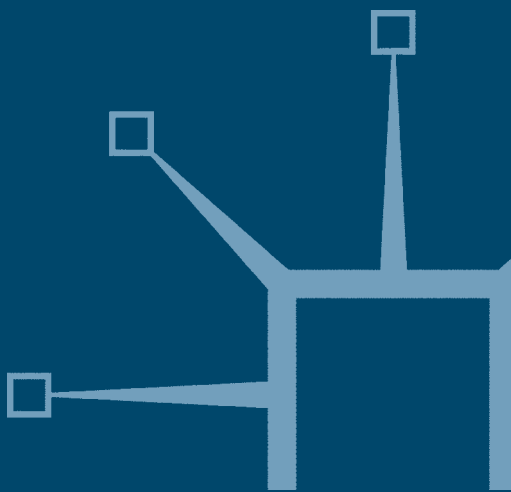
palgrave
macmillan

Business Statistics for non-mathematicians

Sonia Taylor



BUSINESS STATISTICS

‘Sonia Taylor’s book is a well presented, easy to read text underpinned with worked examples of statistical analysis relevant to the world of business. The book acknowledges and explores the different tools with which students are required to analyse data – Excel, SPSS and Minitab. For lecturers it offers a systematic approach, supported with a wealth of worked examples and student questions.’

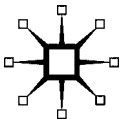
Charles Leatherbarrow, Senior Lecturer,
University of Wolverhampton Business School, UK

Business Statistics

for non-mathematicians

Sonia Taylor

palgrave
macmillan



© Sonia Ann Taylor 2007

All rights reserved. No reproduction, copy or transmission of this publication may be made without written permission.

No paragraph of this publication may be reproduced, copied or transmitted save with written permission or in accordance with the provisions of the Copyright, Designs and Patents Act 1988, or under the terms of any licence permitting limited copying issued by the Copyright Licensing Agency, 90 Tottenham Court Road, London W1T 4LP.

Any person who does any unauthorised act in relation to this publication may be liable to criminal prosecution and civil claims for damages.

The author has asserted her right to be identified as the author of this work in accordance with the Copyright, Designs and Patents Act 1988.

First edition 2001
Reprinted 3 times
Second edition 2007
Published by PALGRAVE MACMILLAN
Houndmills, Basingstoke, Hampshire RG21 6XS and
175 Fifth Avenue, New York, N.Y. 10010
Companies and representatives throughout the world

PALGRAVE MACMILLAN is the global academic imprint of the Palgrave Macmillan division of St. Martin's Press, LLC and of Palgrave Macmillan Ltd. Macmillan® is a registered trademark in the United States, United Kingdom and other countries. Palgrave is a registered trademark in the European Union and other countries.

ISBN-13: 978-0-230-50646-6
ISBN-10: 0-230-50646-1

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources.

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

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

10 9 8 7 6 5 4 3 2 1
16 15 14 13 12 11 10 09 08 07

Printed and bound in China

Contents

Preface

Students:

- Are you a student, either undergraduate or postgraduate, who is studying statistics for the first time?
- Are you planning to do a course including statistics in the near future?
- Are you a researcher who will need to collect and analyse data?
- Do you lack confidence in mathematics or are you 'rusty'?
- Are you nervous about statistics?
- Are you looking for a user-friendly textbook?
- Do you want a textbook which keeps statistical theory, long formulae and mathematical calculations to a minimum?
- Would you like a website giving extra help and practice?

If you answer 'yes' to most of these questions then this is definitely the textbook for you.

Lecturers:

- Are you feeling overworked?

Don't despair: here is a publication which will cut down your preparation and ease your teaching considerably. *Business Statistics for non-mathematicians* is accompanied by lecturers' materials on a companion website, www.palgrave.com/business/taylor, which use the core material from each chapter of the textbook and provide masters for all the paperwork, overhead slides and PowerPoint presentations needed by the lecturer on a weekly basis. Revision questions, multi-choice questions suitable for in-class tests and examination type questions are provided in addition to the questions supplied to the students for revision purposes. (See Appendix F for a full list.)

Business Statistics for non-mathematicians has been written as a practical response to the needs of the present-day student on a business studies or related course, who needs to obtain a reasonable grasp of basic statistics in the limited time available. This book is also suitable for

students or researchers at any level who are not confident in the use of statistics or who are meeting the subject for the first time. It addresses the limitations imposed by modular courses for large teaching groups of students with uncertain mathematical ability. It is organised so that students can gradually build up the knowledge of statistics required, starting from a base of simple arithmetic. Many students experience problems initially, so the emphasis throughout is on understanding through practice, interpretation of results and their application rather than on depth of knowledge. Statistical theory, unhelpful jargon, the use of formulae and long mathematical manipulations are deliberately kept to a minimum. This encourages students who lack confidence in their own mathematical ability.

The main themes of this book are:

- Descriptive statistics: graphical and numerical.
- Inferential statistics: confidence intervals and hypothesis tests.
- Pairwise relationships between variables: correlation, regression and chi-square tests.
- Forecasting: modelling and use of time series.

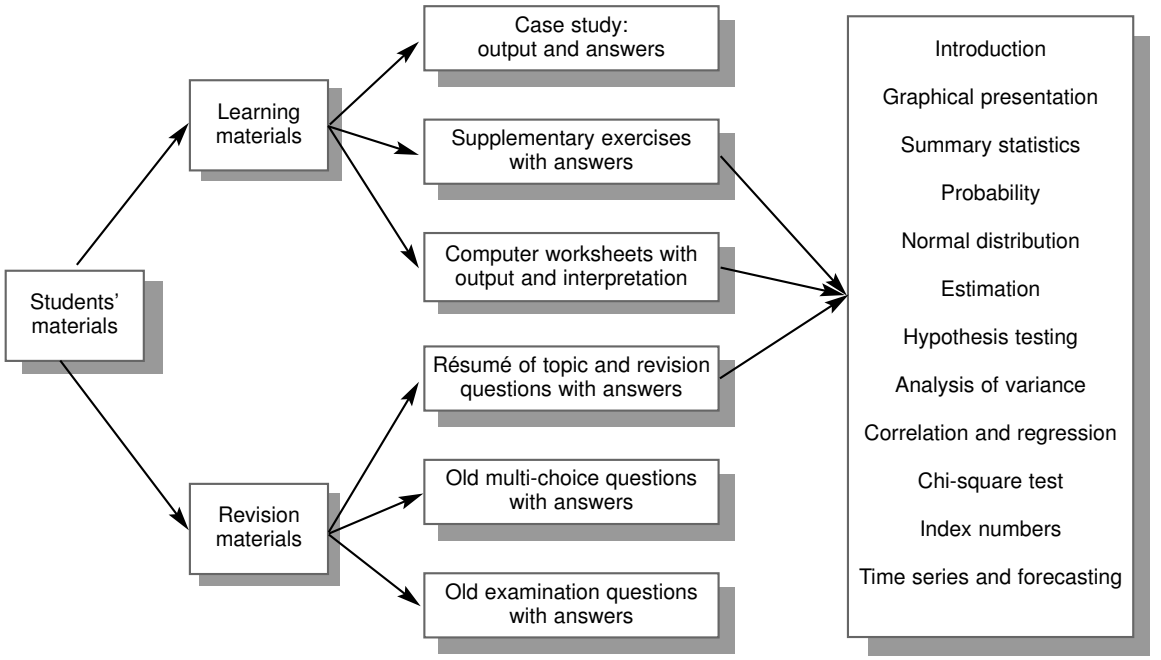
More specifically:

- Chapter 1 introduces the topic of **statistics** generally, defines certain concepts with which you may or may not already be familiar, and concludes with a self check of the basic arithmetic required for this course.
- **Graphical and numerical description** of different types of data sets are covered in Chapters 2 and 3 with the emphasis on data **summarisation** as a means of communicating the information they contain to others.
- **Uncertainty** is a concept often first met in statistics. It is introduced in Chapter 4 as **probability** and enlarged upon in Chapter 5. This chapter describes and uses the **normal distribution** which fits much business data.
- **Inferential statistics** are introduced in Chapter 6 with the use of **estimation** in the form of **confidence intervals**. This subject is further developed with the **hypothesis testing of various parameters**. Chapter 7 is mainly concerned with a variety of t-tests and Chapter 8 with **analysis of variance**.
- Bivariate relationships between variables are analysed in the form of **correlation** and **regression** in Chapter 9 and **chi-square tests** in Chapter 10.
- **Index numbers** are studied in Chapter 11 to monitor and compare any changes over time. Various **time series** are modelled in Chapter 12 and are then extended into the future to produce **forecasts** in Chapter 13.
- Chapter 14 deals with the production of the **computer analysis** of data, in **SPSS**, **Minitab** and **Excel**, using all the methods covered in the previous chapters. The versions used are the current ones at the time of writing: SPSS 12.0.1, Minitab 14 and Microsoft Office Excel 2003.
- A continuous **case study** using real data runs through most of the chapters.

Each chapter includes the necessary theory but stresses the reasons for, and methods of, carrying out the various techniques and analyses. Plenty of practice is provided with tutorial exercises. Computer analysis by all methods using SPSS, Minitab and some Excel is described separately in Chapter 14.

A companion website

www.palgrave.com/business/taylor provides additional materials for the student:



A complete set of lecturer's materials

For the lecturer there is provided on the companion website, www.palgrave.com/business/taylor, a very comprehensive set of all the materials necessary to run the course (see below).

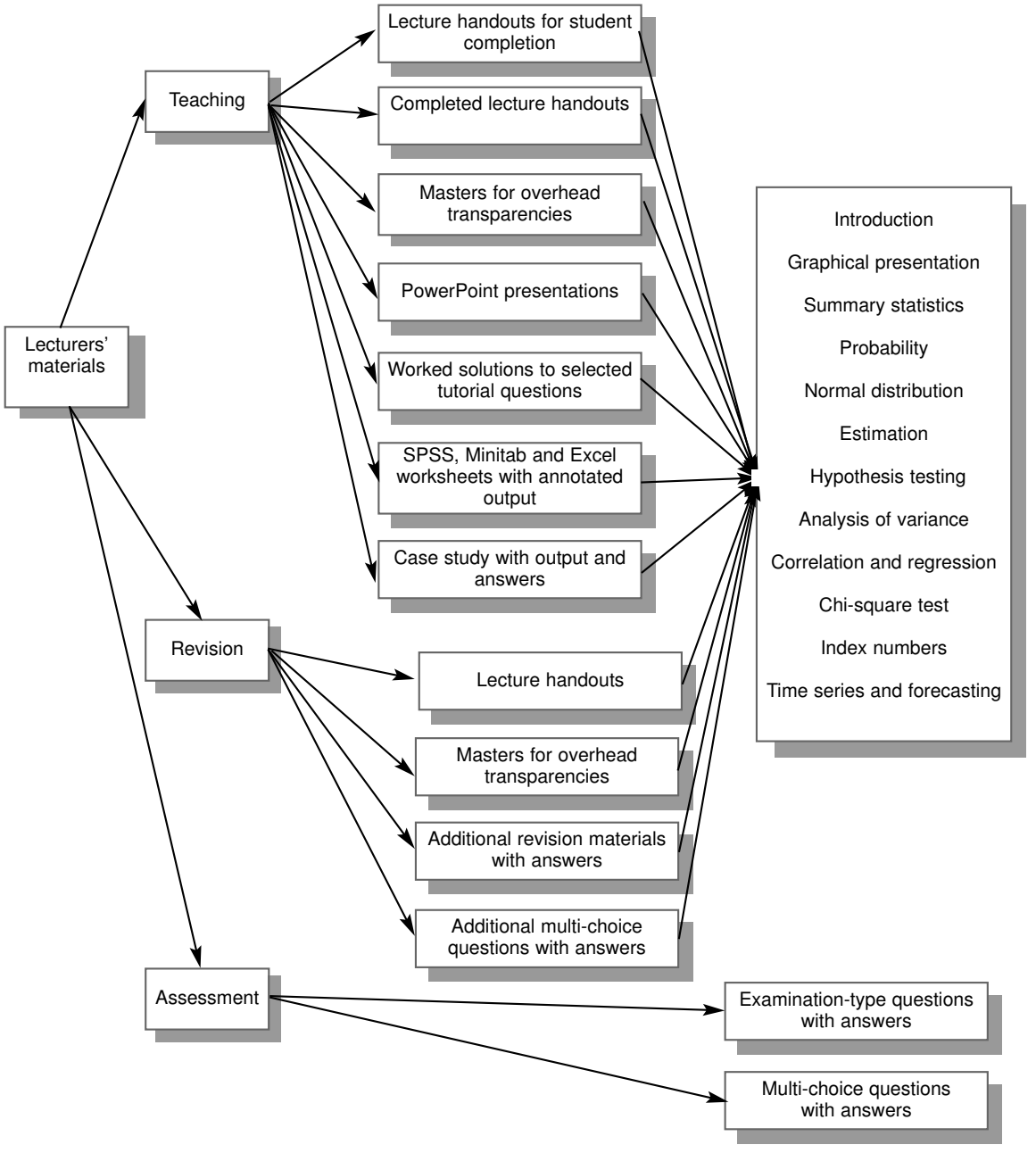
Assuming that *Business Statistics for non-mathematicians* is the set textbook for your course, each student will already have all the subject material, tutorial sheets with answers for each topic and computer worksheets (both SPSS and Minitab for most topics, Excel for some). In addition, they have access to the students materials listed above on the Companion website.

The folders and files providing extra materials on the companion website www.palgrave.com/business/taylor (plus a user name and password) for the lecturer are shown overleaf. These materials are listed in full in Appendix F of this book. It is intended that lecture preparation time should be reduced to a minimum so that as much time as possible can be spent on communicating the subject matter to the students rather than on the increasing burden of administration.

This course material was originally written, in varying forms, for business, accountancy and computing students at the start of modularisation. It has been refined in order to allow for the increasing size of student groups and a lower standard of numeracy. The course has been run with the inclusion of Minitab or SPSS or neither and has always been well received by the students over many years. The use of Excel and the comprehensive website materials have also been included in this book.

It can be seen, therefore, that this book is intended as a practical approach to the problem of the restricted time available for both the student and the lecturer which limits both the range and the depth of topics taught. This book is an honest attempt to solve the very real problem created by increasing student numbers, increasing demands on lecturers' time and students' decreasing mathematical capability. From this point of view I freely admit that the basic approach is quite different to that of most text books!

I am indebted to many colleagues, particularly Linda Pryce, with whom I developed the



initial course, Jenny Kromer, Jon Blacktop and Jane Parkin who gave generously of valuable time for proofreading the first edition. Others have contributed to this book both willingly and, possibly, inadvertently as the origin of a few of the examples is lost in history!

I also thank Martin Drewe from Palgrave Macmillan for his prompt friendly assistance when required during the production of this second edition.

Mainly my thanks go to my husband, Geoff, for his tolerance during the writing of this book!

Sonia Ann Taylor (Formerly of the University of Huddersfield)

1 Introduction to statistics

Objectives of this chapter

In this introductory chapter no assumptions are made about any prior statistical knowledge. The main aim is to introduce you gently to the subject of statistics and its use of various types of data.

Having studied this chapter you should be aware, in general, of:

- different types of data
- methods of data collection
- reasons for displaying, summarising and analysing data.

Throughout the book, terms printed in **bold** are either occurring for the first time or are particularly important. They are described in the glossary in Appendix B.

1.1 What do we mean by statistics?

Statistics are numerical facts or figures. Therefore statistics, as a science, essentially deals with numbers. It is generally taken to include the systematic:

- collecting
- classifying
- analysing
- presenting

of data in order to get a better understanding of some given situation.

This may mean summarising the data in tabulated, graphical or numerical form.

A statistical study might range from simple exploration enabling us to gain an insight into a virtually unknown situation to a sophisticated analysis designed to produce numerical confirmation, or rejection, of some widely held belief.

1.2 Why do we need statistics?

In business, we may be interested in a set of data in its own right. In this case we could describe it both numerically and graphically in the most appropriate manner – **descriptive statistics**.

Alternatively, the set of data we have may be a sample drawn from a larger population. This population would be our target of interest. In this case we need to use the information held by the sample to tell us something about its parent population – **inferential statistics**. An example is a Gallup poll on a sample of the electorate before a general election.

In another instance we may be interested in the future and so use the data we have up to the present time to estimate the value of a quantity in the future – **forecasting**.

Statistics play a wide role in most aspects of the competitive business world, where they provide an essential tool in decision making. Any decision-making process should be supported by some quantitative measures produced by the analysis of collected data.

Useful data may be on:

- a firm's products, costs, sales or services
- its competitors' products, costs, sales or services
- measurement of industrial processes
- a firm's workforce.

Once collected, this data needs to be summarised and displayed in a manner which helps its communication to, and understanding by, its recipients. Only when fully understood can it profitably become part of the decision-making process.

During this course you will learn to:

- describe data, such as profits, in order to assist decision makers
- estimate a particular property of a large population of data from a comparatively small sample
- seek out relationships between pairs of variables such as advertising and sales
- use known data to forecast a quantity, such as future demand.

First we shall look at some basic considerations which we must always take into account when collecting or handling data. These considerations do not fit into any specific topic area but are applicable throughout.

1.3 Types of data and scales of measurement

The word **data** describes, in general, a collection of observations.

Any data you use can take a variety of **values** or belong to various **categories**, either numerical or non-numerical. The 'thing' being described by the data is therefore known as a **variable**. The values or descriptions you use to measure or categorise this 'thing' are the **measurements**. These are of different types, each with its own appropriate **scale of measurement** which has to be considered when deciding on appropriate methods of graphical display or numerical analysis.

A variable is therefore simply something whose 'value' can vary.

For example a car could be red, blue, green, etc. It would be identified by a registration number. It could be classed as small, medium or large. Its petrol consumption in mpg could be 30, 40, 50, etc. Its year of manufacture could be 1991, 1999, 2006, etc. It would have a particular length. These values all describe the same car but are measured on different 'scales'.

1.3.1 Categorical data

These are generally non-numerical data which are placed into **exclusive** categories and then counted rather than measured. People are often categorised by their occupation or sex. The car mentioned above can be categorised by its make or colour.

1.3.1.1 Nominal data

The scale of measurement for a variable is **nominal** if the data describing it are simple names or labels which cannot be ordered. This is the lowest level of measurement. Numbers may represent nominal data, such as 'codes' for computer analysis, but these can only be used as labels. Vest numbers identify athletes but make no value statements about them. A car registration number only serves to identify the vehicle.

Numbers representing nominal data cannot be used in any arithmetic. Your 'PIN number' allows you access to your bank account, as does your friend's to his, but the sum of them doesn't allow either, never mind both, of you access to either. All nominal data are placed in a limited number of exhaustive categories and any analysis is carried out on the **frequencies** within these categories.

1.3.1.2 Ordinal data

If the categories of data can be placed in a meaningful order without any measurements, then the data are classed as **ordinal**. This is one level up from nominal. We know that the members of one category are more, or less, than the members of another but we cannot say by how much. For example, the results of a race are decided by the finishing order of the athletes

without any reference to their actual times. The vests they wear could be ordered as: 'small', 'medium' and 'large' without the aid of a tape measure.

Degree classifications are only ordinal because the difference between first and second-class degrees is not the same as the difference between a second and a third.

Questionnaires are often used to collect opinions using the categories: 'Strongly agree', 'Agree', 'No opinion', 'Disagree' or 'Strongly disagree'. The responses may be coded as 1, 2, 3, 4 and 5 for the computer, but the differences between these numbers are not claimed to be equal so the categories are only ordinal.

In order to analyse ordinal data, all individuals are placed in their relevant categories, the categories then ordered and calculations are performed on their frequencies.

1.3.2 Interval and ratio data

In interval and ratio scales of measurement all numbers are defined by standard units of measurement, such as metres or grams, so equal difference between numbers genuinely means equal distance between measurements. If there is **also** a meaningful zero, then the fact that one number is twice as big as another means that the measurements are also in that ratio. This data is known as **ratio data**. If, on the other hand, the zero has no mathematical meaning the data is **interval** only. Don't worry too much about this distinction as both sets of data are treated the same.

1.3.2.1 Interval data

There are very few examples of genuine interval scales. Temperature in degrees Celsius provides one example with the 'zero' on this scale being arbitrary. The difference between 30 °C and 50 °C is the same as the difference between 40 °C and 60 °C but we cannot claim that 60 °C is twice as hot as 30 °C. They are therefore **interval data** but not ratio data. Dates are measured on an interval scale as again the zero is arbitrary and not meaningful.

1.3.2.2 Ratio data

Ratio data must have a meaningful zero as their lowest possible value. For example, the time taken for athletes to complete a race would be measured on this scale. If we consider ages, a child at 12 years old is twice as old as his 6-year-old brother, and the age difference between them is the same as between his sisters who are 15 years old and 9 years old respectively. Their ages are ratio as well as interval but their dates of birth are only interval.

Suppose Bill earns £80,000, Ben earns £60,000 and Bob earns £40,000. The intervals of £20,000 between Bill and Ben and also between Ben and Bob genuinely represent equal amounts of money. Also the ratio of Bob's earnings to Bill's earnings is genuinely in the same ratio (1:2), as are the numbers which represent them. The value of £0 represents 'no money'. This data set is therefore **ratio** as well as interval.

The distinction between interval and ratio data is more theoretical than practical, as the same numerical and graphical methods are appropriate for both. They are usually referred to as **'at least interval'** data. Much of the data, such as money, that you will study will be measured on this scale.

We have therefore identified three measurement scales – 'nominal', 'ordinal' and 'at least interval'. The data measured on these scales are referred to in the same way, and this classification determines which methods of display and analysis are appropriate.

Any type of data may be analysed using methods appropriate to lower levels. For example: interval data may be analysed as ordinal but useful information is lost. If we know that Bill earns £80,000 and Ben earns £60,000 we are throwing information away by only recording that Bill earns 'more than' Ben.

Data cannot be analysed using methods which are only appropriate for higher-level data as the results will be either invalid or meaningless. For example it makes no sense to code the sex of students as 'male' = 1, 'female' = 2 and then report 'the mean value is 1.7'. It is however quite appropriate to report that '70 per cent of the students are female'.

1.3.3 Qualitative and quantitative data

Various definitions exist for the distinction between qualitative and quantitative data. Non-numerical (nominal) data are always described as being **qualitative** (non-metric) data as they describe some qualities without measuring them. **Quantitative** (metric) data which describe some measurement (or quantity) are always numerical. All definitions agree that interval or ratio data are quantitative. Some textbooks, however, use the term qualitative to refer to words only, while others also include nominal or ordinal numbers. Problems of definition could arise with numbers, such as house numbers, which identify or rank rather than measure. You don't need to worry about the 'grey areas'.

In the next two chapters we shall study the graphical and numerical methods appropriate to each type of data. The statistical techniques you will meet later are often divided into **parametric statistics,** which require data to be interval or ratio, and **non-parametric statistics,** which are appropriate for use at the lower levels.

1.3.4 Discrete and continuous data

Quantitative data may be **discrete** or **continuous**. If the values that can be taken by a variable change in steps, the data is **discrete**. These discrete values are often, but not always, whole numbers. If the variable can take any value within a range, so that you can imagine other values between those given, it is **continuous**. The number of people shopping in a supermarket is discrete but the amount they spend is continuous. The number of children in a family is discrete but a baby's birth weight is continuous. This is usually given to the nearest ounce but could be measured more precisely in half or quarter ounces. Some variables, such as money, are considered to be continuous even though they are measured in small discrete amounts.

1.4 Populations and samples

The **population** is the **entire group** of interest. This definition is not confined to people, as is usual in the non-statistical sense. A statistical population may include objects such as all the houses in a local authority area rather than the people living in them.

It is usually neither possible nor practical to examine every member of a **population**, so we use a **sample** – a smaller selection taken from that population – to **estimate** some value or characteristic of the whole population. Care must be taken when selecting the sample, as it must be representative of the whole population under consideration, otherwise it doesn't tell us anything relevant to that particular population.

Occasionally the whole population is investigated by a **census**, such as is carried out every ten years in the United Kingdom. The data are gathered from everyone in the population. A more usual method of collecting information is by a **survey** in which only a sample is selected from the population of interest and its data examined. Examples of this are the Mori polls produced from a sample of the electorate to forecast the result of a general election.

Analysing a sample instead of the whole population has many advantages such as the obvious saving of both time and money. It is often the only possible method, as the collecting of data may sometimes destroy the article of interest, for example the quality control of loaves of bread.

The ideal method of sampling is **random sampling,** in which every member of the population has an equal chance of being selected and each selection is independent of all the others. This ideal might not be achievable for a variety of reasons, and many other methods can be used (see Section 6.2).

1.5 Descriptive statistics

Descriptive statistics cover the analysis of the whole population of interest. The facts and figures usually referred to as 'statistics' in the media are very often a numerical and graphical summary of data from a specific group, for example unemployment figures. Much of the data generated by a business will be descriptive in nature, as will be the majority of sporting statistics.

1.6 Inferential statistics

If the information we have available is from a sample of the whole population of interest, we analyse it to produce the **sample statistics** from which we can infer (estimate) values for the parent population. This branch of statistics is usually referred to as **inferential statistics**. For example, we use the proportion of faulty items in a sample taken from a production line to estimate what proportion of all the items from the whole line are expected to be faulty. Pharmaceutical research is an example of the use of inferential statistics. Tests are necessarily limited to a small sample of patients, but inferences are applied to the whole relevant patient population.

A descriptive measure from the sample is usually referred to as a **sample statistic,** and the corresponding measure estimated for the population is referred to as a **population parameter**. The problem with using samples is that each sample produces a different sample statistic, giving us a different estimate for the population parameter. They cannot all be correct so a margin of error is generally quoted with any estimations.

1.7 Summary

In this short chapter you have been introduced to the different types of data we shall be using throughout the book. We shall be dealing mainly with interval data, such as money, which we shall learn to display, summarise and analyse. It is, however, important to be able to distinguish between the different types of data as this determines which methods of display and analysis are the most appropriate (see Table 1.1).

You have also been introduced briefly to the two main branches of statistics, descriptive and inferential statistics.

- Descriptive statistics result from gathering data about a whole group, or population, and reaching conclusions about that group only.
- Inferential statistics result from gathering data from a sample taken from a population and then reaching conclusions about the whole population from an analysis of the sample data.

Table 1.1 Scales of measurement

Scale of measurement	Non-numeric data	Numeric data
Nominal	Name or label only	Numbers only identify groups which cannot be ordered
Ordinal	Names or labels can be ranked	These numbers allow ranking but no arithmetic
Interval	Always numeric	Intervals between numbers are meaningful
Ratio	Always numeric	Intervals between numbers are meaningful and also their ratios as the lowest value is a meaningful zero.

1.8 Case study

A case study requiring you to investigate data in order to provide a quantitative basis for decision making runs throughout this book. Partial studies are to be found near the end of each chapter, requiring analysis from you as appropriate.

1.8.1 Scenario

You work for a company, Restful Restaurants, which is exploring the possibility of opening new premises in the expanding city of Lonbridge. The directors of your company need some quantitative evidence on which to base any decisions. If they open the company's standard type of restaurant, how big should it be? Might it be preferable to diversify and open a different type of food outlet, such as a large café or a takeaway? The directors have obtained survey data from the University of Lonbridge about existing food outlets in the city, and are asking you to analyse it in order to provide them with background information on the current situation regarding restaurants, cafés and takeaways.

More specifically, at the end of the case study, they would like to know for each type of establishment:

- its present frequency in Lonbridge
- how optimistic the owners feel about future sales
- how the estimated value of a business relates to (a) its gross sales, (b) its number of employees, and (c) the amount it spends on advertising
- how gross sales are related to the amount of new capital invested each year
- the proportion of sales revenue spent on (a) buying goods and (b) staff wages

- how effective advertising is in increasing sales
- how the number of employees is related to gross sales
- whether the size of an establishment relates to its type.

In order to build up a picture of the situation in Lonbridge you will be asked to carry out analysis in most chapters after you have studied specific topics, using your preferred computer package.

1.8.2 The data

This data originated as a survey of restaurants in Wisconsin, Canada, and was originally reproduced as a Minitab data set (Restrnt.MTW). The data set has been reduced in size and slightly adapted for this case study. This version will be found on the companion website in the folder 'Case study' in a file called 'Restaurants'.

It includes the following information on 255 eating establishments:

```
Variable N      Label
OUTLOOK  255    Business Outlook
SALES    245    Gross sales (£'000)
NEWCAP   217    New capital invested (£'000)
VALUE    239    Estimated market value of business (£'000)
COSTGOOD 233    Cost of goods sold as % of sales
WAGES    233    Wages as % of sales
ADVERT   231    Advertising as % of sales
TYPE     254    Type of Outlet
OWNER    250    Type of ownership
FULL     248    Number of full-time employees
PART     248    Number of part-time employees
SIZE     246    Size of establishment in FTE equivalent
```

1.9 Check your course prerequisites

There is no exercise connected to the material in this introductory chapter. You should check, or revise, the basic mathematical knowledge that will be assumed when working through this book. This knowledge is fairly basic. In addition to a sound working knowledge of simple arithmetic you should be able to:

- work with fractions, decimals and percentages
- handle large and small numbers
- carry out simple algebraic manipulations
- solve fairly simple equations.

The following tutorial should enable you to check these prerequisites. If you find any question particularly difficult you should get some practice from any basic quantitative analysis or mathematics textbook.

Tutorial 1: Basic mathematics revision

1 Evaluate:

a) $\frac{3}{7}-\frac{4}{21}+\frac{5}{3}$ b) $1\frac{2}{3}+2\frac{4}{5}-3\frac{1}{2}$ c) $\frac{2}{7}\times\frac{14}{25}\times\frac{15}{24}$ d) $\frac{16}{21}\div\frac{4}{7}$

e) $\frac{4}{9}\times\frac{3}{16}\div\frac{7}{12}$ f) $4\frac{2}{5}\div\frac{11}{12}\times1\frac{7}{8}$

2 a) Give 489 267 to 3 significant figures
b) Give 489 267 to 2 significant figures
c) Give 0.002 615 to 2 significant figures
d) Give 0.002 615 to 1 significant figure
e) Give 0.002 615 to 5 decimal places
f) Give 0.002 615 to 3 decimal places

3 Retail outlets in a town were classified as small, medium and large and their numbers were in the ratio 6 : 11 : 1. If there were 126 retail outlets altogether, how many were there of each type?

4 Convert:
a) 28% to a fraction in its lowest terms
b) 28% to a decimal

c) $\frac{3}{8}$ to a decimal

d) $\frac{3}{8}$ to a percentage

e) 0.625 to a fraction in its lowest terms
f) 0.625 to a percentage

5 Express the following in standard form:
a) 296 000 b) 0.000 296 c) 0.4590 d) 459.0 e) $\frac{1}{25\,000}$ f) $\frac{1}{0.000\,25}$

6 Reduce the following expressions to their simplest form, expanding brackets if appropriate.
a) $3a + b + 2a - 4b$ b) $2a + 4ab + 3a^2 + ab$ c) $a^2(3a + 4b + 2a)$
d) $(x + 2)(x + 4)$ e) $(x + 2)^2$ f) $(x + 1)(x - 1)$

7 Make x the subject of the formula and evaluate when $y = -3$.

a) $y = 5x - 4$ b) $y = x^2 - 7$ c) $y = 2(3 + 6x)$ d) $y = \frac{3}{x}$

8 Find the value of x in the formulae in Question 7 when $y = -3$.

9 Evaluate the following when $x = -2$, $y = 5$ and $z = 4$.
a) xy b) $(xy)^2$ c) $(xy + z)^2$
d) $zy - x^2$ e) $(x + z)(2y - x)$ f) $x^2 + y^2 + z^2$

10 Solve for x:
a) $3x - 1 = 4 - 2x$ b) $2(x - 3) = 3(1 - 2x)$ c) $\frac{3}{x-1} = \frac{1}{2}$

Answers

1 a) $1\frac{19}{21}$ b) $\frac{29}{30}$ c) $\frac{1}{10}$ d) $1\frac{1}{3}$ e) $\frac{1}{7}$ f) 9

2 a) 489 000 b) 490 000 c) 0.0026 d) 0.003
e) 0.002 62 f) 0.003

3 42 small, 77 medium, 7 large

4 a) $\frac{7}{25}$ b) 0.28 c) 0.375 d) 37.5%

e) $\frac{5}{8}$ f) 62.5%

5 a) 2.96×10^5 b) 2.96×10^{-4} c) 4.59×10^{-1} d) 4.59×10^2
e) 4.0×10^{-5} f) 4.0×10^3

6 a) $5a - 3b$ b) $3a^2 + 5ab + 2a$ c) $5a^3 + 4a^2b$ d) $x^2 + 6x + 8$
e) $x^2 + 4x + 4$ f) $x^2 - 1$

7 a) $x=\frac{1}{5}(y+4)$ b) $x=\sqrt{y+7}$ c) $x=\frac{y-6}{12}$ d) $x=\frac{3}{y}$

8 a) 0.2 b) +2 or –2 c) –0.75 d) –1

9 a) –10 b) 100 c) 36 d) 16 e) 24 f) 45

10 a) $x = 1$ b) $1\frac{1}{8}$ c) 7

2 Graphical representation of data

Objectives of this chapter

So far you have considered four different types of data: nominal, ordinal, interval and ratio. In this chapter we shall investigate the different ways of presenting this data graphically in a meaningful manner. You are probably already familiar with **bar charts** and **pie charts** so they will be considered fairly briefly. The emphasis will be on **histograms** and **cumulative frequency diagrams**, plus an introduction to **stem-and-leaf plots** and **box plots**.

Graphs are potentially very good tools for communication, but they must be kept as simple as possible and never be presented so as to mislead the reader. If they can't be easily understood they are neither use nor ornament!

After studying this chapter you should be able to:

- draw appropriate graphs of given data
- interpret a variety of types of graph
- understand graphs presented in the media.

2.1 Introduction: why do we represent data by graphs?

What is your reaction if you are presented with a table full of figures? I'm sure that you are neither filled with delight nor an immediate understanding of the situation described by them! On the other hand, a fairly simple graph, if well presented, can quickly convey a general summary of a set of data to its reader. Further examination can then reveal more detail and produce a deeper understanding. If you always keep in mind that the purpose of drawing a graph is to convey information to its reader as easily and quickly as possible, you will not go far wrong. Simple graphs make for easy comparisons, so never put too much information in any one picture.

2.2 Tabulation

Immediately after collection, all new data (**raw data**) are in the form of individual figures. There may be pages of these, often far too many to convey any useful information as they stand, or they may be presented as a table. These individual data are ungrouped, so the first step in organising them is usually to produce **grouped data**: that is, collect like with like in order to reduce the total volume. This grouping can take a variety of forms – some students like to use tally charts – which all result in a **frequency distribution**. Various statistics can be calculated from the frequency distribution and suitable graphs produced.

First determine the range of the data: that is, the largest value minus the smallest value. Then decide on suitable **class intervals** to give a reasonable number of classes or groups. Somewhere between 5 and 15 is generally acceptable – too few classes will cause the loss of detail but too many classes may obscure the overall picture. In order to keep the work simple it is advisable to group in fives or tens, and in the first grouping you should use equal intervals. Data values must not fall into more than one class, so the class interval is usually described as, for example, '20 and under 30' so that 30 would go in the next class.

Now group the data by constructing a tally chart. In Example 2.1, record the first number, 55, next to the relevant interval, '55 and under 60', with a '|'. Then allocate the second, 64, to the '60 and under 65' interval, and so on until all the numbers have been recorded. After four strokes it is conventional to draw a diagonal through them for the fifth so that the number of 'fives' can easily be totalled. Total each 'tally' for the frequency in each interval.

Example 2.1

The following figures represent the examination marks (per cent) for 60 students on a business studies course. We shall first find the range, then decide the number of intervals to use, define the class intervals, and finally draw up a frequency table.

Table 2.1 gives Range: 84 – 37 = 47. These figures would give us either 6 class intervals of 10 or 10 class intervals of 5. Either would be acceptable so we shall first group in classes of 5 and then combine into intervals of 10 and compare the results.

Table 2.1

55	64	74	53	66	40	52	39	70	59	53	57
62	60	40	45	54	72	47	42	38	60	43	37
61	65	41	54	69	47	80	66	52	78	43	72
44	84	61	67	74	57	60	61	60	56	66	49
54	59	59	60	57	70	61	54	67	54	65	56

Frequency distribution table

Class interval	Tally	Frequency
35 & under 40	|||	3
40 & under 45	𝍸 ||	7
45 & under 50	||||	4
50 & under 55	𝍸 ||||	9
55 & under 60	𝍸 ||||	9
60 & under 65	𝍸 𝍸 |	11
65 & under 70	𝍸 |||	8
70 & under 75	𝍸 |	6
75 & under 80	|	1
80 & under 85	||	2

This gives us an overall idea of the 'shape' of the distribution. Does it look better if grouped in tens?

Frequency distribution table

Class interval	Tally	Frequency
30 & under 40	|||	3
40 & under 50	𝍸 𝍸 |	11
50 & under 60	𝍸 𝍸 𝍸 |||	18
60 & under 70	𝍸 𝍸 𝍸 ||||	19
70 & under 80	𝍸 ||	7
80 & under 90	||	2

It is probably easier to interpret this second distribution: that is, it shows that 'most' of the students pass (40 per cent) but get less than 70 per cent. Have we lost any detail of interest? Probably not, in this case. There is no set rule about the number of classes, so just use a 'reasonable' number.

A frequency distribution already gives a good indication of the 'shape' of the data, but a well-drawn graph, such as a bar chart or histogram, communicates it better.

2.3 Graphs of non-metric (non-measurable) data

2.3.1 Bar charts

Bar charts are drawn as a pictorial summary of categorical data. Each bar, which is separated from its neighbours, represents one category, and the length of the bar represents the

frequency in that category. Alternatively the length of the bar is proportional to the size of the items being considered, for example the Sales in Example 2.2. For nominal data the bars may be arranged in any order, but with ordinal data the categories are usually presented in ascending order. For comparative purposes the bars may be grouped, the frequencies stacked, or 'percentage charts' drawn, as in Example 2.2.

Example 2.2

A shop had sales from four departments for the last four quarters (£'000):

Table 2.2

Department	Spring	Summer	Autumn	Winter	Total
Food	160	180	180	200	720
Clothing	280	300	200	300	1080
Furniture	860	560	500	240	2160
Electrical	60	60	100	140	360
Total	1360	1100	980	880	4320

Figure 2.1 Bar charts for comparing the total and departmental quarterly sales (diagrams from Microsoft Excel 2003 – see Section 14.8, Excel Worksheets)

2.3.2 Pie charts

Pie charts are also drawn as a summary of categorical data. The total count of all the data represented is equivalent to 360° on the 'pie', and the frequency within each category is represented by the size of the angle of its sector: that is, its 'slice'. If pie charts are to be used for comparing **relative frequencies** between variables, then the area of the 'pie' describing each can be drawn so that it represents the total frequency of that variable.

Calculating the sector angles (refer to Table 2.2 for total sales):

$$\text{Food: } \frac{720}{4320} \times 360° = 60° \qquad \text{Clothing: } \frac{1080}{4320} \times 360° = 60° \text{ and so on}$$

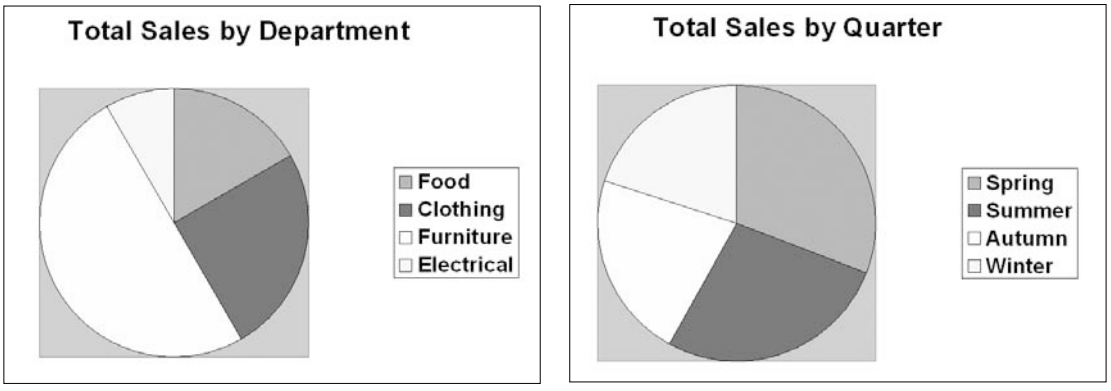


Figure 2.2

By hand, and some computer packages, all this information can be put onto one 'pie' by placing a second ring around, say, the simple departmental pie shown in the first diagram and then subdividing each departmental sector radially between the different quarters. These more complicated 'pies' can be difficult to interpret, and should be used with caution.

(For production of pie charts by computer see Sections 14.2.1, 14.5.1 and 14.8.1.)

2.3.3 Pictograms

You are not expected to draw pictograms but it useful to know how to interpret any shown in the media. You are probably familiar with the type of graph produced by the motor industry in which each little car drawn represents, for example, 1000 cars actually produced. Ten tiny cars might represent 10,000 real cars from one factory, and 20,000 real cars from another factory might be represented by 20 tiny cars. That method of pictorial comparison is fine as its meaning is clear. Other methods can be misleading.

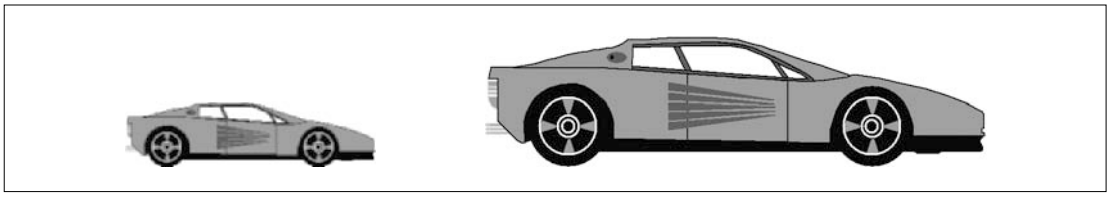


Figure 2.3

If the first car in Figure 2.3 represents 1000 real cars, how many does the second car represent? Its length is twice, its area four times and its volume eight times that of the first car. Does it represent 2000, 4000 or 8000 real cars? Diagrams of this type are open to misinterpretation and should be avoided. Published diagrams in the press need very careful scrutiny if they are not explained numerically.

2.4 Graphs of metric (measurable) data

2.4.1 Histograms

In Section 2.2 we grouped raw data to form a frequency distribution table. This gave a much clearer picture than did the individual data values. The most usual presentation of metric data is in the form of a **histogram**.

A histogram is a pictorial method of presenting frequency data. It appears similar to a bar chart but has two fundamental differences:

- The data must be measurable on a **continuous** scale, for example, lengths rather than colours.
- The **area** of a rectangle rather than its height is proportional to the frequency, so if one column is twice the width of another its height must be halved for the same frequency.

Histograms are produced by all statistical software packages but these often do not give you as much choice in presentation as is available when you draw them by hand.

Using the same data as in Example 2.1, we can look at the output from Minitab, one of the most commonly used educational packages, which is shown in Figure 2.4(a). Figure 2.4(b) has been grouped in 'tens' and the labelling has been moved to the class boundaries, which is always preferable. SPSS produces similar diagrams but does not allow the labelling to be moved. Excel does not produce genuine histograms but only contrived bar charts.

The histograms in Figure 2.4 both show equal class intervals, which is the default format in all packages. Neither Minitab nor SPSS will allow for the use of unequal intervals, so this will be described below in the method for hand-drawing histograms.

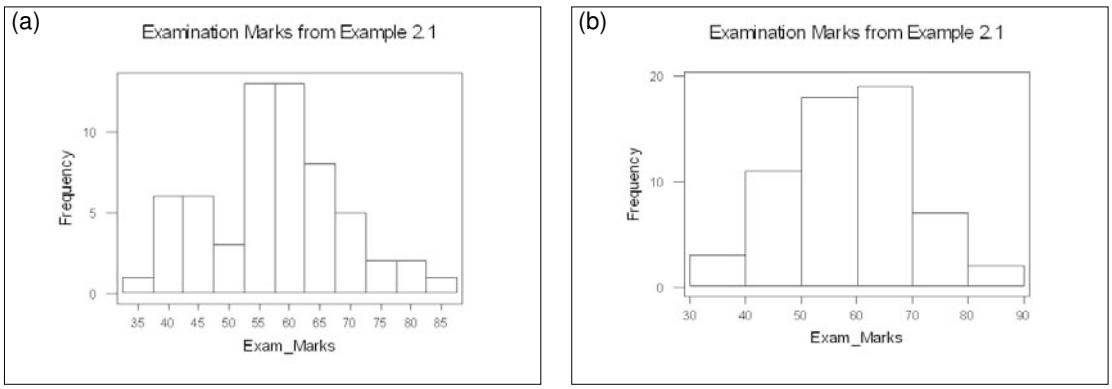


Figure 2.4

The construction of a frequency distribution table is shown again in Example 2.3. The data are first grouped in a frequency distribution table:

- Decide on sensible limits if the first or last class interval is left open, for example, 'less than 20', and also decide into how many classes you intend to group your data. Too few classes may hide information about the data; too many classes may hide its overall shape.
- Construct a frequency distribution table, grouping the data into a reasonable number of classes, (somewhere in the order of 10). Intervals are usually of the same width for the first summary.

Example 2.3

You are working for the transport manager of a large chain of supermarkets which leases cars for the use of its staff. She is interested in the weekly distances covered by these cars. Mileages recorded for a sample of hired vehicles during a given week yielded the following data for Fleet 1:

Table 2.3

138	164	150	132	144	125	149	157	161	150
146	158	140	109	136	148	152	144	145	145
168	126	138	186	163	109	154	165	135	156
146	183	105	108	135	153	140	135	142	128

Minimum = 105, Maximum = 186, Range = 186 – 105 = 81

Nine intervals of 10 miles width seams reasonable, but the first and last interval may be wider if data proves to be scarce at the extremes.

Frequency distribution table

Class interval		Frequency
100 & less than 110	||||	4
110 & less then 120		
120 & less than 130	|||	3
130 & less than 140	𝍸 ||	7
140 & less than 150	𝍸 𝍸 |	11
150 & less than 160	𝍸 |||	8
160 & less than 170	𝍸	5
170 & less than 180		
180 & less than 190	||	2
	Total	40

It might be preferable to combine the two intervals at each end of the table and work with frequency density (per 10 mile interval).

Frequency distribution table

Class interval		Frequency	Freq/10 miles
100 & less than 120	||||	4	2.0
120 & less than 130	|||	3	3.0
130 & less than 140	𝍸 ||	7	7.0
140 & less than 150	𝍸 𝍸 |	11	11.0
150 & less than 160	𝍸 |||	8	8.0
160 & less than 170	𝍸	5	5.0
170 & less than 190	||	2	1.0
	Total	40	

The histogram is then constructed:

- Frequencies are plotted in proportion to the area of each rectangle, so if the intervals (the rectangle bases) are not all the same width, their heights need to be calculated. These heights are known as the **frequency densities**, that is, frequency per constant interval. The most commonly occurring interval is often used.
- Construct the histogram, labelling each axis carefully. Hand-drawn histograms usually show the frequency vertically. (Computer output may be horizontal because it is more convenient for line printers.)

We shall return to histograms in the next chapter for the estimation of modal values.

Example 2.3 continued

If some intervals are wider than others care must be taken that the areas of the blocks are proportional to the **frequencies**, so heights are proportional to **frequency densities**. Figures 2.5 (a) and (b) illustrate the difference; (a) has the **frequency** plotted on the vertical axis and (b) has the **frequency density** with the two outside intervals combined at both of the extremes. It looks more aesthetically pleasing. (See Sections 14.2.1, 14.5.1, 14.8.1 for computer production of histograms).

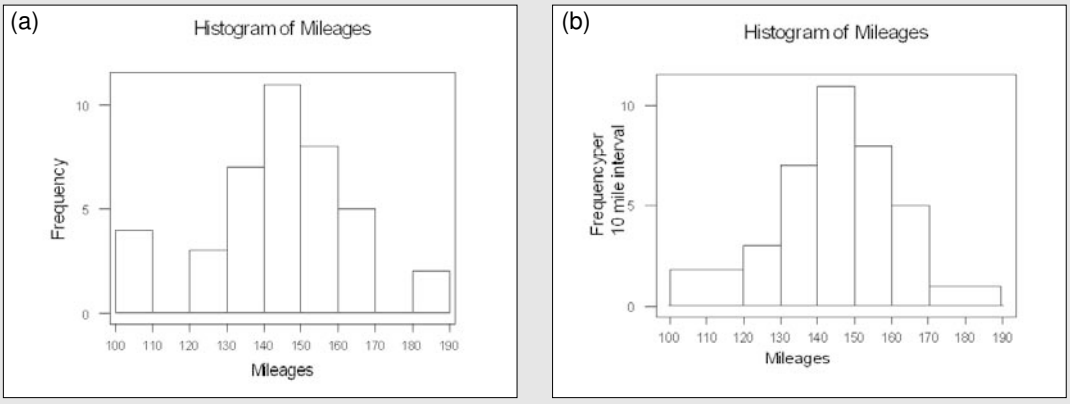


Figure 2.5

2.4.2 Frequency polygons

A **frequency polygon** is constructed by joining the midpoints at the top of each column of the histogram. The area under the polygon is the same as that under the histogram. Polygons can also be drawn without using a histogram. It is often easier to compare two frequency polygons than to interpret than a pair of histograms which tend to obscure each other. The polygons are drawn without the histograms, giving a clearer comparison.

For example if we wished to compare another fleet of 40 cars, Fleet 2, with the one in Example 2.3, Fleet 1, the diagram might look like Figure 2.6(b).

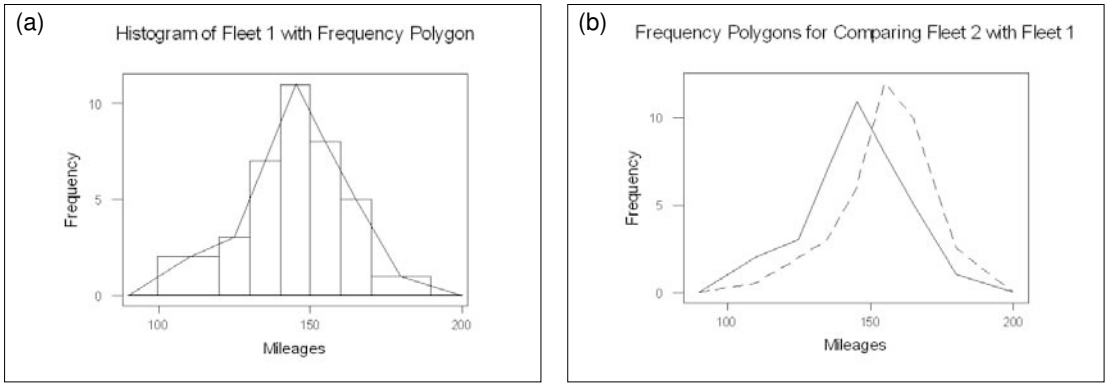


Figure 2.6

If we wish to compare two fleets of different sizes, comparing frequencies directly would not give a clear picture. Instead it would be advisable to compare **relative frequencies**. These measure the percentage of the total fleet which lies within each class interval.

For example for the interval 110 to 120 in Example 2.4, the relative frequency for Fleet 2 is 1/40 = 2.5% and for Fleet 3 is 9/200 = 4.5%.

Example 2.4

Table 2.4

	Fleet 2		Fleet 3	
Class Interval	Frequency	Relative freq.%	Frequency	Relative freq.%
less than 100	0	0.0	0	0.0
100 & less than 110	0	0.0	9	4.5
110 & less than 120	1	2.5	9	4.5
120 & less than 130	2	5.0	35	17.5
130 & less than 140	4	10.0	70	35.0
140 & less than 150	6	15.0	40	20.0
150 & less than 160	12	30.0	23	11.5
160 & less than 170	10	25.0	12	6.0
170 & less than 180	4	10.0	2	1.0
180 & less than 190	1	2.5	0	0.0
Total	40	100.0	200	100.0

If we wished to compare our fleet, Fleet 2, with a much larger fleet, Fleet 3, we would calculate the relative frequencies for each fleet and plot those values instead of the frequencies (see Table 2.4). This makes the small and large fleets easier to compare as the areas under the graphs are equal.

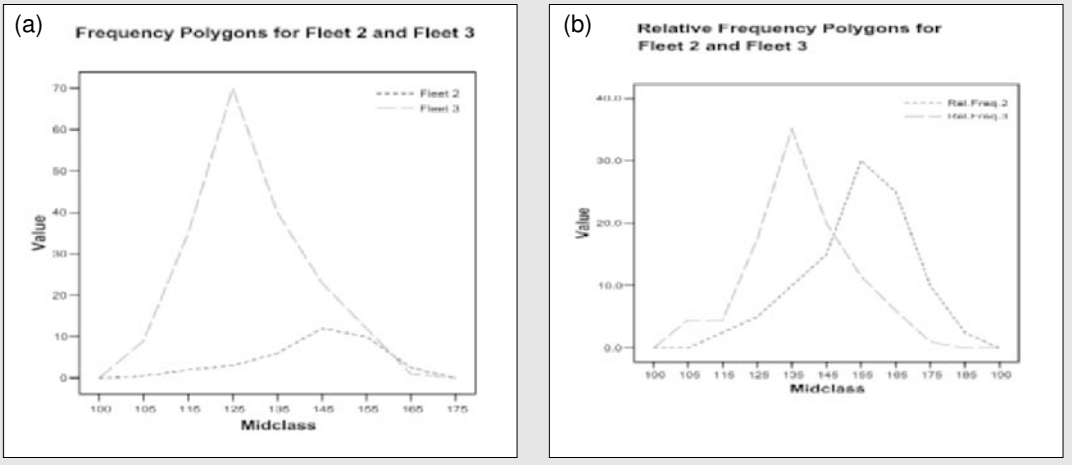


Figure 2.7

2.4.2.1 Shape of a distribution

We can also see whether the data distributions are symmetrical or not. If the 'peak' is towards the left and the longer tail towards the right, the data is referred to as 'positively [or right] skewed', and conversely if the 'peak' is towards the right and the tail towards the left, it is 'negatively [or left] skewed'. If neither is true, it is 'symmetrical'. The distribution of the Fleet 2 data set is seen to be slightly negatively skewed, and Fleet 3 positively, but both are nearly symmetrical. You will meet this concept again later in the course.

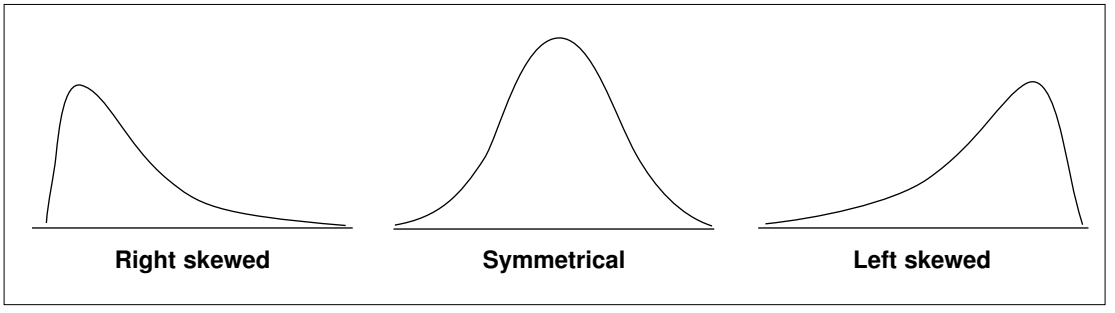


Figure 2.8

2.4.3 Stem-and-leaf plots

The **stem-and-leaf plot** displays the data in the same 'shape' as the histogram, although it tends to be shown horizontally. The main difference is that it retains all the original numerical information. The values themselves are included in the diagram so no information is 'lost'.

The stem-and-leaf plot in Figure 2.9 represents the Fleet 1 vehicles. The 'stem', on the left of the vertical line of dots, contains the first two digits (hundreds and tens), and the 'leaf' shows the units. The 'stem' contains the most significant digits. These might be all different, or grouped depending on the spread of the data (see 'Stem width' in Figure 2.9). The 'leaves', on the right of the vertical line, display the less significant digits. For small numbers these may be units (as in mileages), but for larger numbers they may represent tens, hundreds, etc. For

```
MILEAGES Stem-and-Leaf Plot

 Frequency     Stem &  Leaf
     4.00        10 .   5899
      .00        11 .
     3.00        12 .   568
     7.00        13 .   2555688
    11.00        14 .   00244556689
     8.00        15 .   00234678
     5.00        16 .   13458
      .00        17 .
     2.00        18 .   36

 Stem width:         10
 Each leaf:        1 case(s)
```

Figure 2.9 (from SPSS)

```
Salary Stem-and-Leaf Plot

 Frequency     Stem &  Leaf
     3.00         3 .   699
    12.00         4 .   000023333344
    17.00         4 .   55556666688888889
    17.00         5 .   01111122224444444
    13.00         5 .   5557777777788
    37.00         6 .   00000000000000003333333333333333333
    15.00         6 .   666666666666999
     4.00         7 .   2222
     6.00         7 .   558889
     2.00         8 .   14
     5.00         8 .   57777
     1.00         9 .   3
     1.00         9 .   9
    23.00 Extremes     (>=102000)

 Stem width:    10000
 Each leaf:        1 case(s)
```

Figure 2.10 (from SPSS)

small data sets a leaf will represent a single number, but for larger sets it might represent many. (See 'Each leaf' in Figure 2.9.)

The first row in Figure 2.9 represents the values 105, 108, 109 and 109. We can see that the mileages travelled range from 105 to 186. Notice that the 'leaves' are arranged in numerical order so that it is easy to find the value of the middle mileage or the third largest, and so on. If you draw stem-and-leaf plots by hand, it is necessary to group the values in tens but also to arrange carefully the order of the leaves. Fortunately Minitab and SPSS do this for us. Excel does not produce these plots.

In the larger data set in Figure 2.10, describing the salaries earned by a firm's workers, each stem width of £10,000, as indicated by 'Stem width', has been split into two rows. The leaves will be one number less significant, in this case measured to the nearest thousand, so the first value is £36,000 to two significant figures. Extreme values may often be grouped together. In this case we have 23 salaries which are greater than or equal to £102,000. For larger data sets the leaves may represent more than one case, but we are informed in Figure 2.10 that each leaf in this stem-and-leaf plot represents just one case.

2.4.4 Dot plots

A useful quick picture of the data can be formed by keeping a running total of the situation by means of a **dot plot**. This would be suitable for the 'straw polls' which are invariably taken from voters entering polling stations on election days. All the candidates' names would be displayed on the horizontal axis and a 'dot' plotted against the appropriate name for each voter who favours them. After about the first hour of opening at one particular polling station the picture might look something like Figure 2.11.

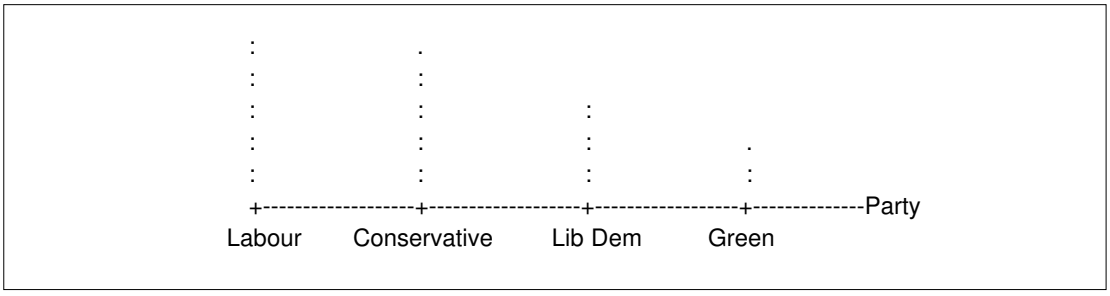


Figure 2.11

The advantage of this method, as opposed to using a bar chart or a pie chart is that the picture is built up gradually and does not need to wait for all the data to be collected before the diagram can be drawn.

2.4.5 Cumulative frequency polygons (ogives)

A **cumulative frequency polygon** is a graphical method of representing the accumulated frequencies up to and including a particular value. Think of it as a running total less than a stated value; for example, the proportion of a workforce earning up to £60,000 or cars doing up to 155 miles per week. These cumulative frequencies are often calculated as percentages of the total frequency. This method is used for estimating median and quartile values, and