UNIVERSITY OF LONDON
18 58
EXTERNAL PROGRAMME

Statistics 1

Economics, Management, Finance and the Social Sciences

C. Phillips

2002

279004a

This guide was prepared for the University of London by:

C. Phillips, (BSc Econ.) PhD, Senior Fellow in Statistics, London School of Economics.

This is one of a series of subject guides published by the University. We regret that due to pressure of work the author is unable to enter into any correspondence relating to, or arising from, the guide. If you have any comments on this subject guide, favourable or unfavourable, please use the form at the back of this guide.

This subject guide is for the use of University of London External students registered for programmes in the fields of Economics, Management, Finance and the Social Sciences (as applicable). The programmes currently available in these subject areas are:

Access route

Diploma in Economics

BSc Accounting and Finance

BSc Accounting with Law/Law with Accounting

BSc Banking and Finance

BSc Business

BSc Development and Economics

BSc Economics

BSc Economics and Management

BSc Information Systems and Management

BSc Management

BSc Management with Law/Law with Management

BSc Politics and International Relations

BSc Sociology.

The External Programme
Publications Office
University of London
34 Tavistock Square
London WC1H 9EZ
United Kingdom.

Web site: www.londonexternal.ac.uk

Published by: University of London Press

Printed by: Central Printing Service, University of London, England

Table of contents

Notes

Introduction

The material in this subject is necessary as a preparation for some subjects you may study later on in your degree. In particular it has links with *Sociology* and *Marketing and Market Research.* You may also choose to take *Statistics 2* or *Management Mathematics* so that you can study the concepts introduced here in more detail or develop your economic statistics by taking *Elements of Econometrics* and *Economic Statistics.* You may wish to build on your interest in social research and take *Methods of Social Research.* These subjects will require you to have a basic understanding of the ideas introduced in this guide, which can only be absorbed by seeing how they emerge in the techniques you will learn here.

Aims and objectives

An understanding of statistics and its uses and limitations is one of the most important skills you can acquire for modern life and the modern workplace, so we hope you take full advantage of this opportunity to master the early stages.

After successfully completing this subject you will:

- be familiar with the key ideas of statistics that are accessible to a student with a moderate mathematical competence
- understand the ideas of randomness and variability, and the way in which these link to probability theory to allow the systematic and logical collection of statistical techniques of great practical importance in many applied areas
- have a grounding in probability theory and some grasp of the most common statistical methods.

How to study statistics

For statistics, you need some familiarity with abstract mathematical concepts and yet enough common sense to see how to use those ideas in real-life applications. The concepts needed for probability and for statistical inference are hard to absorb by just reading them in a book. You need to read, then think a little, then try some problems, and then read and think some more. This procedure should be repeated until the problems are easy to do; **you should not spend a long time reading and then, as a result, forget about the problems.**

You will also need to be able to use basic arithmetic notation and understand some basic ideas of algebra. These are introduced in Chapter 1.

Calculators and computers

You will need to provide yourself with a good calculator that has built-in routines for means, standard deviations and regression. It is best if it is not programmable, because such machines are not allowed in the exams by the University. The models change all the time, so it is hard to recommend one, but something as good as the Casio Scientific Calculator fx-570s is fine for the built-in routines. More expensive graphical calculators, with the capacity to carry out symbolic algebra and to plot data, are in the shops but are not necessary for this subject. Many students do not find it easy to remember how to use even the simplest calculators. Whatever your choice, do make sure that you are comfortable with the calculator and with all the functions you need.

Those students aiming to carry out serious statistical analysis (beyond the level of this subject) will probably use some statistics package such as MINITAB or SPSS. It is not necessary for this subject to have such software available, but those who may have it could profit by sometimes using it in this subject.

The subject guide and how to use it

This guide does not attempt to offer a complete treatment. There are very many well-written textbooks that cover this subject, and it would be foolish to compete with them. You will need to buy at least one textbook and consult others from time to time. The choice of the main textbook is your personal choice, though some students will have a teacher to guide them. There are many good textbooks besides those recommended in this guide and you should be prepared to look at several different texts just to see a lot of extra examples on some tricky topics.

The purpose of the guide is to describe the syllabus in some detail and to show what level of understanding is expected. It should not be used as a main source of help but as a preliminary to more detailed work with your chosen textbooks. However, in order to aid reading and understanding you will find activities for you to try in the text. In addition there are past examination questions at the end of each chapter. Complete answers to all these may be found on the Web.

The subject guide is divided into 11 chapters which should be worked in the order given. There is little point in rushing past half understood material to reach the later chapters. The presentation is sequential, and not a series of self-contained topics. You should be familiar with the earlier chapters, and have some understanding of them, before moving to the later ones.

The following procedure is recommended for each chapter:

1. Read the introductory comments.
2. Read the appropriate section of your text.
3. Work on the appropriate self-test where given.
4. Study the notes and further examples in your textbook.
5. At the end of the chapter go through the Learning outcomes carefully.
6. Then attempt some of the examination questions given at the end of the chapter.
7. Refer back to this subject guide, or to the text, or to supplementary texts if necessary, to improve understanding to the point where you can work confidently through the problems.

The last two steps are the most important. It is easy to think that one has understood the text after reading it, but **working through problems is the crucial test of understanding. Problem solving should take most of your study time.**

There are also specimen examination questions at the end of most chapters and a sample examination paper at the end. Some supplementary material, including comments on the Activities will be found at the url:

http://stats.lse.ac.uk/knott

If you do not have access to the Internet, please contact the LSE External Study Office, Houghton Street, London WC2 2AE to request a paper copy.

Try to be disciplined about this: Don’t look up the answers until you have done your best. Statistical ideas may seem unfamiliar at first, but your attempts at the questions, however dissatisfied you feel with them will help you understand the material far better than reading and rereading the prepared answers!

Time management

About one-third of your private study time should be spent reading textbooks and the other two-thirds doing problems. If you were following a lecture course in this subject you might expect 60 hours of formal teaching and another 200 hours of private study to be enough to cover and understand the material. Of the 200 hours of private study about 120 hours should be spent on trying problems (which may well require more reading) and about 80 hours on initial study of the textbook and subject guide.

As a help to your time management, we have converted the chapters and topics of this subject into **approximate** weeks of a typical 15-week university half unit or semester course. Your study of this subject might well take a period of time other than 15 weeks. There is nothing magical about the 15 weeks figure: it is purely for indicative purposes. What you should gain from the following breakdown is an indication of the relative amounts of time to be spent on each topic. Bear in mind, however, that some of you may not need to spend as much time on Chapter 1 if the concepts and techniques of basic arithmetic and algebra (in particular the use of summation signs, the equation of a straight line, and the idea of a uniform distribution) are familiar to you.

Chapter 1	–	2 weeks
Chapter 2	–	1 week
Chapter 3	–	2 weeks
Chapter 4	–	1 week
Chapter 5	–	1 week
Chapter 6	–	1 week
Chapter 7	–	1 week
Chapter 8	–	1 week
Chapter 9	–	2 weeks
Chapter 10	–	1 week
Chapter 11	–	2 weeks.

Examination paper

Important: the information and advice given in the following section is based on the examination structure used at the time this guide was written. However, the University can alter the format, style or requirements of an examination paper without notice. Because of this, we strongly advise you to check the instructions on the paper you actually sit.

The examination is by a two-hour, unseen, question paper. No books may be taken into the examination, but the use of calculators is permitted, and statistical tables are provided.

You will be asked to attempt most of the questions in the examination so you are encouraged to study the whole syllabus. A specimen examination paper is given at the end of this guide. All the ideas in this subject guide will be important to your further work in business, management, economics, and the social sciences.

There is not much that can be helpfully said about examination techniques specific to this paper. As always it is important to manage time carefully and not to get stuck on one question. If English language is a problem it may be easier when tackling questions involving explanation or definition to give examples rather than attempt an abstract description. Sample examination questions are given at the end of each chapter and a specimen examination paper is given on the web site (http://stats.lse.ac.uk/knott).

Reading

Main text and statistical tables

You should be able to assimilate the main ideas of the **Statistics 1** course using two main texts and statistical tables, as follows:

Lindley, D.V. and W.F. Scott. *New Cambridge Statistical Tables.* (Cambridge: Cambridge University Press, 1995) second edition [ISBN 0 5214 8485 5].

Moser, C.A. and G. Kalton. *Survey Methods in Social Investigation.* (Aldershot: Dartmouth.1979) second edition [ISBN 0 4358 2604 2].

Newbold, P. *Statistics for Business and Economics.* (London: Prentice-Hall, 1995) fourth edition [ISBN 0 1385 5549 0].

First, note that the New Cambridge Statistical Tables are those distributed for use in the examination. It is essential that you get familiar with these tables rather than those at the end of your textbook.

The Moser and Kalton text gives you background for the material on surveys and experimental design. You will also find it useful if you are studying for Sociology or, perhaps later, when you take *Marketing and market research.*

We refer to Newbold as a main text, but there are many other books that are as good. One looks at the range of textbooks that cover this subject with admiration for their excellence. You are encouraged to look at those given below, and at any you find. It may be necessary to look at several texts for any topic, and you may find it helpful to study texts for which additional examples are available on CD-ROM. There is even more computer-based teaching material available freely on the Web. One example of a computer-based approach with lively demonstrations is:

Doane, D.P., K. Mathieson and R.L. Tracy. *Visual Statistics 2.0.* (Boston: McGraw-Hill/Irwin, 2001) [ISBN 0 0724 0014 5 and 0 0724 0012 9 (CD)].

Other texts

These texts are recommended as alternatives if you have difficulty obtaining Newbold. There are, however, many good introductory texts on the market, so do look for one which suits you as long as it covers the subject guide main topics. Remember though that the references we give later in the guide refer to Newbold.

Huff's book is a useful introduction to the whole area. You may find it useful to read it before embarking on this guide:

Huff, D. *How to lie with Statistics* (London: Penguin, 1994) [ISBN 0 1401 3629 0].

Two texts supplement the work in Moser and Kalton and will be useful to you when you reach Chapter 10. They are:

Douglas, J.W. *The Home and the School, a Study of Ability and Attainment in the Primary School.* (St Albans: Panther, 1964) [Note: you will have to look for this in your library. It is also described in Moser and Kalton].

Shipman, M. *The Limitations of Social Research.* (London: Longman, 1997) fourth edition [ISBN 0 5823 1103 9].

The book by Douglas describes a particular longitudinal survey and both it and Shipman, which gives examples of the use of social research, will be very helpful to you in Sociology.

In addition, Social Trends (a compendium of UK official statistics and surveys) is useful in Chapter 9.

The Stationery Office, *Social Trends*. (London: HMSO, 2002) [ISBN 011-621472-4].

The other books are alternatives to Newbold:

Aczel, A.D. *Complete Business Statistics*. (London: Irwin/McGraw Hill, 1999) [ISBN 0 0728 9302 8].

Anderson, D.R., D.J. Sweeney, and T.A. Williams. *Statistics for Business and Economics.* (Cincinatti: South-Western Thomson Learning, 2002) eighth edition [ISBN 0 3240 6671 6].

Hanke, J.E. and A.G. Reitsch. *Understanding Business Statistics.* (Burr Ridge Ill: Irwin, 1994) second edition [ISBN 0 2561 1219 3].

Mason, R.D. and D.A. Lind. *Statistical Techniques in Business and Economics.* (Boston: McGraw Hill, 2001) eleventh edition [ISBN 0 0724 0282 2].

Moskowitz H. and G.P. Wright. *Statistics for Management and Economics.* (London: Charles Merrill Publishers, 1985) [ISBN 0 6752 0211 6].

Wonnacott, T.H. and R.J. Wonnacott. *Introductory Statistics.* (Chichester: Wiley, 1990) fifth edition [ISBN 0 4715 1733 X].

To help you with the basic mathematics you will need for the course and which you will cover in Chapter 1, we recommend:

Anthony, M. and N. Biggs. *Mathematics for Economics and Finance.* (Cambridge: Cambridge University Press, 1996) [ISBN 0 5215 5913 8 (pbk)]. Chapters 1, 2 and 7.

Notes

Chapter 1

Simple algebra and co-ordinate geometry

Essential reading

Anthony, M. and N. Biggs. *Mathematics for Economics and Finance.* (Cambridge: Cambridge University Press, 1996) [ISBN 0 5215 5913 8 paperback] Chapters 1, 2 and 7.

You may find Anthony and Biggs or any other appropriate non-specialist mathematics textbook helpful as background for the mathematics you will need for this course. You should not need all the material in **Mathematics 1**, but if you are working for that, this will support your work for **Statistics 1**.

Introduction

This chapter discusses some of the very basic aspects of the subject, aspects on which the rest of the subject builds. It is essential to have a firm understanding of these topics before the more advanced topics can be understood.

You should be able to add, subtract, multiply and divide, and handle at least the simple operations of an electronic calculator. This chapter should hopefully be used mainly as a reminder of the algebraic and arithmetic rules you covered at school. You should make sure you are comfortable with them. Some material, such as the use of summation signs or drawing graphs of functions, may be new to you. If so, you should master these new ideas before going on to the next chapter.

Remember that although it is most unlikely that an examination question would only test you on the subjects in this chapter alone, the material used may well be an important part of your answer!

Arithmetic operations

Make sure you understand the rules of BODMAS, which stands for:

Brackets

Of

Division

Multiplication

Addition, and

Subtraction.

This shows the order in which you should prioritise arithmetic operations.

Activity

A1.1

Work out the following:

a) ⅓ of 12 – (4 × 2)

b) 1/3 of 12 – 4 ÷ 2

c) (1 + 4)/5 × (100 – 98)

Remember that you can find answers to these, and all activities, on the web site[1]. If you find these difficult at all, go to Anthony and Biggs, Jaques, or your old school textbook and work on some more examples before you do anything else.

You should also know that the:

- **sum** of *a* and *b* means $a + b$
- the **difference** between *a* and *b* means either $a–b$ or $b–a$
- the **product** of a and b means $a \times b$, and
- the **quotient** of *a* and *b* means *a* divided by *b*, i.e. *a/b*.

Squares and square roots

Remind yourself about these. When a number *x* is multiplied by itself we write x^2, (that is $x \times x$, the square of *x*). (In fact you can extend this to cubes, $x^3 = x \times x \times x$ and x^n ($x \times x$ n times) but you will not need this information in *Statistics 1*).

Remember that x^2 is always non-negative. This is important when you finish a long calculation of s^2 (see Chapter 3) or r^2 (see Chapter 11) and find you have a negative number. This tells you that you have made a mistake!

It may help you to think of the square root of *x* (written as $\sqrt{x}$) as the opposite of the square, so that $\sqrt{x} \times \sqrt{x} = x$. In practice the main problems you will have are likely to lie in taking square roots of numbers with decimal points (you will probably have to do this a lot in the subject matter of Chapters 7 and 8 when you are dealing with proportions). Be careful that you understand that .9 is the square root of .81 and that .3 is the square root of .09 (and **not** .9).

Activity

A1.2

Work out the following (use a calculator where necessary):

a) $\sqrt{16}$

b) $(.07)^2$

c) $\sqrt{.49}$

Proportions and percentages

Make sure you are at home with these. If you are not, take a look at the textbooks and practise all of the exercises until you feel comfortable.

Activity

A1.3

a) What is 98% of 200?

b) Give 17/25 as a percentage

c) What is 25% of 98/144?

Some new notations

The absolute value in statistics

One useful sign in statistics is | | (the absolute value of). Statisticians sometimes want to indicate that they only want to use the positive value of a number. For example you might find that it was 5 miles from town *x* to town *y*. The distance from town *y* to town *x* may well be represented mathematically as – 5m (i.e. minus 5 miles), but you will probably only be interested in the absolute amount. So, for example, $|–d| = d$ and $|(3-5)|$ is + 2.

[1] *See introduction for details.*

Activity

A1.4

Give the absolute values for:

a) $| -8 |$

b) $|(15-9)|$

‘Greater than’ and ‘less than’ signs

Learn the following:

$>$ means ‘is greater than’

$\geq$ means ‘is greater than or equal to’

$<$ means ‘is less than’

$\leq$ means ‘is less than or equal to’

$\approx$ or $\simeq$ means ‘is approximately equal to’

Activity

A1.5

a) For which of the following is $x > 3$?

2, 3, 7, 9

b) For which is $x < 3$?

c) For which is $x \leq 3$?

d) For which is $x^2 \geq 49$?

The use of summation signs $\sum$

Summation signs are likely to be new to most of you. They are very commonly used in most statistics. You will see summation signs in most of the following chapters! So it is most important you get to grips with them now.

One of the basic quantities you will meet is the arithmetic mean (it is sometimes called the ‘average’ but there are in fact other measures of average apart from the mean which you will encounter in the next chapter of the guide). It is made up by adding the measures of the number of observations in which you are interested and dividing by the number of observations. So, if there are n items, we write them as:

x_1, x_2, x_3 ... x_n (i.e. the first, second, third ... nth objects)

and describe their total as the sum of the x_i:

i.e. Total = $\sum_{i=1}^{i=n} x_i$ (the sum of the x’s includes all from x_1 to x_n)

To find the mean the total is divided by n:

i.e. Mean = $\sum_{i=1}^{i=n} x_i / n$

We may not want to include all our xi in our summation. Perhaps the x1 really belongs to another group of observations. In that case we would write:

$\sum_{i=2}^{i=n} x_i$ i.e. $x_2 + x_3 + \dots x_n$.

We can do the same if we want to sum the x^2. We will do this for some measures of distribution as you will see in Chapter 4.

So, if we want to add all the x_i^2's from x_3 to x_6 i.e. $x_3^2 + x_4^2 + x_5^2 + x_6^2$, we write:

$$\text{Sum} = \sum_{i=3}^{i=6} x_i^2$$

Activity

A1.6

Given $x_1 = 3, x_2 = 1, x_3 = 4, x_4 = 6, x_5 = 8$ find:

a) $\sum_{i=1}^{i=5} x_i$

b) $\sum_{i=3}^{i=4} x_i^2$

Given also that $p_1 = 1/4, p_2 = 1/8, p_3 = 1/8, p_4 = 1/3, p_5 = 1/6$ find:

c) $\sum_{i=1}^{i=5} p_i x_i$ (Hint – write out the $p_i x_i$ and multiply first)

d) $\sum_{i=3}^{i=5} p_i x_i^2$

If you find these difficult, go back to an elementary text and do some more work on this. It is most important that you deal with this before you embark on the topic of descriptive means in Chapter 3.

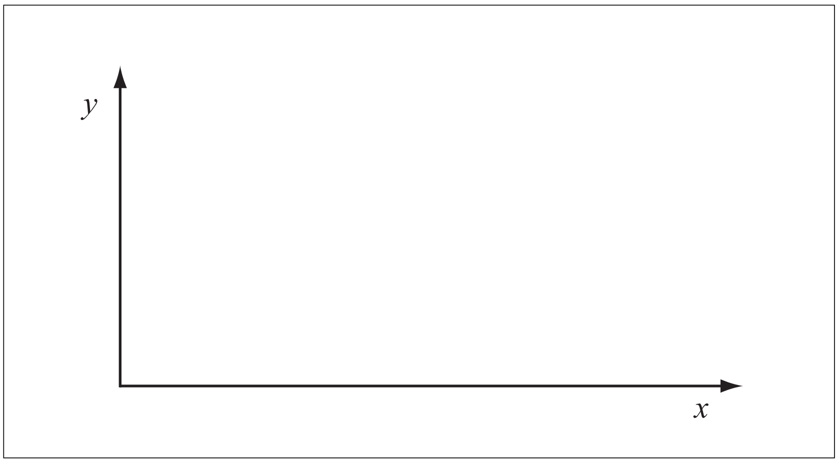
You should also note that when all the x_i's are summed we sometimes write the short cuts $\sum x_i$ or $\sum_1^n x_i$ or sometimes even $\sum_1^n x_i$ instead of $\sum_{i=1}^{i=n} x_i$ in full.

Graphs

You will spend some time learning how to present material in graphical form and also in the representation of the normal distribution. You should make sure you have assimilated the following material.

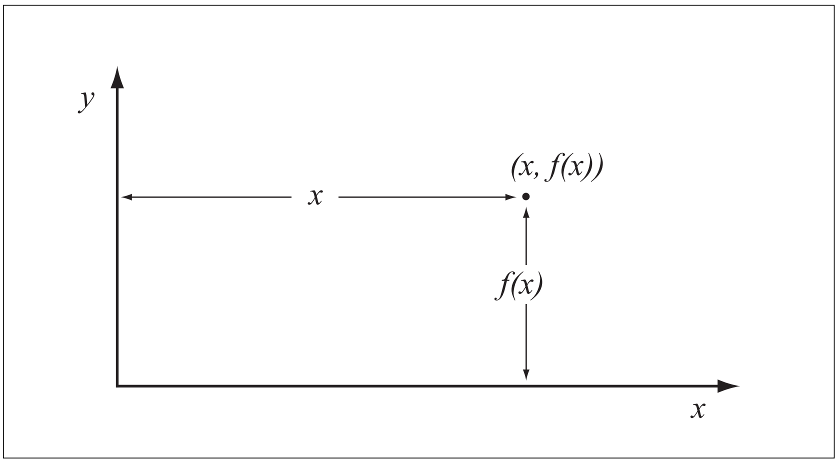
The graph of a function $y = f(x)$ is the set of all points in the plane of the form $(x, f(x))$. Sketches of graphs can be very useful. To sketch a graph, we start with the x-axis and y-axis, as in Figure 1.1.

Figure 1.1 x and y axes



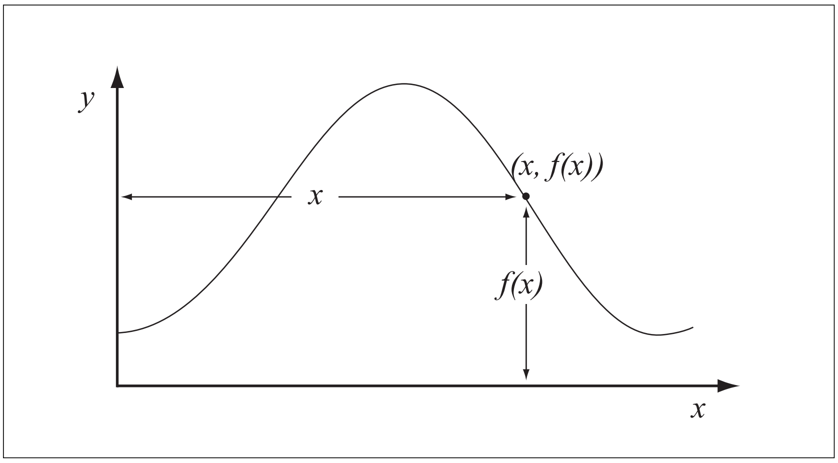
We then plot all points of the form (*x*, *f* (*x*)). Thus, at *x* units from the origin (the point where the axes cross), we plot a point whose height above the *x*-axis (this is, whose *y*-co-ordinate) is *f* (*x*). (Figure 1.2)

Figure 1.2 Plotting a point on a graph



Joining together all points of the form (*x*, *f* (*x*)) results in a **curve**, called the **graph** of *f* (*x*). This is often described as the **curve of equation** *y* = *f (x)*. Figure 1.3 gives an example of what a typical curve might look like.

Figure 1.3 The curve y = f



These figures indicate what is meant by the graph of a function, but you should not imagine that the correct way to sketch a graph is to plot a few points of the form (*x*, *f* (*x*)) and join them up; this approach rarely works well and more sophisticated techniques are needed.

There are two functions you need to know about for this course:

- the linear function (i.e. the graph of a straight line), and
- the Normal[2] function (which we will meet very often in later chapters).

The graph of a linear function

We start with the easiest graph of all. The linear functions are those of the form $f(x) = mx + c$ and their graphs are straight lines, with **gradient** *m*, which cross the *y*-axis at the point (0,*c*). Figure 1.4 illustrates the graph of the function $f(x) = 2x + 3$ and Figure 1.5 the graph of the function $f(x) = – x + 2$.

Figure 1.4 The curve y = 2x + 3, a straight line

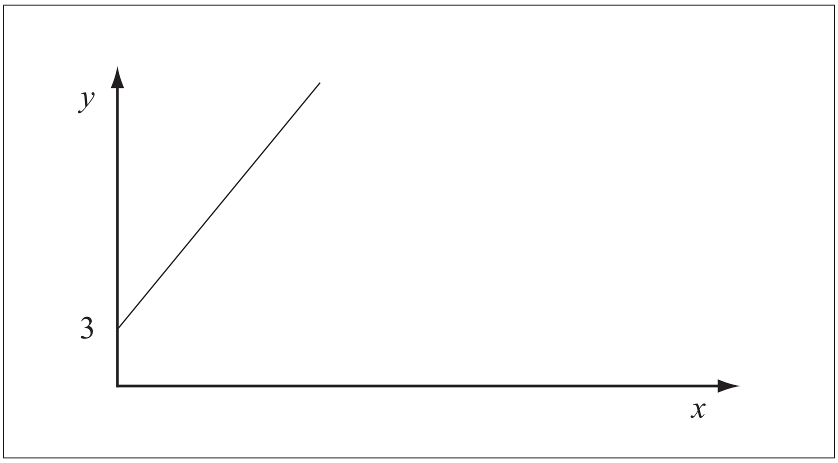
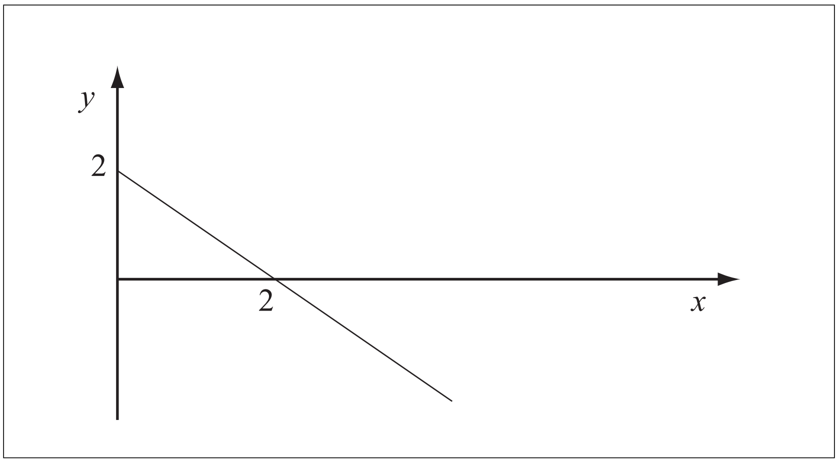


Figure 1.5 The curve y = –x + 2, a straight line



Activities

A1.7

Sketch the following:

a) y = x + 3

b) y = 3x – 2

You are going to need equations like this for all the material on **regression** in Chapter 11.

Parameters

At this point it is useful to think a little about what you have learned in the discussion of graphs, and in particular the new idea of a **parameter**.

We can define a parameter (or set of parameters) as that measure (or the set of measures) which completely describe a function.

Thus the straight line $y(x) = mx + c$ is completely determined by its slope, *m* and the point *c* at which the line cuts the *y*-axis, *c*. So we call *m* and *c* the **parameters** of the straight line. Note that the mathematical texts you will use for revision of this topic use *m* & *c*. In our guide, and in statistics generally, we use the *b* and *a* of the regression line. You will be interested in estimating these parameters when considering the topic of **regression** in Chapter 11 (where we call *m* and *c*, *b* and *a* respectively).

Activity

A1.8

What parameters could you find for a circle?

One of the reasons statisticians like to use the Normal distribution is that, despite its complicated appearance, it is completely determined by two parameters – the mean and standard deviation. If we know those two variables, we can draw a unique Normal curve. As you will see, this is extremely useful for working out probabilities and confidence intervals, and for testing hypotheses. We will look at the Normal distribution in Chapter 5 in detail.

Summary

Much of this material should be familiar to you, but some is almost bound to be new. Although it is only a language or set of rules to help you deal with statistics, without it you will not be able to make sense of the following chapters.

Before you continue, make sure you have completed all the activities, understood what you have done and, if necessary, worked through additional examples in the basic texts.

Learning outcomes

After completing this chapter and the relevant reading you should be able to:

- manipulate arithmetic and algebraic expressions using the simple rules
- understand and be able to use common signs, square, square root, greater than and less than and absolute value
- be able to use the summation sign and understand the use of the ‘i’
- draw the straight line for a linear function
- see what the parameter(s) of a function are.

Sample examination questions

1. $x_1 = 4, x_2 = 1, x_3 = 2$

For the above figures give:

$$\sum_{i=1}^{i=2} x_i^{\ 3}$$

(2 marks)

2. If $n = 4, x_1 = 2, x_2 = 3, x_3 = 5, x_4 = 7$, find:

a) $\displaystyle\sum_{i=1}^{i=3} x_i$

b) $\displaystyle\frac{1}{n}\sum_{i=1}^{i=4} (x_i)^2$

(3 marks)

Notes

Chapter 2

The nature of statistics

Essential reading

Newbold, P. *Statistics for Business and Economics.* (London: Prentice-Hall, 1995) fourth edition [ISBN 0 1385 5549 0]. Sections 6.1, 18.1, 18.2, non-mathematical parts of 18.3 and 18.4. The material on quota sampling at the end of Chapter 18.

Moser, C.A. and G. Kalton. *Survey Methods in Social Investigation.* (Aldershot: Dartmouth, 1979) second edition [ISBN 0 4358 2604 2]. Chapters 5, 6 and 9.

Note that Moser and Kalton cover random sampling, quota methods and experimental design more completely than Newbold does.

Further reading

Wonnacott, T.H. and R.J. Wonnacott. *Introductory Statistics.* (Chichester: Wiley, 1990) fifth edition [ISBN 0 4715 1733 X]. Chapter 1.

Introduction

This chapter sets the framework for the subject guide. Read it carefully, because the ideas introduced are fundamental to this subject, and, though superficially easy, give rise to profound difficulties both of principle and of methodology. **It is important that you think about these ideas as well as being thoroughly familiar with the techniques explained in Chapter 1 before you begin work on later chapters.**

Coverage of the subject

This subject is in six main sections.

1. Data Presentation (Chapter 3)

   Ways of summarising data, measures of location and dispersion, graphical presentation.

2. An Introduction to Probability (Chapter 4)

   Probability theory, probability trees.

3. The Normal distribution and ideas of sampling (Chapter 5. But note that you need to assimilate Chapter 3 and 4 before you can understand this).

   Definitions of random variables, their expectations and variances. The Normal distribution. The Central Limit theorem.

   These lead directly to the subject matter of:

4. Sampling Design (Chapter 9)

   The key features of different data sampling techniques.

   Chapters 4–6 also need to be mastered before you tackle:

5. Decision Analysis (Chapters 6, 7 and 8)

   Estimation and confidence intervals for means, proportions, sums and differences. The Student’s distribution. Hypothesis testing. Type I and Type II errors. Contingency Tables.

6. Modelling for Decision Making (Chapters 10 and 11)

   Measures of correlation, spurious correlation. Model selection and model testing.

   You may find that D.Huff's book 'How to Lie with Statistics' gives you a light-hearted introduction to all this.

   So much for the material you will meet; you might wish to review the suggested timetable that was given in the Introduction. What about the ideas?

Concepts

Before you go on to Chapter 3, it is wise to make sure you understand the following words which underpin the ideas and concepts you will need both in this module and *Statistics 2*.

Population

As is so often the case in statistics, some words have technical meanings that overlap with their common use but are not the same. 'Population' is one such word. It is often difficult to decide which population should be sampled. For instance, if we wished to sample 500 listeners to an FM radio station specialising in music should the population be of listeners to that radio station in general, or of listeners to that station's classical music programme, or perhaps just the regular listeners, or any one of many other possible populations that you can construct for yourself? In practice the population is often chosen by finding one that is easy to sample from, and that may not be the population of first choice. In medical trials (which are an important statistical application) the population may be those patients who arrive for treatment at the hospital carrying out the trial, and this may be very different from one hospital to another.

If you look at any collection of official statistics (which are most important for state planning) you will be struck by the great attention that is given to defining the population that is surveyed or sampled, and to the definition of terms. For instance, it has proved difficult to get consensus on the meaning of 'unemployed' in recent years but a statistician must be prepared to investigate the population of unemployed. Think about this carefully. It should help you with your *Sociology* and *Marketing and Market Research* modules as well as this one:

Bias

In addition to the common-sense meaning of bias, there is also a more technical meaning for the word in statistics. This will be found both in Chapter 10 of this guide and in the work on estimators in **Statistics 2**. It seems natural enough to wish to avoid bias, but it is not helpful to be swayed by the value judgements inherited from the use of a word outside the limits of academic discussion.

Sampling

Although it seems sensible to sample a population to avoid the cost of a total enumeration (or census) of that population, it is possible to make a strong argument against the practice. One might well consider that sampling is fundamentally unfair because a sample will not accurately represent the whole population, and it allows the units selected for the sample to have more importance than those not selected. This might be thought undemocratic. Many countries continue to take a full census of their population, even though sampling might be cheaper. It is less obvious, but true, that sampling might well be more accurate, because more time can be spent verifying the information collected for a sample.

Random sampling

Though it may be clear that random sampling should avoid a sample biased by the prejudices of the sampler, you should think carefully whether that is always a good idea. In other times it was thought that a sampler could produce a better sample by using personal judgement than by randomising, because all the relevant facts could be taken into consideration. Randomisation is popular because this belief seems to be contradicted by experience. Remember that a random sample may (though not very often) come out to look very biased just by chance – should one then reject it and try again? If the population has both men and women, would you accept a random sample that just by chance included only men? These difficulties are usually dealt with by some form of restricted randomisation. One might, for instance in the example above, use a stratified random sample and select half the sample from the men and half from the women.

To carry out a random sample one needs a sample frame. This is the list of all the population units. For instance, if you want to investigate UK secondary schools, the sample frame would be a list of all the secondary schools in the UK. The key advantages of random sampling are that it avoids systematic bias and it allows an assessment of the size of sampling error (as we shall see in Chapter 9).

Random sampling is carried out, for many populations, by choosing at random without replacement a subset of the very large number of units in the population. If the sample is a small proportion of the population, then there is no appreciable difference between the inferences possible for sampling with replacement and sampling without replacement.

Quota sampling

Quota sampling avoids the need for a sample frame. The interviewer seeks out units to ensure that the sample contains given quotas of units meeting specified criteria (known as **quota controls**).

Quota sampling is cheap, but it may be biased systematically by the choices made by the interviewer. For instance, interviewers avoid anyone who looks threatening, or too busy, or too strange. Quota sampling does not allow an accurate assessment of sampling error.

Observational studies and designed experiments

In industrial and agricultural applications of statistics it is possible to control the levels of the important factors that affect the results. Bias from the factors that cannot be controlled is dealt with by randomisation. These investigations are **designed experiments.** In medical, economic and other social science applications of statistics one usually just observes a sample of the population available, without control of any of the factors that may influence the measures observed. These studies are **observational studies**. Much of the methodology of statistics is devoted to disentangling the effects of different factors in such observational studies, but it is scientifically better to have a designed experiment. In this course we will only look at regression and correlation as ways of studying the connections between variables, but analysis of variance (covered in *Statistics 2*) is also very relevant in this context.

Even if it is possible to experiment, there may be ethical objections. The ethical questions raised by medical trials are still a lively area of debate. There can be strong objections from patients (or their relatives) to their treatment by an established method when a new method is available. After the trial is complete, if the new method is shown to be worse than the established one, there can be complaints from those who were given the new treatment. Some of these difficulties can be resolved by the use of sequential methods, which try to discontinue a trial at the earliest opportunity when one treatment has been shown to give better results than the others.

It is easy to be too optimistic about the use of regression (see Chapter 11) for observational studies. There is a big difference between a randomised experiment and an observational study, even if regression is used to allow for some of the distorting factors in the latter. The long-standing difficulties in proving that smoking causes lung cancer or more recently in demonstrating the good effects of seat-belts in cars, or of investigating the possible increase in leukaemia cases in children of workers at nuclear power plants should be enough to show how the problems arise. It is difficult to allow for all the distorting factors; even to say whether a factor **should** be eliminated is hard.

As an exercise in thinking through these matters, consider the fact that in the UK there are more cases of Creutzfeld-Jakob Disease among workers on dairy farms than for the population as a whole. What factors might influence this finding? How would you investigate whether Bovine Spongiform Encephalopathy in cattle, in the farms where the sufferers were working, is responsible for the excess?

Activity

A2.1

Why might you use a quota rather than a random sample in an interview survey? What problems would you then have? How would you deal with them?

(Points relating to these questions are discussed below.)

When answering those questions you will have realised that to make a random sample we need some kind of list, frame or map. If this does not exist then we might have to use a quota sample. Another reason might be speed. We may be in a hurry and not want to spend time on office work organising interviewers for a random sample. It is much quicker to set target numbers to interview.

The problem with using a quota design, for whatever reason, is that there is no way to assess the accuracy of the estimates (whether they are means, or proportions or anything else). We have no statistically acceptable way of measuring variability. In particular, bias may have been caused by the interviewers' choice of interviewee and their willingness to reply.

These problems are generally dealt with by:

- assigning quota controls so that particular kinds of interviewee are included (e.g. sex, age, economic status, office location).
- recording non-respondents by these quotas.

If this is done, market research firms feel justified in assuming that the random sample formulae may be used. There is fairly general acceptance of this point of view, though, of course, the more accurate the researchers try to be through the use of detailed controls and careful adjustments for non-response, the longer the quota sampling takes and the more expensive it is to administer.

Activity

A2.2

Explain the difference between a simple random sample and a random sample. Name two kinds of non-simple random samples.

(Points relating to these questions are discussed on page 19.)

In a simple random sample, each unit in the population has an equal non-zero chance of being selected in the sample. In a random sample each unit has a known chance of being selected for the sample.

The two most common kinds of non-simple random samples are:

- stratified (where the precision of estimates should be improved if the stratification factors are appropriate)
- clustered (where precision may not be better, but time and money may be saved on interview surveys involving travel to the site of interview).

Learning outcomes

After completing this chapter and the relevant reading, you should be able to:

- explain the difference between a population, a sample and a census
- discuss the idea of systematic bias
- describe simple random sampling and justify its use
- distinguish between stratified random sampling and quota sampling
- explain why observational studies and designed experiments differ

and you should be ready to start studying statistics!

It is rare for examination questions to focus solely on the topics covered in this chapter. However, a good understanding of the ideas covered is essential when answering the sample examination questions given with later chapters.

Chapter 3

Data presentation

Essential reading

Newbold, P. *Statistics for Business and Economics*. (London: Prentice-Hall, 1995) fourth edition [ISBN 0 1385 5549 0]. Chapter 2.

Further reading

Aczel, A.D. *Complete Business Statistics.* (London: Irwin/McGraw Hill, 1999) [ISBN 0 0728 9302 8]. Chapter 1.

Anderson, D.R., D.J. Sweeney, and T.A. Williams. *Statistics for Business and Economics.* (Cincinatti: South-Western Thomson Learning, 2002) eighth edition [ISBN 0 3240 6671 6]. Chapters 1 to 3.

Hanke, J.E. and A.G. Reitsch. *Understanding Business Statistics.* (Burr Ridge Ill: Irwin, 1994) second edition [ISBN 0 2561 1219 3]. Chapters 1, 3 and 4.

Mason, R.D. and D.A. Lind. *Statistical Techniques in Business and Economics.* (Boston: McGraw Hill, 2001) eleventh edition [ISBN 0 0724 0282 2]. Chapters 3 and 4.

Moskowitz, H. and G.P. Wright. *Statistics for Management and Economics.* (London: Charles Merrill Publishers, 1985) [ISBN 0 6752 0211 6]. Chapters 1 and 2.

Wonnacott, T.H. and R.J. Wonnacott. *Introductory Statistics.* (Chichester: Wiley, 1990) fifth edition [ISBN 0 4715 1733 X]. Chapter 2.

Introduction

This chapter contains two separate but related themes, both to do with the understanding of data. The first idea is to find graphical representations for the data, which allow one to easily see its most important characteristics. The second idea is to find simple numbers, like mean or inter-quartile range, which will summarise those characteristics. Both the ideas could be applied to the data[2] for a whole population, but in most texts they are consistently applied to a sample. The notation would change a bit if a population was being represented. Most of the graphical representations are very tedious to construct without the aid of a computer. However, one understands much more if one tries a few with pencil and paper when completing the activities in this chapter. You should also be aware that spreadsheets do not always use correct terminology when discussing and labelling graphs.

[2] *Note that the word 'data' is plural, but is very often used as if it were singular. One should say 'these data', but you may well find the use of 'this data'.*

Data collection

Examples of data collection include the following:

a) A pre-election opinion poll asks 1,000 people their voting intentions

b) A market research survey asks housewives how many hours of television they watch per week.

c) A census interviewer asks each householder how many of their children are receiving full-time education.

Definitions

The key terms used in data collection can be defined as follows:

- A **variable** is the phenomenon being measured in the experiment or observational study.
- A **continuous variable** takes any value on a range of real numbers.
- A **discrete variable** takes only distinct values, usually often integers (analogous to ‘counting’).

Activity

A3.1

Identify and describe the variables in the above examples of data collection (a–c).

Statistical summary of data

In the next part of this chapter we are concerned with summarising a set of data using different descriptive measures. The case we will consider is that of a software house about to market a new spreadsheet for IBM compatible computers. In order to decide the price of their product, the software house surveys the current price of competing products.

The data (showing prices are in £) is:

Borland QUATTRO	115	Plan Perfect	195
Kuma K-Spread II	52	Sage Planner	65
Legend Twin Level 3	155	SuperCalc 3.21	82
Logistix	64	SuperCalc v4	107
Lotus 1-2-3	265	SuperCalc v5	195
Microsoft Excel	239	VP Planner Plus	105
Multiplan IV	110		

Measures of location

Averages

Averages are an attempt to summarise a set of data in one number in order to give an idea of the general size of the variables in the data. There are three averages (or **measures of central tendency**) in common statistical use, these are:

- the arithmetic mean
- the median
- the mode.

The arithmetic mean

The arithmetic mean for a set of *n* numbers $x_1, x_2, x_3,\ldots\ldots, x_n$ is given by:

$$\bar{x} = \frac{(x_1 + x_2 + \ldots\ldots + x_n}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i$$

This is very appealing to use as a measure for the centre of a set of data, particularly when the quantities measured are natural to add together. It is however, affected greatly by extreme observations, and may better be replaced by the median (the ‘halfway mark’) when the distribution of the data is very one sided (we call this skew).

Notice that, if some observations have equal values, so that there are only m distinct values present (we might say there are in ‘groups’ or ‘classes’) then we can use a slightly different formula. Calling the groups $w_1, w_2, \ldots w_m$ with corresponding frequencies of occurrence $f_1, f_2 \ldots f_m$, the formula for the mean becomes:

$$\bar{x} = \sum_{i=1}^{i=m} f_i w_i \Big/ n \quad \text{or} \quad \sum_{i=1}^{i=m} f_i w_i \Bigg/ \sum_{i=1}^{i=m} f_i$$

This looks different, but just clusters the equal values together in the same sum that defined the mean before. We always use this formula with grouped data.

Activity

A3.2

Find the mean of the numbers of hours of television watched per week by 10 housewives:

Number of hours watched   2, 2, 2, 5, 5, 10, 10, 10, 10, 10

(The solution to this problem is discussed below.)

Do this first by adding the 10 observations and dividing by *n* (=10). Then use the alternative second method, using the grouped data formula. Here your w_1 is 2 hours, w_2 is 5 hours and w_3 is10 hours, and the frequencies are 3, 2 and 5 respectively. Make sure you understand this and get the same result for $\bar{x}$ whichever method you use[3].

The median

The median (we write this x_m) is the halfway mark; 50% of observations are greater than the median and 50% are less than it.

The set of n numbers is arranged in ascending order, say as $x_{(1)}, x_{(2)}, x_{(3)}, \ldots\ldots x_{(n)}$.

If *n* is odd $x_m = x_{(n+1)/2}$

or if *n* is even $x_{(m)} = \dfrac{x_{n/2} + x_{(n+1)/2}}{2}$

For example, the median of the numbers 1, 7, 3, 5 and 4 is 4 (check this by rearranging the numbers in order!) and for 1, 3, 4 and 5, the median is (3+4)/2, that is 3 ½.

This measure is easy to use when you don’t have a calculator, but has the disadvantage that it is less easy to use when testing results which cover estimation and hypothesis testing, as we will do in Chapters 6 and 7.

The mode

The mode is the most frequently occurring value. There is not necessarily only one such value. For example, the figures 1, 2, 2, 3, 5, 9, 9, 11 have two modes: the numbers 2 and 9.

[3] *If you have trouble with this example, go back to Chapter 2 and rework the examples about the use of summation signs.*

Activity

A3.3

Calculate the mean, median and mode for the prices of spreadsheets (using the data below).

Spreadsheet data – Measure of central tendency

Name	**Price (in £)**
Kuma K-Spread II	52
Logistix	64
Sage Planner	65
SuperCalc 3.21	82
VP Planner Plus	105
SuperCalc v4	107
Multiplan IV	110
Borland QUATTRO	115
Legend Twin Level 3	155
SuperCalc v5	195
Plan Perfect	195
Microsoft Excel	239
Lotus 1-2-3	265
Total	1749
Number of observations (n)	13

You should be able to show that the arithmetic mean is 134.54, the median is 110, and the mode is 195 or 110.

(Note: a mode exists at 195 (with 2 values) in this raw data. However, rounding to nearest 10 gives 3 values at 110.)

Generally speaking, the arithmetic mean considers the most information. The median considers less information, but it is consequently less affected by extreme values (outliers). The mode considers even less information and should be used with care (but can be a useful short cut if you are in a hurry and want to get a preliminary idea about a data set).

Measures of dispersion

It is unlikely we will only be interested in the average value of our data, we will want to know how large the spread or dispersion of values is about it.

The most common measures of dispersion are:

- Variance
- Standard deviation
- Mean Absolute Deviation (MAD)
- Range
- Inter-quartile range.

The first three are related to the use of the arithmetic mean.

Variance

Variance is the mean of the squares of the deviation from the mean:

$$\text{Population variance, } \sigma^2 = \frac{1}{n}\sum_{i=1}^{i=n}\left(x_i - \bar{x}\right)^2$$

If we are dealing with data from a sample we divide by (n–1) rather than n and use the symbol s^2, rather than σ^2:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{i=n}\left(x_i - \bar{x}\right)^2$$

If you are using a non-programmable calculator, there is an easy-to-use alternative formula:

$$s^2 = \left(\sum_{i=1}^{i=n} x_i^2 - \frac{\left(\sum_{i=1}^{i=n} x_i\right)^2}{n}\right) \Big/ n-1$$

Activity

A3.4

Work out s^2 for a sample of 9 observations of the number of minutes students took to complete a statistics problem.

Minutes to complete problem: 2, 4, 5, 6, 6, 7, 8, 11, 20

Use both the formulae and make sure you get the same answer in each case.

Standard deviation

The standard deviation, σ for the population or s for the sample is the square root of the variance.

Both the variance and the standard deviation are used, with the arithmetic mean, because they are easy to use in the theoretical problems you will meet in Chapters 6 and 7.

The mean absolute deviation (MAD) uses the absolute values of the deviations from the mean and gives us a more intuitively understandable measure of deviation than variance and standard deviation.

It is written:

$$\text{MAD of } x_i = \frac{1}{n}\sum_{i=1}^{i=n}\left|x_i - \bar{x}\right|$$

Unfortunately for us, it is not so helpful with inference and hypothesis testing!

Range

The range, which looks at the difference between the largest and smallest values, is another easy-to-understand measure, but it will clearly be very affected by a few extreme values. It is written:

Range = max (xi) – min (xi)

but very rarely used.

Inter-quartile range

The inter-quartile range is used when you have measured location using the median. It has the same advantages and disadvantages as the median.

The lower quartile, x_L (or Q_1), is a point such that 25% of observations are less than x_L, and the upper quartile, x_U (or Q_3), is a point such that 75% of observations are less than x_U . The inter-quartile range is $(x_U – x_L)$ (or $Q_3 – Q_1$).

Activity

A3.5

Calculate the range, variance, standard deviation, mean absolute deviation and inter-quartile range of the spreadsheet prices shown below. Check against the answers given after the data.

Spreadsheet data – Measures of Dispersion

Name	Price	Price–Mean	(Price–Mean)^2	∣Price–Mean∣
Kuma K-Spread II	52	–82.54	6812.60	82.54
Logistix	64	–70.54	4975.67	70.54
Sage Planner	65	–69.54	4835.60	69.54
SuperCalc 3.21	82	–52.54	2760.29	52.54
VP Planner Plus	105	–29.54	872.52	29.54
SuperCalc v4	107	–27.54	758.37	27.54
Multiplan IV	110	–24.54	602.14	24.54
Borland QUATTRO	115	19.54	381.75	19.54
Legend Twin Level	3155	20.46	418.67	20.46
SuperCalc v5	195	60.46	3655.60	60.46
Plan Perfect	195	60.46	3655.60	60.46
Microsoft Excel	239	104.46	10912.21	104.46
Lotus 1-2-3	265	130.46	17020.21	130.46
Total		0.00	57661.23	752.62
Total/13	**Variance _ 4435.48**		**57.89 _ MAD**	
	Std.Dev._ 66.60			

Range of data is (265–52) or a price difference of 213 between dearest and cheapest.

			% Observations
Name	**Price**		**Less than price**
Kuma K-Spread II	52		7.69
Logistix	64		15.38
Sage Planner	65		23.08
SuperCalc 3.21	82	_ Lower Quartile	30.77
VP Planner Plus	105		38.46
SuperCalc v4	107		46.15
Multiplan IV	110	_ Median	53.85
Borland QUATTRO	115		61.54
Legend Twin Level	3155		69.23
SuperCalc v5	195	_ Upper Quartile	76.92
Plan Perfect	195		84.62
Microsoft Excel	239		92.31
Lotus 1-2-3	265		100.00

Inter-quartile range is (195–82) or 113.

(Here the value corresponding to the proportion ≥ 25% or 75% is taken as the appropriate quartile. A more detailed approach is discussed in the literature. The more detailed calculation of inter-quartile range and median for grouped data is not required). See Example 3.3 for an acceptable alternative approach using the ungrouped cumulative frequency diagram.

Graphical representations of data

The main representations we use in this subject are **histograms, stem and leaf** diagrams, and **boxplots. We** also use **scatter plots** for two variables. There are many other pretty representations available from software packages, in particular pie charts and standard bar charts. You should look at your text for these. The following example demonstrates the use of the main graphics referred to above and the measures we have been looking at.

Example 3.1

The following are the scores that 200 students obtained in a twelfth-grade achievement test in American History:

623 454 643 585 719 693 571 646 613 655 662
585 580 648 405 506 669 558 577 487 682 565
552 567 745 610 493 571 682 600 740 593 488
520 630 586 610 695 539 490 509 667 597 662
566 597 604 519 643 606 500 460 717 592 752
695 610 620 682 524 552 703 584 550 659 585
578 533 532 708 537 635 591 552 557 599 540
752 726 630 558 646 643 606 682 565 578 488
361 560 630 666 719 669 571 520 571 539 580
629 545 558 544 646 655 585 634 759 532 653
682 641 547 634 609 620 634 585 558 689 780
448 523 571 680 550 544 580 626 617 578 430
662 494 520 760 604 523 484 584 613 696 649
649 578 585 610 641 465 667 578 564 578 539
495 537 558 564 648 673 666 571 487 659 649
675 552 636 580 643 688 620 523 727 502 686
547 481 600 604 573 558 586 597 545 547 601
659 544 507 641 585 630 613 710 509 480 487
526 532

For these data we are asked to find the mean, median, quartiles and standard deviation. We should also make a box plot and a histogram and describe the characteristics of these data in words.

The tally method

With a large data set like this, it is best to tabulate using the **tally** method, and either use the grouped data to find the statistics asked for, or to use the tabulation as a guide as to where to look for the median and the quartiles. Figure 3.1 shows a tabulation of the data such as would result from tallying. The asterisks opposite, for instance, 475.0, show the grades between 475 and 500. The only observation on a class boundary is 675 which is included in the class starting with 675. The median lies in the class starting at 575, and so one can find the two central observations by ordering the grades from 575 onwards. These observations in order are 577, 578, 578, 578, 578, 578, 578, 580, 580, 580, 580, 584, 584, 585, 585, 585, 585, 585, 585, 585, 586, 586, …….

Figure 3.1 Tally of grades

```
350.0    1    *
375.0    0
400.0    1    *
425.0    2    **
450.0    4    ****
475.0   11    ***********
500.0   13    *************
525.0   20    ********************
550.0   27    ***************************
575.0   29    *****************************
600.0   21    *********************
625.0   26    **************************
650.0   16    ****************
675.0   14    **************
700.0    6    ******
725.0    4    ****
750.0    4    ****
775.0    1    *
```

Let us go through this in detail, step by step:

- The 100th and 101st ordered observations are both 586, so this is the median of the sample.
- Similarly the first quartile is between the 50th and 51st ordered observation. This lies in the class starting at 525. The 50th and 51st ordered observations are both 547, so this is the first quartile.
- The third quartile is between the 150th and 151st ordered observation. These lie in the class starting at 625, and are identified as 646 and 648. The observation 646 is the (150–1)/(200–1) = 0.74874 (or 74.87%) quartile, and the observation 648 is the (151–1)/(200–1) = 0.75377 (or 75.38%) quartile. The third quartile is therefore 646+ (0.75 – 0.74874)(648–646)/(0.75377 – 0.74874) = 646.5.
- The mean of the (ungrouped) data is easily calculated to be 595.65.

A stem and leaf diagram

More informative than the tally plot is the stem-and-leaf diagram found in Figure 3.2. Here instead of stars the tens digit (leaf) of the observations is given. The diagram contains more information than the other histogram-type plots. The first figure on the left (the stem) is the hundreds digit. The first line is from the number 361 rounded to the nearest 10 with 5 rounded downwards, the second line comes in the same way from the numbers 405, 430.

Figure 3.2 Stem-and-leaf for grades of 200 students

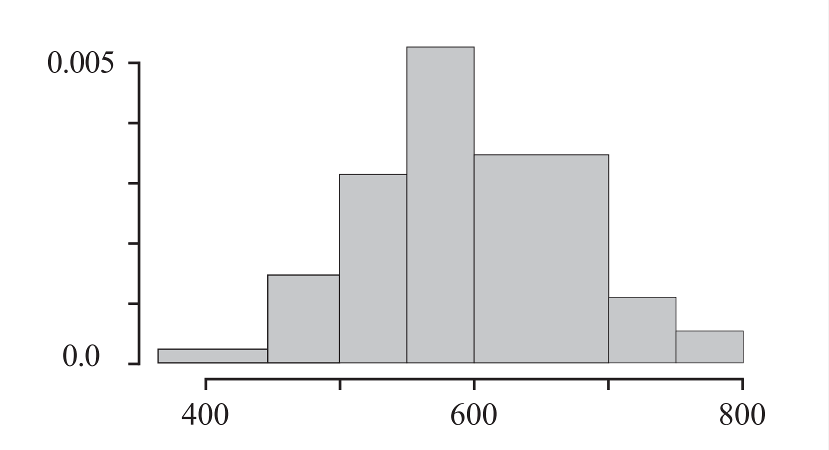
```
Decimal point is 2 places to the right of the colon.
    3   :   6
    4   :   03
    4   :   5566888999999999
    5   :   001111222222233333344444444444
    5   :   5555555555666666666667777777778888888888888888888899999
    6   :   0000000000111111111122222333333333344444444
    6   :   5555555555566666677777777888888999999
    7   :   00112223344
    7   :   55668
```

A histogram

One can compare this with a histogram of the data, deliberately made with unequal class-widths in Figure 4.3. The values along the horizontal axis are the grades, but the vertical axis has a scale measured in units of relative frequency per unit area.

It is the **areas** of the bars in the histogram that are proportional to the frequencies of the corresponding groups of grades. The **heights** of the histogram bars are not proportional to the frequencies of the corresponding groups unless the class-widths are all equal.

Figure 3.3 Histogram of grades of 200 students

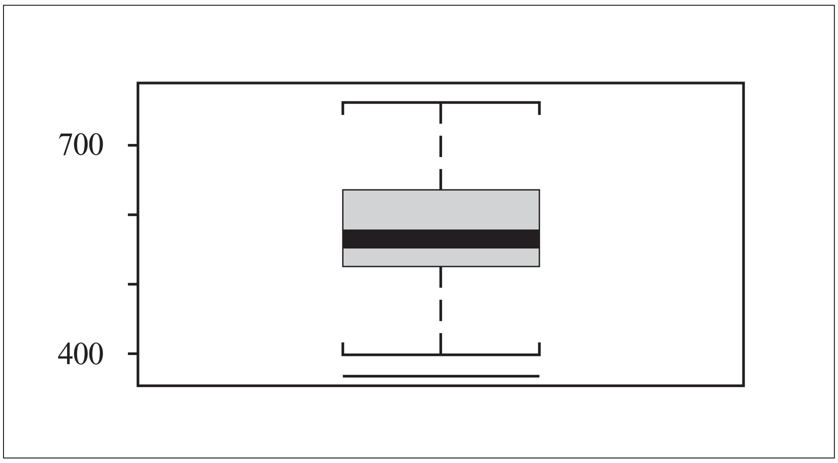


Notice that there are no gaps between the bars of the histogram along the horizontal axis, and that the scale on that axis should allow the upper and lower limits of the histogram to be easily discovered. (Do not make the common error of ignoring the effect of unequal class-widths.)

A box plot

The box plot for the data is shown in Figure 3.4. In a box plot the middle horizontal line is the median and the upper and lower ends of the shaded area are the quartiles. The 'whiskers' are drawn from the quartiles to the observations furthest from the median but not by more than 1.5x the inter-quartile range. Any other points are plotted individually, with the smallest and largest observation marked by horizontal lines.

Figure 3.4 Box plot of grades of 200 students



Calculating the statistics using grouped data

If working by hand, the easiest way to find all the statistics asked for is to use grouped data from a tabulation as in Figure 3.5. Here the numbers *x* are at the centre of the classes, which in the case of the classes right at the lower and upper ends may imply an assumption about the smallest or largest value that needs comment.

Figure 3.5 Calculations for grouped grades given in Figure 3.1

x	f	xf	$x - 597$	$(x - 597)^2$	$(x-597)^2f$
350	1	350	–247	61009	61009
400	1	400	–197	38809	38809
450	5	2250	–147	21609	108045
500	25	12500	–97	9409	235225
550	47	25850	–47	2209	103823
600	50	30000	3	9	450
650	42	27300	53	2809	117978
700	20	14000	103	10609	212180
750	8	6000	153	23409	187272
800	1	800	203	41209	41209
	200	119450			1106000

Mean = 119450/200 = 597.25 = 597 approximately.

Variance = 1106000/199 = 5557.8.

Standard Deviation $= \sqrt{5557.8}$

$= 74.55$

A more accurate value for the standard deviation worked from the ungrouped data without making approximations is 73.21.

Simple interpolation rules are used to find quartiles for grouped data without reference to the full set of data. The median is 575+50(100–79)/50=596, the first quartile is 525+50 (50–32) 47=544, and the third quartile is 625+50 (150–129)/42 = 650. These are all close to the more exact values working from the whole data set.

We can describe these data as having a bell-shaped distribution (as shown in Figure 3.3). To the left the frequencies descend sharply, but then straggle into a long, thin tail. To the right the frequencies decrease more slowly. It is skewed to the right.

Skewed distributions are likely to be found when you tabulate data, so the choice of measure for the centre of the distribution is important. Besides the mean and median there are other measures of where the centre lies. One could average the first and third quartiles, or perhaps take the mean of all observations between the first and third quartiles. The mean of all the observations, which is the measure of the centre used for all elementary work in statistics, is chosen not because it is necessarily the best, but because it is the easiest to work with. It has the serious fault that it is greatly changed by a ‘wild’ observation that is either much too large or much too small. In other words, the mean has little **robustness**.

A second example, introducing cumulative frequencies

Here is a further example of drawing a histogram. This one is with equal intervals. It also introduces the cumulative frequency diagram.

Example 3.2

A stockbroker is interested in the level of trading activity on the New York Stock Exchange. He has collected the following data, which are average daily volumes per week for the first 29 weeks of 1990. Summarise this data in a graphical form.

Average daily volumes per week (millions of shares per day)

1)	172.5	11)	154.6	21)	163.5
2)	161.9	12)	151.6	22)	172.6
3)	172.3	13)	132.4	23)	168.3
4)	181.3	14)	144.0	24)	155.3
5)	169.1	15)	133.6	25)	143.4
6)	155.0	16)	149.0	26)	140.6
7)	148.6	17)	135.8	27)	125.1
8)	159.8	18)	139.9	28)	171.3
9)	161.6	19)	164.4	29)	167.0
10)	153.8	20)	175.6		

(Note that this variate is actually discrete, but because the numbers are so large the variate can be treated as continuous.)

Procedure

Step 1

Decide on class intervals. This is a subjective decision, the objective is to convey information in a useful way. In this case the data lies between (roughly) 120 and 190 million shares/day, so class intervals of 10m will give 7 classes. With 30 observations this choice is probably adequate; more observations might support more classes (a class interval of 5m, say); fewer observations would, perhaps, need a larger interval of 20m.

The class intervals are defined like this:

$120 \leq$ Vol <130, $130 \leq$ Vol <140, etc. or, alternatively [120,130), [130,140) etc.

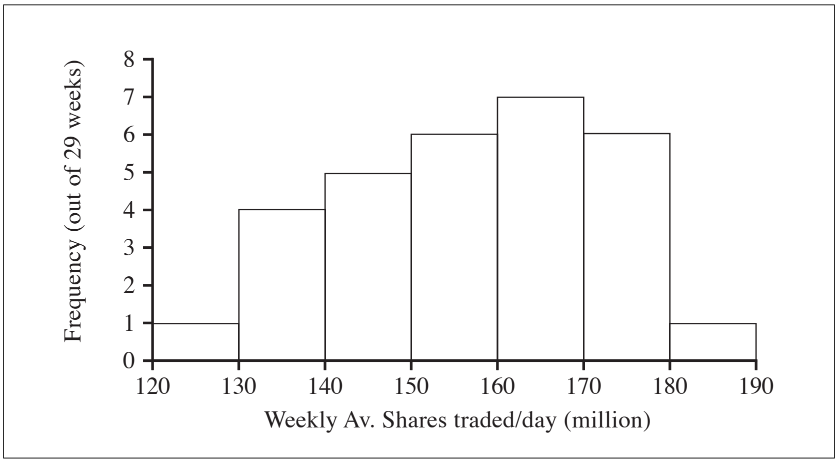
Step 2

Count the observations in each class interval

Interval	[120,130)	[130,140)	[140,150)	[150,160)	[160,170)	[170,180)	[180,190)
No.	1	4	5	6	7	5	1

These results are shown in Figure 3.5 as a histogram.

Figure 3.5 Histogram of trading volume data



Cumulative frequency diagram

A development from the histogram is the cumulative frequency diagram. The **relative frequency** (r.f) in an interval is given by:

$$\text{r.f.} = \frac{\text{Number of observations inside interval}}{\text{Total number of observations}}$$

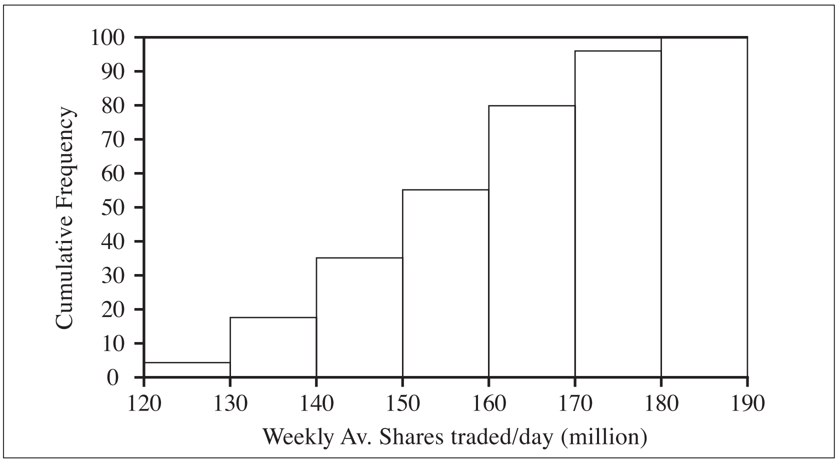
For example in the interval [150,160) the r.f. is (6/29) = 0.207 or 20.7%.

A **cumulative frequency diagram** plots the relative frequency of all observations less than a given point.

Limit point:	120	130	140	150	160	170	180	190
Cumulative frequency:	0	1	5	10	16	23	28	29
% Cumulative relative frequency:	0	3.4	17.2	34.5	55.2	79.3	96.6	100

The lower and upper quartiles and median can be found by linear extrapolation using the ungrouped cumulative frequency diagram. For the trading volume data, x_L is 143, x_M is 155, and x_U is 168. The cumulative frequency diagram is shown in Figure 3.6. Note that a later figure (Figure 3.7) will show the data in an **ungrouped** cumulative frequency diagram.

Figure 3.6 Grouped cumulative frequency diagram



This brings us to a further modification for calculating grouped data (introduced when we calculated the mean). In some circumstances, one only has access to grouped data, for example in the form of a histogram. In this case, the mean and variance can be found by slightly modifying the formulae previously seen.

Consider K classes. In the k$^{\text{th}}$ class there are f_k observations and the midpoint is denoted by x_k. The we can say:

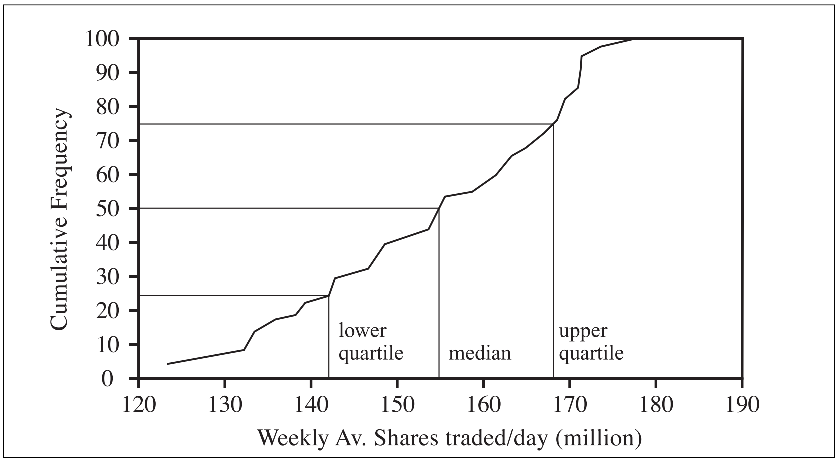
$$\text{Mean} = \bar{X} = \frac{\sum_{k=1}^{K} f_k x_k}{\sum_{k=1} f_k}$$

$$\text{Variance} = \frac{\sum_{k=1}^{K} f_k (x_k - \bar{x})^2}{\sum_{k=1}^{K} f_k} = \frac{\sum_{k=1}^{K} f_k x_k^2}{\sum_{k=1}^{K} f_k} - \left(\frac{\sum_{k=1}^{K} f_k x_k}{\sum_{k=1}^{K} f_k} \right)^2$$

These formulae are used in the following example (Example 3.3).

Example 3.3 Using trading volume data

Figure 3.7 Ungrouped cumulative frequency diagram



There are K=7 classes

K_k	f_k	x_k	f_k/x_k	f_k/x_k^2
1	1	125	125	15625
2	4	135	540	72900
3	5	145	725	105125
4	6	155	930	144150
5	7	165	1155	190575
6	5	175	875	153125
7	1	185	185	34225

The mean is (4535/29) = 156.4 (The ungrouped mean is 156.0)

The variance is (715725/29) – (156.4)2 = 219.2.

This means the standard deviation is 14.8 (that is, the $\left(\sqrt{219.2}\right)$ **.**
(The ungrouped standard deviation is 14.7).

Using a pie chart

Now, how about summarising the data from Example 4.1 as a pie chart?

Activities

A3.6

Make a pie chart of the data on the 200 students grades in a twelfth grade American History test. Use the groupings ‘Under 500’, ‘500 to under 600’, ‘600 to under 700’ and ‘700 and over’. (Figure 3.1 or 3.2 can help you with this). Be careful to label your diagram correctly.

Note that an effective pie chart should not present more than 4 or 5 categories – beyond that, eye and mind become confused, and the whole idea of clarifying your material by using a diagram is lost.

A3.7

Think about why and when you would use each of the following:

- bar chart
- histogram
- stem and leaf
- pie chart.

When would you **not** do so?

Presentational traps

You should be aware when reading articles within newspapers, magazines and, let's admit it, even within academic journals, that it is easy to mislead the reader (either accidentally or deliberately) by careless or poorly defined diagrams. It is an unfortunate practice that bar charts are shown disembowelled (i.e. either with a non-zero start on the '*y*-axis' or with the middle portion of the diagram cut out). See the book by Huff for many more examples. Other 'tricks' are to use pictures when the height of the picture is the key feature but other aspects are attracted by the eye.

Activity

A3.8

Find data presentation diagrams in newspapers or magazines that might be construed as misleading.

Summary

This chapter, although in practice concerned with what you can do with data after you have collected it, serves as a useful introduction to the whole course. It highlights some of the problems with handling data and, furthermore, has introduced many of the fundamental concepts such as mean, variance, discrete and continuous data, etc.

Learning outcomes

After completing this chapter and the relevant reading, you should be able to:

- Calculate the following:
  - arithmetic mean
  - standard deviation
  - variance
  - median
  - quartile
  - range
  - inter-quartile range, and
  - mode.
- Explain the use and limitations of the above quantities.
- Draw and interpret:
  - histograms
  - stem-and-leaf diagrams
  - box plots
  - pie charts, and
  - cumulative frequency distributions.
- Use labels and titles correctly in your diagrams and give the units you have used.

In summary, you should be able to use appropriate measures and diagrams in order to explain and clarify data you have collected or which are presented to you.

At this stage, you need not:

- study skewness or kurtosis
- make detailed estimates of grouped data medians and quartiles; it will be enough to state the median or quartile class or group in examples given (note that the examples given in the text are more detailed to help you understand the process of calculation, you need not be so detailed).

Sample examination questions

1. The data below show the numbers of daily phone calls received by an office supplies company over a 25-hour working day period.

   219 541 58 7 13 476 418 177 175 455 258 312

   164 314 336 121 77 183 133 78 291 138 244 36

   48

   a) Construct a stem and leaf diagram for these data and use this to find the median of the data. (6 marks)

   b) Find the quartiles of the data. (3 marks)

   c) Would you expect the mean to be similar to the median? Explain. (2 marks)

   d) Comment on your figures. (3 marks)

2. Say whether the following statement is true or false and briefly give your reasons. ‘The mean of a data set is always greater than the median.’ (2 marks)

3. For $x_1 = 4$, $x_2 = 1$, $x_3 = 2$, give:

   a) the median

   b) the mean. (2 marks)

4. Briefly state, with reasons, the type of chart which would best convey the data in each of the following:

   a) A country’s total import of wine, by source.

   b) Students in higher education classified by age.

   c) Numbers of students registered for secondary school in years 1998, 1999 and 2000 for areas A, B and C of a country. (6 marks)

5. If $n = 4$ and $x_1 = 1$, $x_2 = 4$, $x_3 = 5$ and $x_4 = 6$, find $\sum_{i=2}^{i=4} x_i \big/ 3$. Why might you use this figure to estimate the mean? (3 marks)

6. State whether the following statement is TRUE or FALSE and briefly give your reasons.

   ‘Three-quarters of the observations in a data set are less than the upper quartile.’ (2 marks)

7. If $x_1 = 4$, $x_2 = 2$, $x_3 = 2$, $x_4 = 5$ and $x_5 = 6$, determine:

   a) the mode

   b) the mean. (2 marks)

Notes

Chapter 4

Probability

Essential reading

Newbold, P. *Statistics for Business and Economics.* (London: Prentice-Hall, 1995) fourth edition [ISBN 0 1385 5549 0]. Chapter 3.

Further reading

Aczel, A.D. *Complete Business Statistics.* (London: Irwin/McGraw Hill, 1999) [ISBN 0 0728 9302 8]. Sections 2.1 to 2.6.

Anderson, D.R., D.J. Sweeney, and T.A. Williams. *Statistics for Business and Economics.* (Cincinatti: South-Western Thomson Learning, 2002) eighth edition [ISBN 0 3240 6671 6]. Sections 4.1 to 4.4.

Hanke, J.E. and A.G. Reitsch. *Understanding Business Statistics.* (Burr Ridge Ill: Irwin, 1994) second edition [ISBN 0 2561 1219 3]. The first half of Chapter 5.

Mason, R.D. and D.A. Lind. *Statistical Techniques in Business and Economics.* (Boston: McGraw Hill, 2001) eleventh edition [ISBN 0 0724 0282 2]. The first half of Chapter 5.

Moskowitz H. and G.P. Wright. *Statistics for Management and Economics.* (London: Charles Merrill Publishers, 1985) [ISBN 0 6752 0211 6]. Sections 3.1 to 3.5.

Wonnacott, T.H. and R.J. Wonnacott. *Introductory Statistics.* (Chichester: Wiley, 1990) fifth edition [ISBN 0 4715 1733 X]. Sections 3.1 to 3.3.

Introduction

Chance is what makes life worth living – if everything was known in advance, imagine the disappointment! If decision-makers had perfect information about the future as well as the present and the past, there would be no need to consider the concepts of probability. However, it is usually the case that uncertainty cannot be eliminated and hence its presence should be recognised and used in the process of decision- making.

Information about uncertainty is often available to the decision-maker in the form of probabilities. This chapter introduces the fundamental concepts of probability. In other subjects (e.g. *Management Science Methods*) you may make full use of probabilities in decision trees and highlight ways in which such information can be used.

Our treatment of probability in this module is quite superficial. The concepts of probability, as you will see, are simple but applying them in some circumstances can be very difficult! As a preliminary we consider the basic ideas concerning sets.

Set theory: random experiments, sample spaces and events

This topic is returned to, and made more substantial use of, in the *Statistics 2* and *Management Mathematics* courses.

Random experiment

- **Sample Space**, S. For a given experiment the sample space, S, is the set of all possible outcomes.
- **Event**, E. This is a subset of S. If an event E occurs, the outcome of the experiment is contained in E.

Example 4.1

a) When tossing a coin we might have the following sets/events:

S = { H, T } E = { H } or E = { T }

(Note: H is the event a head appears, T a tail)

b) When throwing a die:

S = {1,2,3,4,5,6} E = {3,4} F = {4,5,6 }

Example 4.2

Suppose you arrive at a railway station at a random time. There is a train once an hour. The random experiment is to observe the number of (rounded up) minutes that you wait before a train leaves. The elementary outcomes here are the integers (whole numbers) 1 to 60, and the sample space is {1,2,3…60}. The event that ‘you wait less than 10 minutes’ is the subset {1,2,3,4,5,6,7,8,9}.

Example 4.3

A population of interest has four members: Mabel, Belmont, Gertrude and Elsie. A random experiment selects a sample of size two from the population without replacement. The sample space is:

S = {(Mabel, Belmont), (Mabel, Gertrude), (Mabel, Elsie),
(Belmont, Gertrude), (Belmont, Elsie), (Gertrude, Elsie)}.

The event that ‘the sample includes Gertrude’ is the subset:

{(Mabel, Gertrude), (Belmont, Gertrude}, (Gertrude, Elsie)}.

This example shows that the elementary outcomes can themselves be sets.

Some rules and symbols

- **Union**. We write $E \cup F$ to mean the union of E and F. This set consisting of outcomes that belong to at least one of E or F. » is equivalent to ‘either or both’ in English.

  If you look at Example 4.1(b) again, throwing a die, you will see that $E \cup F = \{3,4,5,6\}$.

- **Intersection** – ($E \cap F$ or E.F) We write $E \cap F$ to mean the intersection of E and F. This set consisting of outcomes belonging to E and F. $\cap$ is equivalent to **and** in English.

  Returning to Example 4.1(b), $E \cap F = \{4\}$.

- **Complement**. We write the complement of E as E^c. It indicates all the elements of a set not in event E.

  Looking at Example 4.1(b) again, throwing a die, you can see that E^c is = {1,2,5,6}.

Activity

A4.1

In Example 4.1(b), give

a) F^c.

b) $E^c \cap F^c$

c) $(E \cup F)^c$

d) $E^c \cap F$

Definitions of probability

We now come to the definitions we use in probability theory.

A priori

Assuming all outcomes of an experiment are equally likely then:

Probability of event A = number of outcomes ÷ Total number of outcomes

$= N_A/N$ say.

We write this as P(A).

The table below contains information about the punctuality of the delivery of 300 orders made by three different suppliers:

Supplier		**Delivery time**		
	Early	On time	Late	Total
Jones	20	20	10	50
Smith	10	90	50	150
Robinson	0	10	90	100
Total	30	120	150	300

Suppose we want to know the probability that an order chosen at random is late, given that it came from Jones. This can be approached using the **a priori** method for calculating probabilities. The number of orders coming from Jones is 50 so this is our **Total number of outcomes**. Of these, the number that were late was 10 so this is the **Number of outcomes where A occurs**. Hence the required probability is (10/50) = 0.2.

Activity

A4.2

What are the probabilities associated with a delivery chosen at random for each of the following?

a) Being an early delivery.

b) Being a delivery from Smith.

c) Being both from Jones and late.

Answers

a) Of the total equally likely outcomes (300) there are 30 that are early. Hence required probability is 30/300 = 0.1.

b) Again, of the total equally likely outcomes (300) there are 150 from Smith. Hence required probability is 150/300 = 0.5.

c) Now, of the 300, there are only 10 that are late and from Jones. Hence the probability is 10/300 (or 0.033).

Frequency

There are two common ways of thinking about probability:

- An experiment is repeated many times. The probability of an event is the limiting value of the proportion of experiments in which the event occurs, as the number of experiments tends to infinity.
- The probability of event A which we write $P(A) = (N_A/N)$ as N gets larger and larger, where N_A is the number of experiments where A occurs.

Whichever of these two ways we come to understand probability, we use the following axioms.

Axioms of probability

Given E and F that are events in a sample space S, and that ø is the empty set, the axioms are:

$P(E) \geq 0$ for all events E (probabilities are non–negative) (a)

$P(S) = 1$; in other words the probability of the sample space is 1 (so all the probabilities must be less than or equal to 1) (b)

$P(E \cup F) = P(E) + P(F)$, if $E \cap F = ø$; in other words E and F are mutually exclusive: they cannot occur at the same time. This is the **addition rule for mutually exclusive events**. (c)

Note that, if $E \cap F \neq ø$, then it can be easily shown that
$P(E \cup F) = P(E) + P(F) - P(E \cap F)$

Other useful relationships

Consider two events $E \cap F$ and $E^c \cap F$. Their union is F and their intersection is empty. Thus, using the third axiom (c):

$P(F) = P[(E \cap F) \cup (E^c \cap F)] = P(E \cap F) + P(Ec \cap F)$

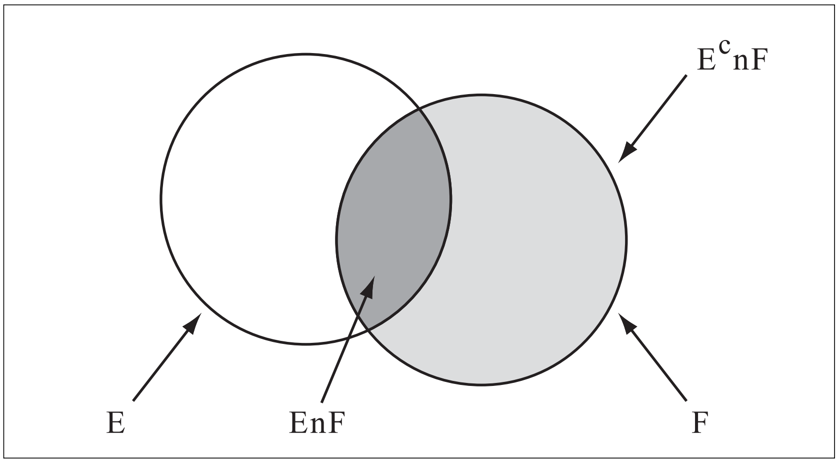
Rearranging, $P(E^c \cap F) = P(F) - P(E \cap F)$, and putting $F = S$, gives:

$P(E^c) = P(S) - P(E) = 1 - P(E)$

In other words, the probability of event E not occurring is (1– probability of event E occurring).

Note that this is also true in reverse: the probability of event E occurring is 1 minus the probability that E does not occur. This can be most useful when it is easier to calculate $(P(E^c)$ and we can find P(E) indirectly. We can draw this using a Venn diagram.

Figure 4.1 Venn diagram showing two intersecting sets



Two sets E and F that are not mutually exclusive (i.e. their intersection is non-empty) can be rearranged into 2 events which are mutually exclusive, namely E and $E^c \cap F$, where:

$(E \cup F) = E \cup (F \cap E^c)$ and $E \cap \ (F \cap E^c) \ = \_$

So, for example $P(E \cup F) = P(E) + P(F \cap E^c)$.

Activity

A4.2

Draw the appropriate Venn diagram to show each of the following in connection with Example 4.1(b):

a) $E \cup F = \{3,4,5,6\}$

b) $E \cap F = \{4\}$

c) $E^c = \{1,2,5,6\}$.

Conditional probability and independent events

You should also know a little about **conditional probability** and **independent events**.

The conditional probability P (A|B) is the probability that A happens given that B has already happened.

If P (A|B) = P (A), we say that A and B are **independent**. (d)

Looking more closely at independent events, you need be careful! Note that:

- P (B) must be greater than zero for this to work
- P (A|B) ≠ P (B|A).

Think about this second point a little. Clearly the probability that you have spots given that you have measles is not the same as the probability that you have measles given that you have spots!

Another way you can tell if A and B are **independent** is that

$P (A \cap B) \ = P (A) \ P (B)$ (e)

and if this is not the case they are not independent.

Activity

A4.3

There are three sites a company may move to: A, B, C. We are told that P(A) (the probability of a move to A) is ½, and P(B) is ⅓. What is P(C)? (Use the information given in the section (b) under **axioms of probability**).

A4.4

Two events A and B are independent with probability (⅓) and (¼) respectively.

What is $P (A \cap B)$?

(Use the information given in (e) at the end of the paragraph above).

Probability trees

The setting out of solutions to problems requiring the manipulation of the probabilities of mutually exclusive and independent events can sometimes be helped by the use of probability tree diagrams. These have useful applications in decision theory.

The best choice of probability tree structure often depends upon the question and the natural order in which events like A and B above occur.

The probability of emerging at the end of any path through the diagram is found by multiplying the probabilities on the branches on the path. Simple tree diagrams are given and clearly explained in Newbold and all sample texts in this area.

Activities

A4.5

A company gets 60% of its supply of a part from manufacturer A, the remainder from manufacturer Z. The quality of the parts delivered is given below:

Manufacturer	**% Good Parts**	**% Bad Parts**
A	97	3
Z	93	7

a) The probabilities of receiving good or bad parts can be represented by a probability tree. Show for example that the probability that a randomly chosen part comes from A and is bad is 0.018.

b) Note also that the sum of the probabilities of all the outcomes is 1.

c) The way the tree is used depends on the information required. For example, the tree can be used to show that the probability of receiving a bad part is 0.028 + 0.018 = 0.046.

A4.6

(Using Set theory and Laws of probability or a probability tree)

A company has a security system comprising four electronic devices (A,B,C and D) which operate independently. Each device has a probability of 0.1 of failure. The four electronic devices are arranged so that the whole system operates if at least one of A or B functions and at least one of C or D functions.

Show that the probability that the whole system functions properly is 0.9801.

Summary

This chapter has introduced the idea of probability, and defined the key terms. You have also seen how Venn diagrams can be used to illustrate probability, and used the three axioms. This should prepare you for the following chapter.

Learning outcomes

After completing this chapter and the relevant reading, you should be able to:

- use the ideas and notation involved in set theory for simple examples
- use the basic axioms of probability
- understand the ideas of conditional and independent probability
- draw and use appropriate Venn diagrams
- draw and use appropriate probability trees.

You do not need to be able to prove Baye’s theorem or use permutations and combinations.

Sample examination questions

1. Say whether the following statement is **true** or **false** and briefly give your reasons.

   'If two events are independent, they must be mutually exclusive.'

   (2 marks)

2. If X can take values of 1, 2 and 4 with P(X=1) = 0.3, P(X=2) = 0.5, P(X=4) = 0.2, what are:

   a) $P(X^2 \leq 4)$?

   b) P(X > 2|X is an even number)? (4 marks)

3. Write down and illustrate the use in probability of:

   a) the addition rule

   b) the multiplication rule. (4 marks)

4. A student can enter a course either as a beginner (73%) or as a transferring student (27%). It is found that 62% of beginners eventually graduate, and that 78% of transfers eventually graduate. Find:

   a) the probability that a randomly chosen student is a beginner who will eventually graduate (2 marks)

   b) the probability that a randomly chosen student will eventually graduate (2 marks)

   c) the probability that a randomly chosen student is either a beginner or will eventually graduate, or both. (2 marks)

   Are the events 'Eventually graduates' and 'Enters as a transferring student' statistically independent?

   If a student eventually graduates, what is the probability that the student entered as a transferring student?

   If two entering students are chosen at random, what is the probability that not only do they enter in the same way but that they also both graduate or both fail? (6 marks)

5. A coffee machine may be defective because it dispenses the wrong amount of coffee (C) and/or it dispenses the wrong amount of sugar (S).

   The probabilities of these defects are:

   P (C) = 0.05, p (S) = 0.04, p (C and S) = 0.01

   What proportions of cups of coffee have:

   a) at least one defect

   b) no defects? (4 marks)

Notes

Chapter 5

The Normal distribution and ideas of sampling

Essential reading

Newbold, P. *Statistics for Business and Economics.* (London: Prentice-Hall, 1995) fourth edition [ISBN 0 1385 5549 0]. Chapters 4.1– 4.4, 5.5 and 6.

Further reading

Wonnacott, T.H. and R.J. Wonnacott. *Introductory Statistics.* (Chichester: Wiley, 1990) fifth edition [ISBN 0 4715 1733 X]. Chapters 4 and 6.

Introduction

So far you have studied basic measures and ways of representing data in diagrams (Chapter 3) and probability theory (Chapter 4). This chapter brings the two together and introduces you to:

- the idea of the random variable
- the idea of the distribution of a random variable and, in particular,
  - – the Normal distribution
  - – sampling distributions.

Random variable

You should read Newbold 4.1 carefully to understand what a random variable is. The use of randomness in sampling and applications to survey methodology is also discussed in Chapter 8 of this guide.

Normal distribution

The Normal distribution, fortunately for statisticians, happens to be mathematically easy to handle. It also describes many natural phenomena – ranging from the height of adult males, for example, to the dimensions of manufactured items. It has the further characteristic that, even when the population distribution is not normal, the means taken from successive samples are distributed normally.

The Normal (or Gaussian) random variable

On the face of it, the Normal distribution does not look simple. But, in fact, if you look closely at the formula, you will see that its shape is completely determined by its mean μ and variance σ^2. We say that a Normal random variable with mean μ and variance σ^2 has a $N\ (\mu,\ \sigma^2)$ distribution. It can be written as:

$$y = \frac{1}{\sqrt{2\pi}\,\sigma}\ \exp\text{-}\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The cumulative distribution function for an *N* (0,1) random variable (standardised normal) is given in Table 4 of the *New Cambridge Statistical Tables* and can be used to find the probabilities for different values under the curve using the fact that if *X* has an *N* (μ, σ^2) distribution, then $\frac{X - \mu}{\sigma}$ has an *N* (0,1) distribution.

You will find all this easier to understand when you have worked through the following examples. The first is straightforward.

Example 5.1

> Suppose that a population of men's heights is normally distributed with a mean of 68 inches, and standard deviation of 3 inches. Find the proportion of men who are:
>
> a) under 66 inches
>
> b) over 72 inches
>
> c) between 66 and 72 inches.

Answer

The two cut-off values are 66 and 72. Converted to standard units these are:

(66 – 68)/3 = –2/3 and (72 – 68)/3 = 4/3.

The right-hand tail probability for $z = 4/3 = 1.33$ is found from Table 4 of the *New Cambridge Statistical Tables.* For $x = 1.33$, the probability is given by $1 - \Phi\ (1.33) = 1 - 0.9082 = 0.092$.

The left-hand tail probability for $z = -2/3 = -0.66$. From Table 4 with $x = 0.66$ the probability is $1 - \Phi\,(0.66) = 1 - 0.7454 = 0.255$[5].

So the answers are:

a) 25.5%

b) 9.2%

c) 100 – 25.5 – 9.2 = 65.3%.

Try to draw the diagram as in Newbold to make sure you understand this.

The second example will make you think a little more!

Example 5.2

> Two statisticians disagree about the distribution of IQ scores for a population under study. Both agree that the distribution is normal, and that the standard deviation is 15, but *A* says that 5% of the population have IQ scores greater than 134.6735, whereas *B* says that 10% of the population have IQ scores greater than 109.224. What is the difference between the mean IQ score as assessed by *A* and that as assessed by *B*?

[5] *Note that* $\phi\ (-\mathrm{k}) = 1 - \phi(\mathrm{k})$

Answer

The standardised *z* value giving 5% in the upper tail is from Table 5 with[6] *P*=5 given by *P(x)* = 1.6449. For 10% the standardised *z* value is given for *P* = 10 by x (*P*) = 1.2816. So converting to the scale for IQ scores, the values are:

1.6449 15 = 24,6735

1.2816 15 = 19,224 .

Write the means according to *A* and *B* as μ_A and μ_B respectively. Then:

$\mu_A + 24.6735 = 134.6735$,

so:

$\mu_A = 110$,

whereas:

$\mu_B + 19.224 = 109.224$

so $\mu_B = 90$.

The difference $\mu_A - \mu_B = 110 - 90 = 20$.

Draw this to make sure you understand it and then try the following activities for yourself.

Activities

A5.1

Check the following using your Normal Table.

If $Z \sim N(0,1)$ then:

Prob. ($Z \geq 1$) = 1 – $\Phi(1)$ = 0.1587

Prob. ($Z \leq 1$) = Φ (1) = 1–01587 = 0.8413

A5.2

Check that (approximately):

a) 68% of Normal random variables fall within 1 standard deviation of the mean

b) 95% of Normal random variables fall within 2 standard deviation of the mean

c) 99% of Normal random variables fall within 3 standard deviation of the mean

Draw the areas concerned on a Normal distribution.

Make sure you make yourselves completely familiar with this kind of work. It leads directly into your work on confidence intervals in the next chapter and, ultimately, to ideas of hypothesis testing in Chapters 7 and 8.

Expected values

Before we continue with this topic, we need to be introduced two new measures which use the summation notation we first met in Chapter 1:

- E (X) – called the **expected value** of X, and defined as:

$$E(X) = \sum_{i=1}^{i=n} X_i P(X_i)$$

- V (X) – called the **variance** of X or σ_X^2, and defined as:

$$V(X) = E[X - E(X)]^2$$

To find out more, look through Sections 4.3 and 4.4 of Newbold. We will use these measures later in this chapter.

[6] *Note that P is already given as a percentage.*

Sampling distributions

We now start on the work that makes statistics a distinct subject. The idea of sampling, and of a sampling distribution for a statistic like the mean, must be understood by all users of statistics. Many students find there is some difficulty in understanding what is meant by a sampling distribution. Students should do a few sampling experiments themselves to get a good intuitive feel for sampling distributions. For those fortunate enough to have access to a computer, there is software available to help[7].

An example of a distribution for sampling without replacement

For the simplest example of a sampling distribution we go back to some material from Chapter 4. We have already seen in Example 4.3 how to describe in a formal way the sample space for a random sample without replacement, of size 2 from a population of size 4. The four members of the population were Mabel, Belmont, Gertrude and Elsie, and the sample space:

Ω = {(Mabel, Belmont), (Mabel, Gertrude), (Mabel, Elsie),
(Belmont, Gertrude), (Belmont, Elsie), (Gertrude, Elsie)}.

We shall suppose that the sampling is random, so that each of these 6 samples is equally likely and so has probability 1/6. Now let us define the random variable X to have as its values the number of children in the families of Mabel, Belmont, Gertrude and Elsie. On investigation we discover:

X (Mabel)	=	1
X (Belmont)	=	2
X (Gertrude)	=	3
X (Elsie)	=	3.

The six samples in Ω give six equally likely pairs of values for X (some of which turn out the same):

(1,2), (1,3), (1,3), (2,3) (2,3), (3.3).

In fact there are repeated pairs, so this reduces to:

(1,2) with probability 1/6
(1,3) with probability 2/6
(2,3) with probability 2/6
(3,3) with probability 1/6.

Now suppose we look at the distribution of the mean number of children per family in the sample, say $\overline{X}$. The results above lead to:

$\overline{X}$ = 1.5 with probability 1/6
$\overline{X}$ = 2.0 with probability 2/6
$\overline{X}$ = 2.5 with probability 2/6
$\overline{X}$ = 3.0 with probability 1/6.

This last result is a detailed description of the sampling distribution of the mean number of children in a family of size 2 sampled at random without replacement from the original population of 4 parents.

We can now describe the sampling distribution in the usual ways. The **expected value** of $\overline{X}$ is:

$E(\overline{X})$ = 1.5 x 1/6 + 2.0 x 2/6 + 2.5 x 2/6 + 3.0 x 1/6 = 13.5/6 = 2.25

[7] *See for instance Boane, D.P., K. Mathieson and E.L. Tracy,* Visual Statistics. *(Irwin McGraw-Hill, 1997) [ISBN 0256-20876-X].*

The **expected value** of $\overline{X}^2$ is:

$E[\overline{X}^2] = 2.25 \times 1/6 + 4.0 \text{ x } 2/6 + 6.25 \text{ x } 2/6 + 9.0 \times 1/6 = 31.75/6 = 5.292.$

The variance of $\overline{X}$ is therefore

$E - (\overline{X}^2) - [E(\overline{X})]^2$

i.e. $5.292 - 2.25^2$

$= 5.292 - 5.063$

$= 0.229.$

An example of a distribution for sampling with replacement

We can continue with the previous example, but consider sampling with replacement. The sample space is now larger:

Ω = {(Mabel, Belmont), (Mabel, Gertrude), (Mabel, Elsie), (Belmont, Gertrude),
(Belmont, Elsie), (Gertrude, Elsie), (Mabel, Mabel), (Belmont, Belmont),
(Elsie, Elsie), (Gertrude, Gertrude)}.

The last four samples are each half as likely as any one of the first six samples. Think about this carefully. The reason for this is that we take the order of picking our sample into account. The sample made up of Mabel & Belmont could have been picked by choosing Mabel first and then Belmont or the other way round, making two possible samples which include each of them. Where the sample includes one person twice, this is a unique event. In other words there are 16 possible samples. So the first six samples have each a probability 2/16, and the last four have each a probability 1/16. The 10 samples give rise to pairs of values for *X*:

(1,2), (1,3), (1,3), (2,3), (2,3), (3,3), (1,1), (2,2), (3,3), (3,3).

Again there are repeated pairs, so this reduces to:

(1,1) with probability 1/16

(1,2) with probability 2/16

(1,3) with probability 4/16

(2,2) with probability 1/16

(2,3) with probability 4/16

(3,3) with probability 4/16.

We look again at the sampling distribution of:

$\overline{X}$ = 1.0 with probability 1/16

$\overline{X}$ = 1.5 with probability 2/16

$\overline{X}$ = 2.0 with probability 4/16

$\overline{X}$ = 2.0 with probability 1/16

$\overline{X}$ = 2.5 with probability 4/16

$\overline{X}$ = 3.0 with probability 4/16.

Simplifying gives:

$\overline{X}$ = 1.0 with probability 1/16

$\overline{X}$ = 1.5 with probability 2/16

$\overline{X}$ = 2.0 with probability 5/16

$\overline{X}$ = 2.5 with probability 4/16

$\overline{X}$ = 3.0 with probability 4/16.

This is a complete description of the sampling distribution for $\overline{X}$ for samples of size two taken with replacement. Notice that the distribution is different from that in the previous example. The expected value here is:

$$E(\overline{X}) = 1.0 \times 1/16 + 1.5 \times 2/16 + 2.0 \times 5/16 + 2.5 \times 4/16 + 3.0 \times 4/16.$$

So:

$$E(\overline{X}) = 36/16 = 9/4 = 2.25.$$

Also:

$$\begin{aligned} E(\overline{X}^2) &= 1.0^2 \times 1/16 + 1.5^2 \times 2/16 + 2.0^2 \times 5/16 + 2.5^2 \times 4/16 + 3.0^2 \times 4/16 \\ &= \{1 + 4.5 + 20 + 25 + 36\}/16 = 86.5/16 = 5.40625. \end{aligned}$$

The variance of $\overline{X}$ is:

$$V(\overline{X}) = 5.40625 - 5.06250 = 0.34375.$$

Mean and variance of a sample mean

Note: You may omit this section on first reading and go straight to Activity 5.3. If you cannot or do not wish to work with E($\overline{X}$) and V($\overline{X}$), you can leave it at that! This section is to help you understand, but will not be examined in this way.

It would be very tedious to have to work out sampling distribution in detail as in the examples above. Fortunately there are general results that tell us about sampling distributions. First we can think about $E(\overline{X})$.

The sampling distribution of the sample mean in random sample taken either with or without replacement has a mean μ equal to the population mean. That is

$$E(\overline{X}) = E(X) = \mu$$

Going back to the two previous examples, the mean value of X (the number of children per family) in the population is [1 + 2 + 3 + 3]/4, or 9/4. This agrees with the results of the two examples.

The variance of the sample distribution of the sample mean in random samples of size *n* taken **without** replacement from a population of size *N* is given by:

$$V(\overline{X}) = \frac{\sigma^2}{n}\frac{(N-n)}{(N-1)}$$

where σ^2 is the population variance. The variance of the sampling distribution of the sample mean in random samples of size *n* taken **with** replacement from a population of size *N* is given by:

$$V(\overline{X}) = \frac{\sigma^2}{n}$$

Notice that for samples of size greater than 1, the second variance is greater than the first. The most important consequence of our results is that, for sample size greater than 1, the variance of the sample mean is less than the variance of a single observation. This implies that we can get a better idea of the population mean from a sample mean than from a single observation.

Often, the above result $V(\overline{X}) = \frac{\sigma^2}{n}$ is given in the form:

standard deviation $\overline{X} = \frac{\sigma}{\sqrt{n}}$

and the special name **standard error** is given to this standard deviation of a sample mean.

Continuing the previous example, the population variance is:

$$[(1 - 2.25)^2 + (2 - 2.25)^2 + (3 - 2.25)^2 + (3 - 2.25)^2] /4$$

which is:

$$[1.25^2 + 0.25^2 + 0.75^2 + 0.75^2]/4 = 2.75/4 = 0.6875.$$

For samples of size $n = 2$ without replacement from a population of size $N = 4$, the variance of the sample mean from (6.4.2) is:

$$V(\overline{X}) = \frac{0.6875}{2}\frac{(4-2)}{(4-1)} = \frac{0.6875}{3} = 0.229.$$

This agrees with the previous result. Similarly for sampling with replacement, the formula defining $V(\overline{X})$ gives:

$$V(\overline{X}) = \frac{0.6875}{2} = 0.34375.$$

This agrees with the direct calculation given above.

It is easy to generalise the results in the various examples from first principles. We will not do so here, but you should note that the proof applies to sampling with replacement. This area will be covered more fully in *Statistics 2*.

The Central Limit Theorem

In the introductory remarks to this chapter, we noted that the normal distribution could be used for samples whose parent (population) distribution was not normal. The **Central Limit Theorem** justifies this. It states roughly that for a large enough sample size *n*, the sampling distribution of the sample mean $\overline{X}$ from a random sample of size *n* with replacement from a population of values for *X* is close to the normal distribution $N\,(\mu,\sigma^2/n)$, where the population values of *X* have mean μ and variance σ^2.

This is an approximate version of the precise result for samples from $N\,(\mu,\sigma^2)$, but holding much more generally, for the population is not restricted to be normal. To be more precise, the theorem says that, for each fixed value *x*:

$$P\left[\frac{(\overline{X}-\mu)}{\frac{\sigma}{\sqrt{n}}} \le x\right] \to \Phi(x)$$

as $n \to \infty$.

Some the examples above may seem a little much to cope with on first reading. Try returning to them after you have worked on Chapters 6, 7 and 8. In the meanwhile, here is a simple activity which may help.

Activity

A5.3

The following six observations give the time taken to complete a 100 metre sprint by six individuals:

A	15
B	14
C	10
D	12
E	20
F	15

a) Find μ, the mean, for the population and the S.D. of the X_i.

b) Find the $\overline{X}$ for each possible sample:

- of **two** individuals
- of **three** individuals, and
- **four** individuals.

Work out the mean for each set of samples (it must come to μ) and compare the S.D. of the means of the $\overline{X}$ about μ.

This may take some time, but, after you have done it, you should have a clearer idea about sampling distributions!

Summary

This chapter covered the key points relating to the Normal distribution and the Central Limit Theorem. You should now be ready to embark on Chapters 6, 7 and 8, and work on ideas of statistical **estimation** and **inference**. Don't worry if you find some later sections of this chapter difficult to understand. Work at the three activities and the sample examination questions at the end.

Learning outcomes

After working through this chapter and the relevant reading, you should be able to:

- understand the use of E($\overline{X}$) and V($\overline{X}$) and work out sample expected values
- work out areas under the curve for a normal distribution
- understand the application of the central limit theorem
- understand the relationship between size of sample and the standard deviation of the sample mean.

You are not required to know about the binomial and Poisson distributions, or Bernouilli trials, or demonstrate the Central Limit Theorem. (These are both the subject matter of Statistics 2.)

Sample examination questions

1. Given a normal distribution with mean 20 and variance 4, what proportion of the distribution would be:

   a) above 22.

   b) between 14 and 16? (4 marks)

2. The manufacturer of a new brand of Lithium battery claims that the mean life of a battery is 3,800 hours with a standard deviation of 250 hours.

   a) What percentage of batteries will last for more than 3,500 hours?

   b) What percentage of batteries will last for more than 4,000 hours?

   c) If 700 batteries are supplied, how many should last between 3,500 and 4,000 hours? (6 marks)

3. In an examination, the scores of students who attend schools of type A are normally distributed about a mean of 50 with a standard deviation of 5. The scores of students who attend schools of type B are also normally distributed about a mean of 5.5 with a standard deviation of 6.

   Which type of school would have a higher proportion of students with marks below 45? (4 marks)

Notes

Chapter 6

Estimation

Essential reading

Newbold, P. *Statistics for Business and Economics.* (London: Prentice-Hall, 1995) fourth edition [ISBN 0 1385 5549 0]. Chapters 7 and 8.

Further reading

Hanke, J.E. and A.G. Reitsch. *Understanding Business Statistics.* (Burr Ridge Ill: Irwin, 1994) second edition [ISBN 0 2561 1219 3]. Chapter 8.

Wonnacott, T.H. and R.J. Wonnacott. *Introductory Statistics.* (Chichester: Wiley, 1990) fifth edition [ISBN 0 4715 1733 X]. Chapters 7 and 8.1 – 8.3 and 8.5.

Introduction

This chapter is concerned with **data based decision-making**. It is about making a decision which involves a population. The population is made up of a set of individual items. This could be, for example, a set of individuals or companies which constitute the market for your product. It could consist of the items being manufactured from a production line.

The sort of information needed for a decision may be a mean value, (e.g. How many items does an individual purchase per year, on average?) or a proportion (What proportion of items manufactured have a fault?). The associated decision may range from setting up extra capacity to cope with estimated demand, to stopping the production line for readjustment.

In most cases it is impossible to gather information about the whole population, so one has to collect information about a sample from the population and infer the required information about the population. In *Statistics 1*, we will look at the most commonly used estimators, sample measures $\overline{X}$ and p, for the examples above. If you take *Statistics 2*, you will learn what the properties of a good estimator are and look at other measures to be estimated apart from the mean and proportion.

In order to carry out this type of exercise, one obvious decision needs to be made. How large should the sample be? The answer to this question is – it depends! It depends on how variable the population is, on how accurate the input to the decision needs to be, and on how costly the data collection is.

In this chapter you will look at the idea of sample size. How does it affect accuracy when you estimate population means and proportions? How do you use this information? You will learn to construct confidence intervals.

Note that inferring information about a parent (or theoretical) population using observations from a sample is the primary concern of the subject of statistics.

Estimation

Ideas of randomness and estimation underpin the whole of any statistics course.

A **random sample** of *n* observations of a random variable are taken, for example the salary of 10 MBA students, two years after the end of the course:

Salaries, shown as multipliers of the national average:

1.8	2.2	2.3	1.9	1.7
1.9	2.1	2.4	2.2	2.0

This can be written as follows:

In this example, n=10 and we have

$(x_1,x_2,x_3,.....x_{10})$ (1.8, 2.1, ..., 2.3)

The literal meaning of 'random' is 'without order'. Statisticians, however, have a more precise definition. The person or method which chooses the x_i must choose an unbiased method of selection so that the probability of selection is known and non zero. Random number tables are given in most statistical compilations though sometimes, in practice, we might take every n^{th} item from an alphabetical list (this is called quasi random. Look at the definitions in Chapter 10 for more detail on this).

A function of the random sample, say $T(X_1,X_2,X_3,...X_n)$, which may be a sample mean, sample proportion, sample standard deviation etc. is called a **statistic** (or estimator) if it is used to estimate a population parameter. Thus, if the sample mean for the MBA graduates is 2.05, this figure becomes a statistic if it is used to infer that, on average, the population of MBA graduates can expect to earn (2.05 national average) two years after the completion of the course.

A statistic is an unbiased **estimator** of a population parameter (such as the population mean, _) if it satisfies the following relationship:

The statistic $T(X_1,X_2,...X_n)$ is an unbiased estimator of Θ if $E(T) = \Theta$ for all Θ, where Θ is a general population parameter such as the mean μ or the standard deviation σ.

When looking at the **mean**, we see that the **sample mean**, is an unbiased estimate of the **population mean**, that is:

$$E(\overline{X}) = \mu \quad \text{for all} \quad \mu .$$

However, the statistic, $s^{'2}$ is a biased estimate of the variance of a population:

$$s^{'2} = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n} \qquad E(s^{'2}) = \sigma^2 - \frac{\sigma^2}{n} .$$

This bias ($-\sigma^2/n$), is the reason when an estimate of population variance is required, the statistic s^2 should be used where[8]:

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n - 1} \qquad E(s^2) = \sigma^2 .$$

If you take *Statistics 2*, you will learn about the reasoning behind this.

Confidence intervals and limits

It is important to remember that estimates made from data are bound to be imprecise, and it is essential to indicate the level of imprecision associated with an estimate.

The generally adopted procedure for doing this is to state upper and/or lower limits within which the true value of the parameter is likely to lie. These limits are called **confidence limits**, and the interval between them, is called a **confidence interval**.

[8] *If you are using a standard deviation or variance button on a statistical type calculator, make sure that you can distinguish whether the calculator is producing* s^{12} *or* s^2 *. Some buttons show* σ_n^2 *and* σ_{n-1}^2

As an example, let us consider using $\overline{X}$ as an estimate for μ. From the Central Limit Theorem discussed in Chapter 6 we know that:

$\left(\dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}\right)$ is distributed as N (0,1) .

Thus if we assume n and σ are known, we can use the above result to get limits on the value of μ. The general result is given below:

$$\text{Prob}(-k < \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} < k) = 1 - 2\,(1 - \Phi(k)) .$$

This can be rearranged to give:

$$\text{Prob}(\overline{X} - k\sigma / \sqrt{n} < \mu < \overline{X} + k\sigma / \sqrt{n}) = 1 - 2\,(1 - \Phi(k)) .$$

Thus, for a 95% confidence interval; 1 – 2 (1–Φ (k)) = 0.95, thus 1–Φ (k) = 0.025, thus k = 1.96, from our Normal distribution table.

For a 99% confidence interval; 1 – 2 (1–Φ (k)) = 0.99, thus 1–Φ (k) = 0.005, thus k = 2.58.

The next two examples show you what you can do when the population variance is known, or given. The first shows you how to construct 95% and 99% confidence intervals for large samples.

Example 6.1

Measurements of the diameter of a random sample of 200 ball bearings produced by a machine, showed a mean $\overline{X}$ = 0.824”. The population standard deviation is = 0.042”. Find

a) 95%, and

b) 99% confidence intervals for the true mean value of the diameter of the ball bearings.

We are told that σ = 0.042

So the 95% confidence interval is:

$\overline{X} \pm 1.96\ \sigma / \sqrt{n}\ = 0.824 \pm 1.96\ (.042/\sqrt{200}) = 0.824 \pm .006$.

In other words the interval (0.818, 0.830) covers the true mean with a probability of 95%. We can write this as $0.818 \le \mu \le 0.830$.

In a similar way we can find a 99% confidence interval of (0.816, 0.832) for μ. We can write this as $0.816 \le \mu \le 0.832$.

Example 6.2 demonstrates the relationship between sample size and the required precision of an estimate.

Example 6.2

In measuring the reaction time of a patient to a certain stimulus, a psychologist estimates the standard deviation as 0.05 seconds. How large a sample of measurements must he take in order to be (a) 95% and (b) 99% confident that the error in his estimate of the mean reaction time will not exceed 0.01 seconds?

The 95% confidence interval for μ is $\overline{X} \pm 1.96\ \sigma / \sqrt{n}$ and hence the ‘error’ is $1.96\ \sigma /\sqrt{n}$.

In this example the unknown is *n,* is to be chosen so that $1.96\ \sigma / \sqrt{n}\ =\ 0.01$.

Knowing σ =0.05, and rearranging the above inequality, we find that n = 96.04 (i.e. as n must be a whole number 97, observations are required to achieve an error of 0.01 or less with 95% probability).

For 99% probability we find that n = 166.4 i.e. 167 (as n must be a whole number).

Activity

A6.1

National figures for a blood test result have been collected and the σ for population is 1.2. You take a sample of 100 observations and find a sample mean of 25 units. Give the 95% confidence interval for the mean.

Small sampling theory – the use of Student's *t*

So far, our examples have used the Normal distribution. We have assumed that , the population variance, is **known**. We could also assume, using the Central Limit Theorem, that, for a large sample size, we can treat s (the sample estimate of) as normal. This is common practice.

Activity

A6.2

Look at the ball bearing example again but this time read the description 'Measurements of the diameter of a random sample of 200 ball bearings produced by a machine, showed a mean $\overline{X}$ = 0.824" and a standard deviation s = 0.042".

Find the (a) 95% and (b) 99% confidence intervals from the true mean value of the diameter of ball bearings.

(Note: Although this time you have been told that your figure for variance is a sample estimate, you justify using the normal distribution because of the Central Limit Theorem since the sample size is very large and your confidence interval will be exactly the same as in the previous example.)

What do we do, however, if we have to use the variance estimate from a small sample? The answer is 'use the Student's *t* distribution'. This is a distribution first derived by W.W. Gossett under the pseudonym of 'Student'. The statistic we use is:

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$, where s is the unbiased estimate of population variance.

It can be shown (and you will do this if you take *Statistics 2*) that the *t* distribution density function tends towards the Normal density as *n* tends towards infinity.

Since the distribution is symmetrical about 0 it is used in exactly the same way as the Normal. A general 100 (1 – β)% confidence interval is:

$$\text{Prob}(\,\overline{X} - ks/\sqrt{n} < \mu < \overline{X} + ks/\sqrt{n}\,) = 1 - \beta$$

where *k* is found from *t* tables. Note that the table has columns defined by the probability of falling in the right-hand tail (*a* which, in terms of the above interval, is β/2). The rows of the table are defined by *v* (called the **degrees of freedom** of the sample) which for a single sample is *n*–1.

Thus for a 95% confidence interval with a sample size of 16 observations:

$$k = t_{\alpha;\nu} = t_{\beta/2;\text{n-1}} = t_{0.025;15} = 2.131$$

Activity

A6.3

Open your statistical tables at the student's *t* pages. Note that different probability tails are given for *v* = 1, 2 etc. (*v* is the same as *n*–1 in this case). Now look at the 95% confidence interval for $\overline{X}$ when n = 21 (i.e. *v* = 20). You can see the *t* value is 2.086. However, when n is very large ($\propto$) the *t* value is 1.96, that is, exactly the same as for the normal distribution. Although you can see that *t* values are given for quite large *v*'s, we generally assume that the normal distribution measure can be used instead of *t* if *v* is greater than 30 (some textbooks say 50).

Now look at Example 6.3.

Example 6.3

A sample of 10 measurements of the diameter of the sphere gives a sample mean of 4.38" and a standard deviation *s* = 0.06". Find a 95% confidence interval for the actual diameter, and compare it with one (incorrectly) derived from the Normal distribution.

Correct Answer (using *t* distribution since *n*<30 and σ is estimated by *s*)

The 95% confidence interval for μ is $\overline{X} \pm t_{a/2,v}\ s/\sqrt{n}$ where (from t tables) $t_{0.025,9}$ = 2.262. Hence the confidence interval is 4.38 ± 2.262 (0.06/√10) = (4.337, 4.423).

Be careful not to give the following answer.

Incorrect Answer (making a false assumption about Normality)

The 95% confidence interval for μ is $\overline{X} \pm k_a\ \sigma / \sqrt{n}$ where (from Normal tables) $k_{0.025}$ = 1.96. Hence the confidence interval is 4.38 ± 1.96 (0.06/√10) = (4.343, 4.417) giving a false sense of accuracy!

Confidence limits for proportions

The confidence intervals above have looked at the mean of the population. We can extend the concept to other parameters of interest. In this course, we are most likely to want to look at the **proportion** of a population having a particular attribute, such as voting Conservative, being defective, or buying a particular soap powder. Using the Normal approximation to the binomial, it can be seen that we can make the following substitutions in the confidence interval (this work will be done in *Statistics 2*):

μ	becomes	Π	(population proportion with attribute)
$\overline{X}$	becomes	p	(sample proportion with attribute)
$\sigma/\sqrt{n}$	becomes	$\sqrt{\Pi(1\text{-}\Pi)/n}$	(standard deviation of ρ)

$\sqrt{\Pi(1\text{-}\Pi)/n}$ can be estimated by the sample values $\sqrt{p(1-p)/n}$. In this case student's *t* is used.

Example 6.4

A sample poll of 100 voters chosen at random from all voters in a given district indicated that 55% of them were in favour of a particular candidate. Find

a) 95%, and

b) 99% confidence limits for the proportion of all voters in favour of the candidate.

Answers

a) 95% confidence interval for true proportion π is p $\pm 1.96\sqrt{p(1-p)/\Pi}$

i.e. 0.55 ± 1.96 $\sqrt{0.55\,(0.45)/100}$ = (0.452, 0.647).$\overline{p\,(1p)/n}$

b) Similarly 99% confidence interval is p ± 2.58 × $\sqrt{p\,(1p)/n}$ = (0.422, 0.678).

Estimation of intervals for the differences between means and between proportions

We can extend the concept of a confidence interval to functions of the population mean. Often comparisons are made between two populations or between two random variables, usually with the objective of establishing whether or not they are different.

If the two random variables are denoted as X_1 and X_2 and samples of observations of them:

$(x_{11}, x_{12}, x_{13}, ... x_{1n})$ and $(x_{21}, x_{22}, x_{23}, ... x_{2n})$

of size n_1 , and n_2 respectively.

In this case ,we can make the following substitutions in the confidence interval:

μ	becomes	$\mu_1 \pm \mu_2$	(sum or difference of population means)
$\overline{X}$	becomes	$\overline{X}_1 \pm \overline{X}_2$	(sum or difference of sample means)
σ /√n	becomes	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	(this the standard deviation of $\overline{X}_1 \pm \overline{X}_2$).

You can also look at the difference in proportions in the same way. For large samples, where you have two proportions p_A and p_B and the sample sizes are n_A and n_B respectively, then:

standard deviation of $p_A - p_B$ is $\sqrt{\frac{p_A(1-p_A)}{\pi_A} + \frac{p_B(1-p_B)}{\pi_B}}$

and we get a 100 $(1- \propto_1) \propto$ confidence interval for $\rho_A - \rho_B$ of the form:

$$p_A - p_B \pm Z_{\alpha/2}\sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}$$

Activity

A6.4

A random sample of 200 students is taken; 30 of them say they are ‘really enjoying’ Statistics. Calculate the proportion of students in this sample saying they ‘really enjoy’ Statistics and then construct the 95% confidence interval for this value.

You now take a further random sample, in another institution. This time there are 20 students and 8 say they ‘really enjoy’ statistics. Give the 95% confidence interval for *p* this time.

Think about why the two intervals are different. Create a confidence interval for the difference in proportions.

Summary

The concepts of estimation are obviously extremely important for a manager who wants to collect a reasonable amount of data so as to make a good judgement of the overall situation. Make sure that you understand when student's *t* is required rather than the Normal distribution.

Remember that:

- We use the normal distribution when we know the variance or standard deviation either as given by the researcher or as a population figure.
- If we have to estimate variance or standard deviation from a sample, we will need to use student's *t* whatever the size of the sample.
- If the sample is large, then student's *t* approximates to the Normal distribution (Look at Activity 6.3 again).

Learning outcomes

After working through this chapter and the relevant reading, you should be able to:

- calculate sample means, standard deviations and proportions, and understand their use as estimates
- construct a confidence interval for (a) a sample mean, (b) proportion, (c) the difference between two sample means, and (d) two sample proportions
- know when to use the student's *t* distribution.

You do not need to:

- demonstrate the Central Limit Theorem
- know about the concepts of a 'good' estimator.

Sample examination questions

1. Would you say the following statement is true or false? Give brief reasons.

   'When calculated from the same data set, a 91% confidence interval is wider than a 96% confidence interval.' (2 marks)

2. A factory has 1,200 workers. A simple random sample of 100 of these had weekly salaries with a (sample) mean of £315 and a (sample) standard deviation of £20.

   Calculate a 90% confidence interval for the mean weekly salary of all workers in the factory. (5 marks)

3. a) Write down the formula for the standard error of the sample proportion when sampling is at random from a very large population and the population proportion is equal to p. Give the formula for pooled proportions when comparing the two samples. (Note that you need to check this with your textbook.) (4 marks)

   b) An independent assessment is made of the services provided by two holiday companies. Of a sample of 300 customers who booked through company A, 220 said they were satisfied with the service. Of a random sample of 250 of company B's customers, 200 said they were satisfied. For both companies the total customer base was very large.

   Calculate a 95% confidence interval for the difference in the proportions of satisfied customers between the two companies. Basing your conclusion on this interval, do you believe that one company gives more satisfaction than the other? (4 marks)

Notes

Chapter 7

Hypothesis testing

Essential reading

Newbold, P. *Statistics for Business and Economics.* (London: Prentice-Hall, 1995) fourth edition [ISBN 0 1385 5549 0]. Chapter 9.

Further reading

Aczel, A.D. *Complete Business Statistics.* (London: Irwin/McGraw Hill, 1999) [ISBN 0 0728 9302 8]. Chapters 7, 8 and 14 (taking care to omit topics not in the learning outcomes).

Anderson, D.R., D.J. Sweeney, and T.A. Williams. *Statistics for Business and Economics.* (Cincinatti: South-Western Thomson Learning, 2002) eighth edition [ISBN 0 3240 6671 6]. Chapter 9 and Sections 10.2, 12.2, and 14.9.

Hanke, J.E. and A.G. Reitsch. *Understanding Business Statistics.* (Burr Ridge Ill: Irwin, 1994) second edition [ISBN 0 2561 1219 3]. Chapters 9,10 and 11 (taking care to omit topics not in the learning outcomes).

Mason, R.D. and D.A. Lind. *Statistical Techniques in Business and Economics.* (Boston: McGraw Hill, 2001) eleventh edition [ISBN 0 0724 0282 2]. Chapters 9, 10, 11 and the latter half of 16.

Wonnacott, T.H. and R.J. Wonnacott. *Introductory Statistics.* (Chichester: Wiley, 1990) fifth edition [ISBN 0 4715 1733 X]. Chapter 9.

Introduction

In Chapters 4 and 6 you were introduced to the ideas of the **probability** that a parameter could lie within a range of values and in particular the **confidence interval** (generally 90%, 95% or 99%) for a parameter.

In this chapter we are going to look at the idea of using statistics to see whether we should accept or reject statements about these parameters. The arithmetic you will use is similar to that which you met in the last chapter.

We often need to answer questions about a population such as 'Is the mean of the population greater than 2?', or 'Is there a difference between the performance of two operatives?' As in Chapter 6, and generally in statistics, we try to base our answer to these questions on the information we have been given in the samples. Since the questions asked refer to **populations** we are concerned with ideas of **statistical inference**.

The hypothesis

A statement, which may be true or false, often about a parameter of a population, is called a **hypothesis**.

Example 7.1

Suppose that we are looking at the population of boys between the ages of 7 and 8 years old in the UK. Consider two hypotheses of interest:

a) that the mean weight of the population is less than 50kg

b) that 50% of the population can manage to add ⅓ and ¼ correctly.

If you think about these two hypotheses, you will see they are different. If you were to reject the hypothesis given in the first then you would have to show that the mean weight of the population is equal to or greater than 50kg. In the case of (b), your alternative would be that 50% could not do the addition correctly.

We call the base hypothesis the **null** hypothesis and its negation the **alternative** hypothesis. We also talk about one-sided (as in Example 7a) and two-sided (as in Example 7b) hypotheses.

Null and alternative hypothesis, one- and two-tailed tests

A peculiarity of the theory of testing is that we pick out one hypothesis as our baseline – this is the null hypothesis, which we write as H_0. We than set up a sometimes less precise, but more interesting, hypothesis as its competitor. We call this the alternative hypothesis and write it as H_1. In fact in Example 7.1(a) above, we would have to use the null hypothesis H_0: the mean weight of the population is equal to 50kg. The alternative hypothesis H_1, is that the mean weight is less than 50kg.

We are generally concerned that we should not reject the null hypothesis if it is actually true and the tests you will learn in this chapter will address this problem. Note that the alternative hypothesis could also be rejected when it is true, but we will not learn how to measure the probability of this in *Statistics 1*. This material is covered in *Statistics 2*.

For now, it is important to note that there should be no overlap between the null and alternative hypothesis. They cannot both be true.

Example 7.2

Suppose that we have a long-used and well-tested treatment for stomach ulcers. The average length of treatment using this treatment to cure the condition using this treatment is known to be six months. Now Kwaq Laboratories has a brand-new treatment that it says is better. A suitable null hypothesis might be that the average timeto a cure for the new treatment is six months:

H_0: $\mu = 6$

whereas the Alternative Hypothesis could be that average time to a cure for the new treatment is less than six months:

H_1: $\mu < 6$.

One-sided and two-sided alternative hypotheses

For a parameter Θ, and a given value Θ_0, if H_1 is of the form $\Theta > \Theta_{0,}$ or of the form $\Theta < \Theta_0$, then it is said to be **one-sided**.

If H_1 is of the form $\Theta \neq \Theta_0$ then it is said to be **two-sided**.

Returning to Example 7.1, the fact that the alternative hypothesis is that the mean weight of boys between 7 and 8 is less than 50kg means we have a one-sided hypothesis. In contrast, think about tossing a coin: the alternative hypothesis that the probability of ‘heads’ is not equal to 0.5 means we have a two-sided hypothesis.

To carry out a test we calculate from the data the value T of a **test statistic**. If the test statistics falls in the **critical region** we reject H_0 in favour of H_1. If the test statistic does not fall in this region then we do not reject H_0 (which is retained as our working hypothesis). The critical region is often described by using a **critical value** that is a percentage point from the distribution of the test statistic.

Example 7.3

> Suppose that the null hypothesis is that the population mean μ is 5, and the alternative hypothesis is that μ is greater than 5. First, of course, we have to assume that we are dealing with the Normal distribution. If this is the case, we could use as the test statistic T = $\overline{X}$, the sample mean.

One possible critical region could be all values for T = $\overline{X}$ –5 greater than $3z_a$. The percentage point z_a given in the tables is used to define the critical value $3z_a$.

Activity

> A7.1
>
> Now look at the tables to see what this value of z_a would be.

Using this critical region, we say that if $\overline{X}$ – 5 from the sample is greater than $3z_a$ we would reject H_0 in favour of H_1. If $\overline{X}$ – 5 $3z_a$ we would accept H_0 as a working hypothesis.

One- and two-tailed tests

If the critical region lies in only one tail of a distribution, we have a one-tailed test. A critical region which contains both an upper tail and its lower tail gives a two-tailed test. If we have a one-sided alternative hypothesis ($\overline{X} > 0$) we have a one-tailed test.

A two-sided alternative hypothesis ($\overline{X} \neq 0$) leads to a two-tailed test.

Look at Example 7.3 again. The critical region $\overline{X} - 5 > 3z_a$ represents the upper tail of the distribution of $\overline{X}$, so this is a one-tailed test.

Activity

> A7.2
>
> Think about each of the following statements. Then give the null and alternative hypotheses and say whether they will need one- or two-tail tests:
>
> a) The general mean level of family income in a population is known to be 10,000 ulam a year. You take a random sample in an urban area U and find the mean family income is 6,000 ulam a year in that area. Do the families in the chosen area have a lower income than the population as a whole?
>
> b) You are looking at data from two schools on heights and weights of children by age. Are the mean weights for girls aged 10–11 the same in the two schools?
>
> c) You are looking at reading scores for children before and after a new teaching programme. Have their scores improved?

Type I and Type II errors

The approach taken to testing these statements can be formalised in the following manner:

a) A null-hypothesis: H_0 is assumed to be true unless the sample evidence points against it.

e.g. $H_0 :\mu_A = \mu_A$

b) An alternative hypothesis: H_1 is the hypothesis describing the situation if the null hypothesis is incorrect.

e.g.$H_1 : \mu_A > \mu_B$

or $H_1 : \mu_A \neq \mu_B$

or $H_1 : \mu_A < \mu_B$

The description of the hypothesis depends on the problem, the null hypothesis usually assumes the status quo, equality or no change. The alternative specifies the change required. When testing H_1 against H_0 we can make two possible errors:

- Type I – reject H_0 when it is in fact correct
- Type II – accept H_0 when it is in fact incorrect.

Before testing a hypothesis, the risk of making a type I error has to be specified:

- Prob (reject H_0 | H_0 correct) = α. The probability we reject H_0, given that it is really correct, is called α, the significance level (some texts call it the size of the test) of the test.
- The probability of making a type II error is β. Its complement, $1-\beta$, is called the **power** of the test. You will not be called upon to work out either β or $(1-\beta)$ at this stage. However, you should understand the implications of the different types of error and be able to complete the chart in Activity 7.3.

Activity

A7.3

Complete the following chart:

Result from your test	**Real situation**	
	H_0 true	H_0 false
H_0 true	Correct Probability $(1-\alpha)$ called the **confidence interval** of the test.	Type II error Probability $(1-\beta)$ called the **power** of the test
H_0 false		

Think hard about this and learn it; it will save you problems later!

Some examples of setting and testing hypotheses

The null hypothesis is always an equality, that is:

$$H_0 : \mu = \mu_0$$

or you might prefer to write:

$$H_0 : \mu - \mu_0 = 0.$$

The approach is to assume that the null hypothesis is true until the sample evidence points otherwise. Thus, under H_0, we assume that μ_0 is distributed normally:

$$\overline{X} \text{ is distributed as } N\left(\mu, \frac{\sigma^2}{n}\right).$$

We set the (probability of making a type I error) at an appropriate percentage value and look up the relevant value for *x* in the tables. Here is a worked example.

Example 7.4

The mean lifetime of 100 components in a sample is 1,570 hours and their standard deviation is 120 hours. Is the mean lifetime of all the components produced. Is it likely the sample comes from a population whose mean is 1,600 hours?

We decide to test at a 1% significance level:

$H_0 : \mu = 1600$, against

$H_1 : \mu \neq 1600$.

This is a two-tail test and hence the level or size of test is split equally between the two tails of the distribution.

Since we are using a 1% level of test, then $1 - \Phi(k) = .005$, so $z = 2.58$ from Normal tables. Note that:

$$z = \frac{\bar{x} - 1600}{\sigma / \sqrt{n}}$$

Hence the critical region tells us to reject H_0 if:

$x > \mu + 2.58\,\sigma/\sqrt{n}$, or if $\bar{x} < \mu - 2.58\,\sigma/\sqrt{n}$.

Note that, even though we are estimating from the sample, s can be used here as n > 30. (See Chapter 6 on when it is necessary to use Student's *t*.)

Thus the critical region is:

$\bar{x} < 1600 - 2.58(120)/\sqrt{100} = 1569.0$, or

$\bar{x} > 1600 + 2.58(120)/\sqrt{100} = 1631.0$

Since $\bar{x} = 1570$, which is between these two values, there is not enough evidence to reject H_0.

Activity

A7.4

The manufacturer of a patent medicine claimed that it was 90% effective in relieving an allergy for a period of 8 hours. In a sample of 200 people suffering from the allergy, the medicine provided relief for 160 people.

Determine whether the manufacturer's claim is legitimate. (Be careful. Your parameter here will be *p*.) Is your test one- or two-tail?

Comparing the means and proportions of two populations

Consider the Normal random variables X_1 and X_2. In this case we have samples from both populations. The statistic of interest is the standardised difference between sample means:

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

or, when using the variance estimated from the sample:

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The null hypothesis is that there is no difference, thus the statistic is distributed as N(0, 1) or Student's *t* for small samples with $n_1 + n_2 - 2$ less than 30~ if the population standard deviation is not known..

Here is an example which requires the use of the *t* test.

Example 7.5

Ten patients are given courses of treatment under two different drugs. The benefits derived from each drug can be stated numerically; the readings are given below. Test the hypothesis that there is no difference between the drugs.

Patient no.	**New Drug**	**Old Drug**
1	3.9	2.8
2	2.8	0.6
3	3.6	2.8
4	3.1	1.8
5	1.9	1.9
6	6.4	5.4
7	2.1	0.9
8	6.6	4.0
9	7.2	3.8
10	5.4	4.0
Mean	4.3	2.8
s.d.	1.951	1.524

H_0: $\mu_{new} = \mu_{old}$

H_1: $\mu_{new} = \mu_{old}$

Under H_0:

$$\frac{\bar{x}_{new} - \bar{x}_{old}}{\sqrt{\frac{s^2_{new}}{n_{new}} + \frac{s^2_{old}}{n_{old}}}} - t_{n_{new}+n_{old-2}}$$

The 5% critical region is:

$$| \bar{x}_{new} - \bar{x}_{old} | > t_{n_{new}+n_{old-2}} \sqrt{\frac{s_{new}^2}{n_{new}} + \frac{s_{old}^2}{n_{old}}}$$

The *t* value for 18 (10 + 10 – 2) degrees of freedom with 2.5% in the tail is 2.101 and the term in the square root is 0.783. Thus the critical region is |difference in sample means| > 1.645

The **observed value** $| \bar{x}_{new} - \bar{x}_{old} |$ is 1.5, which does not fall in the **critical region**. There is therefore no reason to reject H_0.

The next example tests the difference between two proportions.

Example 7.6

A postal researcher wanted to test the theory that a higher response rate is achieved when a postal questionnaire is sent out with a personalised covering letter (*A*) than when the covering letter is impersonal (*B*).

Two random samples of 100 people were selected. When the questionnaires were despatched, the first sample received letter *A* and the second received letter *B*. The response rates to letter *A* were 70% and to letter *B*, 55%.

a) Do these results provide evidence in support of the theory?

b) Explain the reasoning behind the test.

Answer

Call the response rate for each sample p_A and p_B respectively and the number in each sample n_A and n_B.

Note that this is a **one-tail** test. We want to know if p_A is greater than p_B (i.e. whether the population proportion responding to letter *A* is greater than that responding to letter *B*)[9].

So H_0 : $p_A - p_B = 0$

H_1 : $p_A - p_B > 0$

From the sample we see that the difference between the two response rates is:

$p_A - p_B = .15.$

We will test the statistic $\dfrac{P_A - P_B}{S.E.(P_A - P_B)}$

where the S.E. (which you may regard as similar to S.D.) of the difference between them is calculated by the following formula (– for each group *A* and *B* we insert the observed proportion p for $\overline{\mathit{\Pi}}$ here).

$$\text{S.E.} = \sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}$$

$$= \sqrt{\frac{0.7 x 0.3}{100} + \frac{0.55 x 0.45}{100}}$$

$$= \sqrt{0.0021 + 0.0025}$$

$$= \sqrt{.0046}$$

$$= .0678$$

So our statistic:

$$\frac{p_{A} - p_{B}}{S.E.(p_{A} - p_{B})} = \frac{.15}{.0678}$$

$$= 2.2124$$

We see, from the tables, the following percentage values for the tails:

10%	=	1.282
5%	=	1.645
1%	=	2.366

Our value lies between 1% and 5%. (It has a p value of 1 – 0.98645 = 0.013555 or 1.355%). So we say the proportions are different at the 5% but not at the 1% level. The evidence that a personal covering letter improves the response rate is clear at the 5% level.

If however, we apply a more stringent criterion, then the evidence of improvement is not convincing. We need more evidence, perhaps by analysing larger samples, before we could be sure of an improvement at the 1% level.

Summary

The idea of testing hypotheses is a central part of statistics, and underpins the development of theories and checking of ideas in management and the social sciences. It is important that you make sure you understand the material introduced in this chapter and Chapters 5 and 6 before you move on to look at the chi-squared distribution in the next chapter.

Learning outcomes

After working through this chapter and the relevant reading, you should be able to:

- set up the null and alternative hypotheses for a problem and state whether the latter is one- or two-tailed
- define and use the terminology of statistical testing
- carry out statistical tests on means and proportions
- construct and explain a simple chart showing the kinds of errors that can be made in hypotheses testing.

You are not expected to understand or make calculations about detail of the ideas of Type II errors or matched or paired tests.

Sample examination questions

1. Say whether the following statement is **true** or **false** and briefly give your reasons.

   'The power function of a test is the probability that the correct hypothesis is chosen.' (2 marks)

2. Explain your attitude to a null hypothesis if a test of hypothesis is significant:

   a) at the 1% level

   b) at the 10% level but not at the 5% level. (4 marks)

3. Measurements of a certain characteristic are normally distributed. What can you say about an individual position (with respect to this characteristic) in the population if their Z score is:

   a) – 0.5

   b) + 8.5

   c) + 1.95? (3 marks)

4. You have been asked to compare the percentages of people in two groups with $n_1 = 16$ and $n_2 = 24$ who are in favour of a new plan. You decide to make a pooled estimate[9] of proportion and make a test. What test would you use? Give the degrees of freedom and the 5% critical value. (6 marks)

5. Look at Question 3 of the sample examination questions in Chapter 6. You were asked in (b) whether you thought one company gave more satisfaction than the other.

   Now give the null and alternative hypotheses for such a one-tailed test. Show how you would accept or reject the null hypothesis (4 marks)

[9] *Look at the work in Newbold before you do this.*

Chapter 8

Contingency tables and the chi-squared test

Essential reading

Newbold, P. *Statistics for Business and Economics.* (London: Prentice-Hall, 1995) fourth edition [ISBN 0 1385 5549 0]. Sections 11.1 and 11.3.

Further reading

Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics.* (Chichester: Wiley, 1990) fifth edition [ISBN 0 4715 1733 X]. Chapter 1.

Introduction

You have now met the main ideas of estimation and inference which you need to know in order to make sense of the data you have collected, or have been given, so that you can make policy decisions. There is one further test which illustrates the uses of hypothesis testing and is also generally used in business management and social science fields.

The figures you will be interested in involve more than one measurement of a variable and deal with them by counting the numbers in a category rather than the absolute values[10].

Examples of a contingency table and a chi-squared test

Imagine, for example, that you were working with the figures given in Example 7.1 about boys of 7 and 8 years old in the UK. Say we had been given a measure of reading ability as well as weight. One question we might ask is: is reading ability related to weight?

We are told that there are:

- four levels of reading ability: very high, high, on target, and low
- three weight groups: above average, average, and below average.

We would be able to make a contingency table and test whether reading ability is related to weight, or not, using the chi-squared test. The contingency table would look like this:

Weight and reading ability of 7 to 8-year-old boys

Reading ability					
Weight	**Very high**	**High**	**On target**	**Low**	
Above average	n_{11}	n_{21}	n_{31}	n_{41}	$\sum n_{i1}$
Average	n_{12}	n_{22}	n_{32}	n_{42}	$\sum n_{i2}$
Below average	n_{13}	n_{23}	n_{33}	n_{43}	$\sum n_{i3}$
	$\sum n_{1j}$	$\sum n_{2j}$	$\sum n_{3j}$	$\sum n_{4j}$	$\sum n_{ij}$

= N

[10] *These also relate to regression and correlation (which you will meet in Chapters*

The n_{ij} are called **observed values**.

We set up our null and alternate hypotheses as we learned in Chapter 7:

H_0 : Weight is not related to reading ability

H_1 : Weight is related to reading ability

But what is the test statistic? Clearly one mean or proportion will not do.

If we think a little, we see that we could imagine an **expected value** for each observation n_{ij} , assuming that weight and reading ability were not related. If this is the case, then the expected value of the $n_{32,}$ for example, can be calculated as the proportion n_{12} of the total N of the number $n_{3j.}$ That is:

$$\text{Expected value for } n_{32} = \frac{\sum n_{i2} \times \sum n_{3j}}{\sum n_{ij}}$$

The chi-squared statistic gives us a formula for using these expected values:

$$\chi^2 = \sum_{ij} \frac{(0_{ij} - E_{ij})^2}{E_{ij}}$$

where the Oij are observed values and the Eij expected values.

If you turn to the chi-squared tables in your book of statistical tables, you will see percentage values given for chi-squared by **degrees of freedom** ν.

All you need to know now is what ν is! In the case of contingency tables we take the number of columns minus one and multiply it by the number of rows minus one:

$\nu = (c - 1)$ (r– 1).

In this case, therefore:

ν = (4–1) (3–1)

= 3 2

= 6

Note that you should then carry out a one-tail test. The 5% value is given in the tables as 12.59.

Activities

A8.1

A survey has been made of levels of satisfaction with housing by people living in different types of accommodation. Levels of satisfaction are:

high, medium, low, very dissatisfied

and housing types are:

public housing apartment, public housing house, private apartment, private detached house, private semi-detached house, miscellaneous (includes boat, caravan etc.!)

Give:

a) the null and alternative hypotheses

b) the degrees of freedom

c) the 5% and 1% critical values for χ^2.

(Note. Remember that you will reject the null hypothesis if your calculated χ^2 were greater than the value given in the tables.)

A8.2

Draw the χ^2 curve and put in the rejection region for 5% and 1% with 6 degrees of freedom. Make sure you understand which calculated values of χ^2 will lead you to reject your H_O.

Now consider the following example.

Example 8.1

In a survey made in order to decide where to locate a factory, samples from five towns were examined to see the numbers of skilled and unskilled workers. The data were as follows

Numbers of workers, by skill and area

Area	Number of skilled workers	Number of unskilled workers
A	80	184
B	58	147
C	114	276
D	55	196
E	83	229

Does the population proportion of skilled workers vary with the area?

Be careful you know what you are doing!

1. Write down H_0 and H_1 first:

   H_O : The proportion of skilled workers does not vary with area

   H_1 : The proportion of skilled workers is related to area.

2. Then write down the degrees of freedom:

   (r – 1) (c – 1) = 4 × 1 = 4.

3. Write down the 10%, 5% and 1% critical values of x^2:

   7.78, 9.49, and 13.28 respectively.

Now look at the calculations, noting that the expected counts are printed below the observed counts.

Area	Number of skilled workers	Unskilled workers	Row total
A	80	184	264
	(72.4)	(191.6)	
B	58	147	205
	(56.2)	(148.8)	
C	114	276	390
	(107.0)	(283.0)	
D	55	196	251
	(68.8)	(182.2)	
E	83	229	312
	(85.6)	(226.4)	
Column total	390	1,032	1,422

Remember that the expected values E_{ij} are easily found from the row and column totals. For instance, for the bottom right-hand corner of the table, the expected value is:

12 1032/1422 = 226.4.

Note that we have put all the expected values in brackets in the chart above.

The value of the statistic is

$$\chi^2 = (80 - 72.4)^2/72.4 + (184 - 191.6)^2/191.6 + \ldots$$

$$= 0.797 + 0.301 + 0.056 + 0.021 + 0.463 + 0.175 +$$

$$2.782 + 1.051 + 0.077 + 0.029 = 5.753$$

If we look at our significance levels (Table 8 of the *New Cambridge Statistical Tables*), we can see that 5.763 is less than all of them (1%, 5% and 10%) and so we will not want to reject the null hypothesis. It seems pretty clear, on this evidence, that there is little difference between the areas in the proportion of skilled workers there are (so, if this was one of your criteria for choosing where to put your new factory, your management team would be no further forward with their factory site decision!).

You could also use Table 7 to see that the *p* value for $\chi^2 = 5.5$ is 0.2 and for 6.0 it is 0.24. In other words these figures are likely to show little difference between the areas in skilled workers.

What do the results of a chi-squared test actually mean?

If you look at the figures in Example 8.1 again, you will see that one area (D) does appear to have a much lower proportion of skilled workers than you might expect (the number of skilled workers observed is only 55, compared with an expected value of 68.8). The chi-squared test only look at the overall differences. If we want to look at individual areas, it might be worth comparing the proportion of partly skilled workers in area D ($p_D =$ 55/251) with others ($p_{ABCE} =$ 335/1171).

Activity

A8.3

Test the hypothesis

$H_0 : p_D = p_{others}$, against

$H_1 : p_{others} < p_{others}$.

Think about your results and what you would explain to your management team who had seen the chi-squared results and want, for other reasons, to site their factory at area D.

Here is an example to try where the chi-squared and test of difference in proportion test an equivalent hypothesis.

Activity

A8.4

Look at the following table taken from a study of gender differences in perception. One of the tests was of scores in verbal reasoning.

Contingency table of verbal reasoning level by gender

	High	Low	Totals
Male	50	150	200
Female	90	210	300

Do these figures show a difference in verbal reasoning by gender?

Try to do this:

a) using chi-squared

b) using the test of differences in proportions.

Make sure you get the same results!!

(Hint: Make sure you think about the H_o and H_1, in each case!)

Other uses of chi-squared

The chi-squared distribution is often used more generally for 'goodness of fit' tests, and tests when population parameters are not known. These cases are not the subject of this course (though you will meet them if you take *Statistics 2*). The examples we have just looked at are typical of the kind of work you might have to do in business, management or the social sciences.

There is a special case, which it is worth looking at, not covered above. This is when you are only dealing with one row, or column. It is in fact a special case of the more general tests you will make in *Statistics 2*. There are two points to note here:

- ν is always 1 times (c–1) or (r–1) (and not 0 as you might expect!)
- you will probably have to think harder about the *E* values.

Example 8.2

A confectionary company is trying out different wrappers for a chocolate bar; its original *A*, and two new ones *B* and *C*. It puts the bars out in a supermarket and looks to see how many of each wrapper type have been sold in the first hour. Here are the results

Numbers of bars chosen in one hour; by wrapper type

Wrapper type	A	B	C	Total
	8	10	15	33

Is there a difference between wrapper types in the choices made?

So H_0 : There is no difference in preference for the wrapper types

H_1 : There is such a difference.

The degrees of freedom will be 1 (3–1) = 2.

10%, 5%, & 1% values are 4.61, 5.99 and 9.21 respectively.

How do we work out the *E*'s? Well, for **equal** preference, with three choices each, the probability will be ⅓, so we expect 11 in each category, as shown below:

	A	B	C	Total
Wrapper type (0)	8	10	15	33
Wrapper type (E)	11	11	11	33

We can now use the formula as in the last example:

$$\chi^2 = \sum_{i=1}^{i=3} \frac{(0_i - E_i)^2}{E_i}$$

$$= \frac{(8-11)^2}{11} + \frac{(10-11)^2}{11} + \frac{(15-11)^2}{11}$$

$$= 2.364.$$

This is less than all our tabular chi-squared values and so we do not reject H_O. It looks as if there are no preferences for a particular wrapper type on the choices so far.

Sometimes the derivation of the *E* value is not quite so obvious.

Activity

A8.5

Set out the null and alternative hypotheses, degrees of freedom, *E* values, and 10%, 5% and 1% values for the following problem. The following figures give live births by season in town X :

Spring	100
Summer	200
Autumn	250
Winter	180

Is there any evidence that births vary over the year?

The number of days per season in this country are spring (93), summer (80), autumn (100), winter (92).

(Hint. You would expect, if the births are regularly distributed over the year, that the number of births would be proportionate to the number of days per season. So work out your *E*'s by taking the number of days per season/number of days in the year and multiplying by the total number of births over the year.)

Summary

You should regard this chapter as an opportunity to revise your work on hypothesis testing in Chapter 7 and also revisit your work on testing proportions. The only new material here is a way of testing the significance of countable figures as opposed to their attributes.

Learning outcomes

After working through this chapter and the relevant reading, you should be able to:

- set up the null and alternative hypothesis appropriate to a contingency table
- work out the degrees of freedom, expected values and appropriate significance levels of chi-squared for a contingency table
- understand the limitations of a chi-squared test
- be able to extend from chi-squared to an appropriate test of proportions if necessary
- be able to work with one row or column contingency table as above.

You are not expected to carry out general goodness of fit tests for distributions or deal with the case where a population parameter is not known.

Sample examination questions

1. You have carried out a χ^2 test on a 3 × 4 contingency table which you have calculated to study whether there is a relationship between advertising and sales of a product. You have four levels of advertising (A, B, C and D) and three levels of sales (low, level, and high).

   Your calculated χ^2 is 13.5. Giving degrees of freedom and an appropriate significance level, set out your hypothesis. What would you say about the result?

   (6 marks)

2. The table below shows a contingency table for a sample of 1,104 randomly selected adults from three types of environment (City, Town, Rural) and classified into two groups by the level of exercise. Test the hypothesis that there is no association between level of exercise and type of environment and draw conclusions.

	Level of exercise	
Environment	High	Low
City	221	256
Town	230	118
Rural	159	120

   (10 marks)

3. Two surveys have collected information on adult and teenager cigarette use, the results of the first survey are given in Table Q3.1 (below). The results of the second survey, carried out two years later on a new sample of 1,000 households, are given in Table Q3.2.

   a) Without doing any further calculations, test for association between rows and columns of Table Q3.1. (4 marks)

   b) Calculate the chi-squared statistic for Table Q3.2 and test for association between rows and columns. (10 marks)

   c) Write a short report explaining what Table Q3.1 shows about the nature of the association, if any, between teenager and adult cigarette use in a household, and (by comparing the two tables) discussing whether or not the extent of any association changed over the years. (6 marks)

Table Q3.1 First survey households classified by cigarette use by teenager and by adults : frequency (column percentage)

		Adult cigarette use		Total
		Yes	No	
Teenage cigarette use	Yes	198 (8.4)	170 (10.4)	368
	No	2,164 (91.6)	1,468 (89.6)	3,632
$\chi^2 = 4.6$	Total	2,362 (100)	1,638 (100)	4,000

Table Q3.2 Second Survey Households classified by cigarette use by teenager and by adult(s) : frequency (column percentage)

		Adult cigarette use		Total
		Yes	No	
Teenage cigarette use	Yes	48 (8.1)	45 (11.0)	93
	No	544 (91.9)	363 (89.0)	907
	Total	592 (100)	408 (100)	1,000

4. You have been given the number of births in a country for each of the four seasons of the year and are asked whether births vary over the year. What would you need to know in order to carry out a chi-squared test of the hypothesis that births are spread evenly between the four seasons? Outline the steps in your work.

(8 marks)

Chapter 9

Sampling design

Essential reading

Moser, C.A. and G. Kalton. *Survey Methods in Social Investigation.* (Aldershot: Dartmouth,1979) second edition [ISBN 0 4358 2604 2]. Background – Chapter 1. Planning of Surveys and their coverage, Chapters 2 and 3, (Sampling and sampling errors 4, 5, 6, 7. Non sampling errors Chapter 11, 12, 15).

Newbold, P. *Statistics for Business and Economics*. (London: Prentice-Hall, 1995) fourth edition [ISBN 0 1385 5549 0]. Chapter 18 all sections apart from 18.5 on determining sample size.

Additional reading

The Stationery Office, *Social Trends*. (London: HMSO, 2002) [ISBN 011-621472-4].

Introduction

This chapter describes the main stages of a survey and the sources of error at each stage. This part of the course is the foundation for your future work in applied Social Science Business and Management. There is not much point in learning the various techniques we have introduced you to in the rest of the guide unless you understand the sources and limitations of the data you are using. This is important to academics and policy-makers alike! The material in this chapter and Chapter 11 is a useful preparation for **Marketing and market research** and **Methods of social research**.

The terms we use

Before you start reading, make sure you have a good grasp of the meaning of the following words.

Census

This is generally used to mean a complete enumeration of the units to be studied whether individuals, families, or firms in an industry. Examples are the population census in the UK and in your own country. Its advantage is that there are no 'sampling errors'. Its disadvantage is expense: there is a limit to the amount of information it is economic to collect from large numbers of firms or individuals. 'Non-sampling errors' may occur because of the necessity to use cheaper interviewers.

Survey or sample survey

'Sample survey' is really the correct term, but both are used currently to describe the study by interview, telephone or mail questionnaire of a subgroup of the population to be covered. Sampling errors will occur but, because of the smaller numbers to be contacted, resources can be should be used to ensure good quality interviews or to check completed questionnaires so that non-sampling errors should be less and consequently researchers can ask more questions.

Types of sample

You should know about the following kinds of sampling for a sample survey.

Simple random

A simple random sample is one in which each unit has an equal, non-zero, chance of being selected. This occurs, for example, when 1 in 10 students are chosen from a school register in random order.

Random

A random sample is one in which each unit has a known (not necessarily equal) non-zero chance of being selected. You may wish, for example, to study equal numbers of girls and boys from the register of Sociology students. If there are 200 boys and 100 girls then, to achieve 20 in each group, you sample 1/10 of the boys and 1/5 of the girls – the chance of selection is known but not equal for the different groups. Stratified samples and cluster samples are both forms of random sampling.

Note that random sampling, including simple random sampling, means that a list (or frame) needs to be available from which you can draw your sample. Once you can select from the frame with known probability, you can use the techniques you learned in Chapters 6, 7 and 8 to assess the accuracy of your results.

The list may not always be as obvious as a school register or voting list (although both are quite common). Telephone interviews often use random digit dialling, where the list is not available to the interviewer but the computer selects random telephone numbers from the data set of phone numbers it has stored.

There are sometimes problems of confidentiality about use of obvious lists. For example you might be asked to make a random sample of people who owed Visa card debts. The list of such debtors would not be generally available and you would not, legally, be allowed to obtain them in most countries! Problems like this lead one sometimes to run an interview survey using a quota sample.

Quota sample

Here the probability of an individual's selection is not known; the sample is a collection of representative individuals. Interviewers are given a quota of numbers they must contact and interview by sex, social class, age and other variables relevant to the investigation being undertaken.

On the face of it, a quota sample sounds an attractive choice, but of course we have no real guarantee that we have achieved a really representative set of respondents to our questionnaire. Were the women we interviewed only those working in the local offices? Were the young adults all students?

Basically, as we do not know the probability that an individual will be selected for the survey, the basic rules of inference which we have been learning to use do not apply.

Example 9.1

You have been asked to make a sample survey of each of the following. Would you use random or quota sampling? Explain.

a) Airline pilots, for their company, about their use of holiday entitlement in order to bring in a new work scheme.

b) Possible tourists, about their holiday destinations and the likely length of time and money they expect to spend on holiday in the next year, for a holiday company planning its holiday schedule and brochure for next year.

c) Household expenditure for government assessment of the effect of different types of taxes.

a) In the case of the airline pilots, as the survey is for the company (and there is therefore no confidentiality issue) it is quite easy to use the company's list of personnel. A quota sample would not be very easy in these circumstances: you would have to send your interviewers to a venue where most pilots would be likely to meet, or you would risk a very unrepresentative sample.

So in this case a random sample would be easy and efficient to use. You would be able to achieve accurate information and use your statistical techniques on it. The subject matter too means that it is likely the pilots would take the survey more seriously if they were contacted through the company's list.

b) The situation for the tourist survey is different. There will be not one, but several, lists of tourists from different holiday companies and data confidentiality might well mean you could not buy lists which do not belong to your company. You might use the register of voters or list of households, but then you would not necessarily target those thinking about holidays in the near future. So a random sample sounds like an expensive option if this is to be a general study for a tourist company assessing its future offers and illustrations for its holiday brochure. Here, a quota sample makes more sense: interviewers can quickly find the right respondent for the company's needs and get a general picture of holidaymakers' preferences.

c) The government survey will require accuracy of information in an important policy area. A random sample will have to be used and the national lists of addresses or voters used.

Activity

A9.1

Think of at least three lists you could use in your country as a basis for sampling. Remember, each list must:

- be generally available
- be up to date
- provide a reasonable target group for the people you might wish to sample.

How errors can occur

It is important to understand the main stages of the survey and the ways in which errors can occur at each of these stages. The following summary of a random survey should help, but try to make your own when you have finished your reading.

Stage of survey	Possible errors in survey
Definition of target sample frame definition	Errors occur if the frame is not representative of the target population. Where the frame is a sensible one it may have other inadequacies (see Moser and Kalton).
Sampling technique chosen – sample size	Errors because of sampling techniques chosen can be estimated and vary with sample size. At this decision stage the researcher can decide how accurate he wishes to be for a given cost. He should carry out a 'pilot' survey to check that he knows the likely variability of the sample.
Questionnaire design	Errors occur if the questionnaire is not clear and easy to complete or reply to. Here the pilot survey is most important. Badly trained or unsuitable interviewers can also affect the results. Note that errors at this stage are difficult to quantify.
Field work	Even well-trained interviewers with good questionnaires, or a well-organised mail questionnaire, may hit problems. People may refuse to reply, they may not be at home over the period of the survey, or the interviewers may not administer the questionnaires properly. Again, these errors are difficult to quantify.
Compilation and analysis of results	Careful computing and intelligent use of statistics are essential for useful results.

Note that sampling errors can be quantified in advance and are really the result of the researchers' planning given a particular expenditure level. Non-sampling errors, on the other hand, can be very difficult to detect once they have occurred and may be caused by very simple things – the misunderstanding of a word in a questionnaire by less educated people, the dislike of a particular social group of one interviewer's manner, or the loss of a batch of questionnaires from one local post office. These could all occur in an unplanned way and bias your survey badly.

Pilot and post-enumeration surveys

Both kinds of error can be controlled or allowed for more effectively by a pilot survey. A pilot survey is used:

1. to find the standard errors which can be attached to different kinds of questions and hence to underpin the sampling design chosen and
2. to sort out non-sampling questions:
   - Do people understand the questionnaires?
   - Are our interviewers working well?
   - Are there particular organisational problems associated with this enquiry?

Sometimes, particularly for government surveys, a post-enumeration survey is used to check the above. Here a sub-sample of those interviewed are reinterviewed to make sure that they have understood the questions and replied correctly. This is particularly useful when technical questions are being asked.

Further information on quota samples

You were introduced to the quota sample as an alternative method to random sampling when no list is available. However, non-random samples are also frequently used by market research organisations or companies when speed is important. They are rarely used by governments.

Example 9.2

> When would you use a quota sample?
>
> (Read the relevant parts of Newbold and Moser and Kalton before reading the explanation!)

You would be likely to use a quota sample:

- **When you are in a hurry.** Clearly an interviewer with a target requiring her to reach a certain **number** of people on a given day (quota) is likely to be quicker than one which requires a **specific** person or household to be contacted (random). Typical quota controls for the interviewer to meet are:
  - age
  - sex
  - socio-economic group or social class.

  Note that the more controls the interviewer is given, the longer it will take to complete the required number of interviews (and hence it will take longer to complete your study).
- **When there is no obvious or convenient list to cover the population to be studied.** Where obtaining a list is likely to be very complicated, then a **sensible** targeting of a population to take a quota sample is required. You might for example wish to contact drivers of coaches and buses over a set of routes. There are a lot of bus companies involved, and some of them will not let you have their list of employees. Besides some of the lists give a driver's home address and some give the route he runs. **One** of the things you could do in these circumstances is carry out a quota study at different times of the day. There are often random alternatives though, using lists you may not have thought of.

  In the case above you might be able to make a list of scheduled journeys on the routes you wish to study and make a random sample of the routes, interviewing the relevant driver as he completes his journey.
- **When you need to reduce cost.** Clearly time saving is an important element in cost saving.
- **When the detailed accuracy of your results is not important.** You may not wish to have an answer to your question to the high and *known* level of accuracy that is possible using a random sample but merely to get an *idea* about a subject. Perhaps you will be carrying out further work on this subject. Perhaps you only need to know if people, on the whole, *like* your new packages for Ooli sweets. In this case, asking a representative group of people (quota) will be quite sufficient for your needs.

Warning. You should be aware of two points however:

- **Omission of non-respondents**. Because you only count the individuals who reply (unlike random sampling where your estimate has to allow for bias though non-response) the omission of non-respondents (see page 85) can lead to serious errors. If all those who did **not** like your new Ooli sweet wrappers, for example, refused to reply when they saw you coming with your samples (and they were all mothers of small children who thought Ooli sweets were very unhealthy and bad for teeth) then your results would be misleading. For this reason, members of the British Market Research association have now agreed to list non-response as it occurs in their quota samples, and this is regarded as good practice.

- **Quota Controls**. One way of dealing with the problem listed above is to insist on more detailed controls. In the example above, for instance, the interviewer could be required to contact a number of mothers of young children and not just 'women aged 25–34'.

  Interviewers may in the end be given many extra controls. In our example, we might ask for age, sex, employment status, marital status and childbearing stage. This can take away a lot of the cost advantages of using a quota, rather than a random sample. Imagine the time you would take locating the last woman for your sample aged 35–44, married with teenage children and a full-time job! There is the additional expense of paying interviewers more for a smaller number of interviews (on the basis of the *time* they spend on the job). If that is not done, the temptation to cheat, and make results completely invalid, will be strong.

Activity

A9.2

Think of three quota controls you might use to make a quota sample of shoppers in order to ask them about the average amount of money they spend on shopping a week. Two will be easy for your interviewer to identify. How can you help them with the third?

Non-response and response bias

Bias caused by response and non-response is worth a special entry. It can cause problems at every stage of a survey, both random and quota and however administered. Going through the steps involved in a survey we can see the points where they can arise:

- The first problem can be in the **frame**. Is an obvious group missing? For example:
  - if the list is of householders, those who have just moved in will be missing
  - if the list is of those aged 18 or over, and the under 20s are careless about registration, then younger people will be missing from the sample.
- In the field, **non-response** is clear. (Note that it occurs in the quota sample but is not necessarily recorded, see the earlier discussion.) It is most important to try to get a picture of any shared characteristics in those refusing to answer or people who are not available at the time of the interview.
- **Response error** is very tricky as it is not so easy to detect. A seemingly clear reply may be based on a misunderstanding of the question asked or a wish to deceive. (A clear example in this country is the reply to the question about the consumption of alcohol in the Family Expenditure Survey. Over the years there is up to a 50% understatement of alcohol use compared with the overall known figures for sales from Customs and Excise!)

In relation to all these problems, pilot work is most important. It may also be possible to carry out a check on the interviewers and methods used after the survey (see Moser and Kalton on the Post Enumeration Survey).

Some surveys

Before reading on, study Figure 9.1 which gives a brief description of the main UK surveys used in *Social Trends*.

Note that different sample frames are used in addition to the Electoral Register described in Moser and Kalton. Respondent type and size of target sample also vary, the latter from 4,500 to over 150,000.

Note also the different response rates. All these factors are related to the subject matter of the survey, the kinds of questions asked and the accuracy required.

Activity

A9.3

Find out about one of the Government Surveys carried out in your own country and write up similar details to those given in Figure 9.1.

This should help you understand the problems involved in designing a useful survey and help you with illustrations for your exam questions. (Remember that your understanding of the general points raised here should be illustrated by examples. The examiners are very happy if you give examples from your own country or area of interest. They are not looking for points memorised from textbooks.)

Figure 9.1 Major surveys used in social trends

	Frequency	Sampling frame	Type of respondent	Location	Effective sample size[1] (most recent survey included in Social Trends)	Response rate (percentages)
Adult literacy in Britain Survey	One-off	Postcode Address File	All aged 16 to 65	GB	3,800 individuals	68
British Crime Survey	Biannual	Postcode Address File	Adult in household	EW	19,808 addresses[2]	83
British Household Panel Survey	Annual	Postal Addresses	All adults in household	GB	5,033 households	95[3]
British Social Attitudes Survey	Annual	Postcode Address File	Adult in household	GB	5,374 addresses	68
Census of Population	Decennial	Detailed local Valuation and Lands Agency Property	Household head	UK	Full count	98
Continuous Household Survey	Continuous	Various	All adults in household	NI	4,170 addresses	70
European Community Household Panel Survey	Annual	Postcode Address File in GB, Valuation and Lands Agency	All adults in household	EU	60,000 households	70[4]
Family Expenditure Survey	Continuous	Property in NI	Household	UK	10,173 addresses[2]	62[5]
Family Resources Survey	Continuous	Postcode Address File	All adults in household	GB	26,435 households	70
General Household Survey	Continuous	Postcode Address File	All adults in household	GB	11,845 households	76
Health Education Monitoring Survey	Annual	Postcode Address File	Adults in household	E	7,000	74
Health Survey for England	Continuous	Postcode Address File	Adults and children over 2 years of age	E	11,700 addresses	79[6]
Infant Feeding Survey	Every 5 yrs	Registration of births	Mothers	UK	12,300	74
International Passenger Survey	Continuous	International passengers at ports and airports	Individual traveller	UK	249,000 individuals	86
Labour Force Survey	Continuous	Postcode Address File	All adults in household[7]	UK	63,000 addresses	81[8]
Longitudinal Study	Continuous	Population	All persons	EW	1%	9
National Food Survey	Continuous	Postcode Address File in GB, Valuation and Lands Agency Property in NI	Person responsible for domestic food arrangements	UK	13,144 addresses	65
National Readership Survey	Continuous	Postcode Address File	Adults aged 15 and over	GB	38,500 individuals	62
National Travel Survey	Continuous	Postcode Address File	Household	GB	4,500-5,000 households per year[2]	73[10]
New Earnings Survey	Annual records	Inland Revenue PAYE	Employee[11]	GB	4,500-5,000 households per year[2]	73[10]
Omnibus Survey	Continuous	Postcode Address File	One adult per household	GB	2,700 individuals[12]	70[12]
Survey of English Housing	Continuous	Postcode Address File	Household	E	25,000 addresses	80
Survey of Personal Incomes	Annual	Inland Revenue	Individuals[13]	UK	80,700 individuals	97
Survey of Psychiatric Morbidity	Ad hoc	Postcode Address File	Adults aged 16 to 64	GB	10,108	80

Key E = Europe EW = England and Wales GB = Great Britain

Notes on Figure 9.1

1. Effective sample size includes non-respondents but excludes ineligible households.
2. Basic sample only.
3. Wave on wave response rate at wave four. This represents 77% of respondents at wave one.
4. Response rate estimated. Response rates vary between EU countries.
5. Response rate refers to Great Britain.
6. Response rate for fully and partially responding households.
7. Includes some proxy information.
8. Response rate to first-wave interviews quoted. Response rate to second to fifth wave interviews 95% of those previously accepting.
9. Linkage rates from Census to Census were 91% for Longitudinal Study members present in both the 1971 and 1981 Censuses and 90% for Longitudinal Study members present for both the 1981 and 1991 Censuses.
10. Response rate for the period January 1994 to January 1997.
11. In the New Earnings Survey employers supply data on a 1% sample of employees who are members of PAYE schemes. For the 1997 sample approximately 219,000 were selected and there was a 92.3% response but some 43,000 returned questionnaires did not contain data.
12. The Omnibus Survey changes from month to month. The sample size and response rate are for September 1997.
13. In the Survey of Personal Incomes local tax offices supply data on individuals to a central point in Inland Revenue.

Further detail on random sampling: the multi-stage sample

Stratification

We use stratification in an attempt primarily to reduce our standard errors. We earlier mentioned the idea of ‘random’ as opposed to ‘simple random’ sampling. Clearly simple random sampling can have a high standard error for the variables in which you might be interested.

For example, consider a classroom situation where we study aptitude in Mathematics. Taking a simple random sample from all the students, regardless of their year of study (although they are following a three-year course), might be very misleading. You could easily draw samples whose students were mainly, or all, from one year of the course and hence obtain an inaccurate idea of students’ overall mathematical ability.

In this situation, you could instead, take a simple random sample of students from separate lists for each of the course and thereby ensure that your estimate of mathematical ability is more representative. In fact by doing this (we call it **stratification by year of course**) you have ensured that there will be no extreme samples. The stratified sample is a form of random sampling – each item still has a known non-zero chance of being selected. It is important to choose stratified factors which relate to your study. If they are irrelevant (and you therefore do not gain in accuracy) then you have raised the cost of your survey without improving your product.

You might also wish to stratify because you are interested in the strata themselves. In the example about the students, you may wish to give information of all kinds about students by year of course to the professor responsible for that year. In these circumstances it would, in any case, be wise to target each year to avoid obtaining a sample in which a particular year is not represented.

A third reason for choice of stratification factor may be administrative and lie in the list itself; it may be organised that way. Your registrar or admissions tutor may only have student lists classified by year on the computer and amalgamating them only wastes time and effort!

Now reread the chapter on this topic in Moser and Kalton and make sure you have a clear idea in your mind as to why you might stratify a sample.

Activity

A9.4

Your company has five sites and a mixture of managerial, clerical and computing, and factory workers. You want to know what kind of help they would like with their travel to work arrangements. One of the trade unions involved has asked for free parking for all its members, another has suggested subsidised train travel cards.

You, as statistician, have been asked to make a sample survey of employees to help them decide what to do. You decide to make a random sample; there is no problem with the list and you have been asked to give statistically reliable advice. You decide to make a stratified sample.

a) Give the strata you will use

b) Explain how they will help you.

Hint: Don’t be afraid to use two sets of strata in these circumstances.

Clustering

Cost is another issue in sampling: in a large-scale random survey, you will have to consider how to cut costs. Using clustering is one way.

The interviewer may be told to interview household members or individuals, in a small area, rather than being required to travel long distances to meet the target required by a simple random sample covering a large area. Individual respondents will be identified in a random manner – but **within** an area. You will save cost (the interviewer will be able to complete a higher number of interviews in a given time, and use less petrol and shoe leather) but will probably have to sacrifice a degree of accuracy.

Clustering is clearly useful in an interviewer-administered survey. It is less important as a design feature for telephone or postal (or mail) interviews unless you are interested in the cluster itself.

To the extent that individuals in a cluster are similar (having **intra-class correlation**) they will be less representative of other clusters. In other words, the variance of your sample estimate will be greater.

A further reason, in addition to cost, for cluster sampling, may arise from your need as a researcher to look at the clusters themselves for their own sake. An interest in income and educational levels for a group living in one area, or a study of the children at a particular school and their reaction to a new television programme, will require you to look at individuals in a cluster.

Note in these cases that you could argue that the cluster itself is the item being studied!

Now reread Moser and Kalton to make sure you understand the following:

- intra class correlation
- variability within a cluster
- variability between clusters.

Activity

> A9.5
>
> You have been asked to design a random sample in order to study the way school children learn in your country. Explain the clusters you might choose, and why.

The multi-stage sample

This describes the final design of your random sample. In a large government survey it will incorporate elements of both stratification and clustering at different stages. A typical national multi-stage sample in the UK might involve the following:

1. Dividing the areas of the country into strata by industrial region.
2. Sampling clusters (local areas) from **each** industrial region.
3. From each local area choosing some areas for which you have lists (say electoral register, or post code) and making a random sample from the chosen lists (clusters).

Activity

> A9.6
>
> Your textbook will have a clear and detailed example of a multi-stage survey.
>
> a) Work through it and then find out about one of the government or other large-scale surveys in your country.
>
> b) Identify the stratification factors and the way in which sampling units are clustered for convenience.
>
> Make sure you have an example clearly in your mind.

Method of contact

A further point you should think about when assessing how to carry out a survey is contact method. The most common methods of contact are face-to-face interview, telephone interview, or postal/mail or self-completion interview. Your textbook should give you a lot of detail on this. Try to keep a clear head!

In most countries you can assume the following:

- An interviewer-administered, face-to-face questionnaire will be the most expensive to carry out.
- Telephone surveys depend very much on whether your interviewer target population is on the telephone (and how good the telephone system is).
- Mail/postal questionnaires can have a low response rate.

Your textbook will elaborate further and it is worth making a list of the advantages and disadvantages of each type of contact. Figure 9.2 gives the generally accepted pros and cons of the type of questionnaire you use. Your list would be similar but remember to add any additional factors which apply specifically to your country.

Figure 9.2: Type of contact – advantages and disadvantages

	Advantages	**Disadvantages**
Face-to-face interview	Good for personal questions, probing detail, explaining difficult concepts. Can show samples (e.g. covers of magazines, new products, etc.).	Expensive. Not always easy to obtain detailed information on the spot.
Telephone interview	Easy to achieve a high number of interviews. Easy to check (central switchboard perhaps) on quality of interviewers.	Not everyone has a telephone so the sample can be biased.You cannot usually show samples.
Mail or other self-completion	Most people can be contacted this way (there will be little non-response due to people not being at home). It allows time for people to look up details e.g. income, tax returns, etc.	People are likely not to reply – it is an effort to fill in a form. The answers to some questions may influence answers to earlier questions. This is important where the order of a questionnaire is important. You have no control over who answers the questionnaire.

Examples of occasions when you might use a particular method are:

- **Interview, face to face – a survey of shopping patterns**

  Here you need to be able to contact a sample of the whole population. You can assume that a large proportion would not bother to complete a postal questionnaire (after all, the subject matter is not very important and it takes time to fill in a form!). Using a telephone would exclude those (for example the poor and the elderly) who either do not have access to a phone or are unwilling to talk to strangers by telephone.

- **Telephone – a survey of businessmen and their attitude to a new item of office equipment**

  All of them will have a telephone, the questions should be simple to ask. Here, booking time for a telephone interview at work (once it has been agreed with the administration) should be much more effective than waiting for a form to be filled in, or sending interviewers to disrupt office routine.

- **Postal or mail? – a survey of teachers about their pay and conditions**

  Here, on the spot interviews will not elicit the level of detail needed. Most people do not remember their exact pay and taxation, particularly if they are needed for earlier years. We would expect a high response rate and good quality data. The recipients are motivated to reply (they may be hoping for a pay rise!) and come from a group of people who find it relatively easy to fill in forms without needing the help or prompting of an interviewer.

Remember that it is always possible to combine methods. The Family Expenditure Survey in the UK, for example, combines the approach of using an interviewer three times over a fortnight (to raise response and explain detail) while the respondent household is required to fill in a self-completion diary (showing expenditure, which could not be obtained by interview alone).

Similarly telephone interviews may be combined with a mail shot sub-sampled individual survey, in the case of offices and business faxing additional information. In the case of the telephone survey of businessmen described above, a description of the new equipment could be faxed to the businessmen as they are telephoned.

Remember also that email surveys are becoming popular, though they will only be appropriate when the population to be studied uses them heavily and are likely to reply to your questions. For example employees at your office.

Note: this part of the course is relatively simple to understand. Try to think of as many examples as you can from your country, job, or area of interest.

Activities

A9.7

What form of contact might you use for your questionnaire in the following circumstances:

a) a random sample of school children about their favourite lessons

b) a random sample of households about their expenditure on non-essential items

c) a quota sample of shoppers about shopping expenditure

d) a random sample of bank employees about how good their computing facilities are

e) a random sample of the general population about whether they liked yesterday's TV programmes.

A9.8

a) Outline the main stages of a random survey. Where do the main dangers of errors lie?

b) Why might you carry out a quota survey rather than a random survey?

c) 'The designing of questionnaires and the training of interviewers is a waste of money'. Discuss.

d) When would you carry out a telephone survey rather than using a face-to -face interview?

e) You have been asked to survey the interest of a population in a new type of audio-tape. How might you stratify your sample? Explain.

Summary

This chapter has described the main stages of a survey and the sources of error at each stage. The various techniques in the rest of the guide are of little use unless you understand the sources and limitations of the data you are using. The contents of this chapter should have helped you to understand how statistical techniques you have learned about so far can be used in practice.

Learning outcomes

After working through this chapter and the relevant reading, you should be able to:

- define random, simple random and quota sampling and describe the implications of using them
- explain the reasons for stratifying and clustering samples
- describe the factors which contribute to errors in surveys, including:
  - inaccurate and poorly judged frames
  - sampling error
  - non-sampling error (non-response, biased response, interviewer error)
- discuss the various contact methods that may be used in a survey and the related implications:
  - interviewer
  - postal
  - other self-administered
  - telephone.

You do not need to know the theoretical and mathematical details of random sampling including detailed estimation of sampling error. This part of the syllabus aims to support your understanding of research methods and is not mathematical in detail.

Sample examination questions

1. a) Define a 'quota' sample.

   b) What are the main reasons you would use such a sample, and what are the alternatives?

   b) What are the main sources of error in a quota sample, and how would you deal with them? (10 marks)

2. Given the data from Chapter 6, Question 3(b) and Chapter 7, Question 5 you decide to look at these results further and contact the customers concerned in each company that you have already selected.

   a) Outline your survey procedure, giving and explaining your preferred method of contact and how you would prevent non-response.

   b) Give examples of the questions you might ask. (10 marks)

3. You are carrying out a random sample survey of leisure patterns for a holiday company, and have to decide whether to use interviews at people's homes and workplaces, postal (mail) questionnaires, or telephones. Explain which method you would use, and why. (10 marks)

4. Discuss the statistical problems you might expect to have in each of the following situations:

   a) Conducting a census of population.

   b) Setting up a study of single parent families.

   c) Establishing future demand for post-compulsory education. (10 marks)

Chapter 10

Some ideas underlying causation: the use of control groups and time order

Essential reading

Moser, C.A. and G. Kalton. *Survey Methods in Social Investigation.* (Aldershot: Dartmouth, 1979) second edition [ISBN 0 4358 2604 2]. Chapters 9 and 6.5.

Additional reading

Douglas, J.W. *The Home and the School, a Study of Ability and Attainment in the Primary School.* (St Albans: Panther, 1964). Chapter summaries and appendix.

Shipman, M. *The Limitations of Social Research.* (London: Longman, 1997) fourth edition [ISBN 0 5823 1103 9]. Part 2.

Introduction

So far we have looked at ways of collecting and describing data. Chapters 2 and 9 introduced you to the basic ideas of sampling from populations. The main ways of describing what you have found were given in Chapter 3. Chapters 4 to 7 dealt with the ways we can assess the relevance or significance of these figures and Chapter 8 looked at the idea of assessing data which is being analysed by more than one category. We finally reached the idea of 'association' between variables. Chapter 11 will complete the process so far as this subject guide is concerned, by covering correlation and regression.

Before you do this, it is important to take stock of the limitations of social research. Anyone who studied science at school will be familiar with the idea of an experiment. Subjects are measured (observations are made), a treatment is administered, and further observations are made. Providing that we can be sure that nothing else has happened between observations apart from the treatment (scientists write 'other things being equal'), the assumption is made that the treatment has caused the change between the first and second set of observations.

If we are dealing with a situation like that, the meaning of our statistical measures and work is very clear. However, in the social science, business and management fields, things are rarely that simple. We are generally faced with figures which show changes in variables but the treatment given is not simple to assess.

Take, for example, an advertising campaign for your new munchy bars with improved wrappers. Your company measures the sales of the old munchy bars (*A*) in two areas *X* & *Y* before introducing the new bars (*B*) and running a four-month campaign.

Imagine your marketing manager's joy when *B* is much higher than *A* in area *X*. Clearly the changes have worked. But, oh dear, *B* is achieving lower than *A* for area *Y*. On closer investigation we find that, while the advertising campaign has been going on, main rivals *M* have withdrawn their product from area *X* and concentrated their effort on area *Y* (where they have figured out there is a larger number of their target population). So your success with product B is not related to your advertising campaign but to your rival's actions. Clearly, whatever measures you use, there is a problem with the measuring of your results. Other things have changed while you were conducting your experiment.

It is a good idea to read the recommended sections of Moser and Kalton carefully at this point.

So how do statisticians try to measure causal corrections in the social sciences? They use two main weapons:

- the control group
- time order.

Use of the control group and matching

We are often limited in the social sciences because we are unable to carry out experiments for ethical or practical reasons.

Imagine for example that you need to assess the likely effect on tooth decay of adding fluoride to the water supply in town *X*. There is no question of being allowed to experiment, as you are afraid that fluoride might have harmful side effects. You know that fluoride occurs naturally in some communities. What can you do, as a statistician?

Observation

Here you can look at the data for your *unfluorided* water population and compare it with one of the communities with naturally occurring fluoride in their water and measure tooth decay in both populations.

But be careful! A lot of other things may be different. Are the following the same for both communities:

- Number of dentists per person?
- Number of sweet shops/per person?
- Eating habits?
- Age distribution?

Think of other relevant attributes which may differ between the two. If you can match in this way (i.e. find two communities which are the same in these characteristics and only differ in the fluoride concentration of their water supply) your results may have some significance.

Activity

A10.1

Your government is assessing whether it should change the speed limit on its motorways or main roads. Several countries in your immediate area have lowered their limit recently by 10 miles an hour.

What control factors might you use in order to examine the likely effect on road accidents of a change for your country?

Experimentation

Occasionally it **is** possible and permissible to carry out experiments in business situations. There are still difficulties caused by the sheer number of variables which will come into consideration, but at least it is possible to distinguish those who had the treatment (the experimental group) from those who did not (the control group). You should read the material on this and familiarise yourself with the ideas of **blind** and **double blind** in an experiment, and the use of placebos. On the whole, these methods are associated with medical statistics more often than with other applied areas, but they are also used in marketing and test marketing when possible.[12]

Time order

Another way we might attempt to disentangle causal relationships is to look at the order in which things occurred. Clearly, if we eat more, we gain weight (other things being equal; for example, if we don't exercise more!).

This underpins work on time series which you will meet if you study *Econometrics and Economic Statistics*. For now, you should know a little about **longitudinal** or **panel** surveys, where the same individuals are resurveyed over time.

Longitudinal surveys

Policy-makers use these surveys over a long period to look at the development of childhood diseases, educational development, and unemployment; there are many long-term studies in these areas. Some longitudinal medical studies of rare diseases have been carried out at an international level over long periods. One such very well known study, which is readily available, is the UK National Child Development Survey. This began with a sample of about 5,000 children born in April 1948. It is still going on! It was initially set up to look at the connections between childhood health and development and nutrition by social groups. The figures produced in the first few years were so useful, that it was extended to study educational development and work experience. There are several books which describe the survey at its different stages. The most useful is probably the first, the *Home and the School*, by J.W.B. Douglas as it explains the methods used.

You should note the advantages and disadvantages of using such methods. The big **advantages** are that you:

- can actually measure **individual** change (not just averages)
- do not depend on people's memories about what they did four years ago.

The **disadvantages** are:

- on the one hand, drop out – if the subject material is trivial, people may not agree to be continually resurveyed
- on the other hand, conditioning – if the researcher manages to persuade participants to continue to co-operate, he or she may have altered their perceptions (the participants may have become too involved in the study).

Nevertheless, such studies are widely regarded as being the best way of studying change over time.

Panel surveys

In business and management research we generally use slightly less ambitious surveys called **panels**. They also involve contacting the same individual over a period but are generally different from longitudinal surveys in the following ways:

- they are more likely to be chosen by quota rather than random methods
- individuals are interviewed every 2-4 weeks (rather than every few years)
- individuals are unlikely to be panel members for longer than two years at a time.

We can use results from such surveys to look at brand loyalty and brand switching. It is particularly useful for assessing the effectiveness of advertising (as you will see if you study *Marketing and Market Research*).

For now, make sure you understand how a longitudinal or panel study is set up.

[12] *These ideas will be discussed more fully in* ***Marketing and Market Research****.*

Activity

A10.2

a) List the advantages and disadvantages of making a panel study.

b) Work out how you might set up a panel of children in order to see whether they like the TV programmes your company is making for under 10s in the after school/before homework time slot.

Causation – smoking and cancer of the lung

In order to clarify the issues raised in the preceding pages, now read through Section 9.5 of Moser and Kalton carefully. You will see how control groups were used in order to see what was connected with lung cancer. You might like to know that initially the researchers expected that pollution in the environment would show the strongest link.

They compared patients in hospital diagnosed as having lung cancer with a control group who did not. (They were likely to have had accidents or non-chest diseases such as appendicitis or, in the case of women, be pregnant.) Each lung cancer patient was matched with a control by:

- age
- sex
- occupation
- home area, and
- being in hospital at roughly the same time.

Using the measures you will meet in Chapter 11, it was found that the main difference between the two group was in smoking behaviour: those with lung cancer were more likely to be heavy smokers than the other. As you can see, this study was carried out some time ago, but similar studies have confirmed the results.

But what did the study tell us? Consider again the introduction to this chapter. Descriptive statistics alone cannot distinguish between three ideas:

a) Smoking causes lung cancer

smoking → lung cancer

b) Smoking is a symptom of developing lung cancer

lung cancer → smoking

c) Your personality factor leads to smoking and lung cancer

personality factor X → smoking + lung cancer

It is important that you understand this. Although animal experimentation may help resolve the conundrum to some extent, we are really at stalemate without an experiment on the people in whom we are interested!

At this point, the original researchers carried out a longitudinal study of 40,000 doctors in the United Kingdom. Their smoking histories were collected and subsequent death rates from lung cancer checked. The results confirmed initial findings.

However, though these results make the causal connection ‘heavy smoking lung cancer’ much more likely, it is still possible that the heavy smokers may have some other characteristic which leads them to contract lung cancer.

Ethically it is not possible to choose individuals to participate in an experiment and ask them to smoke heavily, so a pure experiment could not be made.

Summary

This chapter's focus on causation gives you an opportunity to think of the implications of the work you have covered in the earlier chapters and prepare the way for the material in Chapter 11.

Learning outcomes

After working through this chapter and the relevant reading, you should be able to:

- distinguish between an experiment in the natural sciences and the observations possible in social science and business or management studies
- work out sensible controls for a given experiment or trial
- set up a panel study
- explain the advantages and limitations of a longitudinal/panel survey compared with a cross-sectional survey.

You will not need to design multi-factor experiments or causal models.

Sample examination questions

1. What are the strengths and weaknesses of a longitudinal survey? Describe how you would design such a survey if you were aiming to study the changes in people's use of health services over a 20-year period. Give the target group, survey design, and frequency of contact. You should give examples of a few of the questions you might ask. (12 marks)
2. Write notes on the following:
   a) Blind trials
   b) Control groups
   c) Measuring causation. (12 marks)

Notes

Chapter 11

Correlation and regression

Essential reading

Newbold, P. *Statistics for Business and Economics.* (London: Prentice-Hall, 1995) fourth edition [ISBN 0 1385 5549 0]. Chapter 12.

Further reading

Aczel, A.D. *Complete Business Statistics.* (London: Irwin/McGraw Hill, 1999) [ISBN 0 0728 9302 8]. Chapters 10 and 11.

Anderson, D.R., D.J. Sweeney, and T.A. Williams. *Statistics for Business and Economics.* (Cincinatti: South-Western Thomson Learning, 2002) eighth edition [ISBN 0 3240 6671 6]. Chapters 14 and 15.1.

Hanke, J.E. and A.G. Reitsch. *Understanding Business Statistics.* (Burr Ridge Ill: Irwin, 1994) second edition [ISBN 0 2561 1219 3]. Chapters 13 and 14.

Mason, R.D. and D.A. Lind. *Statistical Techniques in Business and Economics.* (Boston: McGraw Hill, 2001) eleventh edition [ISBN 0 0724 0282 2]. Chapters 13, 14 and 15.

Moskowitz, H. and G.P. Wright. *Statistics for Management and Economics.* (London: Charles Merrill Publishers, 1985) [ISBN 0 6752 0211 6]. Chapters 15 and 16.1.

Wonnacott, T.H. and R.J. Wonnacott. *Introductory Statistics.* (Chichester: Wiley, 1990) fifth edition [ISBN 0 4715 1733 X]. Chapters 12, 13 and 14.

Introduction

In Chapter 8 you were introduced to ideas of testing the relationship between different attributes of a variable using the chi-squared distribution. We did this by looking at the numbers of individuals falling into a category, or experiencing a particular contingency.

Correlation and regression are two techniques which enable us to see the connection between the actual dimensions of two or more variables. The work we will do in this chapter will only involve our looking at two variables at a time, but you should be aware that statisticians use these theories and similar formulae to look at the relationship between many variables. **Factor analysis, discriminant analysis, principal component analysis** and some kinds of **cluster analysis** all use related ideas and techniques.[12]

When we use these techniques we are concerned with using models for prediction and decision making. So, how do we model the relationships between two variables? We are going to look at:

- Correlation – which measures the **strength** of a relationship
- Regression – which is a way of representing that relationship.

It is important you understand what these two techniques have in common, but also the differences between them.

[12] *The use of these is covered in* Marketing and Market Research. *Regression techniques are also emphasized in* Elements of Econometrics and Economic Statistics.

Correlation

Correlation is concerned with measurements of the strength of the linear relationship between two variables. We often begin by drawing a graph plotting this relationship. We call it a scatter diagram. Consider the scatter diagrams below showing pairs of observations of X and Y. Let us imagine that X represents the money spent on marketing a new product (in £) and Y represents the value of sales of that product, by week.

Figure 11.1: Scatter diagrams showing cost of marketing and value of sales, by week.

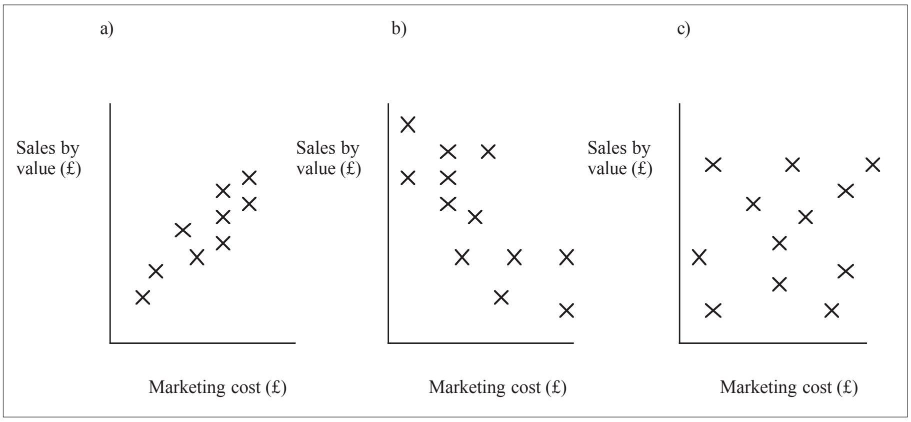


Figure (a) shows a fairly strong direct linear relationship (i.e. as *X* increases *Y* increases).

Figure (b) shows a weaker inverse linear relationship (as *X* increases *Y* decreases).

Figure (c) shows no clear pattern; there is no obvious relationship between *X* and *Y*.

Correlation coefficient

The statistic used to measure the relationship is the **correlation coefficient**. The values of the correlation coefficient for cases (a), (b) and (c) would be something like 0.6, -0.4 and 0.0 respectively.

For random variables *X* and *Y* the **correlation coefficient for a population (*p*)** sometimes also known as the **product moment correlation**) is:

$$\rho = \frac{E(X - E(X))(Y - E(Y))}{\sqrt{V(X)V(Y)}}.$$

We say technically that the population value *p* can only be calculated if we have a perfect knowledge of the bivariate[13] density function of *X* and *Y*. In practice, it is more likely that we will wish to estimate *p* from a set of sample observations of *X* and *Y*:

i.e. the sample $(x_1,y_1),(x_2,y_2)$,.........., (x_n,x_n).

The **sample correlation coefficient (*r*)** is calculated thus:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}.$$

You may find it easier to compute r using this version of the formula:[14]

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)\left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}.$$

[13] A bivariate density function shows the dependence on two factors, in this case x and y.

[14] This is a generally used formula. If the textbook you are using gives a different one, familiarise yourself with that. The answers will be the same!

Both types of correlation coefficient have these properties:

1. it is **independent** of the **scale** of measurement
2. it is **independent** of the **origin** of measurement
3. it is **symmetric** – the correlation of *x* and *y* is the same as the **correlation** of *x* + *y*

and, most important,

4. it only takes the values between zero and plus or minus one; we write $\rho \leq |1|$.

It is important that you understand these basic ideas.

Activity

A11.1

Draw a rough scatter diagram for each of the following situations:

a) r = 0

b) r is very high and positive

c) r is very high and negative

d) r is quite high but negative.

In the case of (ii), (iii) and (iv), think what size r might be. Now let's work through example 12.1.

Example 11.1

A test has been designed to examine a prospective salesman's ability to sell. Some experienced salesmen sit the test and their scores are compared with their actual productivity. Calculate the correlation between test score and productivity. Check the results using your own calculator.

Score (*x*)

(mark out of 50) 41 34 35 40 33 42 37 42 40 43 38 38 46 36 32 43 42 30 41 45

Productivity(*y*)

(in ergos) 32 35 20 24 27 28 31 33 26 41 29 33 36 23 22 38 26 20 30 30

Calculations

n=20 ; Σx =778 ; Σy = 584 ;

$\Sigma\ x^2$ = 30640 ; $\Sigma\ y^2$ = 17704; $\Sigma\ xy$ =23015

$$r = \frac{20\,.\,23015\ -\ 778\,.\,584}{\sqrt{(20\,.\,30640\ -\ 778^2)\,(20\,.\,17704\ -\ 584^2)}} = 0.6012$$

This looks quite a high positive correlation[16].

Spurious correlation

It is important to understand the limitations of correlation as a measure. While we have seen in the previous example a high correlation between test score and productivity, is there a causal connection? External evidence may lead us to think that being an experienced salesman may cause you to have a high score but it is quite possible that people from other occupations (as yet unmeasured) could also have high scores. We have to look at the other evidence and, unless we are carrying out an experiment (as explained in Chapter 10), have no idea what the causal connection is between two variables.

[16] *If you were following **Statistics 2** and were to test for significance, using the kind of techniques and ideas given in Chapters 7 and 8, you would find that the result is significantly high at the 1% value. Note that you are not expected to test for significance for r in **Statistics 1**.*

It is perfectly possible to look for data sets and find high correlations between variables which are unlikely to be causally connected. Here are some examples of high correlations:

a) (Average salary of school teachers) vs (Consumption of alcohol measured by country)

b) (Stork population in Bavaria) vs (Human birth rate)

c) (Size of student population) vs (No. of juvenile offenders by local area in a country)

Would you seriously think there was a causal connection in these cases? Let's look at them in a little more detail.

It would be frivolous to deduce from (a) that respect for teachers causes alcoholism! It is much more likely that buying and consuming alcohol and high salaries are both signs of a flourishing economy.

(b) Is a bit more difficult; can the fairy stories be true? Again there is probably a further variable at play here – are young families living in areas of natural beauty (which encourage storks!)?

As for (c), the more young people there are, the more juvenile offenders, scholarship winners, and students there are likely to be. Connecting these two figures is pretty meaningless.

Activity

A11.2

Think of an example where you feel the correlation is clearly spurious and explain how it might arise.

Now think of a 'clear' correlation and the circumstances in which you might accept causality.

Rank correlation

Just as we saw in Chapter 3 that we can use the median and quartile range for measures of location and dispersion instead of the mean and standard deviation, or variance, it is possible to calculate the **rank correlation coefficient** (generally know as **Spearman's rank correlation coefficient**).

We can do so by ranking the x_i and y_i in ascending order and using the formula given in the last section for the **ranks** of the x_i and y_i rather than their actual values. It may be that we **only** have the rankings in any case.

If there are no ties in the rankings of x_i and y_i, then it is easy to calculate the following:

$$r_s = 1 - 6\frac{\sum_{i=1}^{i=n} d_i^2}{n\left(n^2 - 1\right)}$$

where d_i are the differences in the ranks between each x_i and y_i. This is the same, and equivalent to P for ranks.

Here is a simple example using rank correlation.

Example 11.2

Following the work shown in Example 12.1, the company asks a new research assistant to administer the sales aptitude test on the 10 sales staff recruited in the last year. Instead of putting achieved scores in the computer, the assistant ranks the individuals in order, starting with the lowest scores and productivity measures.

Here are the figures:

Staff members	A	B	C	D	E	F	G	H	I	J
Rank order in test	2	3	5	1	4	9	10	6	7	8
Rank order in productivity	1	2	3	4	5	6	7	8	9	10

Does it look as if the test is a good predictor for less experienced staff?

Using the formula is very simple as there are no ties. We get:

d_i	1	1	2	-3	1	3	3	-2	-2	-2
d_i^2	1	1	4	9	1	9	9	4	4	4

and thus $\sum_{i=1}^{i=8} d_i^2 = 46$.

Using the formula:

$$
\begin{aligned}
r_s &= 1 - 6 \times \frac{46}{10 \times 99} \\
&= 1 - .2788 \\
&= .72
\end{aligned}
$$

which looks quite high.

As with other order statistics, such as the median and quartile, it is helpful to use Spearman's rank correlation coefficient if you are worried about the effect of extreme observations on our sample.

The limits for r_s are the same as for r (it must be between –1 and +1).

Activity

A11.3

The following figures give examination and project results (in percentages) for eight students.

a) Find the Spearman's rank correlation coefficient for the data.

b) Compare it with the Pearson rank correlation coefficient, as calculated in Example 11.1.

Students examination and project marks

	1	2	3	4	5	6	7	8
Examination	95	80	70	40	30	75	85	50
Project	65	60	55	50	40	80	75	70

Regression

So much for correlation. What about regression? It is important to distinguish regression from correlation at the outset. Remember, we are here looking at a description of the **relationship** between *x* and *y* and not just its **strength.**

We start with a scatter diagram between two variables as before. This time, we want to know what line will best fit the data. The theory we learn here assumes we are going to use a straight line, and not a curve of any kind, though in some disciplines (physics or finance, for example) a curve would be more appropriate.

We have to find the line of best fit. Before we can do this, we must assume that one variable is dependent on the other. By convention we call the **dependent variable *y*** and the **independent variable *x***. We have to work out the slope of the line, and the point at which it cuts the *y* (or *x*) axis. Again, by convention, we call these values β and α respectively for the population. (See more about the equation of a straight line in Chapter 1). The basic model is therefore given as follows.

The model of a random variable *Y*, the dependent variable, which is related to random variable *X*, the independent (or predictor or explanatory) variable by the equation:

$$Y = \alpha + \beta X + \varepsilon$$

where α and β are constants and $\varepsilon \sim N(0, \sigma^2)$, a random error term. The coefficients α and β are **theoretical** values (like *p*) and can only be **estimated** from sample data. The estimates are generally written as a and b, rather than α and β .

Given a sample of bivariate data, $(x_1,y_1),(x_2,y_2)$,.........., (x_n,y_n), a and b can be estimated. To fit a line to some data as in this case, an objective must be chosen to define which straight line best describes the data. The most common objective is to minimise the sum of the squared distance between the observed value of y_i and the value predicted by the *y*.

The estimated least squares regression line is written as:

$$y = \alpha + \beta x$$

We can derive the formulae for *b* and *a*:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$, and

$$a = \bar{y} - b\bar{x} .$$

The line is called the **sample regression line** of *y* on *x*.[16]

Example 11.3 demonstrates the calculation of *a* and *b* and the use of the resultant equation to estimate *y* for a given *x*.

Example 11.3

A study was made by a retailer to determine the relation between weekly advertising expenditure and sales (in thousands of pounds). Find the equation of a regression line to predict weekly sales from advertising. Estimate weekly sales when advertising costs are £35,000.

Adv. Costs (in £'000)	40	20	25	20	30	50	40	20	50	40	25	50
Sales (in £'000)	385	400	395	365	475	440	490	420	560	525	480	510

[16] *Note that the derivations of these are given in numerous textbooks. You do not need to be able to reproduce this bookwork.*

Calculations representing sales as *y* and advertising costs as *x* give:

n=12; Σx =410; Σy = 5445 ; Σx^2 = 15650 ; Σy^2 = 2512925; Σxy = 191325

b = (12 × 191325 – 410 × 5445) / (12 × 15650 – 410^2) = 3.221

a = (5445 – 3.221 × 410) / 12 = 343.7

Thus the estimated equation is:

$y = 343.7 + 3.221\,x$

and the estimated sales for £35,000, worth of advertising is:

$(y_i \mid x_i = 35) = 343.71 + 3.221 \times 35 = 456.4$ or £456,400.

Points to watch about linear regression

Non-linear relationships

Note first that you have only learned how to use a straight line for your best fit. So you could be missing quite important non-linear relationships, particularly if you were working in the natural sciences.

Which is the dependent variable?

Note also that it makes a difference which is your dependent variable. In Example 11.2 above, you would have a different line if you had taken advertising costs as *y* and sales as *x*. (Imagine a situation where, when sales were low, more advertising resulted the following week!)

Activity

A11.4

Work out the *b* and *a* in Example 11.3 given advertising cost as the dependent variable. Now predict advertising costs when sales are £460,000. Make sure you understand how and why your results are different from the example in the text!

Extrapolation

In Example 11.3 and Activity 11.3, you used your estimated line of best fit to work out *y* given *x*. This is only acceptable if you are dealing with figures which lie within the data set. You cannot use the figures to go beyond the existing range of the data.

It is not so clear in Example 11.2 that the relationship between advertising expenditure and sales could change, but a moment's thought should convince you that, were you to quadruple expenditure, you would be unlikely to get a nearly 13 times rise in sales! Sometimes it is very easy to see that the relationship must change. For example, consider Example 11.4, showing an anthropologist's figures on years of education of mother and number of children she has, based on a Pacific Island.

Example 11.4

Figures from our anthropologist show a negative relationship between the number of years education of the mother and the number of live births she has. The regression line is:

$y = 8 - .6\,x$

based on figures on women with between 5 and 8 years education who had 0 to 8 live births (*y* is the number of live births and *x* the number of years of education). This looks sensible. We predict 8 – 3 = 5 births for those with 5 years of education and 8 – 6 = 2 births for those with 10 years of education.

This is all very convincing, but say a woman on the island went to university and completed a doctorate and had 15 years of education. She clearly cannot have minus 1 children! And, if someone missed school entirely, is she likely to have 8 children? We have no way of knowing. The relationships shown by our existing figures will probably not obtain beyond their boundaries.

Activity

A11.5

Try to think of a likely linear relationship between *x* and *y* which would probably work over some of the data but then breakdown like that in Example 11.3.

This should make sure you understand the difference between interpolation (which statisticians do all the time) and extrapolation which they should not.

Points to note about correlation and regression

First, a warning!

Do not be tempted to rely on your calculator and not bother to learn the formulae. Examiners frequently give you the $\sum x_i, \sum x_i^2, \sum y_i, \sum y_i^2$ and $\sum x_i y_i$ in order to save you computation time. If you don't know how to take advantage of this, you will waste valuable time which you really need for the rest of the question. Note that if you use your calculator, show no working and get the answer wrong you are unlikely to be given credit. This part of the syllabus leads directly into *Econometrics and Economic Statistics* so it is important you understand it.

What is the relationship between correlation and regression?

As you have seen, all the calculations we do for the two involve similar summation measures so they must be connected in some way. This is indeed the case.

For example a high r^2 means that the standard error of *b* will be low (in other words, if there is a strong connection between *x* and *y*, the points will be close to the line of best fit). A low r^2 means that the points are scattered (a high standard error for *b*).

While you are not expected to carry out hypothesis tests or build confidence intervals for *a, b* and *r* (these are all part of *Statistics 2*), it is important that you understand these ideas.

On the other hand be sure you are clear that a high *r* does not mean that the slope of the regression line *b* is high.

Activity

A11.6

Draw a rough scatter diagram with the line of best fit for the following *b, a* and r^2 .

a) $a = 2$, $b = ½$, $r =\ .9$

b) $a = -3$, $b = -\ 2$, $r = -\ .3$.

Multiple correlation and regression

The ideas you have met could be extended to look at the relationship between more than two variables. Much of the work you will meet in social planning or market research uses such concepts. Such techniques are discussed a little more in *Statistics 2* and their use and limitations are part of the *Marketing and Market Research* syllabus.

Make sure you understand the basic variable cases introduced here so that you can understand the general concepts when you meet them, either further on in this degree, or in the world of work.

Summary

This brings us to the end of our introduction to dealing with the estimation of two variables (bi-variate statistics)

Learning outcomes

After working through this chapter and the relevant reading, you should be able to:

- draw and label a scatter diagram
- calculate r and r^2.
- explain the meaning of a particular value and the general limitations of r and r2 as measures
- calculate b and a for the line of best fit in a scatter diagram
- explain the relationship between a, b and r
- illustrate the problems caused by extrapolation.

You are not required to make confidence intervals or test hypotheses for r, a or b. That is part of the **Statistics 2** syllabus. You are not required to explain techniques such as factor analysis or multiple regression, which are covered in a descriptive manner in **Marketing and Market Research**.

Sample examination questions

1. State whether the following statements are true or false and explain

   a) The correlation between X and Y is the same as the correlation between Y and X.

   b) If the slope is negative in a regression equation Y = a + b X, then the correlation coefficient between X and Y would be negative too.

   c) If two variables have a correlation coefficient of minus 1 they are not related.

   d) A large correlation coefficient means the regression line will have a high slope b. (8 marks)

2. The following table shows the number of computers, in thousands (x) produced by a company each month and the corresponding monthly costs in £'000 (y) for running its computer maintenance department.

Number of computers (in thousands) x	**Maintenance costs (£'000)** y
7.2	100
8.1	116
6.4	98
7.7	112
8.2	115
6.8	103
7.3	106
7.8	107
7.9	112
8.1	111

Note that the following statistics have been calculated from these data:

Sum of x values = 75.5

Sum of y values = 1080

Sum of squares of x values = 573.33

Sum of squares of y values = 116 988

Sum of squares of x values and y values = 8184.9

a) Draw the scatter diagram. (4 marks)

Calculate the correlation coefficient for computers and maintenance costs.

(3 marks)

c) Compute the best fitting straight line for y and x. (3 marks)

d) Comment on your results. How would you check on the strength of the relationships you have found? (4 marks)

Appendix

Sample examination paper

Candidates should attempt **all** questions in **Part A** and **two** of the three questions in **Part A**.

A copy of *New Cambridge Elementary Statistical Tables* (Tables 4 to 10) is provided.

Marks will be deducted for insufficient explanation within your answers.

Part A

1. a) Calculate the mean, median, mode and quartile deviation for the following data:

   6, 14, 18, 6, 22, 17, 18, 19, 16, 17, 16. (5 marks)

   b) Define each of the following briefly:

   – Type I error

   – Level of significance

   – P value. (6 marks)

   c) A tea machine may be defective because it dispenses the wrong amount of tea (T) and/or the wrong amount of sugar (S). The probabilities of these defects are:

   p(T) = 0.025, p(S) = 0.04, p(T and S) = 0.005.

   What proportion of cups of tea have:

   – at least one defect

   – no defects. (4 marks)

   d) If $n = 5$, $x_1 = 5$, $x_2 = 4_{,}$ $x_3 = 1$, $x_4 = 3$, $x_5 = 2$, find:

   1) $\frac{1}{n}\sum_{i=1}^{i=n} x_i$

   2) $\sum_{i=2}^{i=4} x_i^{2}$

   3) If *X* takes the above values with probabilities given by:

   i)=2/5, p(2)=2/15, p(3)=4/15, p(4)=1/10, p(5)=10

   find E(*X*). (5 marks)

   e) For each of the following statements (i)-(iv) state whether they are true or false and give a brief explanation:

   i) A sample mean is always an unbiased estimator of the population mean in simple random sampling.

   ii) The distance between the first quartile and the third quartile is called the standard deviation.

   iii) If the constant *a* in a regression equation Y=*a* + *b*X is high then so is the correlation coefficient *r* between *x* and Y.

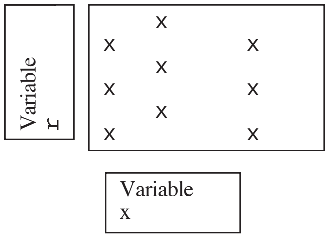
   iv) The higher the calculated chi-squared, the more likely the relationship between the variables measured to be significant. (8 marks)

f) i) State the conditions under which you would use a student's *t* distribution rather than a normal distribution.

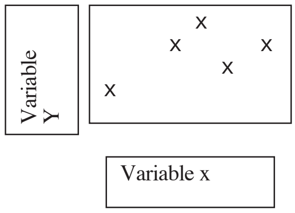
ii) Has a *t* distribution with 10 degrees of freedom a larger standard deviation than one where the degrees of freedom are 4? Explain how you came by your conclusion. (4 marks)

g) For each of the scatter plots below, give an indication of a likely value for the correlation coefficient *r* between *x* and Y:

(i)



(ii)



(2 marks)

h) i) Why does it matter whether you use a one-sided or a two-sided alternative when testing hypotheses?

ii) Would you use a one-sided or a two-sided alternative if asked to find whether families in one area spend more on food than families in another? Explain briefly.

iii) You are told that, when the appropriate test is made, families in area A spend significantly more than those in area B at the 10% level, but not at the 1% level. There are 9 degrees of freedom. Give the 10% and 1% points and comment. (6 marks)

2. a) The data in the following table give the number of parking fines issued in one month in three different areas by five traffic wardens T, R, A, F and C.

	Fine Location		
Warden %	Main Street	Shop Street	Market Street
T	130	300	190
R	1160	250	150
A	210	210	170
F	230	210	240
C	90	190	270

Calculate the percentage of all fines issued by each traffic warden.

What is the probability that a randomly chosen fine was issued by warden F in Market Street?

Given that a fine was issued in Main Street, what is the probability that it was issued by warden C?

Given that a fine was issued by warden F, what is the probability that it was in Shop Street? (10 marks)

b) Looking at these figures, your boss asks you to measure whether there is a connection between traffic wardens and area in the parking fines issued. You decide to carry out a chi-squared test.

i) Outline the stages in your work assuming that you are testing at the 5% level of significance. Give the hypotheses, degrees of freedom and appropriate significance value. Give the formula for Q^2 (the observed chi-squared value) and the method of calculating expected values. (You need not do all the working!)

ii) You find, on first calculating Q^2, that your figures are not significant at the 5% level, but are at the 10% level. What should you tell your boss?

iii) To your embarrassment, when you check your calculations, you find that your figures are significant both at 5% and 1% levels. How does this change the picture? (7 marks)

c) A dice is thrown 180 times. The number of times each face shows is as follows:

Face of dice	1	2	3	4	5	6
Number of throws	28	36	36	30	27	23

Do these data suggest the dice is unfair? (8 marks)

3. a) The output (in thousands) and profit per unit of output, is given in the following table:

Output (X thousands)	**Profit per unit of output (Y)**
5	1.7
7	2.4
9	2.8
11	3.4
13	3.7
15	4.4

$$\left(\sum x = 60, \sum x^2 = 670, \sum xy = 202, \sum y^2 = 61\right)$$

i) Draw the scatter diagram of Y on X.

ii) Calculate b and a and draw the resultant line on your scatter diagram.

iii) Use your regression equation to estimate the profit per unit of output when the output is 10,000.

iv) Why would it be wrong to use the equation to estimate the profit per unit of output for an output of 20,000?

v) Calculate the correlation coefficient r and explain its meaning. (15 marks)

b) You are carrying out a random sample survey of the hours people spend in their workplace and have to decide whether to use interviews at their home or workplace, postal (mail) questionnaire or telephones. Explain which method you would choose and why. (10 marks)

4. a) A survey is carried out to estimate the total number of alcohol units consumed by a population of adults in a week; 150 are sampled, and the mean number of alcohol units consumed in the sample is found to be 15 with standard deviation 4.

   i) Find a 95% confidence interval for the mean number of units consumed in the population.

   ii) A census of the whole population is taken and the population mean and standard deviation are found to be 16 and 4 respectively. Do you think the sample was a random sample?

   iii) You now have population figures for adult alcohol consumption. Government guidelines suggest that it would be desirable for people to drink fewer than 12 units a week. What proportion of the population currently do so?

   iv) It is known that it is dangerous to health to consume more than 25 units a week. How many adults, out of a population of two million, do you think do so? (10 marks)

   b) You take a sample of 250 people from a neighbouring area and get a mean of 12 and standard deviation of 4. Test the hypothesis that the mean adult alcohol consumption for the neighbouring population is less than for your original population. Test at two levels of significance and explain your results.
   (8 marks)

   c) i) The prices of a computer game for a random sample of eight retailers have a sample mean of £15.70 and a sample standard deviation of £2.20. Determine a 95% confidence interval for the mean price of all such retailers.

   ii) You are now told the population variance is known and equals 13. What is now the 95% confidence interval for the mean price of retailers?
   (3 marks)

   iii) And in the same circumstances, what is the 90% confidence interval?
   (7 marks)

Notes

Comment form

We welcome any comments you may have on the materials which are sent to you as part of your study pack. Such feedback from students helps us in our effort to improve the materials produced for the External Programme.

If you have any comments about this guide, either general or specific (including corrections, non-availability of essential texts, etc.), please take the time to complete and return this form.

Name ______________________________

Address ______________________________

____

Email ______________________________

Student number ______________________________

For which qualification are you studying? ______________________________

Comments ______________________________

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

______________________________Please continue on additional sheets if necessary.

Date: ______________________________

Please send your comments on this form (or a photocopy of it) to:
Managing Editor, External Programme, University of London, 34 Tavistock Square, London WC1H 9EZ UK.