Applied Business Analysis

Arun Sukumar; Jayne Revill; Lucian Tipi

ARUN SUKUMAR, LUCIAN TIPI,
JAYNE REVILL

APPLIED BUSINESS ANALYSIS

CONTENTS

AUTHOR'S PREFACE

The reality of every professional's working life these days is the amount of organisational data that they will have to rely on in order to make good, sound, decisions. It is this sheer amount of data that leads to the need to use business analysis methods in order to understand and manipulate many and large data sets with a view of extracting information that will support decision making. To this end, two aspects of business analysis give the ability to extract a significant amount of useful information from numerical data sets:

- A working understanding of data analysis techniques,
- The ability to use data analysis software to implement data analysis techniques.

The two aspects illustrated above will be making the object of this book and you, the reader, will find throughout the following chapters a bland of data analysis techniques, with plentiful examples of implementing these techniques into data analysis software. The data analysis software used in this book is Microsoft Excel, which is the de-facto industry standard and it is used in many organisations. However, given the fairly consistent approach in which the numerical data analysis techniques' functions are implemented across a wide range of data analysis software (e.g. Open Office, Google Apps) the practical examples given in this book can be used directly in a variety of software.

This book relies on a large number of practical examples, with realistic data sets to support the business analysis process. These examples have been refined over time by the authors and have been used to support large numbers of learners that want to forge a professional career in business / marketing / human resources / accounting and finance.

The aims of the book are to:

- Stimulate a structured approach to business analysis processes in organisations,
- Develop an awareness of and competence in using structured business analysis frameworks and models to support information extraction from data sets,
- Illustrate the added value and analytical power of using structured approaches and data analysis software when working with organisational data sets.

Having worked your way through the content of this book, you will be able to:

- Apply appropriate data analysis techniques and models for working on organisational data sets,
- Evaluate your options in terms of which techniques to apply to real world organisational problems,
- Use appropriate software tools to enable data analysis on organisational data sets.

This book has been designed to introduce business analysis concepts and example of increased difficulty throughout, starting with introductory concepts and culminating in financial analysis and that is the way in which the authors recommend that you explore it. However, given the self-contained nature of the examples in the book and depending on your level of ability, feel free to explore it in the way that feels more natural to you, while serving your learning purposes.

At this point the authors wish you an interesting journey through the book, knowing that whether you are a business analysis beginner or a seasoned professional the examples given through the book will enhance your portfolio of data analysis techniques.

1 INTRODUCTION TO FUNDAMENTAL CONCEPTS FOR BUSINESS ANALYSIS

1.1 WHAT IS THE POINT OF THIS BOOK?

The reality of the professional's working environment in the contemporary organisation is that there is an avalanche of data and information that they need to contend with. As such, evidence based decision making is a most valuable skill that any professional or aspiring manager needs to possess in their tool box. As such, making sense of the numerical data that represents the various aspects of an organisation is essential and this is what this book is designed to help achieve.

1.2 GETTING TO GRIPS WITH THE BASICS

In order to use this text book to its full potential for developing your knowledge and skills, we must be able to assume that you are familiar with the relevant material from GCSE mathematics. If you have studied mathematics or statistics beyond GCSE then you will find some aspects of this book easier to tackle. Try to see this bonus as a means of improving yourself and your work, not as a reason for being overconfident. The benefit of post-GCSE study will be in the area of self-managed learning – you should be able to cover the material in this book more quickly. However, you should still work through it carefully, paying particular attention to aspects where you feel that you need further development.

In order to bring your GCSE (or equivalent) knowledge and skills back to your fingertips, work your way through the upcoming concepts and ensure that you understand and can use all of them. If there is anything listed which you do not understand then you should take steps at once to remedy the problem. An extensive bibliography is listed at the end of this chapter, should you need to explore further and in more detail some of the concepts illustrated here.

Remember – the key to understanding the concepts illustrated in this book is practice – the more practice you get the easier to remember and use they will be. The material in the book is presented in such a way that both theoretical and practical illustrations are given along the way (including many Microsoft Excel examples). You, the reader, are strongly encouraged to try your hand at implementing the practical illustrations given along the way in Microsoft Excel, where such implementations are not available in the book. The book will introduce a range of models – illustrated by the use of formulae / equations / Microsoft Excel examples that will enable professionals and managers to effectively manipulate business data sets.

1.3 WORKING THROUGH THE BASICS

In the following sections we will work through a range of essential concepts required for grasping business analysis and applying its concepts to real, organisational problems.

1.3.1 SYMBOLS

The following table contains a range of mathematical symbols and notations that are going to be used throughout this book and that anyone that reads this book will need to conversant in.

+	plus	-	minus
=	equals	≡	same as
>	greater than	±	plus or minus
<	less than	√	square root
≥	greater than or equal to	%	percentage
≤	less than or equal to	a, b, x, y, etc.	various notations use to express variables or other numerical concepts

Table 1.1 Commonly used mathematical symbols

1.3.2 KEY TERMINOLOGY

As well as using a range of mathematical symbols, a range of key words denoting specific business analysis terminology will be used throughout this book. An introduction to these and a brief explanation of each is given in the table below:

arithmetic mean	the mean obtained by adding several numbers together and dividing the sum by the number of numbers added together. Example: the mean of 3, 4 and 5 is (3 + 4 + 5) / 3 = 4.
axis (plural axes)	a line used as a fixed reference, possibly used in conjunction with one or more other similar references for determining the position of a point or of a series of points forming a line, a curve or a surface. Example: x-axis, y-axis, z-axis, etc.
coefficient	a number placed before and multiplying another number (usually a variable). Example: 5 * x, 7 * b, 23 * x * y
constant	a number (quantity) that is assumed to be unchanged throughout a given mathematical process. Example: 3, 5.87, -3.2
co-ordinates	a numerical representation that serves to define the position of a point, line or surface on a system of axes. Example: (3,5) in a (X,Y) two dimensional space
decimal place(s)	the position of a digit after the decimal point. Example: in 0.0247, 4 is situated in the third decimal place
equation	an expression that assumes the equality of two quantities. Example: 2 + 3 * x = 26
formula	a rule expressed in algebraic symbols. Formulae are used to calculate a variety of numerical results, based on a range of input values. Example: x + 3 * y, in practice x and y will need to be given numerical values to allow for calculating the numerical value of the formula
frequency	the number of times that items are occurring in a given category or the number of times that a value occurs in a particular observation. Example: 34 times, 4 times
frequency distribution	the profile of a range of frequencies observed or measured in relation to various categories of items. Example: a list of price ranges for sun screen products paired with the number of products available in each price range
integer (whole number)	a number that does not have any decimal places. Example: 1, 15, 3467, etc.
linear equation	a polynomial equation of the first degree (none of its variables are raised to a power of more than 1). Example: x – y = 25, x + 3 = 68 * y
median	the middle positioned number in a sequence of numbers. If the sequence has an even number of numbers, then the median is taken as the average of the two middle numbers. Example: the median of 1, 5, 6, 7, 9 is 6. The median of 1, 2, 5, 6, 7, 9 is (5 + 6) / 2 = 5.5

Download free eBooks at bookboon.com

percentage	a proportion usually worked out of one hundred, what proportion of a given total is being considered. Example: one quarter of a quantity is worked out as 25 parts of the total 100 parts of that quantity, leading to 25 / 100 times that quantity. This is usually represented as 25%. Sometimes a percentage it may be represented as a value between 0 and 1, for example 25% would be represented as 0.25
range	the numerical difference between the highest and lowest values of a series of numbers. Example: the range of 1, 2, 5, 6, 7, 9 is 9 – 1 = 8
significant figures (digits)	all the nonzero digits of a number and the zero digits that are included between them. Also final zeros that signify the accuracy of a given number. Example: for 0.025600 the significant numbers are 256 and the final two zeroes which indicate 6 places accuracy.
variable	a quantity that may take any given value from the range of values that it belongs to. Example: integer variable x may take any integer value, such as 3, 76, 8, 2456, etc.

Table 1.2 Essential terminology

1.3.3 DECIMAL PLACES AND SIGNIFICANT FIGURES

Writing numbers to an appropriate number of decimal places or significant figures is explained in this section.

If the answer to a calculation was 53.025602, you would probably not write this down. Instead, you would 'round off' the number. There are two ways to do this. You can round off to a certain number of:

- decimal places, or,
- significant figures

Therefore 53.025602 to 1 decimal place (1dp) is 53.0 – write down 1 number after the decimal point. 53.025602 to 4 decimal places (4dp) is 53.0256 – write down 4 numbers after the decimal point.

However, 53.025602 to 1 significant figure (1sf) is 50 – starting from the left, write down 1 number (we must include the 0 to keep the 'place value' – i.e. make it clear that the number is 50 and not 5). 53.025602 to 4 significant figures (4sf) is 53.03 – starting from the left, write down 4 numbers which gives 53.02 but the next digit is a 5, hence it is assumed that the number is closer to 53.03 than 53.02.

The rule is: when rounding a number, if the digit after the place you stop is 5 or above, you add one to the last digit that you write.

If the result of a calculation is 0.00256023164, we would 'round off' the answer. This number, rounded off to 5 decimal places is 0.00256. You write down the 5 numbers after the decimal point.

To round the number to 5 significant figures, write down 5 numbers. However, since the number is less than 1, you do not count any zeros at the beginning. So to 5sf, the number is 0.0025602 (5 numbers from the first non-zero number).

If 4.909 is rounded to 2 decimal places, the answer is 4.91.
If 3.486 is rounded to 3sf, the answer is 3.49

0.0096 to 3dp is 0.010 (add 1 to the 9, making it 10). When rounding to a number of decimal places, always write any zeros at the end of the number. If you say 3dp, write 3 decimal places, even if the last digit is a zero.

Further resources in this area are available at the end of this chapter, should they be needed.

1.3.4 PERCENTAGES

In this section we will look at evaluating percentage increases/decreases and how to work with percentages in a variety of applications.

A reminder of what a percentage is:

"a rate or proportion of a hundred parts"

For example, to work out a percentage of a numerical value, we would need to multiply the percentage with the value and then divide by 100. Of course the calculation of percentages has been made easy by the introduction of calculators and office software (e.g. Microsoft Excel).

Assuming that we want to calculate 3% of 225:

The result is (3 * 225) / 100 = 675 / 100 = 6.75%

An alternative way of calculating percentages would be to divide the numerical value that we want to work out a percentage of by 100, this would obtain the value of 1% of our numerical value and then multiply this with the percentage that we are trying to calculate.

Using this method, our calculation of 3% of 225 would become:

225 / 100 = 2.25 – this represents 1% of 225. Therefore 3% of 225 is 2.25 * 3 = 6.75%

Yet another way to calculate a percentage would be to express something like 3% as 3 / 100 = 0.03 and then multiply this with the value that we are trying to calculate a percentage of. Our example above would become 0.03 * 225 = 6.75%.

In a business environment, as well as calculating a percentage, often we need to calculate changes in percentages, for comparison or decision making purposes.

For example: A person's salary in a company has changed from £20,000 per annum to £25,000 over the course of the year. What is the percentage increase for this person's salary?

Solution: The percentage change is given by the following formula:
              [(New Value – Old Value) / Old Value] * 100

In this example, New Value = £25,000, Old Value = £20,000

Therefore the percentage increase will be:

[(25,000-20,000)/20,000] * 100 = (5,000/20,000) * 100 = 0.25 * 100 = 25%

Therefore, the percentage increase of the person's salary = 25%.

Exercise: An item costs £25.02 inclusive of VAT at 20%. What is the price excluding the VAT?

Solution:

Price excluding VAT + VAT = Price including VAT
Price excluding VAT + Price excluding VAT * 0.2 = Price including VAT
Price excluding VAT * (1+ 0.2) = Price including VAT
Price excluding VAT = Price including VAT / (1 + 0.2)
Price excluding VAT = 25.02 / 1.2
Price excluding VAT = £ 20.85

1.3.5 LINEAR EQUATIONS AND THEIR GRAPHS

In this section linear equations and their graphs will be discussed.

In business analysis it is important to have the ability to:

- recognise the form of a linear equation, i.e. a straight line equation in the $y = a + b * x$ or $y = m * x + n$ format, where a, b, m and n are constants. A linear equation is called as such due to the fact that if the its output values were plotted on a graph, it would produce a straight line.
- plot a graph of a linear equation

In business analysis it is important to identify the linear equations that describe a trend for example; this will be discussed later on in this book.

For now, let us assume that we utilize the following formula for a linear equation:

$$y = 3 + 2 * x$$

In order for us to plot the line generated by this formula, we would need to choose at least two values for x and then calculate the corresponding y values (in fact two values for x are sufficient as a line in a two dimensional space is fully defined by 2 sets of co-ordinates, all other values produced by the same formula will be part of the same line). We will then is a position to plot the line produced by the formula above.

The best way to illustrate this is by producing a table, with the values of x and y. Assuming that x is an integer value in this case we can choose two integer values for x and then obtain the corresponding y values:

Value for x	Value of y
2	3 + 2 * 2 = 3 + 4 = 7
4	3 + 2 * 4 = 3 + 8 = 11

Table 1.3 Example of calculating the values of a line

The resulting line produced by our formula above is therefore obtained by plotting the co-ordinates (2, 7) and (4, 11) from the table above:
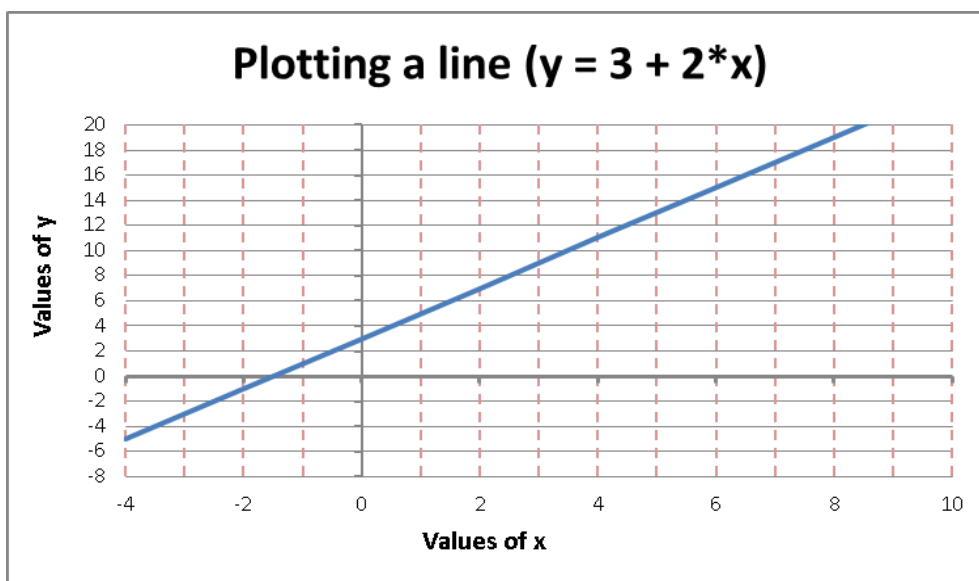
Figure 1.1 Plotting a line

It can be noted that the line crosses the x axis at -1.5, while for x = 0, the value of y = 3. These properties of a line will be studied further in the book, as they are important in determining certain characteristics of a line.

In a line of the y = a + b * x format, a is called the y intercept as it is the value of y where the line crosses the y axis and b is called the slope of the line as it gives a measure of the inclination of the line.

1.3.6 FINDING AVERAGE VALUES

Average values are essential concepts that are widely used in business analysis to evaluate numerical indicators in an organisation. As such it is quite important to be conversant in these concepts. The best way to achieve this is by practising the use of the various average values measures.

Based on the introductory concepts listed in 1.3.2, work your way through the following exercises:

Exercise 1.1: find the median of the following set of numbers: 1, 2, 3, 6, 7
Solution: the median value is 3
Exercise 1.2: find the median of the following set of numbers: 1, 2, 3, 5, 6, 7
Solution: the median value is (3 + 5) / 2 = 8 / 2 = 4
Exercise 1.3: find the (arithmetic) mean of the following set of numbers: 23, 34, 27, 15, 19
Solution: the arithmetic mean is (23 + 34 + 27 + 15 + 19) / 5 = 118 / 5 = 23.6

Exercise 1.4: find the median of a frequency distribution from the following table (data gathered from a sample of families where the number of children per family was observed):

Number of children / family	Frequency
0	11
1	15
2	24
3	17
4	9
5	7

Table 1.4 Frequency distribution

Solution: We need to find out how many data points there are, by adding the values in the Frequency column: 11 +15 + 24 + 17 + 9 + 7 = 83. So the median of this frequency distribution is the value of number of children / family corresponding to value 42 of the frequency:

Number of children / family	Frequency	Frequency values so far
0	11	11
1	15	26
2	24	50
3	17	67
4	9	76
5	7	83

Table 1.5 Frequency distribution – calculations for obtaining median

The number of children corresponding to value 42 is then 2, as value 42 occurs in the 2 children / family row.

Exercise 1.5: use a calculator to find the mean of the following frequency distribution (data gathered from a sample of families where the number of children per family was observed):

Number of children / family	Frequency
0	11
1	15
2	24
3	17
4	9
5	7

Table 1.6 Frequency distribution – calculations for obtaining mean

Solution: we need to calculate a third column that is the multiplication of the values in the first two columns:

Number of children / family (n)	Frequency (f)	n * f
0	11	0 * 11 = 0
1	15	1 * 15 = 15
2	24	2 * 24 = 48
3	17	3 * 17 = 51
4	9	4 * 9 = 36
5	7	5 * 7 = 35

Table 1.7 Frequency distribution – calculations for obtaining mean

We now need to total the values for n * f and to identify the number of data points in the frequency distribution:

Number of children / family (n)	Frequency (f)	n * f
0	11	0
1	15	15
2	24	48
3	17	51
4	9	36
5	7	35
Totals	11 + 15 + 24 + 17 + 9 + 7 = 83	0 + 15 + 48 + 51 + 36 + 35 = 185

Table 1.8 Frequency distribution – calculations for obtaining mean

Therefore the average value of the frequency distribution given here is 185 / 83 = 2.2289 or 2.23 to 2 decimal points.

1.3.7 RAISING A NUMBER TO A POSITIVE INTEGER VALUE

In order to expedite calculations and simplify representation, an expression such as a * a * a may be written as a^3 – this signifies a to the power of 3.

In mathematical expressions powers take a format similar to a^n or x^2. Typically, the index n is a positive integer.

It is important to understand the meaning of an where a is a number and n is a positive whole number (integer), e.g. a^2, a^3, a^4, etc.

For example we might have 5^2 which is $5 \times 5 = 25$ or $2^3 = 2 \times 2 \times 2 = 8$ or 10^6 which is $10 \times 10 \times 10 \times 10 \times 10 \times 10 = 1,000,000$ or a million.

Note that $a^0 = 1$ irrespective of the value of a, so for example $10^0 = 1$.

1.4 REVISION SECTION – EXERCISES

1. If y = 1 / x, write down the value of y, correct to 2 decimal places, when x = 8

   Solution: y = 1 / 8 = 0.125. Therefore, 0.125 to two decimal places is 0.13

2. If the sales of a product increased by 5% in one year and by 8% in the next year, what was the overall percentage increase in sales over the two years? (to one decimal place)

   Solution:
   The sales after the first year = 1.05 × initial sales
   The sales after the second year = 1.08 × (1.05 × initial sales) = 1.134 × initial sales
   This gives an overall increase of 13.4% (0.134 x 100)

3. Find the mean of the set of eleven numbers listed below. Give your answer to 2 decimal places.
   7, 3, 2, 3, 8, 6, 7, 0, 9, 2, 4

   Solution: To find the mean of a set of numbers divide the sum of the numbers by the number of numbers.

   There are eleven numbers and their total (sum) is 51. Therefore the mean is 51 / 11 = 4.636363 = 4.64 to 2 decimal places

4. What is the median of the eleven numbers listed in the exercise 3?

   Solution: To find the median of a set of numbers arrange them in rank order (smallest first) to start with. The median is the value which lies in the middle of the set of numbers.

   In rank order the numbers are: 0, 2, 2, 3, 3, 4, 6, 7, 7, 8, 9
   The value that lies in the middle of the set is therefore 4

5. £25,000 is divided between three people in the ratios 5 : 3 : 2.
   What is the value of the largest share?

   Solution: we need to find out how many equal parts there are in total 5 + 3 + 2 = 10
   Then we divide £25,000 into 10 equal parts £25000 / 10 = £2,500
   Therefore the largest share is £2,500 * 5 = £12,500

6. A total bill, including VAT at 20%, comes to £301.25 – to the nearest penny. What is the bill excluding VAT – to the nearest penny?

    Solution:
    Bill excluding VAT = b
    Bill including VAT = 301.25 = b * (1 + 0.2) = b * 1.2
    Therefore 1.2 * b = 301.25 giving b = 301.25 / 1.2 = 251.041666 = £251.04 to nearest penny

7. A long distance road haulage company is extending its business from within the European Union to the Balkans. One of its new routes runs from Frankfurt (Germany) to Belgrade (Serbia) via Ljubljanas (Slovenia). The distance from Frankfurt to Ljubljana is 725 km and from Ljubljana to Belgrade is 445 km. Express the total distance from Frankfurt to Belgrade in miles given that 1 mile = 1.609 km. Give your answer to the nearest whole number.

    Solution: The total distance in km from Frankfurt to Belgrade is 725 + 445 = 1170 km
    1 mile = 1.609 km so 1 km = 1/1.609 miles
    Therefore the total distance in miles is 1170 km * 1 / 1.609 = 1170 / 1.609 = 727.159726 miles = 727 miles (to nearest mile)

8. Express the distance from Ljubljana to Belgrade as a percentage of the total distance of the journey. Give your answer to 3 significant figures. Answer this exercise using the information given in Exercise 7.

    Solution: The distance from Ljubljana to Belgrade is 445 km. The total distance is 1170 km.
    445 as a percentage of 1170 is 445 / 1170 × 100 = 0.380341 × 100 = 38.0341. Therefore the answer is 38.0 to 3 significant figures.

9. What is the value of the intercept on the y axis of the graph shown below?
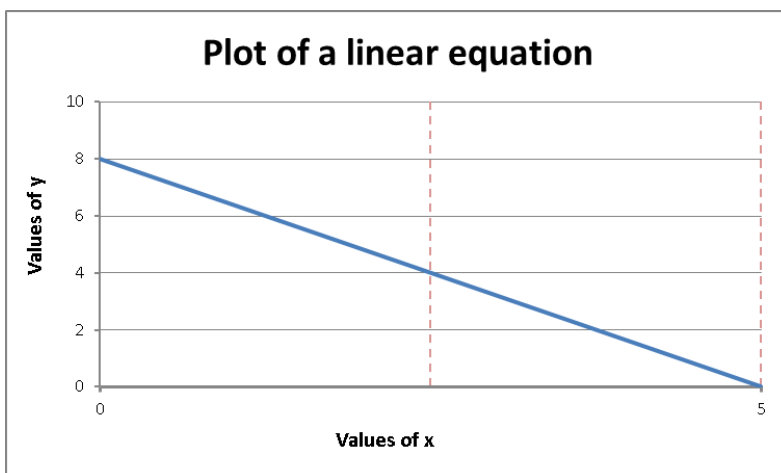
Figure 1.2 Plot of a linear equation

Solution: The intercept on the y axis is the value of y where the graph crosses the y axis (vertical axis). Therefore the intercept is 8.
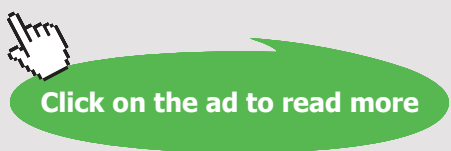
10. For the graph in the previous exercise write down the equation of the line in the form y = a + b * x

Solution: The equation of any straight line can be written in the form y = a + b * x where a and b are constants.

When x = 0 (on the y axis) y = 8 therefore a = 8. This is the intercept of the line. Therefore the equation is so far y = 8 + b * x

When y = 0 (on the x axis) x = 5 therefore 0 = 8 + 5 * b

Hence 5 * b = -8, this gives b = -8 / 5 = -1.6. This is the slope of the line.

Therefore the full equation of the line is y = 8 − 1.6 * x.

1.5 CHAPTER REFERENCES

1. Morris, C, Quantitative Approaches in Business Studies, 8th edition, FT/Prentice Hall, 2012 (eBook, 7th edition 2008), Chapter 1

2. Oakshott, L, Essential Quantitative Methods for Business, Management & Finance, 5th edition, Palgrave Macmillan, 2012, Chapter 1

3. Waters, D, Quantitative Methods for Business, 5th edition, FT/Prentice Hall, 2011 (eBook available), Chapters 1 and 2

4. Wisniewski, M, Quantitative Methods for Decision Makers, 5th edition, FT/Prentice Hall, 2009 (eBook available), Chapter 2 and 10

2 DECISION MAKING, DATA AND INFORMATION

2.1 DECISION MAKING

Running a business or any complex organisation involves making choices. Every day and every minute of the day someone will be making a decision that will affect some aspect of the business and its customers, suppliers, staff or shareholders. Managers within businesses are charged with undertaking a number of roles, many of which involve making decisions.

In general, ***decision making*** involves making a choice between a set of optional courses of action according to a set of criteria or decision rules.

2.1.1 FUNCTIONS OF MANAGEMENT

The functions of management can be grouped into specific areas:

- planning
- organising and co-ordinating
- leading and motivating
- controlling processes

Planning is the process of deciding in advance what is to be done and how it is to be done (e.g. where to build a new plant, how much to spend on marketing). The process results in plans (predetermined courses of action) that reflect organisational objectives and goals. This clearly involves decision making.

Organising and co-ordinating people, resources, materials in order to implement the plan to get things done, will sometimes involve choosing between different courses of action, different people or other resources to do the job.

Leading and motivating people may involve making choices between different styles of management and working to different timeframes.

Controlling the process ensures that things proceed according to the plan. This is often achieved by comparing actual performance with a target and using any difference to guide the adjustment of the operation and thus, to bring about the desired performance (e.g. if the temperature in the office is too low, turn up the heating). Where there is a choice of actions that can be taken, then decisions will have to be made.

2.1.2 INFORMED DECISIONS

Making the right (or at least a good) choice requires judgement but also requires a basis on which to make the choice. Many decisions are based on *information* which relates to the decision and which informs the decision. A decision on how many items to order this week may well depend on how many of the item were sold last week or the same week last year. The choice of person for a job will depend on their prior performance and achievements. The information requirements for making decisions will be explored in more detail later on.

2.1.3 DECISION MAKING WITHIN THE ORGANISATION

Decisions are made at different levels within the organisation:
operational or transactional decisions
made by junior managers or operatives
affect the immediate running of the organisation (or section)

Here problems of a recurring nature are dealt with. For example:

- weekly staff schedule for a particular production line
- weekly machine maintenance schedule
- daily raw material inventory check

The information required is precise, usually not financial and usually related to a policy prescribed by a higher level of management.

tactical decisions
made by middle managers
affect the medium term running of the organisation (or department)

This middle level of management is concerned with decisions which are made on a regular or periodic basis (annually, quarterly, monthly). Tactical decisions will usually be short-range, covering planning cycles of a year or so. These decisions primarily require information of a historical (i.e. company records) or financial nature which is generated within the organisation. For example:

- the budget for personnel recruitment in the next financial year
- expenditure on advertising in the next quarter
- monthly sales targets for the next quarter

strategic decisions
made by senior managers
affect the longer term development of the organisation

This is the executive or top level of management which is concerned largely with issues of long-range planning. For example:

- how large should the organisation be in 10 years' time?
- how many production lines should there be in 5 years' time?
- what kind of research and development policy should be adopted?
- how should the product range be developed over the next 20 years?

For this type of decision-making, management will require access to all internal information, as well as all relevant external information. The information is used irregularly i.e. the decisions made are not routine.

In addition to the information requirements of these management categories, there will be other groups, concerned with an organisation, which have yet other information requirements. For example:

- Employees in general require information about wages, the firm's progress, developments in the provision of staff facilities, etc.
- Shareholders require information about the company's current and expected performance.

2.1.4 TYPES OF DECISIONS

Some decisions are made routinely and others are more novel.

Programmed decisions are routine and repetitive, with clear options and known decision rules (e.g. stop when light is red; always load the largest items into the lorry first). This type of decision tends to be made at the more operational levels within the organisation.

Non-programmed decisions are more novel and unstructured, with complex options and unclear decision rules (e.g. what to do when a machine develops an unusual fault, how to deal best with a new competitor in the market place). This type of decision tends to be made at a more tactical or strategic level.

2.2 DATA AND INFORMATION

Consider what happens at a travel agency.
Some of the *decisions* that must be made are:
how many seats on a charter flight to book for this holiday next year
who to send the latest brochure to
which holidays to promote

Information that will inform these decisions includes:
number of people booked on a holiday
number of un-booked places
comparative demand for the holiday over the past three years
profitability of the package holiday
number of holidays taken by a particular family

This information comes from data that is collected regularly as part of the business:
details of package holidays (price, location, duration)
details of package holiday bookings (no of people, date of departure, holiday selected, amount paid)

Although the terms **data** and **information** are commonly used interchangeably, technically the terms have distinct meanings.

Data are raw facts, unorganised and frequently unrelated to one another. Data are frequently numerical (quantitative) but are not necessarily so. Data can be non-numeric (qualitative).

Examples of data:

- a certain machine broke down 4 times last week,
- there are 13 employees in the Accounts Department,
- last year's budget for the HR Department was £157,000,
- employee satisfaction with working conditions in the factory,
- in December, 1500 wheelbarrows were produced,
- last month 385 expense claims were submitted.

These are examples of ***internal*** data, generated within an organisation.

External data are generated outside the organisation. For example:

- the inflation rate rose last month to 4% pa,
- Parliament has just passed new legislation on pollution controls,
- the imports of Japanese cars rose last year by 5%,
- correspondence from a customer praising the wheelbarrow he purchased.

We will consider sources of external data in more detail later.

Information is obtained by processing data in some way. Information is a collection of related pieces of data. For example:

- the machine that broke down 4 times last week had a major overhaul only 3 weeks ago,
- the 13 employees in the Accounts Department represent 10% of the employees of the company,
- last year's budget of £157,000 for the HR Department was up by 12% on the year before,

The drawing together of relevant pieces of data provides managers with some important information. Information adds context to the data and provides meaning.

Data can be processed into information by:

- bringing together related pieces of data and tabulating, aggregating, filtering or simply rearranging them;
- summarising;
- tabulation and the use of diagrams;
- statistical analysis;
- financial analysis

2.3   SOURCES OF DATA

Organisations of every kind constantly generate large quantities of data which require organising and processing in a variety of ways in order to satisfy the information needs of each function in the organisation. In addition, the 'home grown' data are supplemented by vast quantities of data which are generated externally, but nevertheless may form the basis of useful information for the organisation.

The problem faced by the organisation is the management of these data so that the necessary information can be acquired at the right time and in the correct format.

2.3.1 INTERNAL SOURCES OF DATA

We have already seen that organisations hold significant amounts of transactional data in their databases. Suitably processed, this can provide information for making operational, tactical and strategic decisions.

2.3.2 EXTERNAL SOURCES OF DATA

Organisations frequently use data obtained outside the organisation itself. For example:

- a market research survey to determine customer satisfaction with a particular product
- information, from official publications, on the size and characteristics of the population is useful when estimating the number of potential customers for a new product.
- information, from company reports, on the activities (sales, investments, take-overs, etc.) of competitors is important if a company is to remain competitive.

2.3.3 PRIMARY AND SECONDARY DATA

Data which are used solely for the purpose for which they were collected are said to be *primary data*.

Data which are used for a different purpose to that for which they were originally collected, are called *secondary data*.

The terms 'primary' and 'secondary' refer to the purpose for which the data are used.

2.3.4 THE PROBLEMS OF USING SECONDARY DATA

In most cases it is preferable to use primary data since data collected for the specific purpose is likely to be better, i.e. more accurate and more reliable. However, it is not always possible to use primary data and therefore we need to be aware of the problems of using secondary data. Some of the problems are listed below:

a) The data have been collected by someone else. We have no control over how it was done. If a survey was used, was:
- a suitable questionnaire used?
- a large enough sample taken?
- a reputable organisation employed to carry out the data collection?
- the data recorded to the required accuracy

b) Is the data up-to-date? Data quickly becomes out-of-date since, for example, consumer tastes change. Price increases may drastically alter the market.

c) The data may be incomplete. Certain groups are sometimes omitted from the published figures, for example, unemployment figures do not include everyone who does not have a job. (*Which groups are left out?*) Particular statistics published by the Motor Traders Association, for example, may exclude three-wheeled cars, vans and motor-caravans. We must know which categories are included in the data.

d) Is the information actual, seasonally adjusted, estimated or a projection?

e) The figures may not be published to a sufficient accuracy and we may not have access to the raw data. For example, population figures may be published to the nearest thousand, but we may want to know the exact number.

f) The reason for collecting the data in the first place may be unknown, hence it may be difficult to judge whether the published figures are appropriate for the current use.

If we are to make use of secondary data, we must have answers to these questions. Sometimes the answers will be published with the data itself or sometimes we may be able to contact the people who carried out the data collection. If not, we must be aware of the limitations of making decisions based on information produced from the secondary data.

2.3.5 SOURCES OF SECONDARY DATA

There are numerous sources of secondary data. They can be broadly categorised into two groups:

- a) Those produced by individual companies, local authorities, trade unions, pressure groups etc. Some examples are:
  - i) Bank of England Quarterly Bulletin – reports on financial and economic matters.
  - ii) Company Reports (usually annual) – information on performance and accounts of individual companies.
  - iii) Labour Research (monthly) – articles on industry, employment, trades unions and political parties.
  - iv) Financial Times (daily) – share prices and information on business.

- b) Those produced by Government departments. This is an extensive source of data and includes general digests, such as the Monthly Digest of Statistics, as well as more specific material, such as the New Earnings Survey.

Government Statistics – a brief guide to sources lists all of the main publications and departmental contact points.

Guide to Official Statistics is a more comprehensive list.

There is a comprehensive coverage of government statistical publications and access to the data online at http://www.statistics.gov.uk

2.4 COLLECTING DATA

Every organisation collects what may be called routine data. These include records about staff, customers, invoices, sales, industrial accidents, stock in hand, stock on order, etc. Each organisation will devise some system for the collection and storage of this type of data until it is required.

In addition there will be occasions when the organisation requires data to be collected for some special purpose. For example:

1. a survey of consumer reaction to a recently launched product
2. an assessment of the company's production line efficiency
3. an investigation into the type of faults which have occurred in a particular product
4. a survey of employee reaction to proposed changes in the staff canteen.

In these cases, non-routine methods of collecting data must be employed. The most commonly used methods are:

- a) direct observation or direct inspection (example 2. above)
- b) written questionnaire (includes postal and online) (example 4. above)
- c) personal interview (includes telephone) (example 1. above)
- d) abstraction from records or published statistics (example 3. above)

Direct observation means that the situation under investigation is monitored unobtrusively. This is an ideal method from the point of view of the investigator, since the likelihood of incorrect data being recorded is small. It is, however, an expensive way to collect data.

It is essential that the act of observation does not influence the pattern of behaviour of the observed. For example, if the behaviour of shoppers in a supermarket is being observed, then, many will change their pattern of behaviour if they become aware of the observation. They may even leave the store altogether!

The method is used primarily for scientific surveys, road traffic surveys and investigations such as the determining of customer service patterns.

Direct inspection uses standardised procedures to determine some property or quality of objects or materials. For example, a sample of 5 loaves is taken, from a batch in a bakery, and cut open to test the consistency of the mixture.

Written questionnaire is one of the most useful ways of collecting data if the matter under investigation is straightforward so that short, simple questions can be asked. The questionnaire should comprise the type of questions which require a simple response e.g. YES/NO, tick in a box, ring a preferred choice, etc.

A survey using a questionnaire is relatively cheap to do – the time required (and, hence the cost) is much less for the mailing of 500 questionnaires than it would be if 500 personal interviews were conducted. The main problem is that the response rate for postal/online questionnaires is typically very small, perhaps 10%.

The design of the questionnaire is very important and is by no means as simple as it sounds.

Personal interviewing may be used in surveys about people's attitudes to a particular issue e.g. public opinion poll. It is necessary to have a trained interviewer who remains impartial throughout the interview. The type of question which is asked can be more complicated than that used in a questionnaire since the interviewer is present to help promote understanding of the questions and to record the more complex responses. Tape recording is sometimes used. Clearly, the cost of personal interviewing is high. The use of the telephone reduces the cost but increases the bias in the sample, since not every member of the population can be accessed by phone.

Abstraction from records or published statistics is an extremely cheap and convenient method of obtaining data. However, the data used will usually have been collected for a different purpose and may not be in the format required. All of the problems of using secondary data then arise.

2.5   SAMPLING

2.5.1   POPULATIONS AND SAMPLES – WHY SAMPLE?

The term *population* refers to the entire group of people or items to which the statistical investigation relates. The term *sample* refers to a small group selected from that population. In the same way, we use the term *parameter* to refer to a population measure, and the term *statistic* to refer to the corresponding sample measure. For example, if we consider our population to be the current membership of the Institute of Directors, the mean salary of the full membership is a population parameter. However, if we take a sample of 100 members, the mean salary of this group is referred to as a sample statistic.

In a survey of student opinion about the catering services provided by a college, the target population is all students registered with that college. However, if the survey was concerned with catering services generally in colleges, the target population would be all students in all colleges. Since it is unlikely that in either case, the survey team could collect data from the whole population, a small group of students would be selected to represent the population. The survey team would then draw appropriate inferences about the population from the evidence produced by the sample. It is very important to define the population at the beginning of an investigation in order to ensure that any inferences made are meaningful.

Populations, such as those referred to above, are limited in size, and are known as *finite* populations. Populations which are not limited in size are referred to as *infinite* populations. In practice, if a population is sufficiently large that the removal of one member does not appreciably alter the probability of selection of the next member, then the population is treated statistically as if it were infinite.

On first consideration, we might feel that it would be better to use the whole population for our statistical investigation if at all possible. However, in practice we often find that the use of a sample is better. The advantages of using a sample rather than the population are:

a) **Practicality**

   The population could be very large, possibly infinite. Hence, it would be physically impossible to collect data from the whole population.

b) **Time**

   If the data are needed quickly, then there may not be enough time to cover the whole population. For example, if we are concerned with checking the quality of goods produced by a mass production process, the delay in delivery whilst every item is checked will be unacceptable.

c) **Cost**

The cost of collecting data from the whole population may be prohibitively high. In the example above, the cost of checking every item could make the mass produced item excessively expensive.

d) **Errors**

If data were collected from a large population, then the actual task of collecting, handling and processing the data would involve a large number of people and the risk of error increases rapidly. Hence, the use of a sample, with its smaller data set will often result in fewer errors.

e) **Destructive**

The collection of the data may involve destructive testing. In **tests** for which this is the case, it is obviously undesirable to deal with the entire population. For example, a manufacturer wishes to make a claim about the durability of a particular type of battery. He runs tests on some batteries until they fail to determine what would be a reasonable claim about all the batteries.

When would we use a population rather than a sample?

a) **Small populations**

If the population is small, so that any sample taken would be large relative to the size of the population, then the time, cost and accuracy involved in using the population rather than the sample will not be significantly different.

b) **Accuracy**

If it is essential that the information gained from the data is accurate, then statistical inference from sample data may not be sufficiently reliable. For example, it is necessary for a shop to know exactly how much money has been taken over the counter in the course of a year. It is not sufficient, for the owner to record takings on a sample of days out of the year. The problem of errors is still relevant here, but any errors in the data will be ones of arithmetic rather than unreliability of statistical estimates.

The purpose of a statistical investigation may be to measure a population parameter, for example, the mean age of professional accountants in the UK, or the spread of wages paid to manual workers in the steel industry in France. Alternatively, the purpose may be to verify a belief held about the population, for example, the belief that violence on television contributes to the increase in violence in society, or, that jogging keeps you fit. In most instances the whole population is not available to the investigation and a sample must be used instead.

The use of data from a sample instead of a population has important implications for the statistical investigation since it leads us into the realms of statistical inference; it becomes necessary to know what may be inferred about the population from the sample. What does the sample statistic tell us about the population parameter, or what does the evidence of the sample allow us to conclude about our belief with respect to the population? How does the mean age of a *sample* of professional accountants relate to the mean age of the *population* of professional accountants? If, in a *sample* of adults, the joggers are fitter, can we claim for the *population* as a whole that jogging keeps you fit? With an appropriately chosen sample, it is possible to estimate population parameters from sample statistics, and, to use sample evidence to test beliefs held about the population.

Statistical inference is a large and important aspect of statistics. Information is gathered from a sample and this information is used to make deductions about some aspect of the population. For example, an auditor may check a sample of a company's transactions and, if the sample is satisfactory, he will assume, or infer, that all of the company's transactions are satisfactory. He uses a sample because it is cheaper, quicker and more practical than checking all of the transactions carried out in the company.

2.5.2 THE SELECTION OF A SAMPLE

It is extremely important that the members of a sample are selected so that the sample is as representative of the population as possible, given the constraints of availability, time and money. A biased sample will give a misleading impression about the population.

There are several methods of selecting sample members. These methods may be divided into two categories – *random* sample designs and *non-random* sample designs.

2.5.3 RANDOM SAMPLING

Statistical inference can be used only with this kind of sample. Random sampling means that every member of the population has an equal chance of being selected for the sample.

If the population consists of groups of members with different characteristics, which are important in the investigation, then random sampling should result in a sample which contains members from each of these groups. The larger the group, the more likely it is that its members will be selected for the sample, hence, the larger its representation in the sample. The size of the group is related to the probability that its members will be selected, therefore random sampling has a tendency to lead towards a sample which is representative of the population.

The first step in selecting a random sample from a finite population is to establish a *sampling frame*. This is a list of all members of the population. It does not matter what form the list takes as long as the individual members can be identified. Each member of the population is given a number, then some random method is used to select numbers and the sample members are thus identified. The representativeness of the sample depends on the quality of the sampling frame. It is important that the sampling frame possesses the following properties:

a) **Completeness** – all of the population members should be included in the sampling frame. Incompleteness can lead to defects in the sample, especially if the members which are excluded belong to the same group within the population.

b) **Accuracy** – the information for each member should be accurate and there should be no duplication of members.

This list of the members of the population should not be confused with a census of the population. The purpose of the sampling frame is solely to enable individuals to be identified for selection purposes. The data required have then to be collected from the sample members. If the data are collected from all members of the sample frame, i.e. the whole population, then this is a census. The random selection method can be manual such as picking numbered balls from a bag. Alternatively, a calculator or a computer can be used to generate random numbers. These are then used to select the sample.

2.5.4 RANDOM SAMPLE DESIGNS

Simple Random Sample Design

The basic principle of random sampling has already been described in the previous section and the procedure for the selection of a simple random sample was outlined. A simple random sample is most suitable when the members of the population are similar for the purposes of the investigation. For example, a simple random sample would be suitable for selecting a sample of 20 employees of a company to take part in a survey if we are interested only in the fact that the individuals are employees of this company.

Stratified Random Sampling

If the populations members have different characteristics, which are of interest, then the simple random sampling method may not give the most representative sample for a given sample size. For example, in the survey of employees of a company, it may be important to distinguish between male and female employees. A simple random sample design could result in too many members of one gender, unless a large sample is used.

The Stratified Random Sampling method enables us to produce a sample which more accurately reflects the composition of the population. The procedure requires the sampling frame to be subdivided into the groups of interest. These groups are referred to as strata. In our example, the males and females should be identified. We need to know how many members of the population fall into each category. We then use a simple random sampling method to select sample members from each stratum separately in proportion to their number in the population. If the population contains 60% males and 40% females, and we are selecting a sample of 100 members, then we select 60 men from the male stratum and 40 women from the female stratum. The main advantage of stratified random sampling is that we can use smaller sample sizes to get the same results as with a simple random sample.

Systematic Sample Design

The advantage of systematic sampling is that the selection of the sample is much simpler than the previous methods we have considered. If the population contains 1000 members and we require a sample of 100, then every (1000/100) 10th population member is selected from the sampling frame. The random element is introduced by selecting the starting point at random. For example, if we have a set of invoices, numbered 1 to 1000, we use a random number to select the first invoice number, and then pick out every following tenth invoice. If the random number was 3, for example, we pick out invoice numbers 3, 13, 23, 33 etc. This gives the required sample. We have used only one random number, compared with 100 for the other methods. This is a satisfactory procedure as long as there are no periodic effects or odd groupings in the invoices. If for some reason one customer always has invoices ending in 3, the sample would be biased and the method unacceptable.

Multi-Stage Sample Design

As the name implies, the selection of the sample by this method requires several stages. The method is most commonly used when the population is distributed over a wide geographical area. For example, the population might be the world-wide membership of the SHU Alumni Society. The first stage is to divide the population into a few clearly defined areas. In our example, these could be individual countries. The proportion of the sample allocated to these areas is determined by the proportion of the population in each area. Hence, if 80% of the Alumni membership was in England and Wales, and we were looking for a sample of 1000 members, then we allocate 800 sample members to England and Wales, as with stratified sampling. The next stage is to define some smaller areas, these might be local government districts and then companies within these districts. A sample is taken of the local government districts. Within the chosen districts, a sample of the companies is selected. Finally individual members are sampled from the chosen companies in the selected districts. With this method the actual sampling is quicker and more convenient.

Cluster sampling design

In previous designs we have selected items one at a time. In cluster sampling clusters of items are formed, which it is assumed are reasonably representative of the whole population. Clusters are then randomly selected and all of the items in the cluster are included in the sample. For example, suppose a large firm stores its invoices in batches of 50. If, in a year, there are 10,000 invoices generated, then there will be 200 batches. These batches can be used as clusters. Suppose the firm wants a sample of 300 invoices. This could be achieved by selecting 300 individual invoices randomly from the 10,000. Alternatively, the cluster sample design allows 6 clusters to be randomly selected from the 200 batches. This is a much easier and quicker method, but we must be sure that there is no bias within the batches.

There are a number of other random sample designs, but those described above are the most commonly quoted ones.

2.5.5 NON-RANDOM SAMPLE DESIGNS

For many surveys, especially in the area of Market Research, sampling frames do not exist. For example, if we wish to investigate housewives' views of a new product, it would be difficult to draw up a sampling frame for housewives.

Quota sampling is a commonly used sampling method in this situation. Initially the important characteristics of the target population are identified as for stratified sampling, for example, male/female, age groups, social class, etc. The sample is divided proportionately into these groups as far as it is possible, but from then on, it is left to the individual field workers to decide how to obtain the sample members. There is no question of identifying individuals first or choosing them randomly.

In auditing so called *judgement sampling* can be used, where the accountant uses a mixture of hunch, prior knowledge and judgement to select the sample. There is no attempt at stratification and randomness.

Any non-random procedure invalidates subsequent statistical analysis and should be avoided if possible.

2.5.6 STATISTICAL INVESTIGATIONS AND SURVEYS

If a statistical investigation, of the type referred to in the above, is to be carried out, then the following stages are normally required:

a) Define the objectives of the survey – what information is to be collected and for whom or what.
b) Define the target population
c) Decide on the sampling method
d) Choose an appropriate method of collecting the data – ensure that the method will yield the required information.
e) Carry out a pilot survey – this is a 'dress rehearsal' for the full survey and gives a guide to the suitability of the data collection method; particularly the adequacy of any questionnaire used. Amend the procedure as necessary.
f) Carry out the main survey
g) Analyse and present the results

2.6 INFORMATION REQUIREMENTS FOR EFFECTIVE DECISION MAKING

In order to make effective decisions about actions and strategies, the manager is required to know where the company is now and where it wants to be in the future. For example:

where you are	*where you want to be*
operational level	
location of orders within a manufacturing process	achieve the delivery date for the order
location of service engineers out on call	make the next service call which is required
tactical level	
the department's expenditure to date	work within the department's budget
which orders have been delivered so far this month	calculate the delivery performance this month
strategic level	
annual demand for each product	determine overall capacity
utilisation figures for each factory	set target utilisation figures

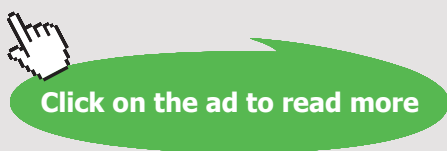**Table 2.1** Examples of decisions at various levels

Information has a hierarchical structure in that higher level requirements can be built up from lower levels. For example, annual total demand is an aggregation of product demands which is an aggregation of individual orders.

2.7   QUALITY OF INFORMATION

In order for decisions to be robust, the information needs to be:

relevant	increases knowledge, reduces uncertainty, usable for the intended purpose.	e.g. ages of passengers, dietary preferences for package holiday guests
accurate	measurement/counting is correct.	weights are accurate, error specified for estimates
complete	contains all of the relevant key aspects.	nothing missing from a record
reliable	data collected by appropriate and consistent means, based on evidence, amenable to cross-checking.	e.g. in customer satisfaction surveys
timely	produced in time for the decision to be effective	hourly, daily, weekly, etc.
communicated to the right person	in the hierarchy of decision making levels.	goes to the most appropriate person

Table 2.2 Useful characteristics of information

2.8   CONCLUSION

Overall the information must be fit for the purpose for which it is intended. In this way it can assist in informing the best possible decision at the correct level within the organisation.

2.9   CHAPTER REFERENCES

1. Morris, C, Quantitative Approaches in Business Studies, 8th edition, FT/Prentice Hall, 2012 (eBook, 7th edition 2008) chapter 3
2. Oakshott, L, Essential Quantitative Methods for Business, Management & Finance, 5th edition, Palgrave Macmillan, 2012 chapter 4
3. Wisniewski, M, Quantitative Methods for Decision Makers, 5th edition, FT/Prentice Hall, 2009 (eBook available) chapter 3

3   EXPLORING DATA; DERIVING INFORMATION

3.1   INTRODUCTION

In the previous chapter we discussed the nature of data and information and investigated ways in which data can be collected and managed by an organisation. We are now moving on to discuss ways of organising and presenting the data so that relevant information can be derived. Many of the ideas and techniques considered here are very simple but they must be applied appropriately if they are to result in useful and timely management information.

In the following sections, it is assumed that the required data have been collected by an appropriate method from an appropriate population or sample. We are now ready to process the data to find the required information.

3.2   STORING AND STRUCTURING DATA

Where there are substantial amounts of data generated within the organisation, it is likely to be stored in an electronic database.

A database is a structured way of holding data so that they can be found (retrieved) when required. The data are stored using a *table* structure. A table is like a labelled drawer in a filing cabinet and everything inside has the same data structure. The storage is designed to be *economic* (as little as possible replication of the data) and *efficient* (the data can be retrieved easily). Commercial database packages (e.g. Microsoft Access) divide the data into categories referred to as *entities.* The database contains a table for each entity. These tables hold data associated with like items. The rows in a table are referred to as **records** and the columns as **fields**. In general each record of the database relates to a different episode or item (a patient, an order, a customer) and holds data for that record only. Each field must be defined in terms of the *type* of data that it holds (numbers, text, dates, etc).

Example 3.1 – Structuring of data

Suppose you want to design a database for your collection of CDs. There are a number of possible entities that you could define. For example in your CD database, you might have a table holding data about albums and one holding data about artists. Your database will contain a number of records**,** one for each CD in your collection. You will have to define the fields that you require – the characteristics of the CD that you wish to identify.

album table

album title

artists' name(s)

date of issue

producer label

tracks

label table

label name

address

other details

artist table

artist name

history (e.g. date of formation, place of origin)

tracks table

track number

track title

type of music

run time

Figure 3.1 shows one way of structuring the data into different tables.

Once the tables have been specified they can be populated with the relevant data and linked together so that all the data for a given CD can be retrieved when required.

The data in one table can be *related* to data in another table by having a field in common – a *relational database.* See **Figure 3.2**.
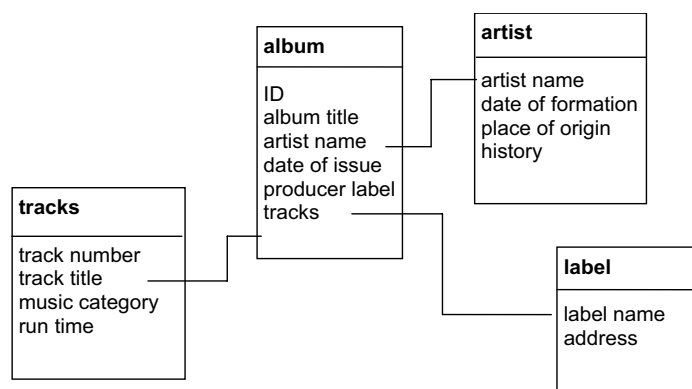
Figure 3.2 Example of relational database

Once the database has been set up and populated then it may be interrogated to retrieve the required information. For example, all the CD's that you have by a particular artist or of a particular type of music.

Although databases are commonly used, not all data will be stored in this way. In addition, it will not be possible to interrogate the database if we do not know what we are looking for. We also require more open-ended ways of exploring our data.

3.3   THE ORGANISATION OF DATA

Since the raw data (not in a database) may consist of a large jumble of unrelated facts, the first step in the processing is to organise the data in some way. There are various formats which can be used. The most appropriate one depends on the nature of the data, the information required and any future analysis which is to be done.

3.4   TABULATION

One of the most commonly used ways in which *numerical* data are organised is, once again, by means of tabulation. However, although this is the most frequently used method of presenting numerical information to the general public, many people find tables difficult to interpret. This factor should always be remembered when designing and using tables to convey information to others.

3.4.1 THE INTERPRETATION OF TABLES

There are some simple rules to follow when using a table prepared by someone else, to ensure that you understand correctly to information given.

1. First look at the row and column headings. Choose any number in the table and describe in words to yourself, as clearly as possible, what that number means. Then question whether or not the explanation makes sense.
2. If any of the figures in the table are percentages, determine of what figure the percentage has been taken.
3. If possible, think out what figures you would expect as entries in the table and compare the given ones with the expected figures.
4. Interpretation is usually helped by comparison. For example, look at changes through time, or differences between areas.

3.4.2 DESIGNING A TABLE

The design of a table depends on its purpose, since different users of the data may require different features to be highlighted. There is no point in giving more detail than the reader needs. There are several stages in the design of a table:

a) Decide on the categories to be used in the table. The simplest is a two-way classification, but there may also be sub-categories.
   Categorisation may be by:
   • attributes (e.g. eye colour – blue, brown, black, …)
   • measured variables (e.g. Less than £100, £100 – £200, …)
   • time periods

   For example:

Sporting interests	Age (years)					
	Under 25		25–40		Over 40	
	Male	Female	Male	Female	Male	Female
Participates						
Watches						
None						

Table 3.1 Examples of decisions at various levels

b) Decide on the level of accuracy required, and the units (e.g. £, £'000, m, 'nearest 10 miles', km, etc.).

c) Plan the framework of the table, with some categories across the page and some down, with appropriate sub-divisions if needed. It is generally thought that the table is easier to read if there are more rows than columns. Each row and column should be briefly but clearly labelled.

d) Row and column sub-totals may be given, possibly in darker print.

e) Percentages may be helpful. It should always be clear of which figure the percentage has been taken.

f) Each table should have a comprehensive title and the source of the data should be quoted.

Tabulations of the type discussed above, are not always appropriate. However, it is almost impossible to see any overall pattern in a large set of data, without using some method of ordering or classification.

We will look next at various methods of organising data into a format from which information can be derived.

3.5   ORDERING DATA

3.5.1 RANKING

If the data set is small, the necessary information may be obtained by simply ranking the values i.e. arranging the data in ascending or descending order of size.

Example 3.2 – Ranking data

Ten people have applied for a basic clerical job in a firm. As part of the interviewing procedure, they are given a standard intelligence test. The results of the test are given below. Rank the values.

Applicant	A	B	C	D	E	F	G	H	I	J
IQ	104	100	105	101	95	103	105	100	99	100

Table 3.2 IQ Data

Solution:

The intelligence test scores for 10 job applicants in rank order:
95, 99, 100, 100, 100, 101, 103, 104, 105, 105
We can now see the pattern, if any, in the data.

3.5.2 STEM-AND-LEAF

A simple way of getting a 'feel' for a set of numbers is Tukey's stem-and-leaf display. This diagram requires the use of the raw data set and enables the specific values to be viewed along with the overall shape of the distribution.

Example 3.3 – A stem-and-leaf display

Mulitpress plc, produces household cleaning products, including washing powder. It is found that, for a reliable product, the % moisture content of washing powder must be carefully controlled. The following figures give the % moisture content in batches of powder produced during different shifts at the factory.

Batch	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
% moisture	7.2	6.5	5.3	3.9	6.7	5.9	6.0	6.5	6.9	4.3	6.3	7.6	5.9	5.7	5.5

Table 3.3 Day shift data (15 batches)

Batch	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
% moisture	5.9	5.6	3.3	4.4	6.9	6.1	5.6	2.9	6.4	8.8	8.1	7.1	7.1	4.8	4.0	2.5	6.4	6.5

Table 3.4 Night shift data (18 batches)

Construct a stem-and-leaf display for the day shift.

Solution:

The 'stem' of the diagram is made up of the first digit of the values. The 'leaf' is then made up of the second digit, **arranged in rank order**.

% moisture content in 15 batches of washing powder produced by day shift workers

Units, %	Tenths, %					
3	9					
4	3					
5	3	5	7	9	9	
6	0	3	5	5	7	9
7	2	6				

Table 3.5 Example of stem and leaf diagram

It is now easy to see which are the biggest and smallest values and where the concentrations of values lie. The overall shape of the distribution may be examined.

Constructing a stem and leaf diagram using only the leading digit on the stem can lead to the loss of valuable information. This may be avoided by using each stem value twice and associating the units 0, 1, 2, 3, 4 with the first section and the units 5, 6, 7, 8, 9 with the second section. The stem values are distinguished by, for example:

2*   for the units 0, 1, 2, 3, 4 and
2.   for the units 5, 6, 7, 8, 9

Example 3.4 – A more detailed stem-and-leaf display

Over the last 30 weeks, the turnover per week of a small shop has been as follows:

4300	4100	7400	6100	7900	6000	7100	6900	6300	7700
7000	6600	6400	7100	7100	7400	5600	7400	4100	7100
6300	5700	5700	6800	6400	6200	5900	5200	4000	7600

Table 3.5 Turnover per week (£)

Construct a stem-and-leaf display.

Solution:

4*	0	1	1	3			
4.							
5*	2						
5.	6	7	7	9			
6*	0	1	2	3	3	4	4
6.	6	8	9				
7*	0	1	1	1	1	4	4	4
7.	6	7	9				

Stem value is in £'000

Leaf value is in £'00

Table 3.6 Turnover (£ per week) of a shop during the last 30 weeks

The nature of the distribution of the weekly turnover of the shop is now apparent.

3.6 FREQUENCY DISTRIBUTION

In larger data sets, it is likely that values will be repeated. If this is the case, then the overall pattern in the data set may be revealed by the construction of a **frequency distribution**.

Example 3.5 – Construction of a frequency distribution

In a certain area, 50 households were surveyed. The following data give the occupancy of each household.

4	7	4	1	4	2	3	6	3	5
6	3	4	9	12	1	3	4	2	2
1	1	3	8	1	1	4	2	3	4
3	2	1	4	6	5	6	1	2	3
4	4	4	1	4	2	3	5	4	4

Table 3.7 50 households occupancy data

Construct a frequency distribution for these data.

Solution: The variable is 'the number of occupants in the household'.

The minimum value is 1 and the maximum value is 12, hence, in this data set, the variable takes values from 1 to 12 inclusive. List these possible values.

The number of households with a given number of occupants is the 'frequency'. Determine the frequency of each value using a tally system[1]. It is necessary (and desirable) to read through the data set once only to reduce the risk of error. See **Table 3.8**.

Number of occupants in the household	Tally			Number of households (frequency)
1	~~1111~~	1111		9
2	~~1111~~	11		7
3	~~1111~~	1111		9
4	~~1111~~	~~1111~~	1111	14
5	111			3
6	1111			4
7	1			1
8	1			1
9	1			1
10				0
11				0
12	1			1
Total				50

Table 3.8 Occupancy of households

The total number of values should always be checked.

We now have a summary of the original data and we can see the overall pattern in the distribution of the values. The long, sparsely populated tail does not add a great deal of information; hence we may prefer to shorten the distribution by combining some of the higher values, **Table 3.9**.

Number of occupants in the household	Number of households
1	9
2	7
3	9
4	14
5 or 6	7
7 and over	4
Total	50

Table 3.9 Occupancy of households

3.7   GROUPED FREQUENCY DISTRIBUTION

In the household occupancy example above, we noted that in some cases, the grouping of values is helpful. This procedure becomes essential if the variable is continuous.

Example 3.6 – Construction of a grouped frequency distribution

Refer to the data in **Example 3.4** (Turnover of a shop)

The minimum value is £4100 and the maximum value is £7900. Choose groups which cover this range as closely as possible. The data set should be read through once only and a tally used to assign the values to the groups. See **Table 3.10**.

Turnover (£ per week)	Tally		Number of weeks (frequency)
4000 but less than 4500	1111		4
4500 but less than 5000			0
5000 but less than 5500	1		1
5500 but less than 6000	1111		4
6000 but less than 6500	1111	11	7
6500 but less than 7000	111		3
7000 but less than 7500	1111	111	8
7500 but less than 8000	111		3
Total			30

Table 3.10 Weekly turnover of a shop

The groups are usually referred to as **classes** or **class intervals**. The range of values included in a class is referred to as the **class width**. The minimum and maximum values included in the class are referred to as **class boundaries**. The numbers used to specify the class are called the **class limits**.

For example: In **Example 3.6**, the first class is '4000 but less than 4500' (£s)

        the class width is £500,   i.e. (£4500 minus £4000)
        the class boundaries are: £4000 and £4500
        the class limits are also: £4000 and £4500

In a different example (e.g. **Example 3.7**), we might have classes specified as:

Petrol consumption (mpg)	Number of cars
30–31	0
32–33	2
etc.	

Table 3.11 Petrol consumption data

In this case, for the first class '30 − 31' mpg:

the class limits are: 30 and 31

the class boundaries are: 29.5 and 31.5 (up to) all values from 29.5 up to but not including 31.5 will be allocated to '30 − 31'

the class width is 2 mpg, 31.5 − 29.5 mpg

Example 3.7 – Constructing a grouped frequency distribution

The petrol consumption of each of 35 cars of a given type was measured using a standard test. The results are shown below.

29.8	39.2	38.3	34.6	37.0	35.9	40.6	35.8	35.3	39.7	38.3	35.5
38.5	37.7	36.4	35.3	38.8	34.7	38.6	32.6	36.9	36.7	38.0	33.3
36.1	37.8	37.3	35.4	38.6	37.1	36.4	38.5	37.9	36.2	37.5	

Table 3.12 Petrol consumption data (mpg)

Construct a grouped frequency distribution.

Solution: The minimum value is 29.8 mpg. The maximum value is 40.6 mpg.

Petrol consumption (mpg)	Tally			Number of cars
under 30	1			1
30 but less than 32				0
32 but less than 34	11			2
34 but less than 36	1111	111		8
36 but less than 38	1111	1111	111	13
38 but less than 40	1111	1111		10
40 and over	1			1
Total				35

Table 3.13 Petrol consumption data (mpg)

Note the use of open ended classes. This is a useful way of avoiding sparsely populated classes at the ends of a distribution, but it must be remembered that information is lost by the use of open ended classes.

General guidelines for the choice of class intervals:

   a) they should cover the whole range of the values, but as closely as possible;
   b) they should be non-overlapping, unambiguous and simple (use multiples of 2 or 5);
   c) they should be equal in width if this is sensible.

If unequal class widths are to be used, then the widths should bear simple numerical relationships to each other. For example:

Class Interval	Class Width	
0 but less than 10	10	
10 but less than 30	20	twice the width of class 1
30 but less than 50	20	twice the width of class 1
50 but less than 100	50	five times the width of class 1
100 but less than 200	100	ten times the width of class 1

Table 3.14 Unequal class widths example

   d) The classes should reflect 'natural groupings' where they exist, e.g. if the variable is age of person, 'natural groupings' might be school children, older teenagers, young adults, …, pensioners. For **Example 3.6**, group together values which constitute a similar level of turnover, e.g. very low, low, average, etc.
   e) The number of classes should be carefully chosen. If there are too many classes, some classes will be too sparsely populated and the overall pattern in the distribution may not become clear. If there are too few classes, important information may be obscured. If in doubt, a very rough rule of thumb is to make the number of classes about equal to the square root of the number of values.

3.8   CUMULATIVE FREQUENCY DISTRIBUTION

In many cases, the information drawn from a data set is required in the form of the number or proportion of observations which are **less than** (or, alternatively, more than) a given value.

Example 3.8 – Construction of a cumulative frequency distribution

Refer to **Example 3.4** (Weekly turnover of a shop). Use the grouped frequency distribution to construct a cumulative frequency distribution.

Solution: See **Table 3.15**.

Turnover (£ per week)	Number of weeks with less than turnover given (Cumulative frequency)
less than 4000	0
less than 4500	4
less than 5000	4
less than 5500	5
less than 6000	9
less than 6500	16
less than 7000	19
less than 7500	27
less than 8000	30

Note: The frequency is accumulated against the upper class boundary.

Table 3.15 Unequal class widths example

Example 3.9 – Constructing a cumulative frequency distribution

Draw up a 'less than' cumulative frequency distribution from the grouped frequency distribution constructed in **Example 3.7** (petrol consumption).

Solution: See **Table 3.16**.

Petrol consumption (miles per gallon)	Number of cars with less than consumption given (Cumulative frequency)
less than 30	1
less than 32	1
less than 34	3
less than 36	11
less than 38	24
less than 40	34
One car achieved 40 or more mpg	

Table 3.16 Unequal class widths example

Note the difficulty with the last open-ended class.

Download free eBooks at bookboon.com

3.9   PERCENTILES

The level below which 'y' percent of the data values fall is called the y percentile. For example, the 30th percentile is the level below which 30% of the data values fall.

Certain percentiles are used regularly. These are:

25th percentile – the first, or lower, quartile, denoted by Q_1

50th percentile – the median, denoted by M

75th percentile – the third, or upper, quartile, denoted by Q_3

Example 3.10 – Determination of percentiles

Return to the **Example 3.5** (Occupancy of households). Construct a cumulative frequency distribution and use it to evaluate the 25th, 50th and 75th percentiles.

Solution: See **Table 3.17**.

Number of occupants per household	Number of households with less than occupants given (cumulative frequency)
less than 1	0
less than 2	9
less than 3	16
less than 4	25
less than 5	39
less than 6	42
less than 7	46
less than 8	47
less than 9	48
less than 10	49
less than 11	49
less than 12	49
less than 13	50

Table 3.17 Occupancy of households – cumulative frequency

The 25th percentile is the value of the item below which 25% of the values fall. In this example, the first 25% of the values gives a cumulative total of 12.5 households. This total is reached between the cumulative frequencies of 9 and 16, therefore the 25th percentile is one of the observations which has the value 2 occupants per household.

Similarly, the 50th percentile is the value below which 50% of the observations fall. From the above table, it can be seen that 50% of the values lie below 4 (the value of the 25th item). The median is therefore 3 occupants per household.

The 75th percentile is the value of the 37.5 observation. This has the value 4 occupants per household.

Percentiles may be estimated from a grouped frequency distribution. The method will be described in a later section

3.9.1 RELATIVE AND PERCENTAGE FREQUENCIES

Relative or percentage frequencies are required when comparisons are to be made between data sets in which there are different *total* numbers of observations.

Example 3.11 – Percentage frequencies

The times taken to satisfy orders for similar types of products by two companies are given below.

Time taken weeks	Number of orders	
	Company A	Company B
less than 1	1	4
1 but less than 2	1	12
2 but less than 3	3	25
3 but less than 4	7	4
4 but less than 5	5	2
5 and over	3	3
Total	20	50

Table 3.18 Orders for two products

Since the total number of orders for each company is different, the frequencies cannot be compared directly. Either the relative or percentage frequencies must be calculated.

$$\text{Relative frequency} = \frac{\text{frequency in the class}}{\text{total frequency}}$$

$$\text{Percentage frequency} = \text{relative frequency} \times 100$$

Time taken weeks	Percentage of orders	
	Company A	Company B
less than 1	5	8
1 but less than 2	5	24
2 but less than 3	15	50
3 but less than 4	35	8
4 but less than 5	25	4
5 and over	15	6
Total	100	100

Table 3.19 Orders for two products

The differences in the patterns of the two distributions can now be assessed.

3.10 PRESENTING INFORMATION PICTORIALLY

The organisation of a data set, using one of the formats discussed above, will in itself yield useful information about the pattern and trends in the data. However, it is often easier to identify the features of a data set if it is displayed pictorially.

There are many different types of pictorial display. The choice is dictated by the type of data, the purpose of the display and the level of statistical sophistication required.

In the following sections, we will discuss a number of different types of display. We will consider the method of construction, when it is appropriate to use the display and how Microsoft Excel can be employed to construct the displays.

3.10.1 CHARTS AND DIAGRAMS TO DISPLAY CATEGORY DATA

The most commonly used charts to display category data are:

    a) pictograms
    b) bar charts – simple, compound and component
    c) pie charts

There are other possibilities but unusual styles should be used with care.

If you are unfamiliar with these charts, use one of the references given to read more about them. A summary only is given below.

You are expected to be able to **use Microsoft Excel** to construct bar charts and pie charts. See the Workshop Worksheet for this Study Block. It will be assumed in the following that the actual charts are produced in Microsoft Excel.

3.10.2 CONSTRUCTION OF CHARTS AND DIAGRAMS

a) **Pictogram** All of the symbols in the diagram must be of the same size. The variation in the frequency of occurrence is shown by varying the **number** of the symbols used, *not* by varying the size of the symbol

b) **Bar chart** A bar chart is suitable for **category data** only. The scale on the frequency axis must *always* include zero i.e. do not suppress the zero.

c) **Pie chart** This diagram is suitable only if we have a single variable which is sub-divided. The pie chart will then illustrate the relative size of the sub-divisions. It is, therefore, not appropriate to a pie chart to illustrate, for example, a time series data, i.e. a variable for which the value varies over time.

Example 3.12 – Illustration of a pictogram, a bar chart and a pie chart

The following table gives the Gross Domestic Product of Wyeland, broken down into three categories, for three separate years.

Category	GDP (£'00 m)		
	2013	2014	2015
Public Expenditure	3	5	15
Manufacturing	12	15	30
Other	18	25	55
Total GDP	33	45	100

Table 3.20 Gross domestic product

a) Illustrate the variation in the total GDP using a pictogram.
b) Illustrate the variation in the total GDP using a bar chart.
c) Illustrate the composition of the GDP using:
   i) a compound bar chart;
   ii) a component bar chart.
d) Illustrate the make-up of the GDP for 2015 using a pie chart.

Solution:

a) Pictogram. See **Figure 3.3**.

Key: (£) = 10 £'00m
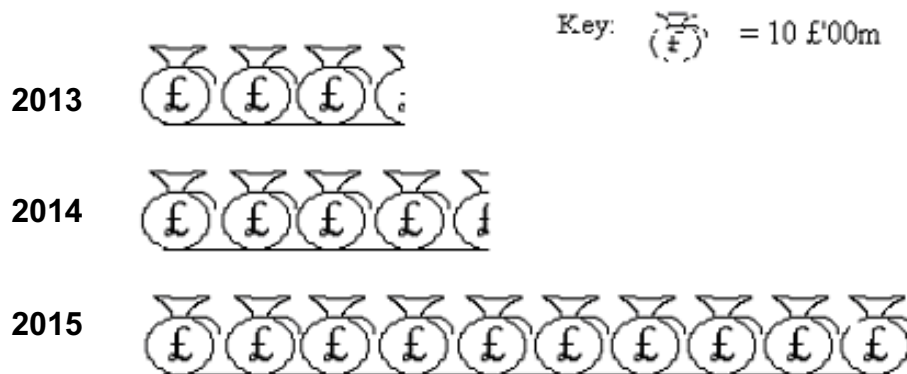
2013 (£)(£)(£)(

2014 (£)(£)(£)(£)(

2015 (£)(£)(£)(£)(£)(£)(£)(£)(£)(£)

Figure 3.3 The total GDP of Wyeland in 2013, 2014 and 2015
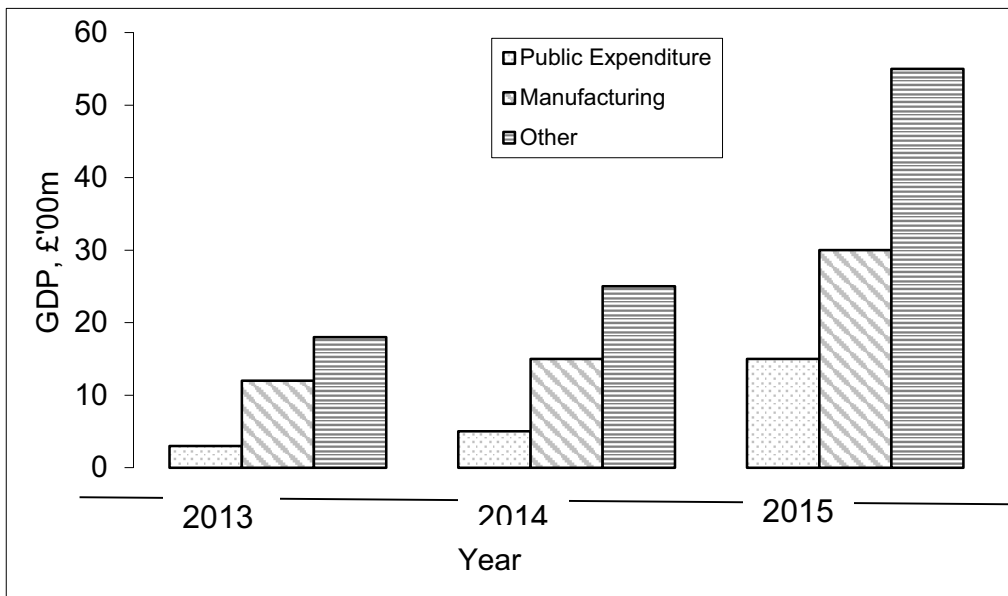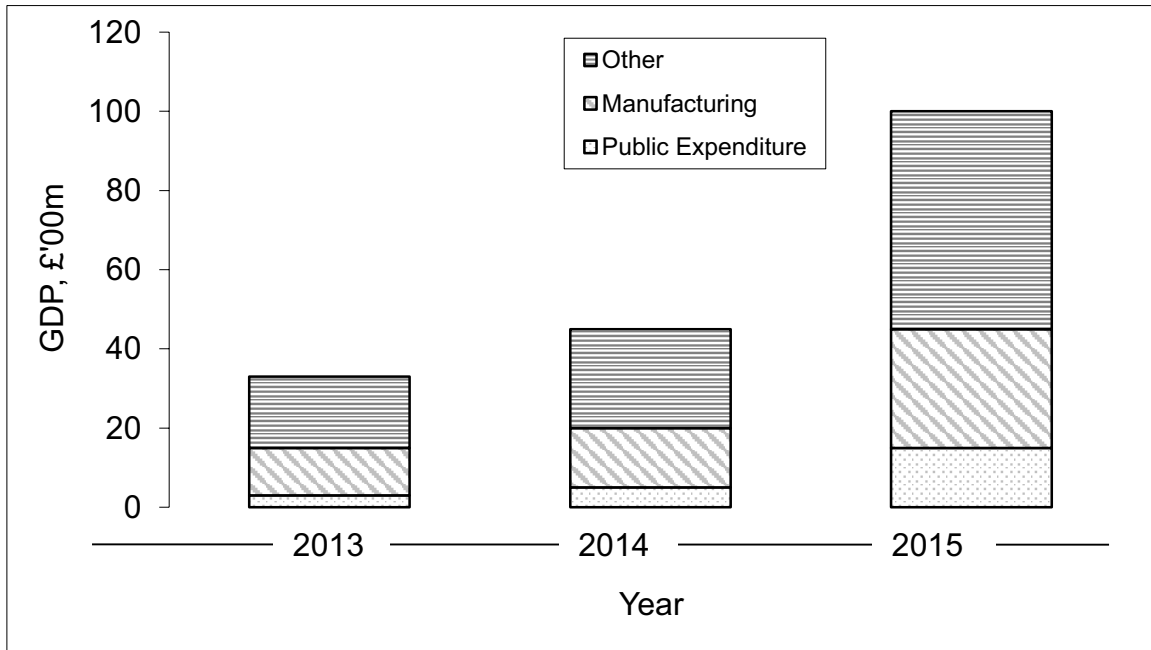
b) Bar chart. See **Figure 3.4**.
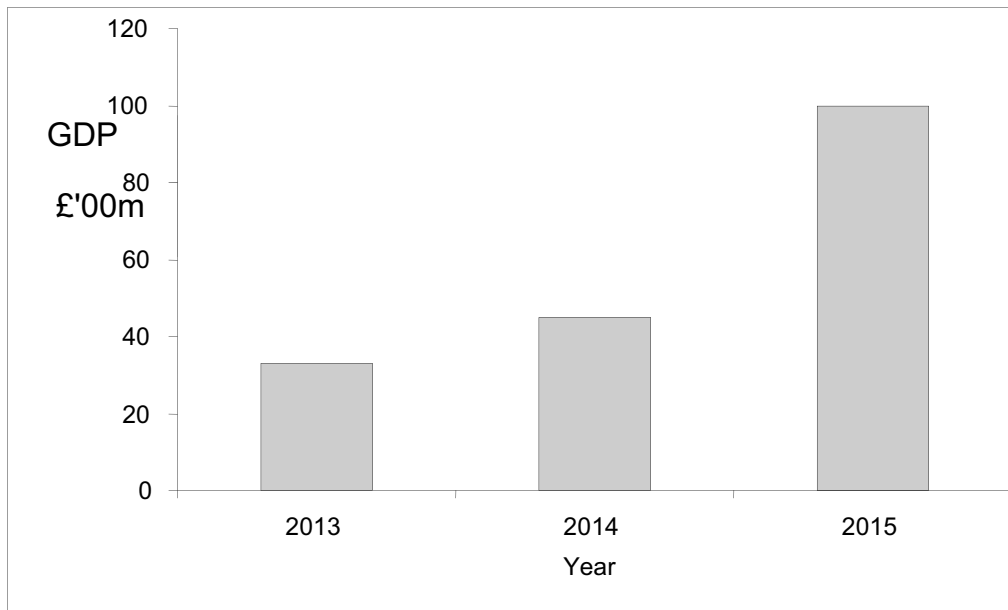
Figure 3.4 The total GDP of Wyeland in 2013, 2014 and 2015
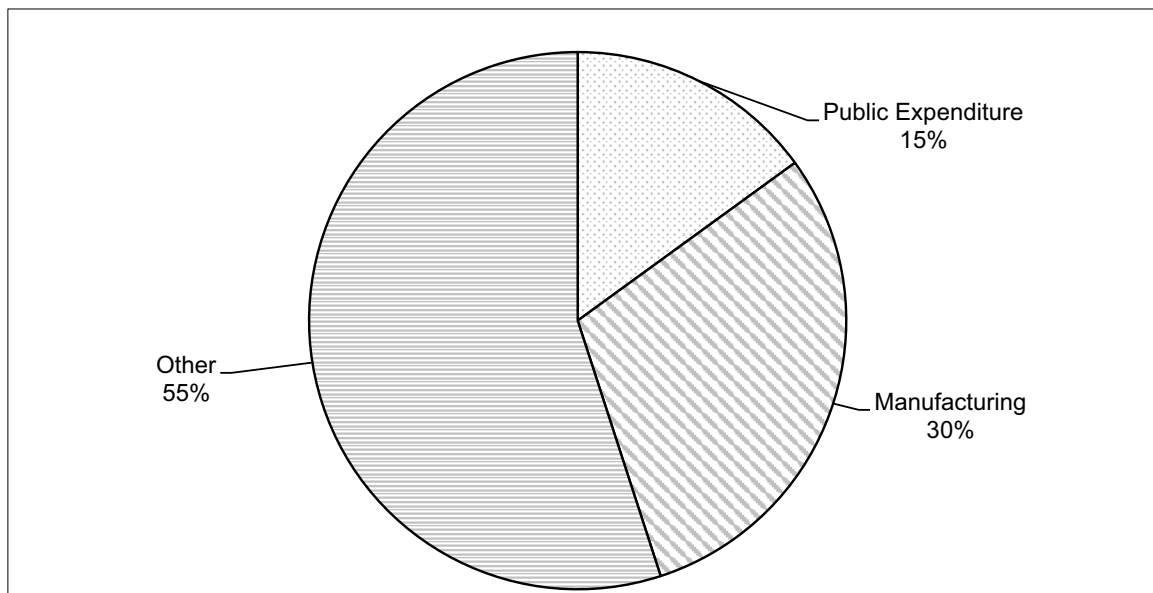
c) i) Compound bar chart. See **Figure 3.5**.

Figure 3.5 The composition of the GDP of Wyeland in 2013, 2014 and 2015

ii) Component bar chart. See **Figure 3.6**.

Figure 3.6 The composition of the GDP of Wyeland in 2013, 2014 and 2015

d) Pie chart. See **Figure 3.7**.

Figure 3.7 The composition of the GDP of Wyeland in 2015

Calculating Angles

Category	2015	Angle Calculation
Public Expenditure	15	15/100 x 360 = 54
Manufacturing	30	30/100 x 360 = 108
Other	55	55/100 x 360 = 198
Total GDP	100	360

Table 3.21 Calculating angles

3.10.3 INTERPRETATION OF CHARTS AND DIAGRAMS

It is very easy to produce diagrams which mislead the user. Great care should always be taken to ensure that any diagram which you draw conveys an accurate impression of the data. By the same token, you should also examine critically, diagrams produced by others, in order to avoid being misled yourself.

3.10.4 DIAGRAMS TO DISPLAY NON-CATEGORY DATA

A frequency distribution may be illustrated in a variety of ways. The method chosen depends on the exact nature of the information required. None of the charts or diagrams considered so far will be suitable. This is particularly the case if the variable concerned is continuous, for example, time or length.

If the data are organised into a frequency distribution or a grouped frequency distribution, then a histogram or a frequency polygon (essentially the same diagram) may be used to illustrate the distribution. See **Figures 3.8–3.10**.

If the data are organised into a cumulative frequency distribution, then an ogive or a cumulative frequency polygon (same thing) may be used to illustrate the distribution. See **Figure 3.11**.

A further useful diagram is the box plot (Box and Whisker plot). This is constructed using the maximum and minimum values of the data and three key percentiles. See **Figure 3.12**.

Apart from the cumulative frequency polygon, these diagrams are not easy to construct using Microsoft Excel, hence the following sections are provided for information only at this stage.

3.10.5 THE HISTOGRAM

The histogram is one of the most commonly used diagrams for the illustration of a frequency distribution. At first glance this diagram is similar to a bar chart, but it differs from a bar chart in two crucial ways.

a) The scale across the page is a **continuous, linear** scale, not a series of categories. The width of each bar is the width of the corresponding class in the frequency distribution.

b) The **area** of the bar (*not* the height) is proportional to the frequency of occurrence in the class.

As a consequence of these differences:

a) The variable plotted up the page, is always a **frequency density** e.g. frequency per unit interval. For example, frequency per £, frequency per £10, frequency per 5 years, etc.

If we denote class frequency by f, the height of the bar by h and the width of the class by w, then:

$$f \propto h \times w \; (\propto \text{ means 'proportional to')}$$

hence the height of the bar (variable plotted up the page)

$$h \propto \frac{f}{w}$$

b) There are no gaps between the bars.
c) When the frequency distribution contains classes of different widths, the height of the bars must be adjusted to keep the area of the bar proportional to the frequency.

Note: A little ingenuity is required in order to use Microsoft Excel to produce a correctly drawn histogram. The 'Histogram' tool on Microsoft Excel does not produce a correct histogram.

Example 3.13 – The construction of a histogram

Refer to **Example 3.6** (weekly turnover of a shop). Illustrate the grouped frequency distribution by a histogram.

Solution: All of the classes are the same width (£500), so we will choose to plot 'Frequency per £500' against 'turnover/week'. This means that, effectively:

$$\text{class frequency} \propto \text{height of bar}$$
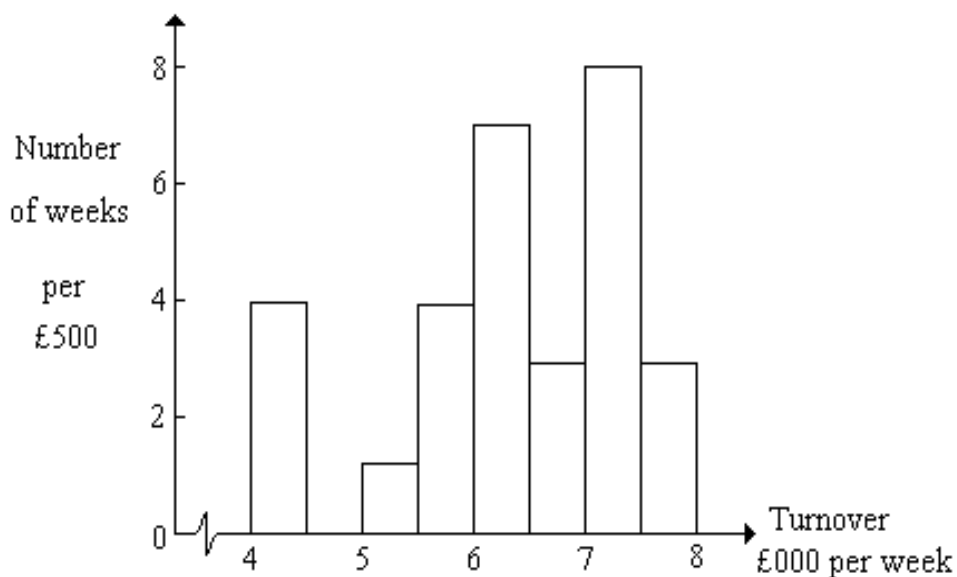
See **Figure 3.8**.

Figure 3.8 Weekly turnover of a shop over the last 30 weeks

Example 3.14 – To construct a histogram

Refer to the **Example 3.7** (petrol consumption). Illustrate the grouped frequency distribution by a histogram.

To illustrate the open-ended classes, simply write on the diagram that there is 1 car with a consumption below 30 mpg and 1 car with a consumption over 40 mpg.
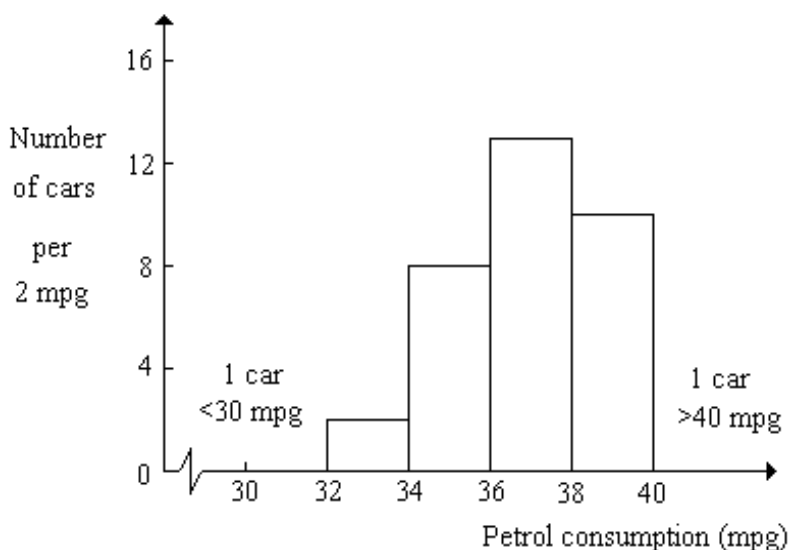
Figure 3.9 Petrol consumption of 35 cars of a given type

3.10.6 FREQUENCY POLYGON

This is a modification of the histogram and may be used in its place. It is constructed by plotting the *frequency density* at the mid-point of the class. The plotted points are joined, dot-to-dot, by straight lines. The polygon should be a closed figure, therefore the first and last points are joined to the axis (zero frequency density) at the adjacent mid-points.

Example 3.15 – To construct a frequency polygon

Refer again to the **Example 3.6** (the weekly turnover of a shop). Illustrate the grouped frequency distribution by a frequency polygon.

Solution:

In this example, the classes are all of width £500, therefore we will choose to plot Frequency per £500 at the class mid-point.

Turnover (£ per week)	Number of weeks (frequency)	Class mid-point, £ per week
4000 but < 4500	4	4250
4500 but < 5000	0	4750
5000 but < 5500	1	5250
5500 but < 6000	4	5750
6000 but < 6500	7	6250
6500 but < 7000	3	6750
7000 but < 7500	8	7250
7500 but < 8000	3	7750

Table 3.22 Turnover data

The polygon is closed at the lower end by joining the first plotted point to the axis at £3750. The polygon is closed at the upper end by joining the last plotted point to the axis at £8250.
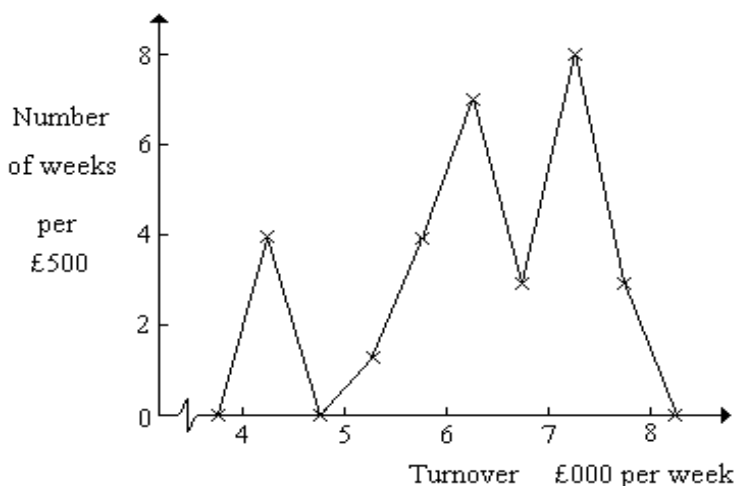
Figure 3.10 Weekly turnover for a shop for the last 30 weeks

The frequency polygon is a useful way of comparing the shapes of two or more distributions. The polygons should be superimposed on the same axes, provided that the total frequencies are the same or the frequencies of each distribution are converted to percentages.

3.10.7 CUMULATIVE FREQUENCY POLYGON OR OGIVE

This diagram is used to illustrate a cumulative frequency distribution. It enables us to estimate the number of observations in the distribution which fall below a certain value.

To construct the ogive, plot the cumulative frequencies against the **upper class boundaries**. The plotted points are then joined, dot-to-dot by straight lines.

The polygon is closed at the lower end by joining the first plotted point to the axis at the lower boundary of the first class. The polygon is closed at the upper end by joining the last plotted point directly to the axis, i.e. at the upper boundary of the last class.

Example 3.16 – To construct a cumulative frequency polygon

Refer to **Example 3.9** (weekly turnover of a shop). Illustrate this cumulative distribution by an ogive. Estimate in how many weeks the turnover was:

  a) less than £5250 per week;
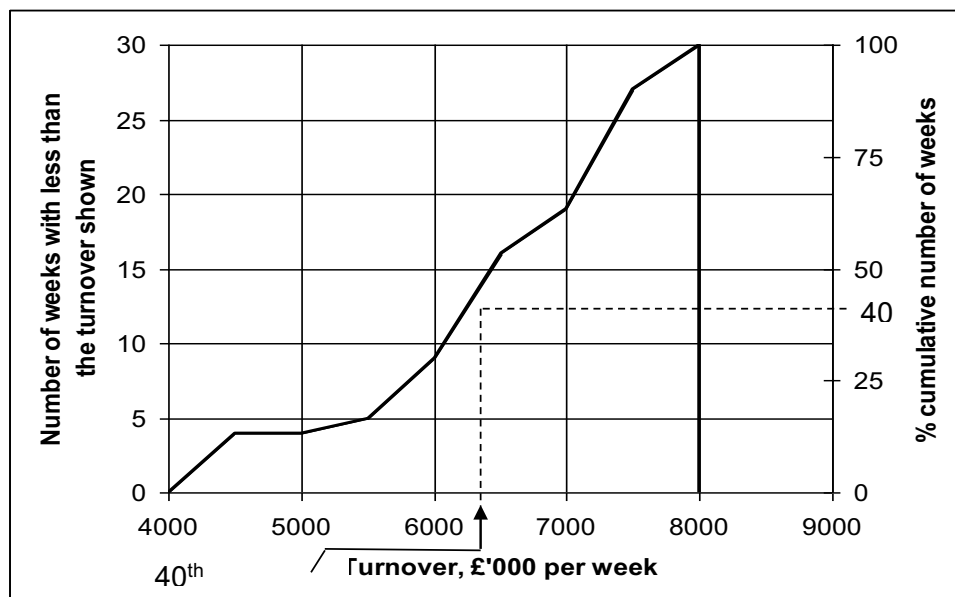 b) more than £6750 per week;
  c) between £5250 and £6750 per week.

Solution: The cumulative frequency distribution is repeated here:

Turnover (£ per week)	Number of weeks with less than turnover given (cumulative frequency)
less than 4000	0
less than 4500	4
less than 5000	4
less than 5500	5
less than 6000	9
less than 6500	16
less than 7000	19
less than 7500	27
less than 8000	30

Table 3.23 Turnover data

The ogive is given in **Figure 3.11**.

Figure 3.11 Ogive of the weekly turnover of a shop over the last 30 weeks

a) From the graph, the turnover was less than £5250 per week in a total of 4.5 weeks.
b) From the graph, the turnover was less than £6750 per week in a total of 17.5 weeks, therefore the turnover was more than £6750 per week in a total of
c) From a) and b), the turnover is between £5250 and £6750 per week in a total of (17.5 − 4.5) = 13 weeks.

It is often useful to construct an ogive using percentage cumulative frequencies. Percentiles can be estimated easily from the percentage ogive.

The graph in **Figure 3.11** can be converted to a percentage ogive by adjusting the scale up the page. See the right hand side of the figure. We can use this scale to estimate, for example, the 40th percentile

The 40th percentile (the value below which 40% of the observations lie) is £6220 (to the nearest £10).

In this type of diagram, no adjustments are needed if the class widths are unequal. However, care is needed to ensure that the scale marked across the page is linear.

3.10.8 BOX PLOT

The box plot is a graphical display which illustrates the behaviour of the data values in the middle of the distribution and also indicates the existence and position of any outliers. The position of the outliers may be seen relative to the middle of the distribution.

The box plot is a useful way of comparing two or more distributions. It enables the general shape of the distribution to be assessed; skewness is immediately apparent. However, if a distribution is bi-modal (i.e. has two peaks), this information is lost.

In this book, we will use an approximate method of construction, based on the median, the two quartiles and the range of the values in the data. In most cases, there will be no discernible error arising from this approximation.

The box plot is constructed by drawing a box, the ends of which correspond to the quartiles (hence the box covers the middle 50% of the distribution); the median is marked within the box. A line is drawn from each end of the box to the minimum and maximum values in the distribution, respectively.

Example 3.17 – To construct a box plot

Construct a box plot for the data in **Example 3.4**. (Weekly turnover of a shop)

Solution: From the original data set, the minimum turnover is £4000 per week and the maximum turnover is £7900 per week.

From the ogive in **Figure 3.11**, the lower quartile, Q1, is estimated as £5800, the upper quartile, Q3, as £7200 and the median as £6400.

We can now construct the box plot. See **Figure 3.12**.
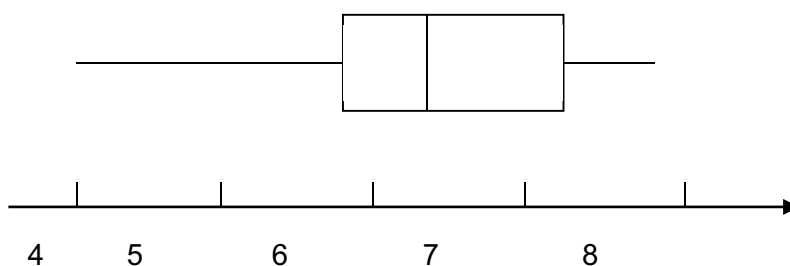
Figure 3.12 Weekly turnover of a shop over the last 30 weeks (£'000 per week)

3.11 CONCLUSION

The appropriate selection of the charts greatly enhances the ability to convey correct information. Especially, in decision making scenarios, it is the selection of correct diagrams and charts that will determine the effectiveness of decisions made at different organisational levels. Therefore it is crucial that care must be taken to select, construct and display correct charts/diagrams when presenting management information.

3.12 CHAPTER REFERENCES

1. Morris, C, Quantitative Approaches in Business Studies, 8th edition, FT/Prentice Hall, 2012 Chapter 5 (eBook, 7th edition 2008)
2. Oakshott, L, Essential Quantitative Methods for Business, Management & Finance, 5th edition, Palgrave Macmillan, 2012, Chapter 5

# 4   SUMMARISING DATA

4.1   INTRODUCTION

In the previous chapter we have looked at ways of organising and presenting raw data in order to extract useful information. This information has been largely descriptive. For example, we can say whether the distribution of the data is symmetrical or skewed, whether there are any extreme values, what the range of values is, etc.

In this chapter, we will look at methods of summarising the data in order to extract further information. It is possible to determine two or three values which represent the distribution of the complete data set. These values are referred to as measures. The summary measures may help us to compare different data sets or to carry out more sophisticated statistical analysis.

4.2   SOME FUNDAMENTAL CONCEPTS FOR SUMMARISING DATA

We will go through a range of fundamental concepts that will build up and consolidate your understanding of the importance of summarising data in general.

Look at the distributions shown in **Figure 4.1**. These distributions might arise from a machine used to fill three different sizes of packets with tea.
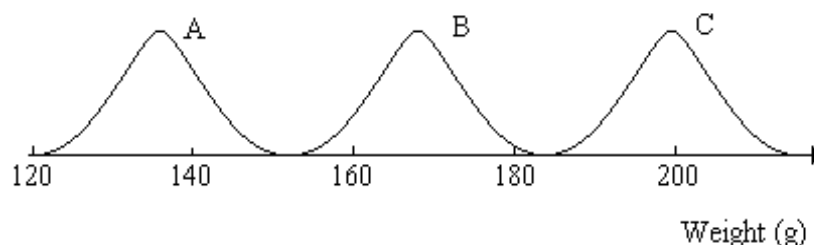
Figure 4.1 Typical distributions of weights delivered by a filling machine

Each distribution is similar in size and shape, but each one covers a different range of values. We could therefore summarise each distribution by finding a measure which describes the size of the values in that distribution – in other words we require a **typical** value; one which is representative of the data set as a whole. This single value is often referred to as a **measure of location**.

Now consider the two distributions shown in **Figure 4.2**. These distributions might arise from two different machines used to fill the *same desired size* of packets with tea.
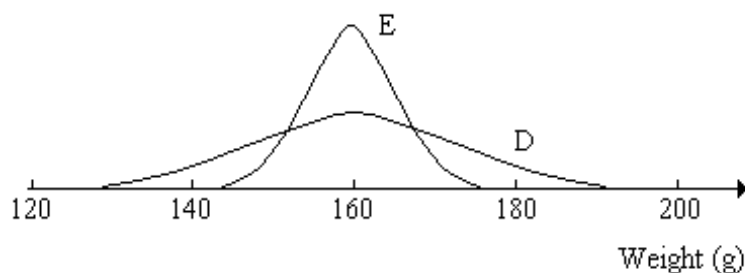
Figure 4.2 Typical distributions of weights delivered by two filling machines

The measure of location which we choose for distribution D would be very similar to the one chosen for distribution E, but the two distributions are clearly very different. As well as a representative measure, we must also consider how spread out the data set is relative to that measure. In other words, we require a **measure of spread** (sometimes called a **measure of dispersion**) as well as a measure of location.
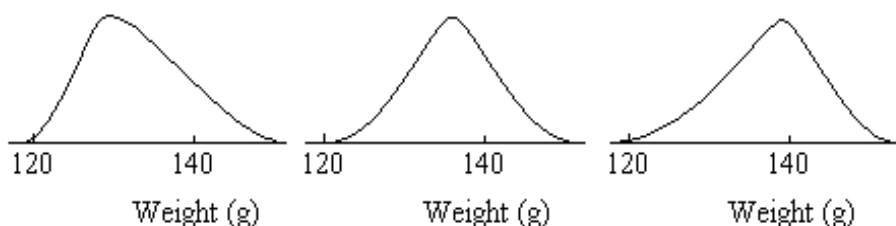
The final aspect of a distribution which is important is the degree of symmetry. Look at the three distributions of **Figure 4.3**:

Figure 4.3 Distributions with varying skewness

It is possible to calculate measures to describe the degree of symmetry in a distribution, but as far as this module is concerned, we will simply describe the symmetry in words – 'symmetrical', 'peak to the left', 'peak to the right'.

We will now look at each of these three properties in more detail.

4.3   MEASURES OF THE TYPICAL VALUE OR AVERAGE (MEASURES OF LOCATION)

There are a number of different ways of describing the 'typical' value of a data set. The method chosen depends on the nature of the data, the way in which we wish to use the measure and the amount of calculation we are prepared to do. The measures used most often are:

The **mode** – the mode is the most frequently occurring value(s) in the data and is denoted by **m** in these notes.

The **median** – this is the 50th percentile, i.e. the value below which 50% of the values fall when they are arranged in order of size. It is denoted in this chapter by **M**.

The (arithmetic) **mean** – the arithmetic mean is the sum of all of the values in the data set divided by the number of data values. It is denoted by $\overline{x}$. It is this measure which is commonly referred to as the 'average'.

4.3.1 THE CALCULATION OF THE MODE, MEDIAN AND MEAN FOR NON-FREQUENCY DATA

Example 4.1 – Calculation of the mode, median and mean

Shefftex plc employs 90 machine operators. The HR Director is interested in looking at the problem of absenteeism. During a period of 25 consecutive working days, the number of absences recorded each day were as follows:

4, 2, 4, 3, 5, 9, 5, 6, 6, 3, 3, 6, 4, 3, 3, 2, 3, 5, 4, 3, 3, 2, 4, 3, 5

Determine: a) the mode; b) the median; c) the mean number of absences per day.

Solution:

a) Mode: Organise the data into a frequency distribution first, as shown below:

No. of absences in a day	No. of days
2	3
3	9
4	5
5	4
6	3
7	-
8	-
9	1
Total	**25**

Table 4.1 Absence data, example 1

The value 3 occurs 9 times. This is the most frequently occurring value. Therefore the modal value is 3 absences in a day.

Note: If there are two or more values with the maximum frequency of occurrence, then the data set has more than one mode, i.e. it is multimodal. For example:

No. of absences in a day	No. of days
2	3
3	6
4	3
5	6
6	2
7	1
Total	**21**

Table 4.2 Absence data, example 2

The values 3 and 5 occur with the maximum frequency therefore, they are both modal values. This data set is **bi-modal**.

If there is more than one modal value, the mode ceases to be useful as a 'typical' value.

b) Median: Arrange the original data in order of size, as shown below:

2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 9

The median is the value of the middle item. In this data set the middle value is the 13th item which has the value 4 absences in a day.

Note: Since the data set in the above example has an odd number of values, it was easy to identify the middle one:

$$\frac{n+1}{2} = \frac{25+1}{2} = 13 \text{ th item}$$

However, if the data had covered 30 working days, then:

$$\frac{n+1}{2} = \frac{30+1}{2} = 15.5 \text{ th item}$$

In other words, there are two 'middle' items – the 15th and the 16th items. It is normal practice to take the median as the average of the two middle values.

For example, if the 15th item has the value 3 absences in a day and the 16th item has the value 4 absences in a day, then the median is taken as:

$$\frac{3+4}{2} = 3.5 \text{ absences in a day}$$

c) The arithmetic mean – this is the sum all of the values in the data set divided by the number of items:

$$\bar{x} = \frac{4+2+4+3+....+3+5}{25} = \frac{100}{25} = 4.0 \text{ absences in a day}$$

The mean can also be found using the statistical functions on your calculator or by using a software package such as Microsoft Excel.

If the data set is reasonably small, it should be possible to estimate the size of the mean by looking at the shape of the frequency distribution. This estimate is a useful check for the calculation. In the example above, we can see that the majority of values are either 2, 3 or 4, therefore we would guess that the mean would around 3 or 4, but probably nearer to 4 than 3 since the single value of 9 will increase the size of the mean.

In practice, as the data sets used in business are likely to be large and stored electronically, spreadsheet functions (such as the ones available in Microsoft Excel) are the obvious way of finding the values of these summary measures. See section 3.3 further in this chapter for more details.

Example 4.2 – The sensitivity of the mode, median and mean

Consider again the problem described in **Example 1**. The HR Director now discovers that the 5 absences recorded for day 18 should actually have been 7 absences. What difference does this change make to the values of the three measures of location?

Solution:

Look back to the determination of the mode and median above. You should be able to see that changing one of the values from 5 to 7 makes no difference to the values of the mode or median.

However, the arithmetic mean becomes:

$$\bar{x} = \frac{102}{25} = 4.1 \text{ absences per day}$$

The mean is the only measure which responds to a small change in the data set by a proportionate small change in the value of the measure (in this case).

4.3.1.1   Advantages and disadvantages of the three types of measure of location

Mode

Advantages:

- The most frequently occurring value may be the most useful /informative measure in certain cases. For example, in marketing research, we are interested in the majority opinion about a product, an event, a policy, etc.
- The evaluation of the mode is very easy.
- It can be used with non-numeric data.
- Open-ended distributions do not affect the finding of the mode.

Disadvantages:

- An appropriate value may not exist, for example, all of the values in a data set may be different.
- If the distribution is bi-modal, there are two values for the mode and the mode ceases to be useful as a 'typical' value.
- The mode may be an extreme value in a highly asymmetric (skewed) distribution and, hence, again may be unsuitable as a typical value.
- The mode cannot be used in further statistical calculations.
- The value of the mode does not involve all of the data points.
- Small changes in the data values are not reflected proportionately by small changes in the value of the mode.
- The above means that the mode is used in a descriptive rather than an analytical way.

Median

Advantages:

- It may be useful / informative to know the data value below which 50% of the distribution lies.
- The value of the median is not affected by out-lying data values or by open-ended distributions.

Disadvantages:

- The value of the median does not involve all of the data points.
- Small changes in the data values are not reflected proportionately by small changes in the value of the median.
- Medians from separate distributions cannot be combined to produce an overall median.
- The median cannot be used in further statistical calculations.
- The above means that the median is used in a descriptive rather than an analytical way.

Mean

Advantages:

- All of the values in the data set contribute to the value of the mean.
- It is possible to 'pool' values from two or more data sets in order to calculate an overall mean for the combined data. See Section 3.4 for more details.
- This measure is suitable for use in further statistical analysis. Its use is not purely for descriptive purposes.
- This measure is capable of reflecting, in a proportionate way, small changes in the data set.

Disadvantages:

- The value of the mean is affected by out-lying values, therefore it is not usually satisfactory as a typical value in highly skewed distributions.
- The values obtained for the mean may be difficult to interpret. For instance, in an investigation of the composition of families in the UK, we could find that the mean number of children per family is 2.47 – what does this mean? Does it matter that the value is not a whole number of children?

4.3.2 OTHER USEFUL MEASURES

In this context the quartiles may be useful in helping to describe the distribution of the values of a variable. The quartiles were introduced in the previous chapter.

The lower quartile, Q1, is the 25th percentile of a set of values. 25% of the values of the variable lie below this level.

The upper quartile, Q3, is the 75th percentile of a set of values. 75% of the values of the variable lie below this level.

Normally the quartiles will be estimated from an ogive or calculated using a spreadsheet function.

4.3.3 THE USE OF RELEVANT MICROSOFT EXCEL FUNCTIONS

Microsoft Excel has a number of in-built functions which can be used to calculate the values of the summary measures discussed in this chapter. The relevant functions so far are:

Microsoft Excel function	Measure
=MODE()	Calculates the mode for a data set
=MEDIAN()	Calculates the median for a data set
=AVERAGE()	Calculates the (arithmetic) mean for a data set
=QUARTILE()	Calculates the quartile for a data set

Table 4.3 Microsoft Excel functions for summary measures

Each function must be used with the raw data, i.e. they cannot be applied to data that have already been organised into a frequency distribution. The range of the data values is inserted into brackets (). For the quartiles, we must also specify whether we want the lower or the upper value. See **Example 4.3**.

Example 4.3 – Use of Microsoft Excel functions

	A	B	C
1	Student	SAT mark /10	
2	Adey	6	
3	Adrian	3	
4	Ashling	5	
5	Ben	8	
6	Catherine	5	
7	Daniel	6	
8	Dipen	6	
9	Emma	7	
10	Frances	2	
11	Hannah	7	

The data shown are the scores achieved by 25 students on a Self-assessment test.

Cells B28:B32 contain the functions to be entered.

Cells C28:C32 contain the values that would appear in cells B28 to B32

12	**Jessica**	7	
13	**Jonathan**	9	
14	**Joseph**	7	
15	**Khalil**	7	
16	**Lauren**	8	
17	**Luke**	5	
18	**Matthew**	5	
19	**Max**	5	
20	**Michael**	4	
21	**Naomi**	7	
22	**Nicole**	8	
23	**Shahid**	6	
24	**Simon**	5	
25	**Susan**	6	
26	**Thomas**	8	
27			
28	mode	=mode(B2:B26)	5
29	median	=median(B2:B26)	6
30	mean	=average(B2:B26)	6.08
31	lower quartile	=quartile(B2:B26,1)	5
32	upper quartile	=quartile(B2:B26,3)	7

Table 4.4 Microsoft Excel example for calculating summary measures

4.3.4 CALCULATING THE MEAN FOR COMBINED VALUES

4.3.4.1   The overall mean of two or more combined groups

You are given the following salary information for the employees below management level, of a large company:

Group of employees	Number	Average^2 salary (£ pa)
Male	74	18,274
Female	57	16,962

Table 4.5 Salary data

Suppose that the information that you require is the average salary of all employees; you do not want the data split by gender. You do not have access to the raw data, i.e. all of the individual salaries.

We can find the average salary of the combined group by remembering how an average is calculated:

Average = sum of all of the values / the total number of values

In the above table, the average salary for the males is derived from:

Average for the males = total of all male salaries / 74 = £18,274

Therefore, the total of all male salaries = £18,274 × 74 = £1,352,276

Similarly for the females, the total of all female salaries = £16,962 × 57 = £966,834

Hence the total of all salaries is (total of male + total of female) salaries

= £1,352,276 + £966,834 = £2,319,110

Therefore the average of all salaries is the total of all salaries / total number of people is £2,319,110 / (74 + 57) = £2,319,110 / 131 = **£17,703** (*to the nearest £1*).

The **general procedure** for this type of calculation is described in the following:

Consider that we have a set of averages (arithmetic means) which relate to sub-groups of the same variable, e.g. salaries by gender as in the above example, or, salaries by age group, or Data Analysis for Business Decisions module exam marks by course, etc.

If we would like to know the average (arithmetic mean) of the whole group, we calculate: (No. in group 1 × mean of group 1 + No. in group 2 × mean of group 2 + …+ No. in group n × mean of group n) / total no. in all groups.2

4.3.4.2   Weighted mean

The calculation methods for the (arithmetic) mean, discussed above, assume that all of the data values are equally important. This will not always be the case. For example, in your modules you are assessed by both coursework and examination but the two components may not be equally weighted. In Data Analysis for Business Decisions the coursework / examination are weighted 60 / 40. If we require an overall average mark for a student in Data Analysis for Business Decisions then we have to take these weights into account.

Download free eBooks at bookboon.com

We do this by calculating:

(60 × coursework mark + 40 × exam mark) / (60 + 40)

Therefore, if a student scores 70% in the coursework and 55% in the exam, the weighted average mark for this student in the module is:

(60 × 70 + 40 × 55) / (60 + 40) = (4200 + 2200) / 100 = 64%

In general, the weighted mean is:

$(\text{weight}_1 \times \text{value}_1 + \text{weight}_2 \times \text{value}_2 + \text{weight}_3 \times \text{value}_3 + ,,, + \text{weight}_n \times \text{value}_n)$ / sum of all weights

Example 4.4 – Weighted mean

The spreadsheet in **Figure 4.4** shows the marks achieved by a student in Data Analysis for Business Decisions.

The marks achieved by the student are in input in cells C5, E5 and F5 and also in cells K5:M5 of **Figure 4.5**.

Cell G5 will contain the weighted average mark for the coursework.

Cell I5 will contain the weighted average mark for the whole module.

The values in cells C3:F3 are the weights for each element of coursework, e.g. the SATs are worth 12% of the total coursework mark, Project 1 is worth 25% of the total coursework mark, etc.

G2 and H2 contain the weights for the total coursework and the exam, respectively.

The mark in D5 is the weighted total of three Microsoft Excel exercises. The individual marks and weights for these are shown in **Figure 4.5**.

	A	B	C	D	E	F	G	H	I
1	**Business Analysis**	**Coursework Portfolio**							
2							0.60	0.40	
3		Weights	0.12	0.26	0.25	0.37	1.00		
4	**Student Name**		**SATs**	**Excel Models**	**Project 1**	**Project 2**	**Total Cwk**	**Exam**	**Overall Total**
5	Bloggs	Jack	67		52	56		42	

Figure 4.4 Marks achieved by a student in Data Analysis for Business Decisions

	K	L	M
1	0.07	0.12	0.07
2	**Excel Ex 1**	**Excel Ex 2**	**Excel Ex 3**
3	84	85	73

Figure 4.5 Further marks achieved by a student in Data Analysis for Business Decisions

The first task is to calculate the entry for cell D5 – the weighted average of the three marks in cells K5:M5. This is done as follows:

Arithmetically: $(0.07 \times 84 + 0.12 \times 85 + 0.07 \times 73) / (0.07 + 0.12 + 0.07) = 81.5\%$

In Microsoft Excel: enter the following formula in cell D5:

=(K1*K2 + L1*L2 + M1*M2) / SUM(K1:M1) The value 82 will appear in cell D5.

The next task is to calculate the entry for cell G5 – the weighted average of all of the coursework elements. This is done as follows:

Arithmetically: $(0.12 \times 67 + 0.26 \times 82 + 0.25 \times 52 + 0.37 \times 56) / (0.12 + 0.26 + 0.25 + 0.37) = 63.1\%$

In Microsoft Excel: enter the following formula in cell G5:

=(C3*C5 + D3*D5 + E3*E5 + F3*F5) / SUM(C3:F3) The value 63 will appear in cell G5.

The final task is to calculate the entry for cell I5 – the weighted average of the total coursework mark and the exam mark. This is done as follows:

Arithmetically: (0.60 × 63 + 0.40 × 42) / (0.60 + 0.40) = 54.6%

In Microsoft Excel: enter the following formula in cell G5:

=(G2*G5 + H2*H5) / SUM(G2:H2) The value 55 will appear in cell I5.

Note: If the weights sum to 1 as they do in the second and final 'tasks' above, then it is not necessary to explicitly divide by the sum of the weights. In the final 'task' if we had used 60 and 40 for the weights (instead of 0.60 and 0.40), then it *would* be necessary to divide by the sum, i.e. divide by 100. **See footnote[3] for further related information.**

4.3.5 THE EFFECT ON THE MEASURES OF LOCATION OF GLOBAL CHANGES TO THE DATA VALUES

When working with data and summary measures on an on-going basis we may find that the data values are changed by adding / subtracting a fixed amount or by increasing / decreasing them by a fixed percentage. If all of the data values are treated in the same way, then we can adjust the values of the summary measures very simply without the need to re-calculate them from scratch (although this is not difficult using Microsoft Excel).

Example 4.5 – Data values increased/decreased by a flat amount

The following table shows a sample of 2 bedroomed cottages in the Bakewell area which are offered as holiday cottages by a Letting Agent, together with the weekly rent up to 30 June.

Name of cottage	Rent (£/ week)	Name of cottage	Rent (£/ week)
Rose	420	Woodside	440
Honeysuckle	475	1 Farm Road	450
Briar	500	3 Chimneys	450
Riverside	450	Bridge	500
3 Top Lane	550	5 Town Mews	450
Bluebell	575	May	600

Table 4.6 Cottage data

Verify for yourselves that the modal rent is £450/week; the median rent is £463/week and the mean rent is £488/week (*the last two values are given to the nearest £1*).

Suppose from 01 July, all rents are to be increased by £30 per week. What effect will this increase have on the three measures of location?

Since all of the data values are being changed in the same way, we do not need to re-calculate any of the measures from the beginning; we simply just add the £30 increase to each of them. All that has happened is that the distribution of rents has been moved £30 to the right along the rent axis.
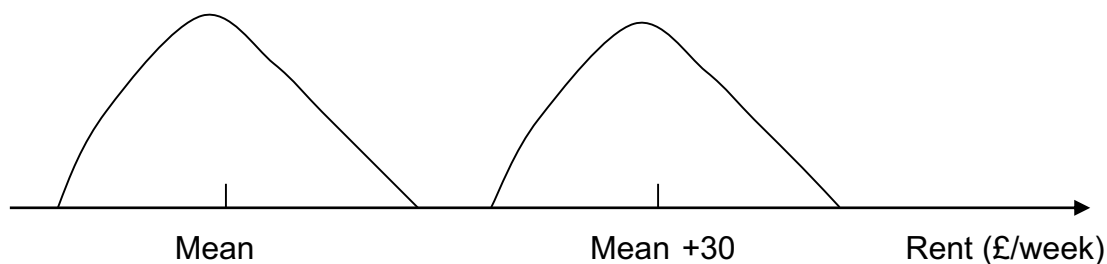
Figure 4.6 Representation of mean after £30 increase

Hence, the new modal rent is £480/week; the new median rent is £493/week and the new mean rent is £518/week (*the last two values are given to the nearest £1*).

Suppose now that for the 2016 season up to 30 June, it is proposed to increase all rents by 10%. What effect will this increase have on the three measures of location?

The same argument applies here as for the flat increase – the summary measures will each increase by 10% since every data value has been increased in this way.

Hence, the new modal rent is £495/week; the new median rent is £509/week and the new mean rent is £537/week (*the last two values are given to the nearest £1; the increases are based on the original figures in the table above*).

The same principle applies for flat rate or percentage decreases. The summary measures are reduced by the decrease in the data values.

4.3.6 MEASURES OF THE VARIABILITY IN THE DATA VALUES (MEASURES OF DISPERSION)

4.3.6.1 Introduction

If we wish to use summary measures to represent or describe a data set, it is not sufficient to simply identify a 'typical' average value (measure of location), we must also consider the degree of spread (dispersion) in the data values.

For example, consider the two distributions shown in **Figure 4.7**. These distributions might arise from two different machines used to fill packets with tea.

We can see that the same measure of location can be used to describe two quite different distributions.

For distribution E, the measure of location (mean, median or mode) is typical of the distribution as a whole. All of the values which have a high frequency of occurrence are clustered over a narrow range around the measure of location. The machine is working consistently.

For distribution D, the measure of location is less typical of the distribution as a whole – the individual values are much more spread out than in E. The amount of tea delivered to a packet is much more variable – this is not a good thing for the company.
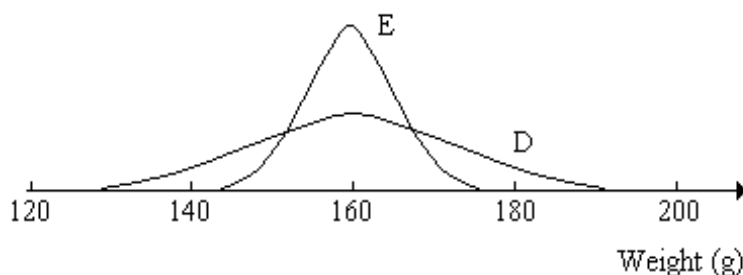
Figure 4.7 Two distributions with the same mean but different dispersion

Example 4.6 – To illustrate the need for a measure of dispersion

Consider two rooms, each containing 5 people. In the blue room the ages of the people are 54, 35, 24, 10 and 2 years. In the red room the ages are 30, 27, 24, 23 and 21 years.

The mean age is the same in each room (25 years), and the median age is the same (24 years). However, the variability of the ages is quite different in the two rooms.

Hence, we require a **measure of spread** (called a measure of **dispersion**) as well as a measure of location. Using the measure of location and the measure of dispersion together, we will be able to distinguish between distributions such as D and E above.

Measures of dispersion are measures of the scatter, spread, variability or unpredictability of the data values relative to a measure of location.

How can we find a suitable measure of the spread of the data values about the mean or the median? (The mode is not usually considered any further.) The measure of dispersion used must provide us with a guide to the reliability of the measure of location as a representative of the distribution as whole.

4.3.6.2   The range as a measure of dispersion

A very simple way of assessing the spread, or dispersion, in a data set is to use the range of the values. By definition, to calculate the range of a set of values:

Range = Largest value – Smallest value

If we consider **Example 4.5**, the range of ages in the blue room is (54 – 2) = 52 years, and the range in the red room is (30 – 21) = 9 years.

Therefore, we can say that in the blue room the mean age is 25 years with a range of 52 years, and in the red room the mean age is 25 years with a range of 9 years.

These values indicate clearly the difference between the two data sets.

However, problems can arise with the range as the measure of spread, if the data set contains outlying values. For example, suppose the ages of the people in the red room were 30, 27, 24, 23 and 2 years. The mean age is 21 years with a range of 29 years.

This information, if it appeared without the raw data, would give a misleading impression of the spread of the values about the mean.

Outlying values should not be ignored altogether, without careful thought, but, at the same time, they should not be allowed to distort the measure of spread unreasonably.

We conclude that using the range as a measure of the dispersion of the data is convenient, in that it is simple to calculate and easy to understand. However, the range has drawbacks in that it uses only two values out of the data set, and it is subject to distortion by outlying values.

4.3.6.3   The quartiles as a measure of dispersion

The problem of outlying values distorting the range can be overcome by using the quartiles to measure the spread in the data. The difference between the lower and upper quartiles measures the range of the middle half of the data values. We have already used this idea when we looked at box plots.

The **quartile deviation** combines the lower and upper quartiles in a single measure which may be used to describe the spread of data values about the median. The quartile deviation is:

$$QD = \frac{Q_3 - Q_1}{2}$$

Whilst it is often useful to be able to describe the spread of the middle portion of the data set in terms of the quartile deviation, this measure is unsuitable for use in further statistical analysis. It is a descriptive measure only.

Example 4.7 – The calculation and use of the quartile deviation

In the previous chapter, **Example 4.3**, we looked at the % moisture content of batches of washing powder, produced by day shift and night shift workers, respectively. The data are repeated below.

Batch	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
% moisture	7.2	6.5	5.3	3.9	6.7	5.9	6.0	6.5	6.9	4.3	6.3	7.6	5.9	5.7	5.5

Table 4.7 Day shift (15 batches)

Batch	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
% moisture	5.9	5.6	3.3	4.4	6.9	6.1	5.6	2.9	6.4	8.8	8.1	7.1	7.1	4.8	4.0	2.5	6.4	6.5

Table 4.8 Night shift (18 batches)

Determine the median and quartile deviation of the % moisture content for the day shift and for the night shift. Compare the two distributions.

Solution:

For the Day Shift, arranging the data values in order of size:

3.9 4.3 5.3 5.5 5.7 5.9 5.9 6.0 6.3 6.5 6.5 6.7 6.9 7.2 7.6

There are 15 values, therefore the median is the value of the 8th item ((n+1) / 2). The median is 6.0%.

The lower quartile is the (n+1) / 4 item, i.e. the 4th item. The lower quartile is therefore 5.5%.Check footnote [4] for further details.

Similarly, the upper quartile is taken as the 3 × (n+1) / 4 item i.e. the 12th item. The upper quartile is 6.7%.

The quartile deviation is:

$$QD = \frac{Q_3 - Q_1}{2} = \frac{6.7 - 5.5}{2} = 0.6\%$$

For the day shift, the median % moisture content is 6.0% with a quartile deviation of 0.6%.

For the Night Shift, arranging the data in order of size as before:

2.5 2.9 3.3 4.0 4.4 4.8 5.6 5.6 5.9 6.1 6.4 6.4 6.5 6.9 7.1 7.1 8.1 8.8

There are 18 values. The median is the average of the 9th and 10th items. The median is:

$$\frac{5.9 + 6.1}{2} = 6.0\%$$

Using the same basis as for the Day Shift, the lower quartile is taken as the 4.75th item which is approximately 4.4%.

The upper quartile is therefore taken as 6.9% approximately.

The quartile deviation is:

$$QD = \frac{Q_3 - Q_1}{2} = \frac{6.9 - 4.4}{2} = 1.25\%$$

For the night shift, the median % moisture content is 6.0% with a quartile deviation of 1.25%.

The two shifts have the same median % moisture content, but the individual batch values are more spread out relative to the median, for the night shift than for the day shift – less consistency in the production at night.

4.3.6.4 The standard deviation as a measure of dispersion

We now require a measure of dispersion which we can use with the arithmetic mean and, which will be suitable for use in further statistical analysis. In other words, we must design a measure of spread which uses all of the data values. The measure which satisfies these conditions is called the **standard deviation**.

Consider **Example 4.5** again. This example concerns the ages of two groups of five people. In the blue room the ages of the people are 54, 35, 24, 10 and 2 years. In the red room the ages are 30, 27, 24, 23 and 21 years.

We have already found that for both rooms the mean age is 25 years. We now wish to look at the spread of values within each group. Let us look at the difference between each value in the data set and the mean. This difference is called the **deviation** from the mean. See **Table 4.9**.

Blue room		Red room	
Age (years)	**Age – mean age (years)**	**Age (years)**	**Age – mean age (years)**
54	29	30	5
35	10	27	2
24	-1	24	-1
10	-15	23	-2
2	-23	21	-4
Total	**0**	**Total**	**0**

Table 4.9 Deviations from the mean

The deviations may be illustrated as shown in **Figure 4.8**:
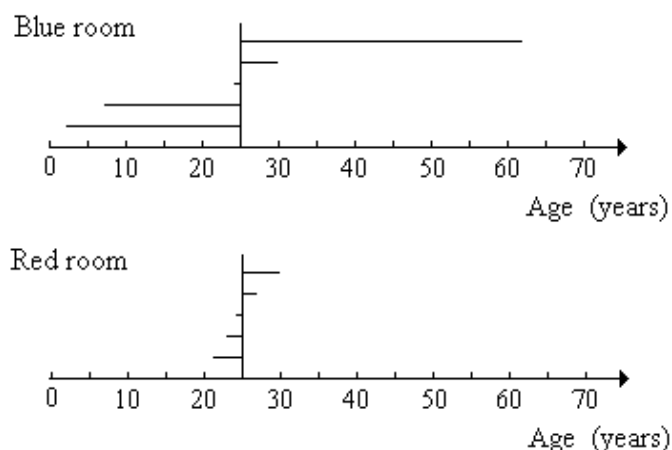
Figure 4.8 The deviations of each age from the mean age

The pattern of the deviations of the individual values from the mean is a good indicator of the dispersion within the data. Ideally, we would like to calculate the arithmetic mean of these deviations and use this as our measure of spread. However, as you can see from the figures in **Table 1**, the deviations from the mean sum to zero for both rooms. This will always happen since the arithmetic mean represents the value of 'equal shares', e.g. in the above example, the total number of years is shared out equally between the 5 people in the room, 'giving' 25 years to each one. In each case, the total negative deviations will exactly balance the total positive deviations.

There are various ways in which we can overcome this problem. We could simply ignore the negative signs and consider only the size of the deviations. The arithmetic mean of the size of the deviations can then be used as a measure of the dispersion in the data set. This measure is called the **mean deviation**.

However, simply ignoring the signs is not entirely satisfactory from a mathematical point of view, therefore we use the alternative tactic of squaring the deviations. We then calculate the arithmetic mean of the squared deviations. This measure is called the **variance**. We will use the data from the **Example 4.5** (Red and Blue Rooms) to illustrate the procedure. The deviation of each age from the mean age is denoted by d.

See **Table 4.10**.

Blue Room			Red Room		
Age (years)	Age – mean age d (years)	d²	Age (years)	Age – mean age d (years)	d²
62	37	1369	30	5	25
30	5	25	27	2	4
24	-1	1	24	-1	1
7	-18	324	23	-2	4
2	-23	529	21	-4	16
Total	**0**	**2248**	**Total**	**0**	**50**

Table 4.10 Calculation of variance

Mean of the squared deviations, i.e. variance = sum of all d2 values / n = $\dfrac{\sum d^2}{n}$

Blue Room: Variance = $\dfrac{2248}{5} = 449.6$ (years)^2

Red Room: Variance = $\dfrac{50}{5} = 10$ (years)^2

The variance is a useful measure in its own right but in the current context, we will take the square root of the variance. The resulting measure is called the **standard deviation**. It has the advantage over the variance of being measured in the same units as the values in the data set.

Standard deviation = $\sqrt{Variance}$

Blue Room: Standard deviation $\sqrt{449.6} = 21.2$ years

Red Room: Standard deviation $= \sqrt{10} = 3.2$ years

Since both data sets have the same mean, 25 years, we can say at once that the ages in the Blue Room are much more spread out about the mean than those in the Red Room.

The important point to remember is that the standard deviation is related to the average deviation from the mean of all of the data values in the set.

4.3.7 THE USE OF RELEVANT MICROSOFT EXCEL FUNCTIONS

Microsoft Excel has an in-built function for the standard deviation and we can use a simple equation to calculate the quartile deviation.

=STDEVP()   Calculates the standard deviation, please check footnote 5.

As for the measures of location, this function must be used with the raw data. We demonstrate its use, using the data in **Example 4.3**, some of which is reproduced here.

Example 4.8 – Use of Microsoft Excel functions

	A	B	C	
1	Student	SAT mark /10		The data shown are the scores achieved by 25 students on a Self-assessment test.
2	Adey	6		
3	Adrian	3		
4	Ashling	5		
	…			Cells B33:B34 contain the functions to be entered.
24	Simon	5		
25	Susan	6		
26	Thomas	8		Cells C33:C34 contain the values that would appear in cells B33 and B34
27				
28	mode	=mode(B2:B26)	5	
29	median	=median(B2:B26)	6	
30	mean	=average(B2:B26)	6.08	
31	lower quartile	=quartile(B2:B26,1)	5	
32	upper quartile	=quartile(B2:B26,3)	7	
33	standard deviation	=STDEVP(B2:B26)	1.62	
34	quartile deviation	=(B32-B31)/2	1.0	

Table 4.11 Microsoft Excel example for calculating variance

4.3.8 INTERPRETATION OF THE STANDARD DEVIATION

We now must consider what information is given by the value of the standard deviation.

Example 4.9 – Interpretation of the standard deviation

Refer to **Example 4.1.** The number of absences each day at Shefftex plc:

4, 2, 4, 3, 5, 9, 5, 6, 6, 3, 3, 6, 4, 3, 3, 2, 3, 5, 4, 3, 3, 2, 4, 3, 5

The mean number of absences per day was found to be 4.0. Using Microsoft Excel and the STDEVP() function, we find that the standard deviation is 1.6 absences per day.

Hence the 25 absence records in the data set may be represented by a mean of 4.0 absences per day. The spread of the individual data values about this mean is measured as 1.6 absences per day. Does this indicate a large or a small spread?

It is difficult to answer this question with any certainty. It all depends. All we can really do is to develop, with practice and experience, a 'feel' about the meaning of particular values of standard deviation.

In this example, 1.6 absences per day is approximately 39% of the mean (4.0 absences per day). This suggests a fairly large spread for this type of data. The mean does not appear to be a good representative of the values in the data set. However, it must be noted that the data set does contain an 'outlier' – 9 absences in a day. This figure increases the value of the measure of spread of the data relative to the mean – as it also increased the value of the mean.

Since the value of the standard deviation is affected by the presence of outlying values, we must always check that large standard deviations do not arise because of one or two outliers, with the rest of the data closely clustered about the mean.

The standard deviation becomes easier to use when we wish to compare two or more distributions.

Suppose that the data in **Example 4.9** refer to August and the corresponding summary measures for September are as follows:

mean number of absences per day is 6.2

standard deviation of the number of absences per day 0.8

These figures are calculated from the equivalent absence records for the working days in September.

We can see at once that the mean number of absences has risen, however, the standard deviation for September is smaller than that for August, indicating that the September absences are more closely clustered around the mean – no evidence of an outlier – the values in the September data set appear to be more consistent than for August. (It would be useful to see the data before coming to a final conclusion.)

4.3.9 RELATIVE DISPERSION

When comparing the means and standard deviations of two or more distributions, it is useful to express the standard deviation as a percentage of the mean. This measure is referred to as the **relative dispersion** or **coefficient of variation**.

$$\text{Coefficient of variation} = \frac{s}{\bar{x}} \times 100\%$$

Using the data from **Example 4.9**, we find:

Absences in August: Relative dispersion $= \dfrac{1.6}{4.0} \times 100\% = 39.4\%$

Absences in September: Relative dispersion $= \dfrac{0.8}{6.2} \times 100\% = 12.9\%$

which confirms the view that the data for the second month is less dispersed relative to the mean.

The use of the coefficient of variation is important when comparing data sets for which the variable values are significantly different in size.

4.3.10   THE EFFECT ON THE STANDARD DEVIATION OF GLOBAL CHANGES TO THE DATA VALUES

As for the measures of location (in 3.5), if all of the data values are treated in the same way, then we can adjust the value of the standard deviation very simply without the need to re-calculate it from scratch (although this is not difficult using Microsoft Excel).

Refer to **Example 4.5**. Verify for yourselves that the standard deviation of the current rents is £55.60 per week (*to the nearest 10p).*

We will find the standard deviation if:

a) all of the rents are increased by £30 per week on 01 July;

b) all of the rents are increased by 10% for the 2008 season.

c) A flat rate increase of £30 per week will increase the mean by £30 but the spread of the data relative to the mean does not change, hence the standard deviation remains at £55.60 per week.

d) A percentage increase of 10% will increase the mean by 10% and this time the spread of the data relative to the mean does change. It can be shown[6] that the standard deviation also increases by 10% to become £61.20 per week. (*to the nearest 10p*)

4.3.11   SYMMETRY

The third aspect of a data set which we must consider is the symmetry, or lack of symmetry, in the distribution of the values. The lack of symmetry is referred to as **skewness**. We will not make any attempt to devise a measure of the symmetry of a distribution. We will simply describe it in words.

Look at the three distributions illustrated in **Figure 4.9**.
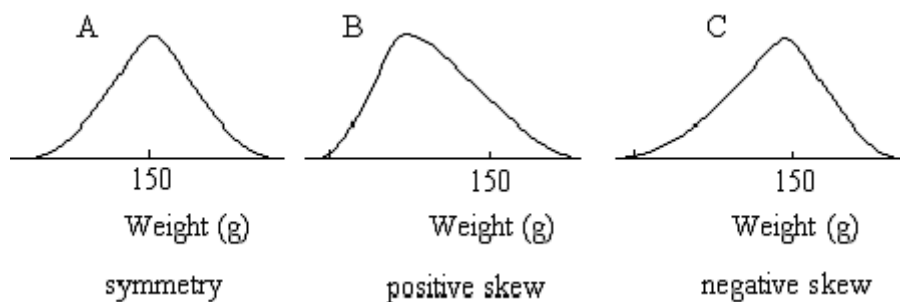
Figure 4.9 Skewed distributions

Distribution A shows symmetry. In this case, mean = median = mode. This is the type of distribution you would expect to obtain if you collected data on:

- the weights of packets of a certain type of biscuit
- the lengths of steel rod cut by a guillotine
- the IQ of a random sample of people

Distribution B has a tail to the right – the higher values. It is said to be **positively skewed**. In this case, mean > median > mode. This is the type of distribution you would expect to obtain if you collected data on:

- the wages paid to production line workers
- the ages of students undertaking first degree courses
- the service times in a garage

Distribution C has a tail to the left – the lower values. It is said to be **negatively skewed**. In this case, mean < median < mode. This is the type of distribution you would expect to obtain if you collected data on:

- the age at which glasses are first needed
- the waiting time in some types of queue
- the lifetime to failure of some types of electrical equipment

4.5   CHAPTER REFERENCES

1. Morris, C, Quantitative Approaches in Business Studies, 8th edition, FT/Prentice Hall, 2012 (eBook, 7th edition 2008), Chapter 6 (in part)
2. Oakshott, L, Essential Quantitative Methods for Business, Management & Finance, 5th edition, Palgrave Macmillan, 2012, Chapter 6
3. Waters, D, Quantitative Methods for Business, 5th edition, FT/Prentice Hall, 2011 (eBook available), Chapters 5 and 6
4. Wisniewski, M, Quantitative Methods for Decision Makers, 5th edition, FT/Prentice Hall, 2009 (eBook available), Chapter 2 (in part) and 3 (in part)

5   MEASURING AND EVALUATING VARIATION IN DATA VALUES

5.1   INTRODUCTION

Coping with variation in the values of key variables such as prices, costs, demand, etc. is one of the key activities within a business. This type of variation arises both internally and externally to the business. Changes in the prices of certain goods may well affect the demand for your product; the way in which inflation or the value of Sterling are controlled may affect the cost of your raw materials, and so on.

The size and direction of the variation in the values of key variables will themselves vary and may have different relative importance. For example, an increase in the price of a raw material of 10 p per kg may not be very important if the price is changing from £20.00 to £20.10 but is likely to represent a more serious situation if it is changing from 10 p to 20 p (i.e. doubling). A manager not only has to accommodate these changes when planning and budgeting but also has to **evaluate** the relative importance of the variation.

In order to evaluate the effect of changes in variable values, a simple percentage is often sufficient. For example, if a price changes from £20.00 to £20 .10, this is a 0.5% increase, but if a price changes from 10 p to 20 p this is a 100% increase.

If the price of an item is recorded on a monthly basis, then we can calculate the percentage change from month to month and form a view on the way in which the price is varying.

Alternatively, it can be helpful to refer the price each month to the price in some fixed month. This chosen reference price forms a basis for comparison of the prices in the other months.

In practice, we do not explicitly calculate the percentage **change** in the price relative to the fixed price, rather we simply calculate the value of the price as a percentage of the price at the base. These relative measures are called **index numbers**. The index number of the reference price is always 100 and we measure all other prices relative to this. This procedure simplifies the calculations.

For example, if the price of an item in the reference month is £5.65 and in the current month is £6.45, the percentage change in price is:

$$\frac{6.45 - 5.65}{5.65} \times 100 = 14.2\%$$

whereas the index number is:

$$\frac{6.45}{5.65} \times 100 = 114.2 \text{ i.e. } 14.2\% \text{ above the base of } 100$$

Both methods give the same information but the index number calculation is more 'streamlined'.

Index numbers allow us to monitor the way in which the value of any variable is changing over time or space and also to assess the relative size and hence importance of the changes. In this way comparisons may be made between values arising at different points in time. Any variable or time interval can be used.

The reference point is called **the base** and the index number of the variable at the base is always **100**.

There is a variety of different ways of constructing index numbers depending on whether the index refers to:

- a single item or a group of items
- prices or quantities, etc.,
- single values or weighted values

In the following we will look at a few of the most commonly used indices, but we will begin with a refresher on the use of percentages to measure change.

5.2 PERCENTAGE CHANGE

It is sometimes difficult to interpret and evaluate the information provide by absolute values. For example:

- The total annual revenue from the restaurant in the Birmingham branch of John Lewis was £3,160,000 in 2014/2015. Is this good?
- The price of a large loaf and 500 g of steak have both increased by 10p over the last month. Is this a lot?

In the first example, £3.2m revenue over a year sounds a lot but we cannot be sure – we need to set the figure in context in order to *evaluate* it.

If we are now told that the Birmingham restaurant revenue comprises 14% of the total revenue from the restaurants in John Lewis we begin to form a clearer view – it is an above average figure for the stores.

In addition, we find that this annual revenue for the Birmingham restaurant has increased by 11% over the previous year which showed a 10% increase over 2012/2013 – there appears to be a small increase in the rate of growth.

These percentages are helping us to put an interpretation on the actual revenue value for 2014/2015. We can evaluate it over time and in comparison to similar units.

In the second example, the 10p represents a 7.5% price increase on a loaf but a 1% price increase on half a kilo of steak. Again, the calculation of the percentage helps us to *evaluate* the absolute value.

Returning to the first example about the annual revenue from the restaurant in the Birmingham store of John Lewis, we said that the figure had increased by 10% from 2012/2013 to 2013/2014 and by 11% from 2013/2014 to 2014/2015.

These percentages provide useful information but sometimes it is beneficial to use a fixed point and refer all changes to that point. Hence in this example, we could say that the percentage change from 2012/2013 to 2013/2014 is 10% and the percentage change from 2012/2013 to 2014/2015 is 22%.

Both percentages relate to the first year 2012/2013. In this situation it is more usual to use an index number rather than a percentage, although the two are closely linked.

5.2.1   INDEX NUMBERS FOR A SINGLE ITEM

The term **price relative** is an alternative name for a simple index number. It is the price of the item at one point in time (a given year, say) expressed as a percentage of the price of the item at the base time.

$$\text{Price relative} = \frac{\text{price at given time, } p_n}{\text{price at base time, } p_0} \times 100$$

Example 5.1 – Evaluating the changes in the value of a single item

Table 5.1 shows the value of UK imports of a given category of goods for the years 2011 – 2015. Construct an index number which will describe the changes in the value of the imports of the good over the period.

Year	Imports (£ billion)
2011	10.89
2012	11.41
2013	12.73
2014	14.46
2015	14.50

Table 5.1 Annual import of goods

Solution: One method of measuring the variation in the value of the imports would be to use percentages.

Table 5.2 shows the change that has occurred in the value of imports relative to 2011, as a number and as a percentage.

Year	Imports (£ billion)	Change since 2011 (£ bn)	Change relative to 2011 (%)
2011	10.89	-	-
2012	11.41	0.52	4.78
2013	12.73	1.84	16.90
2014	14.46	3.57	32.78
2015	14.50	3.61	33.15

Table 5.2 Percentage change in imports relative to 2008

Notice that imports rose a little between 2011 and 2012, but increased very rapidly between 2012 and 2014. From 2014 to 2015 the rate of increase slowed down again.

Alternatively the variation in import values relative to 2011 could be portrayed using index numbers. In **Example 5.1**, the index number for 2013 relative to 2011 is:

Import index for 2013 with 2011 as base = $\dfrac{\text{Imports in 2013 x 100}}{\text{Imports in 2011}}$ = 116.90

The index numbers for 2011 to 2015 are shown in **Table 5.3.**

It is clear that the index number at the base will always have the value 100. Subsequent indexes will be above or below 100 depending on whether there has been an increase or decrease in the variable values compared to the base.

We can say, for example, for 2014 which has an index number of 132.78 that imports have risen by approx. 32.8% since 2011.

Year	Imports (£ billion)	Index No. with 2011 as base
2011	10.89	$\dfrac{10.89}{10.89} \times 100 = 100.00$
2012	11.41	$\dfrac{11.41}{10.89} \times 100 = 104.78$
2013	12.73	$\dfrac{12.73}{10.89} \times 100 = 116.90$
2014	14.46	$\dfrac{14.46}{10.89} \times 100 = 132.78$
2015	14.50	$\dfrac{14.50}{10.89} \times 100 = 133.15$

Table 5.3 Import Index for 2011 – 2015 with 2011 as base year

It should also be clear, that it is vital to specify the base when quoting an index number – the figures are meaningless without this information.

The short hand way of specifying the base is **'2011 = 100'** – using **Example 5.1**. Any year may be chosen as the base; it does not have to be the first in the series.

This example can be translated into Microsoft Excel as following:

	A	B	C	D
1	Year	Imports (£ billion)	Index No. with 2011 as base	
2	**2011**	10.89	=B2/B2*100	100.00
3	**2012**	11.41	=B3/B2*100	104.78
4	**2013**	12.73	=B4/B2*100	116.90
5	**2014**	14.46	=B5/B2*100	132.78
6	**2015**	14.50	=B6/B2*100	133.15

The data shown are the data from **Table 5.3.**

Cells C2:C6 contain the functions to be entered.

Cells D2:D6 contain the values that would appear in cells C2:C6.

Table 5.4 Microsoft Excel example for calculating index numbers

Table 5.5 shows the import index numbers for Example 5.1 using 2013 as the base.

Year	Imports (£ billion)	Index No. with 2013 as base
2011	10.89	$\frac{11.41}{12.73} \times 100 = 85.55$
2012	11.41	$\frac{12.73}{12.73} \times 100 = 89.63$
2013	12.73	$\frac{12.73}{12.73} \times 100 = 100.00$
2014	14.46	$\frac{14.46}{12.73} \times 100 = 113.59$
2015	14.50	$\frac{14.50}{12.73} \times 100 = 113.90$

Table 5.5 Import Index for 2011 – 2015, 2013 = 100

The new index numbers measure the changes over the same period as before but all of the changes are now relative to 2013 and not 2011.

Notice that the index for 2011 is now 85.5. What does this mean? The figure is below 100, which indicates that imports were below the 2013 level. In fact they were:

100 85.55 = 14.45% below the 2013 level.

5.2.2 INDEX NUMBERS FOR A GROUP OF ITEMS

The import index calculated in Example 1, measured the changes in the values of one item only. We also want an index number which will measure the amount by which the value of a group of items has changed. A share price index is an example of this case.

World financial centres have their own share price index – the FTSE100 (London), the Dow Jones (Wall Street), the Nikkei (Tokyo), etc. These indexes are published in the financial columns of newspapers and are valuable indicators of movements of share prices. They are measures of the changes not just of one share price, but of a range of shares.

In order to calculate this type of index, the compilers of the index must first select a sample, or 'basket', of shares. The sample is selected to represent the range of shares normally dealt in on the world's stock exchanges. The price changes in this basket of shares are monitored day to day and the **average change** is calculated.

Example 5.2 – To construct a simple index number for groups of items

Table 5.6 shows price movements of three shares between March and September 2015. Construct an index number which will describe the changes in the average value of the share prices over the period.

Share	Price	Price
	Mar 2015	Sept 2015
UK Enterprises plc	£9.10	£12.15
World Products Co.	£3.40	£3.85
Euroland Holdings	£2.75	£3.55

Table 5.6 Share Price Movements

Solution: Calculate the average share price in March and September 2015.

The average share price for March 2015 is $\dfrac{9.10 + 3.40 + 2.75}{3} = £5.08$

The average share price for September 2015 is $\dfrac{12.15 + 3.85 + 3.55}{3} = £6.52$

Use these two averages to create a share price index for September 2015 (March 2015 = 100):

$$\frac{6.52}{5.08} \times 100 = 128$$

Therefore the share price index is:

March 2015 = 100
September 2015 = 128

Which suggests that average share * prices have risen by 28% over the period.

* Note: Relates to the shares in the 'basket' only.

This method clearly would be less satisfactory in the context of retail prices. The commodities in a 'shopping basket' of, say, food items, are not all the same in the sense that different quantities are purchased of each item in a given period. For example, one unit of butter (250 g say) would probably be sufficient for 3 or 4 units (loaves) of bread.

When we are considering a group of items in 'a basket' where the items are not all equally important, we must consider **weighting** their contribution to the index number.

5.3   WEIGHTED INDEX NUMBERS

If all items in the 'basket' are not of equal importance, one way of weighting the contribution of each one to the index number is through the expenditure on that item, e.g. the cost incurred in the purchase of that item in a given time period. This expenditure depends directly on the **quantity** of the item which is purchased in the period as well as the price.

This is the principle behind the Retail Price Index (RPI) which is used to monitor the 'cost of living' in the UK – amongst other things.

The RPI is based on a well-defined 'shopping basket' of goods and services purchased by a large sample of households in a given period. The index correctly reflects the relative importance of the various items which the shoppers buy in the given period via the expenditure on the items.

$$\text{Price index} = \frac{\text{total expenditure on 'shopping basket' at present prices}}{\text{total expenditure on 'shopping basket' at base year prices}}$$

The total expenditure on an item is the quantity purchased multiplied by the unit price of that item. If these individual expenditures are added together, we have the total expenditure on the 'shopping basket'.

If the price of the item varies over time, then the quantity purchased is likely to vary as well. If an item increases in price, then we may respond by buying less. Alternatively, as time passes we may have become more affluent and hence buy more of the item, regardless of any increase in price.

Hence, for the price index calculation, we have to decide whether to use the quantity which we purchased in the base year or the quantity which we purchased in the current year, in order to work out the expenditure on the item.

There are two approaches which are used. Both of these approaches assume that the quantities being purchased do not change over time even if the prices change.

5.3.1   BASE YEAR WEIGHTING

The first approach is attributed to Laspeyres and is sometime referred to as the Laspeyres' price index. Since the basis of this price index is to assume that people are still buying in the current year the quantities which they bought in the base year, it is more helpful to refer to this price index as the **'baseweighted price index'**.

$$\text{Base weighted price index} = \frac{\text{total cost of base year quantities at current prices}}{\text{total cost of base year quantities at base year prices}} \times 100$$

Using p_n to denote current year unit prices; p_0 to denote base year unit prices; q_n to denote current year quantities purchased; q_0 to denote base year quantities purchased, the base year weighted price index may be written as:

$$\text{Base weighted price index} = \frac{\sum p_n \times q_0}{\sum p_0 \times q_0} \times 100$$

where the $\sum$ symbol means 'sum of'.

Example 5.3 – To construct a base year weighted index number

The prices per unit for bread, milk and butter are given in **Table 5.6** for January 2014 and 2015. Suppose that we have carried out some research and found out that in January 2014 the average shopper distributed spending so that for every £1 spent on butter, £1 was spent on milk and £3 on bread. Calculate a weighted retail price index for these three commodities in January 2015 with January 2014 as the base.

Solution: The numbers 3, 1 and 1 will be used as weights for bread, milk and butter respectively. They are clearly related to the quantities purchased of each item.

The calculations for the weighted index of prices are shown in **Table 5.7**.

Item	Unit price (pence)		Weight, w_0	$w_0 \times p_0$	$w_0 \times p_n$
	Jan 2014	Jan 2015			
Bread	90	99	3	270	297
Milk	60	61	1	60	61
Butter	80	90	1	80	90
Total				**410**	**448**

Table 5.7 Calculation of weighted index number (January 2014 = 100)

The weighted index number $= \dfrac{448}{410} \times 100 = 109$

This figure says that prices have risen by 9% over the year.

5.3.2 CURRENT YEAR WEIGHTING

In the previous section we assumed that shoppers are buying the same quantities of an item in the current year as they were in the base year. An alternative assumption is that people were buying in the base year the same quantities as they are buying now. This second assumption leads us to a current year weighted price index which is attributed to Paasche. Hence:

Current weighted price index $\times \dfrac{\text{total cost of current year quantities at current prices}}{\text{total cost of current year quantities at base year prices}} \times 100$

Using the notation defined above,

Current weighted price index $\times \dfrac{\sum p_n \times q_n}{\sum p_0 \times q_n}\ 100$ where Σ means 'sum of'.

Example 5.4 – To calculate a weighted index number using current year weights

Refer to **Example 5.3**. In 2015 consumer research indicated that for every £5 spent on bread, the average shopper spent £3 on milk and £1 on butter. Calculate the all items price index for 2015 (2014 = 100) using these weights.

Solution: Calculations for the new revised index are shown in **Table 5.7**.

The current year weighted index number = $\dfrac{768}{710} \times 100 = 108$

This figure says that prices have risen by 8% over the year which is slightly lower than the value obtained with the year 2014 weights.

Item	Unit price (pence)		Weight, w_n	$w_n \times p_0$	$w_n \times p_n$
	Jan 2014	Jan 2015			
Bread	90	99	5	450	495
Milk	60	61	3	180	183
Butter	80	90	1	80	90
Total				**710**	**768**

Table 5.8 Calculation of weighted index number (January 2014 = 100)

This example can be translated into Microsoft Excel as following:

	A	B	C	D	E	F
1	Item	Unit price (pence)		Weight, w_n	$w_n \times p_0$	$w_n \times p_n$
2		Jan 2014	Jan 2015			
3	Bread	90	99	5	=D3*B3 (450)	=D3*C3 (495)
4	Milk	60	61	3	=D4*B4 (180)	=D4*C4 (183)
5	Butter	80	90	1	=D5*B5 (80)	=D5*C5 (90)
6	Total				=sum(E3:E5) (710)	=sum(F3:F5) (768)

The data shown are the data from **Table 5.8.**

Cells E3:F6 contain the functions to be entered and the values that would appear in those cells (in brackets).

Table 5.9 Microsoft Excel example for calculating index numbers

5.3.3 BASE YEAR WEIGHTED OR CURRENT YEAR WEIGHTED INDEX NUMBERS

In **Example 5.3** we used weights drawn from the base year. Such indices are known as **baseweighted** indices. In **Example 5.4** we have used weights drawn from the current year. The result is known as a **current-weighted** index.

A baseweighted price index measures the average price change in a fixed basket of goods typically purchased in the base year.

A currentweighted price index also measures average price changes, but in a fixed basket of goods typically purchased in the current year.

The index we use depends very much on our intentions.

For example, base weighting would be useful if we wished to find out how much the cost of a particular basket of goods purchased in some past year has gone up between then and now. There is also the advantage that the value of the weights is fixed.

Current weighting would be the choice if we wished to discover the change in the cost of today's typical basket of goods over past years. We would, however, have to recognise that extra costs would be involved in continually obtaining new data to construct weights for each current period.

Whichever index is chosen, it is essential that weights are kept constant throughout the index number series.

5.4   DEFLATING A TIME SERIES – USING A VALUE INDEX AS A DEFLATOR

Consider the following situation. Between 2013 and 2015 the average weekly wage in a certain industry changed as shown below:

Year	2013	2014	2015
Wage (£/week)	420	460	485

Table 5.10 Average weekly wage data

The employers of the industry claimed that wages had shown real and substantial increases during this period, but the industry trade union leaders were not so sure. In order to shed light on this situation, the union consulted the retail price index for the period:

Year (January)	2013	2014	2015
Price Index, (Jan 2007=100)	186.7	192.0	198.1

Table 5.11 Retail price index data

Since it is known that price inflation erodes wages and salaries, how can we decide whether wages in the above industry have stayed ahead of inflation or not? We wish to know what the wage changes would be like if we allowed for the effect of price changes.

We can use the retail price index to remove the effect of price inflation from the wage figures. The procedure is referred to as deflating the (wages) series, i.e. we are using the retail price index as a 'deflator'.

Before we proceed, we must move the base for the retail price index from 2007 = 100 to 2013 =100. See **Table 5.12**.

Year (January)	2013	2014	2015
Price Index, (Jan 2007=100)	186.7	192.0	198.1
Price Index, (Jan 2013=100)	100.0	$\dfrac{192.0}{186.7} \times 100 = 102.8$	$\dfrac{198.1}{186.7} \times 100 = 106.1$

Table 5.12 Changing the base of an index number

Between 2013 and 2015, the wages rose by $\dfrac{485}{420} \times 100 = 115.5$ i.e. 15.5%

Over the same period prices went up by 6.1%.

In order to adjust the wages to allow for the effects of inflation, we calculate

The deflated wage value (the real value) = $\dfrac{\text{actual wage value}}{\text{retail price index}} \times 100$

In words, we divide the value we wish to deflate by a suitable index for that time period and then multiply by 100

Now that we have a retail price index number series based on 2013 = 100, we are ready to deflate the wage series. See **Table 5.13**.

Year	2013	2014	2015
Actual wage (£/week)	420	460	485
Real wage (£/week)	420	$\dfrac{460}{102.8} \times 100 = 447$	$\dfrac{485}{106.1} \times 100 = 457$

Table 5.13 Deflating a wage series

We have removed from each of the wage figures a factor related to the degree of price inflation. This deflated amount is often referred to as '**real**' wages.

We can see that wages have increased in real terms – the value of the money increase has been reduced but has not been negated by the effect of the price inflation.

5.5   CHAPTER REFERENCES

1. Morris, C, Quantitative Approaches in Business Studies, 8th edition, FT/Prentice Hall, 2012 (eBook, 7th edition 2008) Chapter 7

2. Oakshott, L, Essential Quantitative Methods for Business, Management & Finance, 5th edition, Palgrave Macmillan, 2012, Chapter 2

3. Wisniewski, M, Quantitative Methods for Decision Makers, 5th edition, FT/Prentice Hall, 2009 (eBook available), Chapter 7.

6 ASSOCIATION BETWEEN VARIABLES

6.1 INTRODUCTION

In the previous chapter we discussed the occurrence of variation in the values of business variables. We have seen that many activities in business are subject to variation. Two attempts at the same event (making a sale, taking a journey, producing an item) will generally produce a slightly different result in terms of the time it took or the quality of the outcome. This is quite natural and variation to some extent is inevitable.

A company may notice that its sales revenue varies from month to month. The question is, can the company explain this variation (and maybe control it in some way) or must it just accept it as being 'one of those things'? For the company, some of the variation in sales revenue might be explained by the fact that there are promotions from time to time. This affects the amount earned when an item is sold, but it also affects the number of items sold. Another factor which might have some bearing on the situation is the level of advertising. This also varies from month to month. There is a general belief in the company that spending more on advertising will cause sales to increase in the next month. There are also other factors that might have some bearing on the variation in sales. For example, customers' purchasing patterns might be particularly sensitive to the weather (which varies) and variation in competitors' prices and producer promotions will have some effect as well.

Within this scenario there are a number of factors which all vary to some extent but how far do the variations relate to one another? For example, is there really an association (i.e. a connection) between the variation in advertising expenditure and the variation in sales revenues? If the company can determine the extent of this association (if it does exist) it might be in a better position to decide how best to arrange its advertising in order to increase its sales. On the other hand, there might be a far greater connection between sales revenue and the weather as measured by average monthly temperature or rainfall. If this is the case, changing advertising expenditure may have very little effect on the monthly sales and a lot of money might be being wasted.

Exploring the strength of the association between pairs of factors can be carried out using **correlation analysis**. This analysis measures the degree of association between two quantities that vary (the so called 'variables'). The variation can be over time (as in the sales revenue example) or might be from one situation (e.g. place) to another. However, before we rush into using Microsoft Excel to carry out a correlation analysis for us, we must make a one or two preliminary decisions.

Suppose that we decide that we want to do a correlation analysis on the monthly revenue of one of our products and the monthly advertising expenditure used to promote that product. The first question to ask is:

"Is it reasonable to suppose that there might be an association between these two factors?"

The answer is probably 'yes' – years of experience lead us to believe that the advertising of a product is likely to influence its sales and hence the revenue received.

However, suppose that we are interested in the number of airline tickets booked per week from the UK to Spain and the average temperature in the UK in the preceding week. An association is less obvious here – there will be lots of factors influencing the purchase of airline tickets and current UK temperature, whilst it may have some effect, is possibly not a major influence.

Suppose again, that we have some data on the weekly sales of cars in New York and the weekly petrol consumption in the UK. We could carry out a correlation analysis on the figures but would there be any point? Correlation analysis would produce a figure for all three of the above examples but they would not all be useful figures – a little thought first can save a lot of wasted effort.

6.2   CORRELATION ANALYSIS

Assuming that we do decide that a correlation analysis is worthwhile, the analysis (carried out using for example Microsoft Excel) results in a number (the so-called **correlation coefficient**) which is based on the amount of joint variation shown by the two variables. The value of the correlation coefficient indicates the sense (positive or negative) and the strength of the association. The correlation coefficient is denoted by **r** or **R** and always has a value between -1 and +1.

A value close to +1 (0.9 for example) shows a strong, positive, **linear** association between two factors indicating that high values of one factor (within its range of variation) very commonly occur when there are high values of the other factor (within its range of variation). Similarly low values of the two factors tend to go together.

A value close to -1 (-0.9 for example) shows a strong, negative, **linear** association and high values of one factor commonly occur when there are low values of the other and vice versa.

In both cases, as the value of the correlation coefficient gets closer to zero the degree of the **linear** association diminishes. The scale is not, however, linear – for example a correlation coefficient of 0.8 is not twice as good as one of 0.4, all that you can say is:

If r = 0.8, the value is fairly close to 1, therefore we can assume that there is a fairly strong, positive, linear association between the variables.

If r = 0.4, the value is closer to 0 than to 1, therefore we can assume that there is a little, positive, linear association between the variables.

The stronger the linear association between the variables, the closer the magnitude of r will be to 1. As the strength of the linear association diminishes, the value of r is closer to zero. When r is zero, there is no **linear** relationship between the variables. This does not necessarily mean that there is no relationship of any kind.

Figures 6.1 and **6.2** will give values of the correlation coefficient which are close to zero.

It is important to be clear that the correlation coefficient measures the strength of the **linear** association between the variables – not just any association. Hence the correlation analysis should always be accompanied by a **scatter plot** of the data so that the nature of any relationship between the variables may be judged.
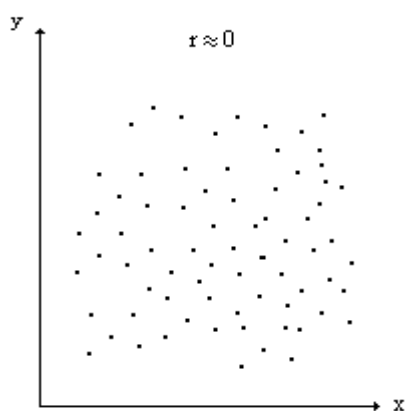
Figure 6.1 Data showing no relationship of any kind between the variables
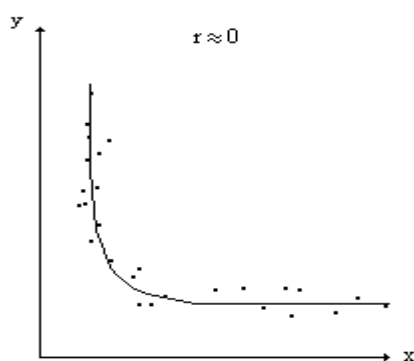
Figure 6.2 Data showing a strong non-linear relationship between the variables

Finally, once the correlation analysis is carried out, care must be taken over the interpretation of the figure. Do not jump to the conclusion that just because two variables are highly correlated, there is necessarily any causality involved. If two variables, A and B, are highly correlated, it could be any of the following:

- A causes B
- B causes A
- A and B move together, since both are affected by a third variable C
- The correlation is co-incidental

It is likely that the correlation analysis is just the beginning of the investigation!

6.3   SCATTER PLOT

Once it has been decided that an investigation of the association between two variables is likely to be worthwhile, the first thing that is useful to do is to plot a scatter diagram.

Although it does not really matter from the point of view of the correlation analysis, we will see later that we should make a sensible choice of which variable to plot on the x and y axes of our diagram. The notion of dependence arises here. We should always plot on the y axis, the variable which is likely to be dependent on, i.e. influenced by or caused by, the other variable – this 'other variable' is plotted on the x axis. The scatter plot is then simply a plot of corresponding pairs of values of the two variables. The resulting pattern of the plotted points can give us a lot of information about the association between the variables and the suitability of the data set before we spend any time on calculations.

When we examine the scatter plot we should consider several issues before reaching a conclusion about the association between the variables. The first issue is obviously whether or not there is evidence of a pattern in the plotted points – a clear pattern indicates an association or relationship between the variables. If we can assume that there is evidence of an association between the variables, the next consideration is the nature of that relationship – is it linear or non-linear.

Finally we should look to see if there are any unusual features in the plotted points. These features include 'outliers', or gaps in the pattern which may be significant, or evidence of more than one distribution being present. **Figure 6.3** illustrates these features.
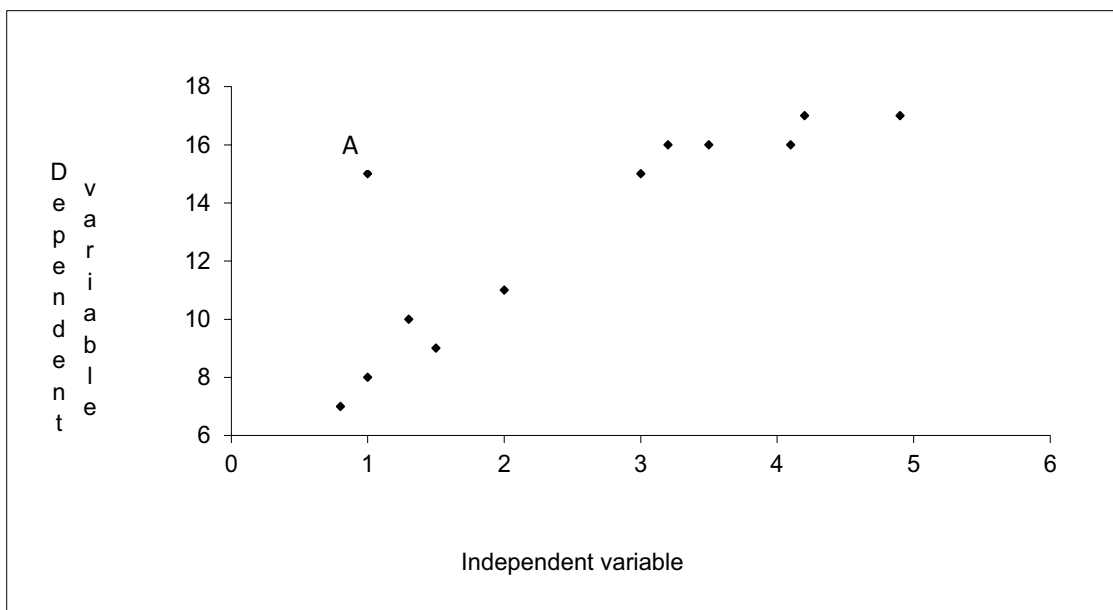
Figure 6.3 A scatter plot which shows unusual features in the data

The point marked A may be regarded as an 'outlier'. It is well away from the main spread of points. A question now arises about why it has occurred.

It may be an error in which case a check of the data will put things right. It may be correct but a non-typical value, i.e. that piece of data was recorded when the conditions were unusual, hence, it may be justifiable to omit the point (extreme care must be taken before data are removed). Alternatively the point may be perfectly legitimate and it appears to be an outlier due to the small size of the sample, i.e. a larger data set would fill in the gap and show that the relationship between these variables is less strong than it appears at present.

We must also consider why there is a gap in the data between the 6 points on the left (low values of the independent variable) and the 6 points on the right (high values of the independent variable). It may be just one of the quirks of sampling or there may be some significant issue about values of the independent variable between 2 and 3. There is the possibility in this case that we have two separate distributions present and that we should not be calculating a single correlation coefficient for the whole data set.

Example 6.1 – To investigate the Association between two variables

The example which we will analyse below is concerned with the time it takes to make deliveries. We are running a special delivery service for short distances in a city. We wish to cost the service and in order to do this we are investigating the association between the time taken for a delivery and the distance travelled.

It is obvious that there are factors other than the distance travelled which will affect the times taken – traffic congestion, time of day, road works, the weather, the road system, the driver, the type of transport, etc. However, following some initial thinking about the problem, we have come to the conclusion that the distance travelled is likely to be a key factor in determining the time taken for the delivery.

To keep the initial investigation as simple as possible, we will consider only distance travelled and its association with the time taken for the delivery. The distance is measured as the shortest practical route in miles, and the time taken is measured in minutes. (Note: it is important to be clear about exactly what is included in the time so that every measurement covers the same activities, e.g. are loading and unloading times included)

The data for the analysis is shown in **Table 6.1**. This shows the values of the two relevant variables for a number of separate journeys. These journeys have been selected at random from a much larger set of journeys for which we have data.

Distance (miles)	3.5	2.4	4.9	4.2	3.0	1.3	1.0	3.0	1.5	4.1
Time (minutes)	16	13	19	18	12	11	8	14	9	16

Table 6.1 Sample data for delivery distances and times

The first step in the correlation analysis is to plot a scatter diagram. In this situation it is the time which is likely to be dependent on the distance. Hence the time is plotted on the y axis and is called the **dependent variable**. Distance is plotted on the x axis and is called the **independent variable**. See **Figure 6.4**.
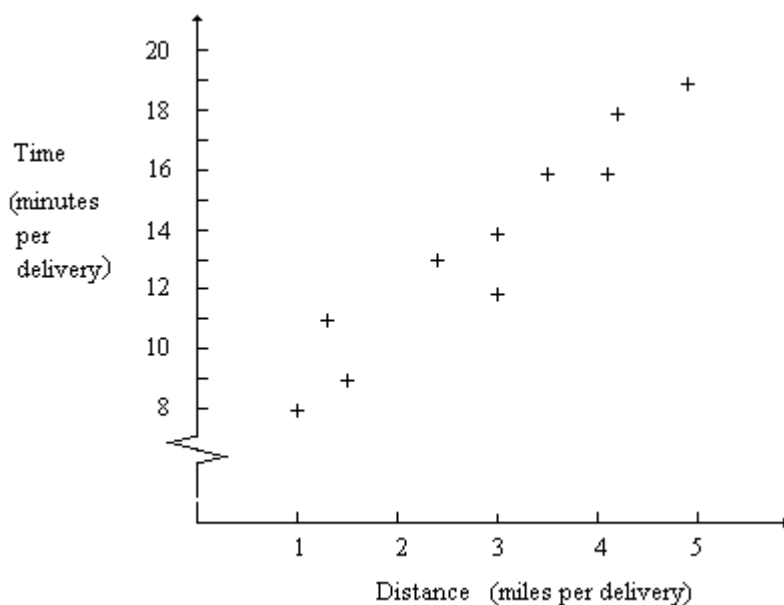
Figure 6.4 Plot of delivery times against distance for a random sample of deliveries

The plot in **Figure 6.4** shows a general increase of time with distance. It also looks as though the plotted points cluster around a straight line. The data do not exhibit any unusual features.

Hence, it would appear that there is strong evidence of a linear association between these two variables. The points are not all exactly on a straight line but they do fall in a general linear pattern. In fact it would be surprising if the points were in a perfect straight line in view of all of the other factors which we know can affect journey time.

We can now carry out the correlation analysis using a Microsoft Excel function in order to find out how strongly correlated the variables are.

6.4 STRENGTH OF THE LINEAR ASSOCIATION – THE CORRELATION COEFFICIENT, R

The strength of the linear association is measured by the **Pearson product moment correlation coefficient, r**. This can be calculated using the Microsoft Excel function correl().

In Microsoft Excel if you type *=correl,* the function will appear in a box then just click on the function or use the function button, fx, to open the *Insert Function* dialogue box then either type *Correl* in the *Search for a function* box and click *Go*, or, select *Statistical* in the *Select a category* box and scroll down to *correl*. When the *Function Arguments* box appears, click in Array 1 and highlight the values of one of the variables (it does not matter which one) and then click in Array 2 and highlight the values of the other variable.

6.4.1 CALCULATION OF THE CORRELATION COEFFICIENT, R

Return to **Example 6.1** above in which we were investigating the relationship between delivery times for journeys of a given distance within a city. Using the *correl()* function we find that the correlation coefficient, r, is 0.958 to 3 decimal places.

This value of the correlation coefficient is very close to +1 which indicates a very strong linear association between the delivery distance and the time taken. This conclusion confirms the subjective assessment made from the scatter diagram.

6.4.2 LINEAR RELATIONSHIPS

In the section above, we said that the scatter plot of delivery times against delivery distances indicated that there was a *linear* association between the variables – the plotted points lie roughly in a linear pattern. We could therefore *describe* the relationship between the two variables in terms of a straight line. This straight line is referred to as a **model** of the relationship between the variables.

Clearly this model will not describe the relationship *exactly* – the plotted points do not lie exactly in a straight line. However, since the value of the correlation coefficient indicates that the linear association between the variables is strong, the linear model will fit the data well – the discrepancy between the actual data points and the corresponding points on the model will be small.

The purpose of setting up the model is twofold:

- it describes the relationship between the variables over the range of the data which we are using;
- it enables us to **estimate** values of the dependent variable (journey time in the case of Example 1) for any given value of the independent variable (journey distance in the example) within the range of the data available.

The analysis used to set up appropriate linear models to describe the relationship between two variables is referred to as **linear regression analysis**.

In the next sections, we will recap the main points covered so far and investigate how to use Microsoft Excel to carry out linear regression analysis.

6.5 REGRESSION ANALYSIS

6.5.1 RECAPITULATION

The term **association** is used to refer to the relationship between variables. For the purposes of statistical analysis two aspects of relationships are defined. The term **regression** is used to describe and model the *shape* of the relationship, and the term **correlation** is used describe the *strength* of the relationship.

The general procedure in the analysis of the relationships between variables is to use a sample of corresponding values of the variables to establish the nature of any relationship. We can then develop a mathematical equation, or **model**, to describe this relationship. From the mathematical point of view, linear equations are the simplest to set up and analyse. Consequently, unless a linear relationship is clearly out of the question, we would normally try to describe the relationship between the variables by means of a linear model. This procedure is called **linear regression**.

A measure of the degree of fit of the linear model to the data is an indicator of the strength of the linear relationship between the variables and, hence, the reliability of any estimates made using this model. The measure used to assess the strength of the linear association between the variables is called the **correlation coefficient**. As well as providing a description of the relationship between two variables, a key purpose of this analysis is to enable us to estimate values of the dependent variable. If we know the value of the independent variable we can use the linear regression model to estimate the value of the dependent variable. So, for instance, if we believed that monthly advertising expenditure was closely linked to sales, we could build an appropriate model and then use advertising expenditure in a month to estimate the level of sales to be expected.

6.5.2 SIMPLE LINEAR REGRESSION MODEL

Simple linear regression is concerned with distributions of two variables. We are interested in whether or not there is any linear relationship between the two variables. The population for these distributions is made up of pairs of variable values, rather than values of a single variable. For example, we may be interested in the heights and weights of a number of people, the price and quantity sold of a product, employees' age and salary, chicken's age and weight, weekly departmental costs and hours worked or distance travelled and the time taken.

The first step in the analysis is to consider the variables qualitatively (i.e. without any numbers). What are the factors with which we are concerned? How do we think they will affect one another? etc. As an example, let us say that a poultry farmer wishes to predict the weights of the chickens he is rearing. Weight is the variable which we wish to predict; therefore weight is the **dependent variable**. We will plot the dependent variable on the y **axis**.

It is suggested to the farmer that weight depends on the chicken's age. Age is said to be the **independent variable**. This is the variable, the value of which we assume is known, which we can use to estimate weight. The independent variable will be plotted on the **x axis**. If we can establish the nature of the relationship between the age and weight of chickens, then we can predict the weight of a chicken at a given age. Any chicken for which the weight differs significantly from the prediction can be investigated.

How would we expect weight to change with age? At first, we expect the weight to increase with age, but when the chicken is fully grown, we would expect its weight to level off with small fluctuations depending on the feed or the time of year. Its weight gain and its adult weight will also depend on the breed of the chicken, and the way it is fed and housed. There are clearly a lot of factors, besides age, which will affect the weight. This process of thinking through the possible relationship, deciding on the dependent (y) and independent (x) variable, and thinking of other factors which could affect the relationship, is a very important part of all modelling. Our purpose is not "to do some linear regression", but, rather, to try to understand and explain the weight variations through the modelling. Can we explain a chicken's weight simply by looking at its age?

The probable conclusion, from the above discussion, is that there will be several factors working together to determine the precise weight of a particular chicken, but a general picture is:
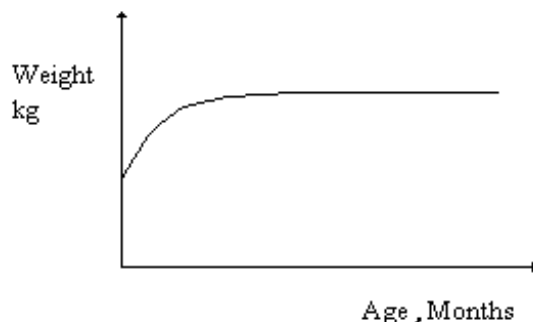
Figure 6.5 Possible relationships between weight and age of chickens of a particular breed

We are now at the stage of collecting a sample of data to test our ideas.

6.5.3 ILLUSTRATING THE SETTING UP OF A SIMPLE LINEAR REGRESSION MODEL

Refer again to **Example 6.1**, described previously. The data are repeated below for convenience. See **Table 6.2**.

Distance (miles)	3.5	2.4	4.9	4.2	3.0	1.3	1.0	3.0	1.5	4.1
Time (minutes)	16	13	19	18	12	11	8	14	9	16

Table 6.2 Sample data for delivery distances and times

We wish to explain variations in the time taken, the dependent variable (y), by introducing distance as the independent variable (x). Generally we would expect the time taken to increase as distance increases. Hence, as explained above, we first plot the sample data to obtain an impression of any relationship which exists between the variables. The scatter plot is repeated in **Figure 6.6**.
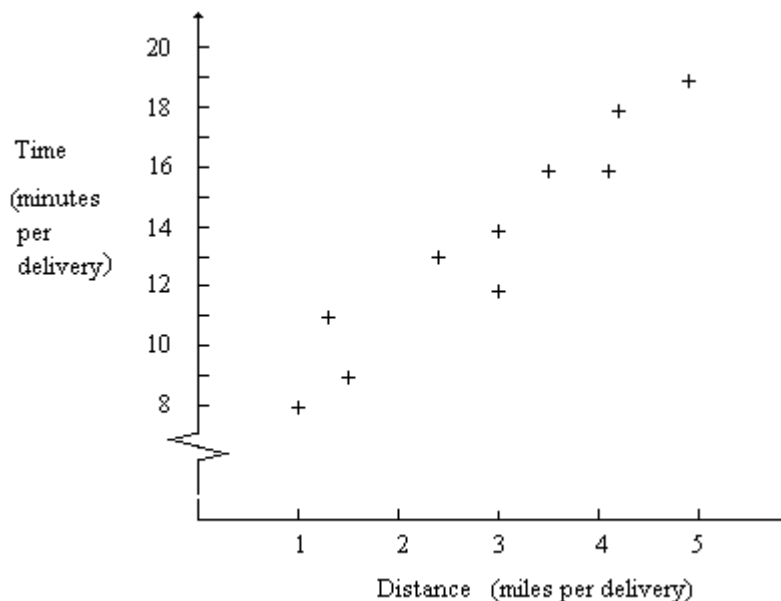
Figure 6.6 Plot of delivery times against distance for a random sample of deliveries

The plot does show a general increase of time with distance. It also looks as though the plotted points cluster around a straight line. This means that we could use a linear model to describe the relationship between these two variables. As mentioned above, the points are not exactly on a line. We do not expect this in view of all the other factors which we know can affect journey time. A linear model will be an approximation only to the true relationship between journey time and distance, but the evidence from the plot is that it is the best available.

We have decided that the best model to describe the relationship between journey time and distance is probably linear. We now require a method of finding the most suitable line to fit through our sample of points. This line is referred to as the line of **best fit** – the line which fits a closely as possible to all of the points. The points should be randomly scattered about the line as shown in **Figure 6.7**.

The line drawn through the data is a possible linear model to describe the relationship between the variables. The equation of this line may be described by:

$$y = a + bx$$

where a is a constant (sometimes, misleadingly in this context, referred to as the intercept on the y axis) and b is the slope of the line. The slope of the line is often called the **regression coefficient**. (not to be confused with the correlation coefficient)
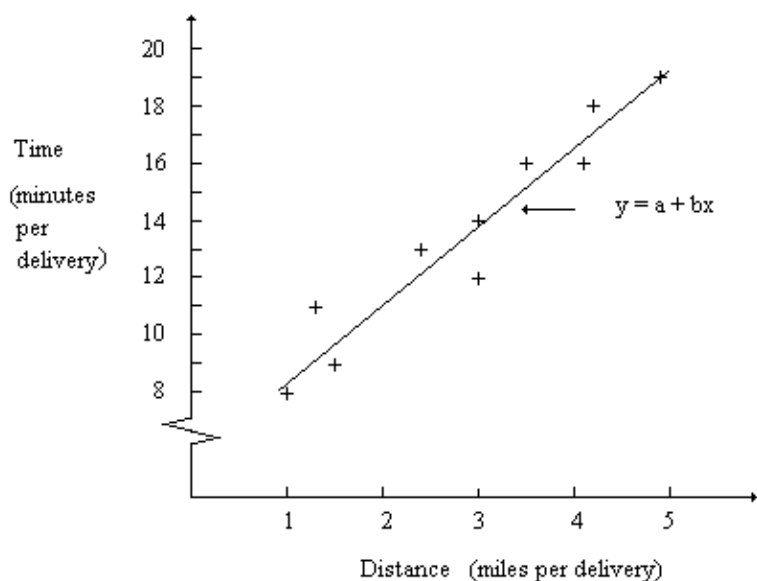
Figure 6.7 Plot of delivery times against distance for a random sample of deliveries

We can use the 'Trend line' command in Microsoft Excel to find the values of a and b.

For the purposes of this discussion, we will assume that we have a way of finding the values and that they are:

a = 5.91
b = 2.66   both to 2 decimal places

We can now insert these values in the linear model, giving:

$$y = 5.91 + 2.66x \text{ or}$$

$$\text{Delivery time (min)} = 5.91 + 2.66 \text{ Delivery distance (miles)}$$

The constant or "a" value may or may not have a sensible physical meaning.

The regression coefficient always does. It is the amount by which the y – variable increases when the x – variable increases by one unit.

In this example, the slope of the regression line (2.66 minutes per mile) can be thought of as the estimated number of minutes per mile needed for a delivery. Similarly, the constant (5.91 minutes) can be thought of as the estimated time to prepare for the journey and to deliver the goods, that is, the time needed for each journey other than the actual travelling time.

6.6 THE COEFFICIENT OF DETERMINATION, R2

We have already discussed how the strength of the linear association can be measured in terms of the correlation coefficient which is denoted by r.

The value of r can sometimes be difficult to interpret with confidence since the scale of measurement is not a linear one. If, however, we square the correlation coefficient then we produce a measure which is a ratio. This ratio is called the **coefficient of determination** and is given the symbol r^2. Since r^2 is a ratio, it can take values from 0 to 1 although it is frequently expressed as a percentage. This percentage tells us the amount of the variation in y which is explained by the linear relationship with x. A perfect linear relationship between x and y would result in $r^2 = 1$ or 100%. No linear relationship at all means $r^2 = 0$ or 0%.

Using Microsoft Excel, r^2 can be found using the RSQ() function or by simply multiplying the value of r by itself.

In our example, the coefficient of determination ($r^2 \times 100\%$) gives the percentage of the total variability in delivery time which we have explained in terms of the linear relationship with the delivery distance. In this example, the coefficient of determination is high:

$$r^2 = 0.958^2 \times 100 = 91.8\%$$

The sample model:

Time (minutes) = 5.91 + 2.66 (Distance in miles)

has explained 91.8% of the variability in the observed times. It has not explained 8.2% of the variability in the journey times. This 8.2% of the variability is caused by all of the other factors which influence the journey time which have not been included in the model.

6.7   ESTIMATION USING THE LINEAR REGRESSION MODEL

We can use the model to estimate the mean journey time for a given distance. If the distance is 4.0 miles, then our estimated mean journey time is:

y = 5.91 + 2.66 × 4.0 = 16.6 i.e. 17 minutes

An important word of caution is required here. It is not good practice to use the model to make estimations for values of the independent variable which are outside the range of the data used to set up the model. In our sample, the distances range from 1.0 mile to 4.9 miles. We have no evidence that the model is valid outside this range. The relationship between time and distance may change as distances increase. For example, a longer journey might include the use of high-speed motorway, but the sample data was drawn only from the slower city journeys. In the same way longer journeys may have to include meal or rest stops which will considerably distort the time taken. The information from which we have been working, is a sample from the population of journey times within the distance range 1 to 4.9 miles. If we wish to extrapolate to distances outside this range, we should collect more data. If we are unable to do this, we must be very careful using the model to predict the journey times. These estimations are likely to be unreliable.

6.7.1   ASSESSING THE RELIABILITY OF THE ESTIMATES

How reliable are our predictions likely to be? The simple answer is that we do not know; but we can say something about how the model would have predicted the data that we do know about. Thus we can say something about the adequacy of the model, or its predictive power. Two aspects are:

a) **The coefficient of determination (r^2)**
   'the percentage of the variation in the dependent variable y explained through knowledge of the independent variable x'
   For the delivery times example r^2 = 0.918. Thus approx. 92% of the variation in times can be explained through knowledge of (variations in) distance. The estimate is likely to be very reliable.

b) **Position of the Estimate in the Data Set**

It has been shown from experience (and common sense) that estimates which are made near the middle of the data are more reliable than those made near the ends of the distribution. The distance of 4.0 miles is towards the upper end of the data range but within the main body of the data so we have no reason to modify our conclusion that this estimate can be expected to be reliable.

6.8 CHAPTER REFERENCES

1. Morris, C, Quantitative Approaches in Business Studies, 8th edition, FT/Prentice Hall, 2012 (eBook, 7th edition 2008) chapter 3

2. Oakshott, L, Essential Quantitative Methods for Business, Management & Finance, 5th edition, Palgrave Macmillan, 2012 chapter 4

3. Wisniewski, M, Quantitative Methods for Decision Makers, 5th edition, FT/Prentice Hall, 2009 (eBook available) chapter 3

7 TIME SERIES AND FORECASTING

7.1 INTRODUCTION

All businesses have to plan their future activities. When making both short and long term plans, managers will have to make forecasts of the future values of important variables such as sales, interest rates, costs, etc.

In order to create a forecast, we must first establish the relationship between the variable which we wish to forecast – the dependent variable – and other variables which influence its value – the independent variables.

In **time series analysis** models are set up which use time as the independent variable. Any data set where values of a variable have been collected, recorded or observed over successive time periods (days, weeks, months, years etc.) is called a **time series**. In the analysis of the time series an attempt is made to identify patterns which exist in the behaviour of the dependent variable over time. These patterns are then used as predictors of future values. This process of using the behaviour patterns of past data to predict future behaviour is called **forecasting**.

For example, we may wish to model the variability in sales by looking at the way in which they have changed with time over the last two years, say, ignoring other factors. If we can identify the past pattern in the sales, we can use it to forecast future values. Care is required since the historical pattern is not always relevant for particular forecasts. We are making the assumption that past patterns will continue. **Note: this is a key assumption, the validity of which needs to be constantly borne in mind**. A company may deliberately plan to change its pattern if, for example, it has been making a loss. There may be large external factors which completely modify the pattern. There may be a major change in raw material prices, world inflation may suddenly increase or a natural disaster may affect the business unpredictably. Even if there are no obvious changes in the conditions relating to the time series, we may find that the patterns it contains are difficult to identify and the resulting models are very complex.

In **section 7.2**, we begin by describing a time series in terms of constituent components, such as:

- **Trend**
- **Seasonal Variation and**
- **Cyclical Variation**.
- **the Irregular or Random element**

The first three components can then be combined in a number of ways to form a model of the time series. By definition we cannot model the Irregular element. We will look at one specific model – the additive component model. As the name implies, the components are added together. You should note that the techniques described are not the only, nor necessarily the best, forecasting methods for any particular forecasting situation. There are many more sophisticated statistical techniques available. As well as the quantitative techniques, there are qualitative methods which must be used when there is little or no past data. The Delphi technique, which uses experts to guess what might happen, and the scenario writing method are examples of these.

7.2   TIME SERIES COMPONENTS

The overall pattern of a time series can be seen most easily if the dependent variable is plotted against time on a scatter diagram. It is usual practice to join the plotted points by straight lines so that the pattern can be seen more easily. A typical time series is illustrated in **Figure 7.1**. We can see that the quarterly sales figures are gradually increasing overall but with some considerable fluctuation.
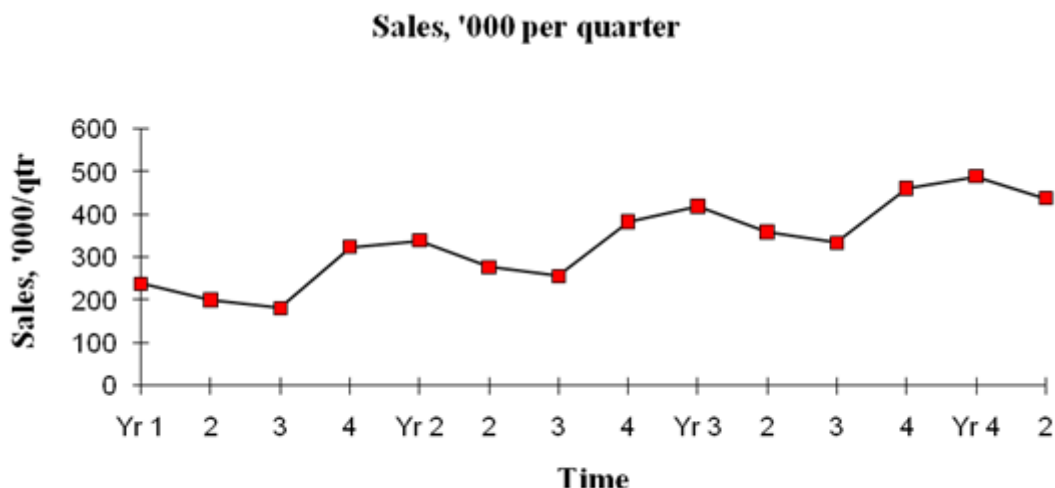
Figure 7.1 A typical Time Series

The value of a variable, such as sales, will change over time due to a number of factors. For example, a company may be expanding the market for a new product hence there will be an upward movement in the sales of the product as time progresses. The general, long-term change in the value of a variable over time is referred to as the **trend**, T. The examples we will use in the following sections are ones in which the trend is linear, i.e. the rate of change of the dependent variable is more or less constant from one time period to the next. In practice, there may be no trend at all, with the dependent variable values fluctuating about a fixed value, or, more likely, there will be a non-linear trend. A non-linear trend is more difficult to model and it is more difficult to obtain a forecast of the trend.

The graphs below represent the trend in sales at different stages of a product's life cycle. There is an underlying upward trend for the newly launched product, and a dying curve for the old product reaching the end of its economic life. It is difficult to fit a model to these trend curves.

Figure 7.2 Life cycle of a new product
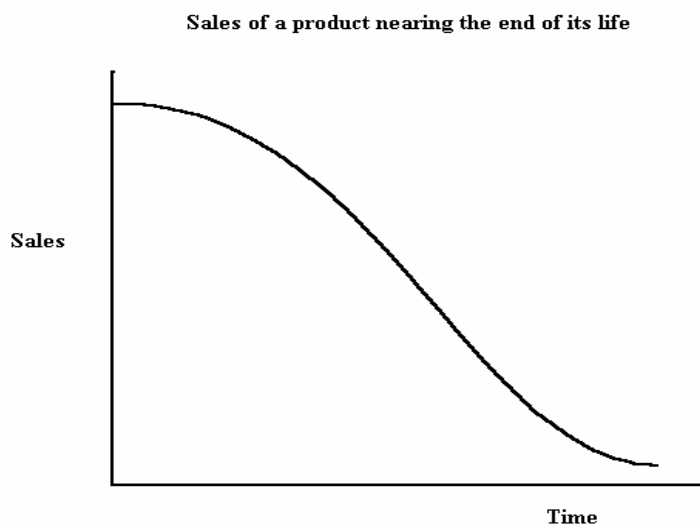
Figure 7.3 Life cycle of a product at the end of its life

Referring again to **Figure 7.1**, we can see that there is also a cyclical variation apparent in the values as well as an upward trend. If these regular fluctuations occur in the short term (usually less than a year), they are referred to as **seasonal variation**, S. Longer term fluctuation is called simply **cyclical variation**.

The **seasonal component** in a time series can be defined as 'a wave like pattern **within** a year which **repeats itself** year after year' e.g. 'quarterly' variations. However, the term 'seasonal' effects can equally be interpreted to encompass monthly, weekly or daily variations. For instance in daily data a "seasonal" effect would be a repeating day-by-day pattern within a week. (*You might like to think of examples of daily patterns within a week; weekly patterns within a month; monthly /quarterly patterns within a year?*)

In **Figure 7.1**, the main time period is a year and we can see that the sales fluctuate in a cycle over the four quarters of the year. For each year, the sales drop from Quarter 1 to Quarter 2 and again to Quarter 3 then they rise steeply to Quarter 4. This pattern is repeated each year on top of a gradually upward drift in the total sales. The drop in sales from Quarter 2 to Quarter 3, for example, should not cause any concern, as long as we know that this happens each year. We might, however, be concerned if the sales in Quarter 3 of one year are less than those of Quarter 3 of the previous year.

If we use annual data, we cannot identify a seasonal pattern. Any fluctuation about the trend in annual data would be described by the cyclical component. The cyclical variation is a gradual cyclical pattern about the trend. It is usually attributable to variation in the macro-economic environment. The effect is seen only in long-term data, covering ten, fifteen or twenty years. Over this length of time large scale economic factors have time to swing back and forth, causing fluctuations about the trend. The length of a cycle is referred to as its period; the time between successive peaks or successive troughs. Economic cyclical patterns tend to repeat in a **yearly** data set every four/five years. To identify such a component usually needs at least three cycles of the data (i.e. at least 12 years). We will not consider how to determine such cyclical effects in time series data in this unit.

The final component in the time series is the **error**, **residual, irregular or random** element; i.e. the part of the actual observation which is 'left over' after we have accounted for the trend, seasonal variation and cyclical variation, i.e.

Error or residual = actual value – forecast value.

These residuals are caused by unpredictable, rare or 'one-off' events. The residuals should, therefore, show no pattern and be relatively small. A large residual could indicate a recording or calculation error and should always be checked. Alternatively, it might indicate an unusual event which we should be able to identify. We would then need to decide whether the occurrence of this event had caused a sufficiently significant jump in the time series that we should eliminate this part of the data set.

7.2.1 MODELS FOR TIME SERIES

In the analysis of a time series, we attempt to identify the components which are present and to build a model which combines them in an appropriate way.

For example, if the time series indicates that the size of the seasonal variation is independent of the size of the actual values, then it would be appropriate to add the components in the model

$$A = T + S + E$$

where A is the actual observation, T is the trend, S is the seasonal variation and E is the residual or error component.

This type of model would be appropriate if we consider the sales of a product which are steadily increasing over time, but the increase in sales above the trend in quarter 4, is always 100 units, regardless of the value of the sales volume. For example, if the sales volume is 1000 units, then the seasonal effect in quarter 4 would be to increase sales to 1100 units. However, if the sales volume was 10,000 units, the same seasonal effect would increase the sales in quarter 4 to 10,100 units.

If, however, the size of the seasonal variation changes as the underlying trend and cyclical variation changes, then it would be more appropriate to multiply the components in the model:

A = T S E

In this case, if we consider the example of the sales of a product, the size of the seasonal variation is a fixed percentage of the sales volume. If the sales volume is 100 units, then the seasonal effect in quarter 4 may be to increase sales by 10% to 110 units. However, if the sales volume was 1000 units, the same seasonal effect would increase the sales in quarter 4 to 1100 units.

These two models are referred to as the **additive component model** and the **multiplicative component mode**l, respectively.

7.2.2 THE ANALYSIS OF AN ADDITIVE COMPONENT MODEL

The first step in the analysis of any time series is to plot the data so that the likely components may be identified and a suitable model selected.

Example 7.1 – Analysis of a time series

The number of employees at Systems plc between 2004 and 2011 is given in **Table 7.1**.

Year	No. of Employees, '000)
2004	1.1
2005	2.4
2006	4.6
2007	5.4
2008	5.9
2009	8.0
2010	9.7
2011	11.2

Table 7.1 The number of employees at Systems plc 2004–2011

Solution: Since the data are annual figures, there will be no seasonal component and there is insufficient data to identify any cyclical variation, therefore the components required in the model are the trend and the residual. Plot the data. See **Figure 7.4.**
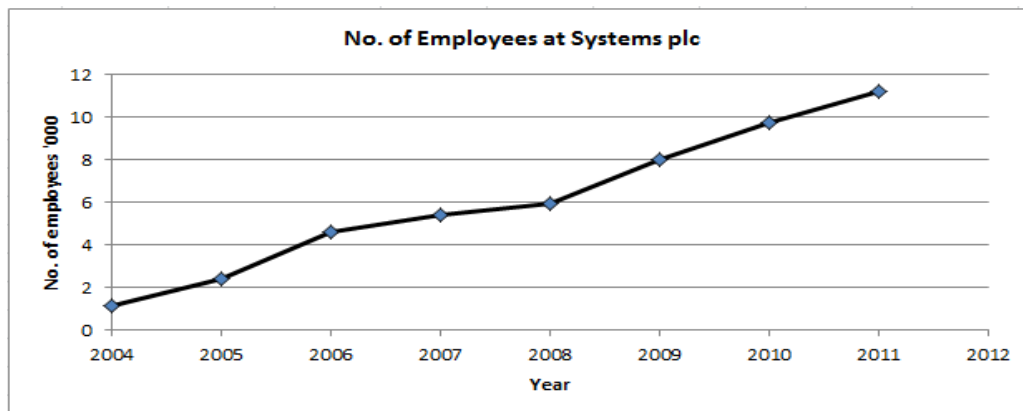
Figure 7.4 Time Series Plot for Example 1

The scatter diagram shows that the number of employees at Systems plc has increased fairly steadily from one year to the next. We may therefore assume that a linear model would be suitable to describe the trend. The residual variation looks fairly uniform about a possible linear model, hence we may assume that a suitable time series model is:

$$A = T + E$$

and that the trend may be modelled by a straight line. The linear trend can be represented by:

$$A = a + bt$$

where A is the estimated number of employees ('000), t is the time in years (t = 1 in 2004) and a and b are constants. (b is the slope of the line)

The appropriate values for a and b can be found by using Microsoft Excel to fit a trend line to the data. The equation of the line is an option in this command. If we use the data in the following format with the year 2004 as year 1,

Time, t (years)	No. of Employees, ('000), A
1	1.1
2	2.4
3	4.6
4	5.4
5	5.9
6	8.0
7	9.7
8	11.2

Table 7.2 The number of employees at Systems plc over time

then the trend line has the equation:

A = -0.279 + 1.404t ('000 of employees). See the figure below.

Figure 7.5 Time Series Plot for Example 1

We can use this model to estimate the number of employees for each year and then we can calculate the residual variation. See **Table 7.3**.

The residual variation, E = actual value – forecast value.

The maximum size of the residuals is approximately 17% with both positive and negative values. This indicates that the model is a reasonable, but not a very good, fit to the data. Estimates made using this model are likely to be fairly reliable – no worse than ± 17%.

Time, t (years)	No. of Employees, A, ('000)	Estimated no. employed Â, ('000)	Residual, E = A – Â ('000)
1	1.1	1.13	-0.03
2	2.4	2.53	-0.13
3	4.6	3.93	0.67
4	5.4	5.34	0.06
5	5.9	6.74	-0.84
6	8.0	8.14	-0.14
7	9.7	9.55	0.15
8	11.2	10.95	0.25

Table 7.3 Calculation of the estimated number of employees and the residual variation

Example 7.2 – Analysis of a time series

The data in **Table 7.4** shows the sales of Product X by Statplan plc over the 13 quarters to March 2014.

Date	Quarter Number	Sales per quarter ('000)
Jan–Mar 2011	1	239
Apr–Jun	2	201
Jul–Sep	3	182
Oct–Dec	4	297
Jan–Mar 2012	5	324
Apr–Jun	6	278
Jul–Sep	7	257
Oct–Dec	8	384
Jan–Mar 2013	9	401
Apr–Jun	10	360
Jul–Sep	11	335
Oct–Dec	12	462
Jan–Mar 2014	13	481

Table 7.4 Sales of Product X over the 13 quarters to March 2014

Solution: The first step is to plot the data. See **Figure 7.6**.

Figure 7.6 Time Series Plot for Example 2

We can see from **Figure 7.6** that the quarterly sales for Statplan plc show an upward trend with a four quarter seasonal pattern. The size of the seasonal fluctuation appears to be unaffected by the size of the sales, therefore an additive model seems appropriate. Hence, we will assume that the appropriate time series model is:

$$A = T + S + E$$

We now need to analyse the data to extract the seasonal components and the trend.

For both the additive and the multiplicative component models the general procedure is the same.

Step 1:   Calculate the seasonal components.

Step 2:   Remove the seasonal component from the actual values. This is called deseasonalising the data. Model the trend from these deseasonalised figures.

Step 3:   Use the trend model to calculate the forecast trend figures and then deduct these trend figures from the deseasonalised figures to leave the errors or residuals.

The seasonal components can be separated from the trend using a technique called **moving averages**. The technique uncovers the historical trend by smoothing away the seasonal fluctuations. However, the moving average trend is not used to forecast future trend values because there is too much uncertainty involved in extending such a series.

7.2.3 CALCULATE THE SEASONAL COMPONENTS FOR AN ADDITIVE MODEL

Step 1: We will continue with **Example 7.2** in the previous section which relates to the quarterly sales of Statplan plc. We have already decided that an additive model is appropriate for these data; therefore the actual sales may be modelled by

$$A = T + S + E$$

To eliminate the seasonal components, we will use the method of moving averages.

If we add together the first four data points, we obtain the total sales for 2011, dividing this by 4 gives the quarterly average for 2011, i.e.

(239 + 201 + 182 + 297)/4 = 229.75.

This figure contains no seasonal component because we have averaged them out over the year. We are left with an estimate of the trend value for the middle of the year, that is, for the mid-point of quarters 2 and 3. If we move on for 3 months, we can calculate the average quarterly figure for April 2011 to March 2012 (251), for July 2011 to June 2012 (270.25), and so on. This process generates the 4 point moving averages for the quarterly sales. The set of moving averages represent the best estimate of the trend in the demand.

We now use these trend figures to produce estimates of the seasonal components. We calculate:

A − T = S + E

Unfortunately, the estimated values of the trend given by the 4 point moving averages, are for points in time which are different to those for the actual data. The first value, 229.75, represents a point in the middle of 2011, exactly between the April–June, and the July–September quarters. 251 fall between the July-September and the October–December actual figures. We require a deseasonalised average which corresponds with the figure for an actual quarter. The position of the deseasonalised averages is moved by averaging pairs of values. The first and second values are averaged, centring them on July–September 2011:

(229.75 + 251)/2 = 240.4

is the deseasonalised average for July–September 2011. This deseasonalised value, called the centred moving average, can be compared directly with the actual value for July to September 2011 of 182. Notice that this means that we have no estimated trend figures for the first two or last two quarters of the time series. The results of these calculations are shown in **Table 7.5**.

Quarter Number	Sales per quarter, A ('000)	4 quarter total, ('000)	4 quarter moving average ('000)	Centred moving average, T ('000)	Estimated seasonal component A-T=S+E ('000)
1	239		-	-	
2	201		-	-	
		919	229.75		
3	182			240.4	-58.4
		1004	251.00		
4	297			260.6	+36.4
		1081	270.25		
5	324			279.6	+44.4
		1156	289.00		
6	278			299.9	-21.9
		1243	310.75		
7	257			320.4	-63.4
		1320	330.00		
8	384			340.3	+43.8
		1402	350.50		
9	401			360.2	+40.8
		1480	370.00		
10	360			379.8	-19.8
		1558	389.50		
11	335			399.5	-64.5
		1638	409.50		
12	462	-	-	-	
13	481	-	-	-	

Table 7.5 Calculation of the centred 4 point moving average trend for the model A = T + S + E

This example can be translated into Microsoft Excel as following:

	A	B	C	D	E	F
1	Quarter Number	Sales per quarter, A ('000)	4 quarter total, ('000)	4 quarter moving average ('000)	Centred moving average, T ('000)	Estimated sea-sonal component A-T=S+E ('000)
2	1	239		-	-	
3	2	201		-	-	
4			= sum(B2:B7) (919)	= C4/4 (229.75)		
5	3	182			= (D4+D6)/2 (240.4)	=B5-E5 (-58.4)
6			= sum(B3:B9) (1004)	= C6/4 (251.00)		

7	4	297			= (D6+D8)/2 (260.6)	=B7-E7 (+36.4)
8			= sum(B5:B11) (1081)	= C8/4 (270.25)		
9	5	324			= (D8+D10)/2 (279.6)	=B9-E9 (+44.4)
10			= sum(B7:B13) (1156)	= C10/4 (289.00)		
11	6	278			= (D10+D12)/2 (299.9)	=B11-E11 (-21.9)
12			= sum(B9:B15) (1243)	= C12/4 (310.75)		
13	7	257			= (D12+D14)/2 (320.4)	=B13-E13 (-63.4)
14			= sum(B11:B17) (1320)	= C14/4 (330.00)		
15	8	384			= (D14+D16)/2 (340.3)	=B15-E15 (+43.8)
16			= sum(B13:B19) (1402)	= C16/4 (350.50)		
17	9	401			= (D16+D18)/2 (360.2)	=B17-E17 (+40.8)
18			= sum(B15:B21) (1480)	= C18/4 (370.00)		
19	10	360			= (D18+D20)/2 (379.8)	=B19-E19 (-19.8)
20			= sum(B17:B23) (1558)	= C20/4 (389.50)		

21	11	335			= (D20+D22)/2 (399.5)	=B21-E21 (-64.5)
22			= sum(B19:B24) (1638)	= C22/4 (409.50)		
23	12	462	-	-	-	
24	13	481	-	-	-	

Table 7.6 Microsoft Excel example for calculating index numbers
The data shown in Table 7.6 are the data from Table 7.5.
Cells C1:F24 contain the functions to be entered and the values that would appear in those cells (in brackets).

For each quarter of the year, we have estimates of the seasonal components, which include some error or residual. There are two further stages in the calculations before we have usable values for the seasonal components.

Firstly, we average the seasonal estimates shown in **Table 7.5** for each of the seasons. This should remove some of the errors. Finally we adjust the averages, moving them all up or down by the same amount, until their total is zero[7]. This is done because we require the seasonal components to average out over the year, i.e. the total seasonal effect should be zero. The correction factor required is: (the sum of the estimated seasonal values)/4
The procedure is shown in **Table 7.7**.

Year	Quarter of the year			
	1	**2**	**3**	**4**
2011	-	-	-58.4	+36.4
2012	+44.4	-21.9	-63.4	+43.8
2013	+40.8	-19.8	-64.5	-
Total	+85.2	-41.7	-186.3	+80.2
Average	+85.22 = +42.6	-41.72 = -20.85	-186.33 = -62.1	+80.22 = +40.1
Estimated seasonal component	+42.6	-20.85	-62.1	+40.1
Adjusted seasonal component	+42.7	-20.8	-62.0	+40.2

Sum = -0.25
Adjustment = +0.0625

Sum = 0*
* allowing for rounding
(the adjustment is so that the seasonal components total zero)

Table 7.7 Calculation of the Average Seasonal Components

The estimated seasonal components are:

Jan–Mar	+42700 items	Apr–Jun	- 20800 items
Jul–Sep	- 62000 items	Oct–Dec	+40200 items

7.2.4 DESEASONALISE THE DATA TO FIND THE TREND

Step 2: is to deseasonalise the basic data. The procedure, set out in **Table 7.7**, is to deduct the appropriate seasonal factor from the actual sales for each quarter, i.e.

$$A - S = T + E.$$

These re-estimated trend values, with residual variation (errors), can be used to set up a model for the underlying trend. The values are superimposed on the original scatter diagram. See **Figure 7.8**.

Date	Quarter Number	Sales per quarter ('000)	Seasonal component, S ('000)	Deseasonalised Sales ('000), A-S=T+E
Jan–Mar 2011	1	239	+42.7	196.3
Apr–Jun	2	201	-20.8	221.8
Jul–Sep	3	182	-62.0	244.0
Oct–Dec	4	297	+40.2	256.8
Jan–Mar 2012	5	324	+42.7	281.3
Apr–Jun	6	278	-20.8	298.8
Jul–Sep	7	257	-62.0	319.0
Oct–Dec	8	384	+40.2	343.8
Jan–Mar 2013	9	401	+42.7	358.3
Apr–Jun	10	360	-20.8	380.8
Jul–Sep	11	335	-62.0	397.0
Oct–Dec	12	462	+40.2	421.8
Jan–Mar 2014	13	481	+42.7	438.3

Table 7.8 Calculation of the deseasonalised data
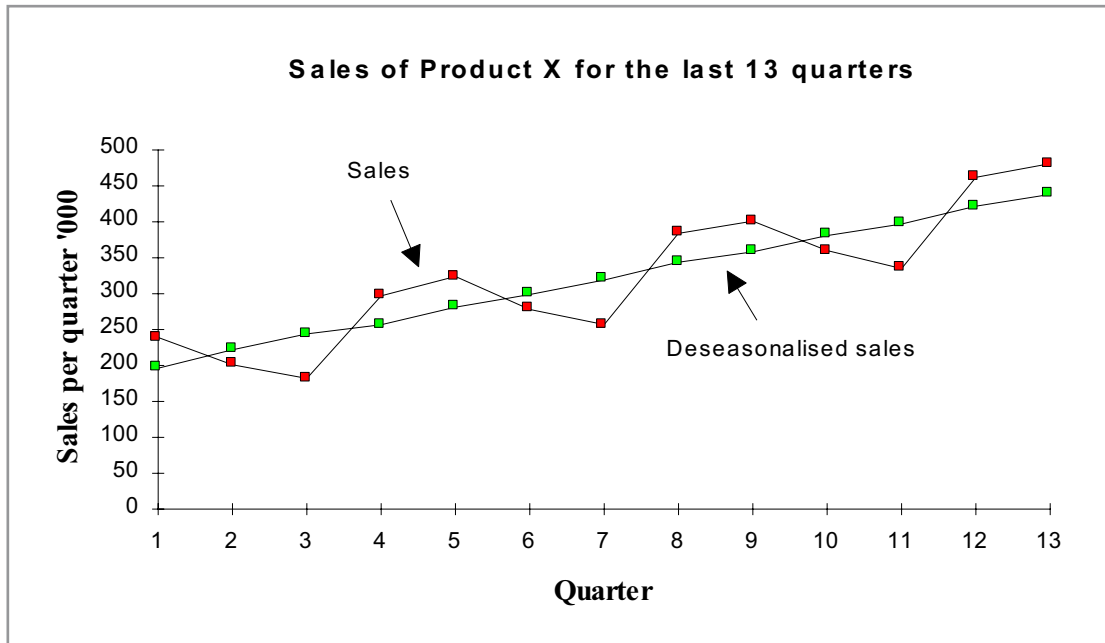
Figure 7.7 Actual and deseasonalised sales for Example 2

The deseasonalised sales show a clear trend which is approximately linear. We can set up a linear model for this trend. The equation of the trend line is:

T = a + b quarter number

where a and b represent the intercept and slope of the line.

Fitting a linear trend line (as in **Example 7.1**) we find that:

b = 19.973   and   a = 180.038

Hence the equation of the trend model may be written:

Trend sales ('000s) = 180.0 + 20.0 quarter number

7.2.5 CALCULATE THE RESIDUAL VARIATION

Step 3 : in the procedure is to calculate the errors or residuals. The model is:

A = T + S + E

We calculated S and T in the previous sections. We can now deduct each of these from A, the actual sales, to find the residuals in the model. See **Table 7.9**.

Date	Quarter Number	Sales per quarter, A ('000)	Seasonal component, S ('000)	Trend component, T ('000) (from model)	Residual, E ('000) A-S-T=E
Jan–Mar 2011	1	239	+42.7	200	-3.7
Apr–Jun	2	201	-20.8	220	+1.8
Jul–Sep	3	182	-62.0	240	+4.0
Oct–Dec	4	297	+40.2	260	-3.2
Jan–Mar 2012	5	324	+42.7	280	+1.3
Apr–Jun	6	278	-20.8	300	-1.2
Jul–Sep	7	257	-62.0	320	-1.0
Oct–Dec	8	384	+40.2	340	+3.8
Jan–Mar 2013	9	401	+42.7	360	-1.7
Apr–Jun	10	360	-20.8	380	+0.8
Jul–Sep	11	335	-62.0	400	-3.0
Oct–Dec	12	462	+40.2	420	+1.8
Jan–Mar 2014	13	481	+42.7	440	-1.7

Table 7.9 Calculation of residuals for the additive component model

The residuals are all small at about 1 or 2% of the actual sales. This implies that the historical pattern is highly consistent and should give a good short-term forecast.

Please note the seasonal components always sum to zero:

i.e. 42.7 +-20.8+-62.0+40.2=0

7.2.6 FORECASTING USING THE ADDITIVE MODEL

The forecast for the additive component model set up in **Example 7.2** is:

F = T + S   '000s units per quarter

Where the trend component T ('000 units) = 180 + 20 × quarter number, and the seasonal components, S, are:

+42.6 for January to March	-20.7 for April to June
-62.0 for July to September	+40.1 for October to December

The quarter number for the next 3 monthly period, April to June 2014, is 14, therefore the forecast trend is:

$$T_{14} = 180 + 20 \times 14 = 460 \text{ '000 units per quarter}$$

The appropriate seasonal component is -20.7 ('000 units). Therefore the forecast for this quarter is:

$$F(\text{April to June 2014}) = 460 - 20.7 = 439.3 \text{ ('000 units)}$$

It is important to remember that the further ahead the forecast, the more unreliable it becomes. We are assuming that the historical pattern continues uninterrupted. This assumption may hold for short periods, but is less and less likely to be true the further we go into the future.

7.3   EVALUATION AND COMPARISON OF MODELS

A key task of the user of a (forecasting) model (or in fact any model) is to evaluate it. This task involves asking "how good would the model have been if we had used it to predict what has actually happened?"

It often happens that we try out several models of forecasting in a particular situation. The question arises "which model would have performed best?", and thus "which one might we adopt for our future forecast?"

We compare models using various '**error statistics**'. These are summary measures of how well/badly the forecast would have done compared to what actually happened.

The summary is a mean (or average) of the errors that would have occurred. In this module we will consider:

the **mean error** (ME)

$$ME \quad = \quad \frac{\text{sum of the errors}}{n}$$

the **mean absolute deviation** (MAD

$$MAD \quad = \quad \frac{\text{sum of the errors ignoring the sign}}{n}$$

the **root mean square error** (RMSE)

$$RMSE = \sqrt{\frac{\text{sum of (error}^2)}{n}}$$

where n is the number of time periods for which we have both observed values and forecast values.

The root mean square error (RMSE)) is a measure of the differences between the values predicted by a model and the values actually observed from the phenomenon being modeled or estimated. RMSE is a good measure of accuracy and acceptable values depend upon the objective of the model and phenomenon being modeled.

For example, consider modeling the response service provided by NHS ambulances, here you want your model to be as accurate as possible and the amount of error tolerance you may wish to accept will be less, say up to 2%. On the other hand if you are modeling the future performance of your favorite football team then you may wish to consider higher tolerance levels, say error levels up to 5 to 10%. Generally, acceptable values depend upon the situation being modeled and the accuracy levels you are willing to accept.

7.4   CHAPTER REFERENCES

1. Hanke, JE and Reitsch, AG Business Forecasting, 9th Edition, Prentice Hall, 2015, chapter 4

2. Oakshott, L, Essential Quantitative Methods for Business, Management & Finance, 5th edition, Palgrave Macmillan, 2012 chapter 4

3. Wisniewski, M, Quantitative Methods for Decision Makers, 5th edition, FT/Prentice Hall, 2009 (eBook available) chapter 3

8   FINANCIAL MODELLING

8.1   INTRODUCTION

We will now turn our attention to some basic aspects of Financial Modelling which may be useful to us in understanding modern financial environments.

In this chapter, we will further develop our knowledge and skills in modelling using financial variables and will focus on developing solutions that may involve simple and complex business variables.

8.2   COSTS, REVENUE, PROFIT AND CONTRIBUTION

8.2.1   COSTS

Costs arise in all organisations in a wide variety of ways. In general, we spend time and effort trying to reduce the value of the costs which occur. At the very least, we have to ensure that they are all properly accounted for. In Business Accounting you will discuss how/when costs arise and how they should be recorded. In Business Analysis we will concentrate on modelling the relationship between some types of cost and other business variables.

Many of the costs which arise can be divided into two parts: the fixed part and the variable part. The cost of producing or selling a product will comprise:

-   **fixed costs**: costs incurred irrespective of how many items you produce or sell – these include premises costs, cost of utilities, staff wages etc.
-   **variable costs**: costs incurred only when you produce or sell an item – these include raw material costs; packaging costs, etc.

Hence we can build a simple model for the total costs, TC:

TC = total fixed costs + total variable costs

If we produce q items per day at £v per item, the total variable cost is £v×q per day.

TC = total fixed costs + v×q

If we produce q items per day at £2 per item, the total variable cost is £2×q per day.

If the production incurs fixed costs of £300 per day, then the total cost per day, TC, may be described by the model:

TC = 300 + 2q   £ per day

We can use this model to calculate the total cost per day which would be incurred for various levels of production.

8.2.2 REVENUE

Revenue, as we know, is the total sum received from the sale of products or services. At its simplest level, the revenue depends on:

the number of items sold, q
the price charged for each item, p

Hence we can model revenue by

R = the price charged for each item × the number of items sold
R = p × q   £ per day

if p is the price in £ per item and q is the quantity sold per day.
If we sell q items per day at £5 per item, the total revenue is £5×q per day.

The revenue model is:

R = 5×q   £ per day

We can use this model to calculate the total revenue per day which would be incurred for various levels of sales.

8.2.3 PROFIT

The objective of many organisations is to make a profit. Even non-profit making organisations usually require to at least break-even.

At the simplest level, profit may be regarded as the excess of revenue over **all** costs (at this level we will ignore all issues relating to taxes).

Hence, a simple model for profit is:

$$P = \text{Total revenue} - \text{Total costs} = R - TC$$
$$P = p \times q_S - (\text{fixed cost} + \text{total variable cost})$$
$$P = p \times q_S - (\text{fixed cost} + v \times q_P)$$

Where q_S is the quantity and q_P is the quantity produced in a given period of time.

In the simplest case $q_S = q_P$ i.e. we sell all that we produce.

If we produce q items per day at £2 per item, and sell them <u>all</u> for £5 per item, the total profit model is:

$$P = R - TC = 5 \times q - (FC + 2 \times q) \qquad \text{£ per day}$$

Where FC denotes the total fixed costs per day

$$P = (5 - 2) \times q - FC = 3 \times q - FC \qquad \text{£ per day}$$

We can use this model to calculate the total profit per day which would be earned for various levels of production and sales.

8.2.4 CONTRIBUTION

Fundamentally, contribution may be regarded as the excess of revenue over the total **variable** costs. This excess is the contribution which our trading activity is making towards covering the fixed costs and creating a profit (if required).

Hence, a simple model for contribution is:

Contribution = total revenue – total variable cost therefore:

$$\text{Contribution} = p \times q_S - v \times q_P$$

Where q_S is the quantity supplied and q_P is the quantity produced in a given period of time.

In the simplest case $q_S = q_P$ i.e. we sell all that we produce.

If we buy-in[8] q items per day at £2 per item, and sell them *all* for £5 per item, the total contribution model is:

Contribution = R – total variable cost = 5×q –2×q   £ per day
Contribution = (5 – 2)×q = 3×q   £ per day

We can use this model to calculate the total contribution per day which would be earned for various levels of purchase/production and sales.

8.3   USING A GRAPH TO ILLUSTRATE FINANCIAL MODELS

It is clearly useful to see how costs, revenues, profits and contributions behave for various levels of production quantity and sales. For example, we may want to know how many items we must sell in order to break-even. In addition, we may want to find the production level which will maximise the profit, or minimise the average cost per item, and so on.

If we can assume that all items which are produced/bought-in are sold, then a simple way of investigating the behaviour of the cost, revenue, profit and contribution models is to plot graphs.

We will discuss these ideas by means of an example.

8.3.1 SIMPLE LINEAR MODELS

Example 8.1 – Plotting a linear equation

Dickson Mills plc manufactures fabrics of various kinds. The production cost of cotton/poplin sheeting is given by:

$$C = 500 + 2q \qquad \text{(£ per day)}$$

Where q is the daily production in metres per day. The equation holds over the range 0 to 200 metres per day.

Solution: q is the independent variable and its range of values is specified i.e. 0 to 200.

What are the fixed costs and unit variable costs for this product?

When you have thought about the answer to this question, look at the footnote on the next page to see if you are correct.

Using Microsoft Excel, we can plot the graph shown in **Figure 8.1**.
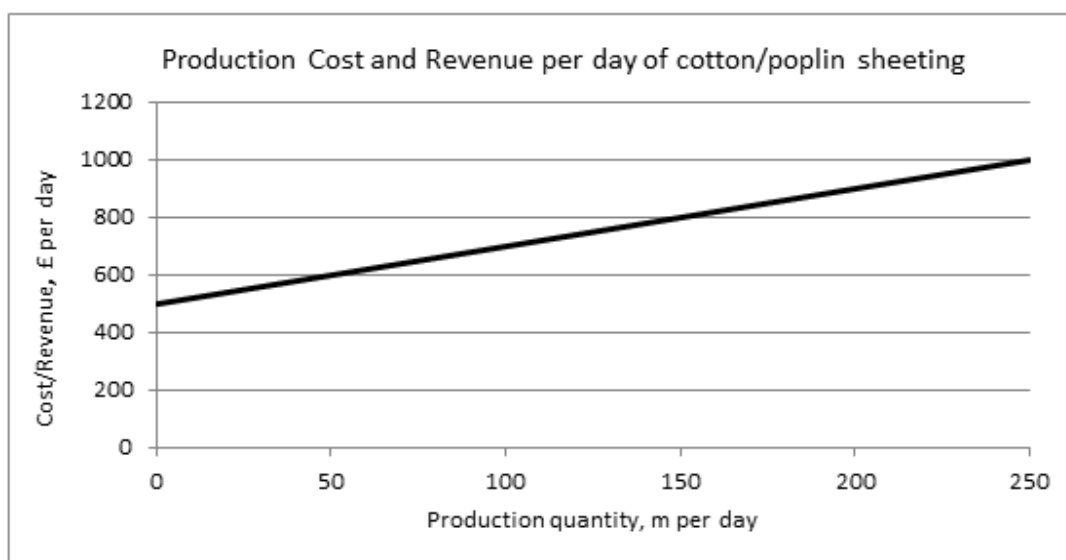
Figure 8.1 Production Cost at Dickson Mills plc

- How can we find the value of the fixed costs from the graph?
- How can we find the value of the unit variable cost from the graph?
        If C = 500 + 2q   (£ per day),
        then C = 500 (the fixed costs) when q = 0.

Hence if no items are produced the total cost is equal to the fixed costs. The fixed cost is therefore given by the point at which the line cuts the Total cost axis. The unit variable cost is the cost of producing 1 m of cloth, i.e. the increase in total cost for each additional metre produced – this is equivalent to the slope of the line. Hence the slope of the line yields the unit variable cost of production.
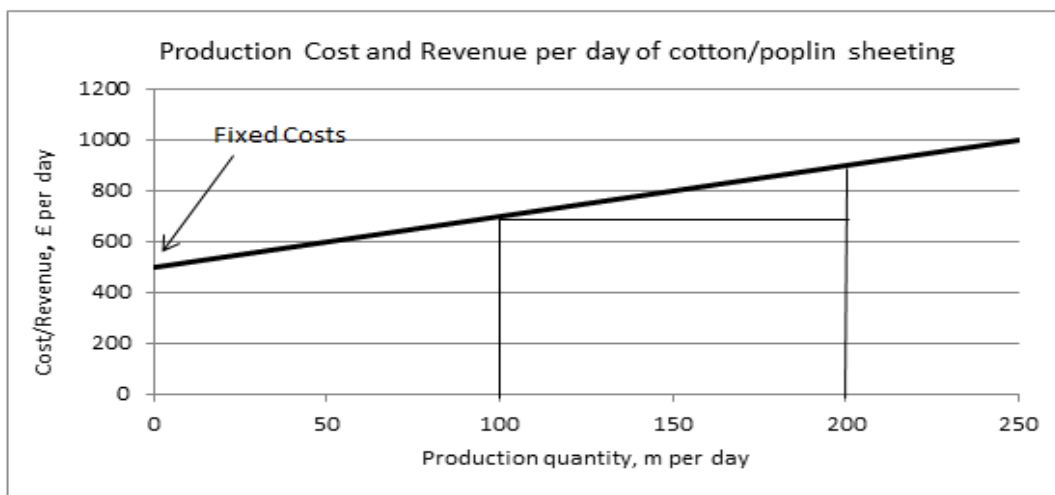
Slope = 200/100 = 2

Figure 8.2 Production Cost at Dickson Mills plc

Linear graphs can be used for a variety of purposes in business decision making. Most people find them easier to use than the equations, although the equations will give the same information. The pictorial display is often helpful in understanding the structure of the decision problem.

Referring to **Example 1**, the graph can be used to determine, for example, the cost of production of 75 m of cloth, or the production level which generates £600 of costs.

8.3.2 BREAK-EVEN CHARTS

Referring to **Example 8.1**, if we know that the sheeting is sold for £7 per m, we can use the graph to determine the production level needed for Dickson's to **break-even,** i.e. to just cover all of the costs and to possibly begin to make a profit. Assume that all of the production is sold.

The revenue, R, from the sheeting is:
R = price per metre × no. of metres sold per day = 7q £ per day
This is a straight line. Add this line to the graph in **Figure 8.2**. See **Figure 8.3**.
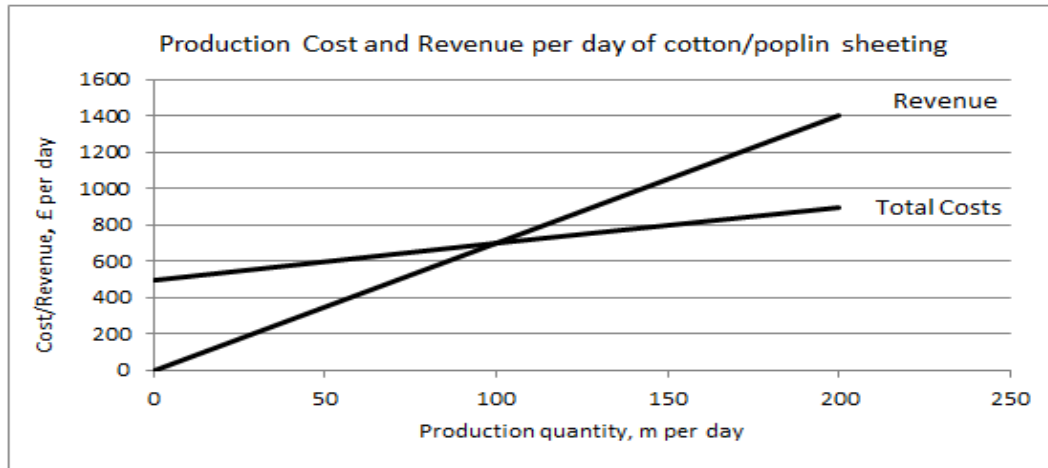
Figure 8.3 Production Cost and Revenue at Dickson Mills plc

How can we tell when the company breaks-even on this product?

The point at which the two lines cross is the break-even point, at this point:

Revenue = Total costs
Profit = 0

The break-even point in this example occurs when q = 100 m per day. Hence Dickson's break-even is achieved if they produce and sell 100 m of sheeting each day. Above 100m per day they make a profit, below they make a loss.

Break-even charts are often constructed with fixed costs shown separately as well.

Example 8.2 – Break-even chart

Consider **Example 8.1** again. The cost equation for Dickson Mills plc is:

$$C = 500 + 2q \text{ £ per day}$$

This cost is made up of a fixed element:

$$FC = 500 \text{ £ per day}$$

and a variable element:

$$VC = 2q \text{ £ per day.}$$

Plotting the total and the fixed costs and the revenue, gives the break-even chart shown in **Figure 8.4**.
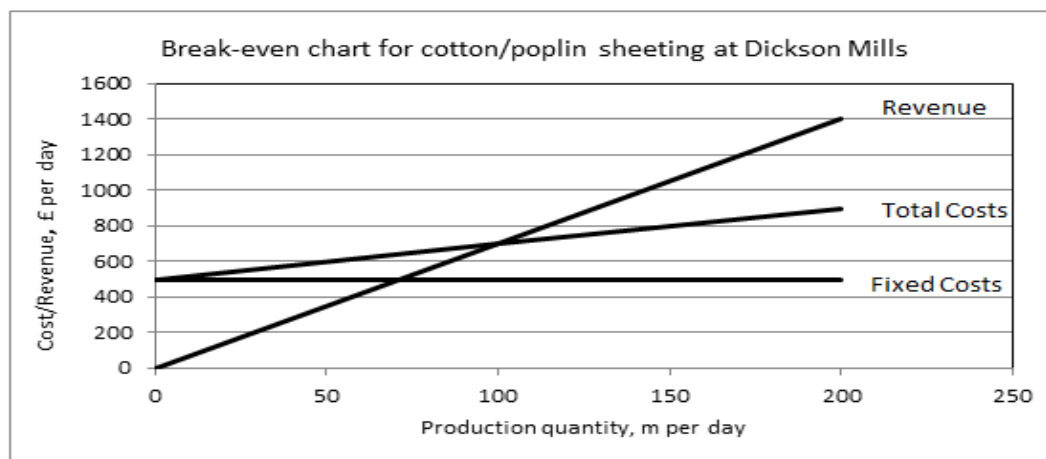
Figure 8.4 Production Cost and Revenue at Dickson Mills plc

A further variation on the break-even chart enables the contribution to be illustrated.

The **contribution chart** illustrates the revenue line, the total cost line and the variable cost line. As before, break-even is the intersection between the revenue line and the total cost line. The total cost line and the variable cost line are parallel to each other. The distance between them is the fixed cost. The distance between the revenue line and the variable cost line is the contribution.

Example 8.3 – Contribution chart

Consider **Example 8.1** again. Plot the revenue line and the total cost line as before and add the variable cost line – **Figure 8.5**.

The break-even point occurs when the contribution is equal to the fixed costs, i.e. at q = 100.

Hence, as before, Dickson's break-even when they produce and sell 100 m of sheeting each day.
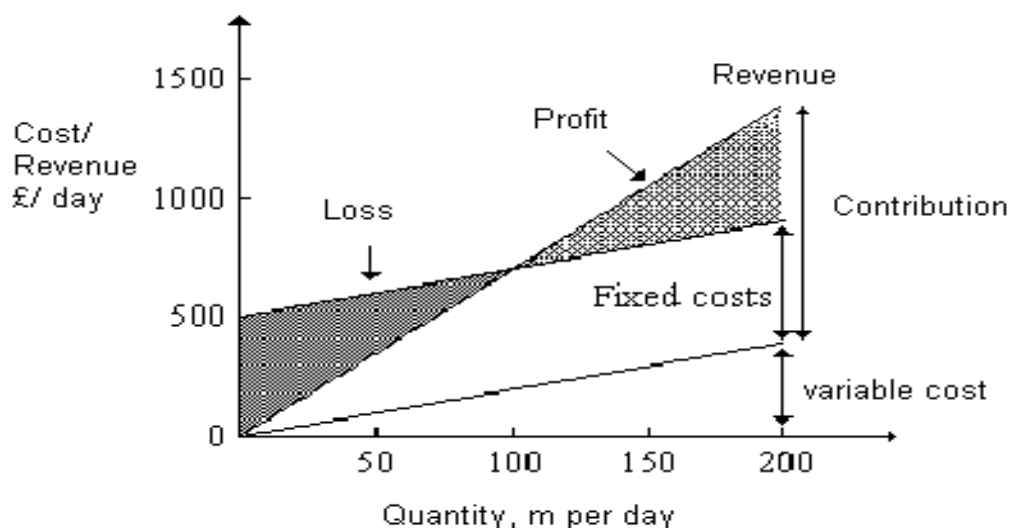
Figure 8.5 Contribution chart for Dickson Mills plc

The final variation on this chart is the **profit-volume** chart. This is commonly used by management since it shows how the basic profit varies with production level.

Example 8.4 – Profit-volume chart

What do you think the profit will be at the break-even point?

If the break even occurs when revenue = total costs, i.e. the revenue just covers the costs incurred, then break-even occurs when:

> Revenue = Total Costs
> Revenue – Total Costs = 0
> Profit = 0 i.e. profit is zero.

Identifying the point of zero profit is an alternative way of identifying the break-even point.

Consider **Example 8.1** again. The profit equation is:

> Profit = Revenue – Total cost
> $P = 7q - (500 + 2q)$ £/day
> $P = 7q - 500 - 2q$ £/day
> $P = 5q - 500$ £/day

Plot the profit line. See **Figure 8.6**.

When the production volume is zero, a loss is made of £500 per day, i.e. the fixed costs are expended but there is no revenue.

When the production volume is 100, the profit is zero, i.e. break-even at q = 100 m per day.

When q > 100, a profit is made.
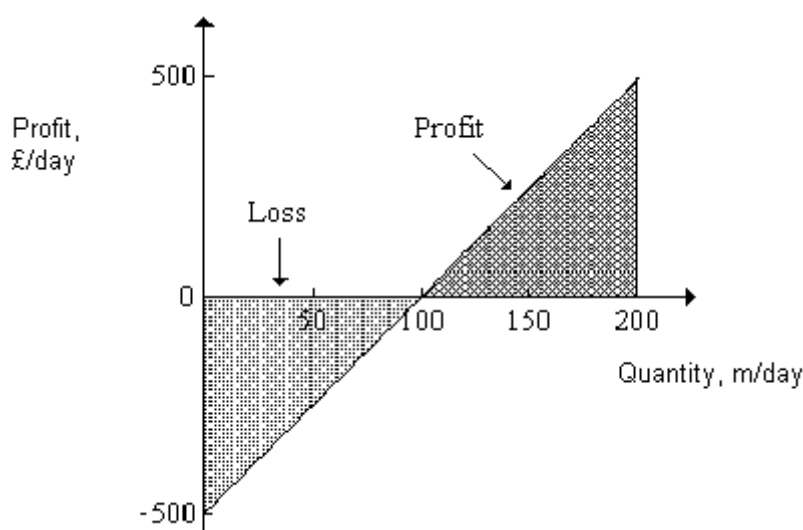
Figure 8.6 Profit-volume chart for Dickson Mills plc

8.4   NON-LINEAR RELATIONSHIPS

Clearly not all business relationships are linear. The task of modelling a business relationship can quickly become quite complex. We will confine ourselves in the following sections to a few basic ideas.

8.4.1 COSTS

In the previous examples, we have assumed that the unit variable cost is constant regardless of the quantity involved. Frequently this is not the case. In a production situation, the unit variable cost often decreases as the quantity produced increases – economy of scale. This effect can be built into the cost model by the inclusion of a third component which depresses the total cost as the quantity produced increases. An example of such a component is given in the following example.

Refer to Example 1 (Dickson Mills plc). The production cost of cotton/poplin sheeting is given by:
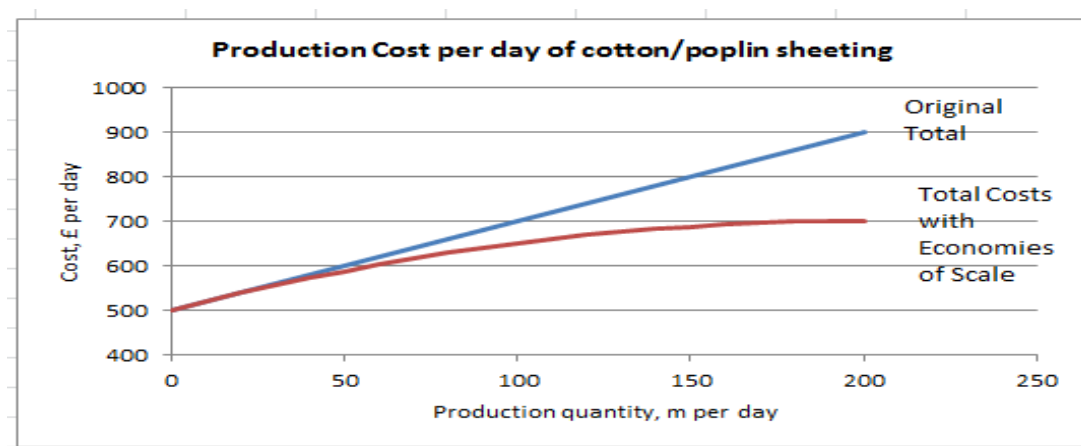
$$C = 500 + 2q \qquad (\text{£ per day})$$

where q is the daily production in metres per day. In order to allow for economy of scale, we might adjust this model to:

$$C = 500 + 2q - 0.005q^2 \qquad (\text{£ per day})$$

Initially, at low production volumes, the q^2 term has very little effect on the total cost, but as q increases, the additional term reduces the total cost more and more.

The effect is illustrated in **Figure 8.7**.

Figure 8.7 Production at Dickson Mills plc

8.4.2 REVENUE

In the previous examples, we have assumed that the price charged per unit is constant and independent of the quantity sold. In fact price charged will often be dependent on the demand for the product. Or, looked at the other way round, the quantity sold will be dependent on the price charged.

Supposing that Dickson Mills had found that the demand for the sheeting and the price charged are related by:

$$q = 700 - 100p \text{ metres/day}$$

or, put the other way round

$$p = 7 - 0.01q \text{ £ per m}$$

Hence the revenue becomes:

$$R = p \times q = (7 - 0.01q) \times q$$
$$= 7q - 0.01q^2 \text{ £ per day or}$$
$$R = p \times q = p \times (700 - 100p) = 700p - 100p^2]$$

Adding this new revenue model to the graph of the new cost model gives **Figure 8.8**.
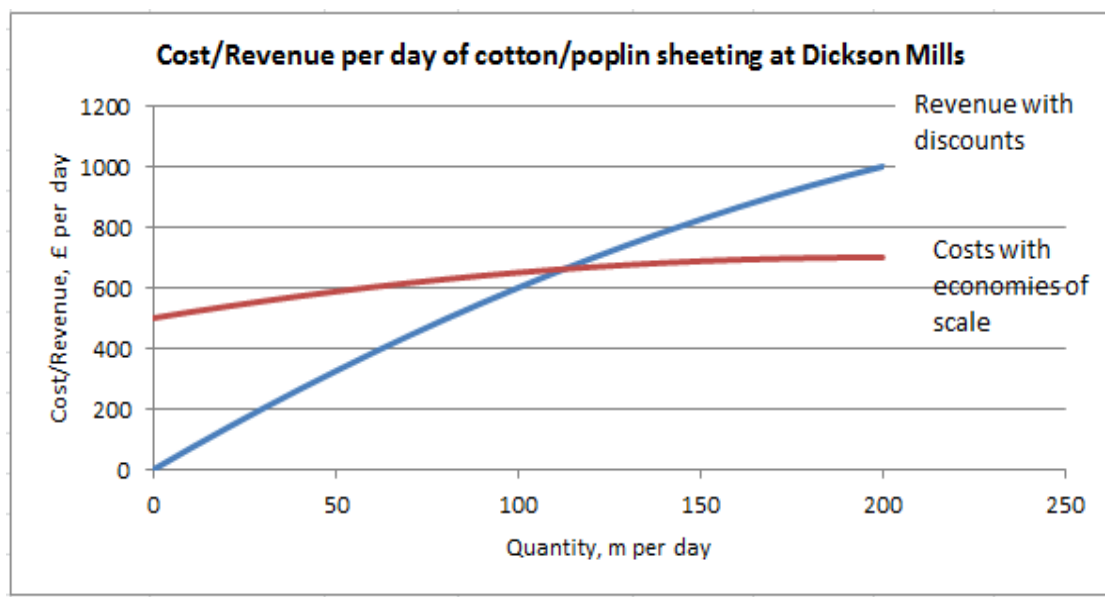
Figure 8.8 Revenue model added to the graph of the new cost model

8.4.3 PROFITS

Suppose that at Dickson Mills plc, the total production costs are linear but the selling price is found to be dependent on demand. Hence:

$$TC = 500 + 2q \qquad £ \text{ per day}$$
$$R = 7q - 0.01q^2 \qquad £ \text{ per day}$$

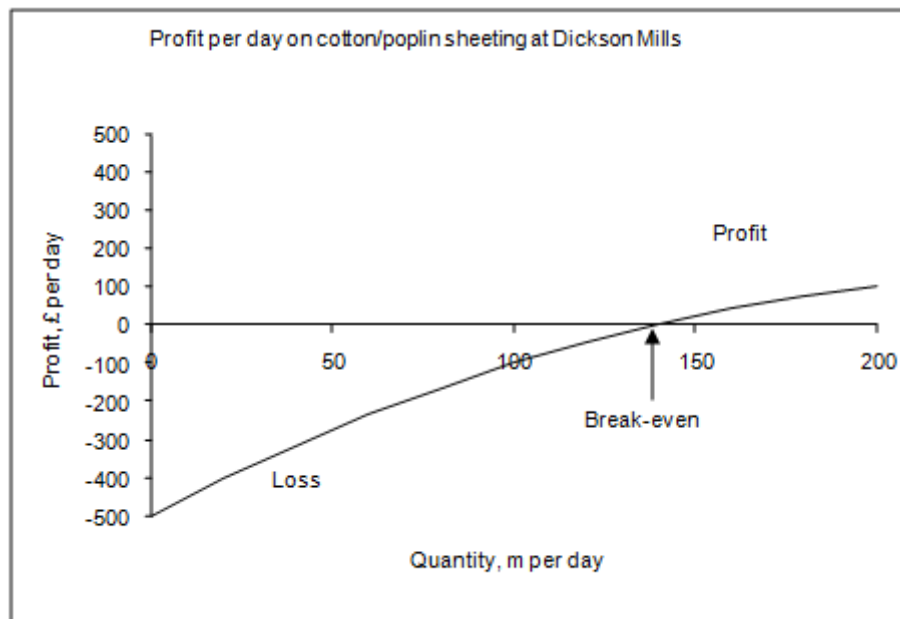The profit is then:
$$P = R - TC$$
$$P = 7q - 0.01q^2 - (500 + 2q)$$
$$P = 7q - 0.01q^2 - 500 - 2q$$
$$P = 5q - 0.01q^2 - 500 \qquad £ \text{ per day}$$

If we now plot profit against quantity produced and sold, we can see how profit varies as quantity varies. See **Figure 8.9**.

As we can see, for low values of production and sales, a loss is made. However, the size of the loss decreases smoothly as quantity produced and sold increases.

Figure 8.9 Profit against quantity produced and sold

The profit/loss is zero when about 140 m are produced and sold per day. Above this level a profit is made.

Example 8.5 – Maximising the profit made

Dickson Mills plc also produce a 'cotton rich' type of sheeting. Using information provided by the management accountants and the marketing department, the following total cost and revenue models have been established for the production range 0 to 200 m per day.

$$TC = 100 + 3q \qquad \text{£ per day}$$
$$R = 13q - 0.05q^2 \qquad \text{£ per day}$$

Hence, the daily profit is:

$$P = R - TC = 13q - 0.05q^2 - (100 + 3q)$$
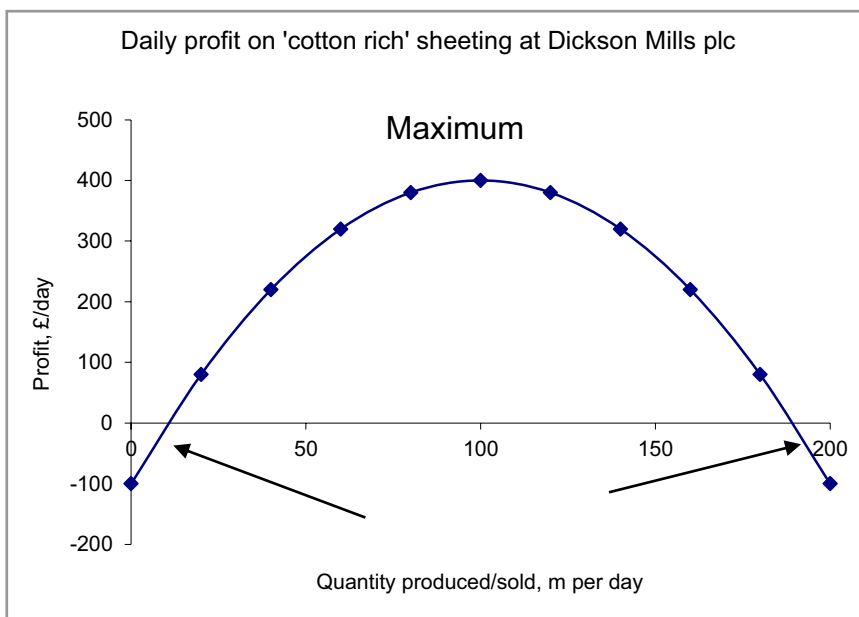$$P = 10q - 0.05q^2 - 100 \qquad \text{£ per day}$$
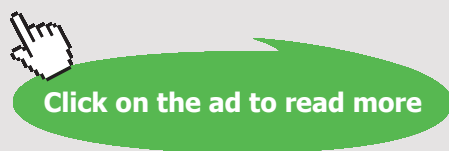
See **Figure 8.10**.

Figure 8.10 Daily profit

We can see that at low and high production levels a loss is made, but over most of the range a profit is achieved.

This profit is a maximum when the production and sales are 100 m per day.

We can also see that the company can break even on this product at two production levels, namely 10.5 m per day and 189.5 m per day. Between these two values, the company will make a profit; outside this range there will be a loss.

Note: For this type of model only, the point of maximum profit is mid-way between the two break-even points.

In general terms, if the graph is plotted from a function of the form

$$aq^2 + bq + c \qquad \text{where a, b and c are numbers}$$

it will always have either

this shape or this shape

(a<0)   (a>0)
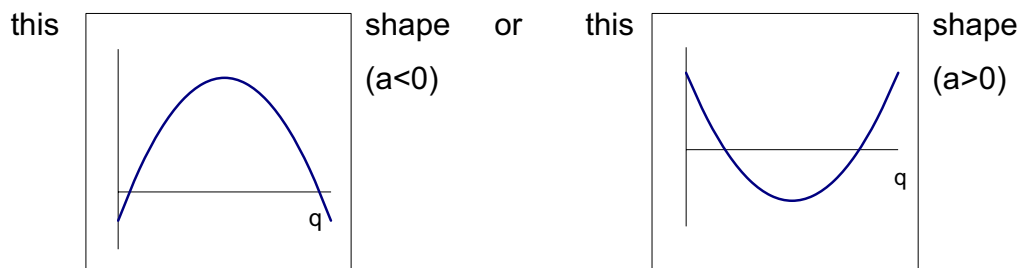
The turning point (i.e. maximum or minimum point) will be midway between the points at which the curve cuts the q axis.

8.5   OTHER USEFUL TYPES OF BUSINESS RELATIONSHIPS

There are lots of different types of models which occur in business contexts. We will consider only one other type.

Suppose a company successfully launches a new product and finds that the sales of this product double every month. The relationship between sales level and time can be modelled by:

$$S = 10 \times 2^t \text{ where S is the number sold in month t}$$

See **Figure 8.11**.
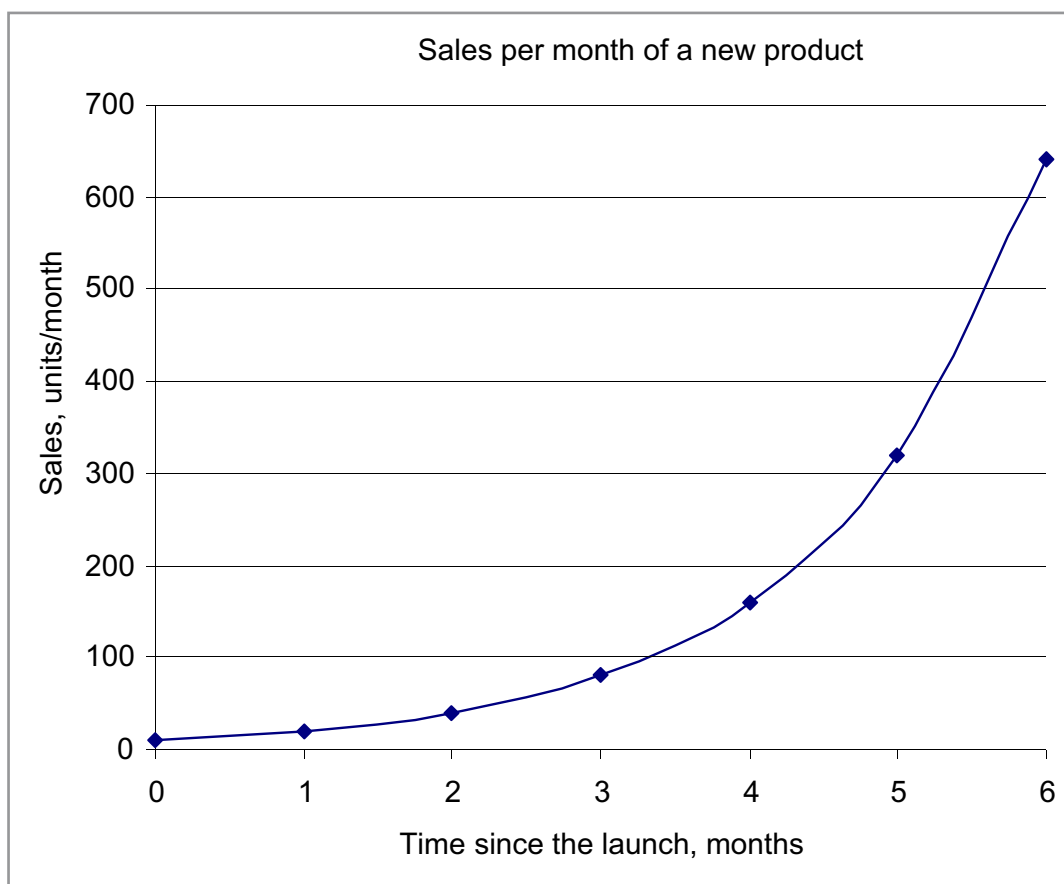
Figure 8.11 Relationship between sales level and time

You can see that the number of units sold is doubling each month.

Suppose another product has reached the end of life and sales are dropping according to:

$S = 1000 \times 2^{-t/3}$ where S is the number sold in month t.

t = 0 when the sales are at their peak. See **Figure 8.12**.
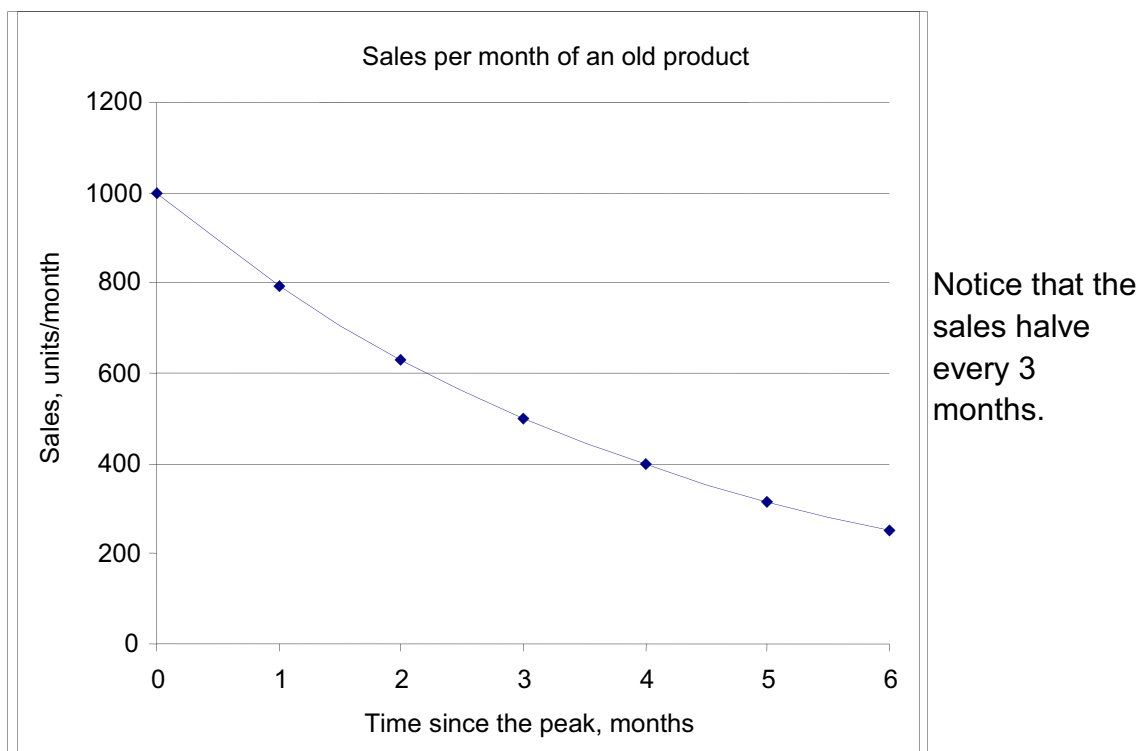
Figure 8.12 Sales of an old product

This type of relationship is called an **exponential**. It can be used when the dependent variable doubles or halves at regular intervals.

8.6   CHAPTER REFERENCES

1. Atrill, P and McLaney, E., Accounting and Finance for Non-Specialists, 5th edition, FT/Prentice Hall, 2006 chapter 7
2. Oakshott, L, Essential Quantitative Methods for Business, Management & Finance, 5th edition, Palgrave Macmillan, 2012
3. Wisniewski, M, Quantitative Methods for Decision Makers, 5th edition, FT/Prentice Hall, 2009 (eBook available)

9 FINANCIAL ANALYSIS MONEY – INTEREST AND TIME VALUE

9.1 INTRODUCTION

Your rich aunt has decided to make you a present of £1000 with the condition that you may not spend the money for three years. What will you do with the money in the meantime?

There are various possibilities. You could put the cheque:

  i) in a sock under the mattress;
 ii) into your ordinary current account at the bank or the building society;
iii) into a deposit account at the bank or a restricted high interest rate account at the building society;
iv) into a stocks and shares investment scheme, ISA, unit trust or such like (your aunt does not count this as 'spending' the money.

Which of these courses of action do you think is the most sensible?

Hopefully, as future business leaders, you have chosen to place your money so that it will earn **interest** – and you have chosen to balance the interest rate obtainable and the risk involved.

Not to be outdone, your uncle says that he will also give you £1000 but not until your birthday in 3 years' time. Assuming that your uncle is sensible, does he actually have to have £1000 now in order to be able to give you this sum in 3 years' time? If your uncle invests £800 now at 10% per year interest, he will accumulate more than the required £1000 in the next three years.

However, from your point of view, will you get as much benefit from £1000 in 3 years' time as you would get if he gave you the £1000 now? Will you be able to buy as much with the money in 3 years' time as you could purchase with it now? In other words, will the value of the money remain the same? The answer to this is almost certainly 'no'. The money will have a smaller purchasing power when you get it than it has now.

These two scenarios illustrate two aspects of money which are of particular concern to managers:

a) the ability of money to earn **interest**,
b) the **time value** of money.

The growth (or decline) in the value of a sum of money is a very important part of many business activities. It concerns all aspects of management, not just the accounting function. The real value of an asset or a sum of money, i.e. the purchasing power of the money, will vary over time. Again, this is a very important problem in business. The lack of an appreciation of the **'time value'** of money is a frequent cause of problems in a business.

In the general accounting function it will frequently be necessary to deal with the value of the asset at different points in time, or to deal with various sums of money at different points in time. For example, if you take out a hire purchase agreement, you will make, say, monthly payments of £20 over the next two years. The value to you, now, of your last payment will be different to the value to you of the first payment.

We will now look in more detail at these two aspects of dealing with money: – interest and 'time value'.

Note: It is assumed that all calculations in the following are carried out on Microsoft Excel, given the fairly tedious nature of the calculations, hence very little, if any, manual workings are shown in the examples.

9.2   INTEREST

When a sum of money is invested it may accumulate interest in a variety of ways. We will look at three of these. In each case, we will take as an example, the investing of £200 for three years at an interest rate of 10% per year.

A_0 - initial sum invested
A_n - the amount accumulated after n time periods, usually years
r   - interest rate per annum, %

9.2.1 COMPOUND INTEREST

The interest is said to be compound when the interest is added to the sum invested and the total amount goes forward to the next time period as the amount upon which the next interest is calculated.

This is the way most interest is calculated in practice. As we have said, there are not many cases in which simple interest is used. In the following notes and examples, assume that the interest is compound, unless stated otherwise.

Example 9.1 – Investment with compound interest

£200 is invested at 10% pa compound interest. How much will have accumulated at the end of three years?

Solution: A general formula for the amount accumulated after n years is:

At the beginning of Year 1,

A0 = 200

At the end of Year 1, the interest, 10% of £200, i.e. £20, is added, giving (200 + 20)

A_1 = £220.

At the end of Year 2, the interest, 10% of £220, i.e. £22, is added, giving (220 + 22)

A_2 = £242.

At the end of Year 3, the interest, 10% of £242, i.e. £24.2, is added, giving (242+24.2)

$$A_3 = £266.20$$

Hence the sum accumulated after three years is £266.20.

9.2.2 THE CALCULATION OF COMPOUND INTEREST

Based on the type of sequence used in Example 1, we can show that in general, the amount accumulated after n years at r% pa compound interest is:

$$A_n = A_0 (1 + r\%)^n$$

This model is referred to as the 'Compound Interest Formula'. We can set up this model quite simply in Microsoft Excel:

	A	B	C	D
1	End of Year	Amount accumulated (£)	Annual interest rate =	
2	0			
3	1	=B2*(1+$D$1)		
4	2	=B3*(1+$D$1)		
5	3	=B4*(1+$D$1)		
6	etc.	etc.		

Table 9.1 Calculation of compound interest

Enter the initial sum invested in cell B2 and the annual interest rate as a percentage in cell D1.

The formula entered in B3 then calculates the amount accumulated at the end of the first year. This goes forward as the initial amount for the second year etc.

This layout is based on the sequencing used in **Example 9.1**.

An alternative layout is:

	A	B	C	D
1	End of Year	Amount accumulated (£)	Annual interest rate =	
2	0			
3	1	=$B$2*(1+$D$1)^A3		
4	2	=$B$2*(1+$D$1)^A4		
5	3	=$B$2*(1+$D$1)^A5		
6	etc.	etc.		

Table 9.2 Calculation of compound interest, alternative example

Enter the initial sum invested in cell B2 and the annual interest rate as a percentage in cell D1 as before.

This layout is based on the general 'Compound Interest Formula'.

Look at the two models carefully and ensure that you understand what each is doing.

9.2.3 NOMINAL AND ACTUAL INTEREST

When the interest is credited to the account at intervals which are shorter than the period for which the interest rate is quoted, then we need to know whether the interest rate quoted is **nominal** or **actual**.

For example, for a particular type of building society account, the interest earned is credited to the account monthly. The interest rate is quoted as 6% per year **nominal**.

This means that the money deposited earns interest at a rate which is 6/12 = 0.5% per month.

The net effect is that the gain over the year is greater than 6% of the sum at the beginning of the year.

This procedure is particularly important when borrowing money, since the debt may accumulate to a larger amount than was expected.

Any organisation paying or charging interest is required by Law to state the **Annual Equivalent Rate (AER)** being used, i.e. the **actual** percentage increase in the sum of money over the year.

If the interest rate is r% per annum **nominal**, credited quarterly, the sum invested accumulates at the rate of (r/4)% per quarter.

If, however, the interest rate is r% per annum **actual**, credited quarterly, the sum invested accumulates at a rate of i% per quarter where i is defined by:

$$(1 + i\%)^4 = 1 + r\%$$

Example 9.2 – Nominal and actual interest rates

Invest £100 at 10% per annum

    i. compound interest;
    ii. compound interest, credited quarterly.

What is the sum accumulated after 5 years?

Solution:

i) Compound Interest: $A_5 = A0(1 + 10\%)^5 = A0(1 + 0.1)^5$

        Hence $A_5 = 100(1.10)^5 = 100 \cdot 1.61051 = 161.051$

The accumulated sum after 5 years is £161.05.

Alternatively, set up the sequence in Microsoft Excel:

	A	B	C	D
1	**End of Year**	**Amount accumulated (£)**	**Annual interest rate =**	**10%**
2	0	100		
3	1	=B2*(1+$D$1)		
4	2	=B3*(1+$D$1)		
5	3	=B4*(1+$D$1)		
6	4	=B5*(1+$D$1)		
7	5	=B6*(1+$D$1)		

Table 9.3 Calculation of compound interest

This sequence gives the following results:

	A	B	C	D
1	**End of Year**	**Amount accumulated (£)**	**Annual interest rate =**	**10%**
2	**0**	100.00		
3	**1**	110.00		
4	**2**	121.00		
5	**3**	133.10		
6	**4**	146.41		
7	**5**	161.05		

Table 9.4 Calculation of compound interest, values

ii) Compound Interest, credited quarterly:

a) Assume that the interest rate is 10% per annum, **nominal**, then the quarterly interest rate is $(10/4) = 2.5\%$.
There will be 4 interest payments per year, giving 20 payments over 5 years.
Hence, $A5 = 100(1 + 2.5\%)^{20} = 100(1.025)^{20} = 163.86$

The sum accumulated after 5 years is £163.86, which is bigger than when we use the same nominal interest rate paid once per year.
Again the sequence can be set up in Microsoft Excel – the entry in cell D1 above is changed to 2.5% and the formula is Auto-filled until the value in column A is 20.

b) Assume the interest rate is 10% per annum **actual**, then the sum accumulated after 5 years is the same as in i) i.e. £161.05, but the interest is credited to the account at the end of each quarter instead of at the end of each year. This is better for the investor since there are now more times in the year when it would be reasonable to withdraw money without losing the interest on the period in which the withdrawal is made.
The quarterly interest rate, i%, is given by:

$$(1 + i\%)^4 \qquad = (1 + 10\%) = 1.10$$

$$(1 + i\%) \qquad = \sqrt[4]{1.10} = 1.02411 \text{ to 5 d.pl.}$$

Therefore i /100 = 0.02411

Hence, the quarterly interest rate is 2.41% to 2 d.pl.

9.3   APPLICATIONS OF INTEREST RATE CALCULATIONS

9.3.1   ANNUAL EQUIVALENT RATE (AER)

The Annual Equivalent Rate (AER) was formerly known as the Annual Percentage rate (APR).

The AER was mentioned briefly in section 9.2.3. Its importance to us should be clearer following the discussion above.

In advertisements for loans to cover the purchase of items such as cars, or in bank or building society literature about savings accounts or mortgage facilities, two interest rates are usually quoted – the one to catch your eye (monthly or quarterly rate or the annual nominal rate) and the AER (annual equivalent rate). The AER is required since the repayments (or payments) are usually made at intervals of less than a year so the debt (or savings) accumulates at a faster rate than the advertiser's interest rate suggests. For example, if the outstanding debt on a credit card bill attracts interest at 2% per month then the debt will grow at the actual rate of nearly 27% per annum.

9.3.2 DEPRECIATION

Depreciation is the loss in value of an asset over time. Generally large capital items depreciate in value due to use but the value will also drop because the item is no longer the latest model. We see this particularly with cars and computing equipment.

When a company buys a piece of equipment, the equipment forms part of the assets of the company and must therefore have a specific value at any point in time. This value will clearly reduce over time as the equipment is used. The company's accountant will make allowance in the accounts for this loss of value. This allowance is referred to as depreciation. The initial expenditure is 'written off' over the lifetime of the equipment in terms of a specific amount of depreciation each year. Depreciation may be counted against profits so that the company can 'save up' to replace the equipment at the end of its life.

There are a number of different ways of dealing with depreciation. The two most widely used methods are:

- the straight line method
- the reducing balance method

The straight line method assumes that the asset depreciates by a fixed sum each year. Hence if the equipment is purchased for V_0 and is expected to be sold n years later for V_n, then the annual depreciation is:

$$\text{annual depreciation} = \frac{\text{cost of equipment - scrap (or re-sale) value}}{\text{estimated life of the equipment}}$$

$$\text{annual depreciation} = \frac{V_0 - V_n}{n}$$

Most electronic equipment loses a bigger proportion of its initial value in the early years, a smaller proportion towards the end of its life. In this case, the straight line method does not reflect the actual value of the asset at each year end.

In the reducing balance method it is assumed that a fixed <u>percentage</u> of the value of the asset is written off each year.

If V_0 is the initial value of the asset and the rate of depreciation is r% pa, then the value of the asset at the end of the first year is:

$$V_1 = V_0 - r\%V_0 = V_0(1 - r\%)$$

The value of the asset at the end of the second year is:

$$V_2 = V_1 - r\%V_1 = V_1(1 - r\%)$$

The value of the asset at the end of the third year is:

$$V_3 = V_2 - r\%V_2 = V_2(1 - r\%)$$

and so on to V_n.

The sequence above can be developed into a general 'reducing balance depreciation' formula:

$$V_n = V_0(1 - r\%)^n$$

Alternatively, we can set up a Microsoft Excel model for the depreciated value of the asset at the end of each year. See **Table 9.5**.

	A	B	C	D	E
1	**End of Year**	**Depreciated value (£)**	**Depreciation (£)**	**Annual depreciation rate =**	
2	**0**				
3	**1**	=B2*(1-$E$1)	=B2-B3		
4	**2**	=B3*(1-$E$1)	=B3-B4		
5	**3**	=B4*(1-$E$1)	=B4-B5		
6	**4**	=B5*(1-$E$1)	=B5-B6		
7	**etc.**	etc.			

Table 9.5 Depreciated value of assets

Enter the initial value of the asset in cell B2 and the rate of depreciation in E1.

Look at the formulae in cells B3 onwards and C3 onwards – make sure that you understand what they are calculating.

Example 9.3 – Calculation of depreciation

A machine costing £6000 is expected to be sold after 5 years for £2000. Evaluate the depreciation which should be charged each year. Assume a) a straight line method; b) the reducing balance method with a depreciation rate of 20%.

Solution:

a) Straight line method: $\text{annual depreciation} = \dfrac{6000 - 2000}{5} = £800$

£800 is deducted from the value of the machine each year for 5 years.

b) The reducing balance method:

	A	B	C	D	E
1	**End of Year**	**Depreciated value (£)**	**Depreciation (£)**	**Annual depreciation rate =**	**20%**
2	**0**	6000			
3	**1**	=B2*(1-$E$1)	=B2-B3		
4	**2**	=B3*(1-$E$1)	=B3-B4		
5	**3**	=B4*(1-$E$1)	=B4-B5		
6	**4**	=B5*(1-$E$1)	=B5-B6		
7	**5**	=B5*(1-$E$1)	=B6-B7		

Table 9.6 Reducing balance method

Or alternatively, we can build the model as follows:

	A	B	C	D	E
1	**End of Year**	**Depreciated value (£)**	**Depreciation (£)**	**Annual depreciation rate =**	**20%**
2	0	6000			
3	1	=$B$2*(1-$E$1)^A3	=B2-B3		
4	2	=$B$2*(1-$E$1)^A4	=B3-B4		
5	3	=$B$2*(1-$E$1)^A5	=B4-B5		
6	4	=$B$2*(1-$E$1)^A6	=B5-B6		
7	5	=$B$2*(1-$E$1)^A7	=B6-B7		

Table 9.7 Reducing balance method, alternative method

These models both yield the following values:

	A	B	C	D	E
1	**End of Year**	**Depreciated value (£)**	**Depreciation (£)**	**Annual depreciation rate =**	**20%**
2	**0**	6000			
3	**1**	4800	1200		
4	**2**	3840	960		
5	**3**	3072	768		
6	**4**	2458	614		
7	**5**	1966	492		

Table 9.8 Reducing balance method, values

Note: The terminal value is not exactly £2000. This is not a difficulty since a final value is likely to be subject to some uncertainty.

If the rate of depreciation is unknown for the reducing balance model, then a suitable value can be found by adjusting the value in cell E1 until the estimated terminal value is achieved.

9.3.3 INFLATION

The effect of inflation is calculated in exactly the same way as compound interest.

Example 9.4 – Calculation of inflation

If coffee, costing £3.30 for a 200g jar in January 1999, is expected to be subject to a constant inflation of 8% per annum, how much will a 200g jar cost in January 2003?

$$V_n = V_0 \left(1 + (r/100)\right)^n$$

Hence, $V_4 = 3.30(1 + 0.08)^4 = 3.30 \; 1.08^4 = 4.49$

Hence, the jar of coffee is expected to cost £4.49 in 2003.

Using Microsoft Excel to model the problem gives:

	A	B	C	D
1	**End of Year**	**Cost (£)**	**Annual inflation rate =**	**8%**
2	**0**	3.30		
3	**1**	=B2*(1+$D$1)		
4	**2**	=B3*(1+$D$1)		
5	**3**	=B4*(1+$D$1)		
6	**4**	=B5*(1+$D$1)		

Table 9.9 Calculation of inflation

The values are:

	A	B	C	D
1	**End of Year**	**Cost (£)**	**Annual inflation rate =**	**8%**
2	**0**	3.30		
3	**1**	3.56		
4	**2**	3.85		
5	**3**	4.16		
6	**4**	4.49		

Table 9.10 Calculation of inflation, values

9.4 PRESENT VALUE

In business we frequently deal with sums of money which arise at different points in time. The time at which the money arises affects its value to us in **'real terms'**, i.e. £100 which we will receive in two years' time, will purchase less (has a smaller 'real value') than £100 which we have now. In order to assess the real value of a sum of money, it is usual to determine the value that it would have at the **present time**.

The value **now** of a sum of money which will arise at some other point in time, is called the **present value** of that sum.

To calculate the present value of a sum of money which will arise in the future means that the actual value in the future has to be reduced to an equivalent value for the present time, using a suitable **discount rate** per annum.

The procedure is referred to as **discounting**.

Consider $A_n = A_0 (1 + r\%)^n$

where An is the sum of money which will arise in n years' time.

A_0 is the present value of this sum
r% pa is the discount rate

It follows that the present value is $A_0 = \dfrac{A_n}{(1+r\%)^n}$

Example 9.5 – Calculating a present value

Your aunt has promised to give you £1000 on your 25th birthday which is 5 years from now. What is the present value of her gift if an average annual discount rate of 6% is assumed?

Solution:

The Present Value, $A_0 = \dfrac{A_n}{(1+r\%)^n} = \dfrac{1000}{(1+0.06)^5} = 747.26$

Hence, the present value of the gift at a discount rate of 6% pa is £747 to the nearest £1. The present value is calculated by multiplying A_n by the following multiplying factor:

$$\dfrac{1}{(1+r\%)^n} \ .$$

This multiplying factor is referred to as the **discount factor**.

The choice of a value for r will depend on the context in which the discounting is being done. In situations such as that described in **Example 5** above, we would probably use the average rate paid on savings accounts. In the two applications discussed below both of which involve the investment of funds (annuity and capital project appraisal), the choice of a discount rate is based on the interest which the investment funds could earn if not used for the purpose specified plus an element to cover any administrative costs, taxation etc.

Example 9.6 – Calculating the total present value of a series of payments

Tom's parents have promised to give him £1000 every 6 months while he is at university. Assume that Tom's course is beginning now and that it lasts for four years.

What is the total present value of the payments if Tom's parents could achieve a rate of 3% per half year if they invested their money elsewhere?

Solution:

The Present Value of one payment is, $A_0 = \dfrac{A_n}{(1+(r/100))^n} = \dfrac{1000}{(1+0.03)^n}$

where n is the number of half years over which the sum is discounted.

Payments are made as follows:

End of

| | | | | | | | |

Now   1yr   2yr   3yr   4yr

P   P   P   P   P   P   P   P

Total Present Value of all of the payments made is:

$$PV_{3\%\ per\ 6\ mths} = 1000 + \frac{1000}{1.03} + \frac{1000}{1.03^2} + \frac{1000}{1.03^3} + \frac{1000}{1.03^4} + \frac{1000}{1.03^5} + \frac{1000}{1.03^6} + \frac{1000}{1.03^7}$$

$$= 1000 \times (1 + \frac{1}{1.03} + \frac{1}{1.03^2} + \frac{1}{1.03^3} + \frac{1}{1.03^4} + \frac{1}{1.03^5} + \frac{1}{1.03^6} + \frac{1}{1.03^7})$$

$$= 1000 \times 7.23028$$

$$= £7230$$

Hence, the total value now of the 8×£1000 payments which Tom receives, is £7230.

This problem may be modelled in Microsoft Excel as follows:

	A	B	C	D	E	F
1	End of half year	Discount Factor	Cash flow (£)	Present value (£)	Discount rate =	3%
2	0	=1/(1+$F$1)^A2	1000	=C2*B2		
3	1	=1/(1+$F$1)^A3	1000	=C3*B3		
4	2	=1/(1+$F$1)^A4	1000	=C4*B4		
5	3	=1/(1+$F$1)^A5	1000	=C5*B5		
6	4	=1/(1+$F$1)^A6	1000	=C6*B6		
7	5	=1/(1+$F$1)^A7	1000	=C7*B7		
8	6	=1/(1+$F$1)^A8	1000	=C8*B8		
9	7	=1/(1+$F$1)^A9	1000	=C9*B9		
10	8					
11			Total =	=SUM(D2:D9)		

Table 9.11 Calculation of present value

Which yields:

	A	B	C	D	E	F
1	End of half year	Discount Factor	Cash flow (£)	Present value (£)	Discount rate =	3%
2	0	1	1000	1000		
3	1	0.97087	1000	970.874		
4	2	0.94260	1000	942.596		
5	3	0.91514	1000	915.142		
6	4	0.88849	1000	888.487		

7	5	0.86261	1000	862.609		
8	6	0.83748	1000	837.484		
9	7	0.81309	1000	813.092		
10	8					
11			Total =	7230.28		

Table 9.12 Calculation of present value, values

Note: It is assumed in the calculation of the present values that the cash flows arise at the end of each period.

9.4.1 NET PRESENT VALUE

In business management will be faced with investment decisions which involve choosing between competing activities. For example, the R&D Department may have a project for developing a product; the Marketing Department may have a project for running a nationwide product promotion; the Information Systems people may have put up a proposal for upgrading the IT system. If the company currently has funds for one of these only, then a choice has to be made. One of the elements which can contribute to the decision making process is the **net present value**.

The **net present value** (NPV) for a project is calculated by summing the year end cash flows which the project is expected to deliver and then deducting the initial cash outlay, i.e. the cost of a capital project is deducted from the total present value of the year end cash flows generated by the project.

In such calculations, the discount rate used is equivalent to what is called the company's 'cost of capital'. This is the minimum rate of return from an investment project which the company expects to get before they will invest. Or, it is the cost to the company of borrowing the capital to be invested and hence the cost of the borrowings must be covered by the returns from the project.

If the NPV is positive it means that the earnings from the project are equivalent to an interest rate larger than the discount rate used to calculate the present values of the cash flows therefore the project is financially attractive. If the NPV is negative it means that the earnings from the project are equivalent to an interest rate less than the discount rate, therefore the project is not financially beneficial.

Example 9.7 – Investment appraisal

A company is considering two possible projects. The initial costs and the expected cash flows are given in the table below.

The company usually expects a 15% pa return on its capital investments.

Project	Initial cost (£'000)	Net cash flow at year end, £'000				
		Year 1	Year 2	Year 3	Year 4	Year 5
A	-1000	500	400	300	200	100
B	-1000	200	200	300	400	400

Table 9.13 Cash flow for two projects

Use the net present value to determine which of these projects, if either, the company should choose.

Solution: The Microsoft Excel model is:

	A	B	C	D	E	F
1	discount rate =	15%	Project A		Project B	
2	End of year	Discount Factor	Year end cash flow (£'000)	Present value (£'000)	Year end cash flow (£'000)	Present value (£'000)
3	0	=1/(1+$B$1)^A3	-1000	=C3*B3	-1000	=E3*B3
4	1	=1/(1+$B$1)^A4	500	=C4*B4	200	=E4*B4
5	2	=1/(1+$B$1)^A5	400	=C5*B5	200	=E5*B5
6	3	=1/(1+$B$1)^A6	300	=C6*B6	300	=E6*B6
7	4	=1/(1+$B$1)^A7	200	=C7*B7	400	=E7*B7
8	5	=1/(1+$B$1)^A8	100	=C8*B8	400	=E8*B8
9						
10			NPV =	=SUM(D3:D8)	NPV =	=SUM(F3:F8)

Table 9.14 Comparison of two projects

	A	B	C	D	E	F
1	discount rate =	15%	Project A		Project B	
2	End of year	Discount Factor	Year end cash flow (£'000)	Present value (£'000)	Year end cash flow (£'000)	Present value (£'000)
3	0	1	-1000	-1000.00	-1000	-1000.00
4	1	0.86957	500	434.78	200	173.91
5	2	0.75614	400	302.46	200	151.23
6	3	0.65752	300	197.26	300	197.26

7	**4**	0.57175	200	114.35	400	228.70
8	**5**	0.49718	100	49.72	400	198.87
9						
10			NPV =	98.56	NPV =	-50.03

Table 9.15 Comparison of two projects, values

This yields Project A generates a positive value for the NPV, therefore it has a return on the initial outlay which is greater than 15% pa. The project is to be preferred to an investment yielding 15% pa. Project A is financially beneficial. Project B generates a negative value for the NPV, therefore it has a return on the initial outlay which is lower than 15% pa. The project does not meet the company's requirement on return on capital invested. In practice, investment decisions will not be taken on the basis of the NPV alone, but it is an important component in the appraisal of capital investment projects.

10 BRIEF CONCLUDING REMARKS AND CONSIDERATIONS

This book has endeavoured to provide a range of data analysis methods while going through a very broad range of examples that include fully worked Microsoft Excel models.

These examples are modelled using data sets which managers in particular are very likely to come across in their day to day activities and will need to manipulate in order to support decision making. It is at this stage where it would be useful to make some comments about the usefulness and applicability of the models illustrated so far.

The first and the most valuable consideration is that the outputs of the various models used for the purpose of decision making are ultimately just a guide – the models cannot make decisions, the ultimate decision maker is the manager/professional using the models. That means that the responsibility for these decisions cannot lie with the models!

Therefore, at this stage it is appropriate to include some considerations in relation to the use of models in business analysis and decision making.

The role of models as part of managerial decision making should be fairly obvious from their use in the numerous examples presented already. However, it will be useful to summarise some points here and to look at some of the disadvantages as well as the benefits. Modelling is one of the ways in which the possible options available for decision making may be investigated.

Here are some of the typical ways for models to be used in decision making:

- Models can help the decision maker to:
  - extract the important elements from a decision problem,
  - investigate variations in conditions / parameters,
  - test the sensitivity of assumptions.
- Models are a supplement to, not a replacement for, managerial creativity and judgement.
- The building of a model encourages debate about the decision and associated issues.
- The model should generate possible options to be considered, not answers. Qualitative information must also be considered.
- The use of a model does not mean that a particular course of action has to be implemented.

The main problems which can arise in the modelling process concern the difficulty of adequately replicating complex situations and also the difficulty of obtaining reliable values of some parameters (we have seen this earlier in the book when forecasting was discussed for example).

However, if the problem to be solved is a general one and a general solution is possible, then the cost-benefits accrued from the model can be high. It offers versatility, i.e. we can ask a range of 'what if' questions and use the model to predict the behaviour of the system under different conditions. Without models, the manager has to rely on experience and intuition alone. These are essential tools but if the manager is operating outside his/her previous experience, more support is necessary – common sense may not be sufficient.

A summary of the advantages and disadvantages of modelling is given below in the following:

Advantages of using models include:

- Stimulation of discussion and creativity
  - requires the decision maker to think in depth about the nature and structure of the situation / system under consideration,
  - requires the decision maker to investigate the properties of the situation / system,
  - allows the decision maker to question 'traditional' views and assumptions.

- Provision of a framework for evaluation of data.
- Provision of a vehicle for communication of the possible business scenarios.
- Enables different courses of action to be evaluated, hence provides supporting evidence for a particular decision / course of action
- Versatility – models allow for 'what if' analysis – in other words what if we change the values of a parameter, what would be the numerical result of the model then?
- Safety – while it is in progress, the investigation does not affect the organisation as no decisions have been made yet.
- Speed – various alternative courses of action can be investigated very quickly.
- Low cost of employing models – a thorough analysis of a problem can help avoid costly mistakes.

Overall, informed decisions can be made; evidence is available to strengthen the arguments for and against a particular decision.

Disadvantages of using models include:

- The model may be too great an approximation of the reality.
- Various parameter values may be estimates only, therefore increasing the margin of error in terms of the model's numerical output.
- The cost of building and solving the model can exceed the benefit gained for small or low risk / cost decisions.

When considering the use of a modelling procedure, the following questions should be considered:

- Is the decision suitable for modelling? In other words is there a numerical method that can be employed to manipulate a data set?
- Will the output from the model improve unaided judgement?
- How long will the model take to build and solve?
- How much will the process of building and solving a model cost?

If the answers to these questions are largely positive then it is worth proceeding to the next stage.

ENDNOTES

1. As you read through the data make a tally mark against the appropriate value. At the end add up all of the tally marks to find the total frequency for each value. It is usual to mark the tally in groups of five – four straight marks and the fifth one across – makes totalling at the end easier.

2. In this section 'average' refers to the 'arithmetic mean'.

3. In this example, we clearly would not normally be doing this calculation for a single student. We would have a list of students and we would auto-fill the formulae over the whole list. However, this will not work unless we create absolute cell references for the weights. Absolute cell referencing, fixes the relevant cell in a formula. To create an absolute cell reference we insert the $ sign in front of the part of the reference we want to fix, i.e. the row or column or both. Hence, the three formulae above for cells D5, G5 and I5 should be amended to the following, respectively:

    =(K$1*K2 + L$1*L2 + M$1*M2) / SUM(K$1:M$1)

    =(C$3*C5 + D$3*D5 + E$3*E5 + F$3*F5) / SUM(C$3:F$3)

    =(G$2*G5 + H$2*H5) / SUM(G$2:H$2)

4. There are often different views about the exact position of the quartiles. This is acceptable as long as you are consistent.

5. There are several standard deviation functions in Excel which have specific uses in various types of statistical analysis. For our purposes in the Business Analysis module, always use =STDEV**P**()

6. The proof is not necessary in this module but if you are interested in what it is, ask your tutor.

7. The seasonal components must always sum to zero