Business Mathematics and Statistics

Andre Francis BSc MSc

Perinatal Institute
Birmingham

Andre Francis works as a medical statistician. He has previously taught Mathematics, Statistics and Information Processing to students on business and professional courses. His teaching experience has covered a wide area, including training students learning basic skills through to teaching undergraduates. He has also had previous industrial (costing) and commercial (export) experience and served for six years in statistical branches of Training Command in the Royal Air Force.

Sixth Edition

THOMSON

Australia • Canada • Mexico • Singapore • Spain • United Kingdom • United States

Acknowledgements

The author would like to express thanks to the many students and teachers who have contributed to the text in various ways over the years.

In particular he would like to thank the following examining bodies for giving permission to reproduce selected past examination questions:

Chartered Association of Certified Accountants (ACCA)
Chartered Institute of Management Accountants (CIMA)
Institute of Chartered Secretaries and Administrators (ICSA)
Chartered Institute of Insurance (CII)
Association of Accounting Technicians (AAT)

Each question used is cross referenced to the appropriate Institute or Association.

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

First Edition 1986

Second Edition 1988; Reprinted 1990; Reprinted 1991
Third Edition 1993; Reprinted 1993
Fourth Edition 1995; Reprinted 1996; Reprinted 1997
Fifth Edition 1998; Reprinted 2003 by Thomson Learning

Sixth Edition 2004; Published by Thomson Learning

ISBN 1-84480-128-4

Typeset in Nottingham, UK by *Andre Francis*

Printed in Croatia by *Zrinski d.d.*

Contents

Preface

1. Aims of the book

The general aim of the book is to give a thorough grounding in basic Mathematical and Statistical techniques to students of Business and Professional studies. No prior knowledge of the subject area is assumed.

2. Courses covered

a) The book is intended to support the courses of the following professional bodies:
   Chartered Association of Certified Accountants
   Chartered Institute of Management Accountants
   Institute of Chartered Secretaries and Administrators

b) The courses of the following bodies which will be supported by the book to a large extent:
   Chartered Institute of Insurance
   Business and Technical Education Council (National level)
   Association of Accounting Technicians

c) The book is also meant to cater for the students of any other courses who require a practical foundation of Mathematical and Statistical techniques used in Business, Commerce and Industry.

3. Format of the book

The book has been written in a standardised format as follows:

a) There are TEN separate parts which contain standard examination testing areas.

b) Numbered chapters split up the parts into smaller, identifiable segments, each of which have their own Summaries and Points to Note.

c) Numbered sections split the chapters up into smaller logical elements involving descriptions, definitions, formulae or examples.

At the end of each chapter, there is a Student Self Review section which contains questions that are meant to test general concepts, and a Student Exercise section which concentrates on the more practical numerical aspects covered in the chapter.

At the end of each part, there is

a) a separate section containing examination examples with worked solutions and

b) examination questions from various bodies. Worked solutions to these questions are given at the end of the book.

4. How to use the book

Chapters in the book should be studied in the order that they occur.

After studying each section in a chapter, the Summaries and Points to Note should be checked through. The Student Self Review Questions, which are cross-referenced to appropriate sections, should first be attempted unaided, before checking the answers with the text. Finally the Student Exercises should be worked through and the answers obtained checked with those given at the end of the book.

After completing a particular part of the book, the relevant section of the examination questions (at the end of the book) should be attempted. These questions should be considered as an integral part of the book, all the subject matter included having been covered in previous chapters and parts. Always make some attempt at the questions before reading the solution.

5. The use of calculators

Examining bodies permit electronic calculators to be used in examinations. It is therefore essential that students equip themselves with a calculator from the beginning of the course.

Essential facilities that the calculator should include are:

a) a square root function, and
b) an accumulating memory.

Very desirable extra facilities are:

c) a power function (labelled 'x^y'),
d) a logarithm function (labelled 'log x'), and
e) an exponential function (labelled 'e^x').

Some examining bodies exclude the use (during examinations) of programmable calculators and/or calculators that provide specific statistical functions such as the mean or the standard deviation. Students are thus urged to check on this point before they purchase a calculator. Where relevant, this book includes sections which describe techniques for using calculators to their best effect.

Andre Francis, 2004

1 Introduction to business mathematics and statistics

1. Introduction

This chapter serves as an introduction to the whole book. It describes the main areas covered under the heading 'Business Mathematics and Statistics' and introduces the idea of a statistical investigation.

2. Differences in terminology

The title of this book is Business Mathematics and Statistics. However, many other terms are used in business and by Professional bodies to describe the same subject matter. For example, Quantitative Methods, Quantitative Techniques and Numerical Analysis.

3. Business mathematics and statistics

A particular problem for management is that most decisions need to be taken in the light of incomplete information. That is, not everything will be known about current business processes and very little (if anything) will be known about future situations. The techniques described in 'Business Mathematics and Statistics' enable structures to be built up which help management to alleviate this problem. The main areas included in the book are: (a) Statistical Method; (b) Management Mathematics; and (c) Probability.

These areas are described briefly in the following sections.

4. Statistical method

Statistical method can be described as:

a) the selection, collection and organisation of basic facts into meaningful data, and then
b) the summarizing, presentation and analysis of data into useful information.

The gap between facts as they are recorded (anywhere in the business environment) and information which is useful to management is usually a large one. (a) and (b) above describe the processes that enable this gap to be bridged. For example, management would find percentage defect rates of the fleets of lorries in each branch more useful than the daily tachometer readings of individual vehicles. That is, management generally require summarized values which represent large areas under their control, rather than detailed figures describing individual instances which may be untypical.

Note that the word 'Statistics' can be used in two senses. It is often used to describe the topic of Statistical Method and is also commonly used to describe values which summarize data, such as percentages or averages.

5. Management mathematics

The two areas covered in this book which can be described as Management Mathematics are described as follows:

a) *The understanding and evaluation of the finances involved in business investments*. This involves considering interest, depreciation, the worth of future cash flows (present value), various ways of repaying loans and comparing the value of competing investment projects;
b) *Describing and evaluating physical production processes in quantitative terms*. Techniques associated with this area enable the determination of the level of production and prices that will minimise costs or maximise the revenue and profits of production processes;

Involved in both of the above are the manipulation of algebraic expressions, graph drawing and equation solving.

6. Probability

Probability can be thought of as the ability to attach limits to areas of uncertainty. For example, company profit for next year is an area of uncertainty, since there will never be the type of information available that will enable management to forecast its value precisely. What can be done however, given the likely state of the market and a range of production capacity, is to calculate the limits within which profit is likely to lie. Thus calculations can be performed which enable statements such as 'there is a 95% chance that company profit next year will lie between £242,000 and £296,000' to be made.

7. Statistical investigations

Management decisions are based on numerous pieces of information obtained from many different sources. They may have used one, some or all of the techniques which have been described as Statistical Method, Management Mathematics or Probability. What the decisions will all have in common however is that they are the final product of a general structure (or set of processes) known as an *investigation* or *survey*. Some significant factors are listed as follows.

a) Investigations can be fairly trivial affairs, such as looking at today's orders to see which are to be charged to credit or cash. Others can be major undertakings, involving hundreds of staff and a great deal of expense over a number of years, such as the United Kingdom Population Census (carried out every ten years).
b) Investigations can be carried out in isolation or in conjunction with others. For example, the calculation of the official monthly Retail Price Index involves a major (ongoing) investigation which includes using the results of the Family Expenditure Survey (which is used also for other purposes). However, the information needed for first line management to control the settings of machines on a production line might depend only on sampling output at regular intervals.
c) Investigations can be regular (routine or ongoing) or 'one-off'. For example, the preparation of a company's trial balance as against a special investigation to examine the calculation of stock re-order levels.
d) Investigations are carried out on populations. A *population* is the entirety of people or items (technically known as *members*) being considered. Thus if a company wanted information on the time taken to complete jobs, the population would consist of all jobs started in the last calendar year say. Sometimes complete populations are investigated, but often only representative sections of

the population, or *samples*, are surveyed due to time, manpower and resource restrictions.

8. Stages in an investigation

However small or large an investigation is, there are certain landmarks or identifiable stages, through which it should normally pass. These are listed as follows.

a) *Definition of target population and objectives of the survey*. Who? (e.g. does the term 'workers' include temporary part-timers?) Why? (Answering this correctly will ensure that unnecessary questions are not asked and essential questions are asked.)
b) *Choice of method of data collection*. Sometimes a survey will dictate which method is used and in other cases there will be a choice. A list of the most common methods of data collection is given in the following chapter.
c) *Design of questionnaire* or the specification of other criteria for data measurement.
d) *Implementation of a pilot (or trial) survey*. A pilot survey is a small 'pre-survey' carried out in order to check the method of data collection and ensure that questions to be asked are of the right kind. Pilot surveys are normally carried out in connection with larger investigations, where considerable expenditure is involved.
e) *Selection of population members to be investigated*. If the whole (target) population is not being investigated, then a method of sampling from it must be chosen. Various sampling methods are covered fully in the following chapter.
f) *Organisation of manpower and resources to collect the data*. Depending on the size of the investigation, there are many factors to be considered. These might include: training of interviewers, transport and accommodation arrangements, organisation of local reporting bases, procedures for non-responses and limited checking of replies.
g) *Copying, collation and other organisation* of the collected data.
h) *Analyses* of data (with which much of the book is concerned).
i) *Presentation* of analyses and preparation of reports.

9. Summary

a) The subject matter of this book, Business Mathematics and Statistics, is sometimes described as Quantitative Methods, Quantitative Techniques or Numerical Analysis.
b) The subject matter attempts to alleviate the problem of incomplete information for management under the three broad headings: Statistical Method, Management Mathematics and Probability.
c) The gap between facts as they are recorded and the provision of useful information for management is bridged by Statistical Method. This covers:
    i. the selection, collection and organisation of basic facts into meaningful data, and
    ii. the summarizing, presentation and analysis of data into useful information.

d) The extent of the Management Mathematics that is covered in this book is concerned with:
   i. the understanding and evaluation of the finances involved in business investments, and
   ii. describing and evaluating physical production processes in quantitative terms.

e) Probability can be thought of as the ability to attach limits to areas of uncertainty.

f) Statistical investigations can be considered as the logical structure through which information is provided for management. They can be: trivial or major; carried out in isolation or in conjunction with other investigations; regular or 'one-off'. Investigations are carried out on populations, which can be described as the entirety of people or items under consideration.

g) The stages in an investigation could be some or all of the following, depending on their size and scope.
   i. Definition of target population and survey objectives.
   ii. Choice of method of data collection.
   iii. Design of questionnaire or the specification of other criteria for data measurement.
   iv. Implementation of a pilot survey.
   v. Selection of population members to be investigated.
   vi. Organisation of manpower and resources to collect the data.
   vii. Copying, collation and other organisation of the collected data.
   viii. Analyses of data.
   ix. Presentation of analyses and preparation of reports.

10. Student self review questions

1. What does management use Business Mathematics and Statistics for? [3]
2. What is Statistical Method and what purpose does it serve? [4]
3. Describe the two main areas covered under the heading of Management Mathematics. [5]
4. What is Probability? [6]
5. What is the particular significance of a statistical investigation to management information? [7]
6. What is meant by the term 'population'? [7]
7. List the stages of a statistical investigation. [8]

Part 1 Data and their presentation

This part of the book deals with the origins, organisation and presentation of statistical data.

Chapter 2 describes methods of selecting data items for investigation (using censuses and samples) and the various ways in which data can be collected.

Data are classified in chapter 3 and some aspects of their accuracy, including rounding, is discussed.

Chapter 4 covers various forms of frequency distributions, which are the main method of organising numerical data into a form which is convenient for either graphical presentation or analysis. Charts used to display frequency distributions include histograms and Lorenz curves.

Chapter 5 describes the many types of charts and graphs that are used to describe non-numeric data and data described over time. These include several types of bar charts, pie charts, and line diagrams.

2 Sampling and data collection

1. Introduction

This chapter is concerned with the various methods employed in choosing the subjects for an investigation and the different ways that exist for collecting data. Primary data sources (censuses and samples) are described in depth and include:

a) advantages and disadvantages in their use, and
b) data collection techniques.

Secondary data sources, mainly official publications, are covered later in the chapter.

2. Primary and secondary data

a) *Primary data* is the name given to data that are used for the specific purpose for which they were collected. They will contain no unknown quantities in respect of method of collection, accuracy of measurements or which members of the population were investigated. Sources of primary data are either censuses or samples and both of these are described in the following sections.
b) *Secondary data* is the name given to data that are being used for some purpose other than that for which they were originally collected. Summaries and analyses of such data are sometimes referred to as *secondary statistics*. The main sources of secondary data are described in later sections of the chapter.

Statistical investigations can use either primary data, secondary data or a combination of the two. An example of the latter follows. Suppose that a national company is planning to introduce a new range of products. It might refer to secondary data on rail and road transport, areas of relevant skilled labour and information on the production and distribution of similar goods from tables provided by the Government Statistical Service to site their new factory. The company might also have carried out a survey to produce their own primary data on prospective customer attitudes and the availability of distribution through wholesalers.

3. Censuses

A *census* is the name given to a survey which examines *every member* of a population.

a) A firm might take a census of all its employees to find out their opinions on the possible introduction of a new incentive scheme.
b) The Government Statistical Service carries out many official censuses. Some of them are described as follows.
    i. A *Population Census* is taken every ten years, obtaining information such as age, sex, relationship to head of household, occupation, hours of work, education, use of a car for travel to work, number of rooms in place of dwelling etc for the whole population of the United Kingdom.
    ii. A *Census of Distribution* is taken every five years, covering virtually all retail establishments and some wholesalers. It obtains information on numbers of employees, type of goods sold, turnover and classification etc.
    iii. A *Census of Production* is taken every five years, covering manufacturing industries, mines and quarries, building trades and public utility produc-

tion services. The information obtained and analysed includes distribution of labour, allocation of capital resources, stocks of raw materials and finished goods and expenditure on plant and machinery.

A census has the obvious advantages of completeness and being accepted as representative, but of course must be paid for in terms of manpower, time and resources. The three government censuses described above involve a great deal of organisation, with some staff needed permanently to answer queries on the census form, check and correct errors and omissions and extensively analyse and print the information collected. Forms can take up to a year to be returned with a further gap of up to two years before the complete results are published.

4. Samples

In practice, most of the information obtained by organisations about any population will come from examining a small, representative subset of the population. This is called a *sample*. For example:

i. a company might examine one in every twenty of their invoices for a month to determine the average amount of a customer order;
ii. a newspaper might commission a research company to ask 1000 potential voters their opinions on a forthcoming election.

The information gathered from a sample (i.e. measurements, facts and/or opinions) will normally give a good indication of the measurements, facts and/or opinions of the population from which it is drawn. The *advantages* of sampling are usually smaller costs, time and resources. A general *disadvantage* is a natural resistance by the layman in accepting the results as representative. Other disadvantages depend on the particular method of sampling used and are specified in later sections, when each sampling method is described in turn.

5. Bias

Bias can be defined as the tendency of a pattern of errors to influence data in an unrepresentative way. The errors involved in the results of investigations that have been subject to bias are known as systematic errors.

The main types of bias are now described.

a) *Selection bias*. This can occur if a sample is not truly representative of the population. Note that censuses cannot be subject to this type of bias. For example, sampling the output from a particular machine on a particular day may not adequately represent the nature and quality of the goods that customers receive. Factors that could be involved are: there may be other machines that perform better or worse; this machine might be manned by more or less experienced operators; this day's production may be under more or less pressure than another day's.

b) *Structure and wording bias*. This could be obtained from badly worded questions.

   For example, technical words might not be understood or some questions may be ambiguous.

c) *Interviewer bias*. If the subjects of an investigation are personally interviewed, the interviewer might project biased opinions or an attitude that might not gain the full cooperation of the subjects.
d) *Recording bias*. This could result from badly recorded answers or clerical errors made by an untrained workforce.

6. Sampling frames

Certain sampling methods require each member of the population under consideration to be known and identifiable. The structure which supports this identification is called a *sampling frame*. Some sampling methods require a sampling frame only as a listing of the population; other methods need certain characteristics of each member also to be known. Sampling frames can come in all shapes and sizes. For example:

i. A firm's customers can be identified from company records.
ii. Employees can be identified from personnel records.
iii. A sampling frame for the students at a college would be their enrolment forms.
iv. The relevant telephone book would form a sampling frame of people who have telephones in a certain area.
v. Stock items can be identified from an inventory file.

Note however that there are many populations that might need to be investigated for which no sampling frame exists. For example, a supermarket's customers, items coming off a production line or the potential users of a new product. Sampling techniques are often chosen on the basis of whether or not a sampling frame exists.

7. Sampling techniques

The sampling techniques most commonly used in business and commerce can be split into three categories.

a) *Random sampling*. This ensures that each and every member of the population under consideration has an equal chance of being selected as part of the sample. Two types of random sampling used are:
   i. Simple random sampling (see section 9), and
   ii. Stratified (random) sampling (see section 12).
b) *Quasi-random sampling*. (Quasi means 'almost' or 'nearly'.) This type of technique, while not satisfying the criterion given in a) above, is generally thought to be as representative as random sampling under certain conditions. It is used when random sampling is either not possible or too expensive to consider. Two types that are commonly used are:
   i. Systematic sampling (see section 13), and
   ii. Multi-stage sampling (see section 14).
c) *Non-random sampling*. This is used when neither of the above techniques are possible or practical. Two well-used types are:
   i. Cluster sampling (see section 15), and
   ii. Quota sampling (see section 16).

Before covering each of the above sampling methods in turn, it is necessary to describe some associated concepts and structures.

8. Random sampling numbers

The two types of random sampling, listed in section 7 above and described in sections 9 and 12 following, normally require the use of *random sampling numbers*. These consist of the ten digits from 0 to 9, generated in a random fashion (normally from a computer) and arranged in groups for reading convenience. The term 'generated in a random fashion' can be interpreted as 'the chance of any one digit occurring in any position in the table is no more or less than the chance of any other digit occurring'.

Appendix 2 shows a typical table of such numbers, blocked into groups of five digits. The table is used to ensure that any random sample taken from some sampling frame will be free from bias. The following section describes the circumstances under which the tables are used.

9. Simple random sampling

Simple random sampling, as described earlier, ensures that each member of the population has an equal chance of being chosen for the sample. It is necessary therefore to have a sampling frame which (at the least) lists all members of the target population. Examples of where this method might be used are:

a) by a large company, to sample 10% of their orders to determine their average value;
b) by an auditor, to sample 5% of a firm's invoices for completeness and compatibility with total yearly turnover;
c) by a professional association, to sample a proportion of its members to determine their views on a possible amalgamation with another association.

Each of these three would have obvious, ready-made sampling frames available.

It is generally accepted that the best method of drawing a simple random sample is by means of random sampling numbers. Example 1, which follows, demonstrates how the tables are used.

The *advantages* of this method of sampling include the selection of sample members being unbiased and the general acceptance by the layman that the method is fair.

Disadvantages of the method include:

i. the need for a population listing,
ii. the need for each chosen subject to be located and questioned (this can take time), and
iii. the chance that certain significant attributes of the population are under or over represented.

For example, if the fact that a worker is part-time is considered significant to a survey, a simple random sample might only include 25% part-time workers from a population having, say, a 30% part-time work force.

10. Example 1 (Use of *random sampling numbers*)

An auditor wishes to sample 29 invoices out of a total of 583 received in a financial year. The procedure that could be followed is listed below.

1. Each invoice would be numbered, from 001 through to 583.
2. Select a starting row or column from a table of random sampling numbers and begin reading groups of three digits sequentially. For example, using the random sampling numbers at Appendix 2, start at row 6 (beginning 34819 80011 17751 03275 ...etc). This gives the groups of three as: 348 198 001 117 751 032 ... etc.
3. Each group of three digits represents the choice of a numbered invoice for inclusion in the sample. Any number that is greater than 583 is ignored as is any repeat of a number. Using the illustration from 2. above, invoice numbers 348, 198, 001, 117, 032, etc would be accepted as part of the sample, while number 751 would be rejected as too large.
4. As many groups of three digits as necessary are considered until 29 invoices have been identified. This forms the required sample.

Notes:

a) Random sampling numbers can be generated by a computer or pre-printed tables can be obtained.

b) The number of digits to be read in groups will always depend upon how many members there are in the population. If there were 56,243 members, then digits would need to be read in fives; groups of four digits would be read if a population being sampled had 8771 members.

c) The choice of a starting row or column for reading groups of digits should be selected randomly.

11. Stratification of a population

Stratification of a population is a process which:

i. identifies certain attributes (or strata levels) that are considered significant to the investigation at hand;

ii. partitions the population accordingly into groups which each have a unique combination of these levels.

For example, if whether or not heavy goods vehicles had a particular safety feature was thought important to an investigation, the population would be partitioned into the two groups 'vehicles with the feature' and 'vehicles without the feature'. On the other hand, if whether an employee was employed full or part-time, together with their sex, was felt to be significant to their attitudes to possible changes in working routines, the population would be partitioned into the four groups: male/full-time; female/full-time; male/part-time and female/part-time.

Populations that are stratified in this way are sometimes referred to as *heterogeneous*, meaning that they are composed of diverse elements or attributes that are considered significant.

12. Stratified sampling

Stratified random sampling extends the idea of simple random sampling to ensure that a heterogeneous population has its defined strata levels taken account of in the sample. For example, if 10% of all heavy goods vehicles have a certain safety feature, and this is considered significant to the investigation in hand, then 10% of a sample of such vehicles must have the safety feature.

The general procedure for taking a stratified sample is:

a) Stratify the population, defining a number of separate partitions.
b) Calculate the proportion of the population lying in each partition.
c) Split the total sample size up into the above proportions.
d) Take a separate sample (normally simple random) from each partition, using the sample sizes as defined in (c).
e) Combine the results to obtain the required stratified sample.

Stratification of a population can be as simple or complicated as the situation demands. Some surveys might warrant that a population be split into many strata. A major investigation into car safety could identify the following significant factors having some bearing on safety: saloon and estate cars; radial and cross-ply tyres; two and four-door models; rear passenger safety belts (or not). The sampling frame in this case would have to be split into sixteen separate partitions in order to take account of all the combinations possible from (i) to (iv) above (for example, saloon/radial/2-door/belts and saloon/radial/2-door/no belts are just two of the partitions).

Advantages of this method of sampling include the fact that the sample itself (as well as the method of selection) is free from bias, since it takes into account significant strata levels (attributes) of a population considered important to the investigation.

Disadvantages of stratified sampling include:

i. an extensive sampling frame is necessary;
ii. strata levels of importance can only be selected subjectively;
iii. increased costs due to the extra time and manpower necessary for the organisation and implementation of the sample.

13. Systematic sampling

Systematic sampling is a method of sampling that can be used where the population is listed (such as invoice values or the fleet of company vehicles) or some of it is physically in evidence (such as a row of houses, items coming off a production line or customers leaving a supermarket). The technique is to choose a random starting place and then systematically sample every 40th (or 12th or 165th) item in the population, the number (40, say) having been chosen based on the size of sample required. For example, if a 2% sample was needed from a population, every 50th item would be selected, after having started at some random point.

This is because 2% = 2 in 100 = 1 in 50.

Systematic sampling is particularly useful for populations that (with respect to the investigation to hand) are of the same kind or are uniform. These are referred to as *homogeneous* populations. For example, the invoices of a company for one financial

year would be considered as a homogeneous population by an auditor, if their value or relationship to type of goods ordered was of no consequence to the investigation. Thus, a systematic sample could be used.

Care must be taken however, when using this method of sampling, that no set of items in the population recur at set intervals. For example, if four machines are producing identical products at the same rate and these are being passed to a single conveyer, it could happen that the products form natural sets of four (one from each machine). A systematic sample, examining every n-th item (where n is a factor of 4), might well be selecting products from the same machine and therefore be biased.

Advantages of this method include:

i. ease of use;
ii. the fact that it can be used where no sampling frame exists (but items are physically in evidence).

The main *disadvantage* of systematic sampling is that bias can occur if recurring sets in the population are possible.

This method of sampling is not truly random, since (once a random starting point has been selected) all subjects are pre-determined. Hence the use of the term 'quasi-random' to describe the technique.

14. Multi-stage sampling

Where a population is spread over a relatively wide geographical area, random sampling will almost certainly entail travelling to all parts of the area and thus could be prohibitively expensive. *Multi-stage sampling*, which is intended to overcome this particular problem, involves the following.

a) Splitting the area up into a number of regions;
b) Randomly selecting a small number of the regions;
c) Confining sub-samples to these regions alone, with the size of each sub-sample proportional to the size of the area. For example, the United Kingdom could be split up into counties or a large city could be split up into postal districts;
d) The above procedure can be repeated for sub-regions within regions... and so on.

Once the final regions (or sub-regions etc) have been selected, the final sampling technique could be (simple or stratified) random or systematic, depending on the existence or otherwise of a sampling frame.

The main *advantage* of this method is that less time and manpower is needed and thus it is cheaper than random sampling.

Disadvantages of multi-stage sampling include:

i. possible bias if a very small number of regions is selected;
ii. the method is not truly random, since, once particular regions for sampling have been selected, no member of the population in any other region can be selected.

15. Cluster sampling

Cluster sampling is a non-random sampling method which can be employed where no sampling frame exists, and, often, for a population which is distributed over

some geographical area. The technique involves selecting one or more geographical areas and sampling *all* the members of the target population that can be identified.

For example, suppose a survey was needed of companies in South Wales who use a computerized payroll. First, three or four small areas would be chosen (perhaps two of these based in city centres and one or two more in outlying areas). Each company, in each area, might then be phoned, to identify which of them have computerized systems. The survey itself could then be carried out.

The *advantages* of cluster sampling include:

i it is a good alternative to multi-stage sampling where no sampling frame exists;

ii. it is generally cheaper than other methods since little organisation or structure is needed in the selection of subjects.

The main *disadvantage* of the method is the fact that sampling is not random and thus selection bias could be significant. (Non-response is not normally considered to be a particular problem.)

16. Quota sampling

A sampling technique much favoured in market research is *quota sampling*. The method uses a team of interviewers, each with a set number (quota) of subjects to interview. Normally the population is stratified in some way and the interviewer's quota will reflect this. This method places a lot of responsibility onto interviewers since the selection of subjects (and there could be many strata involved) is left to them entirely. Ideally they should be well trained and have a responsible, professional attitude.

The *advantages* of quota sampling include:

i. stratification of the population is usual (although not essential);

ii. no non-response;

iii. low cost and convenience.

The main *disadvantages* of this method are:

i. sampling is non-random and thus selection bias could be significant;

ii. severe interviewer bias can be introduced into the survey by inexperienced or untrained interviewers, since all the data collection and recording rests with them.

17. Precision

Clearly the best way of obtaining information about a population is to take a census. This will ensure (barring any bias and clerical errors) that the information obtained about the population is accurate. However, sampling is a fact of life and the information about a population that is derived from a sample will inevitably be imprecise. The error involved is sometimes known as sampling error. One technique that is often used to compensate for this is to state limits of error for any sample statistics produced. Particular precision techniques are just outside the scope of this book.

18. Sample size

There is no universal formula for calculating the size of a sample. However, as a starting point, there are two facts that are well known from statistical theory and should be remembered.

1. The larger the size of sample, the more precise will be the information given about the population.
2. Above a certain size, little extra information is given by increasing the size.

All that can be deduced from the above two statements, together with some other points made in earlier sections of the chapter, is that a sample need only be large enough to be reasonably representative of the population. Some general factors involved in determining sample size are listed below.

a) *Money and time available*.

b) *Aims of the survey*. For example, for a quick market research exercise, a very small sample (perhaps just 50 or 100 subjects) might suffice. However if the opinions of the workforce were desired on a major change of working structures, a 20 or 30% sample might be in order.

c) *Degree of precision required*. The less precise the results need to be, the smaller the sample size.

For example, to gauge an approximate market reaction to one of their new products, a firm would only need a very small sample. On the other hand, if motor vehicles were being sampled for exhaustive safety tests at a final production stage, the sample would need to be relatively large.

d) *Number of sub-samples required*. When a stratified sample needs to be taken and many sub-samples are defined, it might be necessary to take a relatively large total sample in order that some smaller groups contain significant numbers.

For example, suppose that a small sub-group accounted for only 0.1% of the population. A total sample size as large as 10,000 would result in a sample size of only 10 (0.1%) for this sub-group, which would probably not be large enough to gain any meaningful information.

19. Methods of primary data collection

Data collection can be thought of as the means by which information is obtained from the selected subjects of an investigation. There are various data collection methods which can be employed. Sometimes a sampling technique will dictate which method is used and in other cases there will be a choice, depending on how much time and manpower (and inevitably money) is available. The following list gives the most common methods.

a) *Individual (personal) interview*.

This method is probably the most expensive, but has the advantage of completeness and accuracy. Normally questionnaires will be used (described in more detail in the following section).

Other factors involved are:

i. interviewers need to be trained;

ii. interviews need arranging;

iii. can be used to advantage for pilot surveys, since questions can be thoroughly tested;
iv. uniformity of approach if only one interviewer is used;
v. an interviewer can see or sense if a question has not been fully understood and it can be followed-up on the spot.

This form of data collection can be used in conjunction with random or quasi-random sampling.

b) *Postal questionnaire*.

This is a much cheaper method than the personal interview since manpower (one of the most expensive resources) is not used in the data collection. However, much more effort needs to be put into the design of the questionnaire, since there is often no way of telling whether or not a respondent has understood the questions or has answered them correctly (both of these are generally no problem in a personal interview).

Other factors involved are:

i. low response rates (although inducements, such as free gifts, often help);
ii. convenience and cheapness of the method when the population is scattered geographically;
iii. no prior arrangements necessary (unlike the personal interview);
iv. questionnaires sent to a company may not be filled in by the correct person.

This method can be used in conjunction with most forms of sampling.

c) *Street (informal) interview*.

This method of data collection is normally used in conjunction with quota sampling, where the interviewer is often just one of a team. Some factors involved are:

i. possible differences in interviewer approach to the respondents and the way replies are recorded;
ii. questions must be short and simple;
iii. non-response is not a problem normally, since refusals are ignored and another subject selected;
iv. convenient and cheap.

d) *Telephone interview*.

This method is sometimes used in conjunction with a systematic sample (from the telephone book). It would generally be used within a local area and is often connected with selling a product or a service (for example, insurance). It has an in-built bias if private homes are being telephoned (rather than businesses), since only those people with telephones can be contacted and interviewed. It can cause aggravation and the interviewer needs to be very skilled.

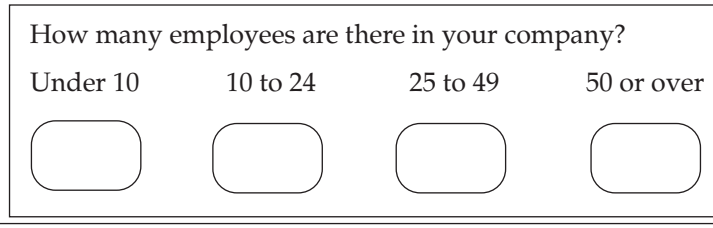
e) *Direct observation*.

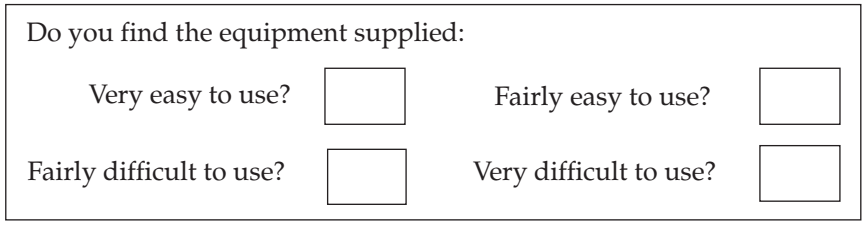
This method can be used for examining items sampled from a production line, in traffic surveys or in work study. It is normally considered to be the most accurate form of data collection, but is very labour-intensive and cannot be used in many situations.

20. Questionnaire design

If a questionnaire is used in a statistical survey, its design requires careful consideration. A badly designed questionnaire can cause many administrative problems and may cause incorrect deductions to be made from statistical analyses of the results. One of the major reasons why pilot surveys are carried out is to check typical responses to questions. Some important factors in the design of questionnaires are given below.

a) The questionnaire should be as short as possible.
b) Questions should:
   i. be simple and unambiguous.
   ii. not be technical.
   iii. not involve calculations or tests of memory.
   iv. not be personal, offensive or leading.
c) As many questions as possible should have simple answer categories (so that the respondent has only to choose one). For example:





d) Questions should be asked in a logical order.

A useful check on the adequacy of the design of a questionnaire can be given by conducting a *pilot survey*.

21. The use of secondary data

Secondary data are generally used when:

a) the time, manpower and resources necessary for your own survey are not available (and, of course, the relevant secondary data exists in a usable form), or
b) it already exists and provides most, if not all, of the information required.

The *advantages* of using secondary data are savings in time, manpower and resources in sampling and data collection. In other words, somebody else has done the 'spade work' already.

The *disadvantages* of using secondary data can be formidable and careful examination of the source(s) of the data is essential. Problems include the following.

i. Data quality might be questionable. For example, the sample(s) used may have been too small, interviewers may not have been experienced or any questionnaires used may have been badly designed.
ii. The data collected might now be out-of-date.
iii. Geographical coverage of the survey may not coincide with what you require. For example, you might require information for Liverpool and the secondary data coverage is for the whole of Merseyside.
iv. The strata of the population covered may not be appropriate for your purposes. For example, the secondary data might be split up into male/female and full-time/part-time workers and you might consider that, for your purposes, whether part-time workers are permanent or temporary is significant.
v. Some terms used might have different meanings. Common examples of this are:
    - Wages (basic only or do they include overtime?)
    - Level of production (are rejects included?)
    - Workers (factory floor only or are office staff included?)

22. Sources of secondary data and their use

Secondary data sources fall broadly into two categories: those that are *internal* and those *external* to the organisation conducting the survey.

Some examples of *internal* secondary data sources and uses are:

a) a customer order file, originally intended for standard accounting purposes, could have its addresses and typical goods amounts used for route planning purposes;
b) using information on raw material type and price (originally collected by the purchase department to compare manufacturers) for stock control purposes;
c) information on job times and skills breakdown, originally compiled for job costing, used for organising new pay structures.

Some examples of *external* secondary data sources are:

d) the results of a survey undertaken by a credit card company, to analyse the salary and occupation of its customers, might be used by a mail order firm for advertising purposes;
e) a commercially produced car survey giving popularity ratings and buying intentions, might be used by a garage chain to estimate stock levels of various models.

Without doubt, the most important external secondary data sources are official statistics supplied by the Central Statistical Office and other government departments. These are listed and briefly described in the next section.

23. Official secondary data sources

The following list gives the major publications of the Central Statistical Office.

a) *Annual Abstract of Statistics*. This publication is regarded as the main general reference book for the United Kingdom and has been published for nearly 150 years. Its tables cover just about every aspect of economic, social and industrial life. For example: climate; population; social services; justice and crime;

education; defence; manufacturing and agricultural production; transport and communications; finance.

b) *Monthly Digest of Statistics*. A monthly abbreviated version of the Annual Abstract of Statistics. Gives the facts on such topics as: population; employment; prices; wages; social services; production and output; energy; engineering; construction; transport; retailing; finance and the weather. It has runs on monthly and quarterly figures for at least two years in most tables and annual figures for longer periods. An annual supplement gives definitions and explanatory notes for each section. An index of sources is included.

c) *Financial Statistics*. A monthly publication bringing together the key financial and monetary statistics of the United Kingdom. It is the major reference document for people and organisations concerned with government and company finance and financial markets generally. It usually contains at least 18 monthly, 12 quarterly or 5 annual figures on a wide variety of topics. These include: financial accounts for sectors of the economy; Government income and expenditure; public sector borrowing; banking statistics; money supply and domestic credit expansion; institutional investment; company finance and liquidity; exchange and interest rates. An annual explanatory handbook contains notes and definitions.

d) *Economic Trends*. Published monthly, this is a compilation of all the main economic indicators, illustrated with charts and diagrams. The first section (Latest Developments) presents the most up-to-date statistical information available during the month, together with a calendar of recent economic events. The central section shows the movements of the key economic indicators over the last five years or so. Finally there is a chart showing the movements of four composite indices over 20 years against a reference chronology of business cycles. In addition, quarterly articles on the national accounts appear in the January, April, July and October issues, and on the balance of payments in the March, June, September and December issues. Occasional articles comment on and analyse economic statistics and introduce new series, new analyses and new methodology. *Economic Trends* also publishes the release dates of forthcoming important statistics. An annual supplement gives a source for very long runs, up to 35 years in some cases, of key economic indicators. The longer runs are annual figures, but quarterly figures for up to 25 years or more are provided.

e) *Regional Trends*. An annual publication, with many tables, maps and charts, it presents a wide range of government statistics on the various regions of the United Kingdom. The data covers many social, demographic and economic topics. These include: population, housing, health, law enforcement, education and employment, to show how the regions of the United Kingdom are developing and changing.

f) *United Kingdom National Accounts (The Blue Book)*. Published annually, this is the essential data source for those concerned with macro-economic policies and studies. The principal publication for national accounts statistics, it provides detailed estimates of national product, income and expenditure. It covers industry, input and output, the personal sector, companies, public corporations,

central and local government, capital formation and national accounts. Tables of statistical information, generally extending over eleven years, are supported by definitions and detailed notes. It is a valuable indicator of how the nation makes and spends its money.

g) *United Kingdom Balance of Payments (The Pink Book)*. This annual publication is the basic reference book of balance of payments statistics, presenting all the statistical information (both current and for the preceding ten years) needed by those who seek to assess United Kingdom trends in relation to those of the rest of the world.

h) *Social Trends*. One of the most popular and colourful annual publications, it has (for over ten years) provided an insight into the changing patterns of life in Britain. The chapters provide accurate analyses and breakdowns of statistical information on population, households and families, education and employment, income and wealth, resources and expenditure, health and social services and many other aspects of British life and work.

i) *Guide to Official Statistics*. A periodically produced reference book for all users of statistics. It indicates what statistics have been compiled for a wide range of commodities, services, occupations etc, and where they have been published. Some 1000 topics are covered and about 2500 sources identified with an index for easy use. It covers all official and significant non-official sources published during the last ten years.

The following publication is compiled by the Department of Employment.

j) *Employment Gazette*. Published monthly, it is a summary of statistics on: employment, unemployment, numbers of vacancies, overtime and short time, wage rates, retail prices, stoppages. Each publication includes one or more 'in-depth' article and details of arbitration awards, notices, orders and statutory instruments.

The following publication is compiled by the Department of Industry.

k) *British Business*. Published weekly, the main topics are production, prices and trade. It includes information on: the Census of Production, industrial materials, manufactured goods, distribution, retail and service establishments, external trade, prices, passenger movements, hire purchase, entertainment.

Other important business publications include: HSBC Holdings plc Annual Review, NatWest Bank Quarterly Review, Lloyds TSB Annual Report, Barclays Review (quarterly), International Review (Barclays, quarterly), Three Banks Review (quarterly), Journal of the Institute of Bankers (bi-monthly), Financial Times (daily), The Economist (weekly) and The Banker (monthly).

24. Summary

a) Data that are used for the specific purpose for which they are collected are called primary data. Secondary data is the name given to data that are being used for some purpose other than that for which they were originally collected.

b) A census is a survey which examines every member of the population. Three important official censuses are the Population Census, the Census of Distribution and the Census of Production.

c) A sample is a relatively small subset of a population with advantages over a census that costs, time and resources are much less. The main disadvantage is that of acceptability by the layman.
d) Bias is the tendency of a pattern of errors to influence data in an unrepresentative way. Bias can be due to selection procedures, structure and wording of questions, interviewers or recording.
e) A sampling frame is a structure which lists or identifies the members of a population.
f) Random sampling numbers are tables of randomly generated digits, used to ensure that the selection of the members of a sample is free from bias.
g) Simple random sampling is a technique which ensures that each and every member of a population has an equal chance of being chosen for the sample.
h) Stratified random sampling ensures that every significant group in the population is represented in proportion in the sample using a stratification process. An extensive sampling frame is needed with this method.
i) Systematic (quasi-random) sampling involves selecting a random starting point and then sampling every n-th member of the population. The value of n is chosen based on the size of sample required. It can be biased if certain recurring cycles exist in the population, but can often be used where no sampling frame exists.
j) Multi-stage (quasi-random) sampling is normally used with homogeneous populations spread over a wide area. It involves splitting the area up into regions, selecting a few regions randomly and confining sampling to these regions alone. It is cheaper than random sampling.
k) Cluster (non-random) sampling involves exhaustive sampling from a few well chosen areas. It is a cheap method, useful for populations spread over a wide geographical area where no sampling frame exists.
l) Quota (non-random) sampling normally involves teams of interviewers who obtain information from a set quota of people, the quota being based on some stratification of the population. It is commonly used in market research.
m) The precision of some statistic obtained from a sample can be measured by describing the limits of error with a given degree of confidence.
n) Some factors involved in determining the size of a sample are: money and time available, survey aims, degree of precision or number of sub-samples required. Generally, the larger the sample the better, but small samples can give relatively accurate information about a population.
o) Main methods of primary data collection are:
    i. Individual (personal) interview.
    ii. Postal questionnaire.
    iii. Street (informal) interview.
    iv. Telephone interview.
    v. Direct observation.
p) The main points in questionnaire design are: questionnaire to be as short as possible; questions to be simple, non-ambiguous, non-technical, not to be

personal or offensive and not to involve calculations or tests of memory; answer categories to be given where possible; questions asked in a logical order.

q) Secondary data can be used where the facilities for your own survey are not available or where the secondary data gives all the information you require. Disadvantages are: data might not be of an acceptable quality or out-of-date; geographical or strata coverage may not be appropriate; there may be differences in the meaning of terms.

r) Some of the main sources of external secondary data are contained in the following official publications:

Annual Abstract of Statistics; Monthly Digest of Statistics;
Financial Statistics; Economic Trends;
Regional Trends;
United Kingdom National Accounts (Blue Book);
United Kingdom Balance of Payments (Pink Book);
Social Trends; Employment Gazette;
British Business.

25. Student self review questions

1. Explain the difference between primary and secondary data. [2]
2. Give the meaning of a census and give some examples of official censuses. [3]
3. What are the major factors involved when deciding between a sample and a census? [3,4]
4. Describe what bias is and give some examples of how it can arise. [5]
5. Give at least four examples of a sampling frame. [6]
6. What is a random sample? [7]
7. What is quasi-random sampling and under what conditions might it be used? [7]
8. What are random sampling numbers and how are they used in simple random sampling? [8,10]
9. What does the term 'stratification of a population' mean and how is it connected with stratified sampling? [11,12]
10. What are the advantages and disadvantages of stratifed sampling when compared with simple random sampling? [12]
11. What is the difference between homogeneous and heterogeneous populations? [11,13]
12. Give an example of a situation where a systematic sample could be taken:
    a) where a sampling frame exists;
    b) where no sampling frame exists. [13]
13. What are the differences between multi-stage and cluster sampling methods? [14,15]
14. In what type of situation is quota sampling most commonly used and what are its main merits? [16]
15. How can the precision of a sample estimate be expressed? [17]
16. What are the factors involved in determining the size of a sample? [18]
17. List the main methods of collecting primary data [19]

18. What are the advantages and disadvantages of a postal questionnaire over a personal interview? [19]
19. Give some important considerations in the design of a questionnaire. [20]
20. Under what conditions might secondary data be used and what are its possible disadvantages compared with the use of primary data? [21]
21. Name some of the major official statistical publications. [22]

26. Student exercises

1. *MULTI-CHOICE*. Which one of the following is NOT a type of random sampling technique:
   a) Quota sampling
   b) Systematic sampling
   c) Stratified sampling
   d) Multi-stage sampling
2. *MULTI-CHOICE*. A 2% random sample of mail-order customers, each with a numeric serial number, is to be selected. A random number between 00 and 49 is chosen and turns out to be 14. Then, customers with serial numbers 14, 64, 114, 164, 214, ... etc are chosen as the sample. This type of sampling is:
   a) simple random b) stratified c) quota d) systematic.
3. A large company is considering a complete reshaping of its pay structures for production workers. What data might be collected and analysed, other than technical details, to help the management come to a decision? Consider both primary and secondary sources.
4. What factors would govern the use of a sample enquiry rather than a census if information was required about shopping facilities throughout a large city.
5. *MULTI-CHOICE*. A sample of 5% of the employees working for a large national company is required. Which one of the following methods would provide the best simple random sample?
   a) Wait in the car park in a randomly selected branch and select every tenth employee driving in to work .
   b) Use random number tables to select 1 in 20 of the branches and then select all the employees.
   c) Select a branch randomly and use personnel records to choose 1 in 20 randomly.
   d) Select 5% of all employees from personnel records at head office randomly.
6. Suggest an appropriate method of sampling that could be employed to obtain information on:
   a) passengers' views on the adequacy of a local bus service;
   b) the attitudes to authority of the workforce of a large company;
   c) the percentage of defects in finished items from a production line;
   d) the views of Welsh car drivers on the wearing seat belts;
   e) the views of schoolchildren on school meals.
7. A national survey has revealed that 40% of non-manual workers travel to work by public transport while one-half use their own transport. For all workers, 47.5% use public transport and one in every ten use methods other than their own or public transport. A statistical worker in a large factory (which is known to have about

three times as many manual workers as non-manual workers) has been asked to sample 200 employees for their views on factory-provided transport. He decides to take a quota sample at factory gate B at five o'clock one evening.

a) How many manual workers will there be in the sample?
b) How many workers who travel to work by public transport will be interviewed?
c) Calculate the quota to be interviewed in each of the six sub-groups defined.
d) Point out the limitations of the sampling technique involved and suggest a better way of collecting the data.

8. The makers of a brand of cat food 'Purrkins' wish to obtain information on the opinions of their customers and include a short questionnaire on the inside of the label as follows:

    1. Do you like Purrkins?
    2. Why do you buy Purrkins?
    3. Have you tried our dog food?
    4. What amount of Purrkins do you normally buy?
    5. When did you start using Purrkins?
    6. What type of house do you live in?

    Criticise the questions.

9. Design a short questionnaire to be posted to a sample of customers to obtain their views on your company's delivery service.

10. A proposal was received by the Local Authority Planning Office for a motel, public house and restaurant to be built on some private land in the city suburbs. Following an article by the builder in the local paper, the office received 300 letters of which only 28 supported the proposal. What conclusions can the Planning Officer draw from these statistics? Describe what action could be taken to gauge people's views further.

3 Data and their accuracy

1. Introduction

This chapter is concerned with the forms that data can take, how data are measured and the errors and approximations that are often made in their description.

2. Data classification

In order to present and analyse data in a logical and meaningful way, it is necessary to understand some of the natural forms that they can take. There are various ways of classifying data and these are now listed.

a) *By source*. Data can be described as either primary or secondary, depending on their source. This area has already been covered in the previous chapter.

b) *By measurement*. Data can be measured in either numeric (or quantitative) or non-numeric (qualitative) terms. This might sound very obvious, but the difference is important since the forms of both presentation and analysis differ markedly in these two cases. Presentation of numeric data is covered in chapter 4 and non-numeric data in chapter 5. For the business and accounting courses that this book covers, only numeric data is analysed and this is done from chapter 7 onwards.

c) *By preciseness*. Data can either be measured precisely (described as *discrete*) or only ever be approximated to (described as *continuous*). The differences between the two are described more fully in sections 3 and 4 in this chapter.

d) *By number of variables*. Data can consist of measurements of one or more variables for each subject or item. *Univariate* is the name given to a set of data consisting of measurements of just one variable, *bivariate* is used for two variables, and for two or more variables the data is described as *multivariate*. Some examples of these different types are given in example 1.

3. Discrete data

Discrete data can be described as data that can be measured precisely. One way of obtaining discrete data is by counting. For example:

i. the number of components produced from an assembly line over a number of consecutive shifts:

45, 51, 44, 44, 43, 50, 46, 43, ... etc;

ii. the number of employees working in various offices of a company:

12, 32, 8, 13, 8, 6, 11, 24, ... etc.

Discrete data can also be obtained from situations where counting is not involved. For example:

iii. shoe sizes of a sample of people:

8, 10, 10, $6\frac{1}{2}$, 9, 9, $9\frac{1}{2}$, $8\frac{1}{2}$ ... etc;

iv. weekly wages (in £) for a set of workers:

121.45, 162.85, 133.37, 108.32, ... etc.

A particular characteristic of discrete data is the fact that possible data values progress in definite steps, i.e. shoe sizes are measured as 6 or $6\frac{1}{2}$ or 7 or $7\frac{1}{2}$ ... etc or there are 1 or 2 or 3 ... etc people (and not 3.5 or 4.67).

4. Continuous data

The most significant characteristic of continuous data is the fact that they cannot be measured precisely; their values can only be approximated to. Examples of continuous data are dimensions (lengths, heights); weights; areas and volumes; temperatures; times.

How well continuous values are approximated to depends on the situation and the quality of the measuring instrument. It might be adequate to measure peoples' heights to the nearest inch, whereas spark plug end gaps would need to be measured to perhaps the nearest tenth of a millimetre. Time card punching machines only record times in hours and minutes while sophisticated process control computers, dealing with volatile chemicals, would need to measure both time and temperature very finely.

Although continuous values cannot be identified exactly, they are often recorded as if they were precise and this is normally acceptable. For example:

i. clock-in times of the workers on a particular shift:
   8:23, 8:28, 8:28, 8:32, 8:28, 8:26, ... etc;
ii. diameters (in mm) of a sample of screws from a production run:
   4.11, 4.10, 4.10, 4.10, 4.15, 4.09, 4.12, ... etc;
iii. weights (in gm) of the contents of a selection of cans of baked beans:
   446.8, 447.0, 446.8, 447.2, 447.0, 447.1, ... etc.

5. Example 1 (Demonstrations of various *classifications* of data)

a) Table 1 shows the ages and annual salaries of a sample of qualified certified accountants. Since each member of the sample is being measured in terms of two variables, age and salary, this is an example of bivariate data. Both variables are numeric with salary discrete (since each is an exact value) and age continuous (since age is really a particular type of time measurement) and approximated to years only.

Table 1 Age and salary for a sample of certified accountants

Sample member	1	2	3	4	5	6	7	8	9
Age	32	30	25	28	25	30	49	26	56
Salary	8800	11900	6000	8200	5800	12500	9650	7200	16450

b) Number of defects found in samples of 100 items taken by a quality control section from batches of finished products.

Batch	1	2	3	4	5	6	7	8	9	10
Number of defectives	2	0	2	5	1	3	0	0	1	0

The data are being described in terms of one variable (number of defectives in a sample of 100) and thus are univariate. They are also discrete, since the values have been obtained by counting, and numeric.

c) Policies handled by an agent for a particular insurance company:

Policy	Type	Annual premium (£)
A	Motor (3rd party)	86
B	Motor (3rd party)	124
C	Life	185
D	Motor (comprehensive)	120
E	Disability	24
F	House contents	49
G	Motor (comprehensive)	252

The data given for the various policies is described in terms of two variables, type of policy and annual premium, and thus is bivariate. Type of policy is non-numeric, annual premium is numeric and both variables are discrete.

6. Rounding and its conventions

a) Data are normally rounded for one of two reasons.
   i. If they are continuous, rounding is the only way to give single values which will represent the magnitude of the data.
   ii. If they are discrete, the values given may be too detailed to use as they stand. For example, the annual profits of a plc might have been calculated precisely as £14,286,453.88, but could be quoted on the stock exchange as £14 million.

b) As should already be familiar, *fair rounding* is the technique of cutting off particular digits from a given numeric value and, depending on whether the first digit discarded has value 5 or more (or not), adding 1 to the last of the remaining digits or not (known as rounding up or down).

   There are two conventions used for displaying the results of fair rounding.

   i. *By decimal place*. This is the most common form of rounding. For example, if the price of a car was given as £4684.45, it could be rounded as:

      £4684.5 (to 1 decimal place or 1D)

      £4684 (to the nearest whole number or n1)

      £4680 (to the nearest 10 or n10) - note the final zero

      £4700 (to the nearest 100 or n100)

      £5000 (to the nearest 1000 or n1000)

   ii. By *number of digits*. This convention is sometimes used as an alternative to decimal place rounding. For example, if a company's profit for the past financial year was £682,056.39, it could be rounded as:

      £682,056.4 (to 7 significant digits or 7S)

      £682,060 (to 5 significant digits or 5S)

      £682,000 (3S) and, finally, £700,000 (1S)

7. Errors and their causes

Any data, whatever their source (international, national, company or personal), can be subject to errors. The causes of errors are numerous, but some of the more important of them can be classified into two main groups.

a) *Unpredictable errors*. These are errors that occur due to:
   i. Incomplete or incorrect records.
   ii. Ambiguous or over-complicated questions asked as part of questionnaires.
   iii. Data being obtained from samples.
   iv. Mistakes in copying data from one form to another.
   All that can be done to minimise this type of error is to ensure that: investigation procedures are carried out in a professional, logical and consistent way; questionnaires are well designed and tested; samples are as representative as possible; and so on.

b) *Planned (predictable) errors*. These are errors that were referred to in section 6 (a) and are due to:
   i. Measuring continuous data.
   ii. Rounding discrete data for the purposes of overall perspective.
   This is the type of error that can be taken account of and is discussed in the rest of the chapter.

8. Maximum errors

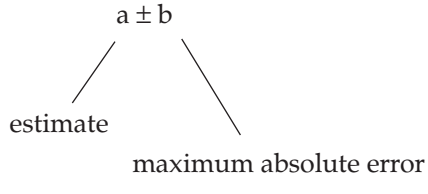
If a number is given and known to be subject to an error, then the error must be unknown (otherwise there would be no need to consider it). However, what can be determined often is the largest value that the error could possibly take. This is known as the *maximum error*. From now on, unless otherwise stated, any error referred to will be assumed to be a maximum error.

9. Methods of describing errors

A number that is subject to some unknown error is sometimes called an *approximate number*. Approximate numbers can be written in three different forms, described as follows.

a) *An interval*. This takes the form of a range of values within which the true number being represented lies. The range is normally given as a low and high value, separated by a comma and written within square brackets. For example: [19.5,20.5]. This would mean that the true value of the number being represented lies between 19.5 and 20.5.

b) An *estimate with a maximum absolute error.* In this form the real value of the number being represented is given as an estimate, together with the maximum error. It takes the form:



For example: 20 ± 0.5. This can be translated as 'the true value of the number being represented lies within 0.5 either side of 20'. Of course this is exactly equivalent to the example given in a) above.

c) *An estimate with a maximum relative error*. This is a slight adaptation of b) above, in that the maximum error is expressed in relative (i.e. proportional or percentage) terms. That is:

$$\text{maximum relative error} = \frac{\text{maximum absolute error}}{\text{estimate}} \times 100\%$$

For example: 20 ± 2.5%. This can be translated as 'the true value of the number being represented lies within 2.5% either side of 20'. Since 2.5% of 20 is just 0.5, it should now be realised that the examples of the three forms of expressing an approximate number shown above are equivalent.

10. Example 2 (Expressing a number *subject to error* in different forms)

If the annual output of a mine is given as 15 million tonnes, subject to a maximum error of 1.5 million tonnes, the real value of the output can be written as:

15 million tonnes ± 1.5 million tonnes

or: $15 \pm \frac{5}{236} \times 100\% = 15 \text{ million tonnes} \pm 10\%$

or: [15–1.5,15+1.5] million tonnes = [13.5,16.5] million tonnes.

11. Rounding errors

As soon as a numeric value is subjected to fair rounding, an error is introduced and an approximate number is thus defined. As an example, suppose the length of a bolt is measured as 14 mm to the nearest mm. This means that the true length of the component must lie between the values 13.5 and 14.5 mm (because any value that lies between these two would have been rounded to 14 mm). Hence, the maximum error can be expressed as ± 0.5 mm.

In fact, any value that is rounded to the nearest whole number will have a maximum error of ± 0.5. Similarly, any value that is rounded to 1D will have a maximum error of ± 0.05. Any value rounded to the nearest 10 will have a maximum error of ± 5. This pattern is tabulated in Table 2.

Table 2 Maximum errors in fair rounded numbers

Degree of rounding	Maximum error
...	...
...	...
3D	± 0.0005
2D	± 0.005
1D	± 0.05
nearest whole number (n1)	± 0.5
nearest 10 (n10)	± 5
nearest 100 (n100)	± 50
nearest 1000 (n1000)	± 500
...	...
...	...

Thus, for example:

i. 64.23 (2D) = 64.23 ± 0.005 = [64.225,64.235];
ii. 12 (n1) = 12 ± 0.5 = [11.5,12.5];
iii. 78400 (n100) = 78400 ± 50 = [78350,78450];
iv. 24.6 (3S) = 24.6 (1D) = 24.6 ± 0.05 = [24.55,24.65];
v. 1530 (3S) = 1530 (n10) = 1530 ± 5 = [1525,1535].

12. Errors in expressions

An *arithmetic expression* is any combination of numbers and arithmetic operations (+, -, x and ÷).

For example: 12+6, (42–6)×3 and $\frac{72 - 12 \times 4}{4 + 8}$ are examples of arithmetic expressions.

If all the numbers involved in an expression are exact, then the expression will be able to be evaluated exactly. If, however, at least one of the numbers involved in an expression is subject to an error, then the value of the expression itself will be subject to error.

As an example of identifying the range of error involved in the value of an expression, suppose the first of the above expressions, 12+6, is to be evaluated, where both 12 and 6 have been *rounded to the nearest whole number*.

Now, 12 can be represented as [11.5, 12.5] (where 11.5 is the least value and 12.5 is the greatest value).

Similarly, 6 can be represented as [5.5, 6.5] (where 5.5 and 6.5 are the least and greatest values respectively).

So that: the *least* value of 12+6 = least value of 12 + least value of 6
= 11.5 + 5.5 = 17

Similarly, the *greatest* value of 12 + 6 is 12.5 + 6.5 = 19

Therefore the range of possible values for 12 + 6 is [17,19].

Finding the range of values for the multiplication of two numbers (each subject to an error) follows a similar line, but the rule for subtraction and division needs special consideration. The following section gives the rules for these four basic operations.

13. Error rules

The range of error for the addition, subtraction, multiplication and division of two numbers, each of which is subject to error, is now given.

Ranges of error

If the value of the number X lies in the range $[a, b]$ and Y lies in the range $[c, d]$, then:

$X + Y$ lies in the range $[a + c, b + d]$
$X \times Y$ lies in the range $[a \times c, b \times d]$
$X - Y$ lies in the range $[a - d, b - c]$
$X \div Y$ lies in the range $[a \div d, b \div c]$

As examples of the use of the above, suppose that X has value 12 (n1) and Y has value 5 (± 1). We can write X = [11.5,12.5] and Y = [4,6].

Therefore, $X + Y$ = [11.5,12.5] + [4,6] = [11.5+4,12.5+6] = [15.5,18.5]

$X \times Y$ = [11.5,12.5] × [4,6] = [11.5×4,12.5×6] = [46,75]

$X - Y$ = [11.5,12.5] – [4,6] = [11.5-6,12.5-4] = [5.5,8.5]

$X \div Y$ = [11.5,12.5] ÷ [4,6] = [11.5÷6,12.5÷4] = [1.92,3.13] (2D)

The least and greatest values of any expression can now be calculated. The technique is to work through the expression, bit by bit, obeying the above rules. Remember, however, to obey also the usual rules of arithmetic when working through the expression (i.e. '×' and '÷' are considered before '+' and '-', and brackets to be evaluated first), sometimes known as the 'BODMAS' rule. Examples 3 and 4 demonstrate the technique.

14. Example 3 (The *range* in which the value of an arithmetic expression lies)

Question

Calculate the range of possible values for the expression: $\dfrac{4.12 - 8.3}{0.8}$, where each term has been rounded.

Answer

4.12 is measured to 2D = 4.12 ± 0.005 = [4.115,4.125];

8.3 is measured to 1D = 8.3 ± 0.05 = [8.25,8.35];

0.8 is measured to 1D = 0.8 ± 0.05 = [0.75,0.85].

Thus the expression can be written as:

$$\frac{[4.115,4.125] + [8.25,8.35]}{[0.75,0.85]} = \frac{[4.115+8.25,4.125+8.35]}{[0.75,0.85]}$$

$$= \frac{[12.365,12.475}{[0.75,0.85]}$$

$$= \left[\frac{12.365}{0.85}, \frac{12.475}{0.75}\right] \text{ (division rule)}$$

$$= [14.55,16.63] \text{ (2D).}$$

15. Example 4 (The *value range* for an expression in a business situation)

Question

A machine can produce 2000 (± 25) items per day, each of which can weigh between 5 and 7 grammes. At the end of each day, the production from eight similar machines is loaded into at least 10 (but no more than 15) equally weighted shipping crates. Find the lower and upper limits of the weight of one loaded crate (to the nearest gramme).

Answer

Number of items produced per machine = 2000 ± 25 = [1975,2025].

Total weight of production per machine per day

= [1975,2025] × [5,7]

= [1975×5,2025×7] (using multiplication rule)

= [9875,14175] grammes.

Total weight of production per day from eight machines

= 8 × [9875,14175]

= [79000,113400] grammes.

Therefore, the weight range of one loaded crate

$$= \frac{[79000,113400]}{[10,15]}$$

$$= \left[\frac{79000}{15}, \frac{113400}{10}\right] \text{ (using division rule)}$$

= [5267,11340] grammes (n1).

16. Fair and biased rounding

Only fair rounding has been considered so far. However, sometimes rounding is performed in one direction only. For instance:

i. When people's ages are quoted (in years), they are usually rounded down. Thus, if someone's age was given as 31 years, their actual age could be as low as 31 years and 0 days or as high as 31 years and 364 days.

ii. Suppose a job in the factory needed 83 bolts and the stores only issue bolts in sets of 10. Clearly the 83 would be rounded up to 90. In this, and similar situations, rounding would be performed upwards, since the tendency is to slightly overstock rather than to understock (and run the risk of not being able to satisfy the requirements of a job or a customer order).

This is called *biased* rounding. Maximum errors involved in biased rounding could be up to twice the size of the errors involved when using fair rounding. For example, an age quoted as 25 (which will have been rounded down) could be representing an actual age which is as high as 25 years and 364 days. In other words, the maximum error (to all intents and purposes) is 1 year. Compare this with a maximum error of only 0.5 years (either + or -) which would have resulted from fair rounding (i.e. to the nearest year).

A similar table to that shown for fair rounded numbers (in section 11) can be drawn up for biased rounded numbers and is shown in Table 3.

Table 3 Maximum errors in biased rounded numbers

Maximum errors in biased rounded numbers

Degree of rounding	*Maximum error*
...	...
...	...
3D	± 0.001
2D	± 0.01
1D	± 0.1
lowest or highest whole number	± 1
lowest or highest 10	± 10
lowest or highest 100	± 100
lowest or highest 1000	± 1000
...	...
...	...

17. Compensating and systematic errors

a) When numbers are rounded fairly, the errors involved are known as *compensating errors*. This name is used because, in the long run, we would expect half the errors to be on one side (i.e. negative) and half on the other (positive), compensating for each other. For example:

						Total
Real value	15123	23375	32914	76089	23547	171048
Rounded value (nearest 1000)	15000	23000	33000	76000	24000	171000
Error	+123	+375	-86	+89	-453	+48

When numbers subject to compensating errors are added, the errors should (roughly speaking) cancel each other out, leaving the total relatively error-free. This can be seen from the above data, where the relative error in the rounded total

$$= \frac{48}{171000} \times 100\%$$

$$= 0.3\%.$$

b) When numbers are subject to biased rounding, the errors involved are known as *systematic* (or *biased* or *one-sided*). The example below shows the numbers used in a) rounded down.

						Total
Real value	15123	23375	32914	76089	23547	171048
Rounded value (lowest 1000)	15000	23000	32000	76000	23000	169000
Error	+123	+375	+914	+89	+547	+2048

When numbers subject to systematic errors are added, the errors quickly accumulate in relative terms. The relative error in the total of rounded values above is $\frac{2048}{169000} \times 100\% = 12.1\%$ (much higher than the 0.3% obtained in a)).

Types of errors in numbers

Compensating errors are the errors involved when numbers are subject to fair rounding. Totals of these type of numbers will be relatively error-free.

Systematic errors are errors involved when numbers are subject to biased rounding. Totals of these type of numbers will have a relatively high error.

18. Example 5 (Effects of adding numbers subject to *compensating / systematic* errors)

Question

The number of minor industrial accidents of a particular type reported per month over six successive months were 41, 62, 87, 96, 32, 39. Calculate the absolute and relative errors for the six-monthly totals of rounded figures, if rounding is performed: (a) to the nearest 10 (b) to the highest 10 (c) to the lowest 10.

Answer

The calculations are shown in Table 4.

Note: Relative error $= \dfrac{\text{Absolute error}}{\text{Total of rounded values}} \times 100\%$

The relative errors in the totals of the two columns subject to systematic errors can be seen to be approximately ten times the relative error in the total of the column subject to compensating errors.

Table 4

	True values	Rounded to the nearest 10 (subject to compensating errors)	Rounded to the highest 10	Rounded to the lowest 10
			(subject to systematic errors)	
	41	40	50	40
	62	60	70	60
	87	90	90	80
	96	100	100	90
	32	30	40	30
	39	40	40	30
Totals	357	360	390	330
Absolute errors		+3	+33	–27
Relative errors		+0.8%	+8.5%	–8.2%

19. Avoiding errors when using percentages

Percentages are normally used with business statistics in order to eliminate actual units so that comparisons can be made. They are particularly useful for measuring increases (or decreases) in sets of values, but care must be taken when calculating and interpreting them. The following three notes highlight areas where errors in calculation or interpretation often occur.

a) *Do not confuse percentage and actual values*. As an example:

	Number of new orders		Absolute increase	Percentage increase
	Period 1	Period 2		
Firm A	10	50	40	400
Firm B	350	451	101	29

Notice that although firm A has a smaller actual increase in orders, its percentage increase is much larger than that of firm A since it is based on a very small period 1 value.

b) *When calculating percentage increases involving a set of values over time, the base time period for the increase must be clearly stated*.
As an example:

Year	1	2	3	4
Number unemployed	200,000	252,000	310,000	376,000

i. % age increase in year 2 (based on year 1) $= \dfrac{252{,}000 - 200{,}000}{200{,}000} \times 100\%$
$= 26\%$
% age increase in year 3 (based on year 2) $= 23\%$
% age increase in year 4 (based on year 3) $= 21\%$

ii. % age increase in year 2 (based on year 1) $= 26\%$
% age increase in year 3 (based on year 1) $= 55\%$
% age increase in year 4 (based on year 1) $= 88\%$

c) *Percentages should not be added (or averaged) in the usual way*, since each is normally derived from a different base. An overall percentage must be calculated *using grand total figures*. Table 5 gives an example of this.

Table 5 Calculating an average percentage

	Units of output		Percentage increase
	Year 1	Year 2	
Factory A	30,000	30,500	1.7
Factory B	15,000	16,000	6.7
Factory C	16,000	20,000	25.0
Total	61,000	66,500	9.0

Notice that the overall percentage increase in output has been calculated on 66,500 compared with the previous 61,000, and NOT a combination of 1.7, 6.7 and 25.0.

20. Summary

a) Data can be classified:
   i. by source - primary or secondary
   ii. by measurement - numeric or non-numeric
   iii. by preciseness - discrete or continuous
   iv. by number of variables - univariate (one variable) or bivariate (two variables) or multivariate (many variables).

b) Discrete data can be measured precisely and progresses from one value to another in definite steps. Continuous data cannot be measured precisely.

c) 'Fair' rounding is the process of rounding either up or down according to the value of the first digit of those that are ignored. Rounding can be performed by decimal place or number of significant digits.

d) Errors in data can be uncontrollable (and only minimized by organisational methods) or planned (by rounding).

e) Errors in numbers can be described using:
   i. intervals, giving the lowest and highest possible values that a number can take, in the form [a,b];
   ii. estimates, giving the maximum absolute error, in the form estimate ± error;
   iii. estimates, giving the maximum relative error, in the form estimate ± error%.

f) A number that has been rounded can be put into any one of the forms given in e) above. Any arithmetic expression involving at least one number subject to an error will itself be subject to error.

g) If X = [a,b] and Y = [c,d], then:

$$X+Y = [a+c, b+d]; \qquad X \times Y = [a \times c, b \times d];$$

$$X-Y = [a-d, b-c]; \qquad \frac{X}{Y} = \left[\frac{a}{d}, \frac{b}{c}\right].$$

h) Biased rounding is where the rounding is always one-sided (i.e. always up or always down). Maximum errors involved here are always twice those involved with fair rounding.

i) Compensating errors are made when numbers are rounded fairly and always have relatively small totals. Systematic errors are made if the rounding is biased and they have relatively large totals.

21. Student self review questions

1. What is meant by discrete data? Give some examples. [3]
2. What is meant by continuous data? Give some examples. [4]
3. Give an example each of univariate and bivariate data. [5]
4. What is fair rounding and what are the two main conventions used to display rounded values? [6]

5. Give some examples of 'unpredictable' errors in data and say how they might be minimized. [7]
6. Describe an 'approximate' number and give its three main forms. [9]
7. Explain what biased rounding means and give some examples. [16]
8. What is the difference between a compensating and systematic error and how do they affect the totals of approximate numbers? [17]
9. Why cannot percentages be added or averaged in the normal way? [19]

22. Student exercises

1. *MULTI-CHOICE*. Which one of the following would constitute a set of discrete data?
   a) Time taken to travel to work each day over one year.
   b) Weights of a consignment of tins of plum tomatoes.
   c) Number of cars passing a census point each minute over a 3-hour period.
   d) Age of applicants applying for catering jobs over a 3-month period in a large hotel chain.
2. *MULTI-CHOICE*. The value 8.2 has been rounded to 1 decimal place while 16 has been rounded to the nearest whole number. What is the largest value that the sum of the two numbers could possibly be?
   a) 24.75 b) 24.50 c) 24.30 d) 24.25
3. Classify the following sets of data in as many ways as possible.
   a) The times that a number of separate jobs have taken to complete.
   b) The job title and number of years experience of a sample of office workers.
   c) The location and number of employees of a set of textile firm's head offices.
   d) The departments to which a set of employees belong.
   e) The average weekly wages of manual and non-manual workers broken down by sex and industry over three separate years.
4. Round the following numbers to the level stated.
   a) £148,356.78 (nearest £10)
   b) 23,345 tons (to nearest 1000 tons)
   c) 3.245 mm (1D)
   d) £16.42 (to nearest £)
   e) 23 months (to highest 10 months)
   f) £18,625 (to lowest £100).
5. *MULTI-CHOICE*. Systematic errors are made when:
   a) numbers have been rounded fairly
   b) numbers have been rounded in a biased fashion
   c) numbers have been obtained by stratified sampling
   d) numbers have been obtained by systematic sampling
6. *MULTI-CHOICE*. Two groups of stock products, Kappa and Lambda, are valued. Kappa is valued at £100,000 ± 5% and Lambda at £200,000 ± 10%. The *maximum* percentage error in the combined stock valuation of £300,000 is closest to:
   a) 7% b) 8% c) 10% d) 15%

7. Give a range of error, in the form [a,b], for the value of the following expressions:
   a) 143 (± 4) + 56 (± 3).
   b) 12 (± 1) × 4 (± 1) + 27 (± 4).
   c) 31.4 (1D) + 12.23 (2D) × 11 (n1) to 2D.
   d) {31.4 (1D) + 12.23 (2D)} × 11 (n1) to 2D.
   e) $\dfrac{140\ (n10) - 130\ (n10)}{14\ (n1) + 13.2\ (1D)}$ to 2D.
8. A businessman estimates that he can buy 200 (± 20) tables at a cost of £15 (± £1) each and sell them at £18 (± £2) each.
   a) Calculate his estimated (expected) profit on the whole deal.
   b) Give a range of values within which his total profit will lie.
9. A firm works a nominal 38-hour week but with overtime and short time its actual working week varies by as much as 1 hour from the nominal figure. The firm produces 50 (± 2) articles per hour. The production cost and selling price are £2 and £3 per unit respectively, rounded off to the nearest 10 pence. Assuming that all production is sold, calculate the following.
   a) The range of production per week.
   b) The range of:
      i. weekly production costs
      ii. weekly revenue.
   c) The range of weekly profit.
   d) The expected weekly costs and hence the minimum and maximum percentage profit to 1D (based on expected costs).
10. In a single financial year, a building society lends £75 million (2S) to house buyers. The ratio of lending to income for the society for the year is 71.4% (1D). Calculate the range of income (to the nearest £1000) for the society over the year.
11. The stock movements of a particular stores item over a quarter is given below:
    Given that deliveries have been rounded to the lowest 10, issues have been rounded to the highest 10 and the balance at January 1st is exact, calculate the range of values for the balance of stock at April 1st. What is the expected balance at April 1st?

    Balance at January 1st: 33

	Deliveries	Issues
January	80	90
February	110	90
March	140	130