%	170	1.9741	1.9600	0.72%
75	1.9925	1.9600	1.66%	175	1.9737	1.9600	0.70%
80	1.9905	1.9600	1.56%	180	1.9733	1.9600	0.68%
85	1.9886	1.9600	1.46%	185	1.9729	1.9600	0.66%
90	1.9870	1.9600	1.38%	190	1.9726	1.9600	0.64%
95	1.9855	1.9600	1.30%	195	1.9723	1.9600	0.63%
100	1.9842	1.9600	1.24%	200	1.9720	1.9600	0.61%

Figure 7.7 As the sample size increases the value of  $t$  approaches  $z$ .

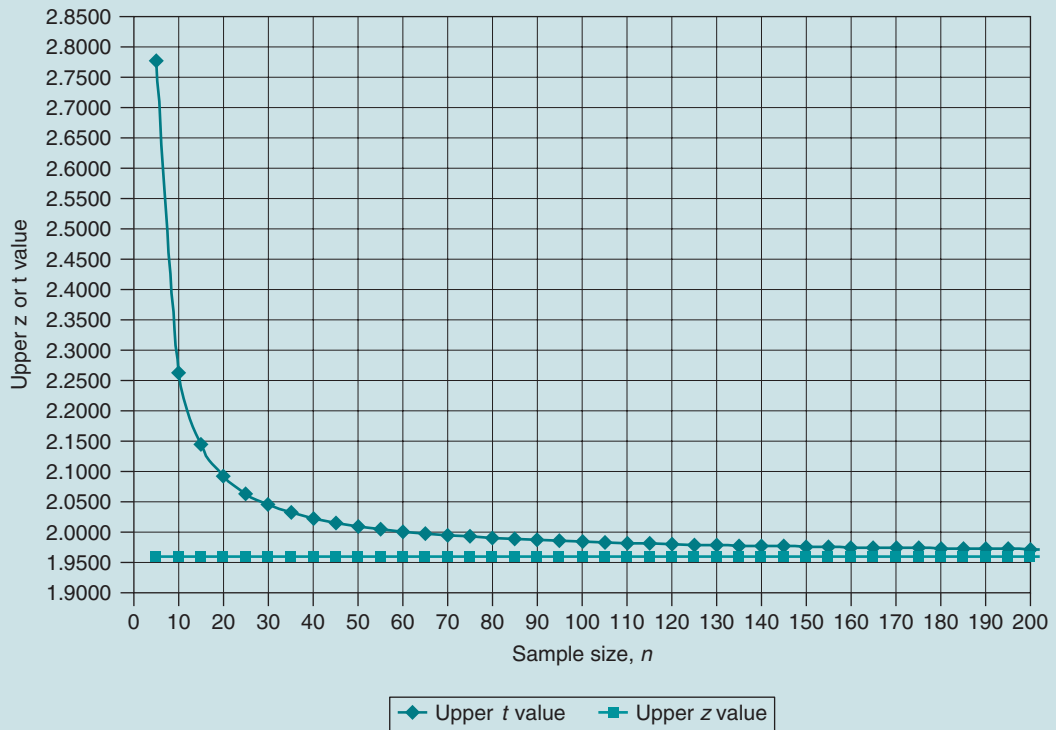


Figure 7.7. Here we have the Student- $t$  value for a confidence level of 95% for sample sizes ranging from 5 to 200. The value of  $z$  is also shown and this is constant at the 95% confidence level since  $z$  is not a function of sample size. In the column  $(t - z)/z$  we see that the difference between  $t$  and  $z$  is 4.35% for a sample size of 30. When the sample size increases to 120 then the difference is just 1.03%. Is this difference significant? It really depends on what you are sampling. We have to remember that we are making estimates so that we must expect errors. In the medical field small differences may be important but in the business world perhaps less so. Let

us take another look at the kiwi fruit example from above using  $z$  rather than  $t$  values.

### Re-look at the example *Kiwi fruit* using the normal distribution

Here all the provided data and the calculations are the same as previously but we are going to assume that we can use the normal distribution for our analysis.

Required confidence level (given) is 95%.

Area outside of confidence interval,  $\alpha$ ,  $(100\% - 95\%)$  is 5%, which means that there is an area of 2.5% in both tails for a symmetrical



distribution. Using [function NORMSINV] in Excel for a value  $P(x)$  of 2.5% the value of  $z$  is  $\pm 1.9600$ .

From equation 7(i),

$$\begin{aligned}\text{Lower confidence level, } \bar{x} - z \frac{\hat{\sigma}_x}{\sqrt{n}} \\ &= 100.24 - 1.9600 * 2.5346 \\ &= 100.24 - 4.9678 = 95.27\end{aligned}$$

$$\begin{aligned}\text{Upper confidence level, } \bar{x} + z \frac{\hat{\sigma}_x}{\sqrt{n}} \\ &= 100.00 + 1.9600 * 2.5346 \\ &= 100.24 + 4.9678 = 105.21\end{aligned}$$

The corresponding values that we obtained by using the Student- $t$  distribution were 95.01 and 105.47 or a difference of only some 0.3%. Since in reality we would report probability of our confidence for the vitamin level of kiwis between 95 and 105 mg, the difference between using  $z$  and  $t$  in this case is insignificant.

## Estimating and Auditing

Auditing is the methodical examination of financial accounts, inventory items, or operating processes to verify that they confirm with standard practices or targeted budget levels.

### Estimating the population amount

We can use the concepts that we have developed in this chapter to estimate the total value of goods such as, for example, inventory held in a distribution centre when, for example, it is impossible or very time consuming to make an audit of the population. In this case we first take a random and representative sample and determine the mean financial value  $\bar{x}$ . If  $N$  is the total number of units, then the point estimate for the population total is the size of the population,  $N$ , multiplied by the sample mean, or,

$$\text{Total} = N\bar{x} \quad 7(x)$$

It is unlikely we know the standard deviation of the large population of inventory and so we would estimate the value from the sample. If the sample size is less than 30 we use the Student- $t$  distribution and the confidence intervals are given as follows by multiplying both terms in equation 7(viii) to give,

$$\text{Confidence intervals: } N\bar{x} \pm Nt \frac{\hat{\sigma}}{\sqrt{n}} \quad 7(\text{xi})$$

Alternatively, if the population is considered finite, that is the ratio of  $n/N \geq 5\%$ , then the standard error has to be modified by the estimated finite population multiplier to give,

$$\text{Estimated standard error: } \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad 7(\text{xii})$$

Thus the confidence intervals when the standard deviation is unknown, the sample size is less than 30, and the population is finite, are,

$$\begin{aligned}\text{Confidence intervals:} \\ N\bar{x} \pm Nt \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad 7(\text{xiii})\end{aligned}$$

The following two applications illustrate the use of estimating the total population amount for auditing purposes.

### Application of auditing for an infinite population: tee-shirts

A store on Duval Street in Key West Florida, wishes to estimate the total retail value of its tee-shirts, tank tops, and sweaters that it has in its store. The inventory records indicate that there are 4,500 of these clothing articles on the shelves. The owner takes a random sample of 29 items and Table 7.3 gives the prices in dollars indicated on the articles.



Table 7.3 Tee shirts – prices in \$US.

16.50	25.00	25.50	42.00	37.00	22.00
21.00	20.00	21.00	9.50	24.50	11.50
52.50	15.50	32.50	18.00	18.50	19.00
29.50	16.00	21.00	44.00	17.50	50.50
27.00	29.50	12.50	32.00	23.00	

1. Estimate the total retail value of the clothing items within a 99% confidence limit.

Using Excel [function AVERAGE] the sample mean value,  $\bar{x}$ , is \$25.31.

Population size,  $N$ , is 4,500

Estimated total retail value is  $N \bar{x} = 4,500 * 25.31$  or \$113,896.55.

Sample size,  $n$ , is 29.

Ratio  $n/N$  is  $29/4,500$  or 0.64%.

Since this value is less than 5% we do not need to use the finite population multiplier.

Sample standard deviation,  $s$ , is \$11.0836.

Estimated population standard deviation,  $\hat{\sigma}$ , is \$11.0836.

Estimated standard error of the sample distribution,  $\hat{\sigma}_x/\sqrt{n} = 11.0836/\sqrt{29}$ , is 2.0582.

Since we do not know the population standard deviation, and the sample size is less than 30 we use the Student- $t$  distribution.

Degrees of freedom ( $n - 1$ ) is 28.

Using Excel [function TINV] for a 99% confidence level, Student- $t$  value is 2.7633.

From equation 7(xi) the lower confidence limit for the total value is,

$$N\bar{x} - Nt \frac{\hat{\sigma}}{\sqrt{n}} = \$113,896.55 - 4,500 * 2.7633 * 2.0582 \text{ or } \$88,303.78$$

and the upper confidence limit is,

$$N\bar{x} + Nt \frac{\hat{\sigma}}{\sqrt{n}} = \$113,896.55 + 4,500 * 2.7633 * 2.0582 \text{ or } \$139,489.33$$

Thus the owner estimates the average, or point estimate, of the total retail value of the clothing items in his Key West store as \$113,897 (rounded) and he is 99% confident that the value lies between \$88,303.78 (say \$88,304 rounded) and \$139,489.33 (say \$139,489 rounded).

## Application of auditing for a finite population: paperback books

A newspaper and bookstore at Waterloo Station wants to estimate the value of paper backed books it has in its store. The owner takes a random sample of 28 books and determines that the average retail value is £4.57 with a sample standard deviation of 53 pence. There are 12 shelves of books and the owner estimates that there are 45 books per shelf.

1. Estimate the total retail value of the books within a 95% confidence limit.

Estimated population amount of books,  $N$ , is  $12 * 45$  or 540.

Mean retail value of books is £4.57.

Estimated total retail value is  $N \bar{x} = 540 * 4.57$  or £2,467.80.

Sample size,  $n$ , is 28.

Ratio  $n/N$  is  $28/540$  or 5.19%.

Since this value is greater than 5% we use the finite population multiplier

Finite population multiplier  $\sqrt{\frac{N-n}{N-1}} =$

$$\sqrt{\frac{540-28}{540-1}} = \sqrt{\frac{512}{539}} = 0.9746.$$

Sample standard deviation,  $s$ , is £0.53.

Estimated population standard deviation,  $\hat{\sigma}$ , is £0.53.

From equation 7(xii) the estimated standard error is,

$$\frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{0.53}{\sqrt{28}} * 0.9746 = 0.0976$$

Degrees of freedom ( $n - 1$ ) is 27.

Using Excel [function **TINV**] for a 95% confidence level, Student- $t$  value is 2.0518.

From equation 7(xiii) the lower confidence limit is,

$$\begin{aligned} N\bar{x} - Nt \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \\ = £2,467.80 - 540 * 2.0518 * 0.0976 \\ = £2,359.64 \end{aligned}$$

From equation 7(xiii) the upper confidence limit is,

$$\begin{aligned} N\bar{x} + Nt \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \\ = £2,467.80 + 540 * 2.0518 * 0.0976 \\ = £2,575.96 \end{aligned}$$

Thus the owner estimates the average, or point estimate, of the total retail value of the paper back books in the store as £2,467.80 (£2,468 rounded) and that she is 95% confident that the value lies between £2,359.64 (say £2,360 rounded) and £2,575.96 (say £2,576 rounded).

## Estimating the Proportion

Rather than making an estimate of the mean value of the population, we might be interested to estimate the proportion in the population. For example, we take a sample and say that our point estimate of the proportion expected to vote conservative in the next United Kingdom election is 37% and that we are 90% confident that the proportion will be in the range of 34% and 40%. When dealing with proportions then the sample proportion,  $\bar{p}$ , is a point estimate of the population proportion  $p$ . The value  $\bar{p}$  is determined by taking a sample of size  $n$  and measuring the proportion of successes.

## Interval estimate of the proportion for large samples

When analysing the proportions of a population then from Chapter 6 we developed the following equation 6(xi) for the standard error of the proportion,  $\sigma_{\bar{p}}$ :

$$\sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}} \quad 6(\text{xi})$$

where  $n$  is the sample size and  $p$  is the population proportion of *successes* and  $q$  is the population proportion of *failures* equal to  $(1 - p)$ . Further, from equation 6(xv),

$$z = \pm \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Reorganizing this equation we have the following expression for the confidence intervals for the estimate of the population proportion as follows:

$$\bar{p} = p \pm z \sqrt{\frac{p(1-p)}{n}} \quad 7(\text{xiv})$$

Thus, analogous to the estimation for the means, this implies that the confidence intervals for an estimate of the population proportion lie in the range given by the following expression:

$$\bar{p} - z \sqrt{\frac{p(1-p)}{n}} \leq p \leq \bar{p} + z \sqrt{\frac{p(1-p)}{n}} \quad 7(\text{xv})$$

If we do not know the population proportion,  $p$ , then the standard error of the proportion can be estimated from the following equation by replacing  $p$  with  $\bar{p}$ :

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad 7(\text{xvi})$$

In this case,  $\hat{\sigma}_{\bar{p}}$  is the **estimated standard error of the proportion** and  $\bar{p}$  is the sample proportion of successes. If we do this then equation 7(xv) is modified to give the expression,

$$\bar{p} - z \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \leq p \leq \bar{p} + z \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad 7(\text{xvii})$$

## Sample size for the proportion for large samples

In a similar way for the mean, we can determine the sample size to take in order to estimate the population proportion for a given confidence level. From the relationship of 7(xiv) the intervals for the estimate of the population proportion are,

$$\bar{p} - p = \pm z \sqrt{\frac{p(1-p)}{n}} \quad 7(\text{xviii})$$

Squaring both sides of the equation we have,

$$(\bar{p} - p)^2 = z^2 \frac{p(1-p)}{n}$$

Making  $n$ , the sample size the subject of the equation gives,

$$n = z^2 \frac{p(1-p)}{(\bar{p} - p)^2} \quad 7(\text{xix})$$

If we denote the sample error,  $(\bar{p} - p)$  by  $e$  then the sample size is given by the relationship,

$$n = z^2 \frac{p(1-p)}{e^2} \quad 7(\text{xx})$$

While using this equation, a question arises as to what value to use for the true population proportion,  $p$ , when this is actually the value that we are trying to estimate! One possible approach is to use the value of  $\bar{p}$  if this is available. Alternatively, we can use a value of  $p$  equal to 0.5 or 50% as this will give the most conservative sample size. This is because for a given value of the confidence level say 95% which defines  $z$ , and the required sample error,  $e$ , then a value of  $p$  of 0.5 gives the maximum possible value of 0.25 in the numerator of equation 7(xx). This is shown in Table 7.4 and illustrated by the graph in Figure 7.8. The following is an application of the estimation for proportions including an estimation of the sample size.

Table 7.4 Conservative value of  $p$  for sample size.

$p$	$(1 - p)$	$p(1 - p)$
0.00	1.00	0.0000
0.10	0.90	0.0900
0.20	0.80	0.1600
0.30	0.70	0.2100
0.40	0.60	0.2400
0.50	0.50	0.2500
0.60	0.40	0.2400
0.70	0.30	0.2100
0.80	0.20	0.1600
0.90	0.10	0.0900
1.00	0.00	0.0000

## Application of estimation for proportions: *Circuit boards*

In the manufacture of electronic circuit boards a sample of 500 is taken from a production line and of these 15 are defective.

1. What is a 90% confidence interval for the proportion of all the defective circuit boards produced in this manufacturing process?

Proportion defective,  $\bar{p}$ , is  $15/500 = 0.030$ .

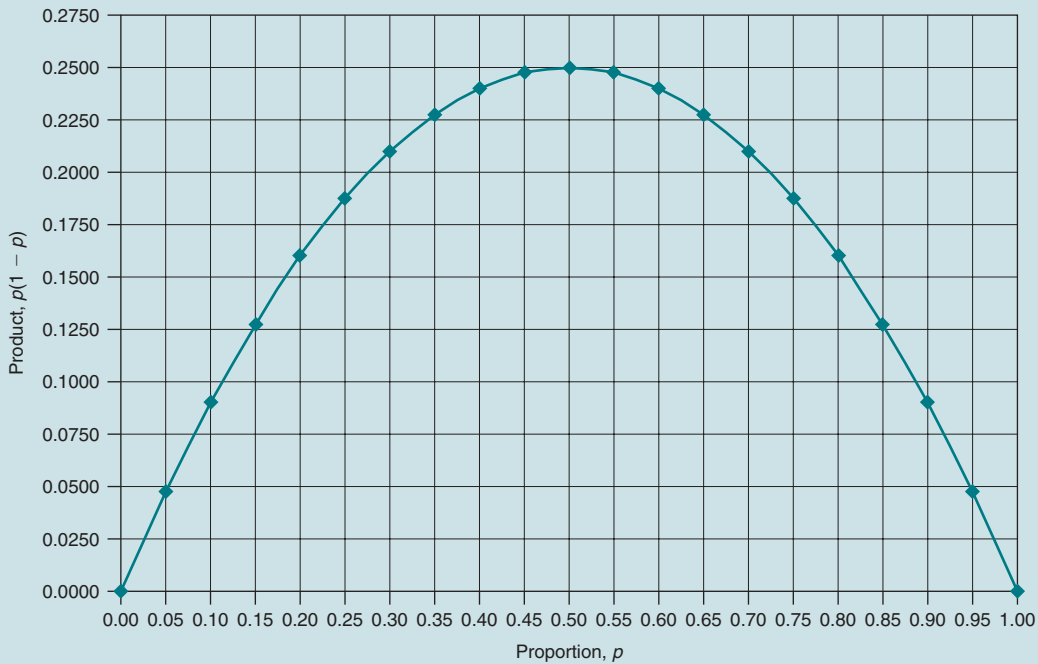
Proportion that is good is  $1 - 0.030$  or also

$$\frac{500 - 15}{500} = 0.97.$$

From equation 7(xvi) the estimate of the standard error of the proportion is,

$$\begin{aligned} \hat{\sigma}_{\bar{p}} &= \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \sqrt{\frac{0.03 * 0.97}{500}} \\ &= \sqrt{\frac{0.0291}{500}} = 0.0076 \end{aligned}$$

When we have a 90% confidence interval, and assuming a normal distribution, then the area of the distribution up to the lower confidence level is  $(100\% - 90\%)/2 = 5\%$

Figure 7.8 Relation of the product,  $p(1 - p)$  with the proportion,  $p$ .

and the area of the curve up to the upper confidence level is  $5\% + 90\% = 95\%$ .

From Excel [function NORMSINV], value of  $z$  at the area of 5% is  $-1.6449$ .

From Excel [function NORMSINV], value of  $z$  at the area of 95% is  $+1.6449$ .

From equation 7(xvii) the lower confidence limit is,

$$\begin{aligned}\bar{p} - z\hat{\sigma}_{\bar{p}} &= 0.03 - 1.6449 * 0.0076 \\ &= 0.03 - 0.0125 = 0.0175\end{aligned}$$

From equation 7(xvii) the upper confidence limit is,

$$\begin{aligned}\bar{p} + z\hat{\sigma}_{\bar{p}} &= 0.03 + 1.6449 * 0.0076 \\ &= 0.03 + 0.0125 = 0.0425\end{aligned}$$

Thus we can say that from our analysis, the proportion of all the manufactured circuit

boards which are defective is 0.03 or 3%. Further, we are 90% confident that this proportion lies in the range of 0.0175 or 1.75% and 0.0425 or 4.25%.

- If we required our estimate of the proportion of all the defective manufactured circuit boards to be within a margin of error of  $\pm 0.01$  at a 98% confidence level, then what size of sample should we take?

When we have a 98% confidence interval, and assuming a normal distribution, then the area of the distribution up to the lower confidence level is  $(100\% - 98\%)/2 = 1\%$  and the area of the curve up to the upper confidence level is  $1\% + 98\% = 99\%$ . From the Excel normal distribution function we have the following.

From Excel [function NORMSINV], value of  $z$  at the area of 1% is  $-2.3263$ .

From Excel [function **NORMSINV**], value of  $z$  at the area of 99% is +2.3263.

The sample error,  $e$ , is 0.01.

The sample proportion  $\bar{p}$  is used for the population proportion  $p$  or 0.03.

Using equation 7(xx),

$$\begin{aligned} n &= z^2 \frac{p(1-p)}{e^2} \\ &= 2.3263^2 * 0.03 * 0.97 / 0.01^2 \\ &= \frac{0.1575}{0.0001} = 1,575 \end{aligned}$$

It does not matter which value of  $z$  we use,  $-2.3263$  or  $+2.3263$ , since we are squaring  $z$  and the negative value becomes positive.

Thus the sample size to estimate the population proportion of the number of defective circuits within an error of margin of error of  $\pm 0.01$  from the true proportion is 1,575.

An alternative, more conservative approach is to use a value of  $p = 0.5$ . In this case the sample size to use is,

$$\begin{aligned} n &= z^2 \frac{p(1-p)}{e^2} \\ &= 2.3263^2 * 0.50 * 0.50 / 0.01^2 \\ &= \frac{0.2500}{0.0001} = 2,500 \end{aligned}$$

This value of 2,500 is significantly higher than 1,575 and would certainly add to the cost of the sampling experiment with not necessarily a significant gain in the accuracy of the results.

## Margin of Error and Levels of Confidence

When we make estimates the question arises (or at least it should) “How good is your estimate?” That is to say, what is the margin of error? In addition, we might ask, “Why don’t we always

use a high confidence level of say 99% as this would signify a high degree of accuracy?” These two issues are related and are discussed below.

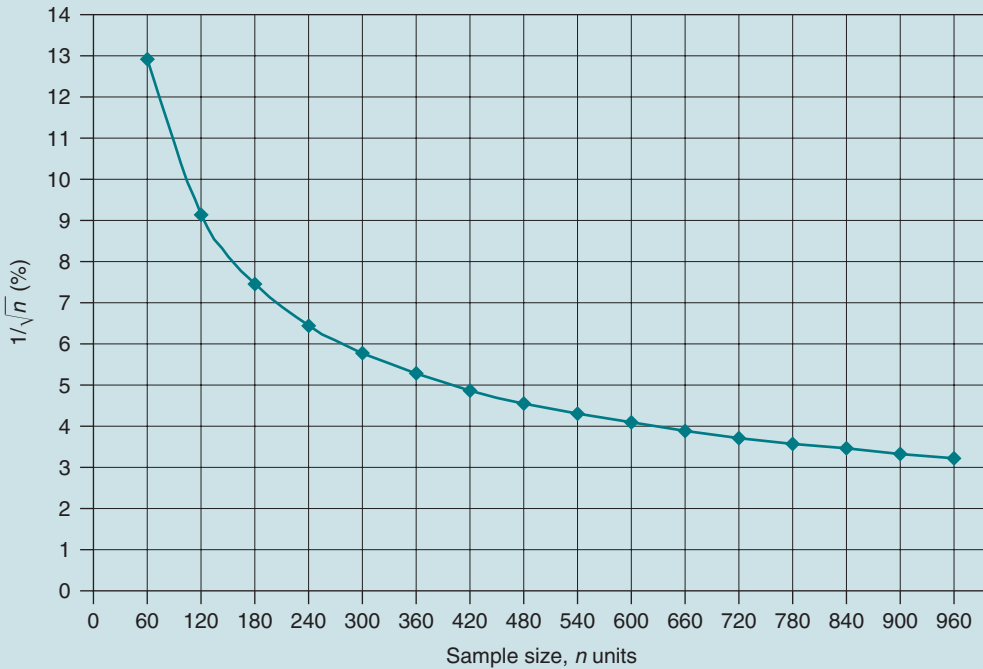
## Explaining margin of error

When we analyse our sample we are trying to estimate the population parameter, either the mean value or the proportion. When we do this, there will be a **margin of error**. This is not to say that we have made a calculation error, although this can occur, but the margin of error measures the maximum amount that our estimate is expected to differ from the actual population parameter. The margin of error is a plus or minus value added to the sample result that tells us how good is our estimate.

If we are estimating the mean value then,

$$\text{Margin of error is } \pm z \frac{\sigma_x}{\sqrt{n}} \quad 7(\text{xxi})$$

This is the same as the confidence limits from equation 7(i). In the worked example paper, at a confidence level of 95%, the margin of error is  $\pm 1.9600 * 0.0013$  or  $\pm 0.0025$  cm. Thus, another way of reporting our results is to say that we estimate that the width of all the computer paper from the production line is 20.9986 cm and we have a margin of error of  $\pm 0.0025$  cm at a 95% confidence. Now if we look at equation 7(xxi), when we have a given standard deviation and a given confidence level the only term that can change is the sample size  $n$ . Thus we might say, let us analyse a bigger sample in order to obtain a smaller margin of error. This is true, but as can be seen from Figure 7.9, which gives the ratio of  $1/\sqrt{n}$  as a percentage according to the sample size in units, there is a diminishing return. Increasing the sample size does reduce the margin of error but at a decreasing rate. If we double the sample size from 60 to 120 units the ratio of  $1/\sqrt{n}$  changes from 12.91% to 9.13% or a difference

Figure 7.9 The change of  $1/\sqrt{n}$  with increase of sample size.

of 3.78%. From a sample size of 120 to 180 the value of  $1/\sqrt{n}$  changes from 9.13% to 7.45% or a difference of 1.68% or, if we go from a sample size of 360 to 420 units the value of  $1/\sqrt{n}$  goes from 5.27% to 4.88% or a difference of only 0.39%. With the increasing sample size the cost of testing of course increases and so there has to be a balance between the size of the sample and the cost.

If we are estimating for proportions then the margin of error is from equation 7(xvii) the value,

$$\pm z \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad 7(\text{xxii})$$

Since for proportions we are trying to estimate the percentage for a situation then the margin of error is a plus or minus percentage. In the

worked example circuit boards the margin of error at a 90% level of confidence is,

$$\begin{aligned} \hat{\sigma}_{\bar{p}} &= \pm z \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \pm 1.6449 \sqrt{\frac{0.03 * 0.97}{500}} \\ &= \pm 0.0125 = \pm 1.25\% \end{aligned}$$

This means that our estimate could be 1.25% more or 1.25% less than our estimated proportion or a range of 2.50%. The margin of error quoted in a sampling situation is important as it can give uncertainty to our conclusions. If we look at Figure 7.1, for example, we see that 52% of the Italian population is against Turkey joining the European Union. Based on just this information we might conclude that the majority of the Italians are against Turkey's membership. However, if we then bring in the  $\pm 3\%$  margin of error then this means that we can

Table 7.5 Questions asked in house construction.

Your question	Constructor's response	Implied confidence interval	Implied confidence level
1. Will my house be finished in 10 years?	I am certain	>99%	10 years
2. Will my house be finished in 5 years?	I am pretty sure	>95%	5 years
3. Will my house be finished in 2 years?	I think so	>80%	2 years
4. Will my house be finished in 18 months?	Possibly	About 50%	1.5 years
5. Will my house be finished in 6 months?	Probably not	About 1%	0.50 years

have 49% against Turkey joining the Union (52 – 3), which is not now the majority of the population. Our conclusions are reversed and in cases like these we might hear the term for the media “the results are too close to call”. Thus, the margin of error must be taken into account when surveys are made because the result could change. If the margin of error was included in the survey result of the Dewey/Truman election race, as presented in the Box Opener of Chapter 6, the Chicago Tribune may not have been so quick to publish their morning paper!

## Confidence levels

If we have a confidence level that is high say at 99% the immediate impression is to think that we have a high accuracy in our sampling and estimating process. However this is not the case since in order to have high confidence levels we need to have large confidence intervals or a large margin of error. In this case the large intervals give very broad or fuzzy estimates. This can be illustrated qualitatively as follows.

Assume that you have contracted a new house to be built of 170 m<sup>2</sup> living space on 2,500 m<sup>2</sup> of land. You are concerned about the time taken to complete the project and you ask the constructor various questions concerning the time frame. These are given in the 1st column of Table 7.5. Possible indicated responses to these are given in the 2nd column and the 3rd and 4th columns, respectively, give the implied confidence interval and the implied confidence level. Thus, for a house to be finished in 10 years the constructor is almost certain because this is an inordinate amount of time and so we have put a confidence level of 99%. Again to ask the question for 5 years the confidence level is high at 95%. At 2 years there is a confidence level of 80% if everything goes better than planned. At 18 months there is a 50% confidence if there are, for example, ways to expedite the work. At 6 months we are essentially saying it is impossible. (The time to completely construct a house varies with location but some 18 months to 2 years to build and completely finish all the landscaping is a reasonable time frame.)



This chapter has covered estimating the mean value of a population using a normal distribution and a Student-*t* distribution, using estimating for auditing purposes, estimating the population proportion, and discussed the margin of error and confidence intervals.

### Estimating the mean value

We can estimate the population mean by using the average value taken from a random sample. This is a point estimate. However this single value is often insufficient as it is either right or wrong. A more objective analysis is to give a range of the estimate and the probability, or the confidence, that we have in this estimate. When we do this in sampling from an infinite normal distribution we use the standard error. The standard error is the population standard deviation divided by the square root of the sample size. This is then multiplied by the number of standard deviations in order to determine the confidence intervals. The wider the confidence interval then the higher is our confidence and vice-versa. If we wish to determine a required sample size, for a given confidence interval, this can be calculated from the interval equation since the number of standard deviations,  $z$ , is set by our level of confidence. If we have a finite population we must modify the standard error by the finite population multiplier.

### Estimating the mean using the Student-*t* distribution

When we have a sample size that is less than 30, and we do not know the population standard deviation, to be correct we must use a Student-*t* distribution. The Student-*t* distributions are a family of curves, similar in profile to the normal distribution, each one being a function of the degree of freedom. The degree of freedom is the sample size less one. When we do not know the population standard deviation we must use the sample standard deviation as an estimate of the population standard deviation in order to calculate the confidence intervals. As we increase the size of the sample the value of the Student-*t* approaches the value  $z$  and so in this case we can use the normal distribution relationship.

### Estimating and auditing

The activity of estimating can be extended to auditing financial accounts or values of inventory. To do this we multiply both the average value obtained from our sample, and the confidence interval, by the total value of the population. Since it is unlikely that we know the population standard deviation in our audit experiment we use a Student-*t* distribution and use the sample standard deviation in order to estimate our population standard deviation. When our population is finite, we correct our standard error by multiplying by the finite population multiplier.

### Estimating the proportion

If we are interested in making an estimate of the population proportion we first determine the standard error of the proportion by using the population value, and then multiply this by the number of standard deviations to give our confidence limits. If we do not have a value of the population



proportion then we use the sample value of the proportion to estimate our standard error. We can determine the sample size for a required confidence level by reorganizing the confidence level equation to make the sample size the subject of the equation. The most conservative sample size will be when the value of the proportion  $p$  has a value of 0.5 or 50%.

### Margin of error and levels of confidence

In estimating both the mean and the proportion of a population the margin of error is the maximum amount of difference between the value of the population and our estimated amount. The larger the sample size then the smaller is the margin of error. However, as we increase the size of the sample the cost of our sampling experiment increases and there is a diminishing return on the margin of error with sample size. Although at first it might appear that a high confidence level of say close to 100% indicates a high level of accuracy, this is not the case. In order to have a high confidence level we need to have broader confidence limits and this leads to rather vague or fuzzy estimates.

## EXERCISE PROBLEMS

### 1. Ketchup

#### Situation

A firm manufactures and bottles tomato ketchup that it then sells to retail firms under a private label brand. One of its production lines is for filling 500 g squeeze bottles, which after being filled are fed automatically into packing cases of 20 bottles per case. In the filling operation the firm knows that the standard deviation of the filling operation is 8 g.

#### Required

1. In a randomly selected case, what would be the 95% confidence intervals for the mean weight of ketchup in a case?
2. In a randomly selected case what would be the 99% confidence intervals for the mean weight of ketchup in a case?
3. Explain the differences between the answers to Questions 1 and 2.
4. About how many cases would have to be selected such that you would be within  $\pm 2$  g of the population mean value?
5. What are your comments about this sampling experiment from the point-of-view of randomness?

### 2. Light bulbs

#### Situation

A subsidiary of GE manufactures incandescent light bulbs. The manufacturer sampled 13 bulbs from a lot and burned them continuously until they failed. The number of hours each burned before failure is given below.

342	426	317	545	264	451	1,049	631	512	266	492	562	298
-----	-----	-----	-----	-----	-----	-------	-----	-----	-----	-----	-----	-----

#### Required

1. Determine the 80% confidence intervals for the mean length of the life of light bulbs.
2. How would you explain the concept illustrated by Question 1?
3. Determine the 90%, confidence intervals for the mean length of the life of light bulbs.
4. Determine the 99% confidence intervals for the mean length of the life of light bulbs.
5. Explain the differences between Questions 1, 3, and 4.

### 3. Ski magazine

#### Situation

The publisher of a ski magazine in France is interested to know something about the average annual income of the people who purchase their magazine. Over a period of

three weeks they take a sample and from a return of 758 subscribers, they determine that the average income is €39,845 and the standard deviation of this sample is €8,542.

#### Required

1. Determine the 90% confidence intervals of the mean income of all the magazine readers of this ski magazine?
2. Determine the 99% confidence intervals of the mean income of all the magazine readers of this ski magazine?
3. How would you explain the difference between the answers to Questions 1 and 2?

### 4. Households

#### Situation

A random sample of 121 households indicated they spent on average £12 on take-away restaurant foods. The standard deviation of this sample was £3.

#### Required

1. Calculate a 90% confidence interval for the average amount spent by all households in the population.
2. Calculate a 95% confidence interval for the average amount spent by all households in the population.
3. Calculate a 98% confidence interval for the average amount spent by all households in the population.
4. Explain the differences between the answers to Questions 1–3.

### 5. Taxes

#### Situation

To estimate the total annual revenues to be collected for the State of California in a certain year, the Tax Commissioner took a random sample of 15 tax returns. The taxes paid in \$US according to these returns were as follows:

\$34,000	\$2,000	\$12,000	\$39,000	\$16,000
\$7,000	\$9,000	\$72,000	\$23,000	\$15,000
\$0	\$19,000	\$6,000	\$12,000	\$43,000

#### Required

1. Determine the 80%, 95%, and 99% confidence intervals for the mean tax returns.
2. Using for example the 95% confidence interval, how would you present your analysis to your superior?

- How do you explain the differences in these intervals and what does it say about confidence in decision-making?

## 6. Vines

### Situation

In the Beaujolais wine region north of Lyon, France, a farmer is interested to estimate the yield from his 5,200 grape vines. He samples at random 75 of the grape vines and finds that there is a mean of 15 grape bunches per vine, with a sample standard deviation of 6.

### Required

- Construct a 95% confidence limit for the bunch of grapes for the total of 5,200 grape vines.
- How would you express the values determined in the previous question?
- Would your answer change if you used a Student-*t* distribution rather than a normal distribution?

## 7. Floor tiles

### Situation

A hardware store purchases a truckload of white ceramic floor tiles from a supplier knowing that many of the tiles are imperfect. Imperfect means that the colour may not be uniform, there may be surface hairline cracks, or there may be air pockets on the surface finish. The store will sell these at a marked-down price and it knows from past experience that it will have no problem selling these tiles as customers purchase these for tiling a basement or garage where slight imperfections are not critical. A store employee takes a random sample of 25 tiles from the storage area and counts the number of imperfections. This information is given in the table below.

7	4	1	2	3
4	3	3	2	8
5	5	2	3	1
3	1	6	7	5
8	2	3	4	8

### Required

- To the nearest whole number, what is an estimate of the mean number of imperfections on the lot of white tiles? This would be a point estimate.
- What is an estimate of the standard error of the number of imperfections on the tiles?
- Determine a 90% confidence interval for the mean amount of imperfections on the floor tiles. This would mean that you would be 90% confident that the mean amount of imperfections lies within this range.

4. Determine a 99% confidence interval for the mean amount of imperfections on the floor tiles. This would mean that you would be 99% confident that the mean amount of imperfections lies within this range.
5. What is your explanation of the difference between the limits obtained in Questions 3 and 4?

## 8. World's largest companies

### Situation

Every year Fortune magazine publishes information on the world's 500 largest companies. This information includes revenues, profits, assets, stock holders equity, number of employees, and the headquarters of the firm. The following table gives a random sample of the revenues of 35 of those 500 firms for 2006, generated using the random function in Excel.<sup>2</sup>

Company	Revenues (\$millions)	Country
Royal Mail Holdings	16,153.7	United Kingdom
Rabobank	36,486.5	Netherlands
Swiss Reinsurance	32,117.6	Switzerland
DuPont	28,982.0	United States
Liberty Mutual Insurance	25,520.0	United States
Coca-Cola	24,088.0	United States
Westpac Banking	16,170.5	Australia
Northwestern Mutual	20,726.2	United States
Lloyds TSB Group	53,904.0	United Kingdom
UBS	107,934.8	Switzerland
Sony	70,924.8	Japan
Repsol YPF	60,920.9	Spain
United Technologies	47,829.0	United States
San Paolo IMI	22,793.3	Italy
Vattenfall	19,768.6	Sweden
Bank of America	117,017.0	United States
Kimberly-Clark	16,746.9	United States
State Grid	107,185.5	China
SK Networks	16,733.9	South Korea
Archer Daniels Midland	36,596.1	United States
Bridgestone	25,709.7	Japan
Matsushita Electric Industrial	77,871.1	Japan
Johnson and Johnson	53,324.0	United States
Magna International	24,180.0	Canada
Migros	16,466.4	Switzerland
Bouygues	33,693.7	France
Hitachi	87,615.4	Japan

<sup>2</sup> The World's Largest Corporations, *Fortune*, Europe Edition, 156(2), 23 July 2007, p. 84.

Company	Revenues (\$millions)	Country
Mediceo Paltac Holdings	18,524.9	Japan
Edeka Zentrale	20,733.1	Germany
Unicredit Group	59,119.3	Italy
Otto Group	19,397.5	Germany
Cardinal Health	81,895.1	United States
BAE Systems	22,690.9	United Kingdom
TNT	17,360.6	Netherlands
Tyson Foods	25,559.0	United States

### Required

1. Using the complete sample data, what is an estimate for the average value of revenues for the world's 500 largest companies?
2. Using the complete sample data, what is an estimate for the standard error?
3. Using the complete sample data, determine a 95% confidence interval for the mean value of revenues for the world's 500 largest companies. This would mean that you would be 95% confident that the average revenues lie within this range.
4. Using the complete sample data, determine a 99% confidence interval for the mean value of revenues for the world's 500 largest companies. This would mean that you would be 95% confident that the average revenue lies within this range.
5. Explain the difference between the answers obtained in Questions 3 and 4.
6. Using the first 15 pieces of data, give an estimate for the average value of revenues for the world's 500 largest companies?
7. Using the first 15 pieces of data, what is an estimate for the standard error?
8. Using the first 15 pieces of data, determine a 95% confidence interval for the mean value of revenues for the world's 500 largest companies. This would mean that you would be 95% confident that the average revenue lies within this range.
9. Using the first 15 pieces of data, determine a 99% confidence interval for the mean value of revenues for the world's 500 largest companies. This would mean that you would be 95% confident that the average revenue lies within this range.
10. Explain the difference between in the answers obtained in Questions 8 and 9.
11. Explain the differences between the results in Questions 1 through 4 and those in Questions 6 through 9 and justify how you have arrived at your results.

## 9. Hotel accounts

### Situation

A 125-room hotel noted that in the morning when clients check out there are often questions and complaints about the amount of the bill. These complaints included overcharging on items taken from the refrigerator in the room, wrong billing of restaurant meals consumed, and incorrect accounts of laundry items. On a particular day the hotel

is full and the night manager analyses a random sample of 19 accounts and finds that there is an average of 2.8 errors on these sample accounts. Based on passed analysis the night manager believes that the population standard deviation is 0.7.

#### Required

1. From this sample experiment, what is the correct value of the standard error?
2. What are the confidence intervals for a 90% confidence level?
3. What are the confidence intervals for a 95% confidence level?
4. What are the confidence intervals for a 99% confidence level?
5. Explain the differences between Questions 2, 3, and 4?

## 10. Automobile tyres

#### Situation

An automobile repair company has an inventory of 2,500 different sizes, and different makes of tyres. It wishes to estimate the value of this inventory and so it takes a random sample of 30 tyres and records their cost price. This sample information in Euros is given in the table below.

44	34	66	48	42	36
88	76	68	34	89	73
69	55	72	88	60	74
80	75	57	36	95	50
61	41	32	62	91	65

#### Required

1. What is an estimation of the cost price of the total amount of tyres in inventory?
2. Determine a 95% confidence interval for the cost price of the automobile tyres in inventory.
3. How would you express the answers to Questions 1 and 2 to management?
4. Determine a 99% confidence interval for the cost price of the automobile tyres in inventory.
5. Explain the differences between Questions 2 and 4?
6. How would you suggest a random sample of tyres should be taken from inventory? What other comments do you have?

## 11. Stuffed animals

#### Situation

A toy store in New York estimates that it has 270 stuffed animals in its store at the end of the week. An assistant takes a random sample of 19 of these stuffed animals and determines that the average retail price of these animals is \$13.75 with a standard deviation of \$0.53.

### Required

1. What is the correct value of the standard error of the sample?
2. What is an estimate of the total value of the stuffed animals in the store?
3. Give a 95% confidence limit of the total retail value of all the stuffed animals in inventory.
4. Give a 99% confidence limit of the total retail value of all the stuffed animals in inventory.
5. Explain the difference between Questions 3 and 4.

## 12. Shampoo bottles

### Situation

A production operation produces plastic shampoo bottles for Procter and Gamble. At the end of the production operation the bottles pass through an optical quality control detector. Any bottle that the detector finds defective is automatically ejected from the line. In 1,500 bottles that passed the optical detector, 17 were ejected.

### Required

1. What is a point estimate of the proportion of shampoo bottles that are defective in the production operation?
2. Obtain 90% confidence intervals for the proportion of defective bottles produced in production.
3. Obtain 98% confidence intervals for the proportion of defective bottles produced in production.
4. If an estimate of the proportion of defectives to within a margin of error of  $\pm 0.005$  of the population proportion at 90% confidence were required, and you wanted to be conservative in your analysis, how many bottles should pass through the optical detector? No information is available from past data.
5. If an estimate of the proportion of defectives to within a margin of error of  $\pm 0.005$  of the population proportion at 98% confidence were required, and you wanted to be conservative in your analysis, how many bottles should pass through the optical detector? No information is available from past data.
6. What are your comments about the answer obtained in Question 4 and 5 and in general terms for this sampling process.

## 13. Night shift

### Situation

The management of a large factory, where there are 10,000 employees, is considering the introduction of a night shift. The human resource department took a random sample of 800 employees and found that there were 240 who were not in favour of a night shift.



### Required

1. What is the proportion of employees who are in favour of a night shift?
2. What are the 95% confidence limits for the population who are not in favour?
3. What are the 95% confidence limits for the proportion who are in favour of a night shift?
4. What are the 98% confidence limits for the population who are not in favour?
5. What are the 98% confidence limits for the proportion who are in favour of a night shift?
6. What is your explanation of the difference between Questions 3 and 5?

## 14. Ski trip

### Situation

The Student Bureau of a certain business school plans to organize a ski trip in the French Alps. There are 5,000 students in the school. The bureau selects a random sample of 40 students and of these 24 say they will be coming skiing.

### Required

1. What is an estimate of the proportion of students who say they will not be coming skiing?
2. Obtain 90% confidence intervals for the proportion of students who will be coming skiing.
3. Obtain 98% confidence intervals for the proportion of students who will be coming skiing.
4. How would you explain the difference between the answers to Questions 2 and 3?
5. What would be the conservative value of the sample size in order that the Student Bureau can estimate the true proportion of those coming skiing within plus or minus 0.02 at a confidence level of 90%? No other sample information has been taken.
6. What would be the conservative value of the sample size in order that the Student Bureau can estimate the true proportion of those coming skiing within plus or minus 0.02 at a confidence level of 98%? No other sample information has been taken.

## 15. Hilton hotels

### Situation

Hilton hotels, based in Watford, England, agreed in December 2005 to sell the international Hilton properties for £3.3 billion to United States-based Hilton group. This transaction will create a worldwide empire of 2,800 hotels stretching from the Waldorf-Astoria in New York to the Phuket Arcadia Resort in Thailand.<sup>3</sup> The objective of this new

<sup>3</sup> Timmons, H., "Hilton sets the stage for global expansion", *International Herald Tribune*, 30 December 2005, p. 1.

chain is to have an average occupancy, or a yield rate, of all the hotels at least 90%. In order to test whether the objectives are able to be met, a member of the finance department takes a random sample of 49 hotels worldwide and finds that in a 3-month test period, 32 of these had an occupancy rate of at least 90%.

### Required

1. What is an estimate of the proportion or percentage of the population of hotels that meet the objectives of the chain?
2. What is a 90% confidence interval for the proportion of hotels who meet the objectives of the chain?
3. What is a 98% confidence interval for the proportion of hotels who meet the objectives of the chain?
4. How would you explain the difference between the answers to Questions 2 and 3?
5. What would be the conservative value of the sample size that should be taken in order that the hotel chain can estimate the true proportion of those meeting the objectives is within plus or minus 10% of the true proportion at a confidence level of 90%? No earlier sample information is available.
6. What would be the conservative value of the sample size that should be taken in order that the hotel chain can estimate the true proportion of those meeting the objectives is within plus or minus 10% of the true proportion at a confidence level of 98%? No earlier sample information is available.
7. What are your comments about this sample experiment that might explain inconsistencies?

## 16. Case: Oak manufacturing

### Situation

Oak manufacturing company produces kitchen appliances, which it sells on the European market. One of its new products, for which it has not yet decided to go into full commercialization, is a new computerized food processor. The company made a test market, during the first 3 months that this product was on sale. Six stores were chosen for this study in the European cities of Milan, Italy; Hamburg, Germany; Limoges, France; Birmingham, United Kingdom; Bergen, Norway; and Barcelona, Spain. The weekly test market sales for these outlets are given in the table below. Oak had developed this survey, because their Accounting Department had indicated that at least 130,000 of this food processor need to be sold in the first year of commercialization to break-even. They reasonably assumed that daily sales were independent from country to country, store to store, and from day to day. Management wanted to use a confidence level of 90% in its analysis. For the first year of commercialization after the “go” decision, the food processor is to be sold in a total of 100 stores in the six countries where the test market had been carried out.

Milan, Italy	Hamburg, Germany	Limoges, France	Birmingham, United Kingdom	Bergen, Norway	Barcelona, Spain
3	29	15	34	25	21
8	29	16	22	19	0
20	13	32	31	25	5
8	22	31	28	35	14
17	23	32	23	25	16
11	20	15	20	20	9
12	29	16	26	34	13
3	17	46	39	29	11
6	22	27	24	24	3
13	26	20	35	33	16
12	19	28	37	36	4
13	21	2	20	39	1
15	47	28	27	38	15
0	31	29	30	12	18
15	33	36	34	33	6
5	42	33	25	26	18
2	32	18	21	35	14
17	13	33	26	30	21
19	19	28	16	28	14
18	23	27	31	34	20
17	20	34	23	20	19
12	20	16	25	29	9
17	17	30	12	20	12
6	34	32	22	36	1

### Required

Based on this information what would be your recommendations to the management of Oak manufacturing?

# Hypothesis testing of a single population

## You need to be objective

The government in a certain country says that radiation levels in the area surrounding a nuclear power plant are well below levels considered harmful. Three people in the area died of leukaemia. The local people immediately put the blame on the radioactive fallout.

*Does the death of three people make us assume that the government is wrong with its information and that we make the assumption, or hypothesis, that radiation levels in the area are abnormally high? Alternatively, do we accept that the deaths from leukaemia are random and are not related to the nuclear power facility? You should not accept, or reject, a hypothesis about a population parameter – in this case the radiation levels in the surrounding area of the nuclear power plant, simply by intuition. You need to be objective in decision-making. For this situation an appropriate action would be to take representative samples of the incidence of leukaemia cases over a reasonable time period and use these to test the hypothesis. This is the purpose of this chapter (and the following chapter) to find out how to use hypothesis testing to determine whether a claim is valid. There are many instances when published claims are not backed up by solid statistical evidence.*

## Learning objectives

After you have studied this chapter you will understand the concept of **hypothesis testing**, how to test for the **mean** and **proportion** and be aware of the **risks** in testing. The topics of these themes are as follows:

- ✓ Concept of hypothesis testing • Significance level • Null and alternative hypothesis
- ✓ Hypothesis testing for the mean value • A two-tail test • One-tail, right-hand test • One-tail, left-hand test • Acceptance or rejection • Test statistics • Application when the standard deviation of the population is known: *Filling machine* • Application when the standard deviation of the population is unknown: *Taxes*
- ✓ Hypothesis testing for proportions • Testing for proportions from large samples • Application of hypothesis testing for proportions: *Seaworthiness of ships*
- ✓ The probability value in testing hypothesis •  $p$ -value of testing hypothesis • Application of the  $p$ -value approach: *Filling machine* • Application of the  $p$ -value approach: *Taxes* Application of the  $p$ -value approach: *Seaworthiness of ships* • Interpretation of the  $p$ -value
- ✓ Risks in hypothesis testing • Errors in hypothesis testing • Cost of making an error • Power of a test

### Concept of Hypothesis Testing

A hypothesis is a judgment about a situation, outcome, or population parameter based simply on an assumption or intuition with no concrete backup information or analysis. **Hypothesis testing** is to take sample data and make an objective decision based on the results of the test within an appropriate significance level. Thus like estimating, hypothesis testing is an extension of the use of sampling presented in Chapter 6.

### Significance level

When we make quantitative judgments, or hypotheses, about situations, we are either right, or wrong. However, if we are wrong we may not be far from the real figure or that is our judgment is not significantly different. Thus our hypothesis may be acceptable. Consider the following:

- A contractor says that it will take 9 months to construct a house for a client. The house is

finished in 9 months and 1 week. The completion time is not 9 months however it is not significantly different from the estimated time construction period of 9 months.

- The local authorities estimate that there are 20,000 people at an open air rock concert. Ticket receipts indicate there are 42,000 attendees. This number of 42,000 is significantly different from 20,000.
- A financial advisor estimates that a client will make \$15,000 on a certain investment. The client makes \$14,900. The number \$14,900 is not \$15,000 but it is not significantly different from \$15,000 and the client really does not have a strong reason to complain. However, if the client made only \$8,500 he would probably say that this is significantly different from the estimated \$15,000 and has a justified reason to say that he was given bad advice.

Thus in hypothesis testing, we need to decide what we consider is the **significance level** or the level of importance in our evaluation. This significance level is giving a ceiling level usually in terms of

percentages such as 1%, 5%, 10%, etc. To a certain extent this is the subjective part of hypothesis testing since one person might have a different criterion than another individual on what is considered significant. However in accepting, or rejecting a hypothesis in decision-making, we have to agree on the level of significance. This significance value, which is denoted as alpha,  $\alpha$ , then gives us the **critical value** for testing.

## Null and alternative hypothesis

In hypothesis testing there are two defining statements premised on the binomial concept. One is the **null hypothesis**, which is that value considered correct within the given level of significance. The other is the **alternative hypothesis**, which is that the hypothesized value is not correct at the given level of significance. The alternative hypothesis as a value is also known as the **research hypothesis** since it is a value that has been obtained from a sampling experiment.

For example, the hypothesis is that the average age of the population in a certain country is 35. This value is the null hypothesis. The alternative to the null hypothesis is that the average age of the population is not 35 but is some other value. In hypothesis testing there are three possibilities. The first is that there is evidence that the value is significantly different from the hypothesized value. The second is that there is evidence that the value is significantly greater than the hypothesized value. The third is that there is evidence that the value is significantly less than the hypothesized value. Note, that in these sentences we say there is evidence because as always in statistics there is no guarantee of the result but we are basing our analysis of the population based only on sampling and of course our sample experiment may not yield the correct result. These three possibilities lead to using a **two-tail hypothesis test**, a **right-tail hypothesis test**, and a **left-tail hypothesis test** as explained in the next section.

## Hypothesis Testing for the Mean Value

In hypothesis testing for the mean, an assumption is made about the mean or average value of the population. Then we take a sample from this population, determine the sample mean value, and measure the difference between this sample mean and the hypothesized population value. If the difference between the sample mean and the hypothesized population mean is small, then the higher is the probability that our hypothesized population mean value is correct. If the difference is large then the smaller is the probability that our hypothesized value is correct.

## A two-tail test

A two-tail test is used when we are testing to see if a value is **significantly different** from our hypothesized value. For example in the above population situation, the null hypothesis is that the average age of the population is 35 years and this is written as follows:

$$\text{Null hypothesis: } H_0: \mu_x = 35 \quad 8(i)$$

In the two-tail test we are asking, is there evidence of a difference. In this case the alternative to the null hypothesis is that the average age is not 35 years. This is written as,

$$\text{Alternative hypothesis: } H_1: \mu_x \neq 35 \quad 8(ii)$$

When we ask the question is there evidence of a difference, this means that the alternative value can be significantly lower or higher than the hypothesized value. For example, if we took a sample from our population and the average age of the sample was 36.2 years we might say that the average age of the population is not significantly different from 35. In this case we would accept the null hypothesis as being correct. However, if in our sample the average age was

52.7 years then we may conclude that the average age of the population is significantly different from 35 years since it is much higher. Alternatively, if in our sample the average age was 21.2 years then we may also conclude that the average age of the population is significantly different from 35 years since it is much lower. In both of these cases we would reject the null hypothesis and accept the alternative hypothesis. Since this is a binomial concept, when we reject the null hypothesis we are accepting the alternative hypothesis. Conceptually the two-tailed test is illustrated in Figure 8.1. Here we say that there is a 10% level of significance and in this case for a two-tail test there is 5% in each tail.

### One-tail, right-hand test

A one-tail, right-hand test is used to test if there is evidence that the value is **significantly greater**

than our hypothesized value. For example in the above population situation, the null hypothesis is that the average age of the population is equal to or less than 35 years and this is written as follows:

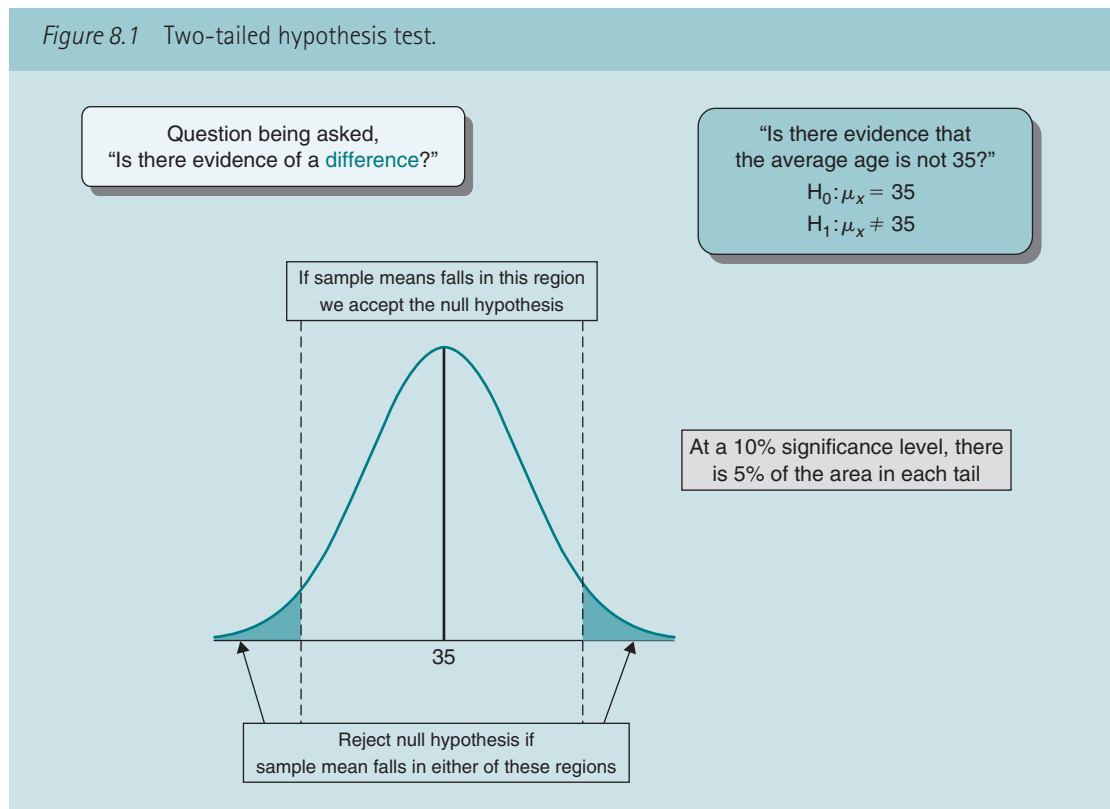
$$\text{Null hypothesis: } H_0: \mu_x \leq 35 \quad 8(\text{iii})$$

The alternative hypothesis is that the average age is greater than 35 years and this is written as,

$$\text{Alternative hypothesis: } H_1: \mu_x > 35 \quad 8(\text{iv})$$

Thus, if we took a sample from our population and the average age of the sample was say 36.2 years we would probably say that the average age of the population is not significantly greater than 35 years and we would accept the null hypothesis. Alternatively, if in our sample the average age was 21.2 years then although this is significantly less than 35, it is not greater than 35. Again we would accept the null hypothesis. However, if in

Figure 8.1 Two-tailed hypothesis test.



our sample the average age was 52.7 years then we may conclude that the average age of the population is significantly greater than 35 years and we would reject the null hypothesis and accept the alternative hypothesis. Note that for this situation we are not concerned with values that are significantly less than the hypothesized value but only those that are significantly greater. Again, since this is a binomial concept, when we reject the null hypothesis we accept the alternative hypothesis. Conceptually the one-tail, right-hand test is illustrated in Figure 8.2. Again we say that there is a 10% level of significance, but in this case for a one-tail test, all the 10% area is in the right-hand tail.

### One-tail, left-hand test

A one-tail, left-hand test is used to test if there is evidence that the value is **significantly less** than our hypothesized value. For example again let

us consider the above population situation. The null hypothesis,  $H_0: \mu_x$ , is that the average age of the population is equal to or more than 35 years and this is written as follows:

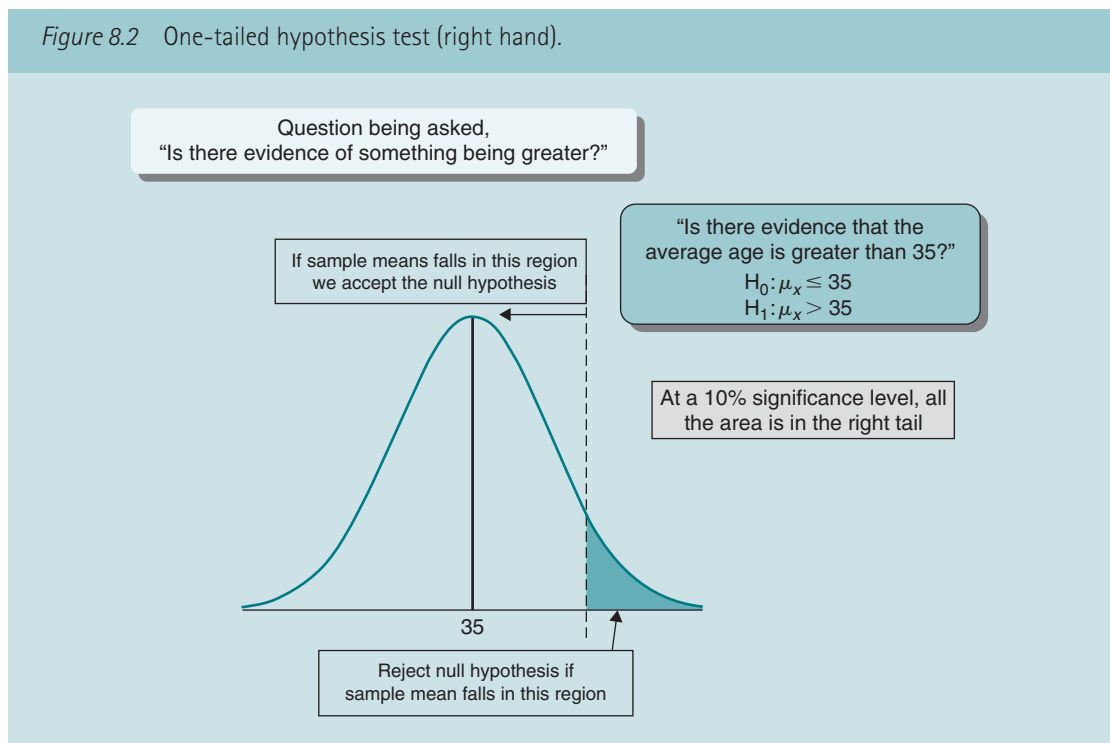
$$H_0: \mu_x \geq 35 \quad 8(v)$$

The alternative hypothesis,  $H_1: \mu_x$ , is that the average age is less than 35 years. This is written,

$$H_1: \mu_x < 35 \quad 8(vi)$$

Thus, if we took a sample from our population and the average age of the sample was say 36.2 years we would say that there is no evidence that the average age of the population is significantly less than 35 years and we would accept the null hypothesis. Or, if in our sample the average age was 52.7 years then although this is significantly greater than 35 it is not less than 35 and we would accept the null hypothesis. However, if in our sample the average age was 21.2 years then we may conclude that the average age of the

Figure 8.2 One-tailed hypothesis test (right hand).





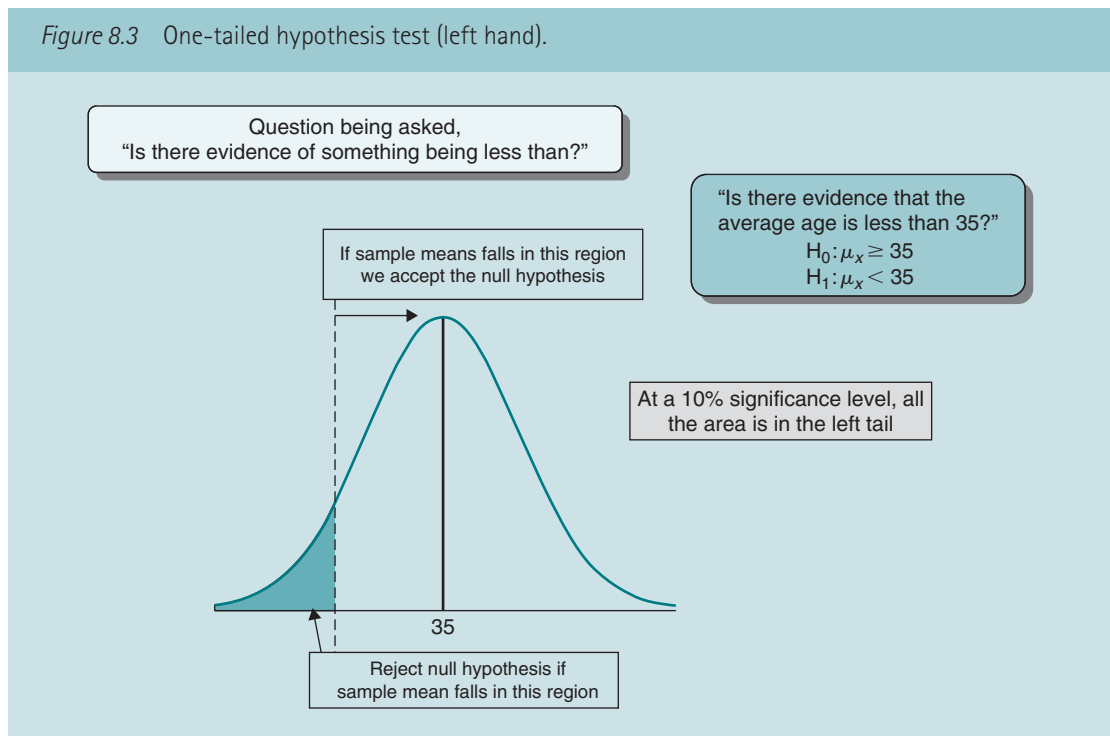
population is significantly less than 35 years and we would reject the null hypothesis and accept the alternative hypothesis. Note that for this situation we are not concerned with values that are significantly greater than the hypothesized value but only those that are significantly less than the hypothesized value. Again, since this is a binomial concept, when we reject the null hypothesis we accept the alternative hypothesis. Conceptually the one-tail, left-hand test is illustrated in Figure 8.3. With the 10% level of significance shown means that for this one-tail test all the 10% area is in the left-hand tail.

### Acceptance or rejection

The purpose of hypothesis testing is not to question the calculated value of the sample statistic, but to make an objective judgment regarding the difference between the sample mean and

the hypothesized population mean. If we test at the 10% significance level this means that the null hypothesis would be rejected if the difference between the sample mean and the hypothesized population mean is so large that it, on average, 10 or fewer times in every 100 samples when the hypothesized population parameter is correct. Assuming the hypothesis is correct, then the significance level indicates the percentage of sample means that are outside certain limits. Even if a sample statistic does fall in the area of acceptance, this does not prove that the null hypothesis  $H_0$  is true but there simply is no statistical evidence to reject the null hypothesis. Acceptance or rejection is related to the values of the test statistic that are unlikely to occur if the null hypothesis is true. However, they are not so unlikely to occur if the null hypothesis is false.

Figure 8.3 One-tailed hypothesis test (left hand).



## Test statistics

We have two possible relationships to use that are analogous to those used in Chapter 7. If the **population standard deviation is known**, then using the central limit theorem for sampling, the test statistic, or the critical value is,

$$\text{test statistics, } z = \frac{\bar{x} - \mu_{H_0}}{\sigma_x / \sqrt{n}} \quad 8(\text{vii})$$

Where,

- $\mu_{H_0}$  is the hypothesized population mean.
- $\bar{x}$  is the sample mean.
- The numerator,  $\bar{x} - \mu_x$ , measures how far, the observed mean is from the hypothesized mean.
- $\sigma_x$  is the population standard deviation.
- $n$  is the sample size.
- $\sigma_x / \sqrt{n}$ , the denominator in the equation, is the standard error.
- $z$ , is how many standard errors, the observed sample mean is from the hypothesized mean.

If the population standard deviation is unknown then the only standard deviation we can determine is the sample standard deviation,  $s$ . This value of  $s$  can be considered an estimate of the population standard deviation sometimes written as  $\hat{\sigma}_x$ . If the sample size is less than 30 then we use the Student- $t$  distribution, presented in Chapter 7, with  $(n - 1)$  degrees of freedom making the assumption that the population from which this sample is drawn is normally distributed. In this case, the test statistic can be calculated by,

$$t = \frac{\bar{x} - \mu_{H_0}}{\hat{\sigma}_x / \sqrt{n}} \quad 8(\text{viii})$$

Where,

- $\mu_{H_0}$  is again the hypothesized population mean.
- $\bar{x}$  is the sample mean.

- The numerator,  $\bar{x} - \mu_{H_0}$ , measures how far, the observed mean is from the hypothesized mean.
- $\hat{\sigma}_x$  is the estimate of the population standard deviation and is equal to the sample standard deviation,  $s$ .
- $n$  is the sample size.
- $\hat{\sigma}_x / \sqrt{n}$ , the denominator in the equation, is the estimated standard error.
- $t$ , is how many standard errors, the observed sample mean is from the hypothesized mean.

The following applications illustrate the procedures for hypothesis testing.

### Application when the standard deviation of the population is known: *Filling machine*

A filling line of a brewery is for 0.50 litre cans where it is known that the standard deviation of the filling machine process is 0.05 litre. The quality control inspector performs an analysis on the line to test whether the process is operating according to specifications. If the volume of liquid in the cans is higher than the specification limits then this costs the firm too much money. If the volume is lower than the specifications then this can cause a problem with the external inspectors. A sample of 25 cans is taken and the average of the sample volume is 0.5189 litre.

1. At a significance level,  $\alpha$ , of 5% is there evidence that the volume of beer in the cans from this bottling line is different than the target volume of 0.50 litre?

Here we are asking the question if there is there evidence of a difference so this means it is a two-tail test. The null and alternative hypotheses are written as follows:

Null hypothesis:  $H_0: \mu_x = 0.50 \text{ litre.}$

Alternative hypothesis:  $H_1: \mu_x \neq 0.50 \text{ litre.}$

And, since we know the population standard deviation we can use equation 8(vii) where,

- $\mu_{H_0}$  is the hypothesized population mean, or 0.50 litre.
- $\bar{x}$  is the sample mean, or 0.5189 litre.
- The numerator,  $\bar{x} - \mu_{H_0}$ , is  $0.5189 - 0.5000 = 0.0189$  litre.
- $\sigma_x$  is the population standard deviation, or 0.05 litre.
- $n$  is the sample size, or 25.
- $\sqrt{n} = 5$ .

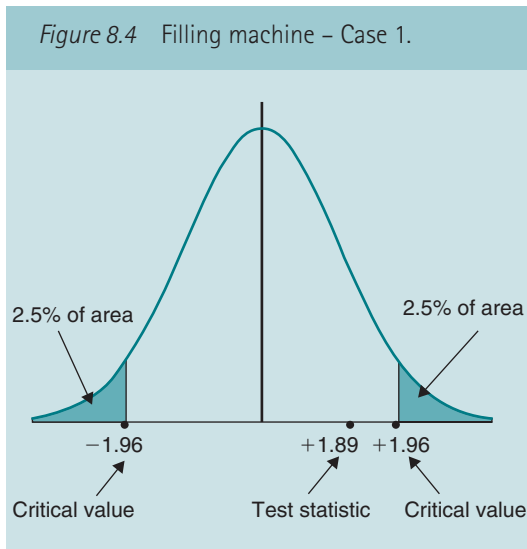
Thus, the standard error of the sample is  $0.05/5 = 0.01$ .

The test statistic from equation 8(vii) is,

$$z = \frac{\bar{x} - \mu_{H_0}}{\sigma_x / \sqrt{n}} = \frac{0.0189}{0.01} = 1.8900$$

At a significance level of 5% for the test of a difference there is 2.5% in each tail. Using [function NORMSINV] in Excel this gives a critical value of  $z$  of  $\pm 1.96$ .

Since the value of the test statistic or 1.89 is less than the critical value of 1.96, or alternatively within the boundaries of  $\pm 1.96$  then there is no statistical evidence that the volume of beer in the cans is significantly different than 0.50 litre. Thus we would accept the null hypothesis. These relationships are shown in Figure 8.4.



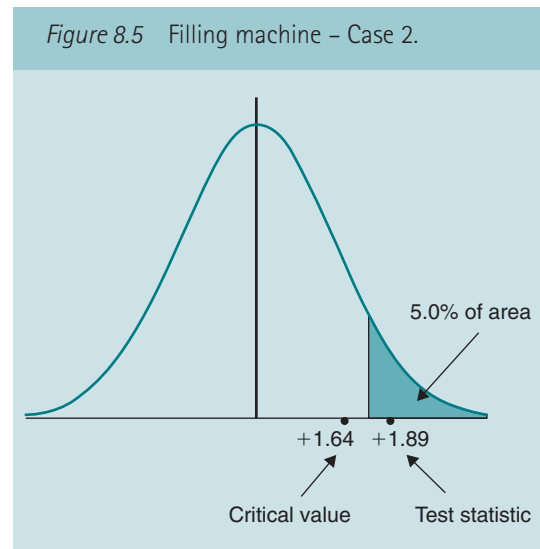
2. At a significance level,  $\alpha$ , of 5% is there evidence that the volume of beer in the cans from this bottling line is greater than the target volume of 0.50 litre? Here we are asking the question if there is evidence of the value being greater than the target value and so this is a one-tail, right-hand test. The null and alternative hypotheses are as follows:

Null hypothesis:  $H_0: \mu_x \leq 0.50$  litre.

Alternative hypothesis:  $H_1: \mu_x > 0.50$  litre.

Nothing has changed regarding the test statistic and it remains 1.8900 as calculated in Question 1. However for a one-tail test, at a significance level of 5% for the test there is 5% in the right tail. The area of the curve for the upper level is  $100\% - 5.0\%$  or  $95.00\%$ . Using [function NORMSINV] in Excel this gives a critical value of  $z$  of +1.64.

Since now the value of the test statistic or 1.89 is greater than the critical value of 1.64 then there is evidence that the volume of beer in all of the cans is significantly greater than 0.50 litre. Conceptually this situation is shown on the normal distribution curve in Figure 8.5.



## Application when the standard deviation of the population is unknown: Taxes

A certain state in the United States has made its budget on the bases that the average individual average tax payments for the year will be \$30,000. The financial controller takes a random sample of annual tax returns and these amounts in United States dollars are as follows.

34,000	12,000	16,000	10,000
2,000	39,000	7,000	72,000
24,000	15,000	19,000	12,000
23,000	14,000	6,000	43,000

1. At a significance level,  $\alpha$ , of 5% is there evidence that the average tax returns of the state will be different than the budget level of \$30,000 in this year?

The null and alternative hypotheses are as follows:

Null hypothesis:  $H_0: \mu_x = \$30,000$ .

Alternative hypothesis:  $H_1: \mu_x \neq \$30,000$ .

Since we have no information of the population standard deviation, and the sample size is less than 30, we use a Student- $t$  distribution.

Sample size,  $n$ , is 16.

Degrees of freedom,  $(n - 1)$  are 15.

Using [function **TINV**] from Excel the Student- $t$  value is  $\pm 2.1315$  and these are the critical values. Note, that since this is a two-tail test there is 2.5% of the area in each of the tails and  $t$  has a plus or minus value.

From Excel, using [function **AVERAGE**].

Mean value of this sample data,  $\bar{x}$ , is \$21,750.00.

$$\begin{aligned}\bar{x} - \mu_x &= 21,750.00 - 30,000.00 \\ &= -\$8,250.00\end{aligned}$$

From [function **STDEV**] in Excel, the sample standard deviation,  $s$ , is \$17,815.72 and

this can be taken as an estimate of the population standard deviation,  $\hat{\sigma}_x$ .

Estimate of the standard error is,

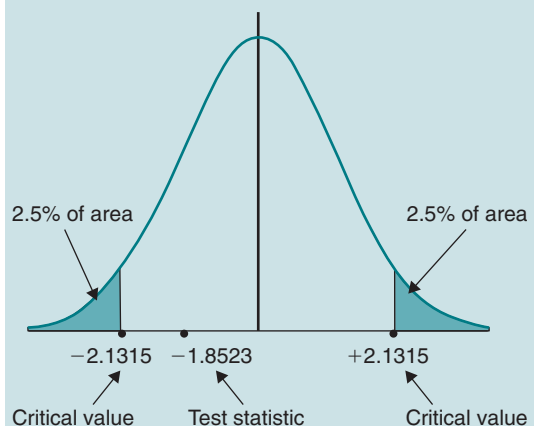
$$\frac{\hat{\sigma}_x}{\sqrt{n}} = \frac{17,815.72}{\sqrt{16}} = \$4,453.93$$

From equation 8(vii) the sample statistic is,

$$t = \frac{\bar{x} - \mu_{H_0}}{\hat{\sigma}_x / \sqrt{n}} = -\frac{8,500}{4,453.93} = -1.8523$$

Since the sample statistic,  $-1.8523$ , is not less than the test statistic of  $-2.1315$ , there is no reason to reject the null hypothesis and so we accept that there is no evidence that the average of all the tax receipts will be significantly different from \$30,000. Note in this situation, as the test statistic is negative we are on the left side of the curve and so we only make an evaluation with the negative values of  $t$ . Another way of making the analysis, when we are looking to see if there is a difference, is to see whether the sample statistic of  $-1.8523$  lies within the critical boundary values of  $t = \pm 2.1315$ . In this case it does. The concept is shown in Figure 8.6.

Figure 8.6 Taxes – Case 1.



2. At a significance level,  $\alpha$ , of 5% is there evidence that the tax returns of the state will be less than the budget level of \$30,000 in this year?

This is a left-hand, one-tail test and the null and alternative hypothesis are as follows:

Null hypothesis:  $H_0: \mu_x \geq \$30,000$ .

Alternative hypothesis:  $H_1: \mu_x < \$30,000$ .

Again, since we have no information of the population standard deviation, and the sample size is less than 30, we use a Student- $t$  distribution.

Sample size,  $n$ , is 16.

Degrees of freedom,  $(n - 1)$  is 15.

Here we have a one-tail test and thus all of the value of  $\alpha$ , or 5%, lies in one tail. However, the Excel function for the Student- $t$  value is based on input for a two-tail test so in order to determine  $t$  we have to enter the area value of 10% (5% in one tail and 5% in the other tail.)

Using [function TINV] gives a critical value of  $t = -1.7531$ .

The value of the sample statistic  $t$  remains unchanged at  $-1.8532$  as calculated in Question 1.

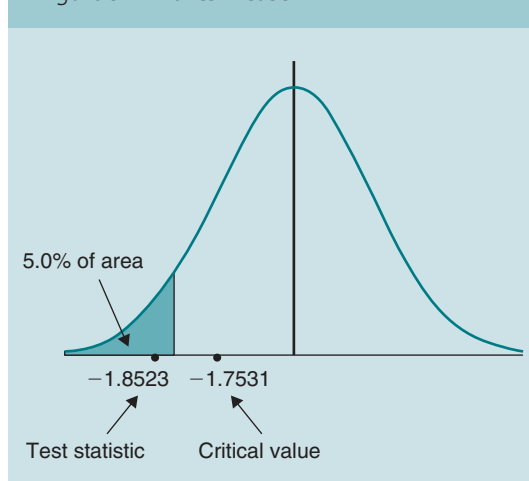
Since now the sample statistic,  $-1.8523$  is less than the test statistic,  $-1.8523$ , then

there is reason to reject the null hypothesis and to accept the alternative hypothesis that there is evidence that the average value of all the tax receipts is significantly less than \$30,000. Note that in this situation we are on the left side of the curve and so we are only interested in the negative value of  $t$ . This situation is conceptually shown on the Student- $t$  distribution curve of Figure 8.7.

## Hypothesis Testing for Proportions

In hypothesis testing for the proportion we test the assumption about the value of the population proportion. In the same way for the mean, we take a sample from this population, determine the sample proportion, and measure the difference between this proportion and the hypothesized population value. If the difference between the sample proportion and the hypothesized population proportion is small, then the higher is the probability that our hypothesized population proportion value is correct. If the difference is large then the probability that our hypothesized value is correct is low.

Figure 8.7 Taxes – Case 2.



## Hypothesis testing for proportions from large samples

In Chapter 6, we developed the relationship from the binomial distribution between the population proportion,  $p$ , and the sample proportion  $\bar{p}$ . On the assumption that we can use the normal distribution as our test reference then from equation 6(xii) we have the value of  $z$  as follows:

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{\bar{p} - p}{\sqrt{p(1-p)/n}} \quad 6(\text{xii})$$

In hypothesis testing for proportions we use an analogy as for the mean where  $p$  is now the

hypothesized value of the proportion and may be written as  $p_{H_0}$ . Thus, equation 6(xii) becomes,

$$z = \frac{\bar{p} - p_{H_0}}{\sigma_{\bar{p}}} = \frac{\bar{p} - p_{H_0}}{\sqrt{p_{H_0}(1 - p_{H_0})/n}} \quad 8(\text{ix})$$

The application of the hypothesis testing for proportions is illustrated below.

### Application of hypothesis testing for proportions: Seaworthiness of ships

On a worldwide basis, governments say that 0.80, or 80%, of merchant ships are seaworthy. Greenpeace, the environmental group, takes a random sample of 150 ships and the analysis indicates that from this sample, 111 ships prove to be seaworthy.

1. At a 5% significance level, is there evidence to suggest that the seaworthiness of ships is different than the hypothesized 80% value?

Since we are asking the question is there a difference then this is a two-tail test with 2.5% of the area in the left tail, and 2.5% in the right tail or 5% divided by 2.

From Excel [function **NORMSINV**] the value of  $z$ , or the critical value when the tail area is 2.5% is  $\pm 1.9600$ .

The hypothesis test is written as follows:

$H_0: p = 0.80$ . The proportion of ships that are seaworthy is equal to 0.80.

$H_1: p \neq 0.80$ . The proportion of ships that are not seaworthy is different from 0.80.

Sample size  $n$  is 150.

Sample proportion  $\bar{p}$  that is seaworthy is  $111/150 = 0.74$  or 74%.

From the sample, the number of ships that are not seaworthy is 39 ( $150 - 111$ ).

Sample proportion  $\bar{q} = (1 - \bar{p})$  that is not seaworthy is  $39/150 = 0.26$  or 26%.

The standard error of the proportion, or the denominator in equation 8(ix).

$$\begin{aligned} \sigma_{\bar{p}} &= \sqrt{\frac{p_{H_0}(1 - p_{H_0})}{n}} = \sqrt{\frac{0.80 * 0.20}{150}} \\ &= \sqrt{\frac{0.16}{150}} = 0.0327. \end{aligned}$$

$$\bar{p} - p_{H_0} = 0.74 - 0.80 = -0.06.$$

Thus the sample test statistic from equation 6(xii) is,

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = -\frac{0.06}{0.0327} = -1.8349$$

Since the test statistic of  $-1.8349$  is not less than  $-1.9600$  then we accept the null hypothesis and say that at a 5% significance level there is no evidence of a significance difference between the 80% of seaworthy ships postulated. Conceptually this situation is shown in Figure 8.8.

2. At a 5% significance level, is there evidence to suggest that the seaworthiness of ships is less than the 80% indicated?

This now becomes a one-tail, left-hand test where we are asking is there evidence that

Figure 8.8 Seaworthiness of ships – Case 1.

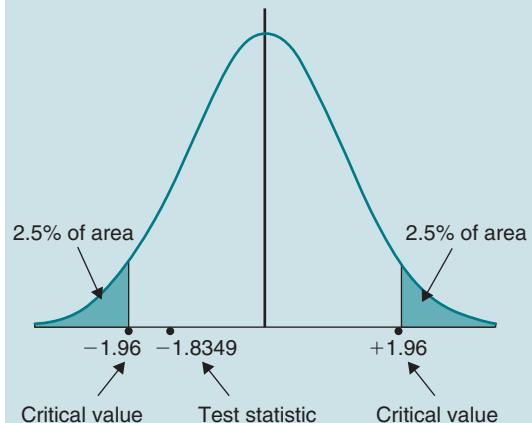
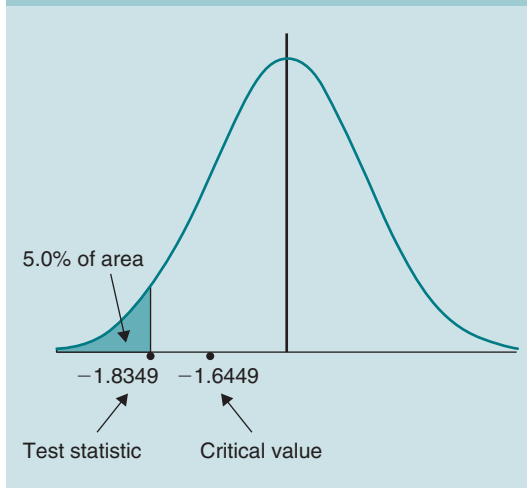


Figure 8.9 Seaworthiness of ships – Case 2.



the proportion is less than 80%. The hypothesis test is thus written as,

$H_0: p \geq 0.80$ . The proportion of ships is not less than 0.80.

$H_1: p < 0.80$ . The proportion of ships is less than 0.80.

In this situation the value of the sample statistics remains unchanged at  $-1.8349$ , but the critical value of  $z$  is different.

From Excel [function **NORMSINV**] the value of  $z$ , or the critical value when the tail area is 5% is  $z = -1.6449$ .

Now we reject the null hypothesis because the value of the test statistic,  $-1.8349$  is less than the critical value of  $-1.6449$ . Thus our conclusion is that there is evidence that the proportion of ships that are not seaworthy is significantly less than 0.80 or 80%. Conceptually this situation is shown on the distribution in Figure 8.9.

## The Probability Value in Testing Hypothesis

Up to this point our method of analysis has been to select a significance level for the hypothesis,

which then translates into a critical value of  $z$  or  $t$ , and then test to see whether the sample statistic lies within the boundaries of the critical value. If the test statistic falls within the boundaries then we accept the null hypothesis. If the test statistic falls outside, then we reject the null hypothesis and accept the alternative hypothesis. Thus we have created a binomial “yes” or “no” situation by examining whether there is sufficient statistical evidence to accept or reject the null hypothesis.

## $p$ -value of testing hypothesis

An alternative approach to hypothesis testing is to ask, what is the minimum probable level that we will tolerate in order to accept the null hypothesis of the mean or the proportion? This level is called the  $p$ -value or the observed level of significance from the sample data. It answers the question that, “If  $H_0$  is true, what is the probability of obtaining a value of  $\bar{x}$ , (or  $\bar{p}$ , in the case of proportions) this far or more from  $H_0$ . If the  $p$ -value, as determined from the sample, is  $\geq \alpha$  the null hypothesis is accepted. Alternatively, if the  $p$ -value is less than  $\alpha$  then the null hypothesis is rejected and the alternative hypothesis is accepted. The use of the  $p$ -value approach is illustrated by re-examining the previous applications, *Filling machine*, *Taxes*, and *Seaworthiness of ships*.

## Application of the $p$ -value approach: *Filling machine*

1. At a significance level,  $\alpha$ , of 5% is there evidence that the volume of beer in the cans from this bottling line is different than the target volume of 0.50 litre?

As before, a sample of 25 cans is taken and the average of the sample volume is 0.5189 litre.

The test statistic,

$$z = \frac{\bar{x} - \mu_{H_0}}{\sigma_x / \sqrt{n}} = \frac{0.0189}{0.01} = 1.8900.$$



From Excel [function NORMSDIST] for a value of  $z$ , the 1.8900 area of the curve from the left is 97.06%.

Thus the area in the right-hand tail is  $100\% - 97.06\% = 2.94\%$ .

Since this is a two-tail test the area in the left tail is also 2.94%.

Since we have a two-tail test, the area in each of the tail set by the significance level is 2.50%.

As  $2.94\% > 2.50\%$  then we accept the null hypothesis and conclude that the volume of beer in the cans is not different from 0.50 litre. This is the same conclusion as before.

- At a significance level,  $\alpha$ , of 5% is there evidence that the volume of beer in the cans from this bottling line is greater than the target volume of 0.50 litre?

The value of the test statistic of 1.8900 gives an area in the right-hand tail of 2.94%.

We now have a one-tail, right-hand test when the significance level is 5%.

Since  $2.94\% < 5.00\%$  we reject the null hypothesis and accept the alternative hypothesis and conclude that there is evidence that the volume of beer in the cans is greater than 0.50 litre. This is the same conclusion as before.

## Application of the $p$ -value approach: Taxes

- At a significance level,  $\alpha$ , of 5% is there evidence that the average tax returns of the state will be different than the budget level of \$30,000 in this year?

The sample statistic gives a  $t$ -value which is equal to  $-1.8523$ .

From Excel [function TDIST] this sample statistic of  $-1.8523$ , for a two-tail test, indicates a probability of 8.38%.

Since  $8.38\% > 5.00\%$  we accept the null hypothesis and conclude that there is no evidence to indicate that the average tax receipts are significantly less than \$30,000.

- At a significance level,  $\alpha$ , of 5% is there evidence that the tax returns of the state will be less than the budget level of \$30,000 in this year?

The sample statistic gives a Student- $t$  value which is equal to  $-1.8523$  and from Excel [function TDIST] this sample statistic, for a one-tail test, indicates a probability of 4.19%.

Since  $4.19\% < 5.00\%$  we reject the null hypothesis and conclude that there is evidence to indicate that the average tax receipts are significantly less than \$30,000. This is the same conclusion as before.

## Application of the $p$ -value approach: Seaworthiness of ships

- At a 5% significance level, is there evidence to suggest that the seaworthiness of ships is different than the 80% indicated?

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = -\frac{0.06}{0.0327} = -1.8349$$

From Excel [function NORMSDIST] this sample statistic, for a two-tail test, indicates a probability of 3.31%.

As this is a two-tail test then there is 2.5% in each tail.

Since  $3.31\% > 2.50\%$  we accept the null hypothesis and conclude that there is no evidence to indicate that the seaworthiness of ships is different from the hypothesized value of 80%.

- At a 5% significance level, is there evidence to suggest that the seaworthiness of ships is less than the 80% indicated?

As this is a one-tail, left-hand test then there is 5% in the tail.

Since now  $3.31\% < 5.00\%$  we reject the null hypothesis and conclude that there is evidence to indicate that the seaworthiness of ships is less than the hypothesized value of 80%. This is the same conclusion as before.



## Interpretation of the $p$ -value

In hypothesis testing we are making inferences about a population based only on sampling. The sampling distribution permits us to make probability statements about a sample statistic on the basis of the knowledge of the population parameter. In the case of the filling machine for example where we are asking is there evidence that the volume of beer in the can is greater than 0.5 litre, the sample size obtained is 0.5189 litre. The probability of obtaining a sample mean of 0.5189 litre from a population whose mean is 0.5000 litre is 2.94% or quite small. Thus we have observed an unlikely event or an event so unlikely that we should doubt our assumptions about the population mean in the first place. Note, that in order to calculate the value of the test statistic we assumed that the null hypothesis is true and thus we have reason to reject the null hypothesis and accept the alternative.

The  $p$ -value provides useful information as it measures the amount of statistical evidence that supports the alternative hypothesis. Consider Table 8.1, which gives values of the sample mean, the value of the test statistic, and the corresponding  $p$ -value for the filling machine situation. As the sample mean gets larger, or moves further away from the hypothesized population mean of 0.5000 litre, the smaller is the  $p$ -value. Values of  $\bar{x}$

far above 0.5000 litre tend to indicate that the alternative hypothesis is true or the smaller the  $p$ -value, the more the statistical evidence there is to support the alternative hypothesis.

Remember that the  $p$ -value is not to be interpreted by saying that it is the probability that the null hypothesis is true. You cannot make a probability assumption about the population parameter 0.5000 litre as this is not a random variable.

## Risks in Hypothesis Testing

In hypothesis testing there are risks when you sample and then make an assumption about the population parameter. This is to be expected since statistical analysis gives no guarantee of the result but you hope that the risk of making a wrong decision is low.

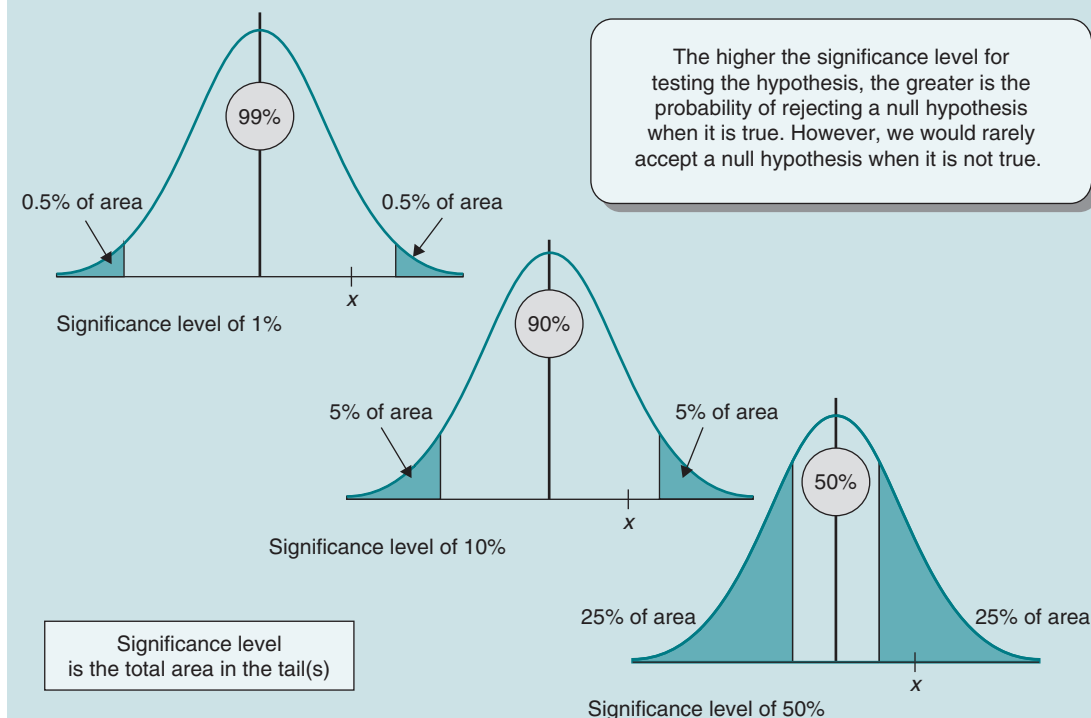
## Errors in hypothesis testing

The higher the value of the significance level,  $\alpha$  used for hypothesis testing then the higher is the percentage of the distribution in the tails. In this case, when  $\alpha$  is high, the greater is the probability of rejecting a null hypothesis. Since the null hypothesis is true, or is not true, then as  $\alpha$  increases there is a greater probability of rejecting the null hypothesis when in fact it is true. Looking at it another way, with a high significance level, that is a high value of  $\alpha$ , it is unlikely we would accept a null hypothesis when it is in fact not true. This relationship is illustrated in the normal distributions of Figure 8.10. At the 1% significance level, the probability of accepting the hypothesis, when it is false is greater than at a significance level of 50%. Alternatively, the risk of rejecting a null hypothesis when it is in fact true is greater at a 50% significance level, than at a 1% significance level. These errors in hypothesis testing are referred to as Type I or Type II errors.

Table 8.1 Sample mean and a corresponding  $z$  and  $p$ -value.

Sample mean $\bar{x}$	Test statistic $z$	$p$ -value %
0.5000	0.0000	50.00
0.5040	0.4000	34.46
0.5080	0.8000	21.19
0.5120	1.2000	11.51
0.5160	1.6000	5.48
0.5200	2.0000	2.28
0.5240	2.4000	0.82

Figure 8.10 Selecting a significance level.



A **Type I error** occurs if the null hypothesis is rejected when in fact it is true. The probability of a Type I error is called  $\alpha$  where  $\alpha$  is also the level of significance. A **Type II error** is accepting a null hypothesis when it is not true. The probability of a Type II error is called  $\beta$ . When the acceptance region is small, or  $\alpha$  is large, it is unlikely we would accept a null hypothesis when it is false. However, at a risk of being this sure, we will often reject a null hypothesis when it is in fact true. The level of significance to use depends on the cost of the error as illustrated as follows.

### Cost of making an error

Consider that a pharmaceutical firm makes a certain drug. A quality inspector tests a sample

of the product from the reaction vessel where the drug is being made. He makes a Type I error in his analysis. That is he rejects a null hypothesis when it is true or concludes from the sample that the drug does not conform to quality specifications when in fact it really does. As a result, all the production quantity in the reaction vessel is dumped and the firm starts the production all over again. In reality the batch was good and could have been accepted. In this case, the firm incurs all the additional costs of repeating the production operation. Alternatively, suppose the quality inspector makes a Type II error, or accepts a null hypothesis when it is in fact false. In this case the produced pharmaceutical product is accepted and commercialized but it does not conform to quality specifications. This may mean

that users of the drug could become sick, or at worse die. The “cost” of this error would be very high. In this situation, a pharmaceutical firm would prefer to make a Type I error, or destroying the production lot, rather than take the risk of poisoning the users. This implies having a high value of  $\alpha$  such as 50% as illustrated in Figure 8.10.

Suppose in another situation, a manufacturing firm is making a mechanical component that is used in the assembly of washing machines. An inspector takes a sample of this component from the production line and measures the appropriate properties. He makes a Type I error in the analysis. He rejects the null hypothesis that the component conforms to specifications, when in fact the null hypothesis is true. In this case to correct this conclusion would involve an expensive disassembly operation of many components on the shop floor that have already been produced. On the other hand if the inspector had made a Type II error, or accepting a null hypothesis when it is in fact false, this might involve a less expensive warranty repairs by the dealers when the washing machines are commercialized. In this latter case, the cost of the error is relatively low and manufacturer is more likely to prefer a Type II error even though the marketing image may be damaged. In this case, the manufacturer will set low levels for  $\alpha$  such as 10% as illustrated in Figure 8.10.

The cost of an error in some situations might be infinite and irreparable. Consider for example a murder trial. Under Anglo-Saxon law the null hypothesis, is that a person if charged with murder is considered innocent of the crime and the court has to prove guilt. In this case, the jury would prefer to commit a Type II error or accepting a null hypothesis that the person is innocent, when it is in fact not true, and thus let the guilty person go free. The alternative would be to accept a Type I error or rejecting the null hypothesis that the person is innocent, when it is in fact true. In this case the person would be found guilty and

risk the death penalty (at least in the United States) for a crime that they did not commit.

## Power of a test

In any analytical work we would like the probability of making an error to be small. Thus, in hypothesis testing we would like the probability of making a Type I error,  $\alpha$ , or the probability of making a Type II error  $\beta$  to be small. Thus, if a null hypothesis is false then we would like the hypothesis test to reject this conclusion every time. However, hypothesis tests are not perfect and when a null hypothesis is false, a test may not reject it and consequently a Type II error,  $\beta$ , is made or that is accepting a null hypothesis when it is false.

When the null hypothesis is false this implies that the true population value, does not equal the hypothesized population value but instead equals some other value. For each possible value for which the alternative hypothesis is true, or the null hypothesis is false, there is a different probability,  $\beta$  of accepting the null hypothesis when it is false. We would like this value of  $\beta$  to be as small as possible. Alternatively, we would like  $(1 - \beta)$  the probability of rejecting a null hypothesis when it is false, to be as large as possible. Rejecting a null hypothesis when it is false is exactly what a good hypothesis test ought to do. A high value of  $(1 - \beta)$  approaching 1.0 means that the test is working well. Alternatively, a low value of  $(1 - \beta)$  approaching zero means that the test is not working well and the test is not rejecting the null hypothesis when it is false. The value of  $(1 - \beta)$ , the measure of how well the test is doing, is called the **power of the test**.

Table 8.2 summarizes the four possibilities that can occur in hypothesis testing and what type of errors might be incurred. Again, as in all statistical work, in order to avoid errors in hypothesis testing, utmost care must be made to ensure that the sample taken is a true representation of the population.

Table 8.2 Sample mean and a corresponding  $z$  and  $p$ -value.

Decision you make	In reality for the population – null hypothesis, $H_0$ is true – what your test indicates	In reality for the population – null hypothesis, $H_0$ is false – what your test indicates
Null hypothesis, $H_0$ is accepted	<ul style="list-style-type: none"> <li>• Test statistic falls in the region <math>(1 - \alpha)</math></li> <li>• Decision is correct</li> <li>• No error is made</li> </ul>	<ul style="list-style-type: none"> <li>• Test statistic falls in the region <math>(1 - \alpha)</math></li> <li>• Decision is incorrect</li> <li>• A Type II error, <math>\beta</math> is made</li> </ul>
Null hypothesis, $H_0$ is rejected	<ul style="list-style-type: none"> <li>• Test statistic falls in the region <math>\alpha</math></li> <li>• Decision is incorrect</li> <li>• A Type I error, <math>\alpha</math> is made</li> </ul>	<ul style="list-style-type: none"> <li>• Test statistic falls in the region <math>\alpha</math></li> <li>• Decision is correct</li> <li>• No error is made</li> <li>• Power of test is <math>(1 - \beta)</math></li> </ul>

This chapter has dealt with hypothesis testing or making objective decisions based on sample data. The chapter opened with describing the concept of hypothesis testing, then presented hypothesis testing for the mean, hypothesis testing for proportions, the probability value in testing hypothesis, and finally summarized the risks in hypothesis testing.

### Concept of hypothesis testing

Hypothesis testing is to sample from a population and decide whether there is sufficient evidence to conclude that the hypothesis appears correct. In testing we need to decide on a significance level,  $\alpha$  which is the level of importance in the difference between values before we accept an alternative hypothesis. The significance level establishes a critical value, which is the barrier beyond which decisions will change. The concept of hypothesis testing is binomial. There is the null hypothesis denoted by  $H_0$ , which is the announced value. Then there is the alternative hypothesis,  $H_1$  which is the other situation we accept should we reject the null hypothesis. When we reject the null hypothesis we automatically accept the alternative hypothesis.

### Hypothesis testing for the mean value

In hypothesis testing for the mean we are trying to establish if there is statistical evidence to accept a hypothesized average value. We can have three frames of references. The first is to establish if there is a significant difference from the hypothesized mean. This gives a two-tail test. Another is to test to see if there is evidence that a value is significantly greater than the hypothesized amount. This gives rise to a one-tail, right-hand test. The third is a left-hand test that decides if a value is significantly less than a hypothesized value. In all of these tests the first step is to determine a sample test value, either  $z$ , or  $t$ , depending on our knowledge of the population. We

then compare this test value to our critical value, which is a direct consequence of our significance level. If our test value is within the limits of the critical value, we accept the null hypothesis. Otherwise we reject the null hypothesis and accept the alternative hypothesis.

### Hypothesis testing for proportions

The hypothesis test for proportions is similar to the test for the mean value but here we are trying to see if there is sufficient statistical evidence to accept or reject a hypothesized population proportion. The criterion is that we can assume the normal distribution in our analytical procedure. As for the mean, we can have a two-tail test, a one-tail, left-hand test, or a one-tail, right-hand test. We establish a significance level and this sets our critical value of  $z$ . We then determine the value of our sample statistic and compare this to the critical value determined from our significance level. If the test statistic is within our boundary limits we accept the null hypothesis, otherwise we reject it.

### The probability value in testing hypothesis

The probability, or  $p$ -value, for hypothesis testing is an alternative approach to the critical value method for testing assumptions about the population mean or the population proportion. The  $p$ -value is the minimum probability that we will tolerate before we reject the null hypothesis. When the  $p$ -value is less than  $\alpha$ , our level of significance, we reject the null hypothesis and accept the alternative hypothesis.

### Risks in hypothesis testing

As in all statistical methods there are risks when hypothesis testing is carried out. If we select a high level of significance, which means a large value of  $\alpha$  the greater is the risk of rejecting a null hypothesis when it is in fact true. This outcome is called a Type I error. However if we have a high value of  $\alpha$ , the risk of accepting a null hypothesis when it is false is low. A Type II error called  $\beta$  occurs if we accept a null hypothesis when it is in fact false. The value of  $(1 - \beta)$ , is a measure of how well the test is doing and is called the power of the test. The closer the value of  $(1 - \beta)$  is to unity implies that the test is working quite well.

## EXERCISE PROBLEMS

### 1. Sugar

#### Situation

One of the processing plants of Béghin Say, the sugar producer, has problems controlling the filling operation for its 1 kg net weight bags of white sugar. The quality control inspector takes a random sample of 22 bags of sugar and finds that the weight of this sample is 1,006 g. It is known from experience that the standard deviation of the filling operation is 15 g.

#### Required

1. At a significance level of 5% for analysis, using the critical value method, is there evidence that the net weight of the bags of sugar is different than 1 kg?
2. If you use the  $p$ -value for testing are you able to verify your conclusions in Question 1? Explain your reasoning.
3. What are the confidence limits corresponding to a significance level of 5%. How do these values corroborate your conclusions for Questions 1 and 2?
4. At a significance level of 10% for analysis, using the critical value method, is there evidence that the net weight of the bags of sugar is different than 1 kg?
5. If you use the  $p$ -value for testing are you able to verify your conclusions in Question 4? Explain your reasoning.
6. What are the confidence limits corresponding to a significance level of 10%. How do these values corroborate your conclusions for Questions 4 and 5?
7. Why is it necessary to use a difference test? Why should this processing plant be concerned with the results?

### 2. Neon lights

#### Situation

A firm plans to purchase a large quantity of neon light bulbs from a subsidiary of GE for a new distribution centre that it is building. The subsidiary claims that the life of the light bulbs is 2,500 hours, with a standard deviation of 40 hours. Before the firm finalizes the purchase it takes a random sample of 20 neon bulbs and tests them until they burn out. The average life of the sample of these bulbs is 2,485 hours. (Note, the firm has a special simulator that tests the bulb and in practice it does not require that the bulbs have to be tested for 2,500 hours.)

#### Required

1. Using the critical value approach, at a 5% significance level, is there evidence to suggest that the life of the light bulbs is different than 2,500 hours?
2. If you use the  $p$ -value for testing are you able to verify your conclusions in Question 1? Explain your reasoning.

3. Using the critical value approach, at a 5% significance level, is there evidence to suggest that the life of the light bulbs is less than 2,500 hours?
4. If you use the  $p$ -value for testing are you able to verify your conclusions in Question 3? Explain your reasoning.
5. If the results from the Questions 3 and 4 what options are open to the purchasing firm?

### 3. Graphite lead

#### Situation

A company is selecting a new supplier for graphite leads which it uses for its Pentel-type pencils. The supplier claims that the average diameter of its leads is 0.7 mm with a standard deviation of 0.05 mm. The company wishes to verify this claim because if the lead is significantly too thin it will break. If it is significantly too thick it will jam in the pencil. It takes a sample of 30 of these leads and measures the diameter with a micrometer gauge. The diameter of the samples is given in the table below.

0.7197	0.7090	0.6600	0.7800	0.7100	0.7030	0.6500	0.7660	0.7200	0.7788
0.7100	0.7100	0.7500	0.6200	0.7000	0.6960	0.7598	0.6900	0.7800	0.7012
0.6600	0.7200	0.6600	0.6900	0.6975	0.7540	0.6888	0.7700	0.7900	0.7600

#### Required

1. At a 5% significance level, using the critical value concept, is there evidence to suggest that the diameter of the lead is different from the supplier's claim?
2. At a 5% significance level, using the  $p$ -value concept, verify your answer obtained in Question 1. Explain your reasoning.
3. What are the confidence limits corresponding to a significance level of 5%. How do these values corroborate your conclusions for Questions 1 and 2?
4. At a 10% significance level, using the critical value concept, is there evidence to suggest that the diameter of the lead is different from the supplier's claim?
5. At a 10% significance level, using the  $p$ -value concept, verify your answer obtained in Question 1. Explain your reasoning.
6. What are the confidence limits corresponding to a significance level of 10%. How do these values corroborate your conclusions for Questions 4 and 5?
7. The mean of the sample data is an indicator whether the lead is too thin or too thick. If you applied the appropriate one-tail test what conclusions would you draw? Explain your logic.

### 4. Industrial pumps

#### Situation

Pumpet Corporation manufactures electric motors for many different types of industrial pumps. One of the parts is the drive shaft that attaches to the pump. An important criterion



for the drive shafts is that they should not be below a certain diameter. If this is the case, then when in use, the shaft vibrates, and eventually breaks. In the way that the drive shafts are machined, there are never problems of the shafts being oversized. For one particular model, MT 2501, the specification calls for a nominal diameter of the drive shaft of 100 mm. The company took a sample of 120 drive shafts from a large manufactured lot and measured their diameter. The results were as follows:

100.23	99.23	99.76	99.22	101.77	99.78
99.76	98.76	98.96	97.77	97.25	99.75
99.56	98.56	97.20	100.76	100.56	99.56
100.56	99.55	99.20	100.18	99.98	98.99
100.15	99.15	101.01	99.39	99.19	98.21
98.78	99.77	100.77	101.77	100.44	99.45
97.50	98.48	99.46	100.45	102.13	101.12
100.78	101.79	100.15	97.78	101.23	100.23
98.99	99.98	100.98	101.99	100.98	101.00
100.20	101.20	102.21	103.24	100.20	99.21
99.77	100.77	98.77	99.23	99.77	98.78
98.99	99.98	98.00	98.45	101.09	100.09
98.76	98.75	97.77	97.24	100.11	99.12
100.65	100.64	99.64	99.10	99.77	98.78
100.45	100.44	99.45	98.90	100.45	102.00
101.45	98.78	100.44	97.27	102.46	101.45
99.00	101.56	98.01	100.01	99.98	98.99
99.87	99.86	98.87	98.33	98.76	97.78
100.78	100.00	99.00	98.47	102.25	101.24
99.94	100.45	101.24	100.69	98.97	99.78

### Required

1. Pumpet normally uses a significance level of 5% for its analysis. In this case, using the critical value method, is there evidence that the shaft diameter of model MT 2501 is significantly below 100 mm? If so there would be cause to reject the lot. Explain your reasoning.
2. If you use the  $p$ -value for testing are you able to verify your conclusions in Question 1? Explain your reasoning.
3. A particular client of Pumpet insists that a significance level of 10% be used for analysis as they have stricter quality control limits. Using this level, and again making the test using the critical value criteria, is there evidence that the drive shaft diameter is significantly below 100 mm causing the lot to be rejected? Explain your reasoning.
4. If you use the  $p$ -value for testing are you able to verify your conclusions in Question 3? Explain your reasoning.
5. If instead of using the whole sample size indicated in the table you used just the data in the first three columns, how would your conclusions from the Questions 1 to 4 change?
6. From your answer to Question 5, what might you recommend?



## 5. Automatic teller machines (ATMs)

### Situation

Banks in France are closed for 2.5 days from Saturday afternoon to Tuesday morning. In this case banks need to have a reasonable estimate of how much cash to make available in their ATMs. For BNP-Paribas in its branches in the Rhone region in the Southeast of France it estimates that for this 2.5-day period the demand from its customers from those branches with a single ATM machine is €3,200 with a population standard deviation of €105. A random sample of the withdrawal from 36 of its branches indicate a sample average withdrawal of €3,235.

### Required

1. Using the concept of critical values, then at the 5% significance level does this data indicate that the mean withdrawal from the machines is different from €3,200?
2. Re-examine Question 1 using the  $p$ -value approach. Are your conclusions the same? Explain your conclusions?
3. What are the confidence limits at 5% significance? How do these values corroborate your answers to Questions 1 and 2?
4. Using the concept of critical values then at the 1% significance level does this data indicate that the mean life of the population of the batteries is different from €3,200?
5. Re-examine Question 4 using the  $p$ -value approach. Are your conclusions the same? Explain your conclusions.
6. What are the confidence limits at 1% significance? How do these values corroborate your answers to Questions 4 and 5?
7. Here we have used the test for a difference. Why is the bank interested in the difference rather than a one-tail test, either left or right hand?

## 6. Bar stools

### Situation

A supplier firm to IKEA makes wooden bar stools of various styles. In the production process of the bar stools the pieces are cut before shaping and assembling. The specifications require that the length of the legs of the bar stools is 70 cm. If the length is more than 70 cm they can be shaved down to the required length. However if pieces are significantly less than 70 cm they cannot be used for bar stools and are sent to another production area where they are re-cut to use in the assembly of standard chair legs. In the production of the legs for the bar stools it is known that the standard deviation of the process is 2.5 cm. In a production lot of legs for bar stools the quality

control inspector takes a random sample and the length of these is according to the following table.

65	74	69	68	68	72	72	73
71	75	74	68	67	69	67	68
67	70	68	68	68	71	67	70
69	69	69	67	72	66	67	72

### Required

1. At a 5% significance level, using the concept of critical value testing, does this sample data indicate that the length of the legs is less than 70 cm?
2. At the 5% significance level, using the  $p$ -value concept, does your answer corroborate the conclusion of Question 1? Give your reasoning.
3. At a 10% significance level, using the concept of critical value testing, does this sample data indicate that the length of the legs is less than 70 cm?
4. At the 10% significance level, using the  $p$ -value concept, does your answer corroborate the conclusion of Question 3? Give your reasoning.
5. Since we know the standard deviation we are correct to use the normal distribution for this hypothesis test. Assume that we did not know the process standard deviation and as the sample size of 32 is close to a cut-off point of 30, we used the Student- $t$  distribution. In this case, would our analysis change the conclusions of Questions 1 to 4?

## 7. Salad dressing

### Situation

Amora salad dressing is made in Dijon in France. One of their products, made with wine, indicates on the label that the nominal volume of the salad dressing is 1,000 ml. In the filling process the firm knows that the standard deviation is 5.00 ml. The quality control inspector takes a random sample of 25 of the bottles from the production line and measures their volumes, which are given in the following table.

993.2	997.7	1,000.0	1,001.0	998.9
999.1	1,000.0	1,000.0	992.5	994.9
994.3	996.0	1,005.2	993.4	1,001.8
995.9	1,002.4	1,005.2	1,002.0	992.7
996.2	997.9	1,002.0	1,001.0	995.0

### Required

1. At the 5% significance level, using the concept of critical value testing, does this sample data indicate that the volume of salad dressing in the bottles is different than the volume indicated on the label?

2. At the 5% significance level, using the  $p$ -value concept, does your answer corroborate the conclusion of Question 1? Give your reasoning.
3. At the 5% significance level, what are the confidence intervals when the test is asking for a difference in the volume? How do these intervals confirm your answers to Questions 1 and 2?
4. At the 5% significance level, using the concept of critical value testing, does this data indicate that the volume of salad dressing in the bottles is less than the volume indicated on the label?
5. At the 5% significance level, using the  $p$ -value concept, does your answer corroborate the conclusion of Question 4? Give your reasoning.
6. Why is the test mentioned in Question 4 important?
7. What can you say about the sensitivity of this sampling experiment?

## 8. Apples

### Situation

In an effort to reduce obesity among children, a firm that has many vending machines in schools is replacing chocolate bars with apples in its machines. Unlike chocolate bars that are processed and thus the average weight is easy to control, apples vary enormously in weight. The vending firm asks its supplier of apples to sort them before they are delivered as it wants the average weight to be 200 g. The criterion for this is that the vending firm wants to be reasonably sure that each child who purchases an apple is getting one of equivalent weight. A truck load of apples arrives at the vendor's depot and an inspector takes a random sample of 25 apples. The following is the weight of each apple in the sample.

198	201	202	186	199
199	208	196	196	196
207	195	187	199	189
195	190	195	197	209
199	205	203	190	199

### Required

1. At the 5% significance level, using the concept of critical value testing, does this sample data indicate that the weight of the truck load of apples is different than the desired 200 g?
2. At the 5% significance level, using the  $p$ -value concept, does your answer corroborate the conclusion of Question 1? Give your reasoning.
3. At the 5% significance level what are the confidence intervals when the test is asking for a difference in the volume. How do these intervals confirm your answers to Questions 1 and 2?
4. At the 5% significance level, using the concept of critical value testing, does this sample data indicate that the weight of the truck load of apples is less than the desired 200 g?
5. At the 5% significance level, using the  $p$ -value concept, does your answer corroborate the conclusion of Question 4? Give your reasoning.

## 9. Batteries

### Situation

A supplier of batteries claimed that for a certain type of battery the average life was 500 hours. The quality control inspector of a potential buying company took a random sample of 15 of these batteries from a lot and tested them until they died. The life of these batteries in hours is given in the table.

350	925	796	689	501
485	546	551	512	589
489	568	685	578	398

### Required

1. Using the concept of critical values, then at the 5% significance level does this data indicate that the mean life of the population of the batteries is different from the hypothesized value?
2. Re-examine Question 1 using the  $p$ -value approach. Are your conclusions the same? Explain your reasoning?
3. Using the concept of critical values then at the 5% significance level does this data indicate that the mean life of the population of the batteries is greater than the hypothesized value?
4. Re-examine Question 3 using the  $p$ -value approach. Are your conclusions the same? Explain your reasoning?
5. Explain the rationale for the differences in the answers to Questions 1 and 3, and the differences in the answers to Questions 2 and 4.

## 10. Hospital emergency

### Situation

A hospital emergency service must respond rapidly to sick or injured patients in order to increase rate of survival. A certain city hospital has an objective that as soon as it receives an emergency call an ambulance is on the scene within 10 minutes. The regional director wanted to see if the hospital objectives were being met. Thus during a weekend (the busiest time for hospital emergencies) a random sample of the time taken to respond to emergency calls were taken and this information, in minutes, is in the table below.

8	14	15	20	7
12	7	8	21	13
9	17	22	10	9

### Required

1. At the 5% significance level, using the concept of critical value testing, does this sample data indicate that the response time is different from 10 minutes?
2. At the 5% significance level, using the  $p$ -value concept, does your answer corroborate the conclusion of Question 1? Give your reasoning.
3. At the 5% significance level what are the confidence intervals when the test is asking for a difference? How do these intervals confirm your answers to Questions 1 and 2?
4. At the 5% significance level, using the concept of critical value testing, does this data indicate that the response time for an emergency call is greater than 10 minutes?
5. At the 5% significance level, using the  $p$ -value concept, does your answer corroborate the conclusion of Question 4? Give your reasoning.
6. Which of these two tests is the most important?

## 11. Equality for women

### Situation

According to Jenny Watson, the commission's chair of the United Kingdom Sex Discrimination Act (SDA), there continues to be an unacceptable pay gap of 45% between male and female full time workers in the private sector.<sup>1</sup> A sample of 72 women is taken and of these 22 had salaries less than their male counterparts for the same type of work.

### Required

1. Using the critical value approach for a 1% significance level, is there evidence to suggest that the salaries of women is different than the announced amount of 45%?
2. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 1. Explain your reasoning.
3. What are the confidence limits at the 1% level? How do they agree with your conclusions of Questions 1 and 2?
4. Using the critical value approach for a 5% significance level, is there evidence to suggest that the salaries of women is different than the announced amount of 45%?
5. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 3. Explain your reasoning.
6. What are the confidence limits at the 5% level? How do they agree with your conclusions of Questions 1 and 2?
7. How would you interpret these results?

## 12. Gas from Russia

### Situation

Europe is very dependent on natural gas supplies from Russia. In January 2006, after a bitter dispute with Ukraine, Russia cut off gas supplies to Ukraine but this also affected other

<sup>1</sup> Overell, S., Act One in the play for equality, *Financial Times*, 5 January 2006, p. 6.

European countries' gas supplies. This event jolted European countries to take a re-look at their energy policies. Based on 2004 data the quantity of imported natural gas of some major European importers and the amount from Russia in billions of cubic metres was according to the table below.<sup>2</sup> The amounts from Russia were on a contractual basis and did not necessarily correspond to physical flows.

Country	Total imports (m <sup>3</sup> billions)	Imports from Russia (m <sup>3</sup> billions)
Germany	91.76	37.74
Italy	61.40	21.00
Turkey	17.91	14.35
France	37.05	11.50
Hungary	10.95	9.32
Poland	9.10	7.90
Slovakia	7.30	7.30
Czech Republic	9.80	7.18
Austria	7.80	6.00
Finland	4.61	4.61

Industrial users have gas flow monitors at the inlet to their facilities according to the source of the natural gas. Samples from 35 industrial users were taken from both Italy and Poland and of these 7 industrial users in Italy and 31 in Poland were using gas imported from Russia.

### Required

1. Using the critical value approach at a 5% significance level, is there evidence to suggest that the proportion of natural gas Italy imports from Russia is different than the amount indicated in the table?
2. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 1. Explain your reasoning.
3. What are the confidence limits at the 5% level? How do they agree with your conclusions of Questions 1 and 2?
4. Using the critical value approach for a 10% significance level, is there evidence to suggest that the proportion of natural gas Italy imports from Russia is different than the amount indicated in the table?
5. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 4. Explain your reasoning.
6. What are the confidence limits at the 10% level? How do they agree with your conclusions of Questions 4 and 5?

<sup>2</sup> White, G.L., "Russia blinks in gas fight as crisis rattles Europe", *The Wall Street Journal*, 3 January 2005, pp. 1–10.

7. Using the critical value approach at a 5% significance level, is there evidence to suggest that the proportion of natural gas Poland imports from Russia is different than the amount indicated in the table?
8. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 7. Explain your reasoning.
9. What are the confidence limits at the 5% level? How do they agree with your conclusions of Questions 7 and 8?
10. Using the critical value approach for a 10% significance level, is there evidence to suggest that the proportion of natural gas Poland imports from Russia is different than the amount indicated in the table?
11. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 10. Explain your reasoning.
12. What are the confidence limits at the 10% level? How do they agree with your conclusions of Questions 10 and 11?
13. How would you interpret these results of all these questions?

### 13. International education

#### Situation

Foreign students are most visible in Australian and Swiss universities, where they make up more than 17% of all students. Although the United States attracts more than a quarter of the world's foreign students, they account for only some 3.5% of America's student population. Almost half of all foreign students come from Asia, particularly China and India. Social sciences, business, and law are the fields of study most popular with overseas scholars. The table below gives information for selected countries for 2003.<sup>3</sup>

Country	Foreign students as % of total	Country	Foreign students as % of total
Australia	19.0	Japan	2.0
Austria	13.5	Netherlands	4.0
Belgium	11.5	New Zealand	13.5
Britain	11.5	Norway	5.5
Czech Republic	4.5	Portugal	4.0
Denmark	9.0	South Korea	0.5
France	10.0	Spain	3.0
Germany	10.5	Sweden	8.0
Greece	2.0	Switzerland	18.0
Hungary	3.0	Turkey	1.0
Ireland	6.0	United States	3.5
Italy	2.0		

<sup>3</sup>Economic and financial indicators, *The Economist*, 17 September 2005, p. 108.

Random samples of 45 students were selected in Australia and in Britain. Of those in Australia, 14 were foreign, and 10 of those in Britain were foreign.

### Required

1. Using the critical value approach, at a 1% significance level, is there evidence to suggest that the proportion of foreign students in Australia is different from that indicated in the table?
2. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 1. Explain your reasoning.
3. What are the confidence limits at the 1% level? How do they agree with your conclusions of Questions 1 and 2?
4. Using the critical value approach, at a 5% significance level, is there evidence to suggest that the proportion of foreign students in Australia is different from that indicated in the table?
5. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 4. Explain your reasoning.
6. What are the confidence limits at the 5% level? How do they agree with your conclusions of Questions 1 and 2?
7. Using the critical value approach, at a 1% significance level, is there evidence to suggest that the proportion of foreign students in Britain is different from that indicated in the table?
8. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 7. Explain your reasoning.
9. What are the confidence limits at the 1% level? How do they agree with your conclusions of Questions 7 and 8?
10. Using the critical value approach, at a 5% significance level, is there evidence to suggest that the proportion of foreign students in Britain is different from that indicated in the table?
11. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 10. Explain your reasoning.
12. What are the confidence limits at the 5% level? How do they agree with your conclusions of Questions 10 and 11?

## 14. United States employment

### Situation

According to the United States labour department the jobless rate in the United States fell to 4.9% at the end of 2005. It was reported that 108,000 jobs were created in December and 305,000 in November. Taken together, these new jobs created over the past 2 months allowed the United States to end the year with about 2 million more jobs than it had

<sup>4</sup> Andrews, E.L., "Jobless rate drops to 4.9% in U.S," *International Herald Tribune*, 7/8 January 2006, p. 17.



12 months ago.<sup>4</sup> Random samples of 83 people were taken in both Palo Alto, California and Detroit, Michigan. Of those from Palo Alto, 4 said they were unemployed and 8 in Detroit said they were unemployed.

### Required

1. Using the critical value approach, at a 5% significance level, is there evidence to suggest that the unemployment rate in Palo Alto is different from the national unemployment rate?
2. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 1. Explain your reasoning.
3. What are the confidence limits at the 5% level? How do they agree with your conclusions of Questions 1 and 2?
4. Using the critical value approach, at a 10% significance level, is there evidence to suggest that the unemployment rate in Palo Alto is different from the national unemployment rate?
5. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 4. Explain your reasoning.
6. What are the confidence limits at the 10% level? How do they agree with your conclusions of Questions 4 and 5?
7. Using the critical value approach, at a 5% significance level, is there evidence to suggest that the unemployment rate in Detroit is different from the national unemployment rate?
8. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 7. Explain your reasoning.
9. What are the confidence limits at the 5% level? How do they agree with your conclusions of Questions 7 and 8?
10. Using the critical value approach, at a 10% significance level, is there evidence to suggest that the unemployment rate in Detroit is different from the national unemployment rate?
11. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 10. Explain your reasoning.
12. What are the confidence limits at the 10% level? How do they agree with your conclusions of Questions 10 and 11?
13. Explain your results for Palo Alto and Detroit.

## 15. Mexico and the United States

### Situation

On 30 December 2005 a United States border patrol agent shot dead an 18-year-old Mexican as he tried to cross the border near San Diego, California. The patrol said the shooting was in self-defence and that the dead man was a *coyote*, or people smuggler. In

2005, out of an estimated 400,000 Mexicans who crossed illegally into the United States, more than 400 died in the attempt. Illegal immigration into the United States has long been a problem and to control the movement there are plans to construct a fence along more than a third of the 3,100 km border. According to data for 2004, there are some 10.5 million Mexicans in the United States, which represents some 31% of the foreign-born United States population. The recorded Mexicans in the United States of America is equivalent to 9% of Mexico's total population. In addition, it is estimated that there are some 10 million undocumented immigrants in the United States of which 60% are considered to be Mexican.<sup>5</sup> A random sample of 57 foreign-born people were taken in the United States and of these 11 said they were Mexican and of those 11, two said they were illegal.

### Required

1. What is the probability that a Mexican who is considering to cross the United States border will die or be killed in the attempt?
2. Using the critical value approach, at a 5% significance level, is there evidence to suggest that the proportion of Mexicans, as foreign-born people, living in the United States is different from the indicated data?
3. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 2. Explain your reasoning.
4. What are the confidence limits at the 5% level? How do they agree with your conclusions of Questions 1 and 2?
5. Using the critical value approach, at a 10% significance level, is there evidence to suggest that the proportion of Mexicans, as foreign-born people, living in the United States is different from the indicated data?
6. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 5. Explain your reasoning.
7. What are the confidence limits at the 10% level? How do they agree with your conclusions of Questions 5 and 6?
8. Using the critical value approach, at a 5% significance level, is there evidence to suggest that the number of undocumented Mexicans living in the United States is different from the indicated data?
9. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 8. Explain your reasoning.
10. What are the confidence limits at the 5% level? How do they agree with your conclusions of Questions 8 and 9?
11. Using the critical value approach, at a 10% significance level, is there evidence to suggest that the number of undocumented Mexicans living in the United States is different from the indicated data?
12. Using the  $p$ -value approach are you able to corroborate your conclusions from Question 11. Explain your reasoning.

<sup>5</sup> "Shots across the border," *The Economist*, 14 January 2006, p. 53.

13. What are the confidence limits at the 10% level? How do they agree with your conclusions of Questions 11 and 12?
14. What are your comments about the difficulty in carrying out this hypothesis test?

## 16. Case: Socrates and Erasmus

### Situation

The Socrates II European programme supports cooperation in education in eight areas, from school to higher education, from new technologies, to adult learners. Within Socrates II is the programme *Erasmus* that was established in 1987 with the objective to facilitate the mobility of higher education students within European universities. The programme is named after the philosopher, theologian, and humanist, Erasmus of Rotterdam (1465–1536). Erasmus lived and worked in several parts of Europe in quest of knowledge and experience believing such contacts with different cultures could only furnish a broad knowledge. He left his fortune to the University of Basel and became a precursor of mobility grants.

The Erasmus programme has 31 participating countries that include the 25 member states of the European Union, the three European Economic area countries of Iceland, Liechtenstein, and Norway, and the current three candidate countries – Romania, Bulgaria, and Turkey. The programme is open to universities for all higher education programmes including doctoral courses. In between the academic years 1987–1988 to 2003–2004 more than 1 million university students had spent an Erasmus period abroad and there are 2,199 higher education institutions participating in the programme. The European Union budget for 2000–2006 is €950 million of which about is €750 million is for student grants. In the academic year 2003–2004, the Erasmus students according to their country of origin and their country of study, or host country is given in the cross-classification Table 1 and the field of study for these students according to their home country is given in Table 2. It is the target of the Erasmus programme to have a balance in the gender mix and the programme administrators felt that the profile for subsequent academic years would be similar to the profile for the academic year 2003–2004.<sup>6</sup>

### Required

A sample of random data for the Erasmus programme for the academic year 2005–2006 was provided by the registrar's office and this is given in Table 3. Does this information bear out the programme administrator's beliefs if this is tested at the 1%, 5%, and 10% significance level for a difference?

<sup>6</sup> <http://europa.eu.int:comm/education/programmes/socrates/erasmus/what-en.html>

Table 1 Students by field of study 2003–2004 according to home country.

Subject	AT	BE	BG	CY	CZ	DK	EE	FI	FR	DE	GR	HU	IS	IE	IT	LV
Agricultural sciences	37	156	51	0	187	18	6	64	398	181	81	136	3	3	317	14
Architecture, Planning	128	163	32	0	168	54	12	30	519	762	149	75	0	30	877	9
Art and design	193	209	42	0	182	60	47	326	651	906	143	114	24	90	756	31
Business studies	1,117	1,089	97	7	584	364	47	1,383	6,573	5,023	306	450	56	593	1,963	88
Education, Teacher training	260	414	12	24	228	74	2	100	320	535	81	126	22	24	267	27
Engineering, Technology	248	384	133	3	481	112	22	487	2,833	1,376	143	147	20	52	1,545	10
Geography, Geology	32	28	12	0	90	27	9	33	259	433	46	66	3	12	206	14
Humanities	147	105	14	0	148	141	9	136	598	1,048	131	64	13	51	1,144	13
Languages, Philological sciences	505	603	73	15	464	346	51	316	3,321	3,528	327	248	47	305	3,346	21
Law	231	357	37	0	185	103	28	117	1,449	1,474	191	159	7	142	1,455	7
Mathematics, Informatics	146	139	86	0	123	20	4	108	570	803	104	64	4	45	392	13
Medical sciences	144	349	60	12	222	115	12	291	399	1,021	172	125	4	46	1,045	8
Natural sciences	143	51	33	0	113	33	4	93	843	879	87	29	3	62	453	6
Social sciences	250	500	48	3	309	171	32	307	1,787	2,067	343	200	15	210	2,220	38
Communication and information science	112	212	19	0	14	44	12	100	295	425	38	23	0	32	723	5
Other areas	28	30	2	0	91	4	8	60	166	227	43	32	0	8	120	4
Total	3,721	4,789	751	64	3,589	1,686	305	3,951	20,981	20,688	2,385	2,058	221	1,705	16,829	308

Table 1 (Continued).

Subject	LI	LT	LU	MT	NL	NO	PL	PT	RO	SK	SI	ES	SE	UK	EUI	Total
Agricultural sciences	0	48	0	0	80	27	112	69	61	37	23	566	19	23	0	2,717
Architecture, Planning	9	37	4	2	109	19	321	264	64	18	24	854	64	96	0	4,893
Art and design	0	63	4	3	145	69	232	205	87	34	38	905	90	489	0	6,138
Business studies	10	241	15	6	1,089	275	1,342	386	290	169	146	3,244	902	1,332	0	29,187
Education, Teacher training	0	56	43	11	354	92	126	215	47	15	17	602	69	163	0	4,326
Engineering, Technology	0	189	6	9	224	112	752	479	604	106	35	3,109	424	269	0	14,314
Geography, Geology	0	25	8	2	84	5	158	66	147	10	6	450	31	88	0	2,350
Humanities	0	33	2	1	81	39	171	60	116	22	12	654	48	206	8	5,215
Languages, Philological sciences	0	92	14	7	253	84	675	334	451	84	97	2,568	121	2,875	0	21,171
Law	0	87	6	31	303	77	429	190	98	25	51	1,413	195	754	1	9,602
Mathematics, Informatics	0	65	0	1	55	35	301	87	176	23	3	674	46	92	0	4,179
Medical sciences	0	85	8	32	219	142	247	407	209	71	6	1,211	176	232	0	7,070
Natural sciences	0	43	7	4	51	22	361	216	206	29	2	1,062	84	220	0	5,139
Social sciences	0	97	19	5	992	137	928	487	355	29	65	1,701	313	585	1	14,214
Communication and information science	0	17	1	5	264	10	68	155	54	3	19	800	56	83	0	3,589
Other areas	0	16	1	0	85	11	53	162	40	7	2	221	29	32	0	1,482
<b>Total</b>	<b>19</b>	<b>1,194</b>	<b>138</b>	<b>119</b>	<b>4,388</b>	<b>1,156</b>	<b>6,276</b>	<b>3,782</b>	<b>3,005</b>	<b>682</b>	<b>546</b>	<b>20,034</b>	<b>2,667</b>	<b>7,539</b>	<b>10</b>	<b>135,586</b>

Table 2 Erasmus students 2003–2007 by home country and host country.

Home Country	Code	AT	BE	BG	CY	CZ	DK	EE	FI	FR	DE	GR	HU	IS	IE	IT	LV
Austria	AT		79	3	5	51	104	7	227	528	262	30	30	15	132	461	5
Belgium	BE	105		11	1	51	84	5	218	768	306	75	28	3	121	467	4
Bulgaria	BG	52	46				14		16	136	227	62			6	39	
Cyprus	CY	1	0	0			2		14	9	4	13			0	3	
Czech Republic	CZ	211	134				103		241	510	931	78			43	180	
Denmark	DK	70	44		2	19		2	5	260	302	13	3	12	36	111	
Estonia	EE	16	10				19		47	42	59	6			2	26	
Finland	FI	229	148	5	9	126	37	35		413	654	72	162	14	111	190	9
France	FR	361	420	9	10	206	500	21	727		2,804	218	169	23	1,081	1,550	3
Germany	DE	387	330	17	7	207	410	25	918	3,997		165	171	47	926	1,755	23
Greece	GR	71	140	6	8	63	45	1	116	420	356		20	2	27	248	1
Hungary	HU	110	98				44		201	276	566	42			15	227	
Iceland	IS	10	4				54		1	26	40	3			2	16	
Ireland	IE	35	47	6	1	26	30	2	40	557	292	12	5			109	
Italy	IT	339	633	8	7	86	357	28	367	2,859	1,994	180	129	29	230		4
Latvia	LV	8	27				13		42	18	111	2			2	9	
Liechtenstein	LI	0	0				2		3		1				1		
Lithuania	LT	49	70				145		180	77	294	18			10	67	
Luxembourg	LU	17	1	0	0	2	2	0	1	27	39		0	0	0	9	0
Malta	MT	4	5	0	0	0	2	0	6	3	6	0			6	52	
Netherlands	NL	98	184	1	0	44	158	7	275	543	391	42	49	11	88	256	6
Norway	NO	50	29	0	0	0	53	0	15	156	190	15	0	0	17	85	0
Poland	PL	159	358				362		310	855	1,870	122			74	481	
Portugal	PT	53	250	8	8	103	63	3	95	325	295	53	59	4	19	713	5
Rumania	RO	38	163				29		33	1,125	457	87			21	448	
Slovakia	SK	44	50				11		52	80	191	24			2	58	
Slovenia	SI	59	30				19		24	62	125	6			1	56	
Spain	ES	298	1,054	11	0	169	573	12	501	3,412	2,553	178	67	21	513	4,250	1
Sweden	SE	142	42	0	0	38	25	10	24	484	426	17	28	9	80	137	3
United Kingdom	UK	143	117	5	4	107	136	8	233	2,303	1,127	60	31	9	21	740	1
EUI*	EUR	2								4	1						
Total		3,161	4,513	90	62	1,298	3,396	166	4,932	20,275	16,874	1,593	951	199	3,587	12,743	65

Table 2 (Continued).

Home Country	Code	LI	LT	LU	MT	NL	NO	PL	PT	RO	SK	SI	ES	SE	UK	Total
Austria	AT	1	12	0	14	215	82	22	60	8	6	16	631	305	410	3,721
Belgium	BE	0	7	3	13	377	40	69	207	30	10	9	1,287	149	341	4,789
Bulgaria	BG					23			34				43	9	44	751
Cyprus	CY								2				3	5	8	64
Czech Republic	CZ					203			189				286	163	317	3,589
Denmark	DK		3		4	117	27	12	15	5		5	259	30	330	1,686
Estonia	EE					10			4				30	26	8	305
Finland	FI		15		16	377	15	60	58	13	22	29	479	101	552	3,951
France	FR		25	6	43	891	246	314	288	167	30	40	5,115	1,062	4,652	20,981
Germany	DE	8	49	1	28	862	463	395	283	27	26	24	4,325	1,653	3,159	20,688
Greece	GR		1	1	5	106	17	14	90	3	0	2	374	109	139	2,385
Hungary	HU					145			42				125	58	109	2,058
Iceland	IS					13			1				36	2	13	221
Ireland	IE		4		5	110	8	10	18			3	291	57	37	1,705
Italy	IT	1	28		71	607	156	174	766	129	29	20	5,688	399	1,511	16,829
Latvia	LV					24			4				9	32	7	308
Liechtenstein	LI					4			2					1	5	19
Lithuania	LT					30			51				61	120	22	1,194
Luxembourg	LU	0	0		0	0	0	1	6	0	0	0	14	3	16	138
Malta	MT					7			2				3	1	22	119
Netherlands	NL	0	10	0	18		140	21	93	14	3	5	907	389	635	4,388
Norway	NO	0	0	0		78		0	36	0	0	0	231	42	159	1,156
Poland	PL					294			222				546	286	337	6,276
Portugal	PT	1	26	0	4	250	38	125		68	7	14	920	95	178	3,782
Rumania	RO					72			119				285	42	86	3,005
Slovakia	SK			3		29			30				59	17	32	682
Slovenia	SI					25			30				63	17	29	546
Spain	ES	0	24	0	9	1,263	200	176	992	59	32	22		670	2,974	20,034
Sweden	SE	0	11	0	11	236	22	24	25	3	0	6	370		494	2,667
United Kingdom	UK	0	3	0	12	365	69	42	97	10	16	6	1,636	238		7,539
EUI*	EUR													1	2	10
Total		11	218	14	253	6,733	1,523	1,459	3,766	536	181	201	24,076	6,082	16,628	135,586

\* European University Institute, Florence.

Table 3 Sample of Erasmus student enrollments for the academic year 2005–2006.

Family name	First name	Home country	Study area	Gender
Algard	Erik	Norway	Business studies	M
Alinei	Gratian	Rumania	Business studies	M
Andersen	Birgitte Brix	Denmark	Engineering, Technology	F
Bay	Hilde	Norway	Social sciences	F
Bednarczyk	Tomasz	Poland	Law	M
Berberich	Rémi	Germany	Engineering, Technology	M
Berculo	Ruwan	Netherlands	Business studies	M
Engler	Dorothea	Germany	Geography, Geology	F
Ernst	Folker	Germany	Business studies	M
Fouche	Elie	France	Education, Teacher training	M
Garcia	Miguel	Spain	Communication and information science	M
Guenin	Aurélié	France	Humanities	F
Johannessen	Sanne Lyng	Denmark	Business studies	F
Justnes	Petter	Norway	Languages, Philological sciences	M
Kauffeldt	Ane Katrine	Denmark	Business studies	F
Keddie	Nikki	United Kingdom	Mathematics, Informatics	F
Lorenz	Jan Sebastian	Germany	Business studies	M
Mallet	Guillaume	France	Business studies	M
Manzo	Margherita	Italy	Business studies	F
Margineanu	Florin	Rumania	Agricultural sciences	M
Miechowka	Anne Sophie	France	Engineering, Technology	F
Mynborg	Astrid	Denmark	Humanities	F
Napolitano	Silvia	Italy	Architecture, Planning	F
Neilson	Alison	United Kingdom	Business studies	F
Ou	Kalvin	France	Education, Teacher training	M
Rachbauer	Thomas	Austria	Engineering, Technology	M
Savreux	Margaux	France	Mathematics, Informatics	F
Seda	Jiri	Czech Republic	Agricultural sciences	M
Semoradova	Petra	Czech Republic	Natural sciences	F
Torres	Maria Teresa	Spain	Humanities	F
Ungerstedt	Malin	Sweden	Law	F
Ververken	Alexander	Belgium	Languages, Philological sciences	M
Viscardi	Alessandra	Italy	Business studies	F
Zawisza	Katarzyna	Poland	Business studies	F



*This page intentionally left blank*

# Hypothesis testing for different populations

## Women still earn less than men

*On 27 February 2006 the Women and Work Commission (WWC) published its report on the causes of the “gender pay gap” or the difference between men’s and women’s hourly pay. According to the report, British women in full-time work currently earn 17% less per hour than men. Also in February, the European Commission brought out its own report on the pay gap across the whole European Union. Its findings were similar in that on an hourly basis, women earn 15% less than men for the same work. In the United States, the difference in median pay between men and women is around 20%. According to the WWC report the gender pay gap opens early. Boys and girls study different subjects in school, and boy’s subjects lead to more lucrative careers. They then work in different sorts of jobs. As a result, average hourly pay for a woman at the start of her working life is only 91% of a man’s, even though nowadays she is probably better qualified.<sup>1</sup> How do we compile this type of statistical information? We can use hypothesis testing for more than one type of population – the subject of this chapter.*

---

<sup>1</sup> “Women’s pay: The hand that rocks the cradle”, *The Economist*, 4 March 2006, p. 33.

## Learning objectives

After you have studied this chapter you will understand how to extend **hypothesis testing for two populations** and to use the **chi-square hypothesis test for more than two populations**. The subtopics of these themes are as follows:

- ✓ **Difference between the mean of two independent populations** • Difference of the means for large samples • The test statistic for large samples • Application of the differences in large samples: *Wages of men and women* • Testing the difference of the means for small sample sizes • Application of the differences in small samples: *Production output*
- ✓ **Differences of the means between dependent or paired populations** • Application of the differences of the means between dependent samples: *Health spa*
- ✓ **Difference between the proportions of two populations with large samples** • Standard error of the difference between two proportions • Application of the differences of the proportions between two populations: *Commuting*
- ✓ **Chi-square test for dependency** • Contingency table and chi-square application: *Work schedule preference* • Chi-square distribution • Degrees of freedom • Chi-square distribution as a test of independence • Determining the value of chi-square • Excel and chi-square functions • Testing the chi-square hypothesis for work preference • Using the  $p$ -value approach for the hypothesis test • Changing the significance level

In Chapter 8, we presented by sampling from a single population, how we could test a hypothesis or an assumption about the parameter of this single population. In this chapter we look at hypothesis testing when there is more than one population involved in the analysis.

### Difference Between the Mean of Two Independent Populations

The difference between the mean of two independent populations is a hypothesis test to sample in order to see if there is a significant difference between the parameters of two independent populations, as for example the following:

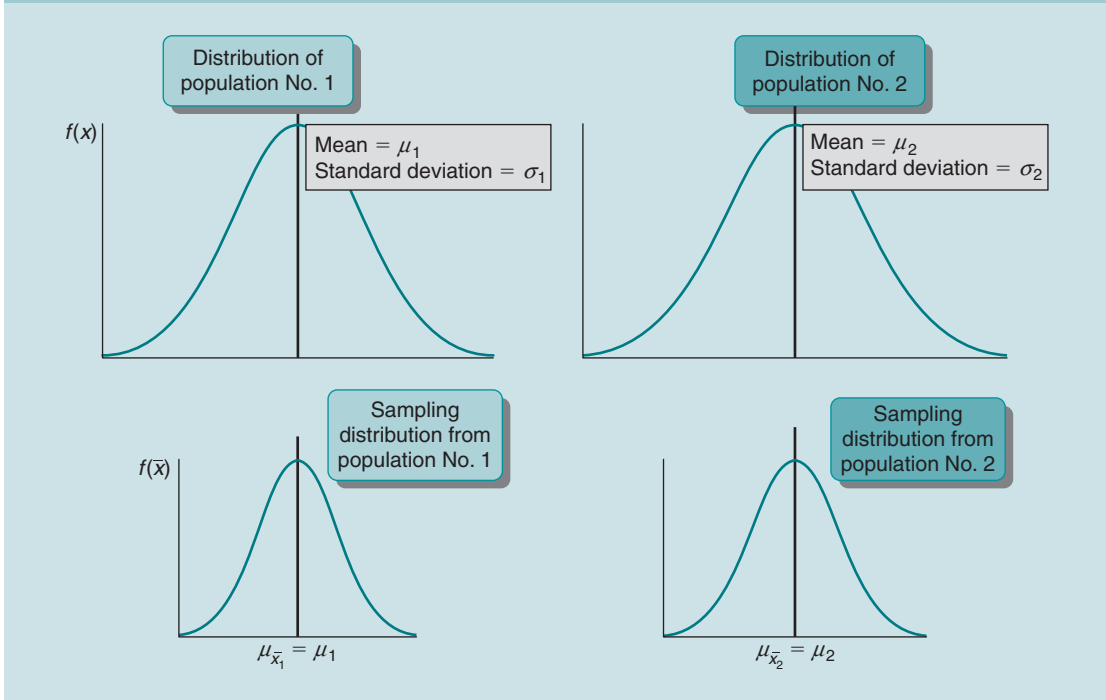
- A human resource manager wants to know if there is a significant difference between the

salaries of men and the salaries of women in his multinational firm.

- A professor of Business Statistics is interested to know if there is a significant difference between the grade level of students in her morning class and in a similar class in the afternoon.
- A company wants to know if there is a significant difference in the productivity of the employees in one country and another country.
- A firm wishes to know if there is a difference in the absentee rate of employees in the morning shift and the night shift.
- A company wishes to know if sales volume of a certain product in one store is different from another store in a different location.

In these cases we are not necessarily interested in the specific value of a population parameter but more to understand something about the relation between the two parameters from the populations. That is, are they essentially the same, or is there a significant difference?

Figure 9.1 Two independent populations.



## Difference of the means for large samples

The hypothesis testing concept between two population means is illustrated in Figure 9.1. The figure on the left gives the normal distribution for Population No. 1 and the figure on the right gives the normal distribution for Population No. 2. Underneath the respective distributions are the sampling distributions of the means taken from that population. From the data another distribution can be constructed, which is then the difference between the values of sample means taken from the respective populations. Assume, for example, that we take a random sample from Population 1, which gives a sample mean of  $\bar{x}_1$ . Similarly we take a random sample from Population 2 and this gives a sample mean of  $\bar{x}_2$ . The difference between the values of the sample means

is then given as,

$$\bar{x}_1 - \bar{x}_2 \quad 9(i)$$

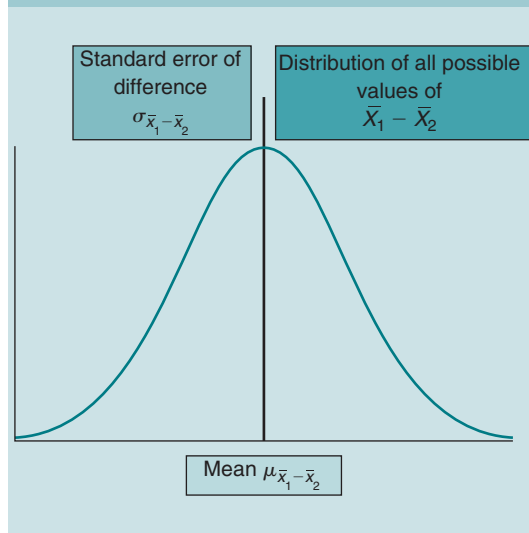
When the value of  $\bar{x}_1$  is greater than  $\bar{x}_2$  then the result of equation 9(i) is positive. When the value of  $\bar{x}_1$  is less than  $\bar{x}_2$  then the result of equation 9(i) value is negative. If we construct a distribution of the difference of the entire sample means then we will obtain a sampling distribution of the differences of all the possible sample means as shown in Figure 9.2.

The mean of the sample distribution of the differences of the mean is written as,

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} \quad 9(ii)$$

When the mean of the two populations are equal then  $\mu_{\bar{x}_1} - \mu_{\bar{x}_2} = 0$ .

Figure 9.2 Distribution of all possible values of difference between two means.



From Chapter 6, using the central limit theory we developed the following relationship for the standard error of the sample mean:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \quad 6(\text{ii})$$

Extending this relationship for sampling from two populations, the **standard deviation of the distribution of the difference between the sample means**, as given in Figure 9.2, is determined from the following relationship:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad 9(\text{iii})$$

where  $\sigma_1^2$  and  $\sigma_2^2$  are respectively the variance of Population 1 and Population 2,  $\sigma_1$  and  $\sigma_2$  are the standard deviations and  $n_1$  and  $n_2$  are the sample sizes taken from these two populations. This relationship is also the **standard error of the difference between two means**. If we do not know the population standard deviations, then we use the sample standard deviation,  $s$ , as an estimate

of the population standard deviation  $\hat{\sigma}$ . In this case the **estimated standard deviation of the distribution of the difference between the sample means** is,

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \quad 9(\text{iv})$$

### The test statistic for large samples

From Chapter 6, when we have just one population, the test statistic  $z$  for large samples, that is greater than 30, is given by the relationship,

$$z = \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}} \quad 6(\text{iv})$$

When we test the difference between the means of two populations then the equation for the test statistic becomes,

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_{H_0}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad 9(\text{v})$$

Alternatively, if we do not know the population standard deviation, then equation 9(v) becomes,

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_{H_0}}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} \quad 9(\text{vi})$$

In this equation,  $(\bar{x}_1 - \bar{x}_2)$  is the difference between the sample means taken from the population and  $(\mu_1 - \mu_2)_{H_0}$  is the difference of the hypothesized means of the population. The following application example illustrates this concept.

Table 9.1 Difference in the wages of men and women.

	Sample mean $\bar{x}$ (\$)	Sample standard deviation, $s$ (\$)	Sample size, $n$
Population 1, women	28.65	2.40	130
Population 2, men	29.15	1.90	140

### Application of the differences in large samples: *Wages of men and women*

A large firm in the United States wants to know, the relationship between the wages of men and women employed at the firm. Sampling the employees gave the information in \$US in Table 9.1.

1. At a 10% significance level, is there evidence of a difference between the wages of men and women? At a 10% significance level we are asking the question is there a difference, which means to say that values can be greater or less than. This is a two-tail test with 5.0% in each of the tails. Using [function NORMSINV] in Excel the critical value of  $z$  is  $\pm 1.6449$ .

The null and alternative hypotheses are as follows:

- Null hypothesis,  $H_0$ :  $\mu_1 = \mu_2$  is that there is no significant difference in the wages.
- Alternative hypothesis,  $H_1$ :  $\mu_1 \neq \mu_2$  is that there is a significant difference in the wages.

Since we have only a measure of the sample standard deviation  $s$  and not the population standard deviation  $\sigma$ , we use equation 9(vi) to determine the test or sample statistic  $z$ :

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_{H_0}}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

Here,  $\bar{x}_1 - \bar{x}_2 = 28.65 - 29.15 = -0.50$  and  $\mu_1 - \mu_2 = 0$  since the null hypothesis

is that there is no difference between the population means.

The standard error of the difference between the means is from equation 9(iv):

$$\begin{aligned}\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} &= \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} = \sqrt{\frac{2.40^2}{130} + \frac{1.90^2}{140}} \\ &= 0.2648\end{aligned}$$

Thus,

$$z = -\frac{0.50}{0.2648} = -1.8886$$

Since the sample, or test statistic, of  $-1.8886$  is less than the critical value of  $-1.6449$  we reject the null hypothesis and conclude that there is evidence to indicate that the wages of women are significantly different from that of men.

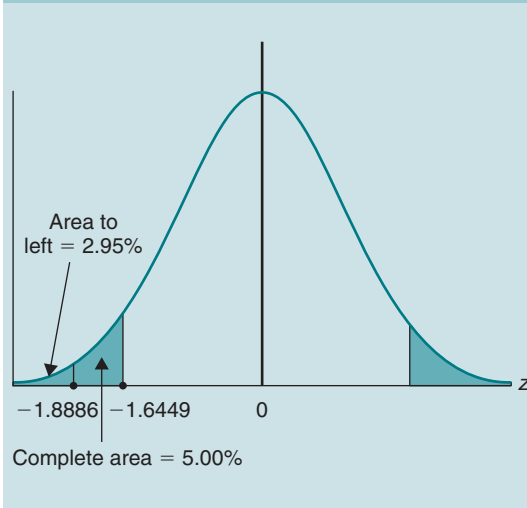
As discussed in Chapter 8 we can also use the  $p$ -value approach to test the hypothesis. In this example the sample value of  $z = -1.8886$  and using [function NORMSDIST] gives an area in the tail of 2.95%. Since 2.95% is less than 5% we reject the null hypothesis. This is the same conclusion as previously.

The representation of this worked example is illustrated in Figure 9.3.

### Testing the difference of the means for small sample sizes

When the sample size is small, or less than 30 units, then to be correct we must use the

Figure 9.3 Difference in wages between men and women.



Student-*t* distribution. When we use the Student-*t* distribution the population standard deviation is unknown. Thus to estimate the standard error of the difference between the two means we use equation 9(iv):

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \quad 9(\text{iv})$$

However, a difference from the hypothesis testing of large samples is that here we make the assumption that the variance of Population 1,  $\sigma_1^2$  is equal to the variance of Population 2,  $\sigma_2^2$ , or  $\sigma_1^2 = \sigma_2^2$ . This then enables us to use a pooled variance such that the sample variance,  $s_1^2$ , taken from Population 1 can be pooled, or combined, with  $s_2^2$ , to give a value  $s_p^2$ . This value of the pooled estimate  $s_p^2$  is given by the relationship,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \quad 9(\text{vii})$$

This value of  $s_p^2$  is now the best estimate of the variance common to both populations  $\sigma^2$ , on the assumption that the two population variances

are equal. Note, that the denominator in equation 9(vii), can be rewritten as,

$$(n_1 - 1) + (n_2 - 1) = (n_1 + n_2 - 2) \quad 9(\text{viii})$$

This is so because we now have two samples and thus two degrees of freedom. Note that in Chapter 8 when we took one sample of size *n* in order to use the Student-*t* distribution we had  $(n - 1)$  degrees of freedom.

Combining equations 9(iv) and 9(vii) the relationship for the estimated standard error of the difference between two sample means, when there are small samples on the assumption that the population variances are equal, is given by,

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad 9(\text{ix})$$

Then by analogy with equation 9(vi) the value of the Student-*t* distribution is given by,

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_{H_0}}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad 9(\text{x})$$

If we take samples of equal size from each of the populations, then since  $n_1 = n_2$ , equation 9(vii) becomes as follows:

$$\begin{aligned} s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \\ &= \frac{(n_1 - 1)s_1^2 + (n_1 - 1)s_2^2}{(n_1 - 1) + (n_1 - 1)} \\ &= \frac{(n_1 - 1)(s_1^2 + s_2^2)}{(n_1 - 1)(1 + 1)} = \frac{(s_1^2 + s_2^2)}{2} \end{aligned} \quad 9(\text{xi})$$

Further, the relationship in the denominator of equation 9(x) can be rewritten as,

$$\left( \frac{1}{n_1} + \frac{1}{n_2} \right) = \left( \frac{1}{n_1} + \frac{1}{n_1} \right) = \frac{2}{n_1} \quad 9(\text{xii})$$

Table 9.2 Production output between morning and night shifts.

Morning (m)	29	24	28	29	31	27	29	28	26	23	25	28	27	27	30	23
Night (n)	22	23	21	25	31	22	28	30	20	22	23	25	26			

Table 9.3 Production output between morning and night shifts.

	Sample mean $\bar{x}$	Sample standard deviation, $s$	Sample size, $n$
Population 1, morning	27.1250	2.3910	16
Population 2, night	24.4615	3.4548	13

Thus equation 9(x) can be rewritten as,

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_{H_0}}{\sqrt{\left( \frac{s_1^2 + s_2^2}{n_1} \right)}} \quad 9(\text{xiii})$$

The use of the Student- $t$  distribution for small samples is illustrated by the following example.

### Application of the differences in small samples: *Production output*

One part of a car production firm is the assembly line of the automobile engines. In this area of the plant, the firm employs three shifts: morning 07:00–15:00 hours, evening 15:00–23:00 hours, and the night shift 23:00–07:00 hours. The manager of the assembly line believes that the production output on the morning shift is greater than that on the night shift. Before the manager takes any action he first records the output on 16 days for the morning shift, and 13 days for the night shift. This information is given in Table 9.2.

1. At a 1% significance level, is there evidence that the output of engines on the morning shift is greater than that on the evening shift?

At a 1% significance level we are asking the question is there evidence of the output on the morning shift being greater than the output on the night shift. This is then a one-tail test with 1% in the upper tail. Using [function **TINV**] gives a critical value of Student- $t = 2.4727$ .

- The null hypothesis is that there is no difference in output,  $H_0: \mu_M = \mu_N$
- The alternative hypothesis is that the output on the morning shift is greater than that on the night shift,  $H_1: \mu_M > \mu_N$ .

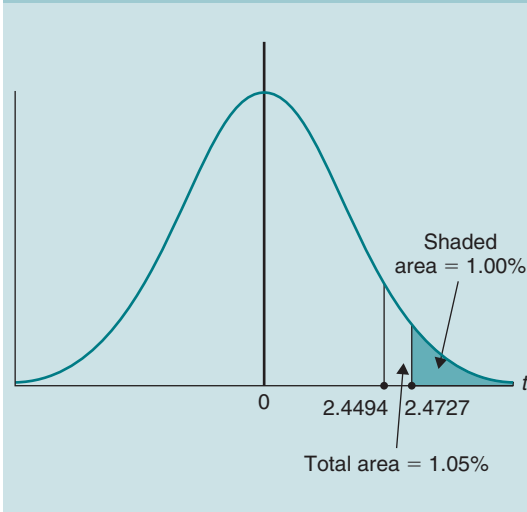
From the sample data we have the information given in Table 9.3.

From equation 9(vii),

$$\begin{aligned}
 s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \\
 &= \frac{(16 - 1) * 2.3910^2 + (13 - 1) * 3.4548^2}{(16 - 1) + (13 - 1)} \\
 &= 8.4808
 \end{aligned}$$



Figure 9.4 Production output between the morning and night shifts.



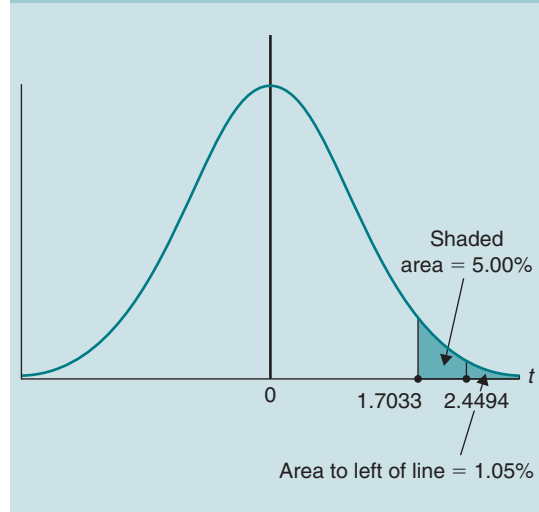
From equation 9(x) the sample or test value of the Student- $t$  value is,

$$\begin{aligned}
 t &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_{H_0}}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \\
 &= \frac{27.1250 - 24.4615 - 0}{\sqrt{8.4808 \left( \frac{1}{16} + \frac{1}{13} \right)}} \\
 &= 2.4494
 \end{aligned}$$

Since the sample test value of  $t$  of 2.4494 is less than the critical value  $t$  of 2.4727 we conclude that there is no significant difference between the production output in the morning and night shifts.

If we use the  $p$ -value approach for this hypothesis test then using [function TDIST] in Excel for a one-tail test then the area in the tail for the sample information is 1.05%. This is greater than 1.00% and so

Figure 9.5 Production output between the morning and night shifts.



our conclusion is the same in that we accept the null hypothesis.

The concept of this worked example is illustrated in Figure 9.4.

2. How would your conclusions change if a 5% level of significance were used?

In this situation nothing happens to the sample or test value of the Student- $t$  which remains at 2.4494. However, now we have 5% in the upper tail and using [function TINV] gives a critical value of Student- $t$  = 1.7033.

Since  $2.4494 > 1.7033$  we concluded that at a 5% level the production output in the morning shift is significantly greater than that in the night shift.

If we use the  $p$ -value approach for this hypothesis test then using [function TDIST] in Excel for a one-tail test, the area in the tail for the sample is still 1.05%. This is less than 5.00% and so our conclusion is the same that we reject the null hypothesis. This new concept is illustrated in Figure 9.5.

Table 9.4 Health spa and weight loss.

Before, kg (1)	120	95	118	92	132	102	87	92	115	98	109	110	95
After, kg (2)	101	87	97	82	121	87	74	84	109	87	100	101	82

## Differences of the Means Between Dependent or Paired Populations

In the previous section we discussed analysis on populations that were essentially independent of each other. In the wage example we chose samples from a population of men and a population of women. In the production output example we looked at the population of the night shift and the morning shift. Sometimes in sampling experiments we are interested in the differences of **paired samples** or those that are dependent or related, often in a before and after situation. Examples might be weight loss of individuals after a diet programme, productivity improvement after an employee training programme, or sales increases of a certain product after an advertising campaign. The purpose of these tests is to see if improvements have been achieved as a result of a new action. When we make this type of analysis we remove the effect of other variables or extraneous factors in our analysis. The analytical procedure is to consider statistical analysis on the difference of the values since there is a direct relationship rather than the values before and after. The following application illustrates.

### Application of the differences of the means between dependent samples: *Health spa*

A health spa in the centre of Brussels, Belgium advertises a combined fitness and diet programme

where it guarantees that participants who are overweight will lose at least 10 kg in 6 months if they scrupulously follow the course. The weights of all participants in the programme are recorded each time they come to the spa. The authorities are somewhat sceptical of the advertising claim so they select at random 13 of the regular participants and their recorded weights in kilograms before and after 6 months in the programme are given in Table 9.4.

1. *At a 5% significance level, is there evidence that the weight loss of participants in this programme is greater than 10 kg?*

Here the null hypothesis is that the weight loss is not more than 10 kg or  $H_0: \mu \leq 10$  kg.

The alternative hypothesis is that the weight loss is more than 10 kg, or  $H_1: \mu > 10$  kg.

We are interested not in the weights before and after but in the difference of the weights and thus we can extend Table 9.4 to give the information in Table 9.5.

The test is now very similar to hypothesis testing for a single population since we are making our analysis just on the difference.

At a significance level of 5% all of the area lies in the right-hand tail. Using **[function TINV]** gives a critical value of Student- $t = 1.7823$ . From the table,

$$\begin{aligned}\bar{x} \text{ (Difference)} &= 11.7692 \text{ kg and} \\ s &= \hat{\sigma} = 4.3999\end{aligned}$$

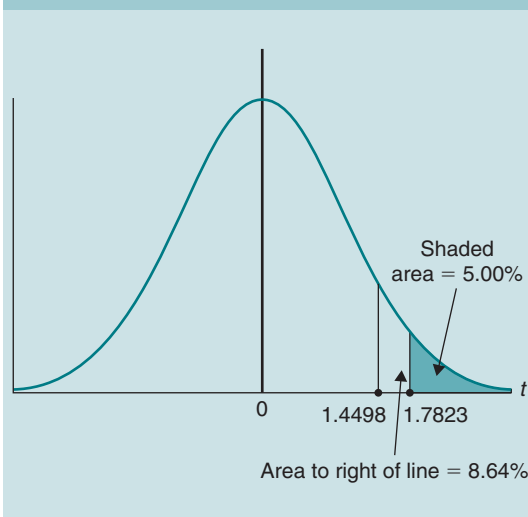
Estimated standard error of the mean is

$$\bar{\sigma}_x = \hat{\sigma}/\sqrt{n} = 4.3999/\sqrt{13} = 1.2203.$$

Table 9.5 Health spa and weight loss.

Before, kg (1)	120	95	118	92	132	102	87	92	115	98	109	110	95
After, kg (2)	101	87	97	82	121	87	74	84	109	87	100	101	82
Difference, kg	19	8	21	10	11	15	13	8	6	11	9	9	13

Figure 9.6 Health spa and weight loss.



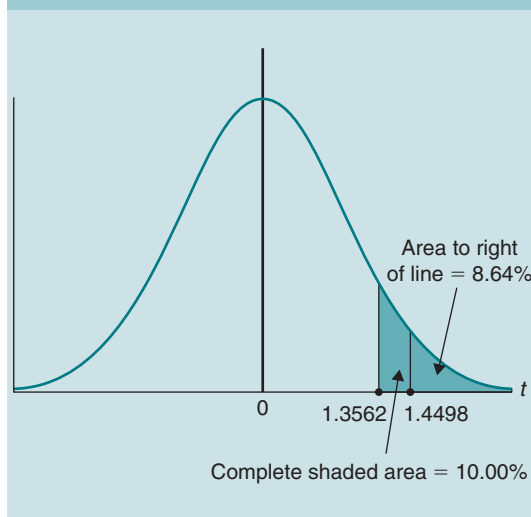
Sample, or test value of Student- $t$  is,

$$t = \frac{\bar{x} - \mu_{H_0}}{\hat{\sigma}/\sqrt{n}} = \frac{11.7692 - 10}{1.2203} = \frac{1.7692}{1.2203} = 1.4498$$

Since this sample value of  $t$  of 1.4498 is less than the critical value of  $t$  of 1.7823 we accept the null hypothesis and conclude that based on our sampling experiment that the weight loss in this programme over a 6-month period is not more than 10 kg.

If we use the  $p$ -value approach for this hypothesis test then using [function TDIST] in Excel for a one-tail test, the area in the tail for sample information is 8.64%. This is greater than 5.00% and so our conclusion is the same in that we accept the null hypothesis.

Figure 9.7 Health spa and weight loss.



The concept for this is illustrated in Figure 9.6.

2. Would your conclusions change if you used a 10% significance level?

In this case at a significance level of 10% all of the area lies in the right-hand tail and using [function TINV] gives a critical value of Student- $t = 1.3562$ . The sample or test value of the Student- $t$  remains unchanged at 1.4498. Now,  $1.4498 > 1.3562$  and thus we reject the null hypothesis and conclude that the publicity for the programme is correct and that the average weight loss is greater than 10 kg.

If we use the  $p$ -value approach for this hypothesis test then using [function TDIST] in Excel for a one-tail test, the area in the tail is still 8.64%. This is less than 10.00% and so our conclusion is the same in that we reject the null hypothesis. This concept is illustrated in Figure 9.7.

Again as in all hypotheses testing, remember that the conclusions are sensitive to the level of significance used in the test.

## Difference Between the Proportions of Two Populations with Large Samples

There are situations we might be interested to know if there is a significant difference between the proportion or percentage of some criterion of two different populations. For example, is there a significant difference between the percentage output of one firm's production site and the other? Is there a difference between the health of British people and Americans? (The answer is yes, according to a study in the Journal of the American Medical Association.<sup>2</sup>) Is there a significant difference between the percentage effectiveness of one drug and another drug for the same ailment? In these situations we take samples from each of the two groups and test for the percentage difference in the two populations. The procedure behind the test work is similar to the testing of the differences in means except rather than looking at the difference in numerical values we have the differences in percentages.

### Standard error of the difference between two proportions

In Chapter 6 (equation 6(xi)) we developed the following equation for the standard error of the proportion,  $\sigma_{\bar{p}}$ :

$$\sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}} \quad 6(\text{xi})$$

<sup>2</sup> "Compared with Americans, the British are the picture of health", *International Herald Tribune*, 22 May 2006, p. 7.

where  $n$  is the sample size,  $p$  is the population proportion of *successes*, and  $q$  is the population proportion of *failures* equal to  $(1 - p)$ . Then by analogy with equation 9(iii) for the difference in the standard error for the means we have the equation for the **standard error of the difference between two proportions** as,

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \quad 9(\text{xiv})$$

where  $p_1, q_1$  are respectively the proportion of *success* and *failure* and  $n_1$  is the sample size taken from the first population and  $p_2, q_2$ , and  $n_2$  are the corresponding values for the second population. If we do not know the population proportions then the **estimated standard error of the difference between two proportions** is,

$$\hat{\sigma}_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\bar{p}_1 \bar{q}_1}{n_1} + \frac{\bar{p}_2 \bar{q}_2}{n_2}} \quad 9(\text{xv})$$

Here,  $\bar{p}_1, \bar{q}_1, \bar{p}_2, \bar{q}_2$  are the values of the proportion of *successes* and *failures* taken from the sample.

In Chapter 8 we developed that the number of standard deviations,  $z$ , in hypothesizing for a single population proportion as,

$$z = \frac{\bar{p} - p_{H_0}}{\sigma_{\bar{p}}} \quad 8(\text{ix})$$

By analogy, the value of  $z$  for the difference in the hypothesis for two population proportions is,

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)_{H_0}}{\hat{\sigma}_{\bar{p}_1 - \bar{p}_2}} \quad 9(\text{xvi})$$

The use of these relationships is illustrated in the following worked example.

### Application of the differences of the proportions between two populations: *Commuting*

A study was made to see if there was a significance difference between the commuting time

of people working in downtown Los Angeles in Southern California and the commuting time of people working in downtown San Francisco in Northern California. The benchmark for commuting time was at least 2 hours per day. A random sample of 302 people was selected from Los Angeles and 178 said that they had a daily commute of at least 2 hours. A random sample of 250 people was selected in San Francisco and 127 replied that they had a commute of at least 2 hours.

1. At a 5% significance level, is there evidence to suggest that the proportion of people commuting Los Angeles is different from that of San Francisco?

Sample proportion of people commuting at least 2 hours to Los Angeles is,

$$p_1 = 178/302 = 0.5894 \quad \text{and}$$

$$q_1 = 1 - 0.5894 = 0.4106$$

Sample proportion of people commuting at least 2 hours to San Francisco is,

$$p_2 = 127/250 = 0.5080 \quad \text{and}$$

$$q_2 = 1 - 0.5080 = 0.4920$$

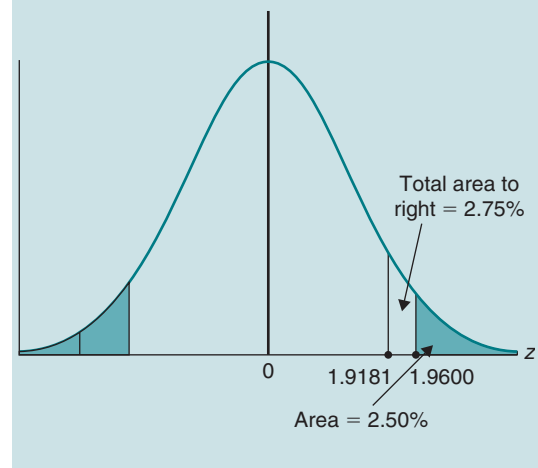
This is a two-tail test since we are asking the question is there a difference?

- Null hypothesis is that there is no difference or  $H_0: p_1 = p_2$
- Alternative hypothesis is that there is a difference or,  $H_1: p_1 \neq p_2$

From equation 9(xv) the estimated standard error of the difference between two proportions is,

$$\begin{aligned} \hat{\sigma}_{\bar{p}_1 - \bar{p}_2} &= \sqrt{\frac{\bar{p}_1 \bar{q}_1}{n_1} + \frac{\bar{p}_2 \bar{q}_2}{n_2}} \\ &= \sqrt{\frac{0.5894 * 0.4106}{302} + \frac{0.5050 * 0.4920}{250}} \\ &= 0.0424 \end{aligned}$$

Figure 9.8 Commuting time.



From equation 9(xvi) the sample value of  $z$  is,

$$\begin{aligned} z &= \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)_{H_0}}{\hat{\sigma}_{\bar{p}_1 - \bar{p}_2}} \\ &= \frac{(0.5894 - 0.5080) - 0}{0.0424} \\ &= 1.9181 \end{aligned}$$

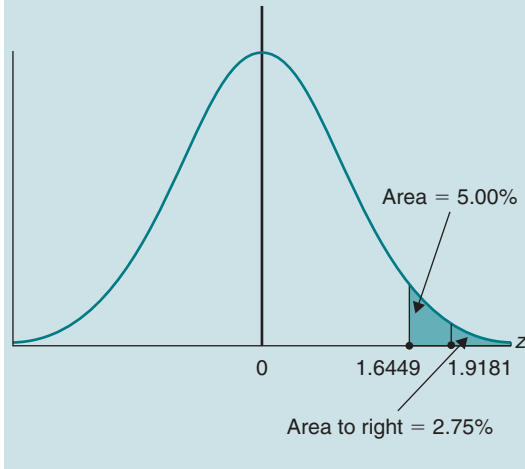
This is a two-tail test at 5% significance, so there is 2.50% in each tail. Using [function NORMSINV] gives a critical value of  $z$  of  $\pm 1.9600$ .

Since  $1.9181 < 1.9600$  we accept the null hypothesis and conclude that there is no significant difference between commuting time in Los Angeles and San Francisco.

We obtain the same conclusion when we use the  $p$ -value for making the hypothesis test. Using [function NORMSDIST] for a sample value  $z$  of 1.9181 the area in the upper tail is 2.75%. This area of 2.75%  $>$  2.50% the critical value, and so again we accept the null hypothesis. This concept is illustrated in Figure 9.8.

2. At a 5% significance level, is there evidence to suggest that the proportion of people commuting

Figure 9.9 Commuting time.



*Los Angeles is greater than those working in San Francisco?*

This is a one-tail test since we are asking the question, is one population greater than the other? Here all the 5% is in the upper tail.

- Null hypothesis is that there is a population not greater or  $H_0: p_1 \leq p_2$
- Alternative hypothesis is that a population is greater than or,  $H_1: p_1 > p_2$

Here we use  $\leq$  since less than or equal is not greater than and so thus satisfies the null hypothesis.

The sample test value of  $z$  remains unchanged at 1.9181. However, using [function NORMSDIST] the 5% in the upper tail corresponds to a critical  $z$ -value of 1.6449. Since the value of  $1.9181 > 1.6449$  we reject the null hypothesis and conclude that there is statistical evidence that the commuting time for Los Angeles people is significantly greater than for those persons in San Francisco.

Using the  $p$ -value approach, the area in the upper tail corresponding to a sample test value of 1.9181 is still 2.75%. Now this value is less than the 5% significant value

and so the conclusion is the same that there is evidence to suggest that the commuting time for those in Los Angeles is greater than for those in San Francisco. This new situation is illustrated in Figure 9.9.

## Chi-Square Test for Dependency

In testing samples from two different populations we examined the difference between either two means, or alternatively, two proportions. If we have sample data which give proportions from more than two populations then a **chi-square test** can be used to draw conclusions about the populations. The chi-square test enables us to decide whether the differences among several sample proportions is significant, or that the difference is only due to chance.

Suppose, for example, that a sample survey on the proportion of people in certain states of the United States who exercise regularly was found to be 51% in California, 34% in Ohio, 45% in New York, and 29% in South Dakota. If this difference is considered significant then a conclusion may be that location affects the way people behave. If it is not significant, then the difference is just due to chance. Thus, assuming a firm is considering marketing a new type of jogging shoe then if there is a significant difference between states, its marketing efforts should be weighted more on the state with a higher level of physical fitness. The chi-square test will be demonstrated as follows using a situation on work schedule preference.

## Contingency table and chi-square application: Work schedule preference

We have already presented a contingency or cross-classification table, in Chapter 2. This table

Table 9.6 Work preference sample data or observed frequencies,  $f_o$ .

Preference	United States	Germany	Italy	England	Total
8 hours/day	227	213	158	218	816
10 hours/day	93	102	97	92	384
Total	320	315	255	310	1,200

presents data by cross-classifying variables according to certain criteria of interest such that the cross-classification accounts for all contingencies in the sampling data. Assume that a large multinational company samples its employees in the United States, Germany, Italy, and England using a questionnaire to discover their preference towards the current 8-hour/day, 5-day/week work schedule and a proposed 10-hour/day, 4-day/week work schedule.

The sample data collected using an employee questionnaire is in Table 9.6. In this contingency table, the columns give preference according to country and the rows give the preference according to the work schedule criteria. These sample values are the observed frequencies of occurrence,  $f_o$ . This is a  $2 \times 4$  contingency table as there are two rows and four columns. Neither the row totals, nor the column totals are considered in determining the dimension of the table. In order to test whether preference for a certain work schedule depends on the location, or there is simply no dependency, we test using a **chi-square distribution**.

## Chi-square distribution

The chi-square distribution is a continuous probability distribution and like the Student- $t$  distribution there is a different curve for each degree of freedom,  $v$ . The  $x$ -axis is the value of chi-square, written  $\chi^2$  where the symbol  $\chi$  is the Greek letter  $c$ . Since we are dealing with  $\chi^2$ , or  $\chi$  to the power of two, the values on the  $x$ -axis are always positive and extend from zero to infinity. The  $y$ -axis is the

frequency of occurrence,  $f(\chi^2)$  where this probability density function is given by,

$$f(\chi^2) = \frac{1}{[(v/2 - 1)!]} \frac{1}{2^{v/2}} (\chi^2)^{(v/2-1)} e^{-\chi^2/2} \quad 9(\text{xvii})$$

Figure 9.10 gives three chi-square distributions for degrees of freedom,  $v$ , of respectively 4, 8, and 12. For small values of  $v$  the curves are positively or right skewed. As the value of  $v$  increases the curve takes on a form similar to a normal distribution. The mode or the peak of the curve is equal to the degrees of freedom less two. For example, for the three curves illustrated, the peak of each curve is for values of  $\chi^2$  equal to 2, 6, and 10, respectively.

## Degrees of freedom

The **degrees of freedom in a cross-classification table** are calculated by the relationship,

$$(\text{Number of rows} - 1) * (\text{Number of columns} - 1) \quad 9(\text{xviii})$$

Consider Table 9.7 which is a  $3 \times 4$  contingency table as there are three rows and four columns.  $R_1$  through  $R_3$  indicate the rows and  $C_1$  through  $C_4$  indicates the columns. The row totals are given by  $TR_1$  through  $TR_3$  and the column totals by  $TC_1$  through  $TC_4$ . The value of the row totals and the column totals are fixed and the “yes” or “no” in the cells indicate whether or not we have the freedom to choose a value in this cell. For example, in the column designated



Figure 9.10 Chi-square distribution for three different degrees of freedom.

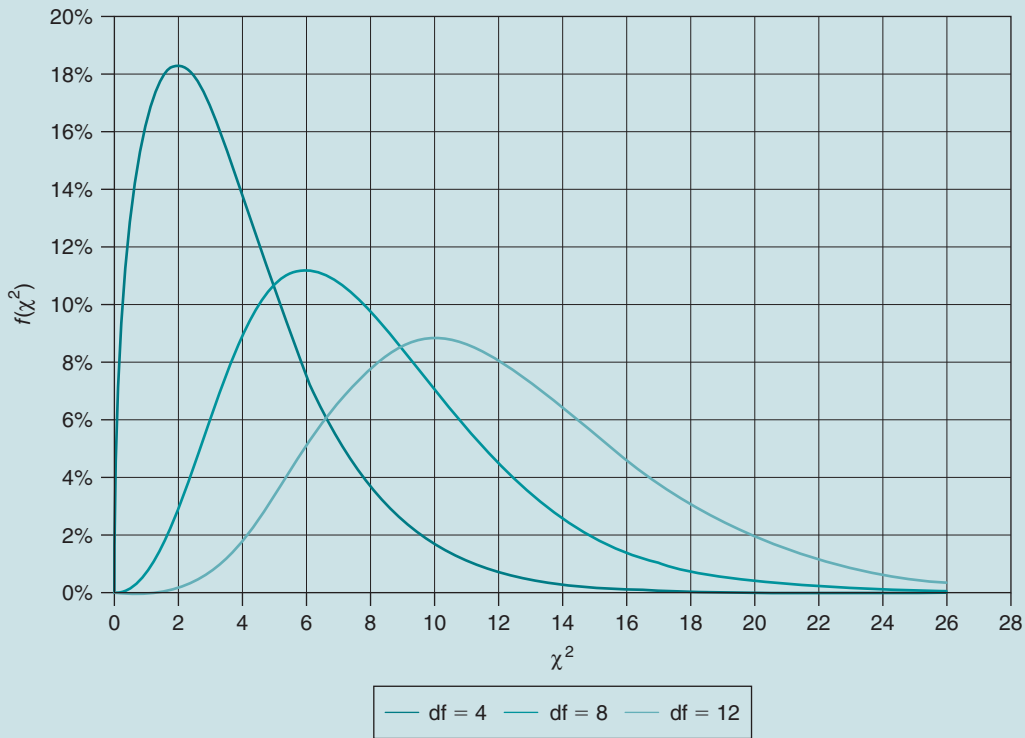


Table 9.7 Contingency table.

	C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	C <sub>4</sub>	Total rows
R <sub>1</sub>	yes	yes	yes	no	TR1
R <sub>2</sub>	yes	yes	yes	no	TR2
R <sub>3</sub>	no	no	no	no	TR3
Total columns	TC <sub>1</sub>	TC <sub>2</sub>	TC <sub>3</sub>	TC <sub>4</sub>	TOTAL

by  $C_1$  we have only the freedom to choose two values, the third value is automatically fixed by the total of that column. The same logic applies to the rows. In this table we have the freedom to choose only six values or the same as determined from equation 9(x).

$$\begin{aligned}\text{Degrees of freedom} &= (3 - 1) * (4 - 1) \\ &= 2 * 3 = 6\end{aligned}$$

### Chi-square distribution as a test of independence

Going back to our cross-classification on work preferences in Table 9.6 let us say that,

$p_U$  is the proportion in the United States who prefer the present work schedule

$p_G$  is the proportion in Germany who prefer the present work schedule



Table 9.8 Work preference – expected frequencies,  $f_e$ .

Preference	United States	Germany	Italy	England	Total
8 hours/day	217.60	214.20	173.40	210.80	816.00
10 hours/day	102.40	100.80	81.60	99.20	384.00
Total	320.00	315.00	255.00	310.00	1,200.00

$p_I$  is the proportion in the Italy who prefer the present work schedule

$p_E$  is the proportion in England who prefer the present work schedule

The null hypothesis  $H_0$  is that the population proportion favouring the current work schedule is not significantly different from country to country and thus we can write the null hypothesis situation as follows:

$$H_0: p_U = p_G = p_I = p_E \quad 9(\text{xix})$$

This is also saying that for the null hypothesis of the employee preference of work schedule is independent of the country of work. Thus, the chi-square test is also known as a test of independence.

The alternative hypothesis is that population proportions are not the same and that the preference for the work schedule is dependent on the country of work. In this case, the alternative hypothesis  $H_1$  is written as,

$$H_1: p_U \neq p_G \neq p_I \neq p_E \quad 9(\text{xx})$$

Thus in hypothesis testing using the chi-square distribution we are trying to determine if the population proportions are independent or dependent according to a certain criterion, in this case the country of employment. This test determines frequency values as follows.

## Determining the value of chi-square

From Table 9.6 if the null hypothesis is correct and that there is no difference in the preference

for the work schedule, then from the sample data.

- Population proportion who prefer the 8-hour/day schedule is  $816/1,200 = 0.6800$
- Population proportion who prefer the 10-hour/day schedule is  $384/1,200 = 0.3200$

We then use these proportions on the sample data to estimate the population proportion that prefer the 8-hour/day or the 10-hour/day schedule. For example, the sample size for the United States is 320 and so assuming the null hypothesis, the estimated number that prefers the 8-hour/day schedule is  $0.6800 * 320 = 217.60$ . The estimated number that prefers the 10-hour/day schedule is  $0.3200 * 320 = 102.40$ . This value is also given by  $320 - 217.60 = 102.40$  since the choice is one schedule or the other. Thus the complete expected data, on the assumption that the null hypothesis is correct is as in Table 9.8. These are then considered expected frequencies,  $f_e$ .

Another way of calculating the expected frequency is from the relationship,

$$f_e = \frac{TR_o * TC_o}{n} \quad 9(\text{xxi})$$

$TR_o$  and  $TC_o$  are the total values for the rows and columns for a particular observed frequency  $f_o$  in a sample of size  $n$ . For example, from Table 9.6 let us consider the cell that gives the observed frequency for Germany for a preference of an 8-hour/day schedule.

Table 9.9 Work preference – observed and expected frequencies.

	$f_o$	$f_e$	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
	227	217.60	9.40	88.36	0.4061
	213	214.20	-1.20	1.44	0.0067
	158	173.40	-15.40	237.16	1.3677
	218	210.80	7.20	51.84	0.2459
	93	102.40	-9.40	88.36	0.8629
	102	100.80	1.20	1.44	0.0143
	97	81.60	15.40	237.16	2.9064
	92	99.20	-7.20	51.84	0.5226
Total	1,200	1,200.00	0.00	757.60	6.3325

$$TR_o = 816 \quad TC_o = 315 \quad n = 1,200$$

Thus,

$$f_e = \frac{TR_o * TC_o}{n} = \frac{816 * 315}{1,200} = 214.20$$

The value of chi-square,  $\chi^2$ , is given by the relationship,

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \quad 9(\text{xxii})$$

where  $f_o$  is the frequency of the observed data and  $f_e$  is the frequency of the expected or theoretical data. Table 9.9 gives the detailed calculations.

Thus from this information in Table 9.9 the value of the sample chi-square as shown is,

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 6.3325$$

Note in order to verify that your calculations are correct, the total amount in the  $f_o$  column must equal to total in the  $f_e$  column and also the total  $(f_o - f_e)$  must be equal to zero.

## Excel and chi-square functions

In Microsoft Excel there are three functions that are used for chi-square testing.

**[function CHIDIST]** This generates the area in the chi-distribution when you enter the chi-square value and the degrees of freedom of the contingency table.

**[function CHIINV]** This generates the chi-square value when you enter the area in the chi-square distribution and the degrees of freedom of the contingency table.

**[function CHITEST]** This generates the area in the chi-square distribution when you enter the observed frequency and the expected frequency values assuming the null hypothesis.

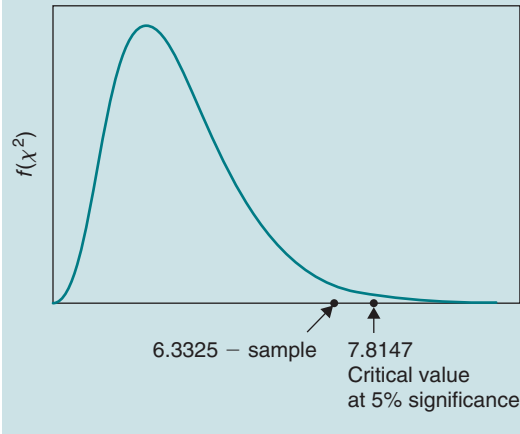
## Testing the chi-square hypothesis for work preference

As for all hypothesis tests we have to decide on a significance level to test our assumption. Let us say for the work preference situation that we consider 5% significance. In addition, for the chi-square test we also need the degrees of freedom. In Table 9.6 we have two rows and four columns, thus the degrees of freedom for this table is,

$$\begin{aligned} \text{Degrees of freedom} &= (2 - 1) * (4 - 1) \\ &= 1 * 3 = 3 \end{aligned}$$

Using Excel **[function CHIINV]** for 3 degrees of freedom, a significance level of 5% gives us a critical value of the chi-square value of 7.8147.

Figure 9.11 Chi-square distribution for work preferences.



The positions of this critical value and the value of the sample or test chi-square value are shown in Figure 9.11. Since the value of the sample chi-square statistic, 6.3325, is less than the critical value of 7.8147 at the 5% significance level given, we accept the null hypothesis and say that there is no statistical evidence to conclude that the preference for the work schedule is significantly different from country to country.

We can avoid performing the calculations shown in Table 9.9 by using first from Excel [function **CHITEST**]. In this function we enter the observed frequency values  $f_o$  as shown in Table 9.6 and the expected frequency values  $f_e$  as given in Table 9.8. This then gives the value 0.0965 or 9.65% which is the area in the chi-square distribution for the observed data. We then use [function **CHIINV**] and insert the value 9.65% and the degrees of freedom to give the sample chi-square value of 6.3325.

### Using the $p$ -value approach for the hypothesis test

In the previous paragraph we indicated that if we use [function **CHITEST**] we obtain the value

Figure 9.12 Chi-square distribution for work preferences.

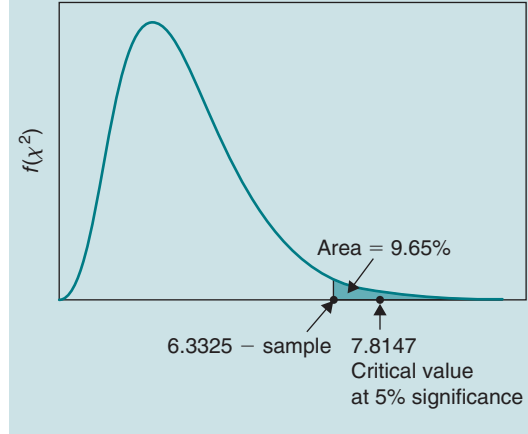
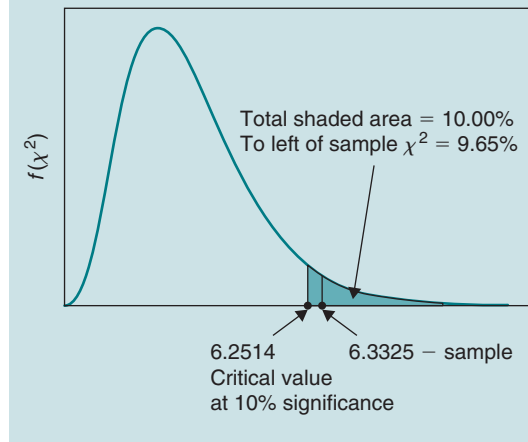


Figure 9.13 Chi-square distribution for work preferences.



9.65%, which is the area in the chi-square distribution. This is also the  $p$ -value for the observed data. Since 9.65% is greater than 5.00% the significance level we accept the null hypothesis or the same conclusion as before. This concept is illustrated in Figure 9.12.

## Changing the significance level

For the work preference situation we made the hypothesis test at 5% significance. What if we increased the significance level to 10%? In this case nothing happens to our sampling data and we still have the following information that we have already generated.

- Area under the chi-square distribution represented by the sampling data is 9.65%.
- Sample chi-square value is 6.3325.

Using [function `CHIINV`] for 10% significance and 3 degrees of freedom gives a chi-square value of 6.2514. Now since,  $6.3325 > 6.2514$  (using chi-square values), alternatively  $9.65\% < 10.00\%$  (the  $p$ -value approach).

We reject the null hypothesis and conclude that the country of employment has some bearing on the preference for a certain work schedule. This new relationship is illustrated in Figure 9.13.

This chapter has dealt with extending hypothesis testing to the difference in the means of two independent populations and the difference in the means of two dependent or paired populations. It also looks at hypothesis testing for the differences in the proportions of two populations. The last part of the chapter presented the chi-square test for examining the dependency of more than two populations. In all cases we propose a null hypothesis  $H_0$  and an alternative hypothesis  $H_1$  and test to see if there is statistical evidence whether we should accept, or reject, the null hypothesis.

## Difference between the mean of two independent populations

The difference between the mean of two independent populations is a test to see if there is a significant difference between the two population parameters such as the wages between men and women, employee productivity in one country and another, the grade point average of students in one class or another, etc. In these cases we may not be interested in the mean values of one population but in the difference of the mean values of both populations. We develop first a probability distribution of the difference in the sample means. From this we determine the standard deviation of the distribution by combining the standard deviation of each sample using either the population standard deviations, if these are known, or if they are not known, using estimates of the population standard deviation measured from the samples. From the sample test data we determine the sample  $z$ -value and compare this to the  $z$ -value dictated by the given significance level  $\alpha$ . Alternatively, we can make the hypothesis test using the  $p$ -value approach and the conclusion will be the same. When we have small sample sizes our analytical approach is similar except that we use a pooled sampled variance and the Student- $t$  distribution for our analytical tool.

## Differences of the means between dependent or paired populations

This hypothesis test of the differences between paired samples has the objective to see if there are measured benefits gained by the introduction of new programmes such as employee training to improve productivity or to increase sales, fitness programmes to reduce weight or increase stamina, coaching courses to increase student grades, etc. In these type of hypothesis

test we are dealing with the same population in a before and after situation. In this case we measure the difference of the sample means and this becomes our new sampling distribution. The hypothesis test is then analogous to that for a single population. For large samples we use a  $z$ -value for our critical test and a Student- $t$  distribution for small sample sizes.

### Difference between the proportions of two populations with large samples

This hypothesis test is to see if there is a significant difference between the proportion or percentage of some criterion of two different populations. The test procedure is similar to the differences in means except rather than measuring the difference in numerical values we measure the differences in percentages. We calculate the standard error of the difference between two proportions using a combination of data taken from the two samples based on the proportion of successes from each sample, the proportion of failures taken from each sample, and the respective sample sizes. We then determine whether the sample  $z$ -value is greater or lesser than the critical  $z$ -value. If we use the  $p$ -value approach we test to see whether the area in the tail or tails of the distribution is greater or smaller than the significance level  $\alpha$ .

### Chi-square test for dependency

The chi-square hypothesis test is used when there are more than two populations and tests whether data is dependent on some criterion. The first step is to develop a cross-classification table based on the sample data. This information gives the observed frequency of occurrence,  $f_o$ . Assuming that the null hypothesis is correct we calculate an expected value of the frequency of occurrence,  $f_e$ , using the sample proportion of successes as our benchmark. To perform the chi-square test we need to know the degrees of freedom of the cross-classification table of our sample data. This is  $(\text{number of rows} - 1) * (\text{number of columns} - 1)$ . The hypothesis test is based on the chi-square frequency distribution, which has a  $y$ -axis of frequency and a positive  $x$ -axis  $\chi^2$  extending from zero. There is a chi-square distribution for each degree of freedom of the cross-classification table. The test procedure is to see whether the sample test value  $\chi^2$  is greater or lesser than the critical value  $\chi^2$ . Alternatively we use the  $p$ -value approach and see whether the area under the curve determined from the sample data is greater or smaller than the significance level,  $\alpha$ .

## EXERCISE PROBLEMS

### 1. Gasoline prices

#### Situation

A survey of 102 gasoline stations in France in January 2006 indicated that the average price of unleaded 95 octane gasoline was €1.381 per litre with a standard deviation of €0.120. Another sample survey taken 6 months later at 97 gasoline stations indicated that the average price of gasoline was €1.4270 per litre with a standard deviation of €0.105.

#### Required

1. Indicate an appropriate null and alternative hypotheses for this situation if we wanted to know if there is a significant difference in the price of gasoline.
2. Using the critical value method, at a 2% significance level, does this data indicate that there has been a significant increase in the price of gasoline in France?
3. Confirm your conclusions to Question 2 using the  $p$ -value approach.
4. Using the critical value method would your conclusions change at a 5% significance level?
5. Confirm your conclusions to Question 4 using the  $p$ -value approach.
6. What do you think explains these results?

### 2. Tee shirts

#### Situation

A European men's clothing store wants to test if there was a difference in the price of a certain brand of tee shirts sold in its stores in Spain and Italy. It took a sample of 41 stores in Spain and found that the average price of the tee shirts was €27.80 with a variance of  $(€2.80)^2$ . It took a sample of 49 stores in Italy and found that the average price of the tee shirts was €26.90 with a variance of  $(€3.70)^2$ .

#### Required

1. Indicate an appropriate null and alternative hypotheses for this situation if we wanted to know if there is a significant difference in the price of tee shirts in the two countries.
2. Using the critical value method, at a 1% significance level, does the data indicate that there is a significant difference in the price of tee shirts in the two countries?
3. Confirm your conclusions to Question 2 using the  $p$ -value approach.
4. Using the critical value method would your conclusions change at a 5% significance level?
5. Confirm your conclusions to Question 4 using the  $p$ -value approach.
6. Indicate an appropriate null and alternative hypothesis for this situation if we wanted to test if the price of tee shirts is significantly greater in Spain than in Italy?

7. Using the critical value method, at a 1% significance level, does the data indicate that the price of tee shirts is greater in Spain than in Italy?
8. Confirm your conclusions to Question 7 using the  $p$ -value criterion?

### 3. Inventory levels

#### Situation

A large retail chain in the United Kingdom wanted to know if there was a significant difference between the level of inventory kept by its stores that are able to order on-line through the Internet with the distribution centre and those that must use FAX. The headquarters of the chain collected the following sample data from 12 stores that used direct FAX and 13 that used Internet connections for the same non-perishable items in terms of the number of days' coverage of inventory. For example, the first value for a store using FAX has a value of 14. This means that the store has on average 14 days supply of products to satisfy estimated sales until the next delivery arrives from the distribution centre.

Stores FAX	14	11	13	14	15	11	15	17	16	14	22	16	
Stores internet	12	8	14	11	6	3	15	8	7	22	19	3	4

#### Required

1. Indicate an appropriate null and alternative hypotheses for this situation if we wanted to show if those stores ordering by FAX kept a higher inventory level than those that used Internet.
2. Using the critical value method, at a 1% significance level, does this data indicate that those stores using FAX keep a higher level of inventory than those using Internet?
3. Confirm your conclusions to Question 2 using the  $p$ -value approach.
4. Using the critical value method, at a 5% significance level, does this data indicate that those stores using FAX keep a higher level of inventory than those using Internet?
5. Confirm your conclusions to Question 4 using the  $p$ -value approach.
6. How might you explain the conclusions obtained from Questions 4 and 5?

### 4. Restaurant ordering

#### Situation

A large franchise restaurant operator in the United States wanted to know if there was a difference between the number of customers that could be served if the person taking the order used a database ordering system and those that used the standard hand-written order method. In the database system when an order is taken from a customer

it is transmitted via the database system directly to the kitchen. When orders are made by hand the waiter or waitress has to go to the kitchen and give the order to the chef. Thus it takes additional time. The franchise operator believed that up to 25% more customers per hour could be served if the restaurants were equipped with a database ordering system. The following sample data was taken from some of the many restaurants within the franchise of the average number of customers served per hour per waiter or waitress.

Standard (S)	23	20	34	6	25	25	31	22	30		
Using database (D)	30	38	43	37	67	43	42	34	50	34	45

### Required

1. What is an appropriate null and alternative hypotheses for this situation?
2. Using the critical value method, at a 1% significance level, does the data support the belief of the franchise operator?
3. Confirm your conclusions to Question 2 using the  $p$ -value approach.
4. Using the same 1% significance level how could you rewrite the null and alternative hypothesis to show that the data indicates better the franchise operator's belief?
5. Test your relationship in Question 4 using the critical value method.
6. Confirm your conclusions to Question 5 using the  $p$ -value approach.
7. What do you think are reasons that some of the franchise restaurants do not have a database ordering system?

## 5. Sales revenues

### Situation

A Spanish-based ladies clothing store with outlets in England is concerned about the low store sales revenues. In an attempt to reverse this trend it decides to conduct a pilot programme to improve the sales training of its staff. It selects 11 of its key stores in the Birmingham and Coventry area and sends these sales staff progressively to a training programme in London. This training programme includes how to improve customer contact, techniques of how to spend more time on the high-revenue products, and generally how to improve team work within the store. The firm decided that it would extend the training programme to its other stores in England if the training programme increased revenues by more than 10% of revenues in its pilot stores before the programme. The table below gives the average monthly sales in £ '000s before and after the training programme. The before data is based on a consecutive 6-month period. The after data is based on a consecutive 3-month period after the training programme had been completed for all pilot stores.



Store number	1	2	3	4	5	6	7	8	9	10	11
Average sales before (£ '000s)	256	202	203	189	302	275	259	358	249	265	302
Average sales after (£ '000s)	302	289	345	259	357	299	368	402	258	267	391

### Required

1. What is the benchmark of sales revenues on which the hypothesis test programme is based?
2. Indicate the null and alternative hypotheses for this situation if we wanted to know if the training programme has reached its objective?
3. Using the critical value approach at a 1% significance level, does it appear that the objectives of the training programme have been reached?
4. Verify your conclusion to Question 3 by using the  $p$ -value approach.
5. Using the critical value approach at a 5% significance level, does it appear that the training programme has reached its objective?
6. Verify your conclusion to Question 5 by using the  $p$ -value approach.
7. What are your comments on this test programme?

## 6. Hotel yield rate

### Situation

A hotel chain is disturbed about the low yield rate of its hotels. It decides to see if improvements could be made by extensive advertising and reducing prices. It selects nine of its hotels and measures the average yield rate (rooms occupied/rooms available) in a 3-month period before the advertising, and a 3-month period after advertising for the same hotels. The data collected is given in the following table.

Hotel number	1	2	3	4	5	6	7	8	9
Yield rate before (1)	52%	47%	62%	65%	71%	59%	81%	72%	91%
Yield rate after (2)	72%	66%	75%	78%	77%	82%	89%	79%	96%

### Required

1. Indicate the null and alternative hypotheses for this situation if we wanted to know if the advertising programme has reached an objective to increase the yield rate by more than 10%?
2. Using the critical value approach at a 1% significance level, does it appear that the objectives of the advertising programme have been reached?
3. Verify your conclusion to Question 2 by using the  $p$ -value approach.
4. Using the critical value approach at a 15% significance level, does it appear that the objectives of the advertising programme have been reached?
5. Verify your conclusion to Question 4 by using the  $p$ -value approach.
6. Should management be satisfied with the results obtained?

## 7. Migraine headaches

### Situation

Migraine headaches are not uncommon. They begin with blurred vision either in one or both eyes and then are often followed by severe headaches. There are medicines available but their efficiency is often questioned. Studies have indicated that migraine is caused by stress, drinking too much coffee, or consuming too much sugar. A study was made on 10 volunteer patients who were known to be migraine sufferers. These patients were first asked to record over a 6-month period the number of migraine headaches they experienced. This was then calculated into the average number per month. Then they were asked to stop drinking coffee for 3 months and record again the number of migraine attacks they experienced. This again was reduced to a monthly basis. The complete data is in the table below.

Patient	1	2	3	4	5	6	7	8	9	10
Average number per month before (1)	23	27	24	18	31	24	23	27	19	28
Average number per month after (2)	12	18	14	5	12	12	15	12	6	14

### Required

1. Indicate the null and alternative hypothesis for this situation if we wanted to show that the complete elimination of coffee in a diet reduced the impact of migraine headaches by 50%.
2. Using the critical value approach at a 1% significance level, does it appear that eliminating coffee the objectives of the reduction in migraine headaches has been reached?
3. Verify your conclusion to Question 2 by using the  $p$ -value approach.
4. Using the critical value approach at a 10% significance level, does it appear that eliminating coffee the objectives of the reduction in migraine headaches has been reached?
5. Verify your conclusion to Question 4 by using the  $p$ -value approach.
6. At a 1% significance level, approximately what reduction in the average number of headaches has to be experienced before we can say that eliminating coffee is effective?
7. What are your comments about this experiment?

## 8. Hotel customers

### Situation

A hotel chain was reviewing its 5-year strategic plan for hotel construction and in particular whether to include a fitness room in the new hotels that it was planning to build. It had made a survey in 2001 on customers' needs and in a questionnaire of 408 people surveyed, 192 said that they would prefer to make a reservation with a hotel that had a fitness room. A similar survey was made in 2006 and out of 397 persons who returned

the questionnaire, 210 said that a hotel with a fitness room would influence booking decision.

#### Required

1. Indicate an appropriate null and alternative hypotheses for this situation.
2. Using the critical value approach at a 5% significance level, does it appear that there is a significant difference between customer needs for a fitness room in 2006 than in 2001?
3. Verify your conclusion to Question 2 by using the  $p$ -value approach.
4. Indicate the null and alternative hypotheses for this situation if we wanted to see if the customer needs for a fitness room in 2006 is greater than that in 2001.
5. Using the critical value approach at a 5% significance level, does it appear that customer needs in 2006 are greater than in 2001?
6. Verify your conclusion to Question 5 by using the  $p$ -value approach.
7. What are your comments about the results?

## 9. Flight delays

#### Situation

A study was made at major European airports to see if there had been a significant difference in flight delays in the 10-year period between 1996 and 2005. A flight was considered delayed, either on takeoff or landing, if the difference was greater than 20 minutes of the scheduled time. In 2005, in a sample of 508 flights, 310 were delayed more than 20 minutes. In 1996 out of a sample of 456 flights, 242 were delayed.

#### Required

1. Indicate an appropriate null and alternative hypothesis for this situation.
2. Using the critical value approach at a 1% significance level, does it appear that there is a significant difference between flight delays in 2005 and 1996?
3. Verify your conclusion to Question 2 by using the  $p$ -value approach.
4. Using the critical value approach at a 5% significance level, does it appear that there is a significant difference in flight delays in 2005 and 1996?
5. Verify your conclusion to Question 4 by using the  $p$ -value approach.
6. Indicate an appropriate null and alternative hypotheses for this situation to respond to the question has there been a significant increase in flight delays between 1996 and 2005?
7. From the relationship in Question 6 and using the critical value approach, what are your conclusions if you test at a significance level of 1%?
8. What has to be the significance level in order for your conclusions in Question 7 to be different?
9. What are your comments about the sample experiment?

## 10. World Cup

### Situation

The soccer World Cup tournament is held every 4 years. In June 2006 it was in Germany. In 2002 it was in Korea and Japan, and in June 1998 it was in France. A survey was taken to see if people's interest in the World Cup had changed in Europe between 1998 and 2006. A random sample of 99 people was taken in Europe in early June 1998 and 67 said that they were interested in the World Cup. In 2006 out of a sample of 112 people taken in early June, 92 said that they were interested in the World Cup.

### Required

1. Indicate an appropriate null and alternative hypotheses for this situation to test whether people's interest in the World Cup has changed between 1998 and 2006.
2. Using the critical value approach at a 1% significance level, does it appear that there is a difference between people's interest in the World Cup between 1998 and 2006?
3. Verify your conclusion to Question 2 by using the  $p$ -value approach.
4. Using the critical value approach at a 5% significance level, does it appear that there is a difference between people's interest in the World Cup between 1998 and 2006?
5. Verify your conclusion to Question 4 by using the  $p$ -value approach.
6. Indicate an appropriate null and alternative hypotheses to test whether there has been a significant increase in interest in the World Cup between 1998 and 2006?
7. From the relationship in Question 6 and using the critical value approach, what are your conclusions if you test at a significance level of 1%?
9. Confirm your conclusions to Question 7 using the  $p$ -value criterion.
10. What are your comments about the sample experiment?

## 11. Travel time and stress

### Situation

A large company located in London observes that many of its staff are periodically absent from work or are very grouchy even when at the office. Casual remarks indicate that they are stressed by the travel time into the City as their trains are crowded, or often late. As a result of these comments the human resource department of the firm sent out 200 questionnaires to its employees asking the simple question what is your commuting time to work and how do you feel your stress level on a scale of high, moderate, and low. The table below summarizes the results that it received.

Travel time	High stress level	Moderate stress level	Low stress level
Less than 30 minutes	16	12	19
30 minutes to 1 hour	23	21	31
Over 1 hour	27	25	12

### Required

1. Indicate the appropriate null hypothesis and alternative hypothesis for this situation if we wanted to test to see if stress level is dependent on travel time.
2. Using the critical value approach of the chi-square test at a 1% significance level, does it appear that there is a relationship between stress level and travel time?
3. Verify your conclusion to Question 2 by using the  $p$ -value approach of the chi-square test.
4. Using the critical value approach of the chi-square test at a 5% significance level, does it appear that there is a relationship between stress level and travel time?
5. Corroborate your conclusion to Question 4 by using the  $p$ -value approach of the chi-square test.
6. Would you say based on the returns received that the analysis is a good representation of the conditions at the firm?
7. What additional factors need to be considered when we are analysing stress (a much overused word today!)?

## 12. Investing in stocks

### Situation

A financial investment firm wishes to know if there is a relationship between the country of residence and an individual's saving strategy regarding whether or not they invest in stocks. This information would be useful as to increase the firm's presence in countries other than the United States. The following information was collected by simple telephone contact on the number of people in those listed countries on whether or not they used the stock market as their investment strategy.

Savings strategy	United States	Germany	Italy	England
Invest in stocks	206	121	147	151
Do not invest in stocks	128	118	143	141

### Required

1. Show the appropriate null hypothesis and alternative hypothesis for this situation if we wanted to test if there is a dependency between savings strategy and country of residence.
2. Using the critical value approach of the chi-square test at a 1% significance level, does it appear that there is a relationship between investing in stocks and the country of residence?
3. Verify your conclusion to Question 1 by using the  $p$ -value approach of the chi-square test.

4. Using the critical value approach of the chi-square test at a 3% significance level, does it appear that there is a relationship between investing in stocks and the country of residence?
5. Verify your conclusion to Question 3 by using the  $p$ -value approach of the chi-square test.
6. What are your observations from the sample data and what is a probable explanation?

### 13. Automobile preference

#### Situation

A market research firm in Europe made a survey to see if there was any correlation between a person's nationality and their preference in the make of automobile they purchase. The sample information obtained is in the table below.

	Germany	France	England	Italy	Spain
Volkswagen	44	27	26	19	48
Renault	27	32	24	17	32
Peugeot	22	33	22	24	27
Ford	37	16	37	25	36
Fiat	25	15	30	31	19

#### Required

1. Indicate the appropriate null and alternative hypotheses to test if the make of automobile purchased is dependent on an individual's nationality.
2. Using the critical value approach of the chi-square test at a 1% significance level, does it appear that there is a relationship between automobile purchase and nationality?
3. Verify your results to Question 2 by using the  $p$ -value approach of the chi-square test.
4. What has to be the significance level in order that there appears a breakeven situation between a dependency of nationality and automobile preference?
5. What are your comments about the results?

### 14. Newspaper reading

#### Situation

A cooperation of newspaper publishers in Europe wanted to see if there was a relationship between salary levels and the reading of a morning newspaper. A survey was made in Italy, Spain, Germany, and France and the sample information obtained is given in the table below.

Salary bracket	≤€16,000	>€16,000 to ≤€50,000	>€50,000 to ≤€75,000	>€75,000 to ≤€100,000	>€100,000
Salary category	1	2	3	4	5
Always read	36	55	65	65	62
Sometimes	44	40	47	47	52
Never read	30	28	19	19	22

### Required

1. Indicate the appropriate null and alternative hypotheses to test if reading a newspaper is dependent on an individual's salary.
2. Using the critical value approach of the chi-square test at a 5% significance level, does it appear that there is a relationship between reading a newspaper and salary?
3. Verify your results to Question 2 by using the  $p$ -value approach of the chi-square test.
4. Using the critical value approach of the chi-square test at a 10% significance level, does it appear that there is a relationship between reading a newspaper and salary?
5. Verify your results to Question 4 by using the  $p$ -value approach of the chi-square test.
6. What are your comments about the sample experiment?

## 15. Wine consumption

### Situation

A South African producer is planning to increase its export of red wine. Before it makes any decision it wants to know if a particular country, and thus the culture, has any bearing on the amount of wine consumed. Using a market research firm it obtains the following sample information on the quantity of red wine consumed per day.

Amount consumed	England	France	Italy	Sweden	United States
Never drink	20	10	15	8	12
One glass or less	72	77	70	62	68
Between one and two	85	65	95	95	48
More than two	85	79	77	85	79

### Required

1. Show the appropriate null hypothesis and alternative hypothesis for this situation if we wanted to test if there is a dependency between wine consumption and country of residence.

2. Using the critical value approach of the chi-square test at a 1% significance level, does it appear that there is a relationship between wine consumption and the country of residence?
3. Verify your conclusion to Question 2 by using the  $p$ -value approach.
4. To the nearest whole number, what has to be the minimum significance level in order to change the conclusion to Question 1? This is the  $p$ -value.
5. What is the chi-square value for the significance level of Question 4?
6. Based on your understanding of business, what is the trend in wine consumption today?

## 16. Case: Salaries in France and Germany

### Situation

Business students in Europe wish to know if there is a difference between salaries offered in France and those offered in Germany. An analysis was made by taking random samples from alumni groups in the 24–26 age group. This information is given in the table below.

<b>France</b>									
52,134	45,294	43,746	55,533	49,263	42,534	65,256	47,070	46,545	42,549
38,550	61,125	49,518	50,589	56,391	49,557	45,006	50,082	57,336	44,592
50,100	53,175	47,487	52,566	54,156	41,841	55,836	52,131	49,683	48,465
50,700	41,493	49,812	47,628	59,586	50,799	54,048	51,198	45,270	48,570
47,451	36,555	52,704	50,787	45,684	45,807	43,578	44,694	52,467	43,665
52,179	50,904	50,379	45,795	45,852	46,767	36,978	41,370	60,240	50,889
50,892	49,398	46,161	46,371	55,125	40,920	40,329	49,728	54,870	52,986
41,934	39,024	38,703	44,583	51,681	53,946	34,923	44,862	44,658	40,800
55,797	46,584	52,278	45,555	46,242	40,164	42,975	50,937	43,461	52,806
40,128	42,717	43,896	56,847	49,086	51,123	44,922	51,615	48,684	44,892
49,326	38,961	32,349	39,465	47,754	53,847	41,094	42,438	53,676	48,330
36,513	54,453	48,276	52,182	48,147	45,066	47,415	54,423	37,263	37,113
44,271	53,349	41,334	59,829	47,202	49,953	56,970	57,261	53,466	56,055
52,608	41,100	53,757	44,787	36,093	42,909	42,018	51,663	52,527	47,457
39,231	44,559	50,775	43,002	47,805	38,358	39,864	43,137	48,870	36,171
52,317	47,790	46,824	47,502	56,235	63,108	43,863	42,129	37,581	49,872
50,481	38,838	52,353	49,941	47,568	48,468	41,319	47,208	51,030	49,056
60,303	40,878	43,305	54,621	44,379	43,359	53,151	51,498	50,346	51,402
36,369	52,821	49,653	43,911	44,181	51,189	44,118	47,382	46,149	46,578
51,921	47,445	46,536	43,863	46,386	52,548	56,001	39,990	54,924	38,013
<b>Germany</b>									
45,716	48,491	53,373	49,169	62,600	44,037	52,574	41,514	46,214	47,847
40,161	48,105	50,279	51,133	52,045	38,961	37,283	47,406	45,609	52,668
43,268	41,976	51,671	53,759	51,382	41,116	51,786	54,738	55,343	48,397
60,469	43,135	44,579	54,939	50,175	43,460	49,829	55,896	59,499	56,091
43,566	41,833	44,384	48,628	46,457	46,758	39,307	54,142	38,292	63,065



52,060	38,322	54,231	37,866	54,185	55,665	56,064	44,822	44,171	58,812
44,159	51,504	53,507	59,012	50,732	55,462	48,613	53,051	50,263	52,467
38,222	42,308	59,265	53,115	35,559	46,020	56,428	40,669	48,856	46,190
41,988	43,651	55,979	40,323	44,335	48,050	43,809	44,530	43,128	45,585
46,989	52,914	57,012	46,278	53,793	59,152	51,440	38,672	42,694	42,916
52,671	52,115	40,240	53,799	55,687	52,586	55,018	49,266	47,533	48,369
45,138	50,999	43,928	46,184	49,056	33,926	43,980	54,322	54,735	59,338
33,507	51,713	57,380	41,262	52,546	44,861	47,184	46,621	50,893	52,856
55,507	45,050	44,044	47,342	58,420	41,751	60,146	43,323	48,278	58,672
41,244	49,148	42,451	47,348	48,424	47,947	41,426	42,128	63,053	41,165
49,354	42,755	43,448	50,342	55,881	53,884	49,938	48,409	50,880	40,800

### Required

1. Using all of the concepts developed from Chapters 1 to 9 how might you interpret and compare this data from the two countries?

# Forecasting and estimating from correlated data

## Value of imported goods into the United States

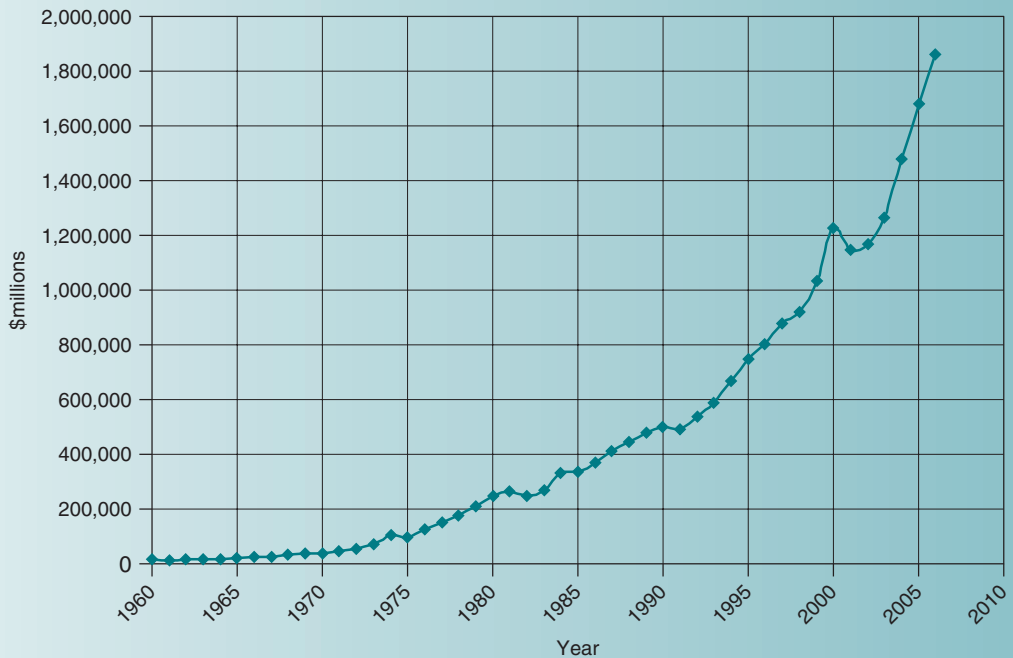
*Forecasting customer demand is a key activity in business. Forecasts trigger strategic and operations planning. Forecasts are used to determine capital budgets, cash flow, hiring or termination of personnel, warehouse space, raw material quantities, inventory levels, transportation volumes, outsourcing requirements, and the like. If we make an optimistic forecast – estimating more than actual, we may be left with excess inventory, unused storage space, or unwanted personnel. If we are pessimistic in our forecast – estimating less than actual, we may have stockouts, irritated or lost customers, or insufficient storage space. In either case there is a cost. Thus business must be accurate in forecasting. An often used approach is to use historical or collected data as the basis for forecasting on the assumption that past information is the bellwether for future activity. Consider the data in Figure 10.1, which is a time series analysis for the amount of goods imported into the United States each year from 1996 to 2006.<sup>1</sup>*

*Consider for example that we are now in the year 1970. In this case, we would say that there has been a reasonable linear growth in imported goods in the last decade from 1960. Then if we used a linear relationship for this*

---

<sup>1</sup> US Census Bureau, Foreign Trade division, [www.census.gov/foreign-trade/statistics/historical\\_goods](http://www.census.gov/foreign-trade/statistics/historical_goods), 8 June 2007.

Figure 10.1 Value of imported goods into the United States, 1960–2006.



period to forecast the value of imported goods for 2006, we would arrive at a value of \$131,050 million. The actual value is \$1,861,380 million or our forecast is low by an enormous factor of 14! As the data shows, as the years progress, there is an increasing or an almost exponential growth that is in part due to the growth of imported goods particularly from China, India, and other emerging countries many of which are destined for Wal-Mart! Thus, rather than using a linear relationship we should use a polynomial relationship on all the data or perhaps a linear regression relationship just for the period 2000–2005. Quantitative forecasting methods are extremely useful statistical techniques but you must apply the appropriate model and understand the external environment. Forecasting concepts are the essence of this chapter.

## Learning objectives

After you have studied this chapter you will understand how to **correlate** bivariate data and use **regression analysis** to make **forecasts** and **estimates** for business decisions. These topics are covered as follows:

- ✓ **A time series and correlation** • Scatter diagram • Application of a scatter diagram and correlation: *Sale of snowboards – Part I* • Coding time series data • Coefficient of correlation • Coefficient of determination • How good is the correlation?
- ✓ **Linear regression in a time series data** • Linear regression line • Application of developing the regression line using Excel: *Sale of snowboards – Part II* • Application of forecasting or estimating using Microsoft Excel: *Sale of snowboards – Part III* • The variability of the estimate • Confidence in a forecast • Alternative approach to develop and verify the regression line
- ✓ **Linear regression and causal forecasting** • Application of causal forecasting: *Surface area and house prices*
- ✓ **Forecasting using multiple regression** • Multiple independent variables • Standard error of the estimate • Coefficient of multiple determination • Application example of multiple regression: *Supermarket*
- ✓ **Forecasting using non-linear regression** • Polynomial function • Exponential function
- ✓ **Seasonal patterns in forecasting** • Application of forecasting when a seasonal pattern exists: *Soft drinks*
- ✓ **Considerations in statistical forecasting** • Time horizons • Collected data • Coefficient of variation • Market changes • Models are dynamic • Model accuracy • Curvilinear or exponential models • Selecting the best model

A useful part of statistical analysis is **correlation**, or the measurement of the strength of a relationship between variables. If there is a reasonable correlation, then **regression analysis** is a mathematical technique to develop an equation that describes the relationship between the variables in question. The practical use of this part of statistical analysis is that correlation and regression can be used to forecast sales or to make other decisions when the developed relationship from past data can be considered to mimic future conditions.

to illustrate the movement of specified variables. Financial data such as revenues, profits, or costs can be presented in a time series. Operating data for example customer service level, capacity utilization of a tourist resort, or quality levels can be similarly shown. Macro-economic data such as Gross National Product, Consumer Price Index, or wage levels are typically illustrated by a time series. In a time series we are presenting one variable, such as revenues, against another variable, time, and this is called **bivariate data**.

### A Time Series and Correlation

A time series is past data presented in regular time intervals such as weeks, months, or years

### Scatter diagram

A **scatter diagram** is the presentation of the time series data by dots on an  $x-y$  graph to see if there is a correlation between the two variables. The time, or independent variable, is presented on

Table 10.1 Sales of snowboards.

Year $x$	Sales, units $y$
1990	60
1991	90
1992	110
1993	320
1994	250
1995	525
1996	400
1997	800
1998	1,200
1999	985
2000	1,600
2001	1,550
2002	2,000
2003	2,500
2004	2,100
2005	2,400

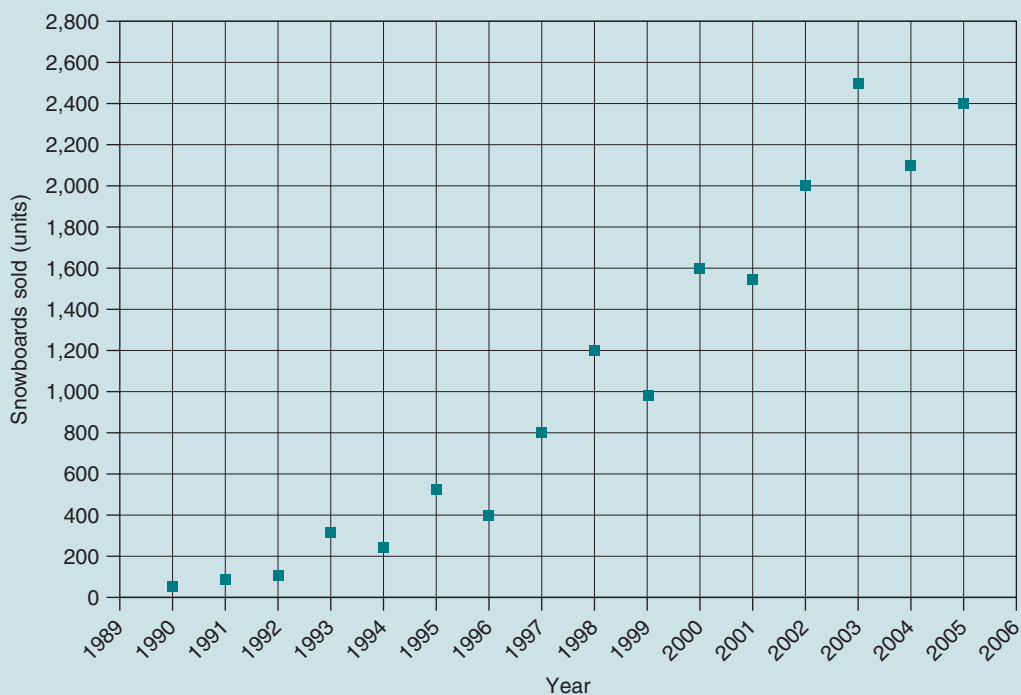
the  $x$ -axis or abscissa and the variable of interest, on the  $y$ -axis, or the ordinate. The variable on the  $y$ -axis is considered the **dependent variable** since it is “dependent”, or a function, of the time. Time is always shown on the  $x$ -axis and considered the **independent variable** since whatever happens today – an earthquake, a flood, or a stock market crash, tomorrow will always come!

### Application of a scatter diagram and correlation: *Sale of snowboards – Part I*

Consider the information in Table 10.1, which is a time series for the sales of snowboards in a sports shop in Italy since 1990.

Using in Excel the graphical command **XY(scatter)**, the scatter diagram for the data from Table 10.1 is shown in Figure 10.2. We

Figure 10.2 Scatter diagram for the sale of snowboards.



can see that it appears there is a relationship, or correlation, between the sale of snowboards, and the year in that sales are increasing over time. (Note that in Appendix II you will find a

guide of how to develop a scatter diagram in Excel.)

## Coding time series data

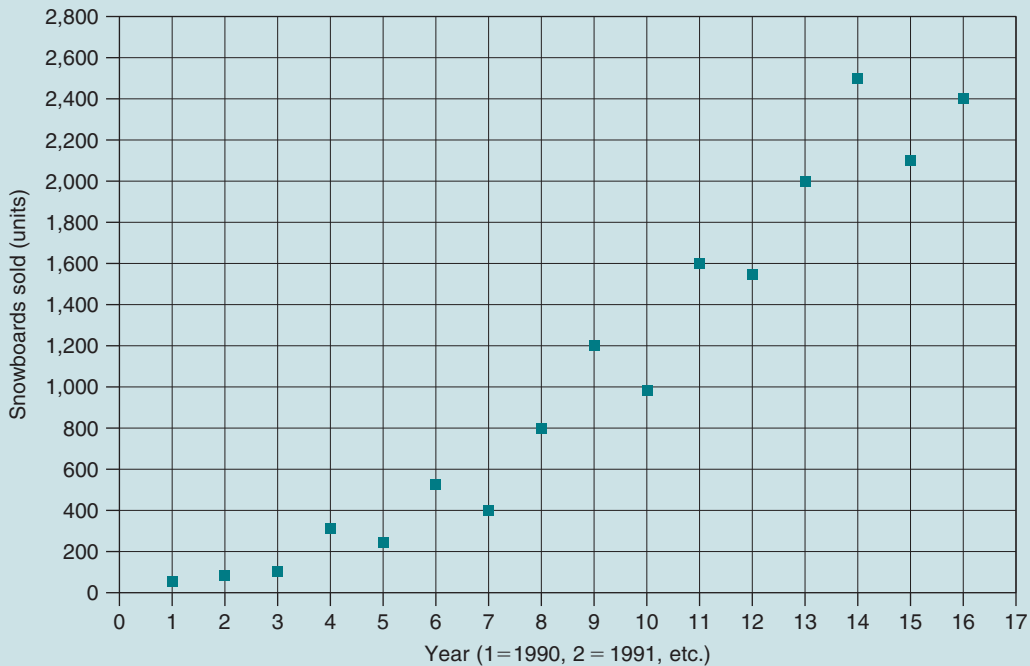
Very often in presenting time series data we indicate the time period by using **numerical codes** starting from the number 1, rather than the actual period. This is especially the case when the time is mixed alphanumeric data since it is not always convenient to perform calculations with such data. For example, a 12-month period would appear coded as in Table 10.2.

With the *snowboard sales* data calculation is not a problem since the time in years is already numerical data. However, the  $x$ -values are large and these can be cumbersome in subsequent calculations. Thus for information, Figure 10.3 gives the scatter diagram using a coded value for  $x$  where 1 = 1990, 2 = 1991, 3 = 1992, etc. The form of this scatter diagram in Figure 10.3 is identical to Figure 10.1.

**Table 10.2** Codes for time series data.

Month	Code
January	1
February	2
March	3
April	4
May	5
June	6
July	7
August	8
September	9
October	10
November	11
December	12

**Figure 10.3** Scatter diagram for the sale of snowboards using coded values for  $x$ .



## Coefficient of correlation

Once we have developed a scatter diagram, a next step is to determine the strength or the importance of the relationship between the time or independent variable  $x$ , and the dependent variable  $y$ . One measure is the **coefficient of correlation**,  $r$ , which is defined by the rather horrendous-looking equation as follows:

Coefficient of correlation,

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{\left[ n \sum x^2 - (\sum x)^2 \right] \left[ n \sum y^2 - (\sum y)^2 \right]}} \quad 10(i)$$

Here  $n$  is the number of bivariate  $(x, y)$  values. The value of  $r$  is either plus or minus and can take on any value between 0 and 1. If  $r$  is negative, it means that for the range of data given the variable  $y$  decreases with  $x$ . If  $r$  is positive, it means that  $y$  increases with  $x$ . The closer the value of  $r$  is to unity, the stronger is the relationship between the variables  $x$  and  $y$ . When

$r$  approaches zero it means that there is a very weak relationship between  $x$  and  $y$ .

The calculation steps using equation 10(i) are given in Table 10.3 using a coded value for the time period rather than using the numerical values of the year. However it is not necessary to go through this complicated procedure as the coefficient of correlation can be determined by using **[function CORREL]** in Excel. You simply enter the corresponding values for  $x$  and  $y$  where  $x$  can either be the indicated period (provided it is in numerical form) or the code value. It does not matter which, as the result is the same. In the case of the *snowboard sales* given in the example,  $r = +0.9652$ . This is close to 1.0 and thus it indicates there is a strong correlation between  $x$  and  $y$ . In Excel **[function PEARSON]** can also be used to determine the coefficient of correlation.

## Coefficient of determination

The **coefficient of determination**,  $r^2$ , is another measure of the strength of the relationship

Table 10.3 Coefficients of correlation and determination for snowboards using coded values of  $x$ .

$x$ (year)	$x$ (coded)	$y$	$xy$	$x^2$	$y^2$		
1990	1	60	60	1	3,600		
1991	2	90	180	4	8,100		
1992	3	110	330	9	12,100	$n$	16
1993	4	320	1,280	16	102,400	$n \sum xy$	3,251,200
1994	5	250	1,250	25	62,500	$\sum x \sum y$	2,297,040
1995	6	525	3,150	36	275,625	$n \sum x^2$	23,936
1996	7	400	2,800	49	160,000	$(\sum x)^2$	18,496
1997	8	800	6,400	64	640,000	$n \sum y^2$	464,912,800
1998	9	1,200	10,800	81	1,440,000	$(\sum y)^2$	285,272,100
1999	10	985	9,850	100	970,225	$n \sum xy - \sum x \sum y$	954,160
2000	11	1,600	17,600	121	2,560,000	$n \sum x^2 - (\sum x)^2$	5,440
2001	12	1,550	18,600	144	2,402,500	$n \sum y^2 - (\sum y)^2$	179,640,700
2002	13	2,000	26,000	169	4,000,000	$r$	0.9652
2003	14	2,500	35,000	196	6,250,000	$r^2$	0.9316
2004	15	2,100	31,500	225	4,410,000		
2005	16	2,400	38,400	256	5,760,000		
Total	136	16,890	203,200	1,496	29,057,050		

between  $x$  and  $y$ . Since it is the square of the coefficient of correlation,  $r$ , where  $r$  can be either negative or positive, the coefficient of determination always has a positive value. Further, since  $r$  is always equal to, or less than 1.0, numerically the value of  $r^2$ , the coefficient of determination, is always equal to or less than  $r$ , the coefficient of correlation. When  $r = 1.0$ , then  $r^2 = 1.0$  which means that there is a perfect correlation between  $x$  and  $y$ . The equation for the coefficient of determination is as follows:

Coefficient of correlation,

$$r^2 = \frac{(n \sum xy - \sum x \sum y)^2}{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]} \quad 10(ii)$$

Again we can obtain the coefficient of determination directly from Excel by using [function RSQ]. For the *snowboard sales* the value of  $r^2$  is 0.9316. Again for completeness, the calculation using equation 10(ii) is shown in Table 10.3.

## How good is the correlation?

Analysts vary on what is considered a good correlation between bivariate data. I say that if you have a value of  $r^2$  of at least 0.8, which means a value of  $r$  of about 0.9 (actually  $\sqrt{0.8} = 0.8944$ ), then there is a reasonable relationship between the independent variable and the dependent variable.

## Linear Regression in a Time Series Data

Once we have developed a scatter diagram for a time series data, and the strength of the relationship between the dependent variable,  $y$ , and the independent time variable,  $x$ , is reasonably strong, then we can develop a linear regression equation to define this relationship. After that,

we can subsequently use this equation to forecast beyond the time period given.

## Linear regression line

The **linear regression line** is the best straight line that minimizes the error between the data points on the regression line and the corresponding actual data from which the regression line is developed. The following equation represents the regression line:

$$\hat{y} = a + bx \quad 10(iii)$$

Here,

- $a$  is a **constant value** and equal to the intercept on the  $y$ -axis;
- $b$  is a **constant value** and equal to the slope of the regression line;
- $x$  is the time and the independent **variable value**;
- $\hat{y}$  is the predicted, or forecast value, of the actual dependent variable,  $y$ .

The values of the constants  $a$  and  $b$  can be calculated by the **least squares method** using the following two relationships:

$$a = \frac{\sum x^2 \sum y - \sum x \sum xy}{n \sum x^2 - (\sum x)^2} \quad 10(iv)$$

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} \quad 10(v)$$

Another approach is to calculate  $b$  and  $a$  using the average value of  $x$  or  $\bar{x}$ , and the average value of  $y$  or  $\bar{y}$  using the two equations below. It does not matter which we use as the result is the same:

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2} \quad 10(vi)$$

$$a = \bar{y} - b\bar{x} \quad 10(vii)$$



Table 10.4 Regression constants for snowboards using coded value of  $x$ .

$x$ (year)	$x$ (coded)	$y$	$xy$	$x^2$	$y^2$		
1990	1	60	60	1	3,600		
1991	2	90	180	4	8,100	$n$	16
1992	3	110	330	9	12,100	$\Sigma x$	136
1993	4	320	1,280	16	102,400	$\Sigma y$	16,890
1994	5	250	1,250	25	62,500	$\Sigma x^2$	1,496
1995	6	525	3,150	36	275,625	$\Sigma xy$	203,200
1996	7	400	2,800	49	160,000	$n \Sigma x^2$	23,936
1997	8	800	6,400	64	640,000	$(\Sigma x)^2$	18,496
1998	9	1,200	10,800	81	1,440,000	$n \Sigma xy$	3,251,200
1999	10	985	9,850	100	970,225	$a$ using equation 10(iv)	-435.2500
2000	11	1,600	17,600	121	2,560,000	$b$ using equation 10(v)	175.3971
2001	12	1,550	18,600	144	2,402,500	$\bar{x}$	8.5000
2002	13	2,000	26,000	169	4,000,000	$\bar{y}$	1,055.6250
2003	14	2,500	35,000	196	6,250,000	$b$ using equation 10(vi)	175.3971
2004	15	2,100	31,500	225	4,410,000	$a$ using equation 10(vii)	-435.2500
2005	16	2,400	38,400	256	5,760,000		
Total	136	16,890	203,200	1,496	29,057,050		
Average	8.5000	1,055.6250					

The calculations using these four equations are given in Table 10.4 for the *snowboard sales* using the coded values for  $x$ . However, again it is not necessary to perform these calculations because all the relationships can be developed from Microsoft Excel as explained in the next section.

### Application of developing the regression line using Excel: *Sale of snowboards – Part II*

Once we have the scatter diagram for the bivariate data we can use Microsoft Excel to develop the regression line. To do this we first select the data points on the scatter diagram and then proceed as follows:

- In the menu select of Excel, select **Chart**
- Select **Add trend line**

- Select **Type**
- Select **Linear**
- Select **Options** and check *Display equation on chart* and *Display R-squared value on chart*

This final window is shown in Figure E-7 of the Appendix II.

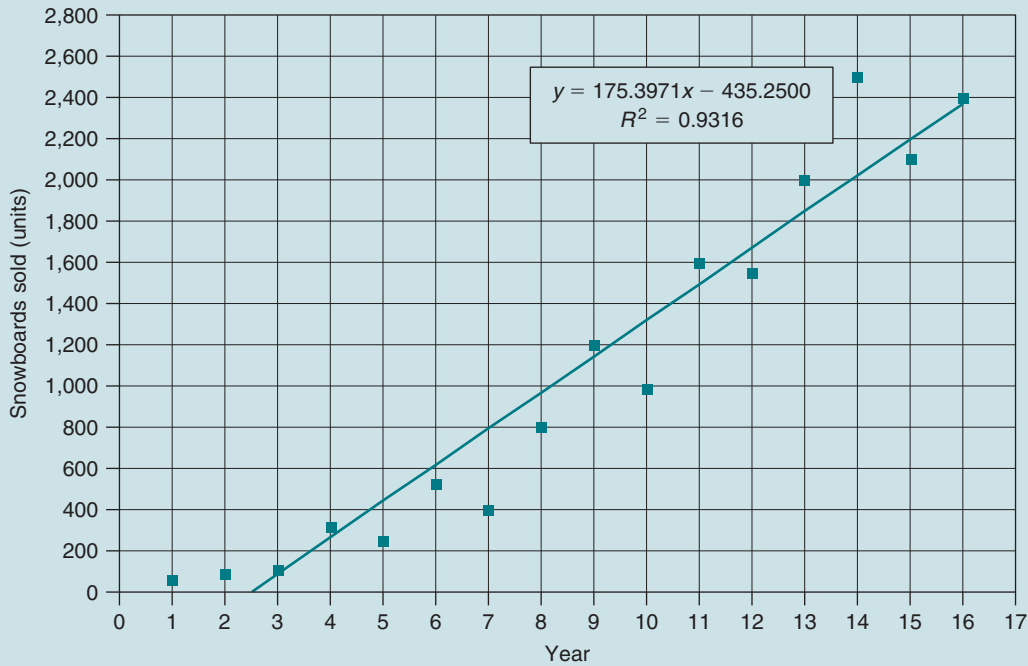
The regression line using the coded values of  $x$  is shown in Figure 10.4.

On the graph we have the regression line written as follows, which is a different form as represented by equation 10(iii). This is the Microsoft Excel format.

$$y = 175.3971x - 435.2500$$

In the form of equation 10(iii) it would be reversed and written as,

$$\hat{y} = -435.2500 + 175.3971x$$

Figure 10.4 Regression line for the sale of snowboards using coded value of  $x$ .

However, the regression information is the same where  $y$  is  $\hat{y}$ , and the slope of the line,  $b$ , is 175.3971 and,  $a$ , the intercept on the  $y$ -axis is  $-435.2500$ . These numbers are the same as calculated and presented in Figure 10.4. The slope of the line means that the sale of snowboards increases by 175.3971 (say about 175 units) per year. The intercept,  $a$ , means that when  $x$  is zero the sales are  $-432.25$  units which has no meaning for this situation. The coefficient of determination, 0.9316, which appears on the graph, is the same as previously calculated though note that Microsoft Excel uses upper case  $R^2$  rather than the lower case  $r^2$ . When the value of  $a$  is negative, but the slope of the curve is positive, it is normal to show the above equations for this example in the form  $\hat{y} = 175.3971x - 435.2500$  rather than

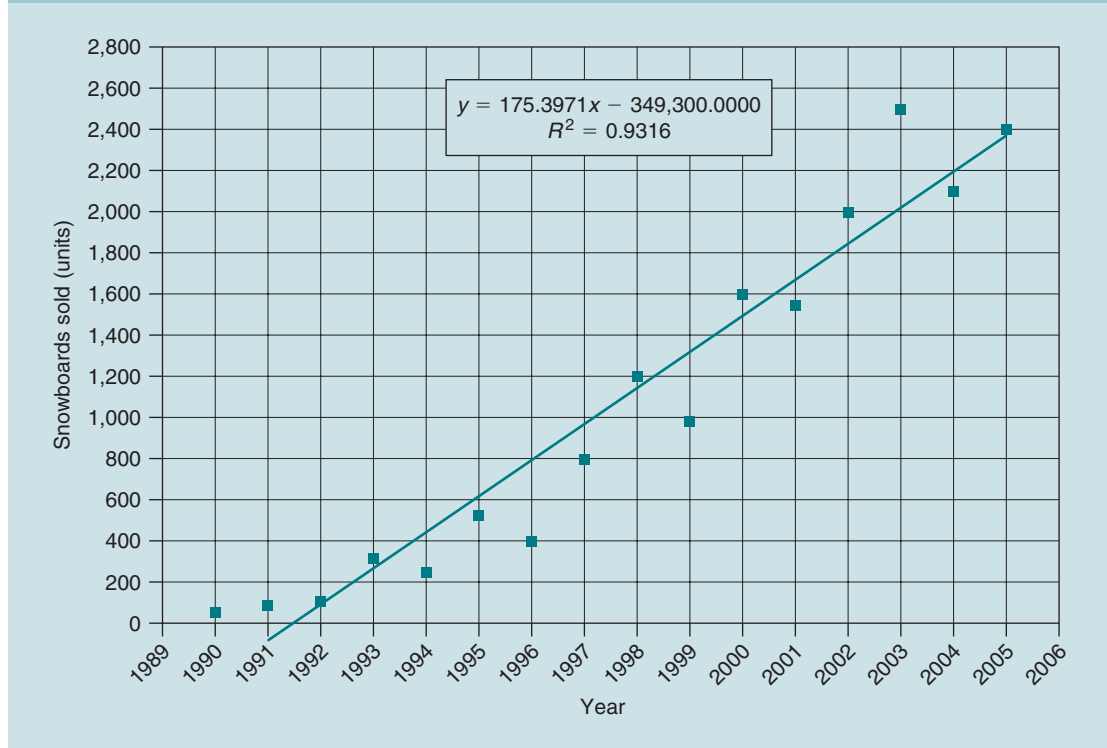
$\hat{y} = -435.2500 + 175.3971x$ . That is, avoid starting an equation with a negative value.

The regression line using the actual values of the year is shown in Figure 10.5. The only difference from Figure 10.4 is the value of the intercept  $a$ . This is because the values of  $x$  are the real values and not coded values.

### Application of forecasting, or estimating, using Microsoft Excel: *Sale of snowboards – Part III*

If we are satisfied that there is a reasonable linear relationship between  $x$  and  $y$  as evidenced by the scatter diagram, then we can forecast or estimate a future value at a given date using in Excel [function FORECAST]. For example, assume that we want to forecast the sale of

Figure 10.5 Regression line for the sale of snowboards using actual year.



snowboards for 2010. We enter into the function menu the  $x$ -value of 2010 and the given values of  $x$  in years and the given value of  $y$  from Table 10.1. This gives a forecast value of  $y$  of 3,248 units. Alternatively, we can use the coded values of  $x$  that appear in the 2nd column of Table 10.2 and the corresponding actual data for  $y$ . If we do this, we must use a code value for the year 2010, which in this case is 21. (Year 2005 has a code of 16, thus year 2010 =  $16 + 5 = 21$ .) Note that in any forecasting using time series data, the assumption is that the pattern of past years will be repeated in future years, which may not necessarily be the case. Also, the further out we go in time, the less accurate will be the forecast. For example, a forecast of sales for next year may be reasonably reliable, whereas a forecast 20 years from now would not.

### The variability of the estimate

In Chapter 2, we presented the sample standard deviation,  $s$ , of data by the equation,

Sample standard deviation,

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x - \bar{x})^2}{(n - 1)}} \quad 2(\text{viii})$$

The standard deviation is a measure of the variability around the sample mean,  $\bar{x}$ , for each random variable  $x$ , in a given sample size,  $n$ . Further, the deviation of all the observations,  $x$ , about the mean value  $\bar{x}$  is zero (equation 2(ix)), or,

$$\sum (x - \bar{x}) = 0 \quad 2(\text{ix})$$

Table 10.5 Statistics for the regression line.

$b$ , slope of the line	175.3971	−435.2500	$a$ , intercept on the $y$ -axis
	12.7000	122.8031	
$r^2$ , coefficient of determination	0.9316	234.1764	$s_e$ standard error of estimate
	190.7380	14	degrees of freedom ( $n - 2$ )
	10,459,803.6029	767,740.1471	

In a similar manner, a measure of the variability around the regression line is the **standard error of the estimate**,  $s_e$ , given by,

Standard error of the estimate,

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}} \quad 10(\text{viii})$$

Here  $n$  is the number of bivariate data points ( $x, y$ ). The value of  $s_e$  has the same units of the dependant variable  $y$ . The denominator in this equation is  $(n - 2)$  or the number of degrees of freedom, rather than  $(n - 1)$  in equation 2(viii). In equation 10(viii) two degrees of freedom are lost because two statistics,  $a$  and  $b$ , are used in regression to compute the standard error of the estimate. Like the standard deviation, the closer to zero is the value of the standard error then there is less scatter or deviation around the regression line. If this is the case, this translates into saying that the linear regression model is a good fit of the observed data, and we should have reasonable confidence in the estimate or forecast made. The regression equation, is determined so that the vertical distance between the observed, or data values,  $y$ , and the predicted values,  $\hat{y}$ , balance out when all data are considered. Thus, analogous to equation 2(ix) this means that,

$$\sum (y - \hat{y}) = 0 \quad 10(\text{ix})$$

Again, we do not have to go through a stepwise calculation but the standard error of the estimate, together with other statistical information, can be determined by using in Excel **[function LINEST]**. To do this we select a cellblock of two columns by five rows and enter the given  $x$ - and  $y$ -values and input 1 both times for the constant data. Like the frequency distribution we execute this function by pressing simultaneously on control-shift-enter (Ctrl-⬆-↵).

The statistics for the regression line for the snowboard data are given in Table 10.5. The explanations are given to the left and to the right of each column. Those that we have discussed so far are highlighted and also note in this matrix that we have again the value of  $b$ , the slope of the line; the value  $a$ , the intercept on the  $y$ -axis, and the coefficient of determination,  $r^2$ . We also have the degrees of freedom, or  $(n - 2)$ . The other statistics are not used here but their meaning in the appropriate format is indicated in Table E-3 of Appendix II.

## Confidence in a forecast

In a similar manner to confidence limits in estimating presented in Chapter 7, we can determine the **confidence limits of a forecast**. If we have a sample size greater than 30 then the confidence intervals are given by,

$$\hat{y} \pm z s_e \quad 10(\text{x})$$

**Table 10.6** Calculating the standard estimate of the regression line using coded values of  $x$ .

Code	$x$	$y$	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	1990	60	-259.85	319.85	102,305.90
2	1991	90	-84.46	174.46	30,434.85
3	1992	110	90.94	19.06	363.24
4	1993	320	266.34	53.66	2,879.58
5	1994	250	441.74	-191.74	36,762.42
6	1995	525	617.13	-92.13	8,488.37
7	1996	400	792.53	-392.53	154,079.34
8	1997	800	967.93	-167.93	28,199.30
9	1998	1,200	1,143.32	56.68	3,212.22
10	1999	985	1,318.72	-333.72	111,369.43
11	2000	1,600	1,494.12	105.88	11,211.07
12	2001	1,550	1,669.51	-119.51	14,283.76
13	2002	2,000	1,844.91	155.09	24,052.36
14	2003	2,500	2,020.31	479.69	230,103.62
15	2004	2,100	2,195.71	-95.71	9,159.62
16	2005	2,400	2,371.10	28.90	835.04
Total				0.00	767,740.15
$s_e$					234.18
$n$	16				

With sample sizes no more than 30, we use a Student- $t$  relationship and the confidence limits are,

$$\hat{y} \pm ts_e \quad 10(\text{xi})$$

For our *snowboard* sales situation we have a forecast of 3,248 units for 2010. To obtain a confidence level, we use a Student- $t$  relationship since we have a sample size of 16. For a 90% confidence limit, using [function TINV], where the degrees of freedom are given in Table 10.5, the value of  $t$  is 1.7613. Then using equation 10(xi) and the standard error from Table 10.5 the confidence limits are as follows:

Lower limit is

$$3,248 - 1.7613 * 234.1764 = 2,836$$

Upper limit is

$$3,248 + 1.7613 * 234.1764 = 3,361$$

Thus to better define our forecast we could say that our best estimate of snowboard sales in

2010 is 3,248 units and that we are 90% confident that the sales will be between 2,836 and 3,361 units.

### Alternative approach to develop and verify the regression line

Now that we have determined the statistical values for the regression line, as presented in Table 10.5, we can use these values to develop the specific values of the regression points and further to verify the standard error of the estimate,  $s_e$ . The calculation steps are shown in Table 10.6. The column  $\hat{y}$  is calculated using equation 10(iii) and imputing the constant values of  $a$  and  $b$  from Table 10.5. The total of  $(y - \hat{y})$  in Column 5 verifies equation 10(ix). And, using the total value of  $(y - \hat{y})^2$  in Column 6, the last column of Table 10.6, and inserting this in equation 10(viii) verifies the value of the standard error of the estimate of Table 10.5.

## Linear Regression and Causal Forecasting

In the previous sections we discussed correlation and how a dependent variable changed with time. Another type of correlation is when one variable is dependent, or a function, not of time, on some other variable. For example, the sale of household appliances is in part a function of the sale of new homes; the demand for medical services increases with an aging population; or for many products, the quantity sold is a function of price. In these situations we say that the movement of the dependent variable,  $y$ , is caused by the change of the dependent variable,  $x$  and the correlation can be used for **causal forecasting** or estimating. The analytical approach is very similar to linear regression for a time series except that time is replaced by another variable. The following example illustrates this.

### Application of causal forecasting: Surface area and house prices

In a certain community in Southern France, a real estate agent has recorded the past sale of houses according to sales price and the square metres of living space. This information is in Table 10.7.

1. Develop a scatter diagram for this information. Does there appear to be a reasonable correlation between the price of homes, and the square metres of living space?

Here this is a causal relationship where the price of the house is a function, or is “caused” by the square metres of living space. Thus, the square metres is the independent variable,  $x$ , and the house price is the dependent variable  $y$ . Using the same approach as for the previous snowboard example in a time series analysis, Figure 10.6 gives the scatter diagram for this causal relationship. Visually

Table 10.7 Surface area and house prices.

Square metres, $x$	Price (€) $y$
100	260,000
180	425,000
190	600,000
250	921,000
360	2,200,000
200	760,500
195	680,250
110	690,250
120	182,500
370	2,945,500
280	1,252,500
450	5,280,250
425	3,652,000
390	3,825,240
60	140,250
125	280,125

it appears that within the range of the data given, the house prices generally increase linearly with square metres of living space.

2. Show the regression line and the coefficient of determination on the scatter diagram. Compute the coefficient of correlation. What can you say about the coefficients of determination and correlation? What is the slope of the regression line and how is it interpreted?

The regression line is shown in Figure 10.7 together with the coefficient of determination. The relationships are as follows:

Regression equation,

$$\hat{y} = -1,263,749.9048 + 11,646.6133x$$

Coefficient of determination,

$$r^2 = 0.8623$$

Coefficient of correlation,

$$r = \sqrt{r^2} = \sqrt{0.8623} = 0.9286$$

Since the coefficient of determination is greater than 0.8, and thus the coefficient of correlation is greater than 0.9 we can say that there is quite a strong correlation

Figure 10.6 Scatter diagram for surface area and house prices.

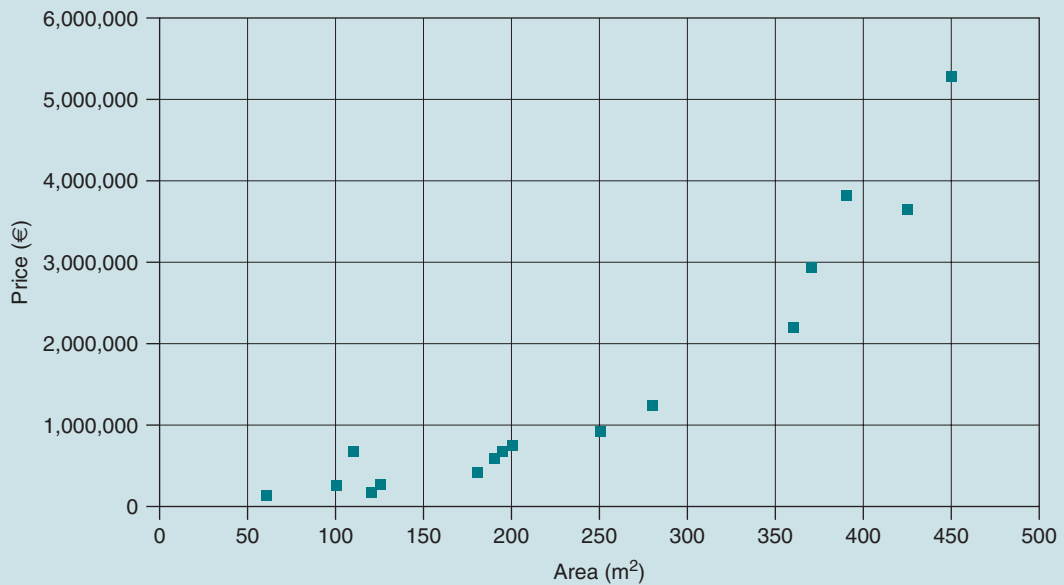


Figure 10.7 Regression line for surface area and house prices.

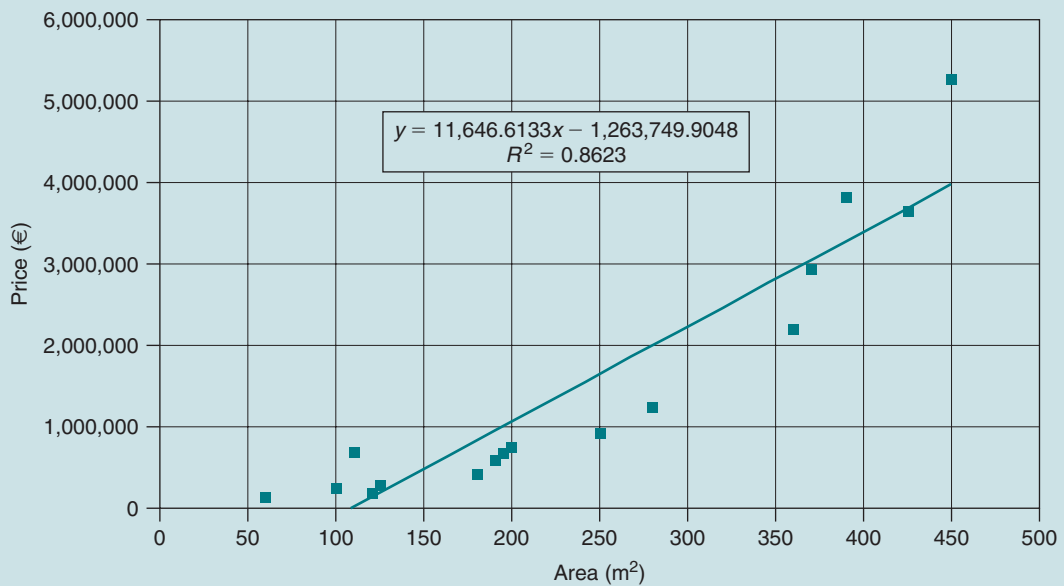


Table 10.8 Regression statistics for surface area and house prices.

$b$ , slope of the line	11,646.6133	-1,263,749.9048	$a$ , intercept on the $y$ -axis
	1,244.0223	332,772.3383	
$r^2$ , coefficient of determination	0.8623	609,442.0004	$s_e$ standard error of estimate
	87.6482	14	degrees of freedom ( $n - 2$ )
	$3.2554 * 10^{13}$	$5.1999 * 10^{12}$	

between house prices and square metres of living space.

The slope of the regression line is 11,646.6133 (say 11,650); this means to say that for every square metre in living space, the price of the house increases by about €11,650 within the range of the data given.

3. If a house on the market has a living space of  $310 \text{ m}^2$ , what would be a reasonable estimate of the price? Give the 85% confidence intervals for this price.

Using in Excel [function FORECAST] for a square metre of living space,  $x$  of  $310 \text{ m}^2$  gives an estimated price (rounded) of €2,346,700.

Using in Excel [function LINEST] we have in Table 10.8 the statistics for the regression line.

Using [function TINV] in Excel, where the degrees of freedom are given in Table 10.8, the value of  $t$  for a confidence level of 85% is 1.5231.

Using equation 10(xi),

$$\hat{y} \pm ts_e$$

Lower limit of price estimate using the standard error of the estimate from Table 10.8 is,

$$\begin{aligned} & \text{€}2,346,700 - 1.5231 * 609,444 \\ & = \text{€}1,418,463 \end{aligned}$$

Upper limit is,

$$\begin{aligned} & \text{€}2,346,700 + 1.5231 * 609,444 \\ & = \text{€}3,274,938 \end{aligned}$$

Thus we could say that a reasonable estimate of the price of a house with  $310 \text{ m}^2$  living space is €2,346,700 and that we are 85% confident that the price lies in the range €1,418,463 (say €1,418,460) to €3,274,938 (say €3,274,940).

4. If a house was on the market and had a living space of  $800 \text{ m}^2$ , what is a reasonable estimate for the sales price of this house? What are your comments about this figure?

Using in Excel [function FORECAST] for a square metre of living space,  $x$  of  $800 \text{ m}^2$  gives an estimated price (rounded) of €8,053,541.

The danger with making this estimate is that  $800 \text{ m}^2$  is outside of the limits of our observed data (it ranges from 60 to  $450 \text{ m}^2$ ). Thus the assumption that the linear regression equation is still valid for a living space area of  $800 \text{ m}^2$  may be erroneous. Thus you must be careful in using causal forecasting beyond the range of data collected.

## Forecasting Using Multiple Regression

In the previous section on causal forecasting we considered the relationship between just one dependent variable and one independent variable.



Multiple regression takes into account the relationship of a dependent variable with more than one independent variable. For example, in people, obesity, the dependent variable, is a function of the quantity we eat and the amount of exercise we do. Automobile accidents are a function of driving speed, road conditions, and levels of alcohol in the blood. In business, sales revenues can be a function of advertising expenditures, number of sales staff, number of branch offices, unit prices, number of competing products on the market, etc. In this situation, the forecast estimate is a causal regression equation containing several independent variables.

## Multiple independent variables

The following is the equation that describes the **multiple regression** model:

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k \quad 10(\text{xii})$$

Here,

- $a$  is a constant and the intercept on the  $y$ -plane;
- $x_1, x_2, x_3$ , and  $x_k$  are the independent variables;
- $b_1, b_2, b_3$  and  $b_k$  are constants and slopes of the line corresponding to  $x_1, x_2, x_3$ , and  $x_k$ ;
- $\hat{y}$  is the forecast or predicted value given by the best fit for the actual data;
- $k$  is a value equal to the number of independent variables in the model.

Since there are more than two variables in the equation we cannot represent this function on a two-dimensional graph. Also note that the more the number of independent variables in the relationship then the more complex is the model, and possibly the more uncertain is the predicted value.

## Standard error of the estimate

As for linear regression, there is a standard error of the estimate  $s_e$  that measures the

degree of dispersion around the multiple regression plane. It is as follows:

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - k - 1}} \quad 10(\text{xiii})$$

Here,

- $y$  is the actual value of the dependant variable;
- $\hat{y}$  is the corresponding predicted value of dependant variable from the regression equation;
- $n$  is the number of bivariate data points;
- $k$  is the number of independent variables.

This is similar to equation 10(viii) for linear regression except that there is now a term  $k$  in the denominator where the value  $(n - k - 1)$  is the degrees of freedom. As an illustration, if the number of bivariate data points  $n$  is 16, and there are four independent variables then the degrees of freedom are  $16 - 4 - 1 = 11$ . In linear regression, with the same 16 bivariate data values, the number of independent variables,  $k$ , is 1, and so the degrees of freedom are  $16 - 1 - 1 = 14$  or the denominator as given by equation 10(xiii). Again, these values of the degrees of freedom are automatically determined in Excel when you use **[function LINEST]**. As before, the smaller the value of the standard error of the estimate, the better is the fit of the regression equation.

## Coefficient of multiple determination

Similar to linear regression there is a coefficient of multiple determination  $r^2$  that measures the strength of the relationship between all the independent variables and the dependent variable. The calculation of this is illustrated in the following worked example.

## Application example of multiple regression: Supermarket

A distributor of Nestlé coffee to supermarkets in Scandinavia visits the stores periodically to

meet the store manager to negotiate shelf space and to discuss pricing and other sales-related activities. For one particular store the distributor had gathered the data in Table 10.9 regarding the unit sales for a particular size of instant coffee, the number of visits made to that store, and the total shelf space that was allotted.

1. From the information in Table 10.9 develop a two-independent variable multiple regression model for the unit sales per month as a function of the visits per month and the allotted shelf space. Determine the coefficient of determination.

As for times series linear regression and causal forecasting, we can use again from Excel [function **LINEST**]. The difference is that we

now select a virgin area of three rows and five columns and we enter two columns for the independent variable  $x$ , visits per month and the shelf space. The output from using this function is in Table 10.10.

The statistics that we need from this table are in the shaded cells and are as follows:

- $a$ , the intercept on the  $y$ -plane = 14,227.67;
- $b_1$ , the slope corresponding to  $x_1$ , the visits per month = 4,827.01;
- $b_2$ , the slope corresponding to the shelf space,  $x_2 = 9,997.64$ ;
- $s_e$ , the standard error of the estimate = 5,938.51;
- Coefficient of determination,  $r^2 = 0.9095$ ;
- Degrees of freedom,  $df = 7$ .

Again, the other statistics in the non-shaded areas are not used here but their meaning, in the appropriate format, are indicated in Table E-4 of Appendix II.

The equation, or model, that describes this relation is from equation 10(xii) for two independent variables:

$$\hat{y} = a + b_1x_1 + b_2x_2$$

$$\hat{y} = 14,227.67 + 4,827.01x_1 + 9,997.64x_2$$

As the coefficient of determination, 0.9095, is greater than 0.8 the strength of the relationship is quite good.

Table 10.9 Sales of Nestlé coffee.

Unit sales/month, $y$	Visits/month, $x_1$	Shelf space, ( $m^2$ ) $x_2$
90,150	9	3.50
58,750	4	1.75
71,250	6	2.32
63,750	5	1.82
39,425	3	1.82
55,487	6	1.50
76,975	7	2.92
74,313	6	2.92
71,813	8	2.35
33,125	2	1.35

Table 10.10 Regression statistics for sales of Nestlé coffee—two variables.

$b_2 = 9,997.64$	$b_1 = 4,827.01$	$a = 14,227.67$
4,568.23	1,481.81	6,537.83
$r^2 = 0.9095$	$s_e = 5,938.51$	#N/A
35.16	$df = 7$	#N/A
2,480,086,663.75	246,861,055.85	#N/A

Table 10.11 Sales of Nestlé coffee with three variables.

Sales, $y$	Visits/month, $x_1$	Shelf space (m <sup>2</sup> ), $x_2$	Price (€/unit), $x_3$
90,150	9	3.50	1.25
58,750	4	1.75	2.28
71,250	6	2.32	1.87
63,750	5	1.82	2.25
39,425	3	1.82	2.60
55,487	6	1.50	2.20
76,975	7	2.92	2.00
74,313	6	2.92	1.84
71,813	8	2.35	2.06
33,125	2	1.35	2.75

2. Estimate the monthly unit sales if eight visits per month were made to the supermarket and the allotted shelf space was 3.00 m<sup>2</sup>. What are the 85% confidence levels for this estimate?

Here  $x_1$  is the estimate of sales of eight visits per month, and  $x_2$  is the shelf space of 3.00 m<sup>2</sup>. The monthly sales are determined from the regression equation:

$$\begin{aligned}\hat{y} &= 14,227.67 + 4,827.01 * 8 \\ &\quad + 9,997.64 * 3.00 \\ &= 82,837 \text{ units}\end{aligned}$$

For the confidence intervals we use equation 10(xi),

$$\hat{y} \pm t_{se} \quad 10(\text{xi})$$

Using [function TINV] in Excel, where the degrees of freedom are 7 as given in Table 10.10, the value of  $t$  for a confidence level of 85% is 1.6166.

The confidence limits of sales using the standard error of the estimate of 5,938.51 from Table 10.10 are,

Lower confidence limit is

$$\begin{aligned}82,837 - 1.6166 * 5,938.51 \\ = 73,237 \text{ units}\end{aligned}$$

Upper confidence limit is,

$$\begin{aligned}82,837 + 1.6166 * 5,938.51 \\ = 92,437 \text{ units}\end{aligned}$$

Thus we can say that using this regression model our best estimate of monthly sales is 82,837 units and that we are 85% confident that the sales will be between 73,237 and 92,437 units.

3. Assume now that for the sales data in Table 10.9 the distributor looks at the unit price of the coffee sold during the period that the analysis was made. This expanded information is in Table 10.11 showing now the variation in the unit price of a jar of coffee. From this information develop a three-independent-variable multiple regression model for the unit sales per month as a function of visits per month, allotted shelf space, and the unit price of coffee. Determine the coefficient of determination.

We use again from Excel [function LINEST] and here we select a virgin area of four rows and five columns and we enter three columns for the three independent variables  $x$ , visits per month, the shelf space, and price. The output from using this function is in Table 10.12.

The statistics that we need from this table are:

- $a$ , the intercept on the  $y$ -plane = 75,658.05;
- $b_1$ , the slope corresponding to  $x_1$  the visits per month = 2,984.28;
- $b_2$ , the slope corresponding to the shelf space,  $x_2$  = 4,661.82;

**Table 10.12** Regression statistics for coffee sales – three variables.

$b_3 = -18,591.50$	$b_2 = 4,661.82$	$b_1 = 2,984.28$	$a = 75,658.05$
12,575.38	5,556.26	1,852.38	41,989.31
$r^2 = 0.9336$	$s_e = 5,491.60$	#N/A	#N/A
28.14	df = 6	#N/A	#N/A
2,546,001,747.28	180,945,972.32	#N/A	#N/A

- $b_3$ , the slope corresponding to the price,  $x_3 = -18,591.50$ ;
- $s_e$ , the standard error of the estimate = 5,491.60;
- Coefficient of determination,  $r^2 = 0.9336$ ;
- Degrees of freedom = 6.

The equation or model that describes this relation is from equation 10(xii) for three independent variables:

$$\begin{aligned}\hat{y} &= a + b_1x_1 + b_2x_2 + b_3x_3 \\ \hat{y} &= 75,658.05 + 2,984.28x_1 \\ &\quad + 4,661.82x_2 - 18,591.50x_3\end{aligned}$$

As the coefficient of determination, 0.9336, is greater than 0.8 the strength of the relationship is quite good.

4. Estimate the monthly unit sales if eight visits per month were made to the supermarket, the allotted shelf space was  $3.00 \text{ m}^2$ , and the unit price of coffee was €2.50. What are the 85% confidence levels for this estimate?

Here  $x_1$  is the estimate of sales of eight visits per month,  $x_2$  is the shelf space of  $3.00 \text{ m}^2$ , and  $x_3$  is the unit sales price of coffee of €2.50. Estimated monthly sales are determined from the regression equation,

$$\begin{aligned}\hat{y} &= 75,658.05 + 2,984.28 * 8 \\ &\quad + 4,661.82 * 3.00 - 18,591.50 * 2.50 \\ &= 67,039 \text{ units}\end{aligned}$$

For the confidence intervals we use equation 10(xi) and [function TINV] in Excel. The degrees

of freedom are 6 from Table 10.12, the value of  $t$  for a confidence level of 85% is 1.6502.

The confidence limits of sales using the standard error of the estimate of 5,491.60 from Table 10.12 are,

Lower limit is,

$$\begin{aligned}67,039 - 1.6502 * 5,491.60 \\ = 57,977 \text{ units}\end{aligned}$$

Upper limit is,

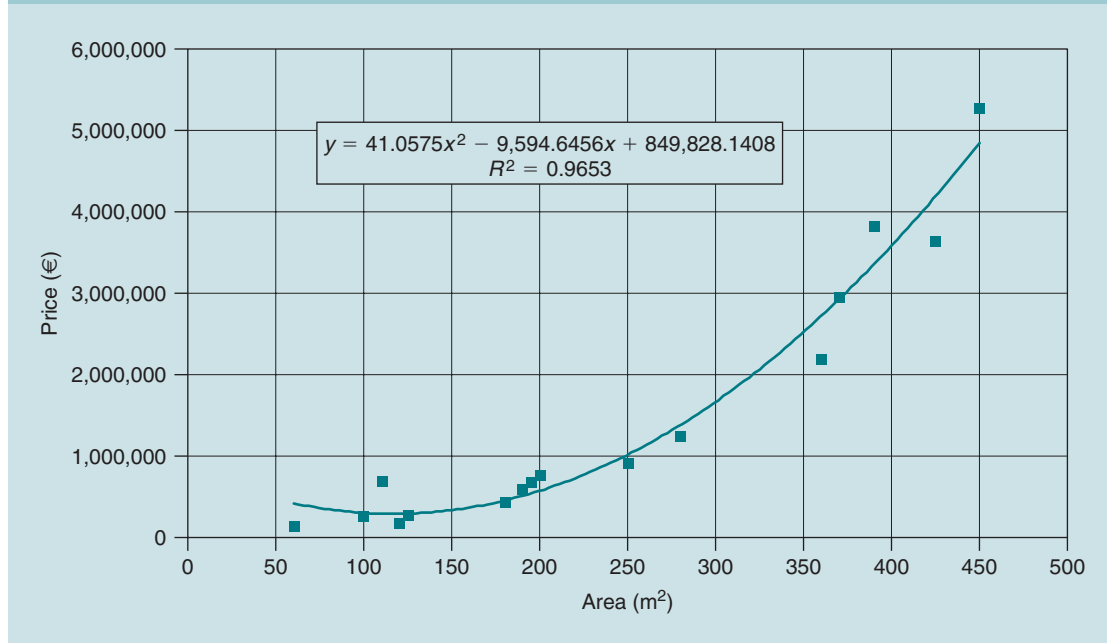
$$\begin{aligned}67,039 + 1.6502 * 5,491.60 \\ = 76,101 \text{ units}\end{aligned}$$

Thus we can say that using this regression model, the best estimate of monthly sales is 67,039 units and that we are 85% confident that the sales will be between 57,977 and 76,101 units.

## Forecasting Using Non-linear Regression

Up to this point we have considered that the dependent variable is a linear function of one or several independent variables. For some situations the relationship of the dependent variable,  $y$ , may be non-linear but a **curvilinear function** of one independent variable,  $x$ . Examples of these are: the sales of mobile phones from about 1995 to 2000; the increase of HIV contamination in

Figure 10.8 Second-degree polynomial for house prices.



Africa; and the increase in the sale of DVD players. Curvilinear relationships can take on a variety of forms as discussed below.

## Polynomial function

A **polynomial function**, takes the following general form where  $x$  is the independent variable and  $a, b, c, d, \dots, k$  are constants:

$$y = a + bx + cx^2 + dx^3 + \dots + kx^n \quad 10(\text{xiv})$$

Since we only have two variables  $x$  and  $y$  we can plot a scatter diagram. Once we have the scatter diagram for this bivariate data we can use Microsoft Excel to develop the regression line. To do this we first select the data points on the graph and then from the **[Menu chart]** proceed sequentially as follows:

- Add trend line
- Type polynomial power

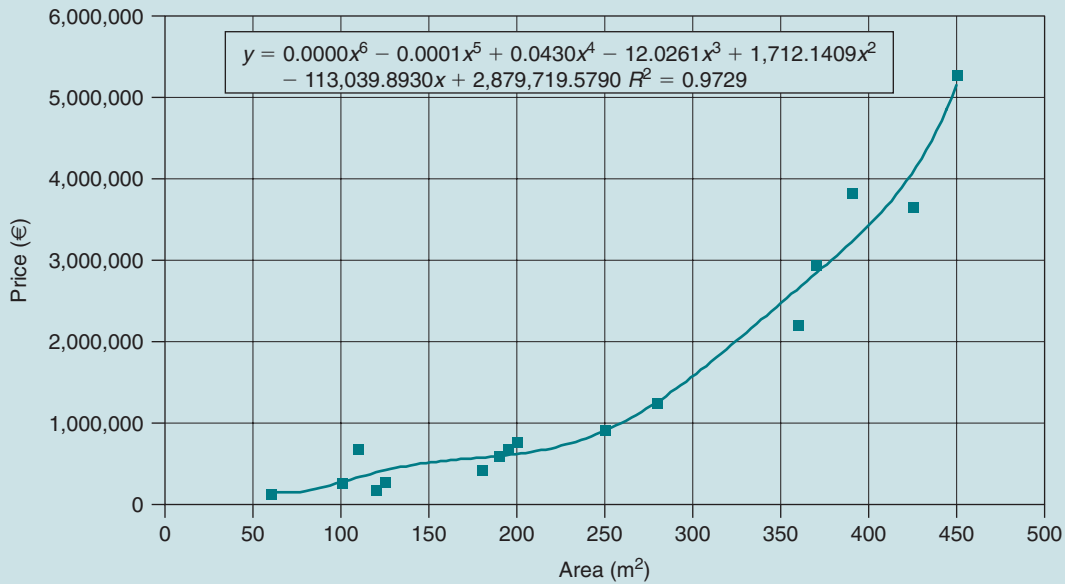
- Options
- Display equation on chart and Display R-squared value on chart.

In Microsoft Excel we have the option of a polynomial function with the powers of  $x$  ranging from 2 to 6. A second-degree or quadratic polynomial function, where  $x$  has a power of 2 for the *surface area and house price* data of Table 10.7 is given in Figure 10.8. The regression equation and the corresponding coefficient of determination are as follows:

$$\hat{y} = 41.0575x^2 - 9,594.6456x + 849,828.1408$$

$$r^2 = 0.9653$$

In Figure 10.9 we have the regression function where  $x$  has a power of 6. The regression equation

Figure 10.9 Polynomial function for house prices where  $x$  has a power of 6.

and the corresponding coefficient of determination are as follows:

$$\hat{y} = -0.0001x^5 + 0.0430x^4 - 12.0261x^3 + 1.712,1409x^2 - 113,039.8930x + 2,879,719.5790$$

$$r^2 = 0.9729$$

We can see that as the power of  $x$  increases the closer is the coefficient of determination to unity or the better fit is the model. Note for this same data when we used linear regression, Figure 10.7, the coefficient of determination was 0.8623.

## Exponential function

An **exponential function** has the following general form where  $x$  and  $y$  are the independent and dependent variables, respectively, and  $a$  and  $b$  are constants:

$$y = ae^{bx} \quad 10(xv)$$

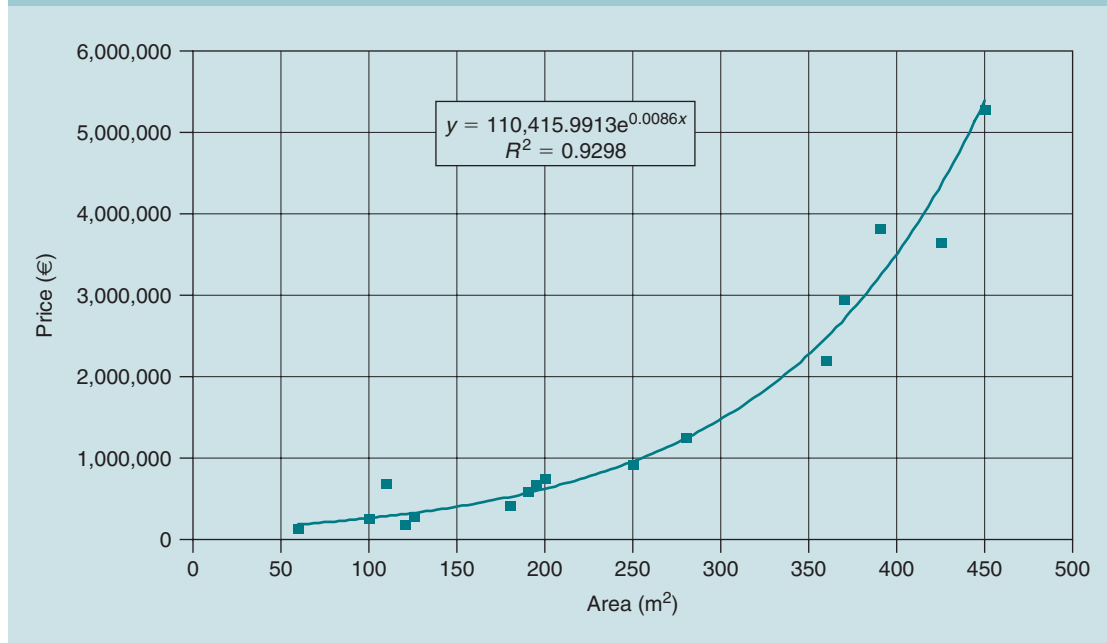
The exponential relationship for the house prices is shown in Figure 10.10 and the following is the equation with the corresponding coefficient of determination:

$$\hat{y} = 110,415.9913e^{0.0086x} \quad r^2 = 0.9298$$

## Seasonal Patterns in Forecasting

In business, particularly when selling is involved, seasonal patterns often exist. For example, in the Northern hemisphere the sale of swimwear is higher in the spring and summer than in the autumn and winter. The demand for heating oil is higher in the autumn and winter, and the sale of cold beverages is higher in the summer than in the winter. The linear regression analysis for a time series analysis, discussed

Figure 10.10 Exponential function for surface area and house prices.



early in the chapter, can be modified to take into consideration seasonal effects. The following application illustrates one approach.

### Application of forecasting when there is a seasonal pattern: *Soft drinks*

Table 10.13 gives the past data for the number of pallets of soft drinks that have been shipped from a distribution centre in Spain to various retail outlets on the Mediterranean coast.

1. Use the information in Table 10.13 to develop a forecast for 2006.

*Step 1. Plot the actual data and see if a seasonal pattern exists*

The actual data is shown in Figure 10.11 and from this it is clear that the data is seasonal.

Note that for the x-axis we have used a coded value for each season starting with winter 2000 with a code value of 1.

*Step 2. Determine a centred moving average*

A **centred moving average** is the average value around a designated centre point. Here we determine the average value around a particular season for a 12-month period, or four quarters. For example, the following relationship indicates how we calculate the centred moving average around the summer quarter (usually 15 August) for the current year  $n$ :

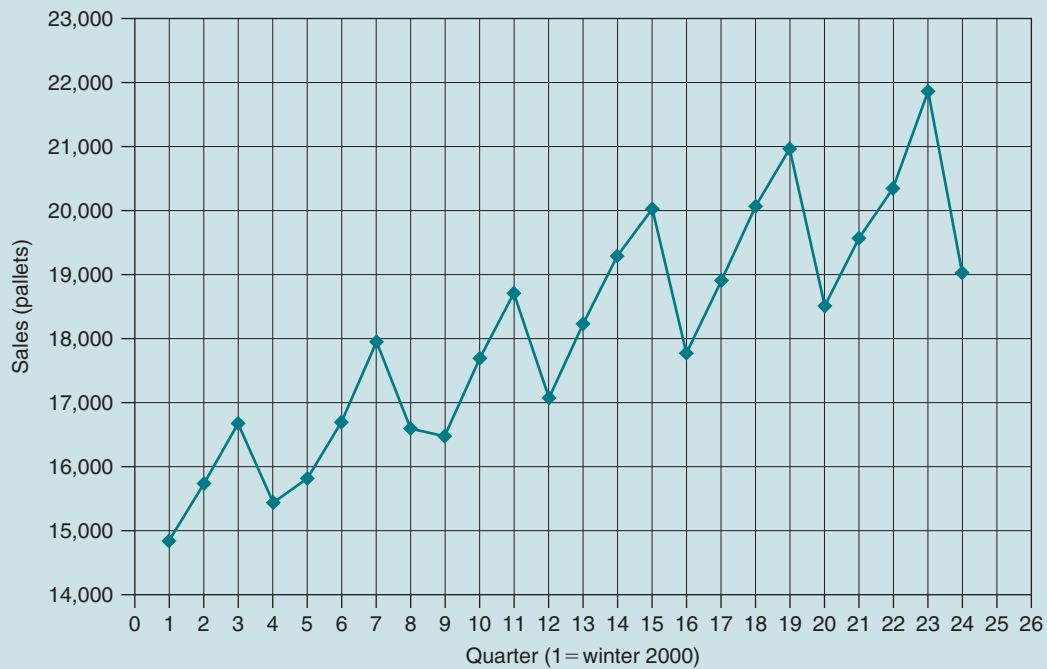
$$\frac{0.5 * \text{winter}(n) + 1.0 * \text{spring}(n) + 1.0 * \text{summer}(n) + 1.0 * \text{autumn}(n) + 0.5 * \text{winter}(n+1)}{4}$$

For example if we considered the centre period as summer 2000 then the centred

Table 10.13 Sales of soft drinks.

Year	Quarter	Actual sales (pallets)	Year	Quarter	Actual sales (pallets)
2000	Winter	14,844	2003	Winter	18,226
	Spring	15,730		Spring	19,295
	Summer	16,665		Summer	19,028
	Autumn	15,443		Autumn	17,769
2001	Winter	15,823	2004	Winter	18,909
	Spring	16,688		Spring	20,064
	Summer	17,948		Summer	19,152
	Autumn	16,595		Autumn	18,503
2002	Winter	16,480	2005	Winter	19,577
	Spring	17,683		Spring	20,342
	Summer	18,707		Summer	20,156
	Autumn	17,081		Autumn	19,031

Figure 10.11 Seasonal pattern for the sales of soft drinks.





moving average around this quarter using the actual data from Table 10.13 is as follows:

$$\frac{0.5 * 14,844 + 1.0 * 15,730 + 1.0 * 16,665 + 1.0 * 15,443 + 0.5 * 15,823}{4} = 15,792.88$$

We are determining a centred moving average and so the next centre period is autumn 2000. For this quarter, we drop the data for winter 2000 and add spring 2001 and thus

the centred moving average around autumn 2000 is as follows:

$$\frac{0.50 * 15,730 + 1.0 * 16,665 + 1.0 * 15,443 + 1.0 * 15,823 + 0.5 * 16,688}{4} = 16,035.00$$

Thus each time we move forward one quarter we drop the oldest piece of data and add the next quarter. The values for the centred moving average for the complete period are in Column 5 of Table 10.14. Note that we

Table 10.14 Sales of soft drinks – seasonal indexes and regression.

1	2	3	4	5	6	7	8	9
Year	Quarter	Code	Actual sales (pallets)	Centred moving average	$SI_p$	Seasonal index $SI$	Sales/ $SI$	Regression forecast, $\hat{y}$
2000	Winter	1	14,844			0.97	15,240.97	15,438.30
	Spring	2	15,730			1.02	15,462.69	15,669.15
	Summer	3	16,665	15,792.88	1.0552	1.06	15,719.93	15,899.99
	Autumn	4	15,443	16,035.00	0.9631	0.95	16,279.60	16,130.84
2001	Winter	5	15,823	16,315.13	0.9698	0.97	16,246.15	16,361.68
	Spring	6	16,688	16,619.50	1.0041	1.02	16,404.41	16,592.53
	Summer	7	17,948	16,845.63	1.0654	1.06	16,930.17	16,823.37
	Autumn	8	16,595	17,052.13	0.9732	0.95	17,494.00	17,054.22
2002	Winter	9	16,480	17,271.38	0.9542	0.97	16,920.72	17,285.06
	Spring	10	17,683	17,427.00	1.0147	1.02	17,382.50	17,515.91
	Summer	11	18,707	17,706.00	1.0565	1.06	17,646.12	17,746.75
	Autumn	12	17,081	18,125.75	0.9424	0.95	18,006.33	17,977.60
2003	Winter	13	18,226	18,492.38	0.9856	0.97	18,713.42	18,208.44
	Spring	14	19,295	18,743.50	1.0294	1.02	18,967.10	18,439.29
	Summer	15	20,028	18,914.88	1.0588	1.06	18,892.21	18,670.13
	Autumn	16	17,769	19,096.38	0.9305	0.95	18,731.60	18,900.98
2004	Winter	17	18,909	19,309.63	0.9793	0.97	19,414.68	19,131.82
	Spring	18	20,064	19,518.50	1.0279	1.02	19,723.03	19,362.67
	Summer	19	20,965	19,693.75	1.0646	1.06	19,776.07	19,593.51
	Autumn	20	18,503	19,812.00	0.9339	0.95	19,505.37	19,824.36
2005	Winter	21	19,577	19,958.13	0.9809	0.97	20,100.55	20,055.20
	Spring	22	20,342	20,135.50	1.0103	1.02	19,996.31	20,286.05
	Summer	23	21,856			1.06	20,616.55	20,516.89
	Autumn	24	19,031			0.95	20,061.97	20,747.74

cannot determine a centred moving average for winter and spring 2000 or for summer and autumn of 2005 since we do not have all the necessary information. The line graph for this centred moving average is in Figure 10.12.

*Step 3. Divide the actual sales by the moving average to give a period seasonal index,  $SI_p$*   
This is the ratio,

$$SI_p = \frac{(\text{Actual recorded sales in a period})}{(\text{Moving average for the same period})}$$

This data is in Column 6 of Table 10.14. What we have done here is compared actual sales to the average for a 12-month period. It gives a specific seasonal index for each month. For example, if we consider 2004 the ratios, rounded to two decimal places, are as in Table 10.15.

We interpret this by saying that sales in the winter 2004 are 2% below the year ( $1 - 0.98$ ), in the spring they are 3% above the year, 6% above the year for the summer, and 10% below the year for autumn 2004 ( $1 - 0.90$ ).

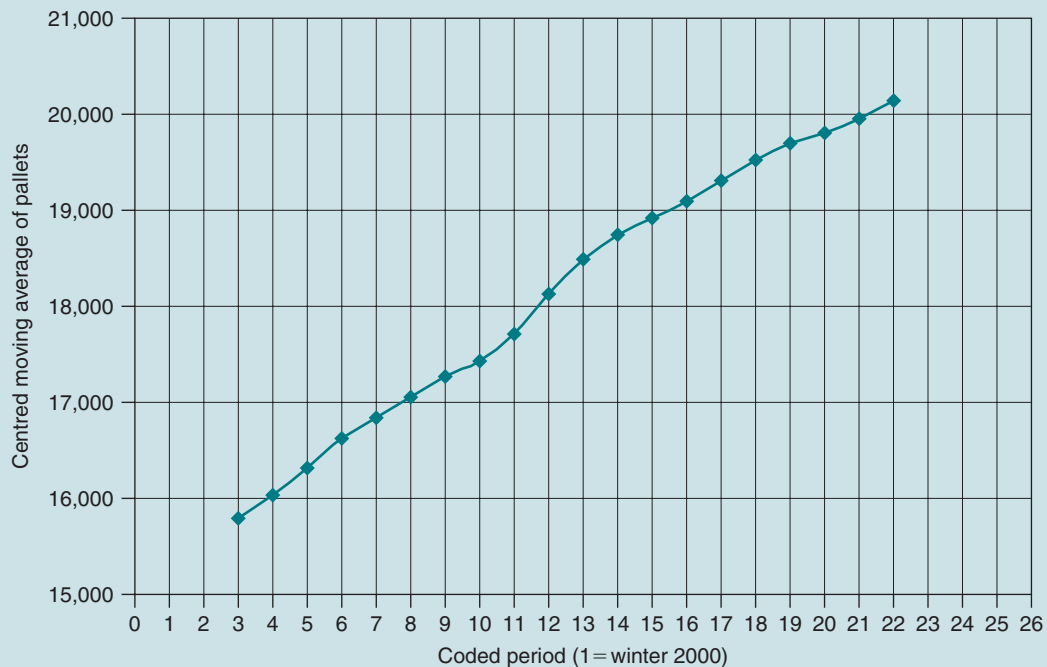
*Step 4. Determine an average seasonal index,  $SI$ , for the four quarters*

This is determined by taking the average of all the ratios,  $SI_p$  for like seasons. For example,

**Table 10.15** Sales of soft drinks – seasonal indexes.

Winter	Spring	Summer	Autumn
0.98	1.03	1.06	0.90

**Figure 10.12** Centred moving average for the sale of soft drinks.



the seasonal index for the summer is calculated as follows:

$$\frac{1.0552 + 1.0654 + 1.0565 + 1.0588 + 1.0646}{5} = 1.0601$$

The seasonal indices for the four seasons are in Table 10.16. Note that the average value of

**Table 10.16** Sales of soft drinks – seasonal indexes.

Season	SI
Summer	1.0601
Autumn	0.9486
Winter	0.9740
Spring	1.0173
Average	1.0000

these indices must be very close to unity since they represent the movement for one year.

These same indices, but rounded to two decimal places, are shown in Column 7 of Table 10.14. Note, for similar seasons, the values are the same.

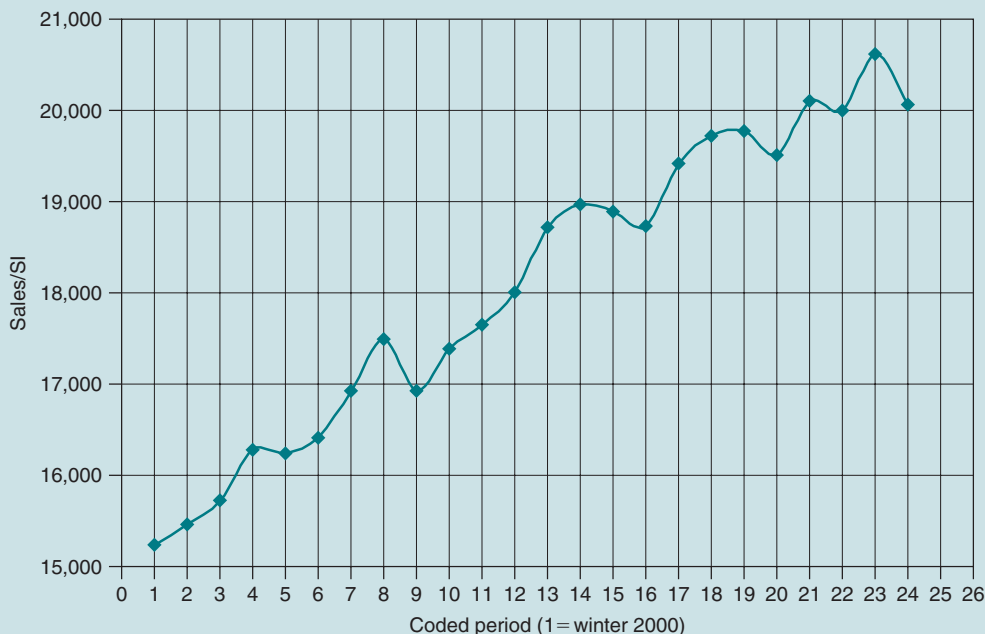
*Step 5. Divide the actual sales by the seasonal index, SI*

This data is shown in Column 8. What we have done here is removed the seasonal effect of the sales, and just showed the trend in sales without any contribution from the seasonal period. Another way to say is that the sales are deseasonalized. The line graph for these deseasonalized sales is in Figure 10.13.

*Step 6. Develop the regression line for the deseasonalized sales*

The regression line is shown in Figure 10.14. The regression equation and the

**Figure 10.13** Sales/SI for soft drinks.



corresponding coefficient of determination are as follows:

$$\hat{y} = 230.8451x + 15,207.4554$$

$$r^2 = 0.9673$$

Alternatively we can use in Excel [function **LINEST**] by entering from Table 10.11 the  $x$ -values of Column 1 and the  $y$ -values from Column 8 to give the statistics in Table 10.17.

Using the corresponding values of  $a$  and  $b$  we have developed the regression line values as shown in Column 9 of Table 10.14.

*Step 7. From the regression line forecast deseasonalized sales for the next four quarters*

This can be done in two ways. Either from the Excel table, continue the rows down for 2006 using the code values of 25 to 28 for the four seasons. Alternately, use [function

Figure 10.14 Deseasonalized sales and regression line for soft drinks.

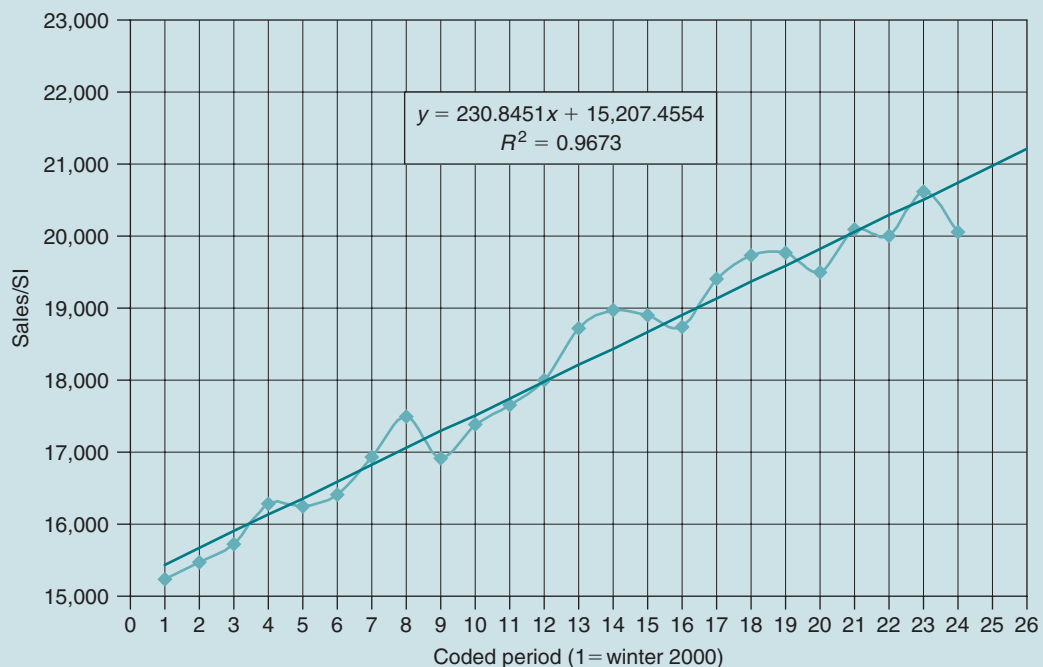


Table 10.17 Sales of soft drinks – seasonal indexes.

$b$ , slope of the line	230.8451	15,207.4554	$a$ , intercept on the $y$ -axis
	9.0539	129.3687	
$r^2$ , coefficient of determination	0.9673	307.0335	$s_e$ , standard error of estimate
	650.0810	22	degrees of freedom ( $n - 2$ )
	61,282,861	2,073,931	

Table 10.18 Sales of soft drinks – forecast data.

1	2	3	4	5	6
Year	Quarter	Code	Forecast sales (pallets)	Seasonal index SI	Regression forecast, $\hat{y}$
2006	Winter	25	20,432	0.97	20,978.58
	Spring	26	21,576	1.02	21,209.43
	Summer	27	22,729	1.06	21,440.27
	Autumn	28	20,557	0.95	21,671.12

**FORECAST]** where the  $x$ -values are the code values 25 to 28 and the actual values of  $x$  are the code values 1 to 24 and the actual values of  $y$  are the deseasonalized sales for these same coded periods. These values are in Column 6 of Table 10.18.

*Step 8. Multiply the forecast regression sales by the SI to forecast 2006 seasonal sales*

The forecast seasonal sales are shown in Column 4 of Table 10.18. What we have done is reversed our procedure by now multiplying the regression forecast by the SI. When we developed the data we divided by the SI to obtain a deseasonalized sale and used the regression analysis on this information.

The actual and forecast sales are shown in Figure 10.15. Although at first the calculation procedure may seem laborious, it can be very quickly executed using an Excel spread sheet and the given functions.

## Considerations in Statistical Forecasting

We must remember that a forecast is just that – a forecast. Thus when we use statistical analysis to forecast future patterns we have to exercise

caution when we interpret the results. The following are some considerations.

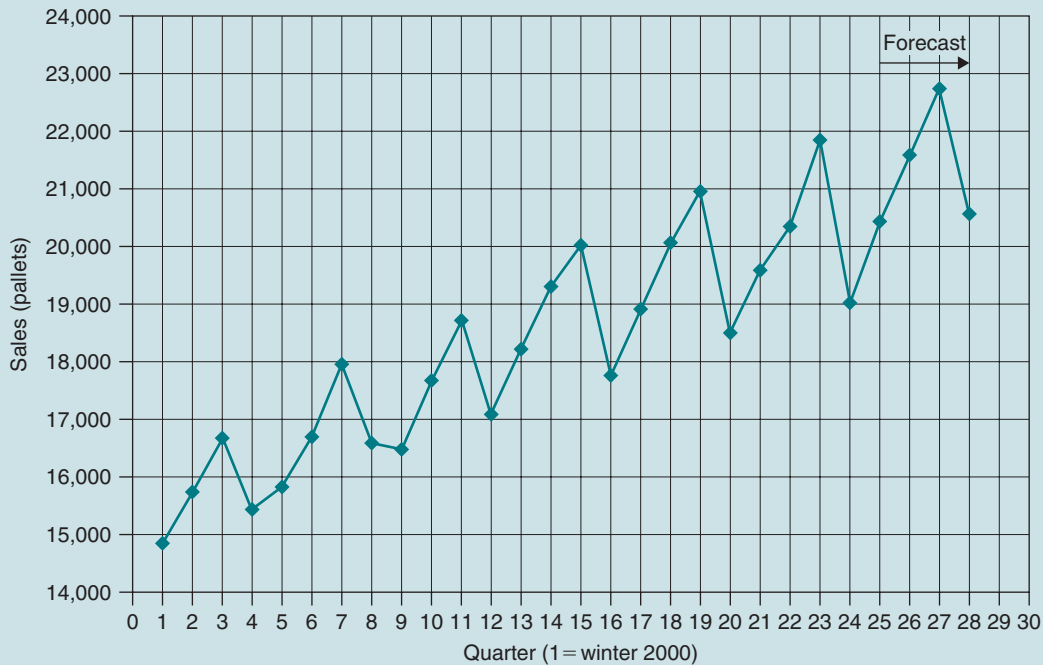
### Time horizons

Often in business, managers would like a forecast to extend as far into the future as possible. However, the longer the time period the more uncertain is the model because of the changing environment – What new technologies will come onto the market? What demographic changes will occur? How will interest rates move? An approach to recognize this is to develop forecast models for different time periods – say short, medium, and long-term. The forecast model for the shorter time period would provide the most reliable information.

### Collected data

Quantitative forecast models use collected or historical data to estimate future outcomes. In collecting data it is better to have detailed rather than aggregate information, as the latter might camouflage situations. For example, assume that you want to forecast sales of a certain product of which there are six different models. You could develop a model of revenues for all of the six models. However, revenues can be distorted by market changes, price increases, or exchange

Figure 10.15 Actual and forecast sales for soft drinks.



rates if exporting or importing is involved. It would be better first to develop a time series model on a unit basis according to product range. This base model would be useful for tracking inventory movements. It can then be extended to revenues simply by multiplying the data by unit price.

### Coefficient of variation

When past data is collected to make a forecast, the coefficient of variation of the data, or the ratio of the standard deviation to the mean ( $\sigma/\mu$ ), is an indicator of how reliable is a forecast model. For example, consider the time series data in Table 10.19.

Table 10.19 Collected data.

Period	Product A	Product B
January	1,100	800
February	1,024	40
March	1,080	564
April	1,257	12
May	1,320	16
June	1,425	456
July	1,370	56
August	1,502	12
September	1,254	954
$s$ (as a sample)	164.02	377.58
$\mu$	1,259.11	323.33
Coefficient of variation, $\sigma/\mu$	0.13	1.17

For product A the coefficient of variation is low meaning that the dispersion of the data relative to its mean is small. In this case a forecast model should be quite reliable. On the other hand, for Product B the coefficient of variation is greater than one or the sample standard deviation is greater than the mean. Here a forecast model would not be as reliable. In situations like this perhaps there is a seasonal activity of the product and this should be taken into account in the selected forecast model. In using the coefficient of variation as a guide, care should be taken as if there is a trend in the data that will of course impact the coefficient. As already discussed in the chapter, plotting the data on a scatter diagram would be a visual indicator of how good is the past data for forecasting purposes. Note that in determining the coefficient of variation we have used the sample standard deviation,  $s$ , as an estimate of the population standard deviation,  $\hat{\sigma}$ .

## Market changes

Market changes should be anticipated in forecasting. For example, in the past, steel requirements might be correlated with the forecast sale of automobiles. However plastic and composite materials are rapidly replacing steel, so this factor would distort the forecast demand for steel if the old forecasting approach were used. Alternatively, more and more uses are being found for plastics, so this element would need to be incorporated into a forecast for the demand for plastics. These types of events may not affect short-term planning but certainly are important in long-range forecasting when capital appropriation for plant and equipment is a consideration.

## Models are dynamic

A forecast model must be a dynamic working tool with the flexibility to be updated or modified as soon as new data become available that might impact the outcome of the forecast. For

example, an economic model for the German economy had to be modified with the fall of the Berlin Wall in 1989 and the fusion of the two Germanys. Similarly, models for the European Economy have been modified to take into account the impact of the Euro single currency.

## Model accuracy

All managers want an accurate model. The accuracy of the model, whether it is estimated at 10%, 20%, or say 50% can only be within a range bounded by the error in the collected data. Further, accuracy must be judged in light of control a firm has over resources and external events. Besides accuracy, also of interest in a forecast is when turning points in situations might be expected such as a marked increase (or decrease) in sales so that the firm can take advantage of the opportunities, or be prepared for the threats.

## Curvilinear or exponential models

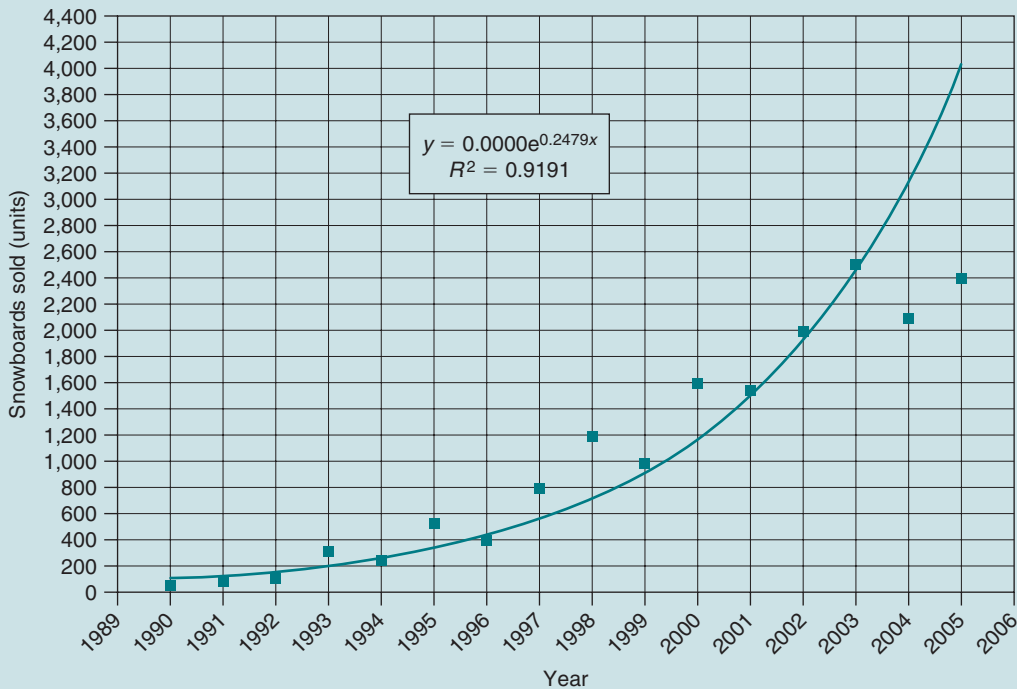
We must exercise caution in using curvilinear functions, where the predicted value  $\hat{y}$  changes rapidly with  $x$ . Even though the actual collected data may exhibit a curvilinear relationship, an exponential growth often cannot be sustained in the future often because of economic, market, or demographic reasons. In the classic life cycle curve in marketing, the growth period for successful new products often follows a curvilinear or more precisely an exponential growth model but this profile is unlikely to be sustained as the product moves into the mature stage.

In the worked example, surface area and house prices, we developed the following two-degree polynomial equation:

$$\hat{y} = 41.0575x^2 - 9,594.6456x + 849,828.1408$$

Using this for a surface area of 1,000 m<sup>2</sup> forecasts a house price of €32.3 million, which is

Figure 10.16 Exponential function for snowboard sales.



beyond the affordable range for most people. Consider also the sale of snowboards worked example presented at the beginning of the chapter. Here we developed a linear regression model that gave a coefficient of determination of 0.9316 and the model forecast sales of 3,248 units for 2010. Now if we develop an exponential relationship for this same data then this would appear as in Figure 10.16. The equation describing this curve is,

$$\hat{y} = e^{0.2479x}$$

The data gives a respectable coefficient of determination of 0.9191. If we use this to make a forecast for sale of snowboards in 2010 we have a value of  $2.62 \times 10^{216}$  which is totally unreasonable.

## Selecting the best model

It is difficult to give hard and fast rules to select the best forecasting model. The activity may be a trial and error process selecting a model and testing it against actual data or opinions. If a quantitative forecast model is used there needs to be consideration of subjective input, and vice-versa.

Models can be complex. In the 1980s, in a marketing function in the United States, I worked on developing a forecast model for world crude oil prices. This model was needed to estimate financial returns from future oil exploration, drilling, refinery, and chemical plant operation. The model basis was a combined multiple regression and curvilinear relationships incorporating variables in the United States economy such as changes in the GNP, interest rates, energy consumption, chemical



production and forecast chemical use, demographic changes, taxation, capital expenditure, seasonal effects, and country political risk. Throughout the development, the model was tested against known situations. The model proved to be a reasonable forecast of future prices.

A series of forecast models have been developed by a group of political scientists who study the United States elections. These models use combined factors such as public opinions in the preceding summer, the strength of the economy, and

the public's assessment of its economic well-being. The models have been used in all the United States elections since 1948 and have proved highly accurate.<sup>2</sup> In 2007 the world economy suffered a severe decline as a result of bank loans to low income homeowners. Jim Melcher, a money manager based in New York, using complex derivative models forecast this downturn and pulled out of this risky market and saved his clients millions of dollars.<sup>3</sup>

## Chapter Summary

This chapter covers forecasting using bivariate data and presents correlation, linear and multiple regression, and seasonal patterns in data.

### A time series and correlation

A time series is bivariate information of a dependent variable,  $y$ , such as sales with an independent variable  $x$  representing time. Correlation is the strength between these variables and can be illustrated by a scatter diagram. If the correlation is reasonable, then regression analysis is the technique to develop an equation that describes the relationship between the two variables. The coefficient of correlation,  $r$ , and the coefficient of determination,  $r^2$ , are two numerical measures to record the strength of the linear relationship. Both of these coefficients have a value between 0 and 1. The closer either is to unity then the stronger is the correlation. The coefficient of correlation can be positive or negative whereas the coefficient of determination is always positive.

### Linear regression in a time series data

The linear regression line for a time series has the form,  $\hat{y} = a + bx$ , where  $\hat{y}$  is the predicted value of the dependent variable,  $a$  and  $b$  are constants, and  $x$  is the time. The regression equation gives the best straight line that minimizes the error between the data points on the regression line and the corresponding actual data from which the regression line is developed. To forecast using the regression equation, knowing  $a$  and  $b$ , we insert the time,  $x$ , into the regression equation to give a forecast value  $\hat{y}$ . The variability around the regression line is measured by the standard error of the estimate,  $s_e$ . We can use the standard error of the estimate to give the confidence in our forecast by using the relationship  $\hat{y} \pm z s_e$  for large sample sizes and  $\hat{y} \pm t s_e$  for sample sizes no more than 30.

<sup>2</sup> Mathematically, Gore is a winner, *International Herald Tribune*, 1 September 2000.

<sup>3</sup> Warnings were missed in US loan meltdown, *International Herald Tribune*, 20 August 2007.

## Linear regression and casual forecasting

We can also use the linear regression relationship for causal forecasting. Here the assumption is that the predicted value of the dependent variable is a function not of time but another variable that causes the change in  $y$ . In causal forecasting all of the statistical relationships of correlation, prediction, variability, and confidence level of the forecast apply exactly as for a time series data. The only difference is that the value of the independent variable  $x$  is not time.

## Forecasting using multiple regression

Multiple regression is when there is more than one independent variable  $x$  to give an equation of the form,  $\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_kx_k$ . A coefficient of multiple determination,  $r^2$ , measures the strength of the relationship between the dependent variable  $y$  and the various independent variables  $x$ , and again there is a standard error of the estimate,  $s_e$ .

## Forecasting using non-linear regression

Non-linear regression is when the variable  $y$  is a curvilinear function of the independent variable  $x$ . The function may be a polynomial relationship of the form  $y = a + bx + cx^2 + dx^3 + \dots + kx^n$ . Alternatively it may have an exponential relationship of the form  $y = ae^{bx}$ . Again with both these relationships we have a coefficient of determination that illustrates the strength between the dependent variable and the independent variable.

## Seasonal patterns in forecasting

Often in selling seasonal patterns exist. In this case we develop a forecast model by first removing the seasonal impact to calculate a seasonal index. If we divide the actual sales by the seasonal index we can then apply regression analysis on this smoothed data to obtain a regression forecast. When we multiply the regression forecast by the seasonal index we obtain a forecast by season.

## Considerations in statistical forecasting

When we forecast using statistical data the longer the time horizon then the more inaccurate is the model. Other considerations are that we should work with specific defined variables rather than aggregated data and that past data must be representative of the future environment for the model to be accurate. Further, care must be taken in using curvilinear models as though the coefficient of determination indicates a high degree of accuracy, the model may not follow market changes.

## EXERCISE PROBLEMS

### 1. Safety record

#### Situation

After the 1999 merger of Exxon with Mobil, the newly formed corporation, ExxonMobil implemented worldwide its Operations Integrity Management System (OIMS), a programme that Exxon itself had developed in 1992 in part as a result of the Valdez oil spill in Alaska in 1989. Since the implementation of OIMS the company has experienced fewer safety incidents and its operations have become more reliable. These results are illustrated in the table below that shows the total incidents reported for every 200,000 hours worked since 1995.<sup>4</sup>

Year	Incidents per 200,000 hours
1995	1.35
1996	1.06
1997	0.98
1998	0.84
1999	0.72
2000	0.82
2001	0.65
2002	0.51
2003	0.38
2004	0.37
2005	0.38
2006	0.25

#### Required

1. Plot the data on a scatter diagram.
2. Develop the linear regression equation that best describes this data.
3. Using the regression information, what is annual change in the number of safety incidents reported by ExxonMobil?
4. What quantitative data indicates that there is a reasonable relationship over time with the safety incidents reported by ExxonMobil?
5. Using the regression equation what is a forecast of the number of reported incidents in 2007?
6. Using the regression equation what is a forecast of the number of reported incidents in 2010? What are your comments about this result?
7. From the data, what might you conclude about the future safety record of ExxonMobil?

<sup>4</sup> Managing risk in a challenging business, *The Lamp*, ExxonMobil, 2007, (2), p. 26.

## 2. Office supplies

### Situation

Bertrand Co. is a distributor of office supplies including agendas, diaries, computer paper, pens, pencils, paper clips, rubber bands, and the like. For a particular geographic region the company records over a 4-year period indicated the following monthly sales in pound sterling as follows.

Month	£ '000s	Month	£ '000s
January 2003	14	January 2005	42
February 2003	18	February 2005	43
March 2003	16	March 2005	42
April 2003	21	April 2005	41
May 2003	15	May 2005	41
June 2003	19	June 2005	42
July 2003	22	July 2005	43
August 2003	31	August 2005	49
September 2003	33	September 2005	52
October 2003	28	October 2005	47
November 2003	27	November 2005	48
December 2003	29	December 2005	49
January 2004	26	January 2006	51
February 2004	28	February 2006	50
March 2004	31	March 2006	52
April 2004	33	April 2006	54
May 2004	34	May 2006	57
June 2004	35	June 2006	54
July 2004	38	July 2006	48
August 2004	41	August 2006	59
September 2004	43	September 2006	61
October 2004	37	October 2006	57
November 2004	37	November 2006	56
December 2004	41	December 2006	61

### Required

1. Using a coded value for the data with January 2003 equal to 1, develop a time series scatter diagram for this information.
2. What is an appropriate linear regression equation to describe the trend of this data?
3. What might be an explanation for the relative increase in sales for the months of August and September?
4. What can you say about the reliability of the regression model that you have created? Justify your reasoning.

5. What are the average quarterly sales as predicted by the regression equation?
6. What would be the forecast of sales for June 2007, December 2008, and December 2009? Which would be the most reliable?
7. What are your comments about the model you have created and its use as a forecasting tool?

### 3. Road deaths

#### Situation

The table below gives the number of people killed on French roads since 1980.<sup>5</sup>

Year	Deaths	Year	Deaths
1980	12,543	1992	9,083
1981	12,400	1993	8,500
1982	12,400	1994	8,333
1983	11,833	1995	8,000
1984	11,500	1996	8,067
1985	10,300	1997	7,989
1986	10,833	1998	8,333
1987	9,855	1999	7,967
1988	10,548	2000	7,580
1989	10,333	2001	7,720
1990	10,600	2002	7,242
1991	9,967		

#### Required

1. Plot the data on a scatter diagram.
2. Develop the linear regression equation that best describes this data.
3. Is the linear equation a good forecasting tool for forecasting the future value of the road deaths? What quantitative piece of data justifies your response?
4. Using the regression information, what is the yearly change of the number of road deaths in France?
5. Using the regression information, what is the forecast of road deaths in France in 2010?
6. Using the regression information, what is the forecast of road deaths in France in 2030?
7. What are your comments about the forecast data obtained in Questions 5 and 6?

<sup>5</sup> Metro-France 16 May 2003, p. 2.

## 4. Carbon dioxide

### Situation

The data below gives the carbon dioxide emissions, CO<sub>2</sub>, for North America, in millions of metric tons carbon equivalent. Carbon dioxide is one of the gases widely believed to cause global warming.<sup>6</sup>

Year	North America
1992	1,600
1993	1,625
1994	1,650
1995	1,660
1996	1,750
1997	1,790
1998	1,800
1999	1,825
2000	1,850
2001	1,800

### Required

1. Plot the information on a time series scatter diagram and develop the linear regression equation for the scatter diagram.
2. What are the indicators that demonstrate the strength of the relationship between carbon dioxide emission and time? What are your comments about these values?
3. What is the annual rate of increase of carbon dioxide emissions using the regression relationship?
4. Using the regression equation, forecast the carbon dioxide emissions in North America for 2010?
5. From the answer in Question 3, what is your 95% confidence limit for this forecast?
6. Using the regression equation, forecast the carbon dioxide emissions in North America for 2020?
7. What are your comments about using this information for forecasting?

## 5. Restaurant serving

### Situation

A restaurant has 55 full-time operating staff that includes kitchen staff and servers. Since the restaurant is open for lunch and dinner 7 days a week there are times that the restaurant does not have the full complement of staff. In addition, there are times when

<sup>6</sup> Insurers weigh moves on global warming, *Wall Street Journal Europe*, 7 May 2003, p. 1.

staff are simply absent as they are sick. The restaurant manager conducted an audit to determine if there was a relationship between the number of staff absent and the average time that a client had to wait for the main meal. This information is given in the table below.

Number of staff absent	Average waiting time (minutes)
7	24
1	5
3	12
8	30
0	3
4	16
2	15
3	20
5	22
9	27

### Required

1. For the information given, develop a scatter diagram between number of staff absent and the average time that a client has to wait for the main meal.
2. Using regression analysis, what is a quantitative measure that illustrates a reasonable relationship between the waiting time and the number of staff absent?
3. What is the linear regression equation that describes the relationship?
4. What is an estimate of the time delay per employee absent?
5. When the restaurant has the full complement of staff, to the nearest two decimal places, what is the average waiting time to obtain the main meal as predicted by the linear regression equation?
6. If there are six employees absent, estimate the average waiting time as predicted by the linear regression equation.
7. If there are 20 employees absent, estimate the average waiting time as predicted by the linear regression equation. What are your comments about this result?
8. What are some of the random occurrences that might explain variances in the waiting time?

## 6. Product sales

### Situation

A hypermarket made a test to see if there was a correlation between the shelf space of a special brand of raisin bread and the daily sales. The following is the data that was collected over a 1-month period.

Shelf space (m <sup>2</sup> )	Daily sales units
0.25	12
0.50	18
0.75	21
0.75	23
1.00	18
1.00	23
1.25	25
1.25	28
2.00	30
2.00	34
2.25	32
2.25	40

### Required

1. Illustrate the relationship between the sale of the bread and the allocated shelf space.
2. Develop a linear regression equation for the daily sales and the allocated shelf space. What are your conclusions?
3. If the allocated shelf space was 1.50 m<sup>2</sup>, what is the estimated daily sale of this bread?
4. If the allocated shelf space was 5.00 m<sup>2</sup>, what is the estimated daily sale of this bread? What are your comments about this forecast?
5. What does this sort of experiment indicate from a business perspective?

## 7. German train usage

### Situation

The German rail authority made an analysis of the number of train users on the network in the southern part of the country since 1993 covering the months for June, July, and August. The Transport Authority was interested to see if they could develop a relationship between the number of users and another easily measurable variable. In this way they would have a forecasting tool. The variables they selected for developing their models were the unemployment rate in this region and the number of foreign tourists visiting Germany. The following is the data collected:

Year	Unemployment rate (%)	No. of tourists (millions)	Train users (millions)
1993	11.5	7	15
1994	12.7	2	8
1995	9.7	6	13
1996	10.4	4	11

(Continued)



Year	Unemployment rate (%)	No. of tourists (millions)	Train users (millions)
1997	11.7	14	25
1998	9.2	15	27
1999	6.5	16	28
2000	8.5	12	20
2001	9.7	14	27
2002	7.2	20	44
2003	7.7	15	34
2004	12.7	7	17

### Required

1. Illustrate the relationship between the number of train users and unemployment rate on a scatter diagram.
2. Using simple regression analysis, what are your conclusions about the correlation between the number of train users and the unemployment rate?
3. Illustrate the relationship between the number of train users and foreign tourists on a scatter diagram.
4. Using simple regression analysis, what are your conclusions about the correlation between the number of train users and the number of foreign tourists?
5. In any given year, if the number of foreign tourists were estimated to be 10 million, what would be a forecast for the number of train users?
6. If a polynomial correlation (to the power of 2) between train users and foreign tourists was used, what are your observations?

## 8. Cosmetics

### Situation

Yam Ltd. sells cosmetic products by simply advertising in throwaway newspapers and by ladies who organize Yam parties in order to sell directly the products. The table below gives data on a monthly basis for revenues, in pound sterling, for sales of cosmetics each month for the last year according to advertising budget and the equivalent number of people selling full time. This data is to be analysed using multiple regression analysis.

Sales revenues	Advertising budget	Sales persons	No. of yam parties
721,200	47,200	542	101
770,000	54,712	521	67
580,000	25,512	328	82
910,000	94,985	622	75
315,400	13,000	122	57

Sales revenues	Advertising budget	Sales persons	No. of yam parties
587,500	46,245	412	68
515,000	36,352	235	84
594,500	25,847	435	85
957,450	64,897	728	81
865,000	67,000	656	37
1,027,000	97,000	856	99
965,000	77,000	656	100

### Required

1. Develop a two-independent-variable multiple regression model for the sales revenues as a function of the advertising budget, and the number of sales persons. Does the relationship appear strong? Quantify.
2. From the answer developed in Question 1, assume for a particular month it is proposed to allocate a budget of £30,000 and there will be 250 sales persons available. In this case, what would be an estimate of the sales revenues for that month?
3. What are the 95% confidence intervals for Question 2?
4. Develop a three-independent-variable multiple regression model for the sales revenues as a function of the advertising budget, the number of sales persons, and the number of Yam parties. Does the relationship appear strong? Quantify.
5. From the answer developed in Question 4, assume for a particular month it is proposed to allocate a budget of \$US 4,000 to use 30 sales persons, with a target to make 21,000 sales contacts. Then what would be an estimate of the sales for that month?
6. What are the 95% confidence intervals for Question 5?

## 9. Hotel revenues

### Situation

A hotel franchise in the United States has collected the revenue data in the following table for the several hotels in its franchise.

Year	Revenues (\$millions)
1996	35
1997	37
1998	44
1999	51
2000	50
2001	58
2002	59

(Continued)

Year	Revenues (\$millions)
2003	82
2004	91
2005	104

### Required

1. From the given information develop a linear regression model of the time period against revenues.
2. What is the coefficient of determination for relationship developed in Question 1?
3. What is the annual revenue growth rate based on the given information?
4. From the relationship in Question 1, forecast the revenues in 2008 and give the 90% confidence limits.
5. From the relationship in Question 1, forecast the revenues in 2020 and give the 90% confidence limits.
6. From the given information develop a two-degree polynomial regression model of the time period against revenues.
7. What is the coefficient of determination for relationship developed in Question 6?
8. From the relationship in Question 6, forecast the revenues in 2008.
9. From the relationship in Question 6, forecast the revenues in 2020.
10. What are your comments related to making a forecast for 2008 and 2020?

## 10. Hershey Corporation

### Situation

Dan Smith has in his investment portfolio shares of Hershey Company, Pennsylvania, United States of America, a Food Company well known for its chocolate. Dan bought a round lot (100 shares) in September 1988 for \$28.500 per share. Since that date, Dan participated in Hershey's reinvestment programme. That meant he reinvested all quarterly dividends into the purchase of new shares. In addition, from time to time, he made optional cash investment for new shares. The share price, and the number of shares held by Dan, at the end of each quarter since the time of the initial purchase, and the 1st quarter 2007, is given in Table 1.

*Table 1* Table Hershey.

End of month	Price (\$/share)	No. of shares	End of month	Price (\$/share)	No. of shares
September 1988	28.500	100.0000	June	31.126	101.9373
December	25.292	100.6919	September	31.500	102.4734
March 1989	26.089	101.3673	December 1989	35.010	102.9584

Table 1 (Continued).

End of month	Price (\$/share)	No. of shares	End of month	Price (\$/share)	No. of shares
March 1990	31.250	103.5043	December 1998	63.000	426.6189
June	36.500	118.5097	March 1999	61.877	428.2736
September	35.722	119.8518	June	55.500	430.1256
December 1990	37.995	120.5615	September	52.539	432.2541
March 1991	38.896	133.9852	December 1999	48.999	434.5491
June	42.375	134.6966	March 2000	41.996	437.2394
September	39.079	135.5411	June	53.967	439.3459
December 1991	39.079	148.6803	September	46.375	441.9986
March 1992	41.317	149.5619	December 2000	59.625	444.0742
June	40.106	150.4756	March 2001	65.250	445.9798
September	44.500	151.3886	June	60.600	448.0405
December 1992	45.500	152.2869	September	66.300	450.0847
March 1993	53.000	153.0627	December 2001	65.440	452.1652
June	49.867	173.0976	March 2002	68.750	454.1548
September	51.824	174.0996	June	64.280	456.2920
December 1993	49.928	175.1457	September	73.280	458.3312
March 1994	49.618	176.2047	December 2002	66.062	460.6034
June	43.971	177.4068	March 2003	63.254	462.9882
September	45.640	178.6702	June	72.100	465.0912
December 1994	48.235	179.8740	September	72.665	467.6194
March 1995	50.210	181.0383	December 2003	77.580	470.0003
June	53.272	191.3792	March 2004	84.939	472.1860
September	62.938	192.4739	June	46.323	948.3983
December 1995	67.170	193.5055	September	48.350	952.7137
March 1996	73.625	194.4516	December 2004	56.239	956.4406
June	71.305	201.9192	March 2005	62.209	959.8230
September	45.261	405.6230	June	64.524	963.0956
December 1996	44.625	407.4409	September	57.600	967.1921
March 1997	49.750	409.0789	December 2005	57.845	971.2886
June	57.081	410.5122	March 2006	52.809	975.7948
September	55.810	412.1304	June	54.001	980.2219
December 1997	63.738	413.5529	September	51.625	985.3484
March 1998	71.233	414.8302	December 2006	50.980	990.5670
June	69.504	416.1432	March 2007	53.928	995.5265
September	67.404	417.6250			

### Required

1. For the data given and using a coded value for the quarter starting at unity for September 1988, develop a line graph for the price per share. How might you explain the shape of the line graph?
2. For the data given and using a coded value for the quarter starting at unity for September 1988, develop a time series scatter diagram for the asset value (value of

the portfolio) of the Hershey stock. Show on the scatter diagram graph the linear regression line for the asset value.

3. What is the equation that represents the linear regression line?
4. What information indicates quantitatively the accuracy of the asset value and time for this model? Would you say that the regression line could be used to reasonably forecast future values?
5. From the linear regression equation, what is the annual average growth rate in dollars per year of the asset value of the portfolio?
6. Dan plans to retire at the end of December in 2020 (4th quarter 2020). Using the linear regression equation, what is a forecast of the value of Dan's assets in Hershey stock at this date?
7. At a 95% confidence level, what are the upper and lower values of assets at the end of December 2020?
8. What occurrences or events could affect the accuracy of forecasting the value of Hershey's asset value in 2020?
9. Qualitatively, would you think there is great risk for Dan in finding that the value of his assets is significantly reduced when he retires? Justify your response.

## 11. Compact discs

### Situation

The table below gives the sales by year of music compact discs by a selection of Virgin record stores.

Year	CD sales (millions)
1995	45
1996	52
1997	79
1998	72
1999	98
2000	99
2001	138
2002	132
2003	152
2004	203

### Required

1. Plot the data on a scatter diagram.
2. Develop the linear regression equation that best describes this data. Is the equation a good forecasting tool for CD record sales? What quantitative piece of data justifies your response?

3. From the linear regression function, what is the forecast for CD sales in 2007?
4. From the linear regression function, what is the forecast for CD sales in 2020?
5. Does a second-degree polynomial regression line have a better fit for this data? Why?
6. What would be the forecast for record sales calls in 2007 using the polynomial relationship developed in Question 5?
7. What would be the forecast for record sales calls in 2020 using the polynomial relationship developed in Question 5?
8. What are your comments regarding using the linear and polynomial function to forecast compact disc sales?

## 12. United States imports

### Situation

The data in Table 1 is the amount of goods imported into the United States from 1960 until 2006.<sup>7</sup> (This is the same information presented in the Box Opener “Value of imported goods into the States” of this chapter.)

*Table 1*

Year	Imported goods (\$millions)	Year	Imported goods (\$millions)	Year	Imported goods (\$millions)
1960	14,758	1976	124,228	1992	536,528
1961	14,537	1977	151,907	1993	589,394
1962	16,260	1978	176,002	1994	668,690
1963	17,048	1979	212,007	1995	749,374
1964	18,700	1980	249,750	1996	803,113
1965	21,510	1981	265,067	1997	876,794
1966	25,493	1982	247,642	1998	918,637
1967	26,866	1983	268,901	1999	1,031,784
1968	32,991	1984	332,418	2000	1,226,684
1969	35,807	1985	338,088	2001	1,148,231
1970	39,866	1986	368,425	2002	1,167,377
1971	45,579	1987	409,765	2003	1,264,307
1972	55,797	1988	447,189	2004	1,477,094
1973	70,499	1989	477,665	2005	1,681,780
1974	103,811	1990	498,438	2006	1,861,380
1975	98,185	1991	491,020		

<sup>7</sup>US Census Bureau, Foreign Trade division, [www.census.gov/foreign-trade/statistics/historical](http://www.census.gov/foreign-trade/statistics/historical) goods, 8 June 2007.

### Required

1. Develop a time series scatter data for the complete data.
2. From the scatter diagram developed in Question 1 develop linear regression equations using just the following periods to develop the equation where  $x$  is the year. Also give the corresponding coefficient of determination: 1960–1964; 1965–1969; 1975–1979; 1985–1989; 1995–1999; 2002–2005.
3. Using the relationships developed in Question 2, what would be the forecast values for 2006?
4. Compare these forecast values obtained in Question 3 with the actual value for 2006. What are your comments?
5. Develop the linear regression equation and the corresponding coefficient of determination for the complete data and show this information on the scatter diagram.
6. Develop the exponential equation and the corresponding coefficient of determination for the complete data and show this information on the scatter diagram.
7. Develop the fourth power polynomial equation and the corresponding coefficient of determination for the complete data and show this information on the scatter diagram.
8. Use the linear, exponential, and polynomial equations developed in Questions 5, 6, and 7 to forecast the value of imports to the United States for 2010.
9. Use the equation for the period, 2002–2005, developed in Question 3 to forecast United States imports for 2010.
10. Discuss your observations and results for this exercise including the forecasts that you have developed.

## 13. English pubs

### Situation

The data below gives the consumption of beer in litres at a certain pub on the river Thames in London, United Kingdom between 2003 and 2006 on a monthly basis.

Month	2003	2004	2005	2006
January	15,000	16,200	16,900	17,100
February	37,500	45,000	47,000	52,500
March	127,500	172,500	210,000	232,500
April	502,500	540,000	675,000	720,000
May	567,500	569,500	697,500	757,500
June	785,000	715,000	765,000	862,500
July	827,500	948,600	1,098,000	1,124,500
August	990,000	978,400	1,042,300	1,198,500
September	622,500	682,500	765,000	832,500
October	75,000	82,500	97,500	105,000
November	15,000	17,500	20,000	22,500
December	7,500	8,500	8,200	9,700

**Required**

1. Develop a line graph on a quarterly basis for the data using coded values for the quarters. That is, winter 2003 has a coded value of 1. What are your observations?
2. Plot a graph of the centred moving average for the data. What is the linear regression equation that describes the centred moving average?
3. Determine the ratio of the actual sales to the centred moving average for each quarter. What is your interpretation of this information for 2004?
4. What are the seasonal indices for the four quarters using all the data?
5. What is the value of the coefficient of determination on the deseasonalized sales data?
6. Develop a forecast by quarter for 2007.
7. What would be an estimate of the annual consumption of beer in 2010? What are your comments about this forecast?

**14. Mersey Store****Situation**

The Mersey Store in Arkansas, United States is a distributor of garden tools. The table below gives the sales by quarter since 1997. All data are in \$ '000s.

Year	Quarter	Sales	Year	Quarter	Sales
1997	Winter	11,302	2001	Winter	13,184
	Spring	12,177		Spring	14,146
	Summer	13,218		Summer	14,966
	Autumn	11,948		Autumn	13,665
1998	Winter	11,886	2002	Winter	13,781
	Spring	12,198		Spring	14,636
	Summer	13,294		Summer	15,142
	Autumn	11,785		Autumn	13,415
1999	Winter	11,875	2003	Winter	14,327
	Spring	12,584		Spring	15,251
	Summer	13,332		Summer	15,082
	Autumn	12,354		Autumn	14,002
2000	Winter	12,658	2004	Winter	14,862
	Spring	13,350		Spring	15,474
	Summer	14,358		Summer	15,325
	Autumn	13,276		Autumn	14,425

**Required**

1. Show graphically that the sales for Mersey are seasonal.
2. Use the multiplication model, predict sales by quarter for 2005. Show graphically the moving average, deseasonalized sales, regression line, and forecast.



## 15. Swimwear

### Situation

The following table gives the sale of swimwear, in units per month, for a sports store in Redondo Beach, Southern California, United States of America during the period 2003 through 2006.

Month	2003	2004	2005	2006
January	150	75	150	75
February	375	450	450	525
March	1,275	1,725	2,100	2,325
April	5,025	5,400	6,750	7,200
May	5,175	5,625	6,975	7,575
June	5,850	6,150	7,650	8,625
July	5,275	5,486	6,980	7,245
August	4,900	5,784	6,523	6,985
September	3,225	3,825	4,650	5,325
October	750	825	975	1,050
November	150	75	150	225
December	75	150	85	175

### Required

1. Develop a line graph on a quarterly basis for the data using coded values for the quarters. That is, winter 2003 has a coded value of 1. What are your observations?
2. Plot a graph of the centred moving average for the data. What is the linear regression equation that describes the centred moving average?
3. Determine the ratio of the actual sales to the centred moving average for each quarter. What is your interpretation of this information for 2005?
4. What are the seasonal indices for the four quarters using all the data?
5. Develop a forecast by quarter for 2007.
6. Why are unit sales as presented preferable to sales on a dollar basis?

## 16. Case: Saint Lucia

### Situation

Saint Lucia is an overseas territory of the United Kingdom with a population in 2007 of 171,000. It is an island of 616 square miles and counts as its neighbours Barbados, Saint Vincent, The Grenadines, and Martinique. It is an island with a growing tourist industry and offers the attraction of long sandy beaches, stunning nature trails, superb diving in deep blue waters, and relaxing spas.<sup>8</sup>

With increased tourism goes the demand for hotel and restaurants. Related to these two hospitality institutions is the volume of wine in thousand litres, sold per month during

<sup>8</sup> Based on information from a Special Advertising Section of *Fortune*, 2 July 2007, p. S1.

2005, 2006, and 2007. This data is given in Table 1. In addition, the local tourist bureau published data on the number of tourists visiting Santa Lucia for the same period. This information is in Table 2.

*Table 1*

Month	Unit wine sales (1,000 litres)		
	2005	2006	2007
January	530	535	578
February	436	477	507
March	522	530	562
April	448	482	533
May	422	498	516
June	499	563	580
July	478	488	537
August	400	428	440
September	444	430	511
October	486	486	480
November	437	502	499
December	501	547	542

*Table 2*

Month	Tourist bookings		
	2005	2006	2007
January	28,700	29,800	30,800
February	23,200	25,200	28,000
March	29,000	28,000	31,000
April	23,500	26,000	28,400
May	21,900	25,000	27,500
June	25,300	31,000	32,000
July	26,000	25,550	31,000
August	20,100	23,200	22,000
September	22,300	24,100	26,000
October	25,100	25,100	27,000
November	22,600	27,000	28,000
December	27,000	31,900	30,200

### Required

Use the data for forecasting purposes and develop and test an appropriate model.

*This page intentionally left blank*

# Indexing as a method for data analysis

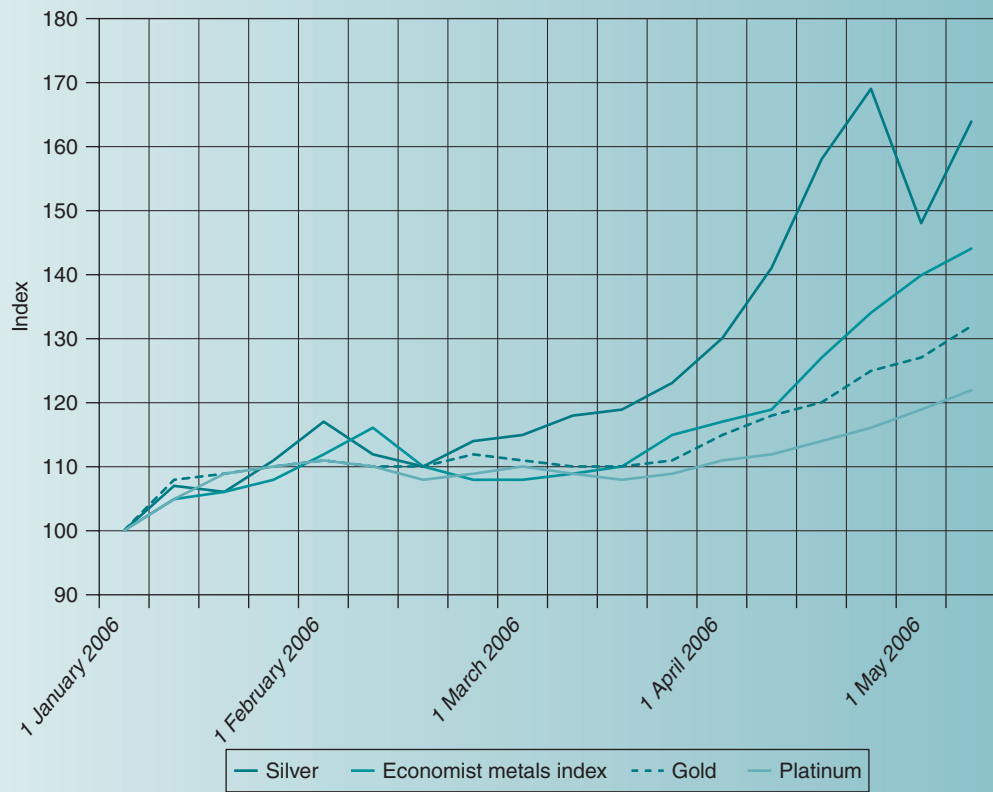
## Metal prices

*Metal prices continued to soar in early 2006 as illustrated in Figure 11.1, which gives the index value for various metals for the first half of 2006 based on an index of 100 at the beginning of the year. The price of silver has risen by some 65%, gold by 32%, and platinum by 21%. Aluminium, copper, lead, nickel, and zinc are included in The Economist metals index curve and here the price of copper has increased by 60% and nickel by 45%.<sup>1</sup> Indexing is another way to present statistical data and this is the subject of this chapter.*

---

<sup>1</sup> Metal prices, economic and financial indicators, *The Economist*, 6 May 2006, p. 105.

Figure 11.1 Metal prices.



## Learning objectives

After studying this chapter you will learn how to **present** and **analyse** statistical data using **index values**. The subjects treated are as follows:

- ✓ **Relative time-based indexes** • Quantity index number with a fixed base • Price index number with a fixed base • Rolling index number with a moving base • Changing the index base • Comparing index numbers • Consumer price index (CPI) and the value of goods and services • Time series deflation.
- ✓ **Relative regional indexes (RRIs)** • Selecting the base value • Illustration by comparing the cost of labour.
- ✓ **Weighting the index number** • Unweighted index number • Laspeyres weighted price index • Paasche weighted price index • Average quantity-weighted price index.

In Chapter 10, we introduced bivariate time-series data showing how past data can be used to forecast or estimate future conditions. There may be situations when we are more interested not in the absolute values of information but how data compare with other values. For example, we might want to know how prices have changed each year or how the productivity of a manufacturing operation has increased over time. For these situations we use an **index number** or **index value**. The index number is the ratio of a certain value to a base value usually multiplied by 100. When the base value equals 100 then the measured values are a percentage of the base value as illustrated in the box opener “Metal prices”.

time period in years and the 2nd column is the absolute values of enrolment in an MBA programme for a certain business school over the last 10 years from 1995. Here the data for 1995 is considered the **index base value**. The 3rd column gives the ratio of a particular year to the base value. The 4th column is the ratio for each year multiplied by 100. This is the index number. The index number for the base period is 100 and this is obtained by the ratio  $(95/95) \times 100$ . If we consider the year 2000, the enrolment for the MBA programme is 125 candidates. This gives a ratio to the 1995 data of  $125/95$  or 1.32.

### Relative Time-Based Indexes

Perhaps the most common indices are quantity and price indexes. In their simplest form they measure the relative change in time respective to a given base value.

#### Quantity index number with a fixed base

As an example of a quantity index consider the information in Table 11.1. The 1st column is the

*Table 11.1* Enrolment in an MBA programme.

Year	Enrolment	Ratio to base value	Index number
1995	95	1.00	100
1996	97	1.02	102
1997	110	1.16	116
1998	56	0.59	59
1999	64	0.67	67
2000	125	1.32	132
2001	102	1.07	107
2002	54	0.57	57
2003	62	0.65	65
2004	70	0.74	74

Table 11.2 Average price of unleaded gasoline in the United States in 2004.

Month	\$/gallon	\$/litre	Ratio to base value	Index number
January	1.5920	0.4206	1.00	100
February	1.6720	0.4417	1.05	105
March	1.7660	0.4666	1.11	111
April	1.8330	0.4843	1.15	115
May	2.0090	0.5308	1.26	126
June	2.0410	0.5392	1.28	128
July	1.9390	0.5123	1.22	122
August	1.8980	0.5015	1.19	119
September	1.8910	0.4996	1.19	119
October	2.0290	0.5361	1.27	127
November	2.0100	0.5310	1.26	126
December	1.8820	0.4972	1.18	118

Thus, the index for 2000 is  $1.32 * 100 = 132$ . We can interpret this information by saying that enrolment in 2000 is 132% of the enrolment in 1995, or alternatively an increase of 32%. In 2004 the enrolment is only 74% of the 1995 enrolment or 26% less ( $100\% - 74\%$ ).

The general equation for this index,  $I_Q$ , which is called the **relative quantity index**, is,

$$I_Q = \frac{Q_n}{Q_0} * 100 \quad 11(i)$$

Here  $Q_0$  is the quantity at the base period, and  $Q_n$  is the quantity at another period. This other period might be at a future date or after the base period. Alternatively, it could be a past period or before the base period.

### Price index number with a fixed base

Another common index, calculated in a similar way to the quantity index, is the price index, which compares the level of prices from one period to another. The most common price index is the **consumer price index**, that is used as

a measure of inflation by comparing the general price level for specific goods and services in the economy. The data is collected and compiled by government agencies such as Bureau of Labour Statistics in the United Kingdom and a similar department in the United States. In the European Union the organization concerned is Eurostat.

Consider Table 11.2 which gives the average price of unleaded regular petrol in the United States for the 12-month period from January 2004.<sup>2</sup> (For comparison the price is also given \$ per litre where 1 gallon equals 3.7850 litres.) In this table, we can see that the price of gasoline has increased 28% in the month of June compared to the base month of January.

In a similar manner to the quantity index, the general equation for this index,  $I_P$ , called the **relative price index** is,

$$I_P = \frac{P_n}{P_0} * 100 \quad 11(ii)$$

Here  $P_0$  is the price at the base period, and  $P_n$  is the price at another period.

<sup>2</sup>US Department of Labor Statistics, <http://data.bls.gov/cgi-bin/surveymost>.

**Table 11.3** Rolling index number of MBA enrolment.

Year	Enrolment	Ratio to immediate previous period	Annual change Rolling index
1995	95		
1996	97	1.0211	102
1997	110	1.1340	113
1998	56	0.5091	51
1999	64	1.1429	114
2000	125	1.9531	195
2001	102	0.8160	82
2002	54	0.5294	53
2003	62	1.1481	115
2004	70	1.1290	113

### Rolling index number with a moving base

We may be more interested to know how data changes periodically, rather than how it changes according to a fixed base. In this case, we would use a **rolling index number**. Consider Table 11.3 which is the same enrolment MBA data from Table 11.1. In the last column we have an index showing the change relative to the previous year. For example, the rolling index for 1999 is given by  $(64/56) * 100 = 114$ . This means that in 1999 there was a 14% increase in student enrolment compared to 1998. In 2002 the index compared to 2001 is calculated by  $(54/102) * 100 = 53$ . This means that enrolment is down 47%  $(100 - 53)$  in 2002 compared to 2001, the previous year. Again the value of the index has been rounded to the nearest whole number.

### Changing the index base

When the base point of data is too far in the past the index values may be getting too high to be meaningful and so we may want to use a more

**Table 11.4** Retail sales index.

Year	Sales index 1980 = 100	Sales index 1995 = 100
1995	295	100
1996	286	97
1997	301	102
1998	322	109
1999	329	112
2000	345	117
2001	352	119
2002	362	123
2003	359	122
2004	395	134

recent index so that our base point corresponds more to current periods. For example, consider Table 11.4 where the 2nd column shows the relative sales for a retail store based on an index of 100 in 1980. The 3rd column shows the index on a basis of 1995 equal to 100. The index value for 1995, for example, is  $(295/295) * 100 = 100$ . The index value for 1998 is  $(322/295) * 100 = 109$ . The index values for the other years are determined in the same manner. By transposing the data in this manner we have brought our index information closer to our current year.

### Comparing index numbers

Another interest that we might have is to compare index data to see if there is a relationship between one index number and another. As an illustration, consider Table 11.5 which is index data for the number of new driving licences issued and the number of recorded automobile accidents in a certain community. The 2nd column, for the number of driving licences issued, gives information relative to a base period of 1960 equal to 100. The 3rd column gives the number of recorded automobile accidents but in this case the base period of 100 is for the year



**Table 11.5** Automobile accidents and driving licenses issued.

Year	Driving licenses issued 1960 = 100	Automobile accidents 2000 = 100	Driving licenses issued 2000 = 100
1995	307	62	76
1996	325	71	80
1997	335	79	83
1998	376	83	93
1999	411	98	101
2000	406	100	100
2001	413	105	102
2002	421	108	104
2003	459	110	113
2004	469	112	116

2000. It is inappropriate to compare data of different base periods and what we have done is converted the number of driving licences issued to a base period of the year 2000 equal to 100. In this case, in 2000 the index is  $(406/406) \times 100 = 100$ . Then for example, the index in 1995 is  $(307/406) \times 100 = 76$  and in 2004 the index is  $(469/406) \times 100 = 116$ . In both cases, the indices are rounded to the nearest whole number.

Now that we have the indices on a common base it is easier to compare the data. For example, we can see that there appears to be a relationship between the number of new driving licenses issued and the recorded automobile accidents. More specifically in the period 1995–2000, the index for automobile accidents went from 62 to 100 or a more rapid increase than for the issue of driving licences which went from 76 to 100. However, in the period 2000–2004, the increase was not as pronounced going from 100 to 112 compared to the number of licenses issued going from 100 to 116. This could have been perhaps because of better police surveillance, a better road infrastructure, or other reasons.

Figure 11.2 gives a graph of the data where we can see clearly the changes. Comparing index numbers has a similarity to causal regression analysis presented in Chapter 10, where we determined if the change in one variable was caused by the change in another variable.

## CPI and the value of goods and services

The CPI is a measure of how prices have changed over time. It is determined by measuring the value of a “basket” of goods in one base period and then comparing the value of the same basket of goods at a later period. The change is most often presented on a **ratio measurement scale**. This basket of goods can include all items such as food, consumer goods, housing costs, mortgage interest payments, indirect taxes, etc. Alternatively, the CPI can be determined by excluding some of these items. When there is a significant increase in the CPI then this indicates an inflationary period. As an illustration, Table 11.6 gives the CPI in the United Kingdom for 1990 for all items.<sup>3</sup> For this period the CPI has increased by 9.34%.  $[(129.9 - 118.8)/118.8]$ . (Note that we have included the CPI for December 1989, in order to determine the annual change for 1990.)

Say now, for example, your annual salary at the end of 1989 was £50,000 and then at the end of 1990 it was increased to £54,000. Your salary has increased by an amount of 8%  $[(£54,000 - 50,000)/50,000]$  and your manager might expect you to be satisfied. However, if you measure your salary increase to the CPI of 9.34% the “real” value or “worth” of your salary has in fact gone down. You have less spending power than you did at the end of 1989 and would not unreasonably be dissatisfied.

Consider now Table 11.7 which is the CPI in the United Kingdom for 2001 for all items. For this period the CPI has increased by only 0.70%

<sup>3</sup> <http://www.statistics.gov.uk> (data, 13 July 2005).

Figure 11.2 Automobile accidents and driving licences issued.

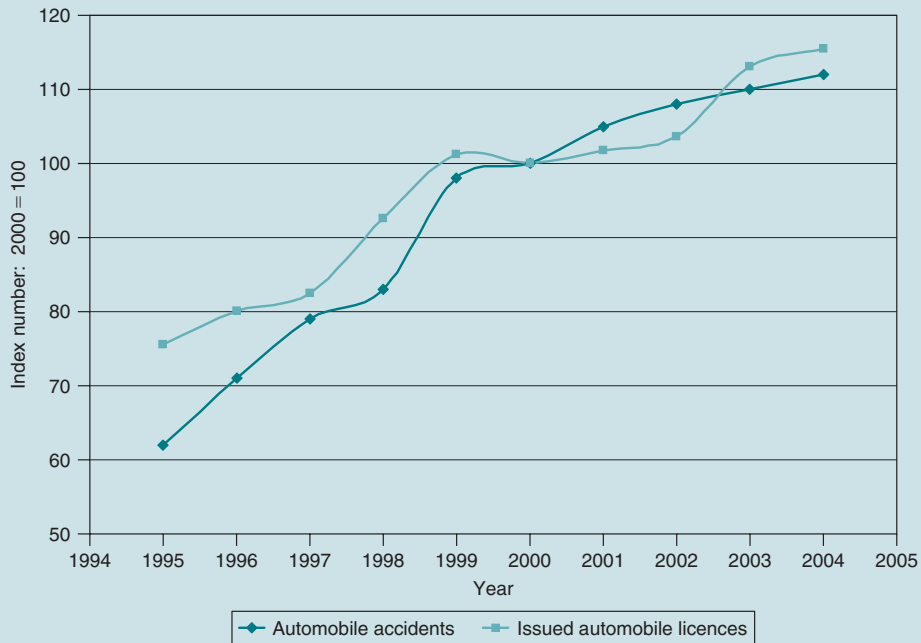


Table 11.6 Consumer price index, 1990.

Month	Index
December 1989	118.8
January 1990	119.5
February	120.2
March	121.4
April	125.1
May	126.2
June	126.7
July	126.8
August	128.1
September	129.3
October	130.3
November	130.0
December 1990	129.9

Table 11.7 Consumer price index, 2001.

Month	Index
December 2000	172.2
January 2001	171.1
February	172.0
March	172.2
April	173.1
May	174.2
June	174.4
July	173.3
August	174.0
September	174.6
October	174.3
November	173.6
December 2001	173.4

[(173.4 – 172.2)/172.2]. Say now a person's annual salary at the end of 2000 was £50,000 and then at the end of 2001 it was £54,000. The salary increase is 8% as before [(£54,000 – 50,000)/50,000]. This person should be satisfied as compared to the CPI increase of 0.70% there has been a real increase in the salary and thus in the spending power of the individual.

## Time series deflation

In order to determine the real value in the change of a commodity, in this case salary from the previous section, we can use **time series deflation**. Time series deflation is illustrated as follows using first the information from Table 11.6:

Base value of the salary at the end of 1989 is

£50,000/year

At the end of 1989, the base salary index is,

$$\frac{50,000}{50,000} * 100 = 100$$

At the end of 1990, the salary index to the base period is,

$$\frac{54,000}{50,000} * 100 = 108$$

Ratio of the CPI at the base period to the new period is,

$$\frac{118.8}{129.9} = 0.9145$$

Multiply the salary index in 1990 by the CPI ratio to give the real value index (RVI) or,

$$108 * 0.9145 = 98.77$$

This means to say that the real value of the salary has in fact declined by 1.23% (100.00 – 98.77).

If we do the same calculation using the CPI for 2001 using Table 11.7 then we have the following:

Base value of the salary at the end of 2000 is

£50,000/year

At the end of 2001, the base salary index is,

$$\frac{50,000}{50,000} * 100 = 100$$

At the end of 2001, the salary index to the base period is,

$$\frac{54,000}{50,000} * 100 = 108$$

Ratio of the CPI at the base period to the new period is

$$\frac{172.2}{173.4} = 0.9931$$

Multiply the salary index in 2001 by the CPI ratio to give the RVI or,

$$108 * 0.9931 = 107.25$$

This means to say that the real value of the salary has increased by 7.25%.

In summary, if you have a time series,  $x$ -values of a commodity and an index series,  $I$ -values, over the same period,  $n$ , then the **RVI of a commodity** for this period is,

$$\text{RVI} = \frac{\text{Current value of commodity}}{\text{Base value of commodity}} * \frac{\text{Base indicator}}{\text{Current indicator}} * 100$$

$$\text{RVI} = \frac{x_n}{x_0} * \frac{I_0}{I_n} * 100 \quad 11(\text{iii})$$

If we substitute in equation 11(iii) the salary and CPI information for 1990 we have the following:

$$\text{RVI} = \frac{54,000}{50,000} * \frac{118.8}{129.9} * 100 = 98.77$$

This means a real decrease of 1.23%.

Similarly, if we substitute in equation 11(iii) the salary and CPI information for 2000 we have the following:

$$\text{RVI} = \frac{54,000}{50,000} * \frac{172.2}{173.4} * 100 = 107.25$$

This means a real increase of 7.25%.

Notice that the commodity ratio and the indicator ratio are in the reverse order since we are deflating the value of the commodity according to the increase in the consumer price.

## Relative Regional Indexes

Index numbers may be used to compare data between one region and another. For example, we might be interested to compare the cost of living in London to that of New York, Paris, Tokyo, and Los Angeles or the productivity of one production site to others. When we use indexes in this manner the time variable is not included.

### Selecting the base value

When we use indexes to compare regions to others, we first decide what our base point is for

comparison and then develop the relative regional index (**RRI**), from this base value.

$$\begin{aligned} \text{Relative regional index} &= \frac{\text{Value at other region}}{\text{Value at base region}} \\ * 100 &= \frac{V_o}{V_b} * 100 \end{aligned}$$

Again, we multiply the ratio by 100 so that the calculated index values represent a percentage change. As an illustration, when I was an engineer in Los Angeles our firm was looking to open a design office in Europe. One of the criteria for selection was the cost of labour in various selected European countries compared to the United States. This type of comparison is illustrated in the following example.

### Illustration by comparing the cost of labour

In Table 11.8 are data on the cost of labour in various countries in terms of the statutory

*Table 11.8* The cost of labour.

Country	Minimum wage plus social security contributions as percent of labour cost of average worker (%)	Index, United States = 100	Index, Britain = 100	Index, France = 100
Australia	46	139	107	85
Belgium	40	121	93	74
Britain	43	130	100	80
Canada	36	109	84	67
Czech Republic	33	100	77	61
France	54	164	126	100
Greece	51	155	119	94
Ireland	49	148	114	91
Japan	32	97	74	59
Luxembourg	50	152	116	93
New Zealand	42	127	98	78
Poland	35	106	81	65
Portugal	50	152	116	93
Slovakia	44	133	102	81
South Korea	25	76	58	46
Spain	37	112	86	69
United States	33	100	77	61

minimum wage plus the mandatory social security contributions as a percentage of the labour costs of the average worker in that country.<sup>4</sup> In Column 3, we have converted the labour cost value into an index using the United States as the base value of 100. This is determined by the calculation  $(33\%/33\%) * 100$ . The base values of the other countries are then determined by the ratio of that country's value to that of the United States. For example, the index for Australia is 139,  $[(46\%/33\%) * 100]$  for South Korea it is 76,  $[(25\%/33\%) * 100]$  and for Britain it is 130,  $[(43\%/33\%) * 100]$ . We interpret this index data by saying that the cost of labour in Australia is 39% more than in the United States; 24% less in South Korea than in the United States  $(100\% - 76\%)$ ; and 30% more in Britain than in the United States.

Column 4 of Table 11.8 gives comparisons using Britain as the base country such that the base value for Britain is 100  $[(43\%/43\%) * 100]$ . We interpret this data in Column 4, for example, by saying that compared to Britain, the labour cost in Australia is 7% more, 16% more in Portugal and 16% less in Canada. Column 5 gives similar index information using France as the base country with an index of 100  $[(54\%/54\%) * 100]$ . Here, for example, the cost of labour in Australia is 15% less than in France, in Britain it is 20% less, and in South Korea it is a whopping 54% less than in France. In fact from Column 5 we see that France is the most expensive country in terms of the cost of labour and this in part explains why labour intensive industries, particularly manufacturing, relocate to lower cost regions.

## Weighting the Index Number

Index numbers may be unweighted or weighted according to certain criteria. The unweighted

index number means that each item in arriving at the index value is considered of equal importance. In the weighted index number, emphasis or weighting is put onto factors such as quantity or expenditure in order to calculate the index.

## Unweighted index number

Consider the information in Table 11.9 that gives the price of a certain 11 products bought in a hypermarket in £UK for the years 2000 and 2005. If we use equation 11(ii) then the price index is,

$$I_P = \frac{P_n}{P_0} * 100 = \frac{96.16}{74.50} * 100 = 129.07$$

To the nearest whole number this is 129, which indicates that in using the items given, prices rose 29% in the period 2000 to 2005. Now, for example, assume that an additional item, a laptop computer is added to Table 11.9 to give the

**Table 11.9** Eleven products purchased in a hypermarket.

Item and unit size (weight, volume, or unit)	2000, $P_0$ (£/unit)	2005, $P_n$ (£/unit)
Bread, loaf	1.10	1.35
Wine, 75 cl	3.45	4.50
Instant coffee, 200 g	5.20	6.90
Cheese, kg	17.50	22.50
Cereals, 750 g	4.50	5.18
Lettuce, each	1.10	1.35
Apples, kg	2.60	3.60
Chicken, kg	20.50	27.00
Milk, litre	0.70	0.93
Fish, kg	18.00	22.50
Petrol, litre	0.95	1.70
Total	74.50	96.16

<sup>4</sup>Economic and financial indicators, *The Economist*, 2 April 2005, p. 88.

revised Table 11.10. Again using equation 11(ii) the price index is,

$$I_p = \frac{P_n}{P_0} * 100 = \frac{1,447.51}{2,925.60} * 100$$

$$= 49.48 \text{ or an index of } 49$$

This indicates that prices have declined by 51% (100 – 49) in the period 2000 to 2005. We know intuitively that this is not the case.

In determining these price indexes using equation 11(ii), we have used an **unweighted aggregate index** meaning that in the calculation each item in the index is of equal importance. In a similar manner we can use equation 11(i) to calculate an unweighted quantity index. This is a major disadvantage of an unweighted index as it neither attaches importance or weight to the quantity of each of the goods purchased nor to price changes of high volume purchased items and to low volume purchased items. For example, a family may purchase

200 lettuces/year but probably would only purchase a laptop computer say every 5 years. Thus, to be more meaningful we should use a **weighted price index**. The concept of weighting or putting importance on items of data was first introduced in Chapter 2.

## Laspeyres weighted price index

The Laspeyres weighted price index, after its originator, is determined by the following relationship:

$$\text{Laspeyres weighted price index} = \frac{\sum P_n Q_0}{\sum P_0 Q_0} * 100 \quad 11(\text{iv})$$

Here,

- $P_n$  is the price in the current period.
- $P_0$  is the price in the base period.
- $Q_0$  is the quantity consumed in the base period.

Note that with this method, the quantities in the base period,  $Q_0$  are used in both the numerator and the denominator of the equation. In addition, the value of the denominator  $\sum P_0 Q_0$  remains constant for each index and this makes comparison of successive indexes simpler where the index for the first period is 100.0. The Table 11.11 gives the calculation procedure for the Laspeyres price index for the items in Table 11.9 with the addition that here the quantities consumed in the base period 2000 are also indicated. Here we have assumed that the quantity of laptop computers consumed is  $\frac{1}{6}$  or 0.17 for the 6-year period between 2000 and 2005. Thus, from equation 11(iv),

Laspeyres price index in 2000 is,

$$\frac{\sum P_n Q_0}{\sum P_0 Q_0} * 100 = \frac{7,466.50}{7,466.50} * 100$$

$$= 100.00 \text{ or } 100$$

**Table 11.10** Twelve products purchased in a hypermarket.

Item and unit size (weight, volume, or unit)	2000, $P_0$ (£/unit)	2005, $P_n$ (£/unit)
Bread, loaf	1.10	1.35
Wine, 75 cl	3.45	4.50
Instant coffee, 200 g	5.20	6.90
Cheese, kg	17.50	22.50
Cereals, 750 g	4.50	5.18
Lettuce, each	1.10	1.35
Apples, kg	2.60	3.60
Chicken, kg	20.50	27.00
Milk, litre	0.70	0.93
Fish, kg	18.00	22.50
Petrol, litre	0.95	1.70
Laptop computer	2,850.00	1,350.00
<b>TOTAL</b>	<b>2,925.60</b>	<b>1,447.51</b>

Table 11.11 Laspeyres price index.

Item and unit size (weight, volume, or unit)	2000, $P_0$ (£/unit)	2005, $P_n$ (£/unit)	Quantity (units) consumed in 2000, $Q_0$	$P_0 * Q_0$	$P_n * Q_0$
Bread, loaf	1.10	1.35	150	165.00	202.50
Wine, 75 cl	3.45	4.50	120	414.00	540.00
Instant coffee, 200 g	5.20	6.90	50	260.00	345.00
Cheese, kg	17.50	22.50	60	1,050.00	1,350.00
Cereals, 750 g	4.50	5.18	25	112.50	129.50
Lettuce, each	1.10	1.35	100	110.00	135.00
Apples, kg	2.60	3.60	25	65.00	90.00
Chicken, kg	20.50	27.00	120	2,460.00	3,240.00
Milk, litre	0.70	0.93	300	210.00	279.00
Fish, kg	18.00	22.50	40	720.00	900.00
Petrol, litre	0.95	1.70	1,500	1,425.00	2,550.00
Laptop computer	2,850.00	1,350.00	0.17	475.00	225.00
TOTAL	2,925.60	1,447.51	2,490.17	7,466.50	9,986.00

Laspeyres price index in 2000 is,

$$\frac{\sum P_n Q_0}{\sum P_0 Q_0} * 100 = \frac{9,986.00}{7,466.50} * 100$$

$$= 133.76 \text{ or } 134 \text{ rounding up.}$$

Thus, if we have selected a representative sample of goods we conclude that the price index for 2005 is 134 based on a 2000 index of 100. This is the same as saying that in this period prices have increased by 34%.

With the Laspeyres method we can compare index changes each year when we have the new prices. For example, if we had prices in 2003 for the same items, and since we are using the quantities for the base year, we can determine a new index for 2003. A disadvantage with this method is that it does not take into account the change in consumption patterns from year to year. For example, we may purchase less of a particular item in 2005 than we purchased in 2000.

## Paasche weighted price index

The Paasche price index, again after its originator, is calculated in a similar manner to the Laspeyres index except that now current quantities in period  $n$  are used rather than quantities in the base period. The Paasche equation is,

$$\text{Paasche price index} = \frac{\sum P_n Q_n}{\sum P_0 Q_n} * 100 \quad 11(v)$$

Here,

- $P_n$  is the price in the current period.
- $P_0$  is the price in the base period.
- $Q_n$  is the quantity consumed in the current period  $n$ .

Thus, in the Paasche weighted price index, unlike, the Laspeyres weighted price index, the value of the denominator  $\sum P_0 Q_n$  changes according to the period with the value of  $Q_n$ . The Paasche price index is illustrated by Table 11.12, which has the same prices for the base period but



Table 11.12 Paasche price index.

Item and unit size (weight, volume, or unit)	2000, $P_0$ (£/unit)	2005, $P_n$ (£/unit)	Quantity consumed in 2005, $Q_n$	$P_0 * Q_n$	$P_n * Q_n$
Bread, loaf	1.10	1.35	75	82.50	101.25
Wine, 75 cl	3.45	4.50	80	276.00	360.00
Instant coffee, 200 g	5.20	6.90	60	312.00	414.00
Cheese, kg	17.50	22.50	20	350.00	450.00
Cereals, 750 g	4.50	5.18	10	45.00	51.80
Lettuce, each	1.10	1.35	200	220.00	270.00
Apples, kg	2.60	3.60	50	130.00	180.00
Chicken, kg	20.50	27.00	200	4,100.00	5,400.00
Milk, litre	0.70	0.93	300	210.00	279.00
Fish, kg	18.00	22.50	80	1,440.00	1,800.00
Petrol, litre	0.95	1.70	800	760.00	1,360.00
Laptop computer	2,850.00	1,350.00	0.17	475.00	225.00
TOTAL	2,925.60	1,447.51	1,875.17	8,400.50	10,891.05

the quantities are for the current consumption period. These revised quantities show that perhaps the family is becoming more health conscious, in that the consumption of bread, wine, coffee, cheese, and petrol (family members walk) is down whereas the consumption of lettuce, apples, fish, and chicken (white meat) is up.

Thus, using equation 11(v),

Paasche price index in 2000 is,

$$\frac{\sum P_n Q_n}{\sum P_0 Q_n} * 100 = \frac{8,400.50}{8,400.50} * 100$$

$$= 100.00 \text{ or } 100$$

Paasche price index in 2005 is,

$$\frac{\sum P_n Q_n}{\sum P_0 Q_n} * 100 = \frac{10,891.05}{8,400.50} * 100$$

$$= 129.65 \text{ or } 130 \text{ rounding up.}$$

Thus, with the Paasche index using revised consumption patterns it indicates that the prices have increased 30% in the period 2000 to 2005.

### Average quantity-weighted price index

In the Laspeyres method we used quantities consumed in early periods and in the Paasche method quantities consumed in later periods. As we see from Tables 11.11 and 11.12 there were changes in consumption patterns so that we might say that the index does not fairly represent the period in question. An alternative approach to the Laspeyres and Paasches methods is to use fixed quantity values that are considered representative of the consumption patterns within the time periods considered. These fixed quantities can be the average quantities consumed within the time periods considered or some other appropriate fixed values. In this case, we have an **average quantity weighted price index** as follows:

Average quantity-weighted price index

$$= \frac{\sum P_n Q_a}{\sum P_0 Q_a} * 100 \quad 11(vi)$$



Here,

- $P_n$  is the price in the current period.
- $P_0$  is the price in the base period.
- $Q_a$  is the average quantity consumed in the total period in consideration.

The new data is given in Table 11.13. From equation 11(vi) using this information,

Average quantity weighted price index in 2000 is,

$$\frac{\sum P_n Q_a}{\sum P_0 Q_a} * 100 = \frac{7,933.50}{7,933.50} * 100$$

$$= 100.00 \text{ or } 100$$

Average quantity weighted price index in 2005 is,

$$\frac{\sum P_n Q_a}{\sum P_0 Q_a} * 100 = \frac{10,438.53}{7,933.50} * 100$$

$$= 131.58 \text{ or } 132.$$

Rounding up this indicates that prices have increased 32% in the period. This average quantity consumed is in fact a fixed quantity and so this approach is sometimes referred to as a **fixed weight aggregate price index**. The usefulness of this index is that we have the flexibility to choose the base price  $P_0$  and the fixed weight  $Q_a$ . Here we have used an average weight but this fixed quantity can be some other value that we consider more appropriate.

Table 11.13 Average price index

Item and unit size (weight, volume, or unit)	2000, $P_0$ (£/unit)	2005, $P_n$ (£/unit)	Average quantity consumed between 2000 and 2005, $Q_a$	$P_0 * Q_a$	$P_n * Q_a$
Bread, loaf	1.10	1.35	112.50	123.75	151.88
Wine, 75 cl	3.45	4.50	100.00	345.00	450.00
Instant coffee, 200 g	5.20	6.90	55.00	286.00	379.50
Cheese, kg	17.50	22.50	40.00	700.00	900.00
Cereals, 750 g	4.50	5.18	17.50	78.75	90.65
Lettuce, each	1.10	1.35	150.00	165.00	202.50
Apples, kg	2.60	3.60	37.50	97.50	135.00
Chicken, kg	20.5	27.00	160.00	3,280.00	4,320.00
Milk, litre	0.70	0.93	300.00	210.00	279.00
Fish, kg	18.00	22.50	60.00	1,080.00	1,350.00
Petrol, litre	0.95	1.70	1,150.00	1,092.50	1,955.00
Laptop computer	2,850.00	1,350.00	0.17	475.00	225.00
TOTAL	2,925.60	1,447.51	1,875.17	7,933.50	10,438.53

This chapter has introduced relative time-based indexes, RRIs, and weighted indexes as a way to present and analyse statistical data.

### Relative time-based indexes

The most common relative time-based indexes are the quantity and price index. In their most common form these indexes measure the relative change over time relative to a given fixed base value. The base value is converted to 100 so that the relative values show a percentage change. An often used price index is the CPI which indicates the change in prices over time and thus is a relative measure of inflation. Rather than having a fixed base we can have rolling index where the base value is the previous period so that the change we measure is relative to the previous period. This is how we would record annual or monthly changes. When the index base is too far in the past the index values may become too high to be meaningful. In this case, we convert the historical sales index to 100 by dividing this value by itself and multiplying by 100. The new relative index values are then the old values divided by the historical index value.

Relative index values can be compared to others to see if there is a relationship between one index and another. This is analogous to causal regression analysis where we establish whether the change in one variable is caused by the change in another variable. A useful comparison of indexes is to compare the index of wage or salary changes to see if they are in line with the change in the CPI. To do this we use a time series deflation which determines the real value in the change of a commodity.

### Relative regional indexes

The goal of relative regional indexes (RRIs) is to compare the data values at one region to that of a base region. Some RRIs might be the cost of living in other locations compared to say New York; the price of housing in major cities compared to say London; or as illustrated in the chapter, the cost of labour compared to France. There can be many RRIs depending on the values that we wish to compare.

### Weighting the index

An unweighted index is one where each element used to calculate the index is considered to have equal value. A weighted price index is where different weights are put onto the index to indicate their importance in calculating the index. The Laspeyres price index is where the index is weighted by multiplying the price in the current period, by the quantity of that item consumed in the base period, and dividing the total value by the sum of the product of the price in the base period and the consumption in the base period. A criticism of this index is that if the time period is long it does not take into account changing consumption patterns. An alternative to the Laspeyres index is the Paasche weighted price index, which is the ratio of total product of current consumption and current price, divided by the total product of current consumption and base price. An alternative to both the Laspeyres and Paasche index is to use an average of the quantity consumed during the period considered. In this way, the index is fairer and more representative of consumption patterns in the period.

## EXERCISE PROBLEMS

### 1. Backlog

#### Situation

Fluor is a California-based engineering and constructing company that designs and builds power plants, oil refineries, chemical plants, and other processing facilities. In the following table are the backlog revenues of the firm in billions of dollar since 1988.<sup>5</sup> Backlog is the amount of work that the company has contracted but which has not yet been executed. Normally, the volume of work is calculated in terms of labour hours and material costs and this is then converted into estimated revenues. The backlog represents the amount of work that will be completed in the future.

Year	Backlog (\$billions)	Year	Backlog (\$billions)	Year	Backlog (\$billions)
1988	6.659	1994	14.022	2000	10.000
1989	8.361	1995	14.725	2001	11.500
1990	9.558	1996	15.800	2002	9.710
1991	11.181	1997	14.400	2003	10.607
1992	14.706	1998	12.645	2004	14.766
1993	14.754	1999	9.142	2005	14.900

#### Required

1. Develop the quantity index numbers for this data where 1988 has an index value of 100.
2. How would you describe the backlog of the firm, based on 1988, in 1989, 2000, and 2005?
3. Develop the quantity index for this data where the year 2000 has an index value of 100.
4. How would you describe the backlog of the firm, based on 2000, in 1989, 1993, and 2005?
5. Why is an index number based on 2000 preferred to an index number of 1988?
6. Develop a rolling quantity index from 1988 based on the change from the previous period.
7. Using the rolling quantity index, how would you describe the backlog of the firm, in 1990, 1994, 1998, and 2004?

<sup>5</sup> Fluor Corporation Annual reports.

## 2. Gold

### Situation

The following table gives average spot prices of gold in London since 1987.<sup>6</sup> In 1969 the price of gold was some \$50/ounce. In 1971 President Nixon allowed the \$US to float by eliminating its convertibility into gold. Concerns over the economy and scarcity of natural resources resulted in the gold price reaching \$850/ounce in 1980 which coincided with peaking inflation rates. The price of gold bottomed out in 2001.

Year	Gold price (\$/ounce)	Year	Gold price (\$/ounce)
1987	446	1997	331
1988	437	1998	294
1989	381	1999	279
1990	384	2000	279
1991	362	2001	271
1992	344	2002	310
1993	360	2003	364
1994	384	2004	410
1995	384	2005	517
1996	388		

### Required

1. Develop the price index numbers for this data where 1987 has an index value of 100.
2. How would you describe gold prices, based on 1987, in 1996, 2001, and 2005?
3. Develop the price index numbers for this data where the year 1996 has an index value of 100.
4. How would you describe gold prices, based on 1996, in 1987, 2001, and 2005?
5. Why is an index number based on 1996 preferred to an index number of 1987?
6. Develop a rolling price index from 1987 based on the change from the previous period.
7. Using the rolling price index, which year saw the biggest annual decline in the price of gold?
8. Using the rolling price index, which year saw the biggest annual increase in the price of gold?

## 3. United States gasoline prices

### Situation

The following table gives the mid-year price of regular gasoline in the United States in cents/gallon since 1990<sup>7</sup> and the average crude oil price for the same year in \$/bbl.<sup>8</sup>

<sup>6</sup>Newmont, 2005 Annual Report.

<sup>7</sup>US Department of Energy, <http://www.doe.gov> (consulted July 2006).

<sup>8</sup><http://www.wtrg.com/oil> (consulted July 2006).

Year	Price of regular grade gasoline (cents/US gallon)	Oil price (\$/bbl)
1990	119.10	20
1991	112.40	38
1992	112.10	20
1993	106.20	19
1994	116.10	18
1995	112.10	19
1996	120.10	20
1997	121.80	22
1998	100.40	19
1999	121.20	12
2000	142.00	15
2001	134.70	30
2002	136.50	25
2003	169.30	25
2004	185.40	27
2005	251.90	35
2006	292.80	62

### Required

1. Develop the price index for regular grade gasoline where 1990 has an index value of 100.
2. How would you describe gasoline prices based on 1990, in 1993, 1998, and 2005?
3. Develop the price index numbers for this data where 2000 has an index value of 100.
4. How would you describe gasoline prices, based on 2000, in 1993, 1998, and 2005?
5. Why might an index number based on 2000 be preferred to an index number of 1990?
6. Develop a rolling price index from 1990 based on the change from the previous period.
7. Using the rolling price index, which year saw the biggest annual increase in the price of regular gasoline?
8. Develop the price index for crude oil prices where 1990 has an index value of 100.
9. Plot the index values of the gasoline prices developed in Question 1 to the crude oil index values developed in Question 8.
10. What are your comments related to the graphs you developed in Question 9?

## 4. Coffee prices

### Situation

The following table gives the imported price of coffee into the United Kingdom since 1975 in United States cents/pound.<sup>9</sup>

<sup>9</sup> International Coffee Organization, <http://www.ico.org> (consulted July 2006).

Year	US cents/1b	Year	US cents/1b
1975	329.17	1990	1,119.13
1976	455.65	1991	1,066.80
1977	1,009.11	1992	872.84
1978	809.51	1993	817.9
1979	979.83	1994	1,273.55
1980	1,011.30	1995	1,340.47
1981	804.84	1996	1,374.08
1982	734.45	1997	1,567.51
1983	730.29	1998	1,477.39
1984	699.54	1999	1,339.49
1985	923.46	2000	1,233.10
1986	965.52	2001	1,181.65
1987	1,103.30	2002	1,273.58
1988	1,102.09	2003	1,421.21
1989	1,027.61	2004	1,530.94

### Required

1. Develop the price index for the imported coffee prices where 1975 has an index value of 100.
2. How would you describe coffee prices based on 1975, in 1985, 1995, and 2004?
3. Develop the price index for the imported coffee prices where 1990 has an index value of 100.
4. How would you describe coffee prices based on 1990, in 1985, 1995, and 2004?
5. Develop the price index for the imported coffee prices where 2000 has an index value of 100.
6. How would you describe coffee prices based on 2000, in 1985, 1995, and 2004?
7. Which index base do you think is the most appropriate?
8. Develop a rolling price index from 1975 based on the change from the previous period.
9. Using the rolling price index, which year and by what amount was the biggest annual increase in the price of imported coffee?
10. Using the rolling price index, which year and by what amount, was the annual decrease in the price of imported coffee?
11. Why are coffee prices not a good measure of the change in the cost of living?

## 5. Boeing

### Situation

The following table gives summary financial and operating data for the United States Aircraft Company Boeing.<sup>10</sup> All the data is in \$US millions except for the earnings per share.

<sup>10</sup>The Boeing Company 2005 Annual Report.

	2005	2004	2003	2002	2001
Revenues	54,845	52,457	50,256	53,831	57,970
Net earnings	2,572	1,872	718	492	2,827
Earnings/share	3.19	2.24	0.85	2.84	3.40
Operating margins (%)	5.10	3.80	0.80	6.40	6.20
Backlog	160,473	109,600	104,812	104,173	106,591

### Required

1. Develop the index numbers for revenues using 2005 as the base.
2. How would you describe the revenues for 2001 using the base developed in Question 1?
3. Develop the index numbers for earnings/share using 2001 as the base?
4. How would you describe the earnings/share for 2005 using the base developed in Question 3?
5. Develop a rolling index for revenues since 2001.
6. Use the index values developed in Question 5, how would you describe the progression of revenues?

## 6. Ford Motor Company

### Situation

The following table gives selected financial data for the Ford Motor Company since 1992.<sup>11</sup>

Year	Revenues automotive (\$millions)	Net income total company (\$millions)	Stock price, high (\$/share)	Stock price, low (\$/share)	Dividends (\$/share)	Vehicle sales North America units 000s
1992	84,407	−7,835	8.92	5.07	0.80	3,693
1993	91,568	2,529	12.06	7.85	0.80	4,131
1994	107,137	5,308	12.78	9.44	0.91	4,591
1995	110,496	4,139	12.00	9.03	1.23	4,279
1996	116,886	4,446	13.59	9.94	1.47	4,222
1997	121,976	6,920	18.34	10.95	1.65	4,432
1998	118,017	22,071	33.76	15.64	1.72	4,370
1999	135,029	7,237	37.30	25.42	1.88	4,787
2000	140,777	3,467	31.46	21.69	1.80	4,933
2001	130,827	−5,453	31.42	14.70	1.05	4,292
2002	134,425	−980	18.23	6.90	0.40	4,402
2003	138,253	495	17.33	6.58	0.40	4,020
2004	147,128	3,487	17.34	12.61	0.40	3,915
2005	153,503	2,024	14.75	7.57	0.40	

<sup>11</sup>Ford Motor Company Annual Reports, 2002 and 2005.

### Required

1. Develop the index numbers for revenues using 1992 as the base.
2. How would you describe the revenues for 2005 using the base developed in Question 1?
3. Develop the rolling index for revenues starting from 1992.
4. Using the rolling index based on the previous period, in which years did the revenues decline, and by how much?
5. Develop the index numbers for North American vehicle sales using 1992 as the base.
6. Based on the index numbers developed in Question 5 which was the best comparative year for vehicle sales, and which was the worst?
7. From the information given, and from the data that you have developed, how would you describe the situation of the Ford Motor Company?

## 7. Drinking

### Situation

In Europe, alcohol consumption rates are rising among the young. The following table gives the percentage of 15- and 16-year olds who admitted to being drunk 3 times or more in a 30-day period in 2003.<sup>12</sup>

Country	Percentage
Britain	23.00
Denmark	26.00
Finland	16.00
France	3.00
Germany	10.00
Greece	3.00
Ireland	26.00
Italy	7.00
Portugal	3.00
Sweden	9.00

### Required

1. Using Britain as the base, develop a relative regional index for the percentage of 15- and 16-year olds who admitted to being drunk 3 times or more in a 30-day period.
2. Using the index for Britain developed in Question 1, how would you describe the percentage of 15- and 16-year olds who admitted to being drunk 3 times or more in a 30-day period in Ireland, Greece, and Germany?
3. Using France as the base, develop a relative regional index for the percentage of 15- and 16-year olds who admitted to being drunk 3 times or more in a 30-day period.
4. Using the index for France developed in Question 3, how would you describe the percentage of 15- and 16-year olds who admitted to being drunk 3 times or more in a 30-day period in Ireland, Greece, and Germany?

<sup>12</sup>Europe at tipping point, *International Herald Tribune*, 26 June 2006, p. 1 and 4.



5. Using Denmark as the base, develop a relative regional index for the percentage of 15- and 16-year olds who admitted to being drunk 3 times or more in a 30-day period.
6. Using the index for Denmark developed in Question 3, how would you describe the percentage of 15- and 16-year olds who admitted to being drunk 3 times or more in a 30-day period in Ireland, Greece, and Germany?
7. Based on the data what general conclusions can you draw?

## 8. Part-time work

### Situation

The following table gives the people working part time in 2005 by country as a percentage of total employment and also the percentage of those working part time who are women. Part-time work is defined as working less than 30 hours/week.<sup>13</sup>

Country	Working part time, percentage of total employment	Percentage of part timers who are women
Australia	27.00	68.30
Austria	16.00	83.80
Belgium	17.80	80.80
Britain	24.50	77.30
Canada	17.90	68.60
Denmark	17.70	64.10
Finland	11.50	63.60
France	14.00	79.10
Germany	22.00	81.40
Greece	6.00	69.60
Ireland	18.00	79.10
Italy	15.00	78.00
Japan	26.00	67.70
The Netherlands	36.00	76.30
New Zealand	22.00	74.80
Norway	21.00	74.60
Portugal	10.00	67.90
Spain	12.00	78.00
Sweden	14.50	69.50
Switzerland	25.50	82.70
Turkey	5.50	59.40
United States	13.00	68.40

### Required

1. Using the United States as the base, develop a relative regional index for the percentage of people working part time.
2. Using the index for the United States developed in Question 1, how would you describe the percentage of people working part time in Australia, Greece, and Switzerland?

<sup>13</sup>Economic and financial indicators, *The Economist*, 24 June 2006, p. 110.

3. Using the Netherlands as the base, develop a relative regional index for the percentage of people working part time.
4. Using the index for the Netherlands developed in Question 3, how would you describe the percentage of people working part time in Australia, Greece, and Switzerland? What can you say about the part-time employment situation in the Netherlands?
5. Using Britain as the base, develop a relative regional index for the percentage of people working part time who are women?
6. Using the index for Britain as developed in Question 5, how would you describe the percentage of people working part time who are women for Australia, Greece, and Switzerland?

## 9. Cost of living

### Situation

The following table gives the purchase price, at medium-priced establishments, of certain items and rental costs in major cities worldwide in 2006.<sup>14</sup> These numbers are a measure of the cost of living. The exchange rates used in the tables are £1.00 = \$1.75 = €1.46.

City	Rent of 2 bedroom unfurnished apartment (£/month)	Bus or subway (£/ride)	Compact disc (£)	International newspaper (£/copy)	Cup of coffee including service (£)	Fast food hamburger meal (£)
Amsterdam	926	1.10	15.08	1.78	1.71	4.46
Athens	721	0.55	13.03	1.23	2.88	4.97
Beijing	1,528	N/A	12.08	2.49	2.42	1.46
Berlin	720	1.44	12.34	1.44	1.71	3.26
Brussels	652	1.03	13.70	1.37	1.51	3.77
Buenos Aires	571	0.15	6.88	2.60	0.84	1.58
Dublin	824	1.03	14.06	1.37	2.06	4.05
Johannesburg	553	N/A	17.01	2.21	1.29	1.84
London	1,700	2.00	11.99	1.10	1.90	4.50
Madrid	892	0.75	13.72	1.71	1.58	4.18
New York	1,998	1.14	10.77	0.93	2.26	3.43
Paris	1,303	0.96	11.65	1.37	1.51	4.12
Prague	754	0.41	14.44	1.20	2.17	2.89
Rome	926	0.69	14.58	1.37	1.51	3.91
Sydney	1,104	1.06	11.03	1.63	1.49	2.74
Tokyo	2,352	1.32	12.25	0.74	1.47	2.99
Vancouver	804	1.13	10.61	1.88	1.63	2.79
Warsaw	754	0.43	13.52	1.80	1.98	2.79
Zagreb	754	N/A	13.60	N/A	2.35	2.58

<sup>14</sup> Global/worldwide cost of living survey ranking, 2006, <http://www.finfacts.com/costofliving.htm>.

### Required

1. Using rental costs as the criterion, how do Amsterdam, Berlin, New York, Paris, Sydney, Tokyo, and Vancouver, compare to London?
2. Using rental costs as the criterion, how do Amsterdam, Berlin, New York, Paris, Sydney, Tokyo, and Vancouver, compare to Madrid?
3. Using rental costs as the criterion, how do Amsterdam, Berlin, New York, Paris, Sydney, Tokyo, and Vancouver, compare to Prague?
4. Using the sum of all the purchase items except rent as the criterion, how do Amsterdam, Berlin, New York, Paris, Sydney, Tokyo, and Vancouver, compare to London?
5. Using the sum of all the purchase items except rent as the criterion, how do Amsterdam, Berlin, New York, Paris, Sydney, Tokyo, and Vancouver, compare to Madrid?
6. Using the sum of all the purchase items except rent as the criterion, how do Amsterdam, Berlin, New York, Paris, Sydney, Tokyo, and Vancouver, compare to Prague?
7. Using rental costs as the criterion, how does the most expensive city compare to the least expensive city? Identify the cities.

## 10. Corruption

### Situation

The Berlin-based organization, Transparency International, defines corruption as the abuse of public office for private gain, and measures the degree to which corruption is perceived to exist among a country's public officials and politicians. It is a composite index, drawing on 16 surveys from 10 independent institutions, which gather the opinions of business people and country analysts. Only 159 of the world's 193 countries are included in the survey due to an absence of reliable data from the remaining countries. The scores range from 10 or squeaky clean, to zero, highly corrupt. A score of 5 is the number Transparency International considers the borderline figure distinguishing countries that do not have a serious corruption problem. The following table gives the corruption index for the first 50 countries in terms of being the least corrupt.<sup>15</sup>

Country	Index	Country	Index
Australia	8.8	Kuwait	4.7
Austria	8.7	Lithuania	4.8
Bahrain	5.8	Luxembourg	8.5
Barbados	6.9	Malaysia	5.1
Belgium	7.4	Malta	6.6
Botswana	5.9	Namibia	4.3
Canada	8.4	The Netherlands	8.6

<sup>15</sup>The 2005 Transparency International Corruption Perceptions Index, <http://www.infioplease.com> (consulted July 2006).

Country	Index	Country	Index
Chile	7.3	New Zealand	9.6
Cyprus	5.7	Norway	8.9
Czech Republic	4.3	Oman	6.3
Denmark	9.5	Portugal	6.5
Estonia	6.4	Qatar	5.9
Finland	9.6	Singapore	9.4
France	7.5	Slovakia	4.3
Germany	8.2	Slovenia	6.1
Greece	4.3	South Africa	4.5
Hong Kong	8.3	Spain	7.0
Hungary	5.0	Sweden	9.2
Iceland	9.7	Switzerland	9.1
Ireland	7.4	Taiwan	5.9
Israel	6.3	Tunisia	4.9
Italy	5.0	United Arab Emirates	6.2
Japan	7.3	United Kingdom	8.6
Jordan	5.7	United States	7.6
South Korea	5.0	Uruguay	5.9

### Required

1. From the countries in the list which country is the least corrupt and which is the most corrupt?
2. What is the percentage of countries that are above the borderline limit as defined by Transparency International, in not having a serious corruption problem?
3. Compare Denmark, Finland, Germany, and England using Spain as the base.
4. Compare Denmark, Finland, Germany, and England using Italy as the base.
5. Compare Denmark, Finland, Germany, and England using Greece as the base.
6. Compare Denmark, Finland, Germany, and England using Portugal as the base.
7. What conclusions might you draw from the responses to Questions 3 to 6?

## 11. Road traffic deaths

### Situation

Every year over a million people die in road accidents and as many as 50 million are injured. Over 80% of the deaths are in emerging countries. This dismal toll is likely to get much worse as road traffic increases in the developing world. The following table gives the annual road deaths per 100,000 of the population.<sup>16</sup>

<sup>16</sup> Emerging market indicators, *The Economist*, 17 April 2004, p. 102.

Country	Deaths per 100,000 people	Country	Deaths per 100,000 people
Belgium	16	Luxembourg	17
Britain	5	Mauritius	45
China	16	New Zealand	13
Columbia	18	Nicaragua	23
Costa Rica	19	Panama	18
Dominican Republic	39	Peru	18
Ecuador	18	Poland	12
El Salvador	42	Romania	11
France	4	Russia	20
Germany	6	Saint Lucia	14
Italy	13	Slovenia	14
Japan	8	South Korea	24
Kuwait	21	Thailand	21
Latvia	25	United States	15
Lithuania	22	Venezuela	24

### Required

1. From the countries in the list, which country is the most dangerous to drive and which is the least dangerous?
2. How would you compare Belgium, The Dominican Republic, France, Latvia, Luxembourg, Mauritius, Russia, and Venezuela to Britain?
3. How would you compare Belgium, The Dominican Republic, France, Latvia, Luxembourg, Mauritius, Russia, and Venezuela to the United States?
4. How would you compare Belgium, The Dominican Republic, France, Latvia, Luxembourg, Mauritius, Russia, and Venezuela to Kuwait?
5. How would you compare Belgium, The Dominican Republic, France, Latvia, Luxembourg, Mauritius, Russia, and Venezuela to New Zealand?
6. What are your overall conclusions and what do you think should be done to improve the statistics?

## 12. Family food consumption

### Situation

The following table gives the 1st quarter 2003 and 1st quarter 2004 prices of a market basket of grocery items purchased by an American family.<sup>17</sup> In the same table is the consumption of these items for the same period.

<sup>17</sup> World Food Prices, <http://www.earth-policy.org> (consulted July 2006).

Product unit amount	1st quarter 2003 (\$/unit)	1st quarter 2004 (\$/unit)	1st quarter 2003 quantity (units)	1st quarter 2004 quantity (units)
Ground chuck beef (1 lb)	2.10	2.48	160	220
White bread (20 oz loaf)	1.32	1.36	60	94
Cheerio cereals (10 oz box)	2.78	3.00	15	16
Apples (1 lb)	1.05	1.22	35	42
Whole chicken fryers (1 lb)	1.05	1.24	42	51
Pork chops (1 lb)	3.10	3.42	96	121
Eggs (1 dozen)	1.22	1.59	52	16
Cheddar cheese (1 lb)	3.30	3.46	37	42
Bacon (1 lb)	2.91	3.00	152	212
Mayonnaise (32 oz jar)	3.14	3.27	19	27
Russet potatoes (5 lb bag)	1.89	1.96	42	62
Sirloin tip roast (1 lb)	3.21	3.52	45	48
Whole milk (1 gallon)	2.80	2.87	98	182
Vegetable oil (32 oz bottle)	2.25	2.76	19	33
Flour (5 lb bag)	1.53	1.62	32	68
Corn oil (32 oz bottle)	2.41	3.09	21	72

### Required

1. Calculate an unweighted price index for this data.
2. Calculate an unweighted quantity index for this data.
3. Develop a Laspeyres weighted price index for this data.
4. Develop a Paasche weighted price index using the 1st quarter 2003 for the base price.
5. Develop an average quantity weighted price index using 2003 as the base price period and the average of the consumption between 2003 and 2004.
6. Discuss the usefulness of these indexes.

## 13. Meat

### Situation

A meat wholesaler exports and imports New Zealand lamb, (frozen whole carcasses) United States beef, poultry, United States broiler cuts and frozen pork. Table 1 gives the prices for these products in \$US/ton for the period 2000 to 2005.<sup>18</sup> Table 2 gives the quantities handled by the meat wholesaler in the same period 2000 to 2005.

*Table 1* Average annual price of meat product (\$US/ton).

Product	2000	2001	2002	2003	2004	2005
New Zealand Lamb	2,618.58	2,911.67	3,303.42	3,885.00	4,598.83	4,438.50
Beef, United States	3,151.67	2,843.67	2,765.33	3,396.25	3,788.25	4,172.75
Poultry, United States	592.08	646.17	581.92	611.83	757.25	847.17
Pork, United States	2,048.58	2,074.08	1,795.58	1,885.58	2,070.75	2,161.17

<sup>18</sup>International Commodity Prices, <http://www.fao.org/es/esc/prices/CIWPQueryServlet> (consulted July 2006).

*Table 2* Amount handled each year (tons).

Product	2000	2001	2002	2003	2004	2005
New Zealand Lamb	54,000	67,575	72,165	79,125	85,124	95,135
Beef, United States	105,125	107,150	109,450	110,125	115,125	120,457
Poultry, United States	118,450	120,450	122,125	125,145	129,875	131,055
Pork, United States	41,254	42,584	45,894	47,254	49,857	51,254

### Required

1. Develop a Laspeyres weighted price index using 2000 as the base period.
2. Develop a Paasche weighted price index using 2005 as the base period.
3. Develop an average quantity weighted price index using the average quantities consumed in the period and 2005 as the base period for price.
4. Develop an average quantity weighted price index using as the base both the average quantity distributed in the period and the average price for the period.
5. What are your observations about the data and the indexes obtained?

## 14. Beverages

### Situation

A wholesale distributor supplies sugar, coffee, tea, and cocoa to various coffee shops in the west coast of the United States. The distributor buys these four commodities from its supplier at the prices indicated in Table 1 for the period 2000 to 2005.<sup>19</sup> Table 2 gives the quantities distributed by the wholesaler in the same period 2000 to 2005.

*Table 1* Average annual price of commodity.

Commodity	2000	2001	2002	2003	2004	2005
Sugar (US cents/lb)	8.43	8.70	6.91	7.10	7.16	9.90
Tea, Mombasa (\$US/kg)	1.97	1.52	1.49	1.54	1.55	1.47
Coffee (US cents/lb)	64.56	45.67	47.69	51.92	62.03	82.76
Cocoa (US cents/lb)	40.27	49.03	80.58	79.57	70.26	73.37

*Table 2* Amount distributed each year (kg).

Commodity	2000	2001	2002	2003	2004	2005
Sugar	75,860	80,589	85,197	94,904	104,759	112,311
Tea	29,840	34,441	39,310	47,887	50,966	59,632
Coffee	47,300	52,429	58,727	66,618	73,427	79,303
Cocoa	27,715	29,156	30,640	35,911	41,219	46,545

<sup>19</sup>International Commodity Prices, <http://www.fao.org/es/esc/prices/CIWPQueryServlet> (consulted July 2006).

**Required**

1. Develop a Laspeyres weighted price index using 2000 as the base period.
2. Develop a Paasche weighted price index using 2005 as the base period.
3. Develop an average quantity weighted price index using the average quantities consumed in the period and 2005 as the base period for price.
4. Develop an average quantity weighted price index using as the base both the average quantity distributed in the period and the average price for the period.
5. What are your observations about the data and the indexes obtained?

**15. Non-ferrous metals****Situation**

Table 1 gives the average price of non-ferrous metals in \$US/ton in the period 2000 to 2005.<sup>20</sup> Table 2 gives the consumption of these metals in tons for a manufacturing conglomerate in the period 2000 to 2005.

*Table 1* Average metal price, \$US/ton.

Metal	2000	2001	2002	2003	2004	2005
Aluminium	1,650	1,500	1,425	1,525	1,700	2,050
Copper	1,888	1,688	1,550	2,000	2,800	3,550
Tin	5,600	4,600	4,250	5,500	7,650	7,800
Zinc	1,100	900	800	900	1,150	1,600

*Table 2* Consumption (tons/year).

Metal	2000	2001	2002	2003	2004	2005
Aluminium	53,772	100,041	86,443	63,470	126,646	102,563
Copper	75,000	93,570	106,786	112,678	79,345	126,502
Tin	18,415	13,302	14,919	22,130	21,916	18,535
Zinc	36,158	48,187	32,788	47,011	49,257	31,712

**Required**

1. Develop a Laspeyres weighted price index using 2000 as the base period.
2. Develop a Paasche weighted price index using 2005 as the base period.
3. Develop an average quantity weighted price index using the average quantities consumed in the period and 2005 as the base period for price.
4. Develop an average quantity weighted price index using as the base both the average quantity consumed in the period and the average price for the period.
5. What are your observations about the data and the indexes obtained?

<sup>20</sup>London Metal Exchange, <http://www.lme.co.uk/dataprices> (consulted July 2006).



## 16. Case study: United States energy consumption

### Situation

The following table gives the energy consumption by source in the United States since 1973 in million British Thermal Units (BTUs).<sup>21</sup>

Year	Coal	Natural gas	Petroleum products	Nuclear	Hydroelectric	Biomass	Geothermal	Solar	Wind
1973	12,971,490	22,512,399	34,839,926	910,177	2,861,448	1,529,068	42,605		
1974	12,662,878	21,732,488	33,454,627	1,272,083	3,176,580	1,539,657	53,158		
1975	12,662,786	19,947,883	32,730,587	1,899,798	3,154,607	1,498,734	70,153		
1976	13,584,067	20,345,426	35,174,688	2,111,121	2,976,265	1,713,373	78,154		
1977	13,922,103	19,930,513	37,122,168	2,701,762	2,333,252	1,838,332	77,418		
1978	13,765,575	20,000,400	37,965,295	3,024,126	2,936,983	2,037,605	64,350		
1979	15,039,586	20,665,817	37,123,381	2,775,827	2,930,686	2,151,906	83,788		
1980	15,422,809	20,394,103	34,202,356	2,739,169	2,900,144	2,484,500	109,776		
1981	15,907,526	19,927,763	31,931,050	3,007,589	2,757,968	2,589,563	123,043		
1982	15,321,581	18,505,085	30,231,314	3,131,148	3,265,558	2,615,048	104,746		
1983	15,894,442	17,356,794	30,053,921	3,202,549	3,527,260	2,831,271	129,339		28
1984	17,070,622	18,506,993	31,051,327	3,552,531	3,385,811	2,879,817	164,896	55	68
1985	17,478,428	17,833,933	30,922,149	4,075,563	2,970,192	2,864,082	198,282	111	60
1986	17,260,405	16,707,935	32,196,080	4,380,109	3,071,179	2,840,995	219,178	147	44
1987	18,008,451	17,744,344	32,865,053	4,753,933	2,634,508	2,823,159	229,119	109	37
1988	18,846,312	18,552,443	34,221,992	5,586,968	2,334,265	2,936,991	217,290	94	9
1989	19,069,762	19,711,690	34,211,114	5,602,161	2,837,263	3,062,458	317,163	55,291	22,033
1990	19,172,635	19,729,588	33,552,534	6,104,350	3,046,391	2,661,655	335,801	59,718	29,007
1991	18,991,670	20,148,929	32,845,361	6,422,132	3,015,943	2,702,412	346,247	62,688	30,796
1992	19,122,471	20,835,075	33,526,585	6,479,206	2,617,436	2,846,653	349,309	63,886	29,863
1993	19,835,148	21,351,168	33,841,477	6,410,499	2,891,613	2,803,184	363,716	66,458	30,987
1994	19,909,463	21,842,017	34,670,274	6,693,877	2,683,457	2,939,105	338,108	68,548	35,560
1995	20,088,727	22,784,268	34,553,468	7,075,436	3,205,307	3,067,573	293,893	69,857	32,630
1996	21,001,914	23,197,419	35,756,853	7,086,674	3,589,656	3,127,341	315,529	70,833	33,440
1997	21,445,411	23,328,423	36,265,647	6,596,992	3,640,458	3,005,919	324,959	70,237	33,581
1998	21,655,744	22,935,581	36,933,540	7,067,809	3,297,054	2,834,635	328,303	69,787	30,853
1999	21,622,544	23,010,090	37,959,645	7,610,256	3,267,575	2,885,449	330,919	68,793	45,894
2000	22,579,528	23,916,449	38,403,623	7,862,349	2,811,116	2,906,875	316,796	66,388	57,057
2001	21,914,268	22,905,783	38,333,150	8,032,697	2,241,858	2,639,717	311,264	65,454	69,617
2002	21,903,989	23,628,207	38,401,351	8,143,089	2,689,017	2,649,007	328,308	64,391	105,334
2003	22,320,928	22,967,073	39,047,308	7,958,858	2,824,533	2,811,514	330,554	63,620	114,571
2004	22,466,195	23,035,840	40,593,665	8,221,985	2,690,078	2,982,342	341,082	64,500	141,749
2005	22,830,007	22,607,562	40,441,180	8,133,222	2,714,661	2,780,755	351,671	64,467	149,490

### Required

Using the concept of indexing, describe the consumption pattern of energy in the United States.

<sup>21</sup>Energy Information Administration, *Monthly Energy Review*, June 2006 (posted 27 June 2006), <http://tonto.eia.doe.gov>.

# Appendix I:

## Key terminology and formula in statistics

Expressions and formulas presented in **bold letters** in the textbook can be found in this section in alphabetical order. In this listing when there is another term in **bold letters** it means it is explained elsewhere in this *Appendix I*. At the end of this listing is an explanation of the symbols used in this equation. Further, if you want to know the English equivalent of those that are Greek symbols, you can find that in *Appendix III*.

**A priori probability** is being able to make an estimate of probability based on information already available.

**Absolute** in this textbook context implies presenting data according to the value collected.

**Absolute frequency histogram** is a vertical bar chart on  $x$ -axis and  $y$ -axis. The  $x$ -axis is a numerical scale of the desired class width, and the  $y$ -axis gives the length of the bar which is proportional to the quantity of data in a given class.

**Addition rule for mutually exclusive events** is the sum of the individual probabilities.

**Addition rule for non-mutually exclusive events** is the sum of the individual probabilities less than the probability of the two events occurring together.

**Alternative hypothesis** is another value when the hypothesized value, or null hypothesis, is not correct at the given level of significance.

**Arithmetic mean** is the sum of all the data values divided by the amount of data. It is the same as the **average value**.

**Asymmetrical data** is numerical information that does not follow a **normal distribution**.

**Average quantity weighted price index** is,

$$\frac{\sum P_n Q_a}{\sum P_0 Q_a} * 100$$

where  $P_0$  and  $P_n$  are prices in the base and current period, respectively, and  $Q_a$  is the average quantity consumed during the period under consideration. This index is also referred to as a **fixed weight aggregate price index**.

**Average value** is another term used for **arithmetic mean**.

**Backup** is an auxiliary unit that can be used if the principal unit fails. In a **parallel arrangement** we have backup units.

**Bar chart** is a type of histogram where the  $x$ -axis and  $y$ -axis have been reversed. It can also be called a Gantt chart after the American engineer Henry Gantt.

**Bayesian decision-making implies** that if you have additional information, or based on the fact that *something has occurred*, certain probabilities

may be revised to give *posterior* probabilities (*post* meaning afterwards).

**Bayes' theorem** gives the relationship for statistical probability under statistical dependence.

**Benchmark** is the value of a piece of data which we use to compare other data. It is the reference point.

**Bernoulli process** is where in each trial there are only two possible outcomes, or **binomial**. The probability of any outcome remains fixed over time and the trials are statistically independent. The concept comes from Jacques Bernoulli (1654–1705) a Swiss/French mathematician.

**Bias** in sampling is favouritism, purposely or unknowingly, present in sample data that gives lopsided, misleading, false, or unrepresentative results.

**Bi-modal** means that there are two values that occur most frequently in a dataset.

**Binomial** means that there are only two possible outcomes of an event such as yes or no, right or wrong, good or bad, etc.

**Binomial distribution** is a table or graph showing all the possible outcomes of an experiment for a discrete distribution resulting from a **Bernoulli process**.

**Bivariate data** involves two variables,  $x$  and  $y$ . Any data that is in graphical form is bivariate since a value on the  $x$ -axis has a corresponding value on the  $y$ -axis.

**Boundary limits of quartiles** are  $Q_0$ ,  $Q_1$ ,  $Q_2$ ,  $Q_3$ , and  $Q_4$ , where the indices indicate the quartile value going from the minimum value  $Q_0$  to the maximum value  $Q_4$ .

**Box and whisker plot** is a visual display of quartiles. The box contains the middle 50% of the data. The 1st whisker on the left contains the first 25% of the data and the 2nd whisker on the right contains the last 25%.

**Box plot** is an alternative name for the **box and whisker plot**.

**Category** is a distinct class into which information or entities belong.

**Categorical data** is information that includes a qualitative response according to a name, label, or category such as the categories of Asia, Europe, and the United States or the categories of men and women. With categorical information there may be no quantitative data.

**Causal forecasting** is when the movement of the **dependent variable**,  $y$ , is caused or impacted by the change in value of the **independent variable**,  $x$ .

**Categories** are the **groups** into which data is organized.

**Central limit theory** in sampling states that as the size of the sample increases, there becomes a point when the **distribution of the sample means**,  $\bar{x}$ , can be approximated by the **normal distribution**. If the sample size taken is greater than 30, then the sample distribution of the means can be considered to follow a normal distribution even though the population is not normal.

**Central moving average** in seasonal forecasting is the linear average of four quarters around a given central time period. As we move forwards in time the average changes by eliminating the oldest quarter and adding the most recent.

**Central tendency** is how data clusters around a central measure such as the mean value.

**Characteristic probability** is that which is to be expected or that which is the most common in a statistical experiment.

**Chi-square distribution** is a continuous probability distribution used in this text to test a hypothesis associated with more than two populations.

**Chi-square test** is a method to determine if there is a dependency on some criterion between the proportions of more than two populations.

**Class** is a grouping into which data is arranged. The age groups, 20–29; 30–39; 40–49; 50–59 years are four classes that can be groupings used in market surveys.

**Class range** is the breadth or span of a given class.

**Class width** is an alternative description of the **class range**.

**Classical probability** is the ratio of the number of favourable outcomes of an event divided by the total possible outcomes. Classical probability is also known as **marginal probability or simple probability**.

**Closed-ended frequency distribution** is one where all data in the distribution is contained within the limits.

**Cluster sampling** is where the population is divided into groups, or clusters, and each cluster is then sampled at random.

**Coefficient of correlation,  $r$**  is a measure of the strength of the relation between the **independent variable**  $x$  and the **dependent variable**  $y$ . The value of  $r$  can take any value between  $-1.00$  and  $+1.00$  and the sign is the same as the slope of the regression line.

**Coefficient of determination,  $r^2$**  is another measure of the strength of the relation between the variables  $x$  and  $y$ . The value of  $r^2$  is always positive and less than or equal to the **coefficient of correlation,  $r$** .

**Coefficient of variation** of a dataset is the ratio of the **standard deviation** to the mean value,  $\sigma/\mu$ .

**Collectively exhaustive** gives all the possible outcomes of an experiment.

**Combination** is the arrangement of distinct items regardless of their order. The number of combinations is calculated by the expression,

$${}^nC_x = \frac{n!}{x!(n-x)!}$$

**Conditional probability** is the chance of an event occurring given that another event has already occurred.

**Confidence interval** is the range of the estimate at the prescribed confidence level.

**Confidence level** is the probability value for the estimate, such as a 95%. Confidence level may also be referred to as the **level of confidence**.

**Confidence limits of a forecast** are given by,  $\hat{y} \pm z s_e$ , when we have a sample size greater than 30 and by  $\hat{y} \pm t s_e$ , for sample sizes less than 30. The values of  $z$  and  $t$  are determined by the desired **level of confidence**.

**Constant value** is one that does not change with a change in conditions. The beginning letters of the alphabet,  $a, b, c, d, e, f$ , etc., either lower or upper case, are typically used to represent a constant.

**Consumer price index** is a measure of the change of prices. It is used as a measure of inflation.

**Consumer surveys** are telephone, written, electronic, or verbal consumer responses concerning a given issue or product.

**Continuity correction factor** is applied to a random variable when we wish to use the normal-binomial approximation.

**Continuous data** has no distinct cut-off point and continues from one class to another. The volume of beer in a can may have a nominal value of 33 cl but the actual volume could be 32.3458, 32.9584, or 33.5486 cl, etc. It is unlikely to be exactly 33.0000 cl.

**Continuous probability distribution** is a table or graph where the variable  $x$  can take any value within a defined range.

**Contingency table** indicates data relationships when there are several categories present. It is also referred to as a **cross-classification table**.

**Continuous random variables** can take on any value within a defined range.

**Correlation** is the measurement of the strength of the relationship between variables.

**Counting rules** are the mathematical relationships that describe the possible outcomes, or results, of various types of experiments, or trials.

**Covariance** of random variables is an application of the distribution of random variables often used to analyse the risk associated with financial investments.

**Critical value** in hypothesis testing is that value outside of which the null hypothesis should be rejected. It is the **benchmark** value.

**Cross-classification table** indicates data relationships when there are several categories present. It is also referred to as a **contingency table**.

**Cumulative frequency distribution** is a display of dataset values cumulated from the minimum to the maximum. In graphical form this is called an ogive. It is useful for indicating how many observations lie above or below certain values.

**Curvilinear function** is one that is not linear but curves according to the equation that describes its shape.

**Data** is a collection of information.

**Data array** is raw data that has been sorted in either ascending or descending order.

**Data characteristics** are the units of measurement that describe data such as the weight, length, volume, etc.

**Data point** is a single observation in a dataset.

**Dataset** is a collection of data either unsorted or sorted.

**Degrees of freedom** means the choices that you have taken regarding certain actions.

**Degrees of freedom in a cross-classification table** are  $(\text{No. of rows} - 1) * (\text{No. of columns} - 1)$ .

**Degrees of freedom in a Student-*t* distribution** are given by  $(n - 1)$ , where  $n$  is the sample size.

**Dependent variable** is that value that is a function or is dependent on another variable. Graphically it is positioned on the  $y$ -axis.

**Descriptive statistics** is the analysis of sample data in order to describe the characteristics of that particular sample.

**Deterministic** is where outcomes or decisions made are based on data that are accepted and can be considered reliable or certain. For example, if sales for one month are \$50,000 and costs \$40,000 then it is certain that net income is \$10,000 ( $\$50,000 - \$40,000$ ).

**Deviation about the mean** of all observations,  $x$ , about the mean value  $\bar{x}$ , is zero.

**Discrete data** is information that has a distinct cut-off point such as 10 students, 4 machines, and 144 computers. Discrete data come from the counting process and the data are whole numbers or integer values.

**Discrete random variables** are those **integer values, or whole numbers**, that follow no particular pattern.

**Dispersion** is the spread or the variability in a dataset.

**Distribution of the sample means** is the same as the **sampling distribution of the means**.

**Empirical probability** is the same as **relative frequency probability**.

**Empirical rule for the normal distribution** states that no matter the value of the mean or the standard deviation, the area under the curve is always unity. As examples, 68.26% of all data

falls within  $\pm 1$  standard deviations from the mean, 95.44% falls within  $\pm 2$  standard deviations from the mean, and 99.73% of all data falls within  $\pm 3$  standard deviations from the mean.

**Estimate** in statistical analysis is that value judged to be equal to the population value.

**Estimated standard error of the proportion** is,

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

where  $\bar{p}$  is the sample proportion and  $n$  is the sample size.

**Estimated standard error of the difference between two proportions** is,

$$\hat{\sigma}_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\bar{p}_1 \bar{q}_1}{n_1} + \frac{\bar{p}_2 \bar{q}_2}{n_2}}$$

**Estimated standard deviation of the distribution of the difference between the sample means** is,

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$$

**Estimating** is forecasting or making a judgment about a future situation using entirely, or in part, quantitative information.

**Estimator** is that statistic used to estimate the population value.

**Event** is the outcome of an activity or experiment that has been carried out.

**Expected value of the binomial distribution**  $E(x)$  or the mean value,  $\mu_x$ , is the product of the number of trials and the **characteristic probability**, or  $\mu_x = E(x) = np$ .

**Expected value of the random variable** is the weighted average of the outcomes of an experiment. It is the same as the **mean value** of the random variable and is given by the relationship,  $\mu_x = \sum xP(x) = E(x)$ .

**Experiment** is the activity, such as a sampling process, that produces an event.

**Exponential function** has the form  $y = ae^{bx}$ , where  $x$  and  $y$  are the **independent** and **dependent** variables, respectively, and  $a$  and  $b$  are constants.

**Exploratory data analysis** (EDA) covers those techniques that give analysts a sense about data that is being examined. A stem-and-leaf display and a box and whisker plot are methods in EDA.

**Factorial rule** for the arrangement of  $n$  different objects is  $n! = n(n-1)(n-2)(n-3) \dots (n-n)$ , where  $0! = 1$ .

**Finite population** is a collection of data that has a stated, limited, or a small size. The number of playing cards (52) in a pack is considered finite.

**Finite population multiplier** for a population of size  $N$  and a sample of size  $n$  is,

$$\sqrt{\frac{N-n}{N-1}}$$

**Fixed weight aggregate price index** is the same as the **average quantity weighted price index**.

**Fractiles** divide data into specified fractions or portions.

**Frequency distribution** groups data into defined classes. The distribution can be a table, polygon, or histogram. We can have an **absolute** frequency distribution or a **relative** frequency distribution.

**Frequency polygon** is a line graph connecting the midpoints of the class ranges.

**Functions** in the context of this textbook are those built-in macros in Microsoft Excel. In this book, it is principally the statistical functions that are employed. However, Microsoft Excel contains financial, logic, database, and other functions.

**Gaussian distribution** is another name for the normal distribution after its German originator, Karl Friedrich Gauss (1777–1855).



**Geometric mean** is used when data is changing over time. It is calculated by the  $n$ th root of the growth rates for each year, where  $n$  is the number of years.

**Graphs** are visual displays of data such as line graphs, histograms, or pie charts.

**Greater than ogive** is a cumulative frequency distribution that illustrates data above certain values. It has a negative slope, where the  $y$ -values decrease from left to right.

**Groups** are the units or ranges into which data is organized.

**Histogram** is a vertical bar chart showing data according to a named category or a quantitative class range.

**Historical data** is information that has occurred, or has been collected in the past.

**Horizontal bar chart** is a **bar chart** in a horizontal form where the  $y$ -axis is the class and the  $x$ -axis is the proportion of data in a given class.

**Hypothesis** is a judgment about a situation, outcome, or population parameter based simply on an assumption or intuition with initially no concrete backup information or analysis.

**Hypothesis testing** is to test sample data and make an objective decision based on the results of the test using an appropriate significance level for the hypothesis test.

**Independent variable** in a time series is the value upon which another value is a function or dependent. Graphically the independent variable is always positioned on the  $x$ -axis.

**Index base value** is the real value of a piece of data which is used as the reference point to determine the index number.

**Index number** is the ratio of a certain value to a base value usually multiplied by 100. When the base value equals 100 then the measured

values are a percentage of the base. The index number may be called as the **index value**.

**Index value** is an alternative for the **index number**.

**Inferential statistics** is the analysis of sample data for the purpose of describing the characteristics of the population parameter from which that sample is taken.

**Infinite population** is a collection of data that has such a large size so that by removing or destroying some of the data elements it does not significantly impact the population that remains.

**Integer values** are whole numbers originating from the counting process.

**Interval estimate** gives a range for the estimate of the population parameter.

**Inter-quartile range** is the difference between the values of the 3rd and the 1st quartile in the dataset. It measures the range of the middle half of an ordered dataset.

**Joint probability** is the chance of two events occurring together or in succession.

**Kurtosis** is the characteristic of the peak of the distribution curve.

**Laspeyres weighted price index** is,

$$\frac{\sum P_n Q_0}{\sum P_0 Q_0} * 100$$

where  $P_n$  is the price in the current period,  $P_0$  is the price in the base period and  $Q_0$  is the quantity consumed in the base period.

**Law of averages** implies that the average value of an activity obtained in the long run will be close to the expected value, or the weighted outcome based on each probability of occurrence.

**Least square method** is a calculation technique in **regression analysis** that determines the best

straight line for a series of data that minimizes the error between the actual and forecast data.

**Leaves** are the trailing digits in a **stem-and-leaf display**.

**Left-skewed data** is when the mean of a dataset is less than the median value, and the curve of the distribution tails off to the left side of the  $x$ -axis.

**Left-tail hypothesis test** is used when we are asking the question, “Is there evidence that a value is less than?”

**Leptokurtic** is when the peak of a distribution is sharp, quantified by a small standard deviation.

**Less than ogive** is a cumulative frequency distribution that indicates the amount of data below certain limits. As a graph it has a positive slope such that the  $y$ -values increase from left to right.

**Level of confidence** in estimating is  $(1 - \alpha)$ , where  $\alpha$  is the proportion in the tails of the distribution, or that area outside of the confidence interval.

**Line graph** shows bivariate data on  $x$ -axis and  $y$ -axis. If time is included in the data this is always indicated on the  $x$ -axis.

**Linear regression line** takes the form  $\hat{y} = a + bx$ . It is the equation of the best straight line for the data that minimizes the error between the data points on the regression line and the corresponding actual data from which the regression line is developed.

**Margin of error** is the range of the estimate from the true population value.

**Marginal probability** is the ratio of the number of favourable outcomes of an event divided by the total possible outcomes. Marginal probability is also known as **classical probability or simple probability**.

**Mean proportion of successes**,  $\mu_{\bar{p}} = p$ .

**Mean value** is another way of referring to the **arithmetic mean**.

**Mean value of random data** is the weighted average of all the possible outcomes of the random variable.

**Median** is the middle value of an ordered set of data. It divides the data into two equal halves. The 2nd quartile and the 50th percentile are also the median value.

**Mesokurtic** describes the curve of a distribution when it is intermediate between a sharp peak, **leptokurtic** and a relatively flat peak, or **platykurtic**.

**Mid-hinge** in quartiles is the average of the 3rd and 1st quartile.

**Midpoint** of a class range is the maximum plus the minimum value divided by 2.

**Midrange** is the average of the smallest and the largest observations in a dataset.

**Mid-spread range** is another term for the **inter-quartile range**.

**Mode** is that value that occurs most frequently in a dataset.

**Multiple regression** is when the dependent variable  $y$  is a function of many independent variables. It can be represented by an equation of the general form,  $\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_kx_k$ .

**Mutually exclusive events** are those that cannot occur together.

**Normal-binomial approximation** is applied when  $np \geq 5$  and  $n(1 - p) \geq 5$ . In this case, substituting for the mean value and the standard deviation of the binomial distribution in the normal distribution transformation relationship we have,

$$z = \frac{x - \mu}{\sigma} = \frac{x - np}{\sqrt{npq}} = \frac{x - np}{\sqrt{np(1 - p)}}$$



**Normal distribution**, or the **Gaussian distribution**, is a continuous distribution of a random variable. It is symmetrical, has a single hump, and the mean, median and mode are equal. The tails of the distribution may not immediately cut the  $x$ -axis.

**Normal distribution density function**, which describes the shape of the **normal distribution** is,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-(1/2)[(x-\mu_x)/\sigma_x]^2}$$

**Normal distribution transformation relationship** is,

$$z = \frac{x - \mu_x}{\sigma_x}$$

where  $z$  is the number of standard deviations,  $x$  is the value of the random variable,  $\mu_x$  is the mean value of the dataset and  $\sigma_x$  is the standard deviation of the dataset.

**Non-linear regression** is when the dependent variable is represented by an equation where the power of some or all the independent variables is at least two. These powers of  $x$  are usually integer values.

**Non-mutually exclusive** events are those that can occur together.

**Null hypothesis** is that value that is considered correct in the experiment.

**Numerical codes** are used to transpose qualitative or label data into numbers. This facilitates statistical analysis. For example, if the time period is January, February, March, etc. we can code these as 1, 2, 3, etc.

**Odds** are the chance of winning and are the ratio of the probability of losing to the chances of winning.

**Ogive** is a frequency distribution that shows data cumulatively. A **less than ogive** indicates data less than certain values and a **greater than**

**ogive** shows data more than certain values. An ogive can illustrate **absolute** data or **relative** data.

**One-arm-bandit** is the slang term for the slot machines that you find in gambling casinos. The game of chance is where you put in a coin or chip, pull a lever and hope that you win a lucky combination!

**One-tail hypothesis test** is used when we are interested to know if something is less than or greater than a stipulated value. If we ask the question, "Is there evidence that the value is greater than?" then this would be a **right-tail hypothesis test**. Alternatively, if we ask the question, "Is there evidence that the value is less than?" then this would be a **left-tail hypothesis test**.

**Ordered dataset** is one where the values have been arranged in either increasing or decreasing order.

**Outcomes of a single type of event** are  $k^n$ , where  $k$  is the number of possible events, and  $n$  is the number of trials.

**Outcomes of different types of events** are  $k_1 * k_2 * k_3 * \dots * k_n$ , where  $k_1, k_2, \dots, k_n$  are the number of possible events.

**Outliers** are those numerical values that are either much higher or much lower than other values in a dataset and can distort the value of the central tendency, such as the average, and the value of the dispersion such as the range or standard deviation.

**P** in upper case or capitals is often the abbreviation used for probability.

**Paired samples** are those that are dependent or related, often in a before and after situation. Examples are the weight loss of individuals after a diet programme or productivity improvement after a training programme.

**Pareto diagram** is a combined histogram and line graph. The frequency of occurrence of the data is indicated according to categories on the

histogram and the line graph shows the cumulated data up to 100%. This diagram is a useful auditing tool.

**Parallel bar chart** is similar to a parallel histogram but the  $x$ -axis and  $y$ -axis have been reversed.

**Parallel arrangement** in design systems is such that the components are connected giving a choice to use one path or another. Which ever path is chosen the **system** continues to function.

**Parallel histogram** is a vertical bar chart showing the data according to a category and within a given category there are sub-categories such as different periods. A parallel histogram is also referred to a **side-by-side histogram**.

**Parameter** describes the characteristic of a population such as the weight, height, or length. It is usually considered a fixed value.

**Percentiles** are fractiles that divide ordered data into 100 equal parts.

**Permutation** is a combination of data arranged in a particular order. The number of ways, or permutations, of arranging  $x$  objects selected in order from a total of  $n$  objects is,

$${}^nP_x = \frac{n!}{(n-x)!}$$

**Pictogram** is a diagram, picture, or icon that shows data in a relative form.

**Pictograph** is an alternative name for the **pictogram**.

**Pie chart** is a circle graph showing the percentage of the data according to certain categories. The circle, or pie, contains 100% of the data.

**Platykurtic** is when the curve of a distribution has a flat peak. Numerically this is shown by a larger value of the coefficient of variation,  $\sigma/\mu$ .

**$p$ -value** in hypothesis testing is the observed level of significance from the sample data or the

minimum probable level that we will tolerate in order to accept the null hypothesis of the mean or the proportion.

**Point estimate** is a single value used to estimate the population parameter.

**Poisson distribution** describes events that occur during a given time interval and whose average value in that time period is known. The probability relationship is,

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

**Polynomial function** has the general form  $y = a + bx + cx^2 + dx^3 + \dots + kx^n$ , where  $x$  is the independent variable and  $a, b, c, d, \dots, k$  are constants.

**Population** is all of the elements under study and about which we are trying to draw conclusions.

**Population standard deviation** is the square root of the **population variance**.

**Population variance** is given by,

$$\sigma^2 = \frac{\sum (x - \mu_x)^2}{N}$$

where  $N$  is the amount of data,  $x$  is the particular data value, and  $\mu_x$  is the mean value of the dataset.

**Portfolio risk** measures the exposure associated with financial investments.

**Posterior probability** is one that has been revised after additional information has been received.

**Power of a hypothesis test** is a measure of how well the test is performing.

**Primary data** is that collected directly from the source.

**Probability** is a quantitative measure, expressed as a decimal or percentage value, indicating the likelihood of an event occurring. The value

$[1 - P(x)]$  is the likelihood of the event not occurring.

**Probabilistic** is where there is a degree of uncertainty, or probability of occurrence from the supplied data.

**Quad-modal is** when there are four values in a dataset that occur most frequently.

**Qualitative data** is information that has no numerical response and cannot immediately be analysed.

**Quantitative** data is information that has a numerical response.

**Quartiles** are those three values which divide ordered data into four equal parts.

**Quartile deviation** is one half of the inter-quartile range, or  $(Q_3 - Q_1)/2$ .

**Questionnaires** are evaluation sheets used to ascertain people's opinions of a subject or a product.

**Quota sampling** in market research is where each interviewer in the sampling experiment has a given quota or number of units to analyse.

**Random** implies that any occurrence or value is possible.

**Random sample** is where each item of data in the sample has an equal chance of being selected.

**Random variable** is one that will have different values as a result of the outcome of a random experiment.

**Range** is the numerical difference between the highest and lowest value in a dataset.

**Ratio measurement scale is** where the difference between measurements is based on starting from a base point to give a ratio. The **consumer price index** is usually presented on a ratio measurement scale.

**Raw data** is collected information that has not been organized.

**Real value index (RVI) of a commodity** for a period is,

$$\text{RVI} = \frac{\text{Current value of commodity}}{\text{Base value of commodity}} \\ * \frac{\text{Base indicator}}{\text{Current indicator}} * 100$$

**Regression analysis** is a mathematical technique to develop an equation describing the relationship of variables. It can be used for forecasting and estimating.

**Relative** in this textbook context is presenting data compared to the total amount collected. It can be expressed either as a percentage or fraction.

**Relative frequency histogram** has vertical bars that show the percentage of data that appears in defined class ranges.

**Relative frequency distribution** shows the percentage of data that appears in defined class ranges.

**Relative frequency probability** is based on information or experiments that have previously occurred. It is also known as **empirical probability**.

**Relative price index**  $I_p = (P_n / P_0) * 100$ ,

where  $P_0$  is the price at the base period, and  $P_n$  is the price at another period.

**Relative quantity index**  $I_Q = (Q_n / Q_0) * 100$

where  $Q_0$  is the quantity at the base period and  $Q_n$  is the quantity at another period.

**Relative regional index (RRI)** compares the value of a parameter at one region to a selected base region. It is given by,

$$\frac{\text{Value at other region}}{\text{Value at base region}} * 100 = \frac{V_o}{V_b} * 100$$

**Reliability** is the confidence we have in a product, process, service, work team, individual, etc. to operate under prescribed conditions without failure.

**Reliability of a series system,  $R_S$**  is the product of the reliability of all the components in the system, or  $R_S = R_1 * R_2 * R_3 * R_4 * \dots * R_n$ . The value of  $R_S$  is less than the reliability of a single component.

**Reliability of a parallel system,  $R_S$**  is one minus the product of all the parallel components not working, or  $R_S = 1 - (1 - R_1)(1 - R_2)(1 - R_3)(1 - R_4) \dots (1 - R_n)$ . The value of  $R_S$  is greater than the reliability of an individual component.

**Replacement** is when we take an element from a population, note its value, and then return this element back into the population.

**Representative sample** is one that contains the relevant characteristics of the population and which occur in the same proportion as in the population.

**Research hypothesis** is the same as the alternative hypothesis and is a value that has been obtained from a sampling experiment.

**Right-skewed data** is when the mean of a dataset is greater than the median value, and the curve of the distribution tails off to the right side of the  $x$ -axis.

**Right-tail hypothesis test** is used when we are asking the question, Is there evidence that a value is greater than?

**Risk** is the loss, often financial, that may be incurred when an activity or experiment is undertaken.

**Rolling index number** is the index value compared to a moving base value often used to show the change of data each period.

**Sample** is the collection of a portion of the population data elements.

**Sampling** is the analytical procedure with the objective to estimate population parameters.

**Sampling distribution of the means** is a distribution of all the means of samples withdrawn from a population.

**Sampling distribution of the proportion** is a probability distribution of all possible values of the sample proportion,  $\bar{p}$ .

**Sampling error** is the inaccuracy in a sampling experiment.

**Sample space** gives all the possible outcomes of an experiment.

**Sample standard deviation,  $s$**  is the square root of the sample variation,  $\sqrt{s^2}$ .

**Sample variance,  $s^2$**  is given by,

$$s^2 = \frac{\sum (x - \bar{x})^2}{(n - 1)}$$

where  $n$  is the amount of data,  $x$  is the particular data value, and  $\bar{x}$  is the mean value of the dataset.

**Sampling from an infinite population** means that even if the sample were not replaced, then the probability outcome for a subsequent sample would not significantly change.

**Sampling with replacement** is taking a sample from a population, and after analysis, the sample is returned to the population.

**Sampling without replacement** is taking a sample from a population, and after analysis not returning the sample to the population.

**Scatter diagram** is the presentation of time series data in the form of dots on  $x$ -axis and  $y$ -axis to illustrate the relationship between the  $x$  and  $y$  variables.

**Score** is a quantitative value for a subjective response often used in evaluating questionnaires.

**Seasonal forecasting** is when in a time series the value of the dependent variable is a function of time but also varies often in a sinusoidal fashion according to the season.

**Secondary data** is the published information collected by a third party.

**Series arrangement** is when in system, components are connected sequentially so that you have to pass through all the components in order that the system functions.

**Shape of the sampling distribution of the means** is about normal if random samples of at least size 30 are taken from a non-normal population; if samples of at least 15 are withdrawn from a symmetrical distribution; or samples of any size are taken from a **normal population**.

**Side-by-side bar chart** is where the data is shown as horizontal bars and within a given category there are sub-categories such as different periods.

**Side-by-side histogram** is a vertical bar chart showing the data according to a category and within a given category there are sub-categories such as different periods. A side-by-side histogram is also referred to as a **parallel histogram**.

**Significantly different** means that in comparing data there is an important difference between two values.

**Significantly greater** means that a value is considerably greater than a hypothesized value.

**Significantly less** means that a value is considerably smaller than a hypothesized value.

**Significance level** in hypothesis testing is how large, or important, is the difference before we say that a null hypothesis is invalid. It is denoted by  $\alpha$ , the area outside the distribution.

**Simple probability** is an alternative for **marginal** or **classical probability**.

**Simple random sampling** is where each item in the population has an equal chance of being selected.

**Skewed** means that data is not symmetrical.

**Stacked histogram** shows data according to categories and within each category there are sub-categories. It is developed from a **cross-classification** or **contingency table**.

**Standard deviation of a random variable** is the square root of the variance or,

$$\sigma = \sqrt{\sum (x - \mu_x)^2 P(x)}$$

Standard deviation of the binomial distribution is the square root of the variance, or  $\sigma = \sqrt{\sigma^2} = \sqrt{npq}$ .

**Standard error of the difference between two proportions** is,

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

**Standard deviation of the distribution of the difference between sample means** is,

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

**Standard deviation of the Poisson distribution** is the square root of the mean number of occurrences or,  $\sigma = \sqrt{(\lambda)}$ .

**Standard deviation of the sampling distribution**,  $\sigma_{\bar{x}}$ , is related to the population standard deviation,  $\sigma_x$ , and sample size,  $n$ , from the **central limit theory**, by the relationship,

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

**Standard error of the estimate** of the linear regression line is,

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$

**Standard error of the difference between two means** is,

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

**Standard error of the proportion**,  $\sigma_{\bar{p}}$  is,

$$\sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$$

**Standard error of the sample means**, or more simply the **standard error** is the error in a sampling experiment. It is the relationship,

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

**Standard error of the estimate** in forecasting is a measure of the variability of the actual data around the regression line.

**Standard normal distribution** is one which has a mean value of zero and a standard deviation of unity.

**Statistic** describes the characteristic of a sample, taken from a population, such as the weight, volume length, etc.

**Statistical dependence** is the condition when the outcome of one event impacts the outcome of another event.

**Statistical independence** is the condition when the outcome of one event has no bearing on the outcome of another event, such as in the tossing of a fair coin.

**Stems** are the principal data values in a **stem-and-leaf display**.

**Stem-and-leaf display** is a frequency distribution where the data has a stem of principal values, and a leaf of minor values. In this display, all data values are evident.

**Stratified sampling** is when the population is divided into homogeneous groups or strata and random sampling is made on the strata of interest.

**Student-t distribution** is used for small sample sizes when the population standard deviation is unknown.

**Subjective probability** is based on the belief, emotion or “gut” feeling of the person making the judgment.

**Symmetrical** in a box and whisker plot is when the distances from  $Q_0$  to the median  $Q_2$ , and the distance from  $Q_2$  to  $Q_4$ , are the same; the distance from  $Q_0$ , to  $Q_1$  equals the distance from  $Q_3$  to  $Q_4$  and the distance from  $Q_1$  to  $Q_2$  equals the distance from the  $Q_2$  to  $Q_3$ ; and the mean and the median value are equal.

**Symmetrical distribution** is when one half of the distribution is a mirror image of the other half.

**System** is the total of all components, pieces, or processes in an arrangement. Purchasing, transformation, and distribution are the processes of the supply chain system.

**Systematic sampling** is taking samples from a homogeneous population at a regular space, time or interval.

**Time series** is historical data, which illustrate the progression of variables over time.

**Time series deflation** is a way to determine the real value in the change of a commodity using the **consumer price index**.

**Transformation relationship** is the same as the **normal distribution transformation relationship**.

**Tri-modal is** when there are three values in a dataset that occur most frequently.

**Type I error** occurs if the null hypothesis is rejected when in fact the null hypothesis is true.



**Type II error** is accepting a null hypothesis when the null hypothesis is not true.

**Two-tail hypothesis test** is used when we are asking the question, “Is there evidence of a difference?”

**Unbiased estimate** is one that on an average will equal to the parameter that is being estimated.

**Univariate data** is composed of individual values that represent just one random variable,  $x$ .

**Unreliability** is when a system or component is unable to perform as specified.

**Unweighted aggregate index** is one that in the calculation each item in the index is given equal importance.

**Variable value** is one that changes according to certain conditions. The ending letters of the alphabet,  $u$ ,  $v$ ,  $w$ ,  $x$ ,  $y$ , and  $z$ , either upper or lower case, are typically used to denote variables.

**Variance of a distribution of a discrete random variable** is given by the expression,

$$\sigma^2 = \sum (x - \mu_x)^2 P(x)$$

**Variance of the binomial distribution** is the product of the number of trials  $n$ , the character-

istic probability,  $p$ , of *success*, and the characteristic probability,  $q$ , of *failure*, or  $\sigma^2 = npq$ .

**Venn diagram** is a representation of probability outcomes where the sample space gives all possible outcomes and a portion of the sample space represents an event.

**Vertical histogram** is a graphical presentation of vertical bars where the  $x$ -axis gives a defined class and the  $y$ -axis gives data according to the frequency of occurrence in a class.

**Weighted average** is the mean value taking into account the importance or weighting of each value in the overall total. The total weightings must add up to 1 or 100%.

**Weighted mean** is an alternative for the **weighted average**.

**Weighted price index** is when different weights or importance is given to the items used to calculate the index.

**What if** is the question asking, “What will be the outcome with different information?”

**Wholes numbers** are those with no decimal or fractional components.

## Symbols used in the equations

Symbol	Meaning
$\lambda$	Mean number of occurrences used in a Poisson distribution
$\mu$	Mean value of population
$n$	Sample size in units
$N$	Population size in units
$p$	Probability of success, fraction or percentage
$q$	Probability of failure = $(1 - p)$ , fraction or percentage
$Q$	Quartile value
$r$	Coefficient of correlation
$r^2$	Coefficient of determination
$s$	Standard deviation of sample
$\sigma$	Standard deviation of population
$\hat{\sigma}$	Estimate of the standard deviation of the population
$s_e$	Standard error of the regression line
$t$	Number of standard deviations in a Student distribution
$x$	Value of the random variable. The independent variable in the regression line
$\bar{x}$	Average value of $x$
$y$	Value of the dependent variable
$\bar{y}$	Average value of $y$
$\hat{y}$	Value of the predicted value of the dependent variable
$z$	Number of standard deviations in a normal distribution

Note: Subscripts or indices 0, 1, 2, 3, etc. indicate several data values in the same series.



*This page intentionally left blank*

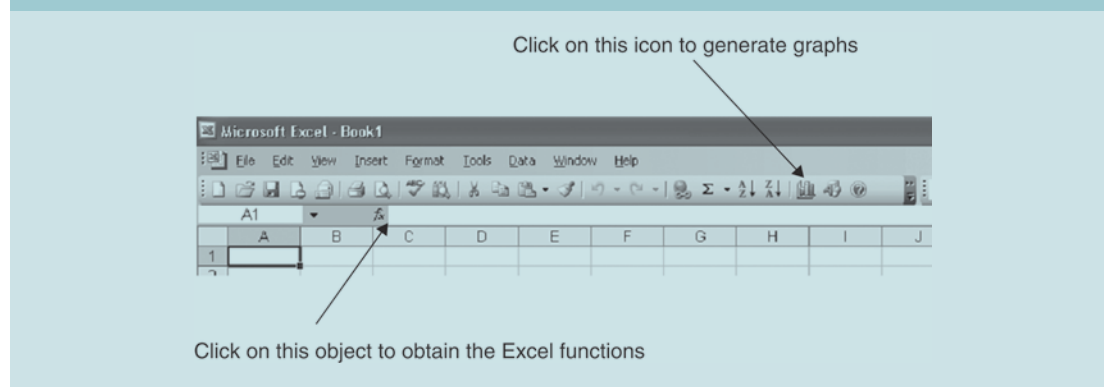
# Appendix II: Guide for using Microsoft Excel in this textbook

(Based on version 2003)

The most often used tools in this statistics textbook are the development of **graphs** and the built-in **functions** of the Microsoft Excel program. To use either of these you simply click on the graph icon, or the object function of the toolbar in the Excel spreadsheet as shown in Figure E-1. The following sections give more

information on their use. Note in these sections the words shown in *italics* correspond exactly to the headings used in the Excel screens but these may not always be the same terms as used in this textbook. For example, Excel refers to *chart type*, whereas in the text I call them graphs.

Figure E.1 Standard tool bar, Excel version 2003.



## Generating Excel Graphs

When you click on the graph icon as shown in Figure E-1 you will obtain the screen that is illustrated in Figure E-2. Here in the tab *Standard Types* you have a selection of the *Chart type* or graphs that you can produce. The key ones that are used in this text are the first five in the list – *Column* (histogram), *Bar*, *Line*, *Pie*, and *XY (Scatter)*. When you click on any of these options you will have a selection of the various formats that are available. For example, Figure E-2 illustrates the *Chart sub-type* for the *Column* options and Figure E-3 illustrates the *Chart sub-types* for the *XY (Scatter)* option.

Assume, for example, you wish to draw a line graph for the data given in Table E-1 that is contained in an Excel spreadsheet. You first select (highlight) this data and then choose the graph option *XY (Scatter)*. You then click on *Next* and this will illustrate the graph you have formed as shown in Figure E-4. This Step 2 of 4 of the chart wizard as shown at the top of the window. If you click on the tab, *Series* at the top of the screen, you can make modifications to the input data. If you then click on *Next* again you will have Step 3 of 4, which gives the various *Chart options* for presenting your graph. This window is shown in Figure E-5. Finally, when you again click on *Next* you will have *Chart Location* according to the screen shown in

Figure E.2 Graph types available in Excel.

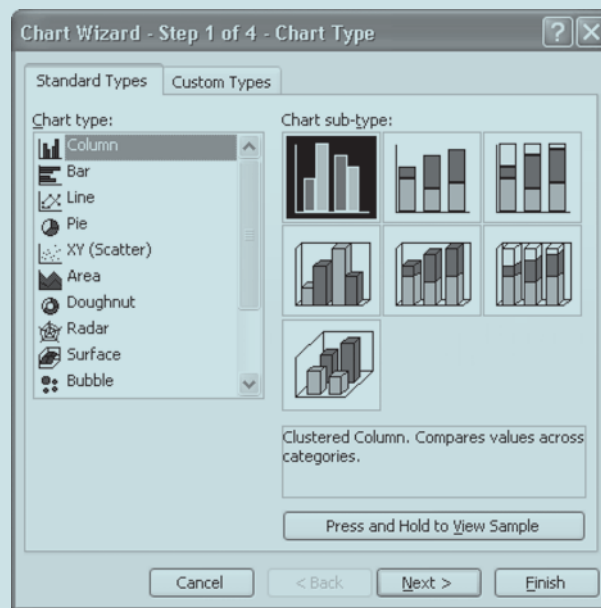


Figure E.3 XY graphs selected.

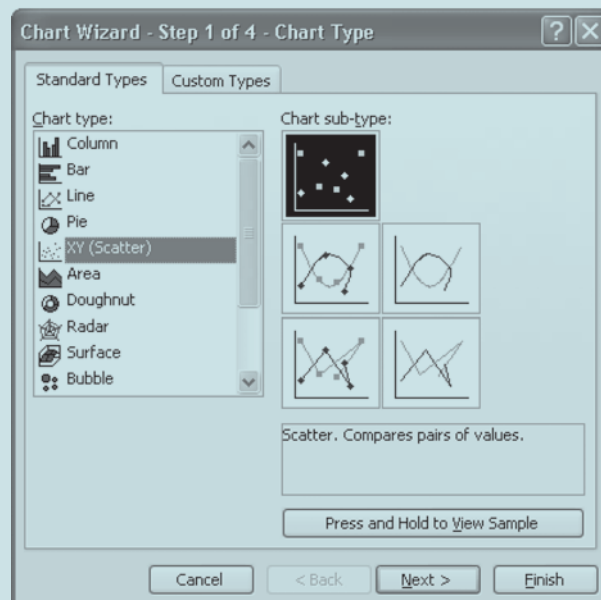


Table E-1  $x, y$  data.

$x$	1	2	3	4	5
$y$	5	9	14	12	21

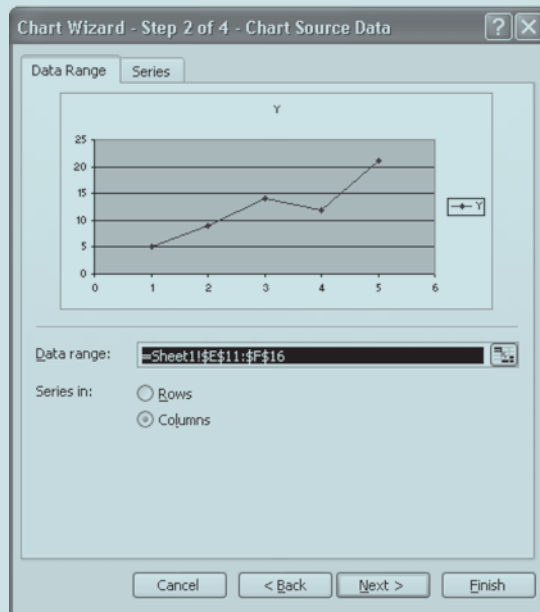
Figure E.4  $X, Y$  line graph.

Figure E.5 Options to present a graph.

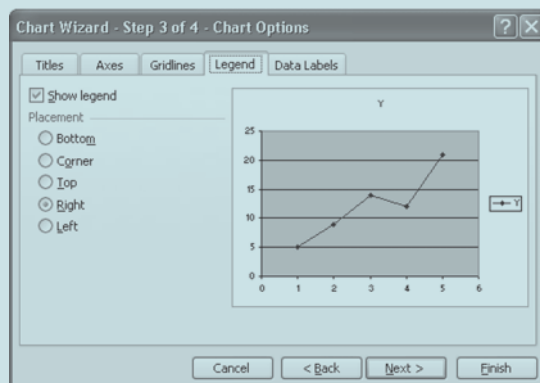


Figure E-6. This gives you a choice of making a graph *As new sheet*, that is as a new file for your graph or *As object in*, which is the graph in your spread sheet. For organizing my data I always prefer to create a new sheet for my graphs, but the choice is yours!

Regardless of what type of graph you decide to make, the procedure is the same as indicated in the previous paragraph. One word of caution is

the choice in the *Standard Types* between using *Line* and *XY (Scatter)*. For any line graph I always use *XY (Scatter)* rather than *Line* as with this presentation the  $x$  and  $y$  data are always correlated.

In Chapter 10, we discussed in detail linear regression or the development of a straight line that is the best fit for the data given. Figure E-7 shows the screen for developing this linear regression line.

Figure E.6 Location of your graph.

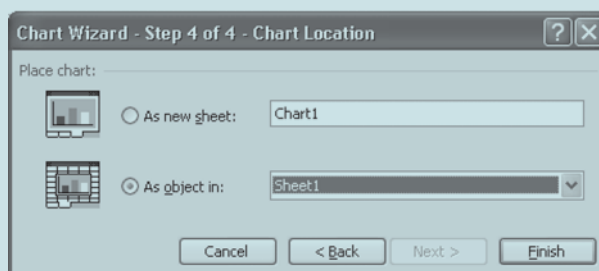
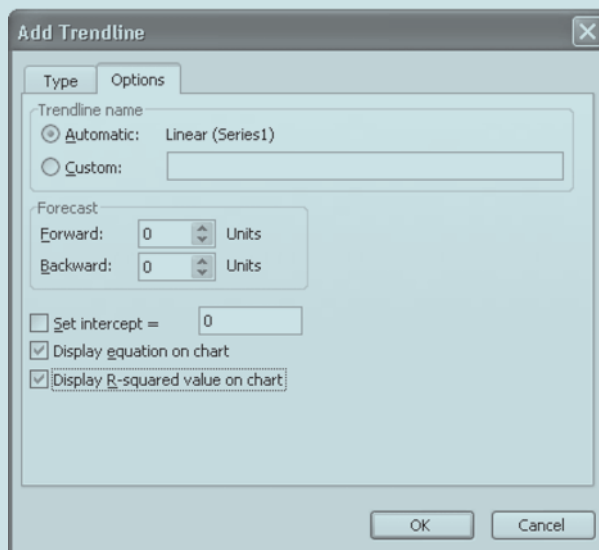


Figure E.7 Adding regression line.



## Using the Excel Functions

If you click on the  $f_x$  object in the tool bar as shown in Figure E-1, and select, *All*, in the command, *Or select a category*, you will have the screen as shown in Figure E-8. This gives a listing of all the functions that are available in Excel in alphabetical order. When you highlight a function it tells you its purpose. For example, here the function *ABS* is highlighted and it says

on the bottom of the screen, *Returns the absolute value of a number; a number without its sign*. If you are in doubt and you want further information about using a particular function you have, “Help on this function” at the bottom of the screen. Table E-2 gives those functions that are used in this textbook and their use. Each function indicated can be found in appropriate chapters of this textbook. (Note, for those living south of the Isle of Wight, you have the equivalent functions in French!)

Figure E.8 Selecting functions in Excel.

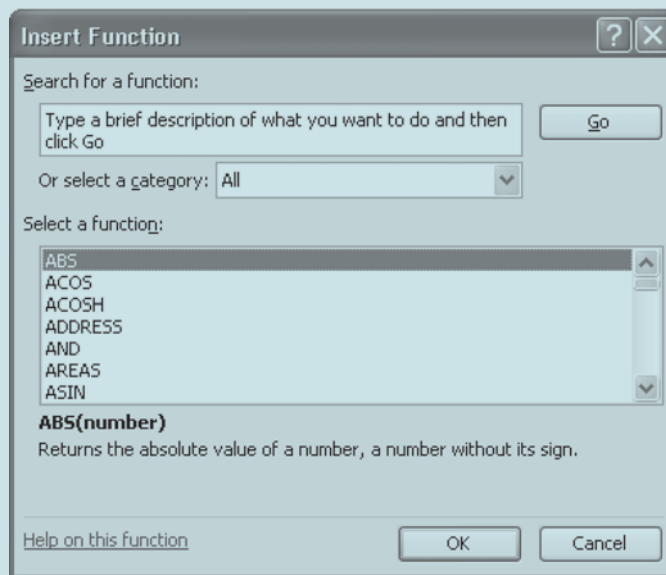


Table E-2 Excel functions used in this book.

English	French	For determining
ABS	ABS	Gives the absolute value of a number. That is the negative numbers are ignored
AVERAGE	MOYENNE	Mean value of a dataset
EXPONDIST	LOI.EXPONENTIELLE	Cumulative distribution exponential function, given the value of the ransom variable $x$ and the mean value $\lambda$ . Use a value of cumulative = 1
CEILING	ARRONDI.SUP	Rounds up a number to the nearest integer value

Table E-2 Excel functions used in this book. (Continued)

English	French	For determining
CHIDIST	LOI.KHIDEUX	Gives the area in the chi-distribution when you enter the chi-square value and the degrees of freedom
CHIINV	KHIDEUX.INVERSE	Gives the chi-square value when you enter the area in the chi-square distribution and the degrees of freedom
CHITEST	TEST.KHIDEUX	Gives the area in the chi-square distribution when you enter the observed and expected frequency values
COMBIN	COMBIN	Gives the number of combinations of arranging $x$ objects from a total sample of $n$ objects
CONFIDENCE	INTERVALLE.CONFIANCE	Returns the confidence interval for a population mean
CORREL	COEFFICIENT;CORRELATION	Determines the coefficient of correlation for a bivariate dataset
COUNT	NBVAL	The number of values in a dataset
CHIINV	KHIDEUX.INVERSE	Returns the inverse of the one-tailed probability of the chi-squared distribution
BINOMDIST	LOI.BINOMIALE	Binomial distribution given the random variable, $x$ , and characteristic probability, $p$ . If cumulative = 0, the individual value is determined. If cumulative = 1, the cumulative values are determined
IF	SI	Evaluates a condition and returns either true or false based on the stated condition
FACT	FACT	Returns the factorial value $n!$ of a number
FLOOR	ARRONDI.INF	Rounds down a number to the nearest integer value
FORECAST	PREVISION	Gives a future value of a dependent variable, $y$ , from known variables $x$ and $y$ , data assuming a linear relationship between the two
FREQUENCY	FREQUENCE	Determines how often values occur in a dataset
GEOMEAN	MOYENNE.GEOMETRIQUE	Gives the geometric mean growth rate from the annual growth rates data. The percentage of geometric mean is the geometric mean growth rate less than 1
GOAL SEEK	VALEUR CIBLE	Gives a value according to a given criteria. This function is in the <i>tools</i> menu
KURT	KURTOSIS	Gives the kurtosis value, or the peakness of flatness of a dataset
LINEST	DROITEREG	Gives the parameters of a regression line
MAX	MAX	Determines the highest value of a dataset
MEDIAN	MEDIANE	Middle value of a dataset
MIN	MIN	Determines the lowest value of a dataset
MODE	MODE	Determines the mode, or that value which occurs most frequently in a dataset
NORMDIST	LOI.NORMALE	Area under the normal distribution given the value of the random variable, $x$ , mean value, $\mu$ , standard deviation, $\sigma$ , and cumulative = 1. If you use cumulative = 0 this gives a point value for exactly $x$ occurring
NORMINV	LOI.NORMALE.INVERSE	Value of the random variable $x$ given probability, $p$ , mean value, $\mu$ , standard deviation, $\sigma$ , and cumulative = 1

Table E-2 Excel functions used in this book.

English	French	For determining
NORMSDIST	LOI.NORMALE.STANDARD	The probability, $p$ , given the number of standard deviations $z$
NORMSINV	LOI.NORMALE.STANDARD. INVERSE	Determines the number of standard deviations, $z$ , given the value of the probability, $p$
OFFSET	DECALER	Repeats a cell reference to another line or column according to the offset required
PEARSON	PEARSON	Determines the Pearson product moment correlation, or the coefficient of correlation, $r$
PERCENTILE	CENTILE	Gives the percentile value of a dataset. Select the data and enter the percentile, 0.01, 0.02, etc
PERMUT	PERMUTATION	Gives the number of permutations of organising $x$ objects from a total sample of $n$ objects
POISSON	LOI.POISSON	Poisson distribution given the random variable, $x$ , and the mean value, $\lambda$ . If cumulative = 0, the individual value is determined. If cumulative = 1, the cumulative values are determined
POWER	PUISSANCE	Returns the result of a number to a given power
RAND	ALEA	Generates a random number between 0 and 1
RANDBETWEEN	ALEA.ENTRE.BORNES	Generates a random number between the numbers you specify
ROUND	ARRONDI	Rounds to the nearest whole number
RSQ	COEFFICIENT.DETERMINATION	Determines the coefficient of determination, $r^2$ or gives the square of the Pearson product moment correlation coefficient
IF	SI	Logical statement to test a specified condition
SLOPE	PENTE	Determines the slope of a regression line
SQRT	RACINE	Gives the square root of a given value
STDEV	ECARTYPE	Determines the standard deviation of a dataset on the basis for a sample
STDEVP	ECARTYPEP	Determines the standard deviation of a dataset on the basis for a population
SUM	SOMME	Determines the total of a defined dataset
SUMPRODUCT	SOMMEPROD	Returns the sum of two columns of data
TDIST	LOI.STUDENT	Probability of a random variable, $x$ , given the degrees of freedom, $\nu$ and the number of tails. If the number of tails = 1, the area to the right is determined. If number of tails = 2, the area in both tails is determined
TINV	LOI.STUDENT.INVERSE	Determines the value of the Student- $t$ given the probability or area outside the curve, $p$ , and the degree of freedom, $\nu$
VALEUR CIBLE	GOAL SEEK	Gives a target value based on specified criteria
VAR	VAR	Determines the variance of a dataset on the basis it is a sample
VARP	VAR.P	Determines the variance of a dataset on the basis it is a population



## Simple Linear Regression

Simple linear regression functions can be solved using the regression function in Excel. A virgin block of cells at least two columns by five rows are selected. When the  $y$  and  $x$  data are entered into the function, the various statistical data are returned in a format according to Table E-3.

## Multiple Regression

As for simple linear regression, multiple regression functions can be solved with the Excel regression function. Here now a virgin block of cells is selected such that the number of columns is at least equal to the number of variables plus one and the number of rows is equal to five. When the  $y$  and  $x$  data are entered into the function, the various statistical data are returned in a format according to Table E-4.

Table E-3 Microsoft Excel and the linear regression function.

$b$	Slope due to variable $x$	$a$	intercept on $y$ -axis
$se_b$	Standard error for slope, $b$	$se_a$	standard error for intercept $a$
$r^2$	coefficient of determination	$s_e$	standard error of estimate
$F$	$F$ -ratio for analysis of variance	df	degree of freedom ( $n - 2$ )
$SS_{reg}$	sum of squares due to regression (explained variation)	$SS_{resid}$	sum of squares of residual (unexplained variation)

Table E-4 Microsoft Excel and the multiple regression function.

$b_k$ , slope due to variable $x_k$	$b_{k-1}$ , slope due to variable $x_{k-1}$	$b_2$ , slope due to variable $x_2$	$b_1$ , slope due to variable $x_1$	$a$ , intercept on $y$ -axis
$se_k$ , standard error for slope $b_k$	$se_{k-1}$ , standard error for slope $b_{k-1}$	$se_2$ , standard error for slope $b_2$	$se_1$ , standard error for slope $b_1$	$se_a$ , standard error for intercept $a$
$r^2$ , coefficient of determination	$s_e$ , standard error of estimate			
$F$ -ratio	df, degree of freedom			
$SS_{reg}$ , sum of squares due to regression (explained variation)	$SS_{resid}$ , sum of squares of residual (unexplained variation)			

# Appendix III: Mathematical relationships

## Subject matter

Your memory of basic mathematical relationships may be rusty. The objective of this appendix is to give a detailed revision of arithmetic relationships, rules, and conversions. The following concepts are covered:

Constants and variables • Equations • Integer and non-integer numbers • Arithmetic operating symbols and equation relationships • Sequence of arithmetic operations • Equivalence of algebraic expressions • Fractions • Decimals • The Imperial and United States measuring system • Temperature • Conversion between fractions and decimals • Percentages • Rules for arithmetic calculations for non-linear relationships • Sigma,  $\Sigma$  • Mean value • Addition of two variables • Difference of two variables • Constant multiplied by a variable • Constant summed  $n$  times • Summation of a random variable around the mean • Binary numbering system • Greek alphabet

Statistics involves numbers and the material in this textbook is based on many mathematical relationships, fundamental ideas, and conversion factors. The following summarizes the basics.

### Constants and variables

A constant is a value which does not change under any circumstances. The straight line distance from the centre of Trafalgar Square in London to the centre of the Eiffel Tower in Paris is constant. However, the driving time from these two points is a variable as it depends on road, traffic, and weather conditions. By convention, constants are represented algebraically by the beginning letters of the alphabet either in lower or upper case.

Lower case ***a, b, c, d, e, .....***

Upper case ***A, B, C, D, E, .....***

A variable is a number whose value can change according to various conditions. By convention variables are represented algebraically by the ending letters of the alphabet again either in lower or upper case.

Lower case ***u, v, w, x, y, z***

Upper case ***U, V, W, X, Y, Z***

The variables denoted by the letters  $x$  and  $y$  are the most commonly encountered. Where two-dimensional graphs occur,  $x$  is the abscissa or horizontal axis, and  $y$  is the ordinate or vertical axis. This is bivariate data. In three-dimensional graphs, the letter  $z$  is used to denote the third axis. In textbooks, articles, and other documents you will see constants and variables written in either upper case or lower case. There seems to be no recognized rule; however, I prefer to use the lower case.

Equations

An equation is a relationship where the values on the left of the equal sign are equal to the values on the right of the equal sign. Values in any part of an equation can be variables or constants. The following is a linear equation meaning that the power of the variables has the value of unity:

$y = a + bx$

This equation represents a straight line where the constant cutting the *y*-axis is equal to *a* and the slope of the curve is equal to *b*.

An equation might be non-linear meaning that the power of any one of the variables has a value other than unity as for example,

$y = a + bx^3 + cx^2 + d$

Integer and non-integer numbers

An integer is a whole number such as 1, 2, 5, 19, 25, etc. In statistics an integer is also known as a discrete number or discrete variable if the number can take any different values. Non-integer numbers are those that are not whole numbers such as the fractions ½, ¾, or 3½, 7¾, etc; or decimals such as 2.79, 0.56, and 0.75.

Arithmetic operating symbols and equation relationships

The following are arithmetic operating symbols and equation relationships:

- + Addition
- Subtraction
- ± Plus or minus
- = Equals
- ≠ Not equal to
- ÷ Divide
- / This means ratio but also divide. For example, ¾ means the ratio of 3 to 4 but also 3 divided by 4.
- > Greater than
- < Less than
- ≥ Greater or equal to

- ≤ Less than or equal to
- ≈ Approximately equal to

For multiplication we have several possibilities to illustrate the operation. When we multiply two algebraic terms *a* and *b* together this can be shown as:

$ab; a.b; a \times b; \text{ or } a * b$

With numbers, and before we had computers, the multiplication or product of two values was written using the symbol  $\times$  for multiplication:

$6 \times 4 = 24$

With Excel the symbol  $*$  is used as the multiplication sign and so the above relationship is written as:

$6 * 4 = 24$

It is for this reason that in this textbook, the symbol  $*$  is used for the multiplication sign rather than the historical  $\times$  symbol.

Sequence of arithmetic operations

When we have expressions related by operating symbols the rule for calculation is to start first to calculate the terms in the **B**rackets, then **D**ivision and/or **M**ultiplication, and finally **A**ddition and/or **S**ubtraction (BDMAS) as shown in Table M-1.

If there are no brackets in the expression and only addition and subtraction operating symbols then you work from left to write. Table M-2 gives some illustrations.

Table M-1 Sequence of arithmetic operations.

Symbol	Term	Evaluation sequence
B	Brackets	1st
D	Division	2nd
M	Multiplication	2nd
A	Addition	Last
S	Subtraction	Last

## Equivalence of algebraic expressions

Algebraic or numerical expressions can be written in various forms as Table M-3 illustrates.

## Fractions

Fractions are units of measure expressed as one whole number divided by another whole number. The common fraction has the numerator on the top and the denominator on the bottom:

$$\text{Common fraction} = \frac{\text{Numerator}}{\text{Denominator}}$$

The common fraction is when the numerator is less than the denominator which means that the number is less than one as for example,  $\frac{1}{7}$ ,  $\frac{3}{4}$  and  $\frac{5}{12}$ . The improper fraction is when the numerator

is greater than the denominator, which means that the number is greater than unity as for example  $\frac{30}{7}$ ,  $\frac{52}{9}$ , and  $\frac{19}{3}$ . In this case these improper fractions can be reduced to a whole number and proper fractions to give  $4\frac{2}{7}$ ,  $5\frac{7}{9}$  and  $6\frac{1}{3}$ . The rules for adding, subtracting multiplying, and dividing fractions are given in Table M-4.

## Decimals

A decimal number is a fraction, whose denominator is any power of 10 so that it can be written using a decimal point as for example:

$$\begin{aligned} 7/10 &= 0.70 & 9/100 &= 0.09 \\ 7,051/1,000 &= 7.051 \end{aligned}$$

The metric system, used in continental Europe, is based on the decimal system and changes in units of 10. Tables M-5, M-6, M-7, and M-8, give

*Table M-2* Calculation procedures for addition and subtraction.

Expression	Answer	Operation
$25 - 11 + 7$	21	Calculate from left to right
$9 * 6 - 4$	50	Multiplication before subtraction
$-22 * 4$	-88	A minus times a plus is a minus
$-12 * -6$	72	Minus times a minus equals a minus
$6 + 9 * 5 - 3$	48	Multiplication then addition and subtraction
$7(9)$	63	A bracket is equivalent to a multiplication operation
$9(5 + 7)$	108	Addition in the bracket then the multiplication
$(7 - 4)(12 - 3) - 6$	21	Expression in brackets, multiplication, then subtraction
$20 * 3 \div 10 + 11$	17	Multiplication and divisions first then addition

*Table M-3* Algebraic and numerical expressions.

Arithmetic rule	Example
$a + b = b + a$	$6 + 7 = 7 + 6 = 13$
$a + (b + c) = a + b + c = (a + b) + c$	$9 + (7 + 3) = 9 + 7 + 3 = (9 + 7) + 3 = 19$
$a - b = -b + a$	$15 - 21 = -21 + 15 = -6$
$a * b = b * a$	$6 * 7 = 7 * 6 = 42$
$a * (b + c) = a * b + a * c$	$3 * (8 + 4) = 3 * 8 + 3 * 4 = 36$

Table M-4 Rules for treating fractions.

$\frac{1}{a} + \frac{1}{b} = \frac{b+a}{ab}$	$\frac{1}{5} + \frac{1}{6} = \frac{6+5}{5*6} = \frac{11}{30}$
$\frac{a}{c} + \frac{b}{c} = \frac{a+b}{c}$	$\frac{4}{5} + \frac{16}{5} = \frac{4+16}{5} = \frac{20}{5} = 4$
$\frac{a}{c} - \frac{b}{c} = \frac{a-b}{c}$	$\frac{4}{7} - \frac{2}{7} = \frac{4-2}{7} = \frac{2}{7}$
$\frac{a}{c} * \frac{b}{d} = \frac{a*b}{c*d}$	$\frac{4}{5} * \frac{7}{6} = \frac{4*7}{5*6} = \frac{28}{30} = \frac{14}{15}$
$\frac{a}{c} \div \frac{b}{d} = \frac{a*d}{c*b}$	$\frac{4}{5} \div \frac{7}{6} = \frac{4*6}{5*7} = \frac{24}{35}$

Table M-5 Length or linear measure.

Micrometre ( $\mu\text{m}$ )	Millimetre (mm)	Centimetre (cm)	Decimetre (dm)	Metre (m)	Decametre (dam)	Hectometre (hm)	Kilometre (km)
$10^9$	1,000,000	100,000	10,000	1,000	100	10	1
$10^8$	100,000	10,000	1,000	100	10	1	0.1
10,000,000	10,000	1,000	100	10	1	0.1	0.01
1,000,000	1,000	100	10	1	0.1	0.01	0.001
100,000	100	10	1	0.100	0.01	0.001	0.0001
10,000	10	1	0.1	0.010	0.001	0.0001	0.00001
1,000	1	0.1	0.01	0.001	0.0001	0.00001	0.000001

Table M-6 Surface or area measure.

Square micrometre ( $\mu\text{m}^2$ )	Square millimetre ( $\text{mm}^2$ )	Square centimetre ( $\text{cm}^2$ )	Square decimetre ( $\text{dam}^2$ )	Square metre ( $\text{m}^2$ )	Are (a)	Hectare (ha)	Square kilometre ( $\text{km}^2$ )
$10^{18}$	$10^{12}$	$10^{10}$	$10^8$	1,000,000	10,000	100	1
$10^{16}$	$10^{10}$	$10^8$	1,000,000	10,000	100	1	0.01
$10^{14}$	$10^8$	1,000,000	10,000	100	1	0.01	0.0001
$10^{12}$	1,000,000	10,000	100	1	0.01	0.0001	0.000001
$10^{10}$	10,000	100	1	0.01	0.0001	0.000001	0.00000001
$10^8$	100	1	0.01	0.0001	0.000001	0.00000001	$10^{-10}$

Table M-7 Volume or capacity measure.

Microlitre ( $\mu\text{l}$ )	Millilitre (ml)	Centilitre (cl)	Decilitre (dl)	Litre (l)	Decalitre (dal)	Hectolitre (hl)	Kilolitre (kl)	Cubic centimetre ( $\text{cm}^3$ )	Cubic decimetre ( $\text{dm}^3$ )	Cubic metre ( $\text{m}^3$ )
$10^9$	$10^6$	100,000	10,000	1,000	100	10	1	$10^6$	1,000	1
$10^8$	100,000	10,000	1,000	100	10	1	0.1	100,000	100	0.1
$10^7$	10,000	1,000	100	10	1	0.1	0.01	10,000	10	0.01
$10^6$	1,000	100	10	1	0.1	0.01	0.001	1,000	1	0.001
100,000	100	10	1	0.1	0.01	0.001	0.0001	100	0.1	0.0001
10,000	10	1	0.1	0.01	0.001	0.0001	0.00001	10	0.01	0.00001
1,000	1	0.1	0.01	0.001	0.0001	0.00001	$10^{-6}$	1	0.001	$10^{-6}$
100	0.1	0.01	0.001	0.0001	0.00001	$10^{-6}$	$10^{-7}$	0.1	0.0001	$10^{-7}$
10	0.01	0.001	0.0001	0.00001	$10^{-6}$	$10^{-7}$	$10^{-8}$	0.01	0.00001	$10^{-8}$
1	0.001	0.0001	0.00001	$10^{-6}$	$10^{-7}$	$10^{-8}$	$10^{-9}$	0.001	$10^{-6}$	$10^{-9}$

**Table M-8** Mass or weight measure.

Microgram ( $\mu\text{g}$ )	Milligram (mg)	Centigram (cg)	Decigram (dg)	Gram (g)	Decagram (dag)	Hectogram (hg)	Kilogram (kg)	Metric ton (t)
$10^{12}$	$10^9$	$10^8$	$10^7$	1,000,000	100,000	10,000	1,000	1
$10^9$	1,000,000	100,000	10,000	1,000	100	10	1	0.001
$10^8$	100,000	10,000	1,000	100	10	1	0.1	0.0001
$10^7$	10,000	1,000	100	10	1	0.1	0.01	0.00001
1,000,000	1,000	100	10	1	0.1	0.01	0.001	0.000001
100,000	100	10	1	0.1	0.01	0.001	0.0001	$10^{-7}$
10,000	10	1	0.1	0.01	0.001	0.0001	0.00001	$10^{-8}$
1,000	1	0.1	0.01	0.001	0.0001	0.00001	0.000001	$10^{-9}$
100	0.1	0.01	0.001	0.0001	0.00001	0.000001	$10^{-7}$	$10^{-10}$
10	0.01	0.001	0.0001	0.00001	0.000001	$10^{-7}$	$10^{-8}$	$10^{-11}$
1	0.001	0.0001	0.00001	0.000001	$10^{-7}$	$10^{-8}$	$10^{-9}$	$10^{-12}$

**Table M-9** Conversions for length or linear measurement.

Inches (in)	Feet (ft)	Yards (yd)	Miles (mi)	Millimetres (mm)	Centimetres (cm)	Metres (m)	Kilometres (km)
1	0.0833	0.0278	$1.5783 \times 10^5$	0.254	2.5400	0.0254	0.0000254
12	1	0.3333	0.0002	3.048	30.4800	0.3048	0.0003048
36	3	1	0.0006	9.144	91.4400	0.9144	0.0009144
63,360	5,280	1,760	1	16,093.44	160,934.40	1,609.344	1.6093
0.3937	0.0328	0.0109	$6.2137 \times 10^6$	0.1	1	0.01	0.00001
39.3701	3.2808	1.0936	0.0006	10	100	1	0.001
39,370.08	3,280.84	1,093.6133	0.6214	10,000	100,000	1,000	1

the relationships for length, surface, volume, and weight.

## The Imperial, US, and metric measuring system

The Imperial measuring system is used in the United States and partly in England though there are efforts to change to the metric system. The Imperial numbering system is quirky with no apparent logic as compared to the metric system. Tables M-9, M-10, M-11, M-12, and M-13

give approximate conversion tables for key measurements. The United States system is not always the same as the Imperial measuring system.

## Temperature

In Europe, usually the Celsius system is used for recording temperature. Here the freezing point of water is measured at  $0^\circ\text{C}$  and the boiling point is  $100^\circ\text{C}$ . In the United States, and sometimes in the United Kingdom, the Fahrenheit system is used where the freezing point of water is given at  $32^\circ\text{F}$  and the boiling point as  $212^\circ\text{F}$ .

Table M-10 Conversions for surface or area measure.

Square inch (in <sup>2</sup> )	Square feet (ft <sup>2</sup> )	Square yard (yd <sup>2</sup> )	Square mile (mi <sup>2</sup> )	Acre (a)	Square centimetre (cm <sup>2</sup> )	Square metre (m <sup>2</sup> )	Area	Hectare (ha)	Square kilometre (km <sup>2</sup> )
1	0.0069	0.0008	$0.3 \times 10^{-9}$	$0.2 \times 10^{-6}$	6.4516	0.0006	$0.7 \times 10^{-5}$	$0.7 \times 10^{-7}$	$0.7 \times 10^{-9}$
144	1	0.1111	$0.4 \times 10^{-7}$	$0.2 \times 10^{-4}$	929.03	0.0929	0.0009	$0.9 \times 10^{-5}$	$0.9 \times 10^{-7}$
1,296	9	1	$0.3 \times 10^{-6}$	0.0002	$8.4 \times 10^3$	0.8361	0.0084	$0.8 \times 10^{-4}$	$0.8 \times 10^{-6}$
$4.01 \times 10^9$	$0.28 \times 10^8$	$3.1 \times 10^6$	1	640	$0.3 \times 10^{11}$	$0.26 \times 10^7$	$0.26 \times 10^5$	259.00	2.59
$0.63 \times 10^7$	43,560	4,840	0.0016	1	$0.4 \times 10^8$	$4.05 \times 10^3$	40.47	0.4047	0.004

Table M-11 Conversions for capacity or volume measure.

USA gallon	USA quart	USA pint	Imperial gallon	Imperial quart	Imperial pint	Cubic inches (in <sup>3</sup> )	Litres (l)
1.0000	2.0000	4.0000	0.8326	1.6652	3.3304	231.0000	3.7850
0.5000	1.0000	2.0000	0.4163	0.8326	1.6652	115.5000	1.8925
0.2500	0.5000	1.0000	0.1041	0.2082	0.4163	28.8750	0.4731
1.2011	2.4021	4.8042	1.0000	2.0000	4.0000	277.4200	4.5460
0.6005	1.2011	2.4021	0.5000	1.0000	2.0000	138.7100	2.2730
0.3003	0.6005	1.2011	0.2500	0.5000	1.0000	69.3550	1.1365

Table M-12 Conversions for mass or weight measure.

Ounce (oz)	Pound (lb)	Short ton	Long ton	Grams (g)	Kilograms (kg)	Metric ton
1	0.0625	0.00003125	0.000027902	28.350	0.02835	0.00002835
16	1	0.0005	0.0004	453.60	0.4536	0.0005
32,000	2,000	1	0.8929	907,200	907.2	0.9072
35,840	2,240	1.1200	1	1,016,064	1,016.064	1.0161

Table M-13 Conversion for pressure.

psi	kg/cm <sup>2</sup>	bar
1	0.07031	0.07031
14.2231	1	1

The following is the relationship between the two scales:

$$C = \frac{5}{9}(F - 32) \quad F = \frac{9}{5}C + 32$$

$$\begin{aligned} \text{When } F = 212^\circ\text{F then } C &= \frac{5}{9}(212 - 32) \\ &= \frac{5}{9} \times 180 = 100^\circ\text{C} \end{aligned}$$



Table M-14 Conversion of fraction and decimals.

Fraction	1/2	1/3	3/4	1/8	7/8	4( <sup>7</sup> / <sub>8</sub> )	6( <sup>2</sup> / <sub>3</sub> )	2( <sup>3</sup> / <sub>16</sub> )
Decimal	0.50	0.33	0.75	0.125	0.875	4.875	6.6667	2.1875

$$\begin{aligned}\text{When } C = 100^{\circ}\text{C then } F &= \frac{9}{5}C + 32 \\ &= \frac{9}{5} * 100 + 32 \\ &= 212^{\circ}\text{F}\end{aligned}$$

$$\begin{aligned}\text{When } C = 35^{\circ}\text{C then } F &= \frac{9}{5}C + 32 \\ &= \frac{9}{5} * 35 + 32 = 95^{\circ}\text{F}\end{aligned}$$

$$\begin{aligned}\text{When } F = 104^{\circ}\text{F then } C &= \frac{5}{9}(104 - 32) \\ &= \frac{5}{9} * 72 = 40^{\circ}\text{C}.\end{aligned}$$

## Conversion between fractions and decimals

To convert from a fraction to a decimal representation you divide the numerator or upper value in the fraction, by the denominator or the lower value in the fraction. To convert from a two-decimal presentation to a fraction, you divide the numbers after the decimal point by 100 and simplify until both the numerator and denominator are the smallest possible integer values. That is, you find the lowest common denominator. To convert from a three-decimal presentation to a fraction you divide to numbers after the decimal point by 1,000 etc. Table M-14 gives some examples.

## Percentages

The term percent means per 100 so that 50 percent, usually written, 50%, is 50 per 100. In order to change from

$$\begin{aligned}\text{A fraction to a percentage multiply by 100} \\ \frac{3}{4} = 75\% \quad (\frac{3}{4} * 100 = 75\%)\end{aligned}$$

$$\begin{aligned}\text{A decimal to a percentage multiply by 100} \\ 0.6712 = 0.6712 * 100 = 67.12\%\end{aligned}$$

$$\begin{aligned}\text{A percentage to a fraction divide by 100} \\ 25\% = 25/100 = \frac{1}{4}\end{aligned}$$

$$\begin{aligned}\text{A percentage to a decimal move the decimal} \\ \text{place two places to the left} \\ 25.86\% = 0.2586\end{aligned}$$

## Rules for arithmetic calculations for non-linear relationships

Table M-15 gives algebraic operations when the power of the variable, or the constant, is non-linear.

## Sigma, $\Sigma$

Very often in statistics we need to determine the sum of a set of data and to indicate this we use the Greek letter sigma,  $\Sigma$ . If we have a set of  $n$  values of data of a variable  $x$  then the sum of these is written as:

$$\Sigma x = x_1 + x_2 + x_3 + x_4 + \dots + x_n \quad \text{M(i)}$$

where  $x_1, x_2, x_3$ , etc. are the individual values in the dataset. If we have a dataset consisting of the values 7, 3, 2, 11, 21, and 9 then the sum of these values is:

$$\Sigma x = 7 + 3 + 2 + 11 + 21 + 9 = 53$$

## Mean value

The mean or average of a dataset is equal to the sum of the individual values,  $\Sigma x$  divided by the number of observations,  $n$ :

$$\bar{x} = \frac{\Sigma x}{n} \quad \text{M(ii)}$$

Table M-15 Algebraic operations involving powers.

Arithmetic rule	Example
$x^a x^b = x^{(a+b)}$	$5^3 * 5^2 = 5^{(3+2)} = 5^5 = 3,125$
$(x^a)^b = x^{ab}$	$(5^3)^2 = 5^6 = 15,625$
$1/x^a = x^{-a}$	$1/2^2 = 2^{-2} = 0.25$
$x^a/x^b = x^{(a-b)}$	$5^3/5^2 = 5^{(3-2)} = 5^1 = 5$
$x^a/x^a = x^{(a-a)} = x^0 = 1$	$6^2/6^2 = 6^{(2-2)} = 6^0 = 1$
$\sqrt{x * y} = \sqrt{x} * \sqrt{y}$	$\sqrt{16 * 25} = \sqrt{16} * \sqrt{25} = 4 * 5 = 20$
$\sqrt{\frac{x}{y}} = \frac{\sqrt{x}}{\sqrt{y}}$	$\sqrt{\frac{64}{144}} = \frac{\sqrt{64}}{\sqrt{144}} = \frac{8}{12} = \frac{2}{3} = 0.6667$

If  $x$  has values of 1, -5, 9, -2, 6, then  $n = 5$  and,

$$\bar{x} = \frac{(1 - 5 + 9 - 2 + 6)}{5} = \frac{9}{5} = 1.80$$

### Addition of two variables

The total of the addition of two variables is equal to the total of the individual sum of each variable:

$$\sum(x + y) = \sum x + \sum y \quad \text{M(iii)}$$

Assume the five values of the dataset  $(x, y)$  are, (1, 6); (-5, -2); (9, -8); (-2, 5); (6, 4):

$$\begin{aligned} \sum(x + y) &= (1 + 6) + (-5 - 2) + (9 - 8) \\ &\quad + (-2 + 5) + (6 + 4) \end{aligned}$$

$$\sum(x + y) = 7 - 7 + 1 + 3 + 10 = 14$$

$$\begin{aligned} \sum x + \sum y &= (1 - 5 + 9 - 2 + 6) \\ &\quad + (6 - 2 - 8 + 5 + 4) \end{aligned}$$

$$\sum x + \sum y = 9 + 5 = 14$$

### Difference of two variables

The sum of the difference of two variables is equal to the sum of the individual differences of each variable:

$$\sum(x - y) = \sum x - \sum y \quad \text{M(iv)}$$

If we have the following five values,

(1, 6), (-5, -2), (9, -8), (-2, 5), (6, 4)

$$\begin{aligned} \sum(x - y) &= (1 - 6) + (-5 + 2) + (9 + 8) \\ &\quad + (-2 - 5) + (6 - 4) \end{aligned}$$

$$\sum(x - y) = -5 - 3 + 17 - 7 + 2 = 4$$

$$\begin{aligned} \sum x - \sum y &= (1 - 5 + 9 - 2 + 6) \\ &\quad - (6 - 2 - 8 + 5 + 4) \end{aligned}$$

$$\sum x - \sum y = 9 - 5 = 4$$

### Constant multiplied by a variable

The sum of a constant times a variable equals to the constant times the sum of the variables:

$$\sum(kx) = k \sum x \quad \text{M(v)}$$

If  $k = 5$ , and  $x$  has values of 1, -5, 9, -2, 6, then,

$$\sum (kx) = 5 \times 1 - 5 \times 5 + 5 \times 9 - 5 \times 2 + 5 \times 6 = 45$$

$$k \sum x = 5(1 - 5 + 9 - 2 + 6) = 45$$

## Constant summed $n$ times

A constant summed  $n$  times is equal to the  $n$  times the constant:

$$\sum k = n * k \quad \text{M(vi)}$$

If  $n = 6$ , and  $k = 5$ ,

$$\sum k = 5 + 5 + 5 + 5 + 5 + 5 = 30 = 5 \times 6$$

## Summation of a random variable around the mean

Summation rules can be used to demonstrate that the summation of a random variable around a mean is equal to zero. Or,

$$\sum (x - \bar{x}) = 0 \quad \text{M(vii)}$$

From equation M(iv) equation M(vii) becomes,

$$\sum (x - \bar{x}) = \sum x - \sum \bar{x} = 0 \quad \text{M(viii)}$$

For any fixed set of data,  $\bar{x}$ , is a constant and thus from equation M(vi),

$$\sum \bar{x} = n\bar{x} \quad \text{M(ix)}$$

Thus from equation M(viii) we have,

$$\sum (x - \bar{x}) = \sum x - \sum \bar{x} = \sum x - n\bar{x} = 0 \quad \text{M(x)}$$

From equation M(ii),

$$\bar{x} = \frac{\sum x}{n}$$

$$\text{Or, } n\bar{x} = \sum x \quad \text{M(xi)}$$

Thus substituting in equation M(x) we have,

$$\begin{aligned} \sum (x - \bar{x}) &= \sum x - \sum \bar{x} = \sum x - n\bar{x} \\ &= \sum x - \sum x = 0 \end{aligned} \quad \text{M(xii)}$$

**Table M-16** Concept of binary numbering system.

0001	Value of 1 in 1st position is equal to	1
0010	Value of 1 in 2nd position is equal to	2
0100	Value of 1 in 3rd position is equal to	4
1000	Value of 1 in 4th position is equal to	8

That is the sum of the random variables about the mean is equal to zero.

## Binary numbering system

In the textbook, we introduced the binomial distribution. Related to this is the binary numbering system or binary code which is a system of arithmetic based on two digits, zero and one. The on/off system of most electrical appliances is based on the binary system where 0 = off and 1 = on.

The arithmetic of computers is based on the binary code. A digit, either the 0 or the 1 in the binary code is called the bit, or binary digit. When a binary digit is moved one space to the left, and a zero is placed after it, the resulting number is **twice** the original number. In the binary code, the actual value of 1 depends on the position of the 1 in a binary number, reading from **right to left**. A digit doubles its value each time it moves one place further to the left as shown in Table M-16.

Table M-17 gives the equivalent between binary numbers, and decimal numbers from 1 to 10. Note, we are reading from the right so that the 1st position is at the extreme right, the 2nd position to the immediate left of the 1st position, etc.

## Greek alphabet

In statistics, letters of the Greek alphabet are sometimes used as abbreviations to denote various terms. Table M-18 gives the Greek letters, both in upper and lower case and their English equivalent, and some of the areas where they are often used. To write these Greek letters in Microsoft Word you write the English letter and then change the font to *symbol*.

Table M-17 Relationship between decimal and binary numbers.

Decimal	Binary	Explanation
0	0000	
1	0001	Value of 1 in the 1st position = 1
2	0010	Value of 1 in the 2nd position = 2
3	0011	Value of 1 in the 1st position + value of 1 in the 2nd position = 3
4	0100	Value of 1 in the 3rd position = 4
5	0101	Value of 1 in the 3rd position + value of 1 in the 1st position = 5
6	0110	Value of 1 in 3rd position + value of 1 in 2nd position = 6
7	0111	Value of 1 in the 3rd position + value of 1 in 2nd position + value of 1 in 1st position = 7
8	1000	Value of 1 in the 4th position = 8
9	1001	Value of 1 in 4th position + value of 1 in 1st position = 9
10	1010	Value of 1 in 4th position + value of 1 in 2nd position = 10

Table M-18 Greek alphabet and English equivalent.

Upper case	Lower case	Name	English equivalent	Use
A	$\alpha$	Alpha	a	Hypothesis testing, exponential smoothing
B	$\beta$	Beta	b	Hypothesis testing
$\Gamma$	$\gamma$	Gamma	g	
$\Delta$	$\delta$	Delta	d	Calculus as the derivative
E	$\varepsilon$	Epsilon	e	
Z	$\zeta$	Zeta	z	
H	$\eta$	Eta	h	
$\Theta$	$\theta$	Theta	th (q)	Angle of a triangle
I	$\iota$	Iota	i	
K	$\kappa$	Kappa	k	
$\Lambda$	$\lambda$	Lambda	l	Poisson distribution, queuing theory
M	$\mu$	Mu	m	Mean value
N	$\nu$	Nu	n	
$\Xi$	$\xi$	Xi	x	
O	$\omicron$	Omicron	o	
$\Pi$	$\pi$	Pi	p	Circle constant
P	$\rho$	Rho	r, rh	
$\Sigma$	$\sigma$	Sigma	s	Total (upper case), standard deviation (lower case)
T	$\tau$	Tau	t	
Y	$\upsilon$	Upsilon	u	
$\Phi$	$\phi, \varphi$	Phi	ph (f, j)	
X	$\chi$	Chi	ch (c)	Chi-square test in hypothesis testing
$\Psi$	$\psi$	Psi	ps (y)	
$\Omega$	$\omega$	Omega	o (w)	Electricity

*This page intentionally left blank*

# Appendix IV: Answers to end of chapter exercises

## Chapter 1: Presenting and Organizing Data

### 1. Solutions – buyout

Question	Answer
6	About 7 (7.45)

### 2. Solutions – closure

Question	Answer
2	\$4.548 million per office
4	About 21
5	\$10.3 million per office remaining

### 3. Solutions – swimming pool

Question	Answer
3	Polygon is symmetrical
5	20%
7	<p>(a) From ogive, probability of more than 900 using the pool is 14.17% – this is greater than 10%. Probability of fewer than 600 people using the pool is 5.00% – this is less than 10%. Thus, both the conditions are satisfied</p> <p>(b) In Part (a) we have indicated that although the criteria was one or the other, we have shown that both conditions apply, or <b>and</b>. Further at the 10% benchmark there will be an estimated 925 people – this is greater than 900. Also, at the 10% level there will be an estimated to be 640 people – this number is greater than the 600 criterion</p> <p>(c) Probability of at least 600 is 95.00% and the probability of at least 900 is 14.17%, or a difference of 80.83%. Alternatively, probability of less than 900 is 85.83% and the probability of less than 600 is 5.00%, or a difference of 85.83%</p>

## 4. Solutions – Rhine river

Question	Answer
2	Logical minimum value is 16 m and logical maximum is 38 m
3	22%
4	31%

## 5. Solutions – purchasing expenditures

Question	Answer
1	Minimum value is €63,680 so use €60,000. Maximum value is €332,923 so use €334,000. In this way there is an even number of limits
3	15.50%
4	31
7	85%
8	€151,000

## 6. Solutions – exchange rates

Question	Answer
2	For all countries, except Canada where there is no change, the US dollar has become stronger. (It takes more of the local currency to buy dollars in 2005 than it did in 2004)

## 7. Solutions – European sales

Question	Answer
4	Best three performing countries are England, Denmark, and Spain with over two-thirds of the profit (67.04%)
5	Worst performing countries are Ireland, Norway, and Czech Republic with less than 10% of the profits (7.47%)

## 8. Solutions – nuclear power

Question	Answer
4	United States (104); France (59); Japan (56)
5	Western Europe (30.52% of world); France (41.84% of Western Europe)

## 9. Solutions – textbook sales

Question	Answer
1	Range of data from maximum to minimum is large so visual presentation of smaller values is not clear
2	Europe (56.44%); Africa (1.27%)
4	England (51.66%); Serbia, Latvia, and Lithuania (each with 0.03%)
6	No sales recorded for the United States. For some reason, the textbook cannot be sold in this country

## 10. Solutions – textile wages

Question	Answer
2	Times more than China, France (40.45); Italy (38.02); United States (32.20); Slovakia (6.67); Turkey (6.22); Bulgaria (2.33); Egypt (1.80); China (mainland) (1.00)
4	Explains why the textile firms (and other industries) are losing out to China

## 11. Solutions – immigration to Britain

Question	Answer
5	Largest proportion of these new immigrants is from Poland (56.74%). They are young with over 80% no more than 34-years old. Over a quarter are in administration, business, and management (27.54%)

## 12. Solutions – pill popping

Question	Answer
3	France 21% (20.69%)
4	France consumes more than twice the amount of pills that is consumed in the United Kingdom

## 13. Solutions – electoral college

Question	Answer
3	Bush 53.16%; Kerry 46.84%
4	California 10.22%
5	39.22%



## 14. Solutions – chemical delivery

Question	Answer
2	Drums damaged (32.19%)
3	Drums damaged (32.19%); documentation wrong (21.46%); delay – bad weather (15.02%); pallets poorly stacked (10.73%)

## 15. Solutions – fruit distribution

Question	Answer
2	Fruit squashed (27.31%). Not necessarily as bacteria on the fruit (3.32%), even though less frequent, is a more serious problem
3	Fruit squashed (27.31%); boxes badly loaded (22.88%); route directions poor (11.07%); fruit not clean (9.23%); client documentation incorrect (8.49%)

## Chapter 2: Characterizing and Defining Data

### 1. Solutions – billing rate

Question	Answer
1	\$50.78/hour
3	\$5,585,763
4	\$46.75/hour

### 2. Solutions – delivery

Question	Answer
1	\$8.28
2	\$3,310,360

### 3. Solutions – investment

Question	Answer
1	6.55%
2	6.04%
3	\$1,885.46
4	\$1,798.23
5	Option 1
6	15.02%

### 4. Solutions – production

Question	Answer
1	6.40%
2	21,760 units

## 5. Solutions – Euro prices

Question	Answer						
1		Milk (1l)	Renault Mégane	Big Mac	Stamp for postcard	Compact disc	Can of coke
	Maximum	1.34	21,700.00	3.10	0.60	22.71	1.18
	Minimum	0.52	12,450.00	2.11	0.38	14.98	0.33
	Range	0.82	9,250.00	0.99	0.22	7.73	0.85
	Average	0.83	16,406.58	2.63	0.50	19.20	0.54
	Midrange	0.93	17,075.00	2.61	0.49	18.85	0.76
	Median	0.81	16,287.50	2.57	0.51	18.97	0.46
	Standard deviation sample	0.23	2,775.18	0.30	0.07	2.78	0.24
	$s/\mu$	28.00%	16.92%	11.56%	13.24%	14.50%	44.82%
2	Finland has some of the highest prices. A Big Mac has the lowest deviation relative to the mean value. Spain's prices are all below the mean and the median value, which is an indicator of the lower cost of living						

## 6. Solutions – students

Question	Answer
1	3.79%
2	4,409

## 7. Solutions – construction

Question	Answer
1	4.49%
2	\$143.45/unit

## 8. Solutions – net worth

Question	Answer
1	8.05%

## 9. Solutions – trains

Question	Answer
1	24.24
2	25.00
3	25. This value occurs 5 times
4	42
5	24
6	147.44
7	12.14
8	50.09%
9	Data is pretty evenly distributed as the mean, median, and mode are quite close

## 10. Solutions – summer Olympics

Question	Answer
1	United States (103); Russia (92); China (63); Australia (49); Germany (48); Japan (37); France (33); Italy (32); Britain (30); South Korea (30)
2	United States (35); China (32); Russia (27); Australia (17); Japan (16); Germany (14); France (11); Italy (10); Britain (9); Cuba (9)
3	United States (35.33); Russia (28.83); China (24.00); Australia (16.50); Germany (15.33); Japan (13.00); France (10.67); Italy (10.50); South Korea (10.00); Britain (9.50)
4	4.01 (gold), 4.01 (silver), 4.36 (bronze)
5	United States (11.63%); China (10.63%); Russia (8.97%)

## 11. Solutions – printing

Question	Answer
1	\$14.23
2	\$15.77

## 12. Solutions – Big Mac

Question	Answer
1	(a) 5.05; (b) 1.38; (c) 2.51; (d) 2.40; (e) 3.67; (f) 3.22; (g) 1.48, 2 modal values; (h) 0.9060; (i) 36.15%
3	1.38, 1.88, 2.40, 2.80, 5.05

(Continued)

## 12. Solutions – (Continued)

Question	Answer
4	0.92
5	The price of the Big Mac in Indonesia is within the 1st quartile indicating a country with a low cost of living; Singapore is within the 2nd quartile and the inter-quartile range – indicating a relative low cost of living. The price for Hungary is within the 3rd quartile and the inter-quartile range and has a relative high cost of living. The price for Denmark is within the 4th quartile and indicates a high cost of living country

## 13. Solutions – purchasing expenditures – Part II

Question	Answer
1	(a) 332,923; (b) 63,680; (c) 269,243; (d) 198,302; (e) 184,767.32; (f) 180,532.00; (g) 194,157; 3 values; (h) 3,105,927,434; (i) 55,730.85; (j) 30,16%
2	63,680.00; 143, 956.75; 180,532.00; 223,039.25; 332,923.00
4	Mean is greater than the median and so the distribution is right-skewed
5	0% = 63,680.00; 81,208.01; 86,209.92; 94,815.19; 99,787.76; 102,413.35; 106,155.00; 108,128.99; 110,167.52; 113,757.08; 115,433.60; 118,273.65; 120,350.32; 123,855.30; 125,097.84; 126,713.95; 128,376.32; 129,940.66; 132,109.36; 133,912.13; 136,249.40; 137,860.00; 139,715.18; 140,972.60; 141,460.64 25% = 143,956.75; 146,154.30; 147,325.86; 148,290.80; 150,005.75; 151,926.70; 153,370.34; 154,557.56; 156,101.46; 156,888.64; 157,823.10; 159,627.44; 161,201.00; 161,600.78; 163,176.04; 165,299.00; 166,890.05; 168,943.82; 171,182.11; 172,934.44; 173,871.50; 175,558.64; 176,472.12; 178,446.68; 179,693.71 50% = 180,532.00; 181,749.50; 182,686.12; 184,333.02; 185,377.00; 187,124.00; 187,173.00; 188,575.71; 189,733.00; 191,227.59; 192,309.40; 194,157.00; 194,696.60; 196,376.94; 198,002.00; 199,767.25; 203,218.60; 204,696.63; 207,498.44; 211,080.78; 214,117.00; 217,105.29; 217,976.56; 219,898.08; 221,607.66 75% = 223,039.25; 224,387.60; 225,153.89; 228,936.48; 231,979.44; 235,516.80; 238,665.04; 241,132.46; 242,755.52; 243,181.64; 246,531.05; 248,332.06; 249,859.00; 253,227.20; 257,571.00; 261,153.00; 263,850.28; 270,766.00; 274,782.53; 276,333.92; 288,613.25; 293,445.00; 295,785.15; 298,360.20; 307,275.88 100% = 332,923.00

## 14. Solutions – swimming pool – Part II

Question	Answer
1	(a) 120; (b) 1,088; (c) 507; (d) 581; (e) 797.50; (f) 782.59; (g) 787.50; (h) 869, 3 times; (i) 109.60; (j) 109.14; (k) 0.1400; (l) 507.00, 711.25, 787.50, 854.00, 1,088.00; (m) 142.75; (n) 782.63
3	It is reasonably symmetrical

(Continued)

## 14. Solutions – (Continued)

Question	Answer							
4	Percentile (%)	Value	Percentile (%)	Value	Percentile (%)	Value	Percentile (%)	Value
	0	507.00						
	1	542.99	26	715.76	51	790.00	76	860.44
	2	561.90	27	724.26	52	790.88	77	866.04
	3	573.55	28	726.64	53	791.07	78	869.00
	4	581.52	29	728.51	54	792.52	79	869.01
	5	607.65	30	731.10	55	794.45	80	870.20
	6	609.14	31	741.79	56	795.64	81	872.17
	7	612.97	32	743.08	57	797.66	82	876.32
	8	619.52	33	744.27	58	806.04	83	878.00
	9	620.71	34	746.38	59	808.42	84	879.92
	10	629.10	35	748.00	60	810.40	85	884.80
	11	650.09	36	748.84	61	816.90	86	897.04
	12	652.96	37	750.03	62	821.78	87	902.06
	13	663.17	38	751.88	63	822.97	88	905.88
	14	669.00	39	755.00	64	825.00	89	914.28
	15	671.55	40	757.40	65	825.35	90	919.70
	16	678.00	41	760.58	66	827.62	91	927.16
	17	678.46	42	762.96	67	829.73	92	930.48
	18	680.84	43	763.17	68	830.00	93	937.03
	19	683.83	44	765.80	69	835.00	94	950.32
	20	689.00	45	771.20	70	835.60	95	956.60
	21	691.98	46	775.22	71	839.94	96	970.88
	22	700.44	47	777.86	72	844.36	97	985.59
	23	707.74	48	779.12	73	846.74	98	999.82
	24	709.00	49	781.55	74	848.24	99	1,016.15
	25	711.25	50	787.50	75	854.00	100	1,088.00

## 15. Solutions – buyout – Part II

Question	Answer
1	(a) 15.1; (b) 5.3; (c) 9.8; (d) 10.20; (e) 10.35; (f) 10.35; (g) 11.10; occurs 3 times; (h) 2.08; (i) 2.06
2	5.30; 9.23; 10.35; 11.58; 15.10
3	0% = 5.30; 5.89; 6.48; 6.59; 6.69; 6.97; 7.26; 7.39; 7.48; 7.54; 7.59; 7.68; 7.78; 7.91; 8.06; 8.21; 8.35; 8.47; 8.56; 8.63; 8.68; 8.76; 8.86; 8.98; 9.13 25% = 9.23; 9.27; 9.30; 9.30; 9.34; 9.44; 9.52; 9.57; 9.62; 9.67; 9.72; 9.76; 9.81; 9.86; 9.92; 10.02; 10.11; 10.16; 10.21; 10.26; 10.30; 10.30; 10.30; 10.30; 10.30 50% = 10.35; 10.40; 10.45; 10.50; 10.50; 10.50; 10.54; 10.59; 10.64; 10.69; 10.70; 10.70; 10.85; 11.05; 11.10; 11.10; 11.10; 11.10; 11.13; 11.18; 11.29; 11.44; 11.50; 11.50; 11.53 75% = 11.58; 11.60; 11.60; 11.64; 11.74; 11.94; 12.28; 12.50; 12.50; 12.50; 12.50; 12.53; 12.63; 12.71; 12.76; 12.81; 12.86; 12.96; 13.36; 13.70; 13.70; 13.73; 14.12; 14.51; 14.81; 100% = 15.10

## Chapter 3: Basic Probability and Counting Rules

### 1. Solutions – gardner's gloves

Question	Answer
1	35.90%
2	51.28%
3	5.69%
4	37.87%
5	47.34%

### 2. Solutions – market survey

Question	Answer
1	20.60%
2	4.95%
3	19.76%
4	3.90%
5	28.16%
6	25.52%
7	40.58%

### 3. Solutions – getting to work

Question	Answer
1	23.50%
2	46.81%
3	11.11%
4	80.00%
5	85.00%

### 4. Solutions – packing machines

Question	Answer
1	38.50%
2	1.93%

(Continued)

## 4. Solutions – (Continued)

Question	Answer
3	8.10%
4	3.30%

## 5. Solutions – study groups

Question	Answer
1	40.63%

## 6. Solutions – roulette

Question	Answer
1	11.11%; 88.89%
2	22.22%; 77.78%; £125.00
3	6; 66.67%; £25.00; 33.33%; £150.00
4	44.44%; 55.56%
5	0%; 11.11%
6	44.44%; 55.56%
7	44.44%; 55.56%

## 7. Solutions – sourcing agents

Question	Answer
2	38.93%
3	12.98%
4	49.62%
5	33.33%
6	53.13%

## 8. Solutions – sub-assemblies

Question	Answer
1	82.08%
2	16.86%
3	17.92%

(Continued)



## 8. Solutions – (Continued)

Question	Answer
4	1.06%
5	0.02%

## 9. Solutions – workshop

Question	Answer
1	30.00%
2	1.80%
3	11.20%
4	1.20%

## 10. Solutions – assembly

Question	Answer
1	78.6931%; 99.7949%; 99.9997%; 99.1469%
2	21.3069%; 0.2051%; 0.0003%; 0.8531%

## 11. Solutions – bicycle gears

Question	Answer
1	Values for each of the three columns are as follows: (A) 1; 1; 1 (B) 1; 2; 2 (C) 2; 3; 6 (D) 2; 4; 8 (E) 2; 5; 10 (F) 3; 4; 12 (G) 3; 7; 21 (H) 4; 7; 28 (I) 4; 8; 32 (J) 4; 9; 36

## 12. Solutions – film festival

Question	Answer
1	6,724,520
2	33,891,580,800
3	(a) 1,184,040; (b) 116,280; (c) 3,432; (d) 1
4	(a) 5,967,561,600; (b) 586,051,200; (c) 17,297,280; (d) 5,040

## 13. Solutions – flag flying

Question	Answer
1	$1.08889 \times 10^{28}$ (There are 27 countries in the European Union as of 2007)
2	8,436,285

(Continued)

## 13. Solutions – (Continued)

Question	Answer
3	80,730
4	$3.04141 \times 10^{64}$
5	1

## 14. Solutions – model agency

Question	Answer
1	54,264
2	1,307,674,368,000
3	70,959,641,905,152,000

## 15. Solutions – thalassothérapie

Question	Answer
1	40,320
2	24
3	479,001,600
4	56

The following is a possible schedule for Question 4. Note, here that the treatment, the enveloppement d'algues and application de boue marine are not put together. Even though they are different treatments, they have similarities regarding their benefits

1. bain hydromassant	6. application de boue marine	4. douche à jet
2. douche oscillante	7. hydrojet	6. application de boue marine
3. massage sous affusion	8. massage à sec	7. hydrojet
4. douche à jet	1. bain hydromassant	8. massage à sec
5. enveloppement d'algues	2. douche oscillante	1. bain hydromassant
5. enveloppement d'algues	6. application de boue marine	8. massage à sec
7. hydrojet	8. massage à sec	2. douche oscillante
8. massage à sec	1. bain hydromassant	3. massage sous affusion
1. bain hydromassant	2. douche oscillante	4. douche à jet
2. douche oscillante	3. massage sous affusion	5. enveloppement d'algues

## Chapter 4: Probability Analysis for Discrete Data

### 1. Solutions – HIV virus

Question	Answer
1	\$3,843,750
2	No change. Remains the same at €3,843,750
3	Table 1 = 1.12%; Table 2 = 0.69%
4	Data from Table 2 as it shows less dispersion

### 2. Solutions – rental cars

Question	Answer
1	27.50 cars
2	\$151,250
3	No change
4	No change
5	Laramie coefficient of variation at 7.7% is smaller than Cheyenne (12.19%)

### 3. Solutions – road accidents

Question	Answer
2	6.60 accidents/day
3	3.74
4	Yes. The coefficient of variation is 56.64% which is high
5	£505,470
6	£1,018,161
7	£1,523,631

### 4. Solutions – express delivery

Question	Answer
2	16.67% for three packages not being delivered on-time
3	4.69/month
4	3.00
5	No. The annual payment is €4,788.33
6	There are high occurrences of non-deliveries in the winter months, which might be explained by bad weather. Also high occurrences in the summer period perhaps because of high volume of air traffic

## 5. Solutions – bookcases

Question	Answer
1	20.25
2	25.00%; 22
3	He would "lose" £312.50 (opportunity cost)
4	Expected values are long-term and may not necessarily apply to the immediate following month

## 6. Solutions – investing

Question	Answer
1	\$137.50; \$56.50
2	\$104.73; \$63.03
3	−6,018.75
4	\$194.00
5	\$97.00
6	9.70%; \$26.94

## 7. Solutions – gift store

Question	Answer
1	99.53%
2	70.31%
3	0.47%
4	83.22%
5	51.55%

## 8. Solutions – European business school

Question	Answer
4	3.08%
5	17.89%
6	41.64%
7	76.25%
8	58.36%

## 9. Solutions – clocks

Question	Answer
2	1.68%
3	12.39%
4	17.37%
5	95.02%
6	82.63%

## 10. Solutions – computer

Question	Answer
2	38.74%
3	73.61%
4	65.13%
5	34.87%
6	26.39%
7	9

## 11. Solutions – bank credit

Question	Answer
2	11.10%
3	93.83%
4	82.73%
5	13.19%
6	24.52%
7	11.32%

## 12. Solutions – biscuits

Question	Answer
2	29.36%
3	33.55%
4	94.37%

(Continued)

## 12. Solutions – (Continued)

Question	Answer
5	83.22%
6	1.04%

## 13. Solutions – bottled water

Question	Answer
2	4.98%
3	16.80%
4	35.28%
5	64.72%
6	81.53%

## 14. Solutions – cash-for-gas

Question	Answer
2	No. The owner should not be satisfied as the percentage of 12 or more using the pump is only 81.52%
3	Only have one attendant at the exit kiosk. This would reduce operating costs by about 50%. If this means a longer wait for customers they may prefer to use the pumps that take credit card purchase, or they may go elsewhere! Provide an incentive to use the cash-for-gas utilization, such as lower prices, a gift. However, this would only be a suitable approach if overall more customers use the service station, and the method does not just siphon off customers who before used credit card for purchases

## 15. Solutions – cashiers

Question	Answer
2	10.63%
3	84.54%
4	4.82%
6	10.67%
7	84.91%
8	4.42%
9	They both tail-off rapidly to the right. The values are close. Since $P < 5\%$ and sample size is greater than 20 it is reasonable to use the Poisson-binomial approximation

## Chapter 5: Probability Analysis in the Normal Distribution

### 1. Solutions – renault trucks

Question	Answer
1	47.40%
2	37.46%
3	12.87%
4	42,526
5	Yes; 126,393 km
6	258,159 km

### 2. Solutions – telephone calls

Question	Answer
1	6.68%
2	86.64%
3	About 134
4	6.62%
5	146.95 seconds

### 3. Solutions – training programme

Question	Answer
1	23.39%
2	6.68%
3	8.08%
4	8.74%
5	About 38 and 74 days

### 4. Solutions – cashew nuts

Question	Answer
1	42.07%
2	8.08%
3	Minimum 123.53 g; maximum 129.97 g
4	124.69 gm

## 5. Solutions – publishing

Question	Answer
1	Almost none (0.1846)
2	Between 1 and 2 (1.5549)
3	About 5 (4.5460)
4	About 13 (12.8765)
5	Just about all of the 19 (18.6055)

## 6. Solutions – gasoline station

Question	Answer
1	4.32%
2	7.66%
3	96.95%
4	4,912 litre
5	35,457 litre

## 7. Solutions – ping-pong balls

Question	Answer
1	40.88%
2	90.50%
3	9.50%
4	About 22,624
5	369.49 mm
6	368.77 and 371.23 mm
7	Slender, with a sharp peak or leptokurtic as the coefficient of variation is small, 0.20%

## 8. Solutions – marmalade

Question	Answer
1	43.80%
2	76.00%
3	19.36%

(Continued)



## 8. Solutions – (Continued)

Question	Answer
4	About 26,755,27
5	336.63 g
6	331.63 and 348.37 g
7	3.78%

## 9. Solutions – restaurant service

Question	Answer
1	Average service time is 125.00 minutes; standard deviation is 19.96 minutes
2	46.02%
3	77.09%
4	22.91%
5	925
6	104.31 minutes

## 10. Solutions – yogurt

Question	Answer
1	13.33%
2	48.23%

## 11. Solutions – motors

Question	Answer
1	56.98%

## 12. Solutions – doors

Question	Answer
1	196 cm
2	About 3% (3.34%)

## 13. Solutions – machine repair

Question	Answer
1	2.28%
2	33.41%
3	58.89%
4	87.75%

## 14. Solutions – savings

Question	Answer
1	29.93%
2	89.03%
3	1.13%
4	Very little chance (0.04%)

## 15. Solutions – buyout – Part III

Question	Answer
1	About 7 (7.45)
2	The values are very close
3	When data follows closely a normal distribution it is not necessary to construct the ogives for this type of information

## Chapter 6: Theory and Methods of Statistical Sampling

### 1. Solutions – credit card

Question	Answer
1	27.76%
2	71.68%
3	88.81%
4	Larger the sample size the data clusters around the mean
5	That the central limit theory applies and the sample means follow a normal distribution

### 2. Solutions – food bags

Question	Answer
1	26.25%
2	22.15%
3	48.81%
4	1.19%
5	Since we have a sample of size 10 rather than just a size 1 from the population, data clusters around the mean. In Questions 1 and 3, the percentage increases as the limits are around the mean. In Questions 2 and 4, the percentage decreases as the limits are away from the mean
6	Normal as the population is normal

### 3. Solutions – telephone calls

Question	Answer
1	7.97%
2	9.87%
3	38.29%
4	39.44%
5	68.27%
6	49.38%
7	With larger sample sizes data lies closer to population mean

## 4. Solutions – soft drinks machine

Question	Answer
1	31.36 cl
2	1.36 cl; 4.11%
3	32.48 cl
4	5.69% of the time
5	33.15 cl; 0.45%

## 5. Solutions – baking bread

Question	Answer
1	11.97%
2	0.93%
3	0.04%
4	76.06%
5	98.14%
6	99.91%
7	With larger sample sizes data lies closer to population mean
8	For Questions 1 to 3 limit values are on the right side of the mean value. For Questions 4 to 6 the limit values are on either side of the mean values

## 6. Solutions – financial advisor

Question	Answer
1	42.79%
2	30.76 minutes
3	23.35%
4	33.94 minutes
5	18.17%
6	34.15 minutes
7	As the sample size increases there are the probability of being beyond 37 minutes declines as more values are clustered around the mean value
8	The population from which the samples are drawn follows a normal distribution

## 7. Solutions – height of adult males

Question	Answer
1	5.48%
2	151.33 and 200.67 cm
3	0.07%
4	163.66 and 188.34 cm
5	About 0%
6	167.78 and 184.22 cm
7	With increase in sample size data lies closer to the mean value

## 8. Solutions – Wal-Mart

Question	Answer
1	11.45%
2	75.22%

## 9. Solutions – automobile salvage

Question	Answer
1	€10.5250
2	87.29%
3	€450.00

## 10. Solutions – education and demographics

Question	Answer
1	65.53%
2	81.86%
3	75.28%
4	89.83%
5	66.07%
6	82.35%
7	The larger the sample size, the closer is the data clustered around the population proportion. Note, that in each case the percentages given are on either side of the population proportion

## 11. Solutions – World Trade Organization

Question	Answer
1	56.74%
2	61.59%
3	71.58%
4	62.26%
5	78.16%
6	86.37%
7	The larger the sample size, the closer is the data clustered around the population proportion. Note, that in each case the percentages given are on either side of the population proportion

## 12. Solutions – female illiteracy

Question	Answer
1	88.43%
2	95.63%
3	68.78%
4	84.71%
5	68.39%
6	84.11%
7	Larger the sample size, the data clusters around the population mean
8	No. Istanbul is the commercial area of Turkey and the people are educated. The sample taken would be biased if it was meant to represent the whole of Turkey

## 13. Solutions – unemployment

Question	Answer
1	4.27%
2	1.03%
3	33.52%
4	45.84%
5	33.70%
6	44.77%

(Continued)

## 13. Solutions – (Continued)

Question	Answer
7	29.65%
8	34.02%
9	As the limits given are on either side of the population proportion the percentages increase with sample size
10	The limits for France are on the left side of the distribution and as the sample size increases the percentage here declines

## 14. Solutions – manufacturing employment

Question	Answer
1	68.66%
2	84.61%
3	31.64%
4	43.56%
5	94.07%
6	99.23%
7	The larger the sample size, the more the data is clustered around the mean
8	You might say that the German data for the population is an underestimation. Perhaps a more palatable reason could be that the sample was taken from an industrial area of Germany and thus the sample is biased

## 15. Solutions – homicide

Question	Answer
1	0.06%
2	0.002%; you have a 30 times more chance of being shot in Jamaica than in Britain
3	0.3 or 30.39%
4	0.4 or 41.93%
5	The larger the sample size, the closer the values are to the mean

## Chapter 7: Estimating Population Characteristics

### 1. Solutions – ketchup

Question	Answer
1	496.71 and 503.29 g
2	495.68 and 504.32 g
3	The higher the confidence intervals, the broader are the limits
4	About 3 cases (between 61 and 62 cases)
5	A case of 20 bottles may not be random since they probably come off the production line in sequence

### 2. Solutions – light bulbs

Question	Answer
1	394.18; 552.74
2	I estimate that the mean life of these light bulbs is about 473 hours (473.46) and I am 80% confident that the mean life will lie in the range 394 (394.18) and 553 (552.74) hours
3	369.28; 577.65
4	294.91; 652.02
5	Higher the certainty, broader the range

### 3. Solutions – ski magazine

Question	Answer
1	€39,334.67 and €40,355.33
2	€39,045.82 and €40,644.18
3	The higher the confidence level, the broader are the limits

### 4. Solutions – households

Question	Answer
1	Lower limit £11.55; upper limit £12.45
2	Lower limit £11.47; upper limit £12.53
3	Lower limit £11.37; upper limit £12.63
4	The higher the confidence level, the wider are the limits



## 5. Solutions – taxes

Question	Answer
1	At 80%, lower limit is 13,923.96 and upper limit is 27,276.04. At 95%, lower limit is 9,954.38 and upper limit is 31,245.62. At 99%, lower limit is 5,824.48 and upper limit is 35,375.520
2	I estimate that the mean value of the tax returns for the year in question is \$20,600.00 and I am 95% confident that they will lie between \$9,954 and \$31,246
3	More confident you are about information (closer to 100%) the wider is the interval

## 6. Solutions – vines

Question	Answer
1	Lower limit is 13.64 and upper limit is 16.36
2	I estimate that in my vineyard this year I will have an average of 15 grape bunches per vine and I am 95% confident that this number will lie between 13.64 and 16.36 bunches.
3	Not really, lower limit is 13.62 and upper limit is 16.38

## 7. Solutions – lemons

Question	Answer
1	91.52 mg
2	86.12 and 96.12
3	85.29 and 97.75
4	The higher the confidence level, the broader are the limits

## 8. Solutions – world's largest companies

Question	Answer
1	\$41,792.57 million
2	\$4,837.83 million
3	\$32,970.82 and \$50,614.32 million
4	\$32,310.61 and \$51,274.53 million
5	Higher the certainty, broader the range
6	\$38,954.66 million
7	\$6,589.85 million
8	\$24,820.85 and \$53,088.47 million

(Continued)

## 8. Solutions – (Continued)

Question	Answer
9	\$19,337.73 and \$58,571.59 million
10	Higher the certainty, broader the range
11	For Questions 1 to 4 we have used a larger sample size than for Questions 6 to 9 and by the central limit theory, the mean values determined are probably closer to the population value. For the first four questions we have a sample size of 35 and a population of 500, which gives a value of $n/N$ of 7% so we should use the finite correction multiplier. In Questions 6 to 9 $n/N$ is 3% so the finite population multiplier is not necessary. However, we have used a Student-t distribution as the sample size is less than 30

## 9. Solutions – hotel accounts

Question	Answer
1	0.1485
2	2.5558 and 3.0442
3	2.5090 and 3.0910
4	2.4175 and 3.1825
5	The higher the confidence level, the broader is the range

## 10. Solutions – automobile tyres

Question	Answer
1	€155,083.00
2	€136,531.81 and €173,634.86 using Student-t; €137,305.22 and €172,861.45 using $z$
3	My best estimate of the total value of the tyres in inventory is €155,083 and I am 95% confident that the value lies between €136,532 and €173,635
4	€130,081.17 and €180,085.50 using Student-t; €131,718.93 and €178,447.74 using $z$
5	The higher the confidence level, the broader are the limits
6	Use random number related to the database of the tyres in inventory. Just selecting tyres from the racks at random may not result in a representative sample

## 11. Solutions – stuffed animals

Question	Answer
1	0.1175
2	\$3,712.50

(Continued)

## 11. Solutions – (Continued)

Question	Answer
3	\$3,645.88 and \$3,779.12
4	\$3,621.22 and \$3,803.78
5	Higher the confidence, broader the limits

## 12. Solutions – shampoo bottles

Question	Answer
1	0.0113
2	0.0068 and 0.0158
3	0.0050 and 0.0177
4	27,056
5	54,119
6	The conservative value requires a very high inspection quantity but since the inspection process is automatic perhaps this is not a problem. Note that this is not a random inspection as the device samples every bottle

## 13. Solutions – night shift

Question	Answer
1	0.70 or 70.00%
2	26.82% and 33.18%
3	66.82% and 73.18%
4	26.23% and 33.77%
5	66.23% and 73.77%
6	The higher the confidence limits, the broader is the range

## 14. Solutions – ski trip

Question	Answer
1	0.40 or 40%
2	Proportion who say yes: 0.6000. Lower limit is 0.4726; upper limit is 0.7274
3	Proportion who say yes: 0.6000. Lower limit is 0.4198; upper limit is 0.7802

(Continued)

## 14. Solutions – (Continued)

Question	Answer
4	The higher the level of confidence, the broader are the limits
5	1,691
6	3,383

## 15. Solutions – Hilton hotels

Question	Answer
1	0.6531 or 65.31%
2	Lower limit is 54.12% and upper limit is 76.49%
3	Lower limit is 49.49% and upper limit is 81.13%
4	The higher the level of confidence, the broader are the limits
5	68
6	136
7	There will be hotels in the Southern and Northern hemisphere and since hotel occupancy is seasonal, the data that is taken can be distorted

## Chapter 8: Hypothesis Testing for a Single Population

### 1. Solutions – sugar

Question	Answer
1	No. Critical value of $z$ is $\pm 1.9600$ and the test value of $z$ is $+1.8762$ , which is within the critical boundaries. This is a two-tail test with 2.5% in each tail
2	Yes. The $p$ -value in each tail is 3.03% which is more than 2.5% Alternatively, the total $p$ -value is 6.06% which is less than the $\alpha$ -value of 5%
3	Lower value is 999.73 and upper value is 1,012.27. The target value of 1,000 g is contained within these limits
4	Yes. Critical value of $z$ is $\pm 1.6449$ and the test value of $z$ is $+1.8762$ , which is outside the critical boundaries. This is a two-tail test with 5.0% in each tail
5	Yes. The $p$ -value in each tail is 3.03% which is less than 5.0%. Alternatively, the total $p$ -value is 6.06% which is less than the $\alpha$ -value of 10%
6	Lower value is 1,000.74 and upper value is 1,011.26. The target value of 1,000 g is not contained within these limits
7	The firms should not be under filling the bags as this would not be abiding by the net weight value printed on the label. In this case it appears that the bags are overfilled and this is costing Béghin Say money. As an illustration assume that 1 million of these bags are sold per year and they in fact contain 6 g more of sugar. This is equivalent to 6,000 kg bags of sugar. At €0.50/bag this would be a loss of an estimated €3,000/year from this production line

### 2. Solutions – neon lights

Question	Answer
1	No. Critical value of $z$ is $\pm 1.9600$ and the test value of $z$ is $-1.6771$ which is inside the critical boundary limits. This is a two-tail test with 2.5% in each tail
2	Yes. The $p$ -value in each tail is 4.68% which is greater than 2.5%. Alternatively the $p$ -value total is 9.35% which is greater than the value of 5.00%
3	Yes. Critical value of $z$ is $-1.6449$ and the test value of $z$ is $-1.6771$ . This is a one-tail test with 5.0% in the tail
4	Yes. The $p$ -value is 4.68% which is less than 5.0%
5	The purchaser can refuse the purchase and go to another supplier. Alternatively the purchaser can negotiate a lower price

### 3. Solutions – graphite lead

Question	Answer
1	No. Accept the null hypothesis. The critical value of $z$ is $\pm 1.9600$ and the test value of $z$ is $+1.8396$ . Thus test value is within the boundaries of the critical value. This is a two-tail test with 2.5% in each tail

(Continued)

## 3. Solutions – (Continued)

Question	Answer
2	$p$ -value is 3.29%. Since this is greater than 2.5% we accept the null hypothesis
3	Lower limit is 0.6989, upper limit is 0.7347. The hypothesis value of 0.7000 mm is contained within these limits
4	Yes. Reject the null hypothesis. The test value of $z$ is +1.8396 and the critical value of $z$ is now $\pm 1.6449$ . Test value is outside boundaries of the critical value
5	$p$ -value is 3.29% and since this is less than the value of 5% we reject the null hypothesis. This is a two-tail test with 5.0% in each tail
6	Lower limit is 0.7018, upper limit is 0.7318. The hypothesis value of 0.7000 mm is not within these limits
7	Since sample mean of 0.7168 mm is greater than the hypothesized value of 0.7000 mm it implies that the diameter is on the high side. Thus we would use a right-hand, one-tail test. At both 5% and 10% limits the null hypothesis should be rejected as test statistic is greater. Test statistic is 1.8396 and $z$ at 5% is 1.6449 and 1.2816 at 10%

## 4. Solutions – industrial pumps

Question	Answer
1	No. Critical value of $z$ is $-1.6449$ and the test value of $z$ is $-1.6366$ . This is a one-tail test with 5.0% in the tail. Sample mean is 99.8154
2	Yes. The $p$ -value is 5.09% which is greater than 5.0% but only just
3	Yes. Critical value of $z$ is $-1.2816$ and the test value of $z$ is $-1.6366$ . This is a one-tail test with 10.0% in the tail
4	Yes. The $p$ -value is 5.09% which is less than 10.0%
5	No change in the conclusions. Sample mean is now 99.7493 and sample test statistic is $-1.5715$ . This gives a $p$ -value of 5.80%, which is greater than 5% but less than 10%
6	The hypothesis testing could be performed with a smaller sample size which would be less expensive. However note that at the 5% level the results between the test and critical value are close

## 5. Solutions – automatic teller machine

Question	Answer
1	Yes. The critical values are $z = \pm 1.9600$ and the sample test statistic is 2.0000, which is greater than +1.9600. Note this is a two-tail test with 2.5% in each tail
2	The $p$ -value is 2.28% which is less than 2.50%. Alternatively we can say that the $p$ -value is 4.55% which is less than 5.00%
3	€3,200.70 and €3,269.30. The value of €3,200.00 is not contained within these limits

(Continued)

## 5. Solutions – (Continued)

Question	Answer
4	No. The sample test value is still 2.0000 but the critical values are $z = \pm 2.5758$ . The test value is within these boundaries. Note this is a two-tail test with 0.50% in each tail
5	The $p$ -value is 2.28% which is greater than 0.50% for one tail. Alternatively, we can say the $p$ -value is 4.55% which is greater than 1.00%
6	€3,189.92 and €3,280.08. The value of €3,200 is contained within these limits
7	If the bank has significantly less in the machines there is a risk of the machines running out of cash and this is poor customer service. Alternatively, if the bank puts significantly too much cash in the machines that is underutilized this money could be invested elsewhere to earn interest

## 6. Solutions – bar stools

Question	Answer
1	No. The critical value of $z$ is $-1.6449$ and the test value of $z$ is $-1.4142$ or inside the critical limit. This is a left-hand, one-tail test with 5.0% in the tail
2	Yes. The $p$ -value is 7.86% which is greater than 5.0%
3	Yes. The critical value of $z$ is $-1.2816$ and the test value of $z$ is $-1.4142$ or outside the critical limit. This is a left-hand, one-tail test with 10.0% in the tail
4	Yes. The $p$ -value is 7.86% which is less than 10.0%
5	From the data, the sample standard deviation is 2.54 cm, which gives a sample statistic of $-1.3937$ and a $p$ -value of 8.67%. At a significance level of 5% the Student- $t$ value is $-1.6955$ and we accept the null hypothesis. At a significance level of 10% the Student- $t$ value is $-1.3095$ and we reject the null hypothesis. Thus there is no change to the conclusions from Questions 1 to 4

## 7. Solutions – salad dressing

Question	Answer
1	No. Critical value of $z$ is $\pm 1.9600$ and the test value of $z$ is $-1.6680$ . This is a two-tail test with 2.5% in each tail
2	Yes. The $p$ -value is 4.77% which is greater than 2.5%
3	996.37 and 1,000.29 ml. The nominal value of 1,000 ml is contained within these intervals
4	Yes. Critical value of $z$ is $-1.6449$ and the test value of $z$ is $-1.6680$ . This is a one-tail test with 5.0% in the left-hand tail
5	Yes. The $p$ -value is 4.77% which is less than 5.0%
6	If as the test suggests there is significantly less than the 1,000 ml as indicated on the label, the manufacturer is not being honest to the customer and they could be in trouble with the control inspectors
7	Very sensitive as small changes in volume will swing the results either side of the barrier limits

## 8. Solutions – apples

Question	Answer
1	No. Critical value of $t$ is $\pm 2.0639$ and the test value of $t$ is $-1.9424$ . This is a two-tail test with 2.5% in each tail
2	Yes. The $p$ -value is 6.39% which is greater than 5.0%
3	195.05 and 200.15 g. The value of 200 g is contained within these intervals
4	Yes. Critical value of $t$ is $-1.7109$ and the test value of $z$ is $-1.9424$ . This is a one-tail test with 5.0% in the left-hand tail
5	Yes. The $p$ -value is 3.20% which is less than 5.0%

## 9. Solutions – batteries

Question	Answer
1	No. Accept the null hypothesis. The test value of $t$ is $\pm 2.0330$ and the critical value of $t$ is $\pm 2.1448$ . Thus test value is within the boundaries of the critical value
2	$p$ -value is 6.15% and since this is greater than the value of 5% we accept the null hypothesis
3	Yes. Accept the alternative hypothesis. The test value of $t$ is $+2.0330$ and the critical value of $t$ is now $+1.7613$ . Thus test value is outside the boundaries of the critical value – it is higher
4	$p$ -value is 3.07% and since this is less than the value of 5% we reject the null hypothesis
5	In Questions 1 and 2 we are testing to see if there is a difference thus the value of $\alpha$ is 5% but there is only 2.5% in each tail which gives a relatively high value of $t$ . In Questions 3 and 4, we are still testing at the 5% significance level but all of this area is the right tail. Thus $t$ is smaller than for the two-tail test and explains the shift from acceptance to rejection

## 10. Solutions – hospital emergency

Question	Answer
1	No. Critical value of $t$ is $\pm 2.1448$ and the test value of $t$ is $+2.0859$ . This is a two-tail test with 2.5% in each tail
2	Yes. The $p$ -value is 5.58% which is greater than 5.0%
3	9.92 and 15.68 minutes. The value of 10 minutes is contained within these intervals
4	Yes. Critical value of $t$ is $+1.7613$ and the test value of $t$ is $+2.0859$ . This is a one-tail test with 5.0% in the right-hand tail
5	Yes. The $p$ -value is 2.79% which is less than 5.0%
6	The test for being greater than 10 minutes. If this is the case, the hospital is not meeting its objectives with the risk to the life of patients



## 11. Solutions – equality for women

Question	Answer
1	Accept the null hypothesis. The critical value of the test is $\pm 2.5758$ and the test value is $-2.4637$ . This is a two-tail test with 0.5% in each tail
2	In each tail the $p$ -value is 0.69% which is greater than the 0.5%. Alternatively the $p$ -value total is 1.38% which is greater than the $\alpha$ -value of 1%
3	Limits are 15.45% and 45.66%. The value of 45% is within these limits so verifies that we should accept the null hypothesis
4	Reject the null hypothesis. The critical value of the test is $\pm 1.9600$ and the test value is $-2.4637$ . This is a two-tail test with 2.5% in each tail
5	In each tail the $p$ -value is 0.69% which is less than the 2.5%. Alternatively the $p$ -value total is 1.38% which is less than the $\alpha$ -value of 5%
6	Limits are 19.06% and 42.05%. The value of 45% is outside these limits and so verifies that we should reject the null hypothesis
7	At the significance level of 1% the test would support the conclusions. At the significance level of 5% indications are that the pay gap is more than 45%

## 12. Solutions – gas from Russia

Question	Answer
1	From the table proportion in Italy imports from Russia is 34.20%. Accept the null hypothesis. The critical value of the test is $\pm 1.9600$ and the test value is $-1.7711$ which is inside the critical boundaries. This is a two-tail test with 2.5% in each tail
2	In each tail the $p$ -value is 3.83% which is greater than 2.5%. Alternatively the $p$ -value total is 7.65% which is greater than the $\alpha$ -value of 5%
3	Limits are 4.28% and 35.72%. The value of 34.20% is within these limits so verifies that we should accept the null hypothesis
4	Reject the null hypothesis. The critical value of the test is $\pm 1.6449$ and the test value is $-1.7711$ which is outside the critical boundaries. This is a two-tail test with 5.0% in each tail
5	In each tail the $p$ -value is 3.83%, which is less than the 5.0%. Alternatively the $p$ -value total is 7.65% which is less than the $\alpha$ -value of 10%
6	Limits are 6.81% and 33.19%. The value of 34.02% is not within these limits so verifies that we should reject the null hypothesis
7	From the table proportion Poland imports from Russia is 86.81%. Accept the null hypothesis. The critical value of the test is $\pm 1.9600$ and the test value is $+0.3074$ which is inside the critical boundaries. This is a two-tail test with 2.5% in each tail
8	In each tail the $p$ -value is 37.93% which is greater than 2.5%. Alternatively the $p$ -value total is 75.85% which is greater than the $\alpha$ -value of 5%

(Continued)

## 12. Solutions – (Continued)

Question	Answer
9	Limits are 77.36% and 99.78%. The value of 86.81% is within these limits so verifies that we should accept the null hypothesis
10	Accept the null hypothesis. The critical value of the test is $\pm 1.6449$ and the test value is $+0.3074$ which is inside the critical boundaries. This is a two-tail test with 5.0% in each tail
11	In each tail the $p$ -value is 37.93% which is greater than 5.0%. Alternatively the $p$ -value total is 75.85% which is greater than the $\alpha$ -value of 10%
12	Limits are 79.16% and 97.98%. The value of 86.81% is within these limits so verifies that we should accept the null hypothesis
13	From this data it appears that Italy is tending to use less than the contractual amount of natural gas from Russia. It perhaps can easily call on imports from say Libya or other North African countries. On the other hand Poland seems to be using close to or more than the contractual amount of natural gas from Russia. This is perhaps because historically and geographically Poland is closer to Russia and is not yet in a position to exploit other suppliers

## 13. Solutions – international education

Question	Answer
1	Accept the published value or the null hypothesis. The critical value of the test is $\pm 2.5758$ and the test value is $+2.0710$ which is inside the critical boundaries. This is a two-tail test with 0.5% in each tail
2	In each tail the $p$ -value is 1.92% which is greater than 0.5%. Alternatively the $p$ -value total is 3.84% which is greater than the $\alpha$ -value of 1%
3	Limits are 16.05% and 46.17%. The value of 19.0% is within these limits so verifies that we should accept the null hypothesis
4	Reject the null hypothesis. The critical value of the test is $\pm 1.9600$ and the test value is $+2.0710$ which is outside the critical boundaries. This is a two-tail test with 2.5% in each tail
5	In each tail the $p$ -value is 1.92%, which is less than 2.5%. Alternatively the $p$ -value total is 3.84% which is less than the $\alpha$ -value of 5%
6	Limits are 19.65% and 42.57%. The value of 19.0% is not within these limits so verifies that we should reject the null hypothesis
7	Accept the published value or the null hypothesis. The critical value of the test is $\pm 2.5758$ and the test value is $+2.2546$ which is inside the critical boundaries. This is a two-tail test with 0.5% in each tail
8	In each tail the $p$ -value is 1.21% which is greater than 0.5%. Alternatively the $p$ -value total is 2.42% which is greater than the $\alpha$ -value of 1%
9	Limits are 9.97% and 34.47%. The value of 11.5% is within these limits so verifies that we should accept the null hypothesis

(Continued)

## 13. Solutions – (Continued)

Question	Answer
10	Reject the null hypothesis. The critical value of the test is $\pm 1.9600$ and the test value is $+2.2546$ which is outside the critical boundaries. This is a two-tail test with 2.5% in each tail
11	In each tail the $p$ -value is 1.21%, which is less than 2.5%. Alternatively the $p$ -value total is 2.42% which is less than the $\alpha$ -value of 5%
12	Limits are 12.90% and 31.54%. The value of 11.5% is not within these limits so verifies that we should reject the null hypothesis

## 14. Solutions – US employment

Question	Answer
1	Accept the published value or the null hypothesis. The critical value of the test is $\pm 1.9600$ and the test value is $-0.0341$ which is inside the critical boundaries. This is a two-tail test with 2.5% in each tail
2	In each tail the $p$ -value is 48.64% which is greater than 2.5%. Alternatively the $p$ -value total is 97.28% which is greater than the $\alpha$ -value of 5%
3	Limits are 0.18% and 9.46%. The value of 4.90% is within these limits so verifies that we should accept the null hypothesis
4	Accept the published value or the null hypothesis. The critical value of the test is $\pm 1.6449$ and the test value is $-0.0341$ which is inside the critical boundaries. This is a two-tail test with 5.0% in each tail
5	In each tail the $p$ -value is 48.64% which is greater than 5.0%. Alternatively the $p$ -value total is 97.28% which is greater than the $\alpha$ -value of 10%
6	Limits are 0.92% and 8.72%. The value of 4.90% is within these limits so verifies that we should accept the null hypothesis
7	Reject the published value or the null hypothesis. The critical value of the test is $\pm 1.9600$ and the test value is $+1.9998$ which is outside the critical boundaries. This is a two-tail test with 2.5% in each tail
8	In each tail the $p$ -value is 2.28% which is less than 2.5%. Alternatively the $p$ -value total is 4.55% which is less than the $\alpha$ -value of 5%
9	Limits are 4.99% and 14.28%. The value of 4.90% is not within these limits so verifies that we should reject the null hypothesis
10	Reject the published value or the null hypothesis. The critical value of the test is $\pm 1.6449$ and the test value is $+1.9998$ which is outside the critical boundaries. This is a two-tail test with 10.0% in each tail
11	In each tail the $p$ -value is 2.28% which is less than 5.0%. Alternatively the $p$ -value total is 4.55% which is less than the $\alpha$ -value of 10%

(Continued)

## 14. Solutions – (Continued)

Question	Answer
12	Limits are 5.74% and 13.54%. The value of 4.90% is not within these limits so verifies that we should reject the null hypothesis
13	The data for Palo Alto indicates low levels of unemployment. This region is a service economy principally in the hi-tech sector. People are young and are more able to move from one company to another. Detroit is manufacturing and this sector is in decline. People are older and do not have flexibility to move from one industry to another

## 15. Solutions – Mexico and the United States

Question	Answer
1	0.10%
2	Accept the indicated value or the null hypothesis. The critical value of the test is $\pm 1.9600$ and the test value is $-1.9102$ which is inside the critical boundaries. This is a two-tail test with 2.5% in each tail
3	In each tail the $p$ -value is 2.81% which is greater than 2.5%. Alternatively the $p$ -value total is 5.61% which is greater than the $\alpha$ -value of 5%
4	Limits are 7.29% and 31.30%. The value of 31.00% is within these limits so verifies that we should accept the null hypothesis
5	Reject the indicated value or the null hypothesis. The critical value of the test is $\pm 1.6449$ and the test value is $-1.9102$ which is outside the critical boundaries. This is a two-tail test with 5.0% in each tail
6	In each tail the $p$ -value is 2.81% which is less than 5.0%. Alternatively the $p$ -value total is 5.61% which is less than the $\alpha$ -value of 10%
7	Limits are 9.22% and 29.37%. The value of 31.00% is not within these limits so verifies that we should reject the null hypothesis
8	Reject the published value or the null hypothesis. The critical value of the test is $\pm 1.9600$ and the test value is $-8.7059$ which is well outside the critical boundaries. This is a two-tail test with 2.5% in each tail
9	In each tail the $p$ -value is 0% which is less than 2.5%. Alternatively the $p$ -value total is 0% which is less than the $\alpha$ -value of 5%
10	Limits indicate $-9.21\%$ (0%) and 16.23%. The value of 60% is far from being within these limits so verifies that we should soundly reject the null hypothesis
11	Reject the published value or the null hypothesis. The critical value of the test is $\pm 1.6449$ and the test value is $-8.7059$ which is well outside the critical boundaries. This is a two-tail test with 10.0% in each tail

(Continued)

## 15. Solutions – (Continued)

Question	Answer
12	In each tail the $p$ -value is 0% which is less than 5.0%. Alternatively the $p$ -value total is 0% which is less than the $\alpha$ -value of 10%
13	Limits indicate $-7.16\%$ (0%) and $14.18\%$ . The value of 60% is not within these limits so verifies that we should reject the null hypothesis
14	Random sampling for verifying that an individual is from Mexico can only be done from census data. Taking a random sample of foreign-born individuals would be biased. To an interviewer, and undocumented alien is unlikely to be truthful and nothing says that the random sample were documented, or undocumented. Exercise underscores difficulty of obtaining reliable information about illegal immigrants

## Chapter 9: Hypothesis Testing for Different Populations

### 1. Solutions – gasoline prices

Question	Answer
1	Null hypothesis is $H_0: \mu_1 = \mu_2$ . Alternative hypothesis is $H_1: \mu_1 \neq \mu_2$ , where $\mu_1$ is the mean value in January 2006 and $\mu_2$ is the mean value in June 2006
2	Yes. At 2% significance, the critical value of $z$ is $-2.3263$ and the sample test statistic is $-2.8816$
3	Significance level is 2% and the $p$ -value is 0.1979%
4	No. The conclusions do not change. At 5% significance, the critical value of $z$ now becomes $-1.9600$ . The sample test statistic remains at $-2.8816$
5	Significance level is 5% and the $p$ -value is still 0.1979%
6	In 2006 the price of crude oil rose above \$70/barrel. Gasoline is a refined product from crude oil

### 2. Solutions – tee shirts

Question	Answer
1	Null hypothesis is $H_0: \mu_S = \mu_I$ . Alternative hypothesis is $H_1: \mu_S \neq \mu_I$ , where $\mu_S$ is the mean value for Spain and $\mu_I$ is the mean value for Italy. This is a two-tail test
2	No. At 1% significance, the critical value of $z$ is $\pm 2.5758$ and the sample test statistic is 2.3733. Accept the null hypothesis
3	Significance level is 1% with 0.5% in each tail. The $p$ -value in each tail is 0.88%
4	Yes. The conclusions change. At 5% significance, the critical value of $z$ is now $\pm 1.9600$ and the sample test statistic is still 2.3733. Reject the null hypothesis
5	Significance level is 5% with 2.50% in each tail. The $p$ -value is still 0.88%
6	Null hypothesis is $H_0: \mu_S \leq \mu_I$ . Alternative hypothesis is $H_1: \mu_S > \mu_I$ , where $\mu_S$ is the mean value for Spain and $\mu_I$ is the mean value for Italy. This is a one-tail test
7	Yes. At 1% significance, the critical value of $z$ is $\pm 2.3263$ and the sample test statistic is 2.3733. Reject the null hypothesis
8	Significance level is 1% with 1.00% in the tail. The $p$ -value is still 0.88%

### 3. Solutions – inventory levels

Question	Answer
1	Null hypothesis is $H_0: \mu_F \leq \mu_I$ . Alternative hypothesis is $H_1: \mu_F > \mu_I$ , where $\mu_F$ is the mean value for stores using FAX and $\mu_I$ is the mean value for those using Internet. Thus, this is a one-tail test
2	No. At 1% significance, the critical value of Student- $t$ is 2.4999 and the sample test statistic $t$ is 2.4320

(Continued)

## 3. Solutions – (Continued)

Question	Answer
3	Significance level is 1% and the $p$ -value is 1.16%
4	Yes. At 5% significance, the critical value of Student- $t$ is 1.7139 and the sample test statistic $t$ is still 2.4320
5	Significance level is 5% and the $p$ -value is still 1.16%
6	The statistical evidence implies that when orders are made by FAX, the delivery is longer. For this reason, those stores using FAX keep a higher coverage of inventory in order to minimize the risk of a stockout. Perhaps FAX orders sit in an in-tray before they are handled. However Internet orders are normally processed immediately as they would go into the supplier's database

## 4. Solutions – restaurant ordering

Question	Answer
1	Null hypothesis is $H_0: \mu_D = \mu_S$ . Alternative hypothesis is $H_0: \mu_D \neq \mu_S$ , where $\mu_D$ is the mean value using database ordering and $\mu_S$ is the mean value for standard method. Since we are asking is there a difference, this is a two-tail test
2	No. At a 1% significance level, accept the null hypothesis. The critical value of Student- $t$ is $-2.8784$ and the sample test statistic $t$ is 4.3436.
3	Significance level is 1% with 0.5% in the tail and the sample $p$ -value is 0.04%
4	$H_0: \mu_D - \mu_S \leq 1.25 \mu_S$ . Alternative hypothesis is $H_0: \mu_D - \mu_S > 1.25 \mu_S$ . This is now a one-tail test
5	Using this new criterion we reject the null hypothesis. The critical value of Student- $t$ is now 2.5524 and the sample test statistic $t$ is still 2.9037. This is a one-tail test
6	Significance level is 1% and the sample $p$ -value is 0.87% (a one-tail test)
7	The investment is too much to install a database system. It could be that they prefer the manual system as they can more easily camouflage real revenues for tax purposes!

## 5. Solutions – sales revenues

Question	Answer
1	Average store sales of those in the pilot programme before is £260,000/store. Thus benchmark is an increase of 10% or £26,000/store
2	Null hypothesis $H_0: \mu_1 - \mu_2 \leq £26,000$ . Alternative hypothesis $H_1: \mu_1 - \mu_2 > £26,000$
3	No. Sample Student- $t$ is 2.7404 and critical Student- $t$ is 2.7638
4	Significance level is 1% and $p$ -value is 1.04%
5	Yes. Sample Student- $t$ is still 2.7404 but critical Student- $t$ is now 1.8125

(Continued)

## 5. Solutions – (Continued)

Question	Answer
6	Significance level is 5% and $p$ -value is 1.04%
7	Even at 1% significance the results are very close to the critical test level. Retail sales are very sensitive to seasons so we need to be sure that the sampling data was carried out in like periods

## 6. Solutions – hotel yield rate

Question	Answer
1	Null hypothesis $H_0: \mu_1 - \mu_2 \leq 10\%$ . Alternative hypothesis $H_1: \mu_1 - \mu_2 > 10\%$
2	No. Sample Student- $t$ is 1.1959 and critical Student- $t$ is 2.8965
3	Significance level is 1% and $p$ -value is 13.30%
4	Yes. Sample Student- $t$ is still 1.1959 but critical Student- $t$ is now 1.1081
5	Significance level is 15% and $p$ -value is still 13.30%
6	That depends if these new yield rate levels are sustainable and if the advertising cost and the reduction in price levels have led to a net income increase for the hotels. Also, note that a high significance level is used before we can say that the objectives have been reached

## 7. Solutions – migraine headaches

Question	Answer
1	Monthly average of the migraine attacks before is 24.4 and 50% of this is 12.20. Null hypothesis $H_0: \mu_1 - \mu_2 \leq 12.20$ . Alternative hypothesis $H_1: \mu_1 - \mu_2 > 12.20$
2	No. Sample Student- $t$ is 0.3509 and critical Student- $t$ is 2.8214
3	Significance level is 1% and sample $p$ -value is 36.69%
4	No. Sample Student- $t$ is still 0.3509 and critical Student- $t$ is 1.3830
5	Significance level is 10% and sample $p$ -value is still 36.69%
6	About a 60% reduction would be required. At a 60.28% reduction this gives a mean value afterwards of 9.7 headaches per month. This gives a sample Student- $t$ of 2.8257 and a $p$ -value of 0.99%
7	Patients may experience less attacks knowing they are undertaking a test (physiological effect). Also eliminating coffee may reduce or increase stress or may reduce or increase consumption of sugar so all variables have not been isolated



## 8. Solutions – hotel customers

Question	Answer
1	Null hypothesis is $H_0: p_6 = p_1$ . Alternative hypothesis is $H_0: p_6 \neq p_1$ . The indices refer to the year the test was made. This is a two-tail test
2	No. Sample z-value is 1.6590 and critical z-value is 1.9600
3	Significance level is 5% with 2.5% in each tail. Sample z gives 4.86% in each tail
4	Null hypothesis is $H_0: p_6 \leq p_1$ . Alternative hypothesis is $H_0: p_6 > p_1$ . The indices refer to the year the test was made. This is a one-tail, right-hand test
5	Yes. Sample z-value is still 1.6590 and critical z-value is now 1.6449
6	Significance level is 5%. Sample z gives 4.86% in the tail
7	The data results are close. The fitness craze seems to have leveled off in this decade. Other variables that need to be considered in the sampling are the gender of customers, and whether the customers at the hotel are for business or pleasure?

## 9. Solutions – flight delays

Question	Answer
1	Null hypothesis is $H_0: p_{05} = p_{96}$ . Alternative hypothesis is $H_0: p_{05} \neq p_{96}$ . The indices refer to the year the test was made. This is a two-tail test
2	No. Accept null hypothesis. Sample z-value is 2.4972 and critical z-value is 2.5758
3	Significance level is 1% with 0.5% in each tail. Sample z gives 0.63% in each tail
4	Yes. Reject null hypothesis. Sample z-value is still 2.4972 and critical z-value is 1.9600
5	Significance level is 5% with 2.5% in each tail. Sample z gives 0.63% in each tail
6	Null hypothesis is $H_0: p_{05} \leq p_{96}$ . Alternative hypothesis is $H_0: p_{05} > p_{96}$ . The indices refer to the year the test was made. This is a one-tail test
7	Reject the null hypothesis. Sample z-value is 2.4972 and critical z-value is 2.3263
8	0.63% or the $p$ -value of the sampling experiment
9	The experiment could be defined better. What is a major European airport? When was the sampling made – winter or summer? In winter, there are weather problems. In summer traffic is heavier. Certainly though airplane traffic has increased significantly and delays have increased. A 20-minute delay may be excessive for some passengers

## 10. Solutions – World Cup

Question	Answer
1	Null hypothesis is $H_0: p_{98} = p_{06}$ . Alternative hypothesis is $H_0: p_{98} \neq p_{06}$ . The indices refer to the year the test was made. This is a two-tail test

(Continued)

## 10. Solutions – (Continued)

Question	Answer
2	No. Accept null hypothesis. Sample z-value is $-2.4385$ and critical z-value is $\pm 2.5758$
3	Significance level is 1% with 0.5% in each tail. Sample z gives 0.74% in each tail
4	Yes. Reject null hypothesis. Sample z-value is still $-2.4385$ and critical z-value is $\pm 1.9600$
5	Significance level is 5% with 2.5% in each tail. Sample z gives 0.74% in each tail
6	Null hypothesis is $H_0: p_{06} \leq p_{98}$ . Alternative hypothesis is $H_0: p_{06} > p_{98}$ . The indices refer to the year the test was made. This is a one-tail test. Note, here we have reversed the position and put the year 2006 first. We could also say, $H_0: p_{98} \geq p_{06}$ . Alternative hypothesis is $H_0: p_{98} < p_{06}$ though this does not fit with the way the question is phrased
7	Reject the null hypothesis. Sample z-value is 2.4385 and critical z-value is 2.3263
8	Significance level is 1.00% and $p$ -value is 0.74%
9	The experiment should be defined. What was the gender of the people sample? (Men are more likely to be enthusiastic about the World Cup than omen.) The country surveyed is also important. Not all European countries have a team in the World Cup

## 11. Solutions – travel time and stress

Question	Answer
1	Null hypothesis $H_0: p_{HS} = p_{MS} = p_{LS}$ . Alternative hypothesis $H_1: p_{HS} \neq p_{MS} \neq p_I \neq p_{LS}$ where the indices refer to the stress level
2	No. Critical chi-square value is 13.2767 and sample chi-square value is 9.5604
3	Significance level is 1% and $p$ -value is 4.85%
4	Yes. Critical chi-square value is 9.4877 and sample chi-square value is still 9.5604
5	Significance level is 5% and $p$ -value is 4.85%
6	Yes. Percentage returns are 93% (186/200)
7	Stress may not only be a result of travel but could be personal reasons or work itself

## 12. Solutions – investing in stocks

Question	Answer
1	Null hypothesis $H_0: p_U = p_G = p_I = p_E$ . Alternative hypothesis $H_1: p_U \neq p_G \neq p_I \neq p_E$ where the indices refer to the first letter of the country
2	No. It seems that the investment strategy of individuals is independent of the country of residence. Sample chi-square is 10.9108 and critical chi-square is 11.3449
3	Significance level is 1% and $p$ -value is 1.22%

(Continued)

## 12. Solutions – (Continued)

Question	Answer
4	Yes. It seems that the investment strategy of individuals is dependent on the country of residence. Sample chi-square is 10.9108 and critical chi-square is 8.9473
5	Significance level is 3% and $p$ -value is still 1.22%
6	According to the sample, nearly 62% of individuals invest in stocks in the United States while in the European countries it is between 50% and 52%. The higher level in the United States is to be expected as Americans are usually more risk takers and the social security programmes are not as strong as in Europe

## 13. Solutions – automobile preference

Question	Answer
1	Null hypothesis $H_0: p_G = p_F = p_E = p_I = p_S$ . Alternative $H_1: p_G \neq p_F \neq p_E \neq p_I \neq p_S$ where the indices refer to the country
2	Yes. Critical chi-square value is 31.9999 and sample chi-square value is 38.7764
3	Significance level is 1% and $p$ -value is 0.12%
4	0.12% the $p$ -value of the sample data
5	Results are not surprising. Even though we have the European Union people are still nationalistic about certain things including the origin of the products they buy

## 14. Solutions – newspaper reading

Question	Answer
1	Null hypothesis $H_0: p_1 = p_2 = p_3 = p_4 = p_5$ . Alternative $H_1: p_1 \neq p_2 \neq p_3 \neq p_4 \neq p_5$ where the indices refer to the salary category
2	No. Critical chi-square value is 15.5073 and sample chi-square value is 14.5675
3	Significance level is 5.00% and $p$ -value is 6.81%
4	Yes. Critical chi-square value is 13.3616 and sample chi-square value is 14.5675
5	Significance level is 10.00% and $p$ -value is 6.81%
6	There are many variables in this sample experiment for example the country of readership, the newspaper that is read, and the profession of the reader. Depending on the purpose of this analysis another sampling experiment should be carried out to take into account these other variables

## 15. Solutions – wine consumption

Question	Answer
1	Null hypothesis $H_0: p_E = p_F = p_I = p_S = p_U$ . Alternative $H_1: p_E \neq p_F \neq p_I \neq p_S \neq p_U$ where the indices refer to the first letter of the country
2	No. It seems that the amount of wine consumed is independent of the country of residence. Sample chi-square is 23.1757 and critical chi-square is 26.2170
3	Significance level is 1% and $p$ -value is 2.63%
4	3.00% since the $p$ -value of the sample is 2.63%
5	22.7418
6	In Europe wine consumption per capita is down but wine exporters such as South Africa, Chile, Australia, and California are up – at the expense of French wine!

## Chapter 10: Forecasting and Estimating from Correlated Data

### 1. Solutions – safety record

Question	Answer
2	$\hat{y} = 179.6883 - 0.0895x$
3	An annual reduction of 0.895 incidents per 200,000 hours worked
4	The coefficient of determination, $r^2$ , which is 0.9391 or the coefficient of correlation of $-0.9691$ . Note that the coefficient of correlation has the same negative sign as the slope of the regression line
5	0.11 reported incidents per 200,000 hours worked
6	$-0.16$ reported incidents per 200,000 hours worked. We cannot have negative reported incidents of safety
7	It suggests that the safety record of ExxonMobil has reached its minimum level and if anything will now decline in an exponential fashion. Using the linear regression equation to forecast for 2009 already shows a negative 0.07 incidents per 200,000 hours worked

### 2. Solutions – office supplies

Question	Answer
2	$\hat{y} = 17.4441 + 0.9002x$
3	This might correspond to increased purchases due to back-to-school period
4	$r^2$ is 0.9388 and $r = +0.9689$ which indicate a good correlation so there is a reasonable reliability for the regression model
5	£2,701 increase per quarter
6	June 2007 is £66,057. This is most reliable as it is closer period to actual data. December 2008 is £82,261 and December 2009 is £93,064
7	Forecasting using revenues has a disadvantage in that it covers up the increases in prices over time. From an operating point of view it may be better to report unit sales in order to see those items which are the major revenue contributors

### 3. Solutions – road deaths

Question	Answer
2	$\hat{y} = 498,620.49 - 245.59x$ – using the years as the x-values
3	The coefficient of determination, $r^2$ . Here the value is 0.9309. Since it is quite close to unity, the regression line seems to be a reasonable predictor for future values. Coefficient of correlation is $-0.9649$
4	A reduction of about 246 deaths per year
5	4,983

(Continued)

## 3. Solutions – (Continued)

Question	Answer
6	71
7	The value of 4,983 for Question No. 5 is achievable with a lot of effort on the part of drivers and national and local authorities. The value of 71 for Question No. 6 is desirable but probably unlikely. These values indicate the danger of making forecasts beyond the collected data and way into the future

## 4. Solutions – carbon dioxide

Question	Answer
1	$\hat{y} = 28.5455x - 55,256.0000$
2	Coefficient of determination is 0.8822 and the coefficient of correlation is 0.9393. For the data given they are greater than 0.8 and indicate a reasonable relationship
3	28.54 millions of metric tons of carbon per year
4	2,120 millions of metric tons of carbon per year
5	Between 2,043 and 2,198 million metric tons of carbon equivalent. To be correct we use the Student- <i>t</i> distribution.
6	2,406 millions of metric tons of carbon per year
7	Worldwide there is a lot of discussion and action on reducing carbon dioxide emissions, so that the rate of increase may decline and thus making forecasts using the given data may be erroneous

## 5. Solutions – restaurant serving

Question	Answer
2	Coefficient of determination, $r^2 = 0.8627$ or coefficient of correlation, $r = +0.9288$
3	$\hat{y} = 5.8088 + 2.7598x$
4	2.8 minutes for every employee absent
5	5.8 minutes
6	22.4 minutes
7	61.0 minutes. A value of 20 for <i>x</i> is beyond the range of the collected data. This would be an unacceptable internal operating situation (36% of employees are absent). Further, externally many clients would not wait over an hour to get served
8	People order different meals, and different cooking conditions (Rare, medium, and well cooked steaks). Planning not optimum in the kitchen or in the dining room. Kitchen staff not polyvalent so bottlenecks for some dishes. Clients change their minds

## 6. Solutions – product sales

Question	Answer
2	Regression equation is $\hat{y} = 10,536x + 11,944$ ; $r^2 = 0,8681$ . The relationship is strong
3	28 units/day
4	65 units/day. The shelf space of $5\text{ m}^2$ is beyond the range of the data collected. Is it realistic for the store?
5	The extent of product exposure has an impact on sales

## 7. Solutions – German train usage

Question	Answer
2	Regression equation is, $\hat{y} = 57.2459 - 3.5570x$ ; $r^2$ is 0.5054, relationship is not strong
4	Regression equation is, $\hat{y} = 2.7829 + 1.7849x$ ; $r^2$ is 0.9156, relationship is strong
5	About 21 million (20.6319)
6	$\hat{y} = 0.0706x^2 + 0.2909x + 8.6512$ ; $r^2 = 0.9483$ a strong relationship. However this profile cannot be maintained indefinitely

## 8. Solutions – cosmetics

Question	Answer
1	$\hat{y} = 230,981.21 + 2.28 * \text{advertising budget} + 745.40 * \text{No. of sales persons}$ . Coefficient of multiple determination $r^2$ is 0.9640 which indicates a strong relationship
2	£485,649
3	Lower limit is £381,835 and upper limit is £589,463
4	$\hat{y} = 205,987.21 + 2.31 * \text{advertising budget} + 732.79 * \text{sales persons} + 379.65 * \text{No. of Yam parties}$ . Coefficient of multiple determination $r^2$ is 0.9650, a strong relationship
5	£477,485
6	Lower limit is £366,799 and upper limit is £588,171

## 9. Solutions – hotel revenues

Question	Answer
1	$\hat{y} = 7.400x - 14,742.60$
2	$r^2 = 0.9099$
3	\$7.4 million per year

(Continued)

## 9. Solutions – (Continued)

Question	Answer
4	\$116.60 million. Lower limit \$102.70 million and upper 130.70 million
5	\$205.40 million. Lower limit €191.50 million and upper 219.30 million
6	$\hat{y} = 0.7652x^2 - 3,053.9712x + 3,047,387.6424$
7	$r^2 = 0.9722$
8	€348.85 million
9	€687.90 million
10	2008 is closer to the period when the data was measured and so is more reliable. The further away in the time horizon that forecasts are made, the more difficult it is to make a forecast

## 10. Solutions – Hershey

Question	Answer
1	The number of shares more than doubles in mid-1996 and mid-2004 indicating a stock split in those periods
3	$\hat{y} = 732.7656x - 5169.7610$
4	The coefficient of determination, $r^2$ . Here the value is 0.8942. Since it is greater than 0.8 the regression line is a reasonable predictor for future values
5	\$2,931/year
6	\$90,090
7	\$79,248,61 and \$100,930,94
8	A severe economic downturn which could depress stock prices
9	Probably not significantly reduced. Hershey has been around for a long time. People will continue to eat chocolate and related foods. Hershey is expanding their market into Asia so that it should help to keep up their revenues

## 11. Solutions – compact discs

Question	Answer
2	$\hat{y} = 15.6727x - 31,230.6182$ ; $r^2$ is 0.9275
3	224.55 million
4	428.29 million
5	$\hat{y} = 1.0114x^2 - 4,028.7705x + 4,012,193.1091$ ; $r^2 = 0.9522$ , closer to unity
6	273.11 million

(Continued)



## 11. Solutions – (Continued)

Question	Answer
7	844.99 million. Note the value obtained depends on how many decimal places are used for the coefficient. Answers to Question 6 and 7 use eight decimal places
8	Using polynomial relationships implies that we can reach infinite values. The answer to Question 7 is almost twice that of Question 4. In fact sales of CDs are now declining with Internet downloading of music and further, there is evidence that the CD, introduced in 1982, will soon be replaced by other technologies. <sup>1</sup>
<sup>1</sup> At 25, compact disc faces retirement, <i>International Herald Tribune</i> , 17 August 2007, p. 12.	

## 12. Solutions – US imports

Question	Answer
2	1960–1964; $\hat{y} = 1,039.50x - 2,023,238.40$ ; $r^2 = 0.9158$ 1965–1969; $\hat{y} = 3,609.20x - 7,070,763.00$ ; $r^2 = 0.9712$ 1975–1979; $\hat{y} = 27,941.80x - 55,088,472.80$ ; $r^2 = 0.9954$ 1985–1989; $\hat{y} = 35,791.80x - 70,710,080.20$ ; $r^2 = 0.9973$ 1995–1999; $\hat{y} = 68,034.40x - 134,988,756.40$ ; $r^2 = 0.9758$ 2002–2005; $\hat{y} = 175,599.60x - 350,416,159.10$ ; $r^2 = 0.9767$
3	Using the actual data for 1960–1964; forecast for 2006 is \$61,998.60 million Using the actual data for 1965–1969; forecast for 2006 is \$169,292.20 million Using the actual data for 1975–1979; forecast for 2006 is \$962,778.00 million Using the actual data for 1985–1989; forecast for 2006 is \$1,088,270.60 million Using the actual data for 1995–1999; forecast for 2006 is \$1,488,250.00 million Using the actual data for 2002–2005; forecast for 2006 is \$1,836,638.50.00 million
4	The actual value for 2006 is \$1,861,380 million. Even though the coefficients of regression determined in Question 2 show a strong correlation for the time series data, only using the data for 2002–2005 is a close agreement (a difference of some 1%). This shows that the closer your actual data is to the forecast there is a better reliability. Also, you must know the business environment – in this case there has been a significant change in the volume of US imports since 1960
5	1960–2006; $\hat{y} = 32,733.7090x - 64,449,298.3373$ ; $r^2 = 0.8368$
6	$\hat{y} = 0.0000e^{0.1087x}$ ; $r^2 = 0.9700$
7	$\hat{y} = 1.2310x^4 - 9,737.5400x^3 + 28,885,675.8297x^2 - 38,083,581,799.6840x + 18,829,057,563,855.8000$ ; $r^2 = 0.9935$
8	Linear: \$1,345,457 million; exponential: 3,860,721; polynomial: 284,264,832
9	\$2,573,675 million
10	As the years move on, the slope of the curve increases. Thus using linear regression based on earlier period data is not useful. The curve developed using the polynomial function indicates a close fit and this is corroborated by the coefficient of determination, which is close to unity. From this we might conclude that the forecast developed from this relation for 2010 is the most reliable

## 13. Solutions – English pubs

Question	Answer
1	There is a seasonal variation with highest sales in the spring and summer which is to be expected. Also, sales are increasing over time
2	$\hat{y} = 32,515.4283x + 1,013,146.9843$
3	Ratios for 2004 are winter 0.1997; spring 1.5298; summer 2.1763; fall 0.0873. This means that for 2004 winter sales are 80.03% below the yearly average, spring 52.98% above, summer 117.63% above and winter 91.27% below
4	Winter 0.2052; spring 1.5606; summer 2.1427; fall 0.0874
5	$r^2 = 0.9031$
6	Winter 328,017; spring 2,551,173; summer 3,580,963; fall 149,200 litre
7	8,402,514 litre in 2010. There is a lot of concern about the amount of beer consumed by the English and the hope is that consumption will decline or at least not increase

## 14. Solutions – Mersey Store

Question	Answer
2	$Q_1: 14,919; Q_2: 15,750; Q_3: 16,544; Q_4: 15,021$

## 15. Solutions – swimwear

Question	Answer
1	There is a seasonal variation and sales are increasing over time
2	$\hat{y} = 6,830.7796 + 364,7968x$
3	Ratios for 2005 are winter 0.2587; spring 1.9720; summer 1.66773; fall 0.1083. This means that for 2005 winter sales are 74.13% below the yearly average, spring 97.20% above, summer 66.73% above and winter 89.12% below
4	Winter 0.2569; spring 1.9627; summer 1.6686; fall 0.1120
5	Winter 3,302; spring 25,905; summer 22,598; fall 1,555 units
6	Unit sales are more appropriate for comparison. If dollar sales are used this masks the impact of price increases over time

## Chapter 11: Indexing as a Method of Data Analysis

### 1. Solutions – backlog

Question	Answer
1	1988 = 100; 126; 144; 168; 221; 222; 211; 221; 237; 216; 190; 137; 2000 = 150; 173; 146; 159; 222; 224
2	In 1989 the backlog was 26% more than in 1988; in 2000 it was 50% more than in 1988; in 2005 it was 124% more than in 1988
3	1988 = 67; 84; 96; 112; 147; 148; 140; 147; 158; 144; 126; 91; 2000 = 100; 115; 97; 106; 148; 149
4	In 1989 the backlog was 33% less than in 2000; in 1993 it was 48% more than in 2000; in 2005 it was 49% more than in 2000
5	1988 is a long way back and with a base of 1988 gives some index values greater than 200 which is sometimes difficult to interpret, 2000 is closer to the present time and all the index values are less than 200 making interpretation more understandable
6	1989 = 126; 114; 117; 132; 100; 95; 105; 107; 91; 88; 72; 2000 = 109; 115; 84; 109; 139; 101
7	In 1990 the backlog increased 14% over 1989. In 1994 the backlog decreased by 5% compared to 1993. In 1998 the backlog decreased by 12% compared to 1997. In 2004 the backlog increased by 39% compared to 2003

### 2. Solutions – gold

Question	Answer
1	1987 = 100; 98; 85; 86; 81; 77; 81; 86; 86; 1996 = 87; 74; 66; 63; 63; 61; 70; 82; 92; 116
2	In 1996 the price was 13% less than in 1987. In 2001 it was 39% less than in 1987. In 2005 it was 16% more than in 1987
3	1988 = 115; 113; 98; 99; 93; 89; 93; 99; 99; 1996 = 100; 85; 76; 72; 72; 70; 80; 94; 106; 133
4	In 1987 the price was 15% more than in 1996. In 2001 it was 30% less than in 1996. In 2005 it was 33% more than in 1996
5	1987 is a long way back. A base of 1996 gives index values for the last 10 years
6	1988 = 98; 87; 101; 94; 95; 105; 107; 100; 101; 85; 89; 95; 100; 97; 114; 117; 113; 126
7	1997 when the price decreased by 15% from the previous year
8	2005 when the price increased by 26% from the previous year

## 3. Solutions – US gasoline prices

Question	Answer
1	1990 = 100.0; 94.4; 94.1; 89.2; 97.5; 94.1; 100.8; 102.3; 84.3; 101.8; 2000 = 119.2; 113.1; 114.6; 142.1; 155.7; 211.5; 2006 = 245.8
2	In 1993 the price was 10.8% less than in 1990. In 1998 it was 15.7% less than in 1990. In 2005 it was 111.5% more than in 1990 or more than double the price
3	1990 = 83.9; 79.2; 78.9; 74.8; 81.8; 78.9; 84.6; 85.8; 70.7; 85.4; 2000 = 100.0; 94.9; 96.1; 119.2; 130.6; 177.4; 2006 = 206.2
4	In 1993 the price was 15.2% less than in 2000. In 1998 it was 29.3% less than in 2000. In 2005 it was 77.4% more than in 2000
5	1990 is in the last century. A base of 2000 gives the benchmark start for the 21st century
6	1991 = 94.4; 99.7; 94.7; 109.3; 96.6; 107.1; 101.4; 82.4; 120.7; 2000 = 117.2; 94.9; 101.3; 124.0; 109.5; 135.9; 2006 = 116.2
7	In 2005 when there was a 35.9% increase over the previous year
8	1990 = 100.0; 190.0; 100.0; 95.0; 90.0; 95.0; 100.0; 110.0; 95.0; 60.0; 2000 = 75.0; 150.0; 125.0; 125.0; 135.0; 175.0; 310.0
10	There is an approximate correlation between the two in that when crude oil prices rise, then the price of refined gasoline rises. This is to be expected. However, there is always a lag when gasoline prices change to correspond to crude oil price changes

## 4. Solutions – coffee prices

Question	Answer
1	1975 = 100; 138.4; 306.6; 245.9; 297.7; 307.2; 244.5; 223.1; 221.9; 212.5; 1985 = 280.5; 293.3; 335.2; 334.8; 312.2; 1990 = 340.0; 324.1; 265.2; 248.5; 386.9; 407.2; 417.4; 476.2; 448.8; 406.9; 2000 = 374.6; 359.0; 386.9; 431.8; 2004 = 465.1
2	In 1985 the price was 180.5% more than in 1975. In 1995 it was 307.2% more than in 1975. In 2004 it was 365.1% more than in 1975
3	1975 = 29.4; 40.7; 90.2; 72.3; 87.6; 90.4; 71.9; 65.6; 65.3; 62.5; 82.5; 86.3; 98.6; 98.5; 91.8; 1990 = 100.0; 95.3; 78.0; 73.1; 113.8; 119.8; 122.8; 140.1; 132.0; 119.7; 2000 = 110.2; 105.6; 113.8; 127.0; 136.8
4	In 1985 the price was 17.5% less than in 1990. In 1995 it was 19.8% more than in 1990. In 2004 it was 36.8% more than in 1990
5	1975 = 26.7; 37.0; 81.8; 65.6; 79.5; 82.0; 65.3; 59.6; 59.2; 56.7; 1985 = 74.9; 78.3; 89.5; 89.4; 83.3; 90.8; 86.5; 70.8; 66.3; 103.3; 1995 = 108.7; 111.4; 127.1; 119.8; 108.6; 2000 = 100.0; 95.8; 103.3; 115.3; 124.2
6	In 1985 the price was 25.1% less than in 2000. In 1995 it was 8.7% more than in 2000. In 2004 it was 24.2% more than in 2000

(Continued)

## 4. Solutions – (Continued)

Question	Answer
7	A base of 2000 is probably the best as it starts the 21st century. A base of 1975 is inappropriate as it is too far back and we are introducing percentage values more than 100% which is not always easy to grasp
8	1976 = 138.4; 221.5; 80.2; 121.0; 103.2; 79.6; 91.3; 99.4; 95.8; 1985 = 132.0; 104.6; 114.3; 99.9; 93.2; 108.9; 95.3; 81.8; 93.7; 155.7; 1995 = 105.3; 102.5; 114.1; 94.3; 90.7; 92.1; 95.8; 107.8; 111.6; 107.7
9	In 1977 when prices increased by 121.5%
10	In 1981 when prices decreased by 20.4%
11	Coffee beans are a commodity and the price is dependent on the weather and the quantity harvested. Also country unrest can have an impact such as in the Ivory Coast in 2005

## 5. Solutions – Boeing

Question	Answer
1	100; 96; 92; 98; 106
2	They were 6% higher than in 2005
3	94; 66; 25; 84; 100
4	They were 6% less than in 2001
5	105; 104; 93; 93
6	There was a 7% annual decline in both 2002 and 2003 and a 4% annual increase in 2004 and a 5% annual increase in 2005

## 6. Solutions – Ford Motor Company

Question	Answer
1	1992 = 100.0; 108.5; 126.9; 130.9; 138.5; 144.5; 139.8; 2000 = 160.0; 166.8; 155.0; 159.3; 163.8; 174.3; 2004 = 181.9
2	They were 81.9% higher than in 1992
3	1993 = 108.5; 117.0; 103.1; 105.8; 104.4; 96.8; 114.4; 2000 = 104.3; 92.9; 102.8; 102.8; 106.4; 2005 = 104.3
4	1998 by 3.2% and 2001 by 7.1%
5	1992 = 100.0; 111.9; 124.3; 115.9; 114.3; 120.0; 118.3; 129.6; 2000 = 133.6; 116.2; 119.2; 108.9; 2004 = 106.0
6	2000 when 33.6% more vehicles were sold than in 1992
7	Not that great; stock price has fallen; dividends are half what they were in 1992; North American vehicle sales look to be on the decline

## 7. Solutions – drinking

Question	Answer
1	Britain = 100; 113; 70; 13; 43; 13; 113; 30; 13; Sweden = 39
2	In Ireland 13% more people than in Britain admit to have been drunk. In Greece the figure is 87% less than in Britain and in Germany it is 57% less
3	Britain = 767; 867; 533; France = 100; 333; 100; 867; 233; 100; Sweden = 300
4	In Ireland 767% more people than in France admit to have been drunk (almost 8 times as many). In Greece the figure is about the same as in France and in Germany it is 233% more or over 3 times as much
5	Britain = 88; Denmark = 100; 62; 12; 38; 12; 100; 27; 12; Sweden = 35
6	In Ireland the drinking level is about the same as in Denmark. In Greece it is about 88% less than in Denmark and in Germany it is 62% less than in Denmark
7	Southern Europeans seem to admit being drunk less than Northern Europeans. In Southern Europe drinking wine with a meal is common and this perhaps suppresses the notion for "binge" drinking

## 8. Solutions – part-time work

Question	Answer
1	Australia = 208; 123; 137; 188; 138; 136; 88; 108; 169; 46; Ireland = 138; 115; 200; 277; 169; 162; 77; 92; 112; 196; 42; United States = 100
2	In Australia there is 108% more people working part time than in the United States. In Greece there is 54% less people working part time. In Switzerland there is 96% more people working part time than in the United States
3	Australia = 75; 44; 49; 68; 50; 49; 32; 39; 61; 17; 50; 42; 72; Netherlands = 100; 61; 58; 28; 33; 40; 71; 15; United States = 36
4	In Australia there is 15% less people working part time than in the Netherlands. In Greece there is 39% less people working part time. In Switzerland there is 29% less people working part time than in the Netherlands
5	Australia = 88; 108; 105; Britain = 100; 89; 83; 82; 102; 105; 90; 102; 101; 88; 99; 97; 97; 88; 101; 90; 107; 77; 88
6	In Australia the proportion of women working part time is 12% less than in Britain. In Greece the proportion of women working part time is 10% less than in Britain. In Switzerland the proportion of women working part time is 7% less than in Britain

## 9. Solutions – cost of living

Question	Answer
1	Amsterdam 46% less; Berlin 58% less; New York 18% more; Paris 23% less; Sydney 35% less; Tokyo 38% more; Vancouver 53% less

(Continued)

## 9. Solutions – (Continued)

Question	Answer
2	Amsterdam 4% more; Berlin 19% more; New York 14% less; Paris 9% less; Sydney 16% less; Tokyo 13% less; Vancouver 10% less
3	Amsterdam 23% more; Berlin 6% less; New York 165% more; Paris 73% more; Sydney 46% more; Tokyo 212% more; Vancouver 16% less
4	Amsterdam 12% more; Berlin 6% less; New York 14% less; Paris 9% less; Sydney 16% less; Tokyo 13% less; Vancouver 16% less
5	Amsterdam 10% more; Berlin 8% less; New York 16% less; Paris 11% less; Sydney 18% less; Tokyo 14% less; Vancouver 18% less
6	Amsterdam 14% more; Berlin 4% less; New York 12% less; Paris 7% less; Sydney 15% less; Tokyo 11% less; Vancouver 15% less
7	Most expensive is Tokyo which is 425% more than the least expensive or over 5 times more expensive

## 10. Solutions – corruption

Question	Answer
1	Iceland is the least corrupt; the countries of Czech Republic, Greece, Slovakia, and Namibia are equally the most
2	78%
3	Spain 100.0; Denmark 135.7; Finland 137.1; Germany 117.1; United Kingdom 122.9
4	Italy 100.0; Denmark 190.0; Finland 192.0; Germany 164.0; United Kingdom 172.0
5	Greece 100.0; Denmark 220.9; Finland 223.3; Germany 190.7; United Kingdom 200.0
6	Portugal 100.0; Denmark 146.2; Finland 147.7; Germany 126.2; United Kingdom 132.3
7	Of the selected eight countries from the European Union those in North appear to be less corrupt than in the South

## 11. Solutions – road traffic deaths

Question	Answer
1	Mauritius is most dangerous and France is the least
2	Belgium is 220% more dangerous; Dominican Republic 680% more; France 20% less; Latvia 400% more; Luxembourg 220% more; Mauritius 800% more; Russia 300% more; Venezuela 380% more
3	Belgium is 6.7% more dangerous; Dominican Republic 160% more; France 73.3% less; Latvia 66.7% more; Luxembourg 13.3% more; Mauritius 200% more; Russia 33.3% more; Venezuela 60% more

(Continued)

## 11. Solutions – (Continued)

Question	Answer
4	Belgium is 23.8% less dangerous; Dominican Republic 85.7% more; France 81% less; Latvia 19% more; Luxembourg 29% less; Mauritius 114.3% more; Russia 4.8% less; Venezuela 14.3% more
5	Belgium is 23.1% more dangerous; Dominican Republic 200% more; France 69.2% less; Latvia 92.3% more; Luxembourg 30.8% more; Mauritius 246.2% more; Russia 53.8% more; Venezuela 84.6% more
6	Emerging countries are more dangerous to drive than developed countries. Improved road conditions, better driving education, better traffic control, and severer penalties when traffic laws are infringed

## 12. Solutions – family food consumption

Question	Answer
1	2003 = 100.0; 2004 = 110.5
2	2003 = 100.0; 2004 = 141.2
3	2003 = 100.0; 2004 = 109.4
4	2003 = 100.0; 2004 = 109.2
5	2003 = 100.0; 2004 = 109.3
6	The unweighted price index measures inflation. The unweighted quantity index measures consumption changes for this family. The other indexes take into account price and quantity. The values are close as although the numerator changes, the denominator changes by almost proportionality the same amount

## 13. Solutions – meat

Question	Answer
1	2000 = 100.0; 2001 = 100.8; 2002 = 100.4; 2003 = 116.3; 2004 = 133.3; 2005 = 138.1
2	2000 = 89.6; 2001 = 87.2; 2002 = 74.8; 2003 = 86.3; 2004 = 112.4; 2005 = 100.0
3	2000 = 71.3; 2001 = 70.9; 2002 = 70.9; 2003 = 83.1; 2004 = 95.4; 2005 = 100.0
4	2000 = 87.1; 2001 = 86.6; 2002 = 86.5; 2003 = 101.4; 2004 = 116.4; 2005 = 122.0
5	The meat prices show a reasonable trend during the period and there is an increasing trend in the quantity of meat handled. This explains a reasonable trend of the indexes



## 14. Solutions – beverages

Question	Answer
1	2000 = 100.0; 2001 = 85.1; 2002 = 100.1; 2003 = 102.9; 2004 = 103.5; 2005 = 114.6
2	2000 = 82.1; 2001 = 75.4; 2002 = 121.0; 2003 = 123.2; 2004 = 108.6; 2005 = 100.0
3	2000 = 87.5; 2001 = 72.4; 2002 = 81.1; 2003 = 84.1; 2004 = 87.0; 2005 = 100.0
4	2000 = 102.5; 2001 = 84.8; 2002 = 95.1; 2003 = 98.6; 2004 = 101.9; 2005 = 117.1
5	The commodity prices show no trend during the period but there is an increasing trend of the consumption of the commodity. This explains the fluctuation of the indexes

## 15. Solutions – non-ferrous metals

Question	Answer
1	2000 = 100.0; 2001 = 84.9; 2002 = 78.4; 2003 = 96.9; 2004 = 129.9; 2005 = 146.5
2	2000 = 116.0; 2001 = 104.5; 2002 = 71.1; 2003 = 90.0; 2004 = 80.2; 2005 = 100.0
3	2000 = 64.8; 2001 = 56.8; 2002 = 52.6; 2003 = 63.5; 2004 = 82.9; 2005 = 100.0
4	2000 = 92.5; 2001 = 81.0; 2002 = 75.1; 2003 = 90.5; 2004 = 118.3; 2005 = 142.6
5	There is neither a trend in the commodity price or the usage of the metals during the 6-year period. This helps to explain the fluctuation of the indexes

# Bibliography

There are many books related to statistics, and numerous sources of statistical information. The following are some that I have used.

- Alreck, P.L. and Settle, R. (2003). *Survey Research Handbook*, 3rd edition, McGraw Hill, New York.
- Barnes, S. (Ed.) (2005). *News of the World, Football Annual 2005–2006*, Invincible Press, London, ISBN 0-00-720582-1.
- Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). *Basic Business Statistics*, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-13-196869-6.
- Bernstein, P.L. (1998). *Against the Gods: The Remarkable Story of Risk*, Wiley, New York, USA, ISBN 0-471-12104-5.
- Buglear, J. (2002). *Stats Means Business: A Guide to Business Statistics*, Butterworth-Heinemann, Oxford, ISBN 0-7506-5364-7.
- Economist (The)*, Editorial offices in major cities. First published in 1843. Weekly publication, [www.economist.com](http://www.economist.com)
- Eurostat*, Statistical information for Europe and elsewhere, <http://epp.eurostat.ec.europa.eu>
- Financial Times*, Printed in London and other world cities. Daily newspaper published Monday through Friday on a distinguishing salmon coloured paper, [www.ft.com](http://www.ft.com)
- Fortune*, Published monthly in the Netherlands by Time Warner, [www.fortune.com](http://www.fortune.com)
- Francis, A. (2004). *Business Mathematics and Statistics*, 6th edition, Thomson, London, ISBN 1-84480-128-4.
- International Herald Tribune*, Published by the New York Times, edited in Paris and Hong Kong and printed in Paris, France. Daily newspaper published Monday through Saturday, [www.iht.com](http://www.iht.com)
- Levin, R.I. and Rubin, D.S. (1998). *Statistics for Management*, 7th edition, Prentice Hall, New Jersey, ISBN 0-13-606716-6.
- Organization for Economic and Development, [www.oecd.org](http://www.oecd.org)
- Tufte, E.R. (1987). *The Visual Display of Quantitative Information*, 2nd edition (May 2001), Graphics Press, Connecticut, USA, ISBN 0-961-39214-2.
- United Kingdom Government Statistics. Statistical information on many aspects of the UK, [www.statistics.gov.uk](http://www.statistics.gov.uk)
- United Nations Conference on Trade and Development, [www.unctad.org](http://www.unctad.org)
- United States Government Statistics. Statistics produced by more than 70 agencies of the United States Federal Government, <http://fedstats.gov>
- Wall Street Journal*, Europe, Printed in Belgium. Daily newspaper published Monday through Friday, [www.wsj.com](http://www.wsj.com)
- Waller, D.L. (2003). *Operations Management: A Supply Chain Approach*, 2nd edition, Thomson, Learning, London, ISBN 1-86152-803-5.

*This page intentionally left blank*

# Index

- 80/20 rule 17
- a priori* probability, concepts 84–6, 99, 104
- ABS function, Excel 433
- absolute frequency histograms, concepts 5–6, 16, 17, 23
- absolute frequency ogives 9–11
- absolute frequency polygons 7–9, 23
- acceptance/rejection issues, hypothesis testing 265–9, 276–80, 305–20
- addition rules
  - classical probability 84–5, 91–2, 103
  - equation 84
- alcohol 348
- algebraic expressions, concepts 439
- alternative hypothesis
  - see also* hypothesis testing
  - concepts 265–80, 305–20
- answers to end of chapter exercises 449–508
- appendices 413–508
  - answers to end of chapter exercises 449–508
  - Excel usage guide 429–36
  - key terminology/formulas in statistics 413–26, 427, 438, 446–7
  - mathematical relationships 437–47
- arithmetic
  - operating symbols list 438
  - rules for calculations 444–5
  - sequence of operations 438
- arithmetic mean
  - see also* mean...
  - coefficient of variation 56–7, 63, 361–2, 365
  - concepts 47–8, 54–7, 59–60, 63, 120–1, 151–73, 187–212, 231–43, 301–20, 444–6
  - definition 47, 120–1, 231
  - deviation about the mean 55–6
  - different populations hypothesis testing 301–32
  - distribution of the sample means 190, 194–6, 211
  - equation 47, 444–5
  - estimates 231–43
  - examples 47–8, 161–2, 164–9
  - hypothesis testing 265–72, 279–80, 302–20
  - median 59–60, 151–2, 161–2, 164–9, 172–3
  - normality of data 161–9, 172–3
  - sampling for the mean 187–212, 231–52
- arrangement of different objects (rule no. 3), counting rules 101, 104
- assembly operation application, system reliability and probability 97–9
- asymmetrical data, concepts 60, 164–9, 172–3
- audits 207, 231, 243–5, 251–2
  - concepts 243–5
  - estimates of population characteristics 231, 243–5, 251–2
  - finite populations 244–5
  - infinite populations 243–4
- automobile accidents 387–9
- AVERAGE function, Excel 47, 240, 244, 271–2, 433
- average quantity-weighted price index
  - see also* indexing
  - concepts 395–7
- average value *see* arithmetic mean
- backup systems, system reliability and probability 93, 95–9, 103
- bar charts 5, 16–19, 24
  - see also* horizontal...; parallel...
- bars of chocolate 149–50, 154–5, 161, 194–5
- base factors, indexing 385–8, 391–2, 397
- basic probability rules
  - see also* probability
  - addition rules 84–5, 91–2, 103
  - Bayesian decision-making (Bayes' Theorem) 88–9, 103
  - bottling machine application 92
  - classical probability 83–6
  - concepts 81–93, 103–4
  - events 82–5
  - exercises and answers 105–17, 458–61
  - hospitality management application 89–92
  - joint probability 84, 85–6, 91–2, 94–9, 103
  - mutually exclusive events 84–5, 89–92, 99–100
  - non-mutually exclusive events 84–5, 89–92
  - odds 93, 103
  - relative frequency probability 83, 100–1, 103
  - risk concepts 82, 92–3, 103
  - statistical dependence 86–8, 103
  - statistical independence 84–8
  - subjective probability 82–3, 103
  - Venn diagrams 89–92, 103
- Bayesian decision-making (Bayes' Theorem), concepts 88–9
- beer cans 149–50, 154–5, 161, 269–70, 274–5
- Bernoulli, Jacques 128, 134
  - see also* binomial distribution
- bi-modal datasets 52

- bias
  - concepts 206–7, 212, 231–2
  - definition 206
  - point estimates 231–2, 245, 251–2
  - samples 55, 206–7, 212, 231–2
- bibliography 509
- Big Mac index 45–6
- binary numbering system, concepts 446–7
- BINOMDIST function, Excel 129, 134, 169, 434
- binomial distribution
  - applications 129–30, 169–73, 203–6, 211, 272–4
  - concepts 120, 125, 127–30, 132–5, 136–47, 169–73, 203–6, 211, 272–4
  - continuity correction factor 170–1
  - definition 127, 134
  - equation 128, 203–6
  - exercises and answers 136–47, 462–5
  - expected values 128
  - normal distribution 150, 169–73, 203–6, 211, 272–4
  - normal–binomial approximation 169–73, 204–6, 211
  - Poisson distribution 132–4, 135
  - sampling distribution of the proportion 203–6, 211
  - standard deviation 128–9, 134, 169–73, 203–6, 211
  - validity conditions 127–8, 129–30, 134–5
  - validity deviations 129–30
  - variance 128–9
- binomial function
  - concepts 82, 94–5, 119, 120, 125, 127–30, 134–5, 136–47, 169–73, 203–6, 211, 265–7, 272–4, 279
  - mathematical expression 128
- bivariate data
  - see also* numerical data; time series
  - concepts 3, 5, 12–15, 23, 335–9
  - definition 3, 23, 335
  - line graphs 12–15, 17, 20
- bonds 124–6
- bottling machine application, basic probability rules 92
- boundary limits of quartiles
  - see also* quartiles
  - concepts 57–8
- box and whisker plots 12, 59–60, 63–4, 161, 164–6, 172
  - concepts 59–60, 63–4, 161, 164–6, 172
  - definition 59
  - EDA 12, 60
  - examples 59–60, 161, 164–6
- casinos 79–80
- categorical data
  - concepts 3, 15–24
  - cross-classification (contingency) tables 20, 21, 24, 86–8, 92–3
  - definition 15, 23–4
  - exercises and answers 25–43, 449–52
  - horizontal bar charts 16, 19, 24
  - overview 3
  - parallel bar charts 16–17, 19, 24
  - parallel histograms 16, 18
  - pareto diagrams 17, 20, 24
  - pictograms 22–3, 24
  - pie charts 15–16, 24
  - questionnaires 15, 24
  - stacked histograms 20–2, 24
  - vertical histograms 16–18, 24
- categories *see* classes
- causal forecasting, linear regression 345–7, 365, 388
- CEILING function, Excel 433
- central limit theorem, concepts 190, 194, 211, 232–3, 269–72, 304
- central tendency of data
  - arithmetic mean 47–8, 54–7, 59–60, 63, 269–72
  - concepts 47–53, 63, 190, 194, 211, 232–3, 269–72, 304
  - definition 47
  - exercises and answers 65–77, 453–7
  - geometric mean 52–3, 63
  - median 49–51, 57–60, 63, 151–2, 161–2, 164–9, 172–3
  - midrange 52, 58, 63, 151–2
  - mode 51–2, 63, 151–2, 165
  - weighted average 48–9, 63
- ceramic plates 169–71
- characteristic probability, concepts 99–100, 128
- characterizing data
  - concepts 45–77
  - exercises and answers 65–77, 453–7
- chi-square hypothesis test 302, 313–32, 447
  - concepts 313–20, 447
  - cross-classification (contingency) tables 313–20
  - definition 313–14
  - degrees of freedom 314–19
  - Excel 317–19
  - p*-value approach 318, 320
  - significance levels 319, 320
  - test of independence 315–16
  - value-determination calculations 316–17, 320
  - work preferences distribution 317–19
- CHIDIST function, Excel 317–18, 434
- CHIINV function, Excel 317–19, 434
- CHITEST function, Excel 317–18, 434
- circuit boards 246–8
- class range/width, frequency distributions 4–5
- classes, frequency distributions 3–5
- classical probability
  - addition rules 84–5, 91–2, 103
  - concepts 83–6, 103
  - definition 83
  - equation 83
  - joint probability 84, 85–6, 91–2, 94–9, 103
  - mutually exclusive events 84–5, 89–92, 99–100
  - non-mutually exclusive events 84–5, 89–92
- closed-ended frequency distributions

- see also* frequency distributions
- concepts 5, 7
- cluster sampling
  - concepts 209, 212
  - definition 209
- coefficient of correlation
  - see also* correlation
  - concepts 338–9, 345–7, 364
  - definition 338
  - equation 338
- coefficient of determination
  - concepts 338–9, 343, 345–7, 351–2, 359–60, 363, 364–5
  - definition 339
  - equation 339
- coefficient of multiple determination
  - see also* multiple regression
  - concepts 348–51
- coefficient of variation
  - concepts 56–7, 63, 361–2, 365
  - definition 56
  - equation 57
  - examples 57
  - forecasting considerations 361–2, 365
- coffee samples 236
- coin-tossing experiments, counting rules 99–100, 125–8, 171
- collected-data considerations, forecasts 360–1, 365
- collectively exhausting outcomes 99–100
- COMBIN function, Excel 102, 434
- combinations 100–1, 102, 104, 188
  - concepts 102, 104, 188
  - definition 102
- combinations of objects (rule no. 5), counting rules 102, 104
- commuting times 311–13
- computer systems, backups 96–7
- conditional probabilities
  - equation 88
  - statistical dependence 86–8, 103
- confidence intervals
  - applications 234–5, 236–7, 243–5
  - concepts 231, 232–52, 343–4
  - definition 233–4
  - examples 233–7, 243–5
  - exercises and answers 253–62, 475–9
  - finite populations 231, 236–7, 244–5, 251–2
  - infinite populations 233–6, 243–5, 251–2
  - student-*t* distribution 231, 238–43, 251–2, 344, 350–1
- confidence levels
  - concepts 231, 232–52, 343–4, 350–1
  - definition 232–3
  - exercises and answers 253–62, 475–9
  - forecasts 343–4, 364
  - margin of error 248–52
  - regression analysis 343–4, 364
- constants, concepts 339, 437, 445–6
- consumer price index (CPI)
  - see also* indexing
  - concepts 386, 388–91, 397
- consumer surveys
  - concepts 209–10, 212
  - definition 209
  - examples 209–10
- contingency tables *see* cross-classification (contingency) tables
- continuity correction factor, normal–binomial approximation 170–1
- continuous random variables
  - concepts 150–73
  - definition 150
- control-shift-enter keys 5
- CORREL function, Excel 338, 434
- correlation 335–81
  - see also* forecasts
  - causal forecasting 345–7, 388
  - concepts 335–9, 345–7, 364–5
  - definition 335, 338
  - equation 338
  - Excel 338
  - good correlation value 339
- costs of errors, hypothesis testing 277–80
- COUNT function, Excel 434
- counting rules
  - arrangement of different objects (rule no. 3) 101, 104
  - combinations of objects (rule no. 5) 102, 104
  - concepts 81, 99–102, 104
  - definition 99
  - different types of events (rule no. 2) 100–1, 104
  - examples 100–2
  - exercises and answers 105–17, 458–61
  - permutations of objects (rule no. 4) 101–2, 104
  - single type of event (rule no. 1) 99–100, 104
- covariance
  - concepts 120, 124–6, 134–5
  - definition 124
  - equation 124
  - examples 124–6
  - portfolio risk 124–6
- critical levels
  - see also* significance levels
  - concepts 264–80, 302–20
- cross-classification (contingency) tables
  - chi-square hypothesis test 313–20
  - concepts 20, 21, 24, 86–8, 92–3, 313–20
  - definition 20
  - degrees of freedom 314–15
  - examples 21, 87, 93, 313–14
- cumulative distributions *see* ogives
- currencies 45–6
- curvilinear functions 351–3, 362–5
  - see also* non-linear regression

## data

*see also* categorical...; numerical...

asymmetrical data 60, 164–9, 172–3

characterizing data 45–77

concepts 3–24

defining data 45–77

indexing 383–412

normality of data 161–9, 172–3

presenting data 1–43, 449–52

primary/secondary data 210, 212

data arrays, concepts 50

Data menu bar 12

## decimals

concepts 439–42, 444, 446–7

fractions 444

## defining data

concepts 45–77

exercises and answers 65–77, 453–7

## degrees of freedom

chi-square hypothesis test 314–19

cross-classification (contingency) tables 314–15

student-*t* distribution 237–43, 251, 269–72, 306–11, 344, 347

dependent populations, hypothesis testing 309–11, 319–20

dependent variables, definition 336, 345

## derivatives 364

descriptive statistics, definition 188

development of percentiles 61–2

*see also* percentiles

deviation about the mean, concepts 55–6

Dewey, Governor 185–6

dice-tossing experiments 100, 104, 120, 121–2

## different populations

exercises and answers 321–32, 489–95

hypothesis testing 301–32

different types of events (rule no. 2), counting rules 100–1, 104

## discrete data

binomial concepts 82, 94–5, 119, 120, 125, 127–30, 132–5, 136–47, 169–73

concepts 120–35, 169–73

definition 120

distribution for random variables 120–7, 134–5, 136–47

exercises and answers 136–47, 462–5

Poisson distribution 120, 130–5, 136–47

probability analysis 119–47, 169–73

discrete random variables, concepts 120–35

dispersion of data 54–7, 63, 121–4, 128–9, 131, 134–5, 151–73

*see also* standard deviation

coefficient of variation 56–7, 63, 361–2, 365

concepts 47, 53–7, 63

definition 53

deviation about the mean 55–6

exercises and answers 65–77, 453–7

range 53–4, 58

variance 54–7, 63, 121, 124, 128–9

## distribution for discrete random variables

application of the random variable 121–4

characteristics 120–1, 134–5

concepts 120–7, 134–5, 136–47

covariance 120, 124–6, 134–5

exercises and answers 136–47, 462–5

expected values 121–4, 125–7, 134–5

## distribution of the sample means

concepts 190, 194–6, 211

sample size 194–6

shape factors 194–5

division of data, percentiles 62

Dow Jones Industrial Average 187

drugs 277–8

e-mail surveys 209–10, 212

EDA *see* exploratory data analysis

80/20 rule 17

empirical probability *see* relative frequency probability

empirical rule, normal distribution 152, 172–3

equal reliability concepts, parallel (backup) systems 96–9

equations, definition 438

## errors

hypothesis testing 276–80

Type I/Type II errors 276–80

## estimates of population characteristics

*see also* forecasts

audits 231, 243–5, 251–2

concepts 229–52

confidence levels 231, 232–62, 343–4, 364

exercises and answers 253–62, 475–9

interval estimates 231, 232, 245–6, 251–2

margin of error 229, 231, 247–52

mean values 231–43

point estimates 231–2, 245, 251–2

population characteristics 229–62

proportions 231, 245–8, 251–2, 272–4, 311–13

sample sizes 231, 235–40, 245–8, 251–2, 272–4, 280, 303–20, 344

student-*t* distribution 231, 237–45, 251–2, 269–72, 306–9, 344, 347, 350–1

ethics 127, 149

European Commission 301

European Union 229–30, 249–50, 301

## events

concepts 82–5, 99–102, 103–4

counting rules 81, 99–102, 104

mutually exclusive events 84–5, 89–92, 99–100

non-mutually exclusive events 84–5, 89–92

Excel 4–5, 11–12, 16, 47–8, 51–3, 60–1, 99, 101–2,

129, 131, 154, 156–7, 161–4, 169–71, 197,

198–9, 201–5, 235, 237, 240–8, 251–2, 270–5,

305–13, 317–19, 336, 344, 347, 350–1, 429–36

*see also* individual functions

- chi-square hypothesis test 317–19
- correlation 338
- forecasts 341–3, 347, 349–53, 359–60
- graphs 336, 429–32
- linear regression 340–3, 432, 436
- list of function 433–5
- multiple regression 436
- random sampling 207, 212
- regression analysis 340–3, 347, 348, 349–53, 359–60, 432, 436
- scatter diagrams 336–7
- student-*t* distribution 240–3, 244–5, 251–2, 308–11, 344, 347, 350–1
- usage guide 429–36
- exercises
  - answers to end of chapter exercises 449–508
  - binomial distribution 136–47, 462–5
  - characterizing/defining data 65–77, 453–7
  - counting rules 105–17, 458–61
  - discrete data 136–47, 462–5
  - distribution for discrete random variables 136–47, 462–5
  - estimates of population characteristics 253–62, 475–9
  - forecasts 366–81, 496–501
  - hypothesis testing 281–99, 321–32, 480–8, 489–95
  - indexing 398–412, 502–8
  - normal distribution 174–83, 466–9
  - Poisson distribution 136–47, 462–5
  - presenting data 25–43, 449–52
  - probability 105–17, 136–47, 174–83, 213–27, 458–74
  - regression analysis 366–81, 496–501
  - samples 213–27, 253–62, 281–99, 321–32, 470–9, 480–8, 489–95
  - system reliability and probability 105–17, 458–61
- expected values
  - binomial distribution 128
  - concepts 121–4, 125–7, 128, 134–5
  - definition 121
  - equation 121, 125, 128
  - examples 121–3, 125–7
  - law of averages 125–7
  - portfolio risk 125
  - two random variables 124, 125
- exploratory data analysis (EDA)
  - box and whisker plots 12, 60
  - concepts 12, 23, 60
  - stem-and-leaf displays 12, 23
- EXPONDIST function, Excel 433
- exponential function 334, 353, 362–5, 447
  - see also* non-linear regression
  - concepts 353, 362–3, 365
  - definition 353
- face-to-face consumer surveys 209–10, 212
- FACT function, Excel 101, 434
- factorial rule, concepts 101, 104
- Fenwick trolleys 132–4
- filling machines 269–70, 274–6
- finite population multiplier, concepts 201–3, 211, 244–5
- finite populations
  - audit estimates 244–5
  - concepts 199–202, 211, 236–7, 244–5
  - confidence intervals 231, 236–7, 244–5, 251–2
  - definition 199–200
  - samples 199–202, 211, 236–7, 244–5
  - standard error 200–3, 211, 236–7, 244–5, 251–2
- fixed base indexes 385–7, 397
  - see also* indexing
- fixed weight aggregate price index, concepts 396–7
- FLOOR function, Excel 434
- FORECAST function, Excel 341–2, 347, 359–60, 434
- forecasts 333–81
  - see also* estimates...; regression analysis
  - applications 336–7, 341–2, 345–7, 348–51, 354–60, 362–3
  - causal forecasting 345–7, 365, 388
  - coefficient of variation 361–2, 365
  - collected-data considerations 360–1, 365
  - concepts 333–65
  - confidence levels 343–4, 364
  - considerations 360–5
  - correlation 335–9, 345–7, 364–5
  - Excel 341–3, 347, 349–53, 359–60
  - exercises and answers 366–81, 496–501
  - exponential function 334, 353, 362–5
  - linear regression 339–47, 353–60, 364–5, 436
  - market-change considerations 362
  - model considerations 362–4
  - moving averages 354–60
  - multiple regression 347–51, 365, 436
  - non-linear regression 351–3, 362–5
  - political forecasts 364
  - polynomial function 352–3
  - scatter diagrams 335–7, 345–6, 364
  - seasonal patterns 353–60, 365, 366–81
  - student-*t* distribution 344, 347, 350–1
  - time series 333–4, 335–44, 353–4, 361–2, 364
  - time-horizon considerations 360, 365
  - variability of estimates 342–4, 364
- foreign exchange *see* currencies
- formulas, key terminology/formulas in statistics 413–26
- fractions
  - concepts 439–40, 444
  - decimal conversions 444
- frequency distributions
  - class range/width 4–5
  - classes 3–5
  - concepts 3–15, 23, 53, 123–4, 190–4
  - definition 3
  - midpoint of class range/width 4–5, 7–8
  - sampling the mean 190–4



- FREQUENCY function, Excel 5, 434
- frequency polygons 5, 6–9, 23, 161, 164–6  
     concepts 6–9, 23, 161, 164–6  
     definition 7  
     examples 8
- gambling 79–104, 120, 125
- Gantt charts *see* horizontal bar charts
- Gauss, Karl Friedrich 150  
     *see also* normal distribution
- gender pay gaps 301, 305–7
- General Electric Company 157
- GEOMEAN function, Excel 53, 434
- geometric mean  
     concepts 52–3, 63  
     definition 52  
     equation 52  
     examples 53
- GOAL SEEK function, Excel 434
- Goldman Sachs 47
- Gossett, William 237–8
- graphs *see* line graphs
- graphs  
     concepts 7, 12–15, 17, 20, 23, 429–32  
     Excel 336, 429–32
- Greek alphabet 444, 446–7
- groups *see* classes
- Hamlet* (Shakespeare) 89
- health spas 309–11
- histograms  
     *see also* absolute...; relative...  
     concepts 5–6, 23, 123–4, 170  
     definition 5
- historical data, relative frequency probability 83
- HIV 351–2
- Hope, Sheila 240
- horizontal bar charts  
     concepts 16, 19, 24  
     definition 16  
     examples 19
- hospitality management, Venn diagram application 89–92
- house prices, surface area 345–6, 362–3
- hurricane Charlie 97
- hypothesis, definition 264–5
- hypothesis testing 240, 263–99, 301–32, 447  
     acceptance/rejection issues 265–9, 276–80, 305–20  
     alternative hypothesis 265–80, 305–20  
     applications 269–72, 273–5, 305–19  
     chi-square hypothesis test 302, 313–32  
     concepts 263–99, 301–32  
     costs of errors 277–80  
     definition 264–5  
     dependent populations 309–11, 319–20  
     different populations hypothesis testing 301–32  
     errors 276–80  
     exercises and answers 281–99, 321–32, 480–8, 489–95  
     independent populations 302–9, 319–20  
     mean values 265–72, 279–80, 302–20  
     null hypothesis 265–80, 305–20  
     one-tail, left-hand hypothesis tests 265, 267–8, 279–80, 307–9  
     one-tail, right-hand hypothesis tests 265, 266–7, 279–80, 307–11  
     p-value approach 274–6, 280, 305–6, 310–11, 318, 320–32  
     power of tests 278–80  
     proportions 272–4, 280, 311–13, 315–16, 319–20  
     risks 276–80  
     significance levels 264–80, 302–32  
     single populations hypothesis testing 263–99  
     test statistics 268–72, 279–80, 304–20  
     two populations 302–32  
     two-tail hypothesis tests 265–6, 279–80, 305–20  
     Type I/Type II errors 276–80
- IF function, Excel 434, 435
- IKEA 97
- Imperial measuring system, concepts 442–3
- imported goods, United States 333–4
- independent events, classical probability 84–6
- independent populations, hypothesis testing 302–9, 319–20
- independent variables, definition 335–6
- index base value, definition 385–6
- index numbers, definition 385
- index values, definition 385
- indexing 383–412  
     average quantity-weighted price index 395–7  
     base factors 385–8, 391–2, 397  
     comparisons 387–8, 397  
     concepts 383–97  
     CPI 386, 388–91, 397  
     definition 385  
     exercises and answers 398–412, 502–8  
     fixed base indexes 385–7, 397  
     Laspeyres weighted price index 393–4, 397  
     moving base indexes 387–8, 397  
     Paasche weighted price index 394–5, 397  
     price index number with fixed base 386, 397  
     quantity index number with fixed base 385–6, 397  
     relative regional indexes 391–2, 397, 403–12  
     relative time-based indexes 385–91, 397  
     rolling index number with moving base 387, 397  
     RVI 390–1, 397  
     time series deflation 390–1, 397  
     unweighted index numbers 392–3, 397  
     weightings 392–7
- inferential statistics 81, 187–8, 211, 229–52, 276  
     concepts 187–8, 276  
     definition 187, 211
- infinite populations

- audit estimates 243–4
- concepts 196–200, 211, 231, 233–6, 243–5, 251–2
- confidence intervals 233–6, 243–5, 251–2
- definition 196
- samples 196–200, 211, 233–6, 243–5, 251–2
- inflation 22–3, 386, 388–91, 392–3, 397
- integers
  - concepts 120, 438
  - definition 438
- inter-quartile range *see* mid-spread
- interval estimates 231, 232, 245–6, 251–2
  - concepts 232, 245–6, 251–2
  - definition 232
  - proportions 245–6
- inventory levels, estimates of population characteristics 243–4
- joint probability
  - concepts 84, 85–6, 91–2, 94–9, 103
  - equation 85
- key terminology/formulas in statistics 413–26, 427, 438, 446–7
- kiwi fruits 240–3
- KURT function, Excel 154, 434
- kurtosis 152, 153–4, 172–3
- labour costs 301, 305–7, 391–2
- lambda, Poisson distribution 131–5, 447
- laptop computers 392–4
- Las Vegas 80
- Laspeyres weighted price index
  - see also* indexing
  - concepts 393–4, 397
- law of averages
  - concepts 100, 125–7
  - definition 125
  - expected values 125–7
- least squares method
  - see also* regression...
  - concepts 339–40
  - equations 339
- left-skewed data, normal distribution 164–9
- left-tail hypothesis tests, concepts 265, 267–8, 279–80, 307–9
- leptokurtic curves, normal distribution 153–4, 172
- leukaemia 263
- license plates, counting rules 100–1
- light bulbs 157–61
- line graphs
  - bivariate data 12–15
  - concepts 7, 12–15, 17, 20, 23, 336, 429–32
  - definition 12–13
  - Excel 336, 429–32
  - pareto diagrams 17, 20, 24
- linear regression
  - see also* regression...
  - alternative approaches 344
  - application 340–2, 345–7
  - causal forecasting 345–7, 365, 388
  - concepts 339–47, 353–60, 364–5, 432, 436
  - definition 339–40
  - equation 339
  - Excel 340–3, 432, 436
  - exercises and answers 366–81, 496–501
  - seasonal patterns 353–60, 365
  - standard error 343–4, 347, 359–60, 364–5
  - time series 339–44, 353–4, 361–2, 364
  - variability of estimates 342–4, 364
- linear relationships, forecasts 333–4
- LINEST function, Excel 343, 347, 348, 349–50, 359–60, 434
- margin of error 229, 231, 247–52
  - concepts 248–52
  - confidence levels 248–52
  - definitions 248–9
  - equations 248–9
- marginal probability *see* classical probability
- market research 209
- market-change considerations, forecasts 362
- marketing life-cycle 362–3
- mathematical relationships 437–47
- MAX function, Excel 4, 434
- mean *see* arithmetic mean
- mean proportion of successes, definition 203–4, 245–6
- mean value of random data
  - concepts 120–4
  - definition 120–1
  - equation 121
- median
  - arithmetic mean 59–60, 151–2, 161–2, 164–9, 172–3
  - concepts 49–51, 57–60, 63, 151–2, 161–2, 164–9, 172–3
  - definition 49–50
  - equation 50
  - examples 50–1, 161–2, 164–9
  - normality of data 161–2, 164–9, 172–3
- MEDIAN function, Excel 51, 434
- Melcher, Jim 364
- mesokurtic curves, normal distribution 172
- metal prices 383–5
- methods, samples 206–12
- metric measuring system, concepts 442–3
- Microsoft Excel *see* Excel
- mid-hinge, properties of quartiles 58
- mid-spread, properties of quartiles 58, 151–2, 161–2, 172–3
- midpoint of class range/width, frequency distributions 4–5, 7–8
- midrange
  - concepts 52, 58, 63, 151–2
  - definition 52
  - examples 52

- MIN function, Excel 4, 434
- mobile phones 351–2
- mode
  - concepts 51–2, 63, 151–2, 165
  - definition 51
  - examples 51–2, 165–9
  - multi-modal datasets 52
- MODE function, Excel 52, 434
- model factors, forecast considerations 362–4
- Monte Carlo 80
- moving averages 354–60
- moving base indexes 387–8, 397
  - see also* indexing
- multiple regression
  - see also* regression...
  - application 348–51
  - coefficient of multiple determination 348–51
  - concepts 347–51, 365, 436
  - definition 348
  - equation 348–51
  - Excel 436
  - standard error 348–51, 364–5, 436
- murders 278
- mutually exclusive events, classical probability 84–5, 89–92, 99–100
- Nestlé 348–9
- Newcastle United Football Club 88–9
- non-integers, definition 438
- non-linear regression
  - see also* regression...
  - concepts 351–3, 362–5
  - exponential function 334, 353, 362–5
  - polynomial function 352–3
- non-mutually exclusive events, classical probability 84–5, 89–92
- normal distribution 149–83, 197–206, 238, 241–3
  - applications 157–61, 169–71, 242–3
  - asymmetrical data 164–9, 172–3
  - binomial distribution 150, 169–73, 203–6, 211, 272–4
  - characteristics 150–1, 172–3
  - concepts 149–73, 197–206, 238, 241–3
  - confidence intervals 231, 233–7, 251–2
  - continuity correction factor 170–1
  - definition 150–1
  - demonstration of normality 161–9, 172–3
  - description 150–61, 172–3
  - different means/standard deviations 152–4, 172–3
  - distribution of the sample means 190, 194–6, 211
  - empirical rule 152, 172–3
  - equation 151–2
  - Excel functions 156–7
  - exercises and answers 174–83, 466–9
  - kurtosis 153–4, 172–3
  - leptokurtic curves 153–4, 172
  - mathematical expression 151–2
  - mesokurtic curves 172
  - normal probability plots 167–8, 172–3
  - normality of data 161–9, 172–3
  - percentiles 167–9, 172–3
  - platykurtic curves 153–4, 172
  - samples 171, 173, 196–212
  - standard normal distribution 155–6, 172–3
  - student-*t* distribution 240
  - symmetrical data 160–9, 172–3, 234–5
  - transformation relationship 154–5, 172–3, 197–9, 211
  - verification of normality 161–4, 172–3
  - z values 154–73, 197–206, 238, 241–3, 247–52, 272–80, 304–5, 311–13, 319–20
- normal probability plots 167–8, 172–3
- normal–binomial approximation
  - concepts 169–73, 204–6, 211
  - continuity correction factor 170–1
  - sample size 171, 173
- NORMDIST function, Excel 156–7, 161–4, 170–1, 275, 305–6, 434
- NORMINV function, Excel 156–7, 434
- NORMSDIST function, Excel 156–7, 198–9, 201–5, 312–13, 435
- NORMSINV function, Excel 156–7, 235, 237, 243, 247–8, 270, 273–4, 305
- null hypothesis
  - see also* hypothesis testing
  - concepts 265–80, 305–20
  - definition 265
  - different populations 305–20
  - Type I/Type II errors 276–80
- numerical codes, time series 337, 344
- numerical data
  - see also* bivariate...; univariate...
  - absolute frequency histograms 5–6, 16, 17, 23
  - concepts 3–15, 23–4
  - definition 3, 23
  - exercises and answers 25–43, 449–52
  - frequency distributions 3–15, 23, 53
  - frequency polygons 6–9, 23, 161, 164–6
  - line graphs 7, 12–15, 17, 20, 23, 336, 429–32
  - ogives 9–11, 23
  - overview 3
  - relative frequency histograms 6–7, 16, 18, 23, 83, 100–1
  - stem-and-leaf displays 10–12, 23, 161
  - types 3, 23
- obesity 348
- odds, concepts 93, 103
- OECD *see* Organization for Economic Cooperation and Development
- OFFSET function, Excel 435
- ogives
  - concepts 9–11, 23, 164

- definition 9
- examples 9–11
- oil prices 363–4
- one-arm bandits 79, 86, 125
- one-tail, left-hand hypothesis tests
  - concepts 265, 267–8, 279–80, 307–9
  - examples 267–8
- one-tail, right-hand hypothesis tests
  - concepts 265, 266–7, 279–80, 307–11
  - examples 266–7, 308–9
- Organization for Economic Cooperation and Development (OECD) 204
- organizing data
  - concepts 1–43
  - exercises and answers 25–43, 449–52
- 'other' category 15, 210
- p*-value approach 274–6, 280, 305–6, 310–11, 318, 320–32
  - chi-square hypothesis test 318, 320
  - concepts 274–6, 280, 305–6, 310–11, 318, 320
  - hypothesis testing 274–6, 280, 305–6, 310–11, 318, 320
  - interpretation 276
- Paasche weighted price index
  - see also* indexing
  - concepts 394–5, 397
- paired populations *see* dependent populations
- paper 234–5
- paperback books 244–5
- parallel (backup) systems
  - assembly operation application 97–9
  - system reliability and probability 93, 95–9, 103
- parallel bar charts
  - concepts 16–17, 19, 24
  - definition 16–17
  - examples 19
- parallel histograms
  - concepts 16, 18, 24
  - definition 16
  - examples 18
- pareto diagrams
  - concepts 17, 20, 24
  - definition 20
  - examples 20
  - line graphs 17, 20, 24
- part-time workers 204–6
- Partouche 80
- pay gaps, gender issues 301, 305–7
- PEARSON function, Excel 338, 435
- percentages, concepts 444
- PERCENTILE function, Excel 61, 435
- percentiles
  - concepts 47, 60–4, 167–9, 172–3
  - definition 60–1
  - development 61–2
  - division of data 62
  - examples 61–2, 167–9
  - exercises and answers 65–77, 453–7
  - normal distribution 167–9, 172–3
  - standard deviations 167–9
- PERMUT function, Excel 102, 435
- permutations
  - concepts 101–2, 104
  - definition 101–2
- permutations of objects (rule no. 4), counting rules
  - 101–2, 104
- pharmaceutical companies 277–8
- pictograms
  - concepts 22–3, 24
  - definition 23
  - examples 22
- pie charts
  - concepts 15–16, 24
  - definition 15
  - examples 16
  - 'other' category 15
- platykurtic curves, normal distribution 153–4, 172
- point estimates, concepts 231–2, 245, 251–2
- Poisson distribution
  - applications 131–3
  - binomial relationship 132–4, 135
  - concepts 120, 130–5, 136–47, 447
  - definition 130–1, 135
  - equation 131
  - exercises and answers 136–47, 462–5
  - mathematical expression 131
  - standard deviation 131, 135
- POISSON function, Excel 131, 132, 435
- poker 79–80
- political forecasts 364
- polynomial function
  - see also* non-linear regression
  - concepts 352–3
- population standard deviation
  - see also* standard deviation
  - concepts 55–7, 240–5, 269–72
- population variance
  - see also* variance
  - concepts 54–7, 121
- populations
  - see also* estimates...
  - concepts 54–7, 63, 81, 127–8, 187–212, 229–52, 263–80
  - different populations 301–32
  - finite populations 199–202, 211, 231, 236–7, 244–5, 251–2
  - hypothesis testing of single populations 263–99, 301–32
  - infinite populations 196–200, 211, 231, 233–6, 243–5, 251–2
  - sample size 188–95, 231, 235–40, 241–3, 245–8, 251–2, 272–4, 280, 303–20, 344
  - single populations 263–99

- portfolio risk, concepts 124–6
- postal surveys 209–10, 212
- posterior probabilities
  - see also* Bayesian decision-making
  - concepts 88–9
- POWER function, Excel 3, 99, 435
- power of tests, hypothesis testing 278–80
- presenting data 1–43, 449–52
  - bad example 1–2
  - concepts 1–24
  - exercises and answers 25–43, 449–52
  - importance 3
- price index number with fixed base
  - see also* indexing
  - concepts 386, 388–91, 397
  - CPI 386, 388–91, 397
- price indexes
  - see also* indexing
  - concepts 385–97
- primary data
  - concepts 210, 212
  - definition 210
- printing 236–7, 244–5
- probability 79–117, 119–47, 169–73, 174–83, 213–27, 231–52, 274–80, 458–74
  - see also* basic probability rules
  - binomial distribution 120, 125, 127–30, 132–5, 136–47, 169–73, 203–6, 211
  - concepts 79–104, 119–35, 231–52, 274–6, 280
  - definition 81–2
  - discrete data 119–47, 169–73
  - distribution for discrete random variables 120–7, 134–5, 136–47
  - exercises and answers 105–17, 136–47, 174–83, 213–27, 458–74
  - law of averages 100, 125–7
  - normal distribution 149–83
  - p*-value approach to hypothesis testing 274–6, 280, 305–6, 310–11, 318, 320–32
  - Poisson distribution 120, 130–5, 136–47
  - system reliability 81, 93–9, 103
  - Type I/Type II errors 277–80
- product costing, weighted averages 48–9
- production output 307–9
- properties of quartiles
  - see also* quartiles
  - concepts 58–60, 161–2, 172
- proportions
  - applications 246–8
  - binomial distribution 203–6, 211
  - concepts 203–6, 211, 231, 245–8, 251–2, 272–4, 280, 311–13, 315–16, 319–20
  - definition 203
  - different populations hypothesis testing 311–13, 319–20
  - equation 203, 272–3
  - estimates of population characteristics 231, 245–8, 251–2, 272–4, 311–13
  - hypothesis testing 272–4, 280, 311–13, 315–16, 319–20
  - interval estimates 245–6
  - mean proportion of successes 203–4, 245–6, 311–13
  - sample sizes 245–8, 272–4, 280, 311–12, 319–20
  - sampling distribution of the proportion 203–6, 211
  - standard error 204–6, 245–6, 311–13, 319–20
- quad-modal datasets 52
- quadratic polynomial functions 352–3
  - see also* non-linear regression
- qualitative data 3, 15–24, 250
  - see also* categorical data
- quantitative data 3–15, 23–4, 156–7, 334–65
  - see also* numerical data
- quantity index number with fixed base
  - see also* indexing
  - concepts 385–6, 397
- quantity indexes
  - see also* indexing
  - concepts 385–97
- quartile deviation, properties of quartiles 58
- QUARTILE function, Excel 57, 60
- quartiles
  - boundary limits 57–8
  - box and whisker plots 12, 59–60, 63–4, 161, 164–6, 172
  - concepts 47, 57–60, 63–4, 161–2, 172
  - definition 57
  - examples 58–60, 161, 164–6
  - exercises and answers 65–77, 453–7
  - properties 58–60, 161–2, 172
- questionnaires
  - concepts 15, 24, 209–10, 212
  - consumer surveys 209–10, 212
  - examples 15
- quota sampling, concepts 209, 212
- RAND function, Excel 207, 435
- RANDBETWEEN function, Excel 207, 435
- random numbers 207–8, 212
- random sampling
  - concepts 207–8, 209, 212
  - definition 207
  - examples 207
  - Excel 207, 212
- random variables
  - see also* distribution for discrete random variables
  - characteristics 120–1, 134–5
  - concepts 86–8, 120–7, 134–5, 150, 207, 446
- range, concepts 53–4, 58
- ratio measurement scale, definition 388
- raw data, definition 3
- real value index (RVI) 390–1, 397

- redundancy issues, backup systems 97
- regression analysis 335, 339–81, 432, 436, 496–501
  - see also* forecasts; linear...; multiple...; non-linear...
  - alternative approaches 344
  - application 340–2, 345–7, 348–51, 354–60, 362–3
  - causal forecasting 345–7, 365, 388
  - concepts 335, 339–65, 432, 436
  - confidence levels 343–4, 364
  - considerations 360–5
  - definition 335, 339–40
  - equations 339, 345, 348
  - Excel 340–3, 347, 348, 349–53, 359–60, 432, 436
  - exercises and answers 366–81, 496–501
  - seasonal patterns 353–60, 365, 366–81
  - standard error 343–4, 347, 348–51, 359–60, 364–5, 436
  - variability of estimates 342–4, 364
- rejection/acceptance issues, hypothesis testing 265–9, 276–80, 305–20
- relative frequency histograms
  - concepts 6–7, 16, 18, 23, 83, 100–1
  - examples 7, 18, 101
- relative frequency ogives 9–11
- relative frequency polygons 8–9, 23
- relative frequency probability, concepts 83, 100–1, 103
- relative price index
  - concepts 386, 397
  - definition 386
- relative quantity index
  - concepts 385–6, 397
  - equation 386
- relative regional indexes (RRIs)
  - see also* indexing
  - base selections 391
  - concepts 391–2, 397
  - equation 391
  - exercises 403–12
- relative time-based indexes
  - see also* indexing
  - base changes 387
  - comparisons 387–8, 397
  - concepts 385–91, 397
  - equations 386
  - exercises and answers 398–412, 502–8
  - price index number with fixed base 386, 397
  - quantity index number with fixed base 385–6, 397
  - rolling index number with moving base 387, 397
  - time series deflation 390–1, 397
- reliability of estimates *see* confidence...
- reliability of systems, concepts 81, 93–9, 103
- research hypothesis *see* alternative hypothesis
- right-skewed data, normal distribution 164–9
- right-tail hypothesis tests, concepts 265, 266–7, 279–80, 307–11
- risk 81, 82–117, 124, 276–80
  - see also* probability
  - attitudes 82–3, 92–3
  - aversion 82–3
  - basic probability rules 82, 92–3, 103
  - concepts 82–3, 92–3, 103, 124, 276–80
  - definition 82
  - hypothesis testing 276–80
  - portfolio risk 124–6
  - takers 82–3
- rolling index number, definition 387
- rolling index number with moving base
  - see also* indexing
  - concepts 387, 397
- ROUND function, Excel 435
- RRIs *see* relative regional indexes
- RSQ function, Excel 339, 435
- RVI *see* real value index
- safety valves 198–200
- sample mean, definition 196, 199
- sample standard deviation
  - see also* standard deviation
  - concepts 55–7, 240–1, 342–3
- sample variance
  - see also* variance
  - concepts 54–7
- samples 54–7, 63, 81, 132–3, 169, 171, 173, 185–227, 229–62, 263–99, 301–32, 470–9, 480–8, 489–95
  - see also* estimates...
  - applications 198–200, 201–3, 204–6, 236, 240–8, 269–72, 273–5, 305–19
  - bias 55, 206–7, 212, 231–2
  - binomial concepts for the proportion 203–6, 211
  - central limit theorem 190, 194, 211, 232–3, 269–72, 304
  - concepts 54–7, 63, 81, 132–3, 169, 171, 173, 185–212, 229–52
  - distribution of the sample means 190, 194–6, 211
  - exercises and answers 213–27, 253–62, 281–99, 321–32, 470–9, 480–8, 489–95
  - finite populations 199–202, 211, 231, 236–7, 244–5, 251–2
  - frequency distributions 190–4
  - hypothesis testing 240, 263–99, 301–32
  - inferential statistics 81, 187–8, 211, 229–52, 276
  - infinite populations 196–200, 211, 231, 233–6, 243–5, 251–2
  - margin of error 229, 231, 247–52
  - methods 206–12
  - normal distribution 171, 173, 196–212
  - normal-binomial approximation 171, 173, 203–6, 211
  - population estimates 229–62
  - proportions 203–6, 211, 245–8, 251–2, 272–4, 311–13, 315–16, 319–20
  - size issues 171, 173, 188–97, 211–12, 231, 235–40, 241–3, 245–8, 251–2, 272–4, 280, 303–20, 344



- samples (*Contd*)
  - standard error 196–212, 232–7, 243–5, 251–2, 269–72, 303–20, 343–4, 347, 348–51, 359–60, 364–5, 436
  - statistical relationships 187–97, 211, 272
  - student-*t* distribution 231, 237–45, 251–2, 306–11, 344, 347, 350–1
  - theory 187–212, 231–52
  - transformation relationship 197–9, 211
  - variability issues 195–7
- sampling distribution of the means *see* distribution of the sample means
- sampling distribution of the proportion
  - see also* proportions
  - definition 203
- sampling methods
  - cluster sampling 209, 212
  - concepts 206–12
  - consumer surveys 209–10, 212
  - exercises and answers 213–27, 470–4
  - primary data 210, 212
  - questionnaires 15, 24, 209–10, 212
  - quota sampling 209, 212
  - random sampling 207–8, 209, 212
  - secondary data 210, 212
  - stratified sampling 208, 209, 212
  - systematic sampling 207–8, 209, 212
- scatter diagrams
  - see also* time series
  - application 336–7, 345–6
  - concepts 335–7, 345–6, 364
  - definition 335–6
  - dependent/independent variables 335–6
  - examples 336–7, 345–6
  - Excel 336–7
- seasonal index (SI), concepts 357–60
- seasonal patterns, forecasts 353–60, 365, 366–81
- seaworthiness of ships 273–4, 275–6
- second-degree polynomials *see* quadratic polynomial functions
- secondary data, concepts 210, 212
- series systems
  - assembly operation application 97–9
  - concepts 93–5, 97–9, 103
- shopping malls 119, 120
- SI *see* seasonal index
- side-by-side bar charts *see* parallel bar charts
- sigma 444, 447
- significance levels 264–80, 302–32
  - see also* hypothesis testing
  - chi-square hypothesis test 319, 320
  - concepts 264–5, 279–80
  - definition 264–5
- simple probability *see* classical probability
- single populations, hypothesis testing 263–99
- single type of event (rule no. 1), counting rules 99–100, 104
- size issues, samples 171, 173, 188–97, 211–12, 231, 235–40, 241–3, 245–8, 251–2, 272–4, 280, 303–20, 344
- skewed data 60, 164–9, 172–3
  - see also* asymmetrical data
- SLOPE function, Excel 435
- snowboard sales 336–7, 340–2, 344
- soft drinks 354–60
- SORT function, Excel 11, 12, 435
- SPC *see* statistical process control
- spread *see* dispersion of data
- SQRT function, Excel 240
- stacked histograms
  - concepts 20–2, 24
  - definition 20
  - examples 21–2
- standard deviation
  - binomial distribution 128–9, 134, 169–73, 203–6, 211
  - coefficient of variation 56–7, 63, 361–2, 365
  - concepts 54–7, 63, 121–4, 128–9, 131, 134–5, 151–73, 196–212, 232–43, 269–72, 304–20, 342–4
  - definitions 55, 121, 124, 131, 342–3
  - equations 55, 121, 124, 128, 131
  - examples 55–6
  - hypothesis testing for different populations 303–20
  - normal distribution 151–73, 197–212, 232–7
  - normality of data 161–9, 172–3
  - percentiles 167–9
  - Poisson distribution 131, 135
  - standard error 196, 198–9, 232–7, 251–2, 269–72
  - sum of two random variables 121–4, 134
  - test statistics 269–72, 304–20
- standard error
  - concepts 196, 198–9, 200–2, 232–7, 243–5, 251–2, 269–72, 303–20, 343–4, 347, 348–51, 359–60, 364–5, 436
  - definition 196, 200–1
  - different populations hypothesis testing 303–20
  - equation 196, 200–1, 343
  - exercises and answers 213–27, 470–4
  - finite populations 200–3, 211, 237–8, 244–5, 251–2
  - linear regression 343–4, 347, 359–60, 364–5, 436
  - multiple regression 348–51, 364–5, 436
  - proportions 204–6, 245–6, 311–13, 319–20
  - regression analysis 343–4, 347, 348–51, 359–60, 364–5
  - standard deviation 196, 198–9, 232–7, 251–2, 269–72
- standard normal distribution
  - see also* normal distribution
  - concepts 155–6, 172–3
- statistical dependence
  - basic probability rules 86–8, 103
  - equation 88
  - examples 87

- statistical independence, basic probability rules 84–8
- statistical process control (SPC) 169, 173
- statistical relationships, samples 187–97, 211, 272
- statistics
  - key terminology/formulas in statistics 413–26, 427, 438, 446–7
  - principal uses 54–5
- STDEV function, Excel 55, 240, 435
- STDEVP function, Excel 55, 435
- steel rods 188–95
- stem-and-leaf displays
  - concepts 10–12, 23, 161
  - definition 10
  - EDA 12, 23
  - examples 12
- stratified sampling
  - concepts 208, 209, 212
  - definition 208
  - examples 208
- student-*t* distribution 231, 237–45, 251–2, 269–72, 275–6, 306–11, 344, 347, 350–1
  - application 240–5, 269–72, 275–6, 307–11, 347, 350–1
  - concepts 237–45, 251–2, 269–72, 275–6, 306–11, 347, 350–1
  - confidence intervals 231, 238–43, 251–2, 344, 350–1
  - definition 237–8
  - degrees of freedom 237–43, 251, 269–72, 306–9, 344, 347
  - examples 238–9, 307–11
  - Excel 240–5, 251–2, 308–11, 344, 350–1
  - forecasts 344, 347
  - normal distribution 240
  - profile 238–9, 251–2
  - sample sizes 231, 241–3, 251–2, 306–9, 344
  - z values 238, 241–3
- subjective probability, concepts 82–3, 103
- SUM function, Excel 435
- summation concepts 444–7
- SUMPRODUCT function, Excel 48, 435
- supermarkets 348–51
- surface area, house prices 345–6, 362–3
- symbols
  - arithmetic operating symbols 438
  - Greek alphabet 444, 446–7
  - Microsoft Word 446–7
  - terminology 427, 438, 446–7
- symmetrical data, normal distribution 160–9, 172–3, 234–5
- system reliability and probability
  - assembly operation application 97–9
  - concepts 81, 93–9, 103
  - examples 94–9
  - exercises and answers 105–17, 458–61
  - parallel (backup) systems 93, 95–9, 103
  - series systems 93–5, 97–9, 103
- systematic sampling
  - concepts 207–8, 209, 212
  - definition 208
- systems concepts 93–9, 103
- t*-distribution *see* student-*t* distribution
- taxes 271–2, 275
- TDIST function, Excel 11, 240, 275, 308, 310, 435
- tee-shirts 243–4
- telephone surveys 209–10, 212
- temperature 442–4
- terminology
  - key terminology/formulas in statistics 413–26, 427, 438, 446–7
  - symbols 427, 438, 446–7
- test of independence, chi-square hypothesis test 315–16
- test statistics
  - concepts 268–80, 304–20
  - definition 269
  - equation 269
- time series
  - see also* scatter diagrams
  - concepts 333–4, 335–44, 353–4, 361–2, 364
  - definition 335
  - linear regression 339–44, 353–4, 361–2, 364
  - numerical codes 337, 344
- time series deflation
  - concepts 390–1, 397
  - definition 390
- time-horizon considerations, forecasts 360, 365
- TINV function, Excel 240, 244–5, 271–2, 308, 309–11, 344, 347, 350–1, 435
- transformation relationship
  - see also* z values
  - normal distribution 154–5, 172–3, 197–9, 211
  - samples 197–9, 211
- tri-modal datasets 52
- Truman, Harold 185–6
- Turkey 229–30, 249–50
- two populations, hypothesis testing 302–32
- two-tail hypothesis tests
  - concepts 265–6, 279–80, 305–20
  - examples 265–6
- Type I errors, concepts 276–80
- Type II errors, concepts 276–80
- unbiased estimates *see* point estimates
- univariate data
  - see also* numerical data
  - concepts 3–12, 23
  - definition 3, 23
- unweighted index numbers
  - see also* indexing
  - concepts 392–3, 397
- US measuring system, concepts 442–3



- VAR function, Excel 55, 435
- variability issues, samples 195–7
- variables, concepts 86–8, 120–7, 134–5, 150, 207, 335–6, 345, 437, 445–7
- variance
  - see also* standard deviation
  - binomial distribution 128–9
  - concepts 54–7, 63, 121, 124, 128–9
  - definitions 54, 121, 124
  - distribution for a discrete random variable 121, 124
  - equations 54, 121, 124, 128
  - examples 55–6, 124–6
  - sum of two random variables 124–5, 134
- variation *see* dispersion of data
- VARP function, Excel 55, 435
- Venn diagrams
  - concepts 89–92, 103
  - examples 90–1
  - hospitality management application 89–92
- vertical histograms
  - concepts 16–18, 24
  - examples 17, 18
- vitamin C 240–3
- wages 301, 305–7, 391–2
- Wal-Mart 334
- weighted average, concepts 48–9, 63
- weighted indexes
  - see also* indexing
  - average quantity-weighted price index 395–7
  - concepts 392–7
  - Laspeyres weighted price index 393–4, 397
  - Paasche weighted price index 394–5, 397
- whole numbers, concepts 120
- wine sellers 121–2
- Women and Work Commission (WWC) 301
- work preferences 313–19
- XY(scatter) function, Excel 336, 429–32
- z values 154–73, 197–206, 238, 241–3, 247–52, 272–80, 304–5, 311–13, 319–20 *see also* transformation relationship