UNIVERSITY OF LONDON | INTERNATIONAL PROGRAMMES

Statistics 1

J.S. Abdey

ST1**04a**

2014

Undergraduate study in
Economics, Management, Finance and the Social Sciences

This is an extract from a subject guide for an undergraduate course offered as part of the University of London International Programmes in Economics, Management, Finance and the Social Sciences. Materials for these programmes are developed by academics at the London School of Economics and Political Science (LSE).

For more information, see: www.londoninternational.ac.uk

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

This guide was prepared for the University of London International Programmes by:

James S. Abdey, BA (Hons), MSc, PGCertHE, PhD, Department of Statistics, London School of Economics and Political Science.

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.

University of London International Programmes
Publications Office
Stewart House
32 Russell Square
London WC1B 5DN
United Kingdom
www.londoninternational.ac.uk

Published by: University of London

Contents

Chapter 1
Introduction

1.1 General introduction to the subject area

Welcome to the world of statistics! This is a discipline with unparalleled applicability whose use can be found in a wide range of areas such as finance, business, management, economics and other fields in the social sciences. **ST104a Statistics 1** provides you with the opportunity to grasp the fundamentals of the subject and will equip you with the vital quantitative skills and powers of analysis which are highly sought-after by employers in many sectors.

The material in this course is necessary as preparation for other courses you may study later on as part of your degree or diploma; indeed, in many cases statistics is a compulsory course on our degrees.

- In particular, it has links with **SC1021 Principles of sociology** and **MN3141 Principles of marketing**.
- You may also choose to take **ST104b Statistics 2** or **MT2076 Management mathematics** so that you can study the concepts introduced here in greater depth. A natural continuation of this course and **ST104b Statistics 2** are the advanced courses **ST3133 Advanced statistics: distribution theory** and **ST3134 Advanced statistics: statistical inference**.
- You may wish to develop your economic statistics by taking **EC2020 Elements of econometrics**.
- You may want to build on your interests in social research and take **SC2145 Social research methods**.
- You will also find these techniques valuable as a geographer by taking this course with **GY1148 Methods of geographical analysis**.

So the reach of statistics is considerable. Hence it rightly forms a core component of the EMFSS programmes since all of the courses mentioned above require an understanding of the concepts and techniques introduced in **ST104a Statistics 1**. The analytical skills which you will develop on this course will thus stand you in very good stead both for your future studies and beyond into the real world of work.

1.2 Aims of the course

The emphasis of this 100 half course is on the application of statistical methods in management, economics and the social sciences. Attention will focus on the

interpretation of tables and results and the appropriate way to approach statistical problems. Treatment is at an elementary mathematical level. Ideas of probability, inference and multivariate analysis are introduced and are further developed in the half course **ST104b Statistics 2**.

1.3 Learning outcomes

At the end of the course, and having completed the essential reading and activities, you should:

- be familiar with the key ideas of statistics that are accessible to a student with a moderate mathematical competence
- be able to routinely apply a variety of methods for explaining, summarising and presenting data and interpreting results clearly using appropriate diagrams, titles and labels when required
- be able to summarise 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
- be able to perform inference to test the significance of common measures such as means and proportions and conduct chi-squared tests of contingency tables
- be able to use simple linear regression and correlation analysis and know when it is appropriate to do so.

1.4 Syllabus

This course introduces some of the basic ideas of theoretical statistics, emphasising the applications of these methods and the interpretation of tables and results.

Basic background: Elementary summation signs, elementary probability, Venn and tree diagrams.

Data collection: Elements of survey design, the stages of a survey, ideas of randomness, observation and experiment.

Data presentation and analysis: Descriptive statistics, measures of location and dispersion, pictorial and graphical representation.

The normal distribution: Estimation of mean, proportion, standard deviation, confidence intervals and hypothesis testing. Ideas of testing for differences between means and proportions. The use of Student's t.

Goodness-of-fit tests: The chi-squared distribution and contingency tables.

Regression and correlation: An introduction to the ideas of regression and correlation, least squares, estimation of a, b and r, scatter diagrams.

1.5 Essential reading list and other learning resources

Numerous titles are available covering the topics frequently covered in foundation statistics courses such as **ST104a Statistics 1**. Due to the doubtless heterogeneity among students taking this course, some may find one author's style easier to comprehend than another's.

That said, the recommended textbook for this course is:

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060].

This textbook contains many additional examples, complete with solutions to selected even-numbered exercises. To ensure you fully understand each topic, you should attempt these, and check your answers against the solutions provided. In addition, the textbook introduces all the topics in an alternative style to this subject guide so if you have difficulty following a specific point, you may wish to consult the textbook for further clarification. A non-exhaustive list of other suggested textbooks is outlined below under 'Further reading'.

Statistical tables will be provided in the examination for **ST104a Statistics 1**, hence you should also purchase:

- Lindley, D.V. and W.F. Scott *New Cambridge Statistical Tables*. (Cambridge: Cambridge University Press, 1995) second edition [ISBN 9780521484855].

These tables form the basis of the tables which are provided for you to use in the examination. It is essential that you familiarise yourself with these tables in advance rather than those printed in your textbook, as the method of navigating through the tables is not uniform across all publications. In order to prepare yourself, you should concentrate in particular on using Tables 4, 5, 7, 8, 9 and 10 of the *New Cambridge Statistical Tables*. These relate to the (standard) normal, Student's t and chi-squared distributions.

Detailed reading references in this subject guide refer to the editions of the set textbooks listed above. New editions may have been published by the time you study this course. You can use a more recent edition of any of the books; use the detailed chapter and section headings and the index to identify relevant readings. Also check the virtual learning environment (VLE) regularly for updated guidance on readings.

1.5.1 Further reading

As previously mentioned, numerous alternative textbooks are available to purchase, although references in this subject guide will be to the eighth edition of Newbold et al. If you do decide to purchase a different/additional textbook, do ensure it covers all the syllabus topics (listed earlier in this introductory chapter). A word of caution at this point: statistical notation can differ between textbooks so be prepared for this.

Please note that as long as you read the Essential reading you are then free to read around the subject area in any textbook, paper or online resource. You will need to

support your learning by reading as widely as possible and by thinking about how these principles apply in the real world. To help you read extensively, you have free access to the VLE and University of London Online Library (see below). Other useful textbooks for this course include:

- Aczel, A.D. *Complete Business Statistics*. (London: McGraw-Hill Higher Education, 2009) seventh edition [ISBN 9780071287531].
- Anderson, D.R., D.J. Sweeney, T.A. Williams, J. Freeman and E. Shoesmith *Statistics for Business and Economics*. (South-Western Cengage Learning, 2010) eleventh edition [ISBN 9780324783247].
- Lind, D.A., W.G. Marchal and S.A. Wathen *Statistical Techniques in Business and Economics*. (Boston: McGraw-Hill Higher Education, 2009) fourteenth edition [ISBN 9780073401768].
- Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics for Business and Economics*. (Chichester: John Wiley & Sons, 1990) fourth edition [ISBN 9780471615170].

Chapters 9 and 10 of the subject guide concentrate on the concepts of **surveys and experimentation**. You will find these topics particularly useful if you are studying **SC1021 Principles of sociology** or **MN3141 Principles of marketing**. For those who wish to research these areas in more detail, we recommend the following:

- Shipman, M. *The Limitations of Social Research*. (London: Longman, 1997) fourth edition [ISBN 9780582311039].

In addition, *Social Trends* (a compendium of UK **official statistics and surveys**) is useful when you work on Chapter 9.

- Office for National Statistics *Social Trends*. (Basingstoke: Palgrave Macmillan, 2009) [ISBN 9780230220508].

If you feel that you need to refresh your **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 9780521559133] Chapters 1, 2 and 7.

1.6 How to study statistics

1.6.1 Mathematical background

To study and understand statistics you will need some familiarity with abstract mathematical concepts combined with the common sense to see how to use these ideas in real-life applications. The concepts needed for probability and statistical inference are impossible to absorb by just reading them in a book – although you may find you need

to do this more than once! You need to read, then think a little, then try some problems, then read and think some more. This process should be repeated until you find the problems easy to do. **Reading without practising with the examples and activities set in this subject guide will be of little help.**

You will also need to be able to use high-school arithmetic and understand some basic algebraic ideas. These are very important. Starting with them should help you feel comfortable with this course from the outset and we introduce you to these ideas in Chapter 2.

1.6.2 Calculators and computers

A calculator may be used when answering questions on the examination paper for **ST104a Statistics 1** and it must comply in all respects with the specification given in the Regulations. You should also refer to the Admission Notice you will receive when entering the examination and the 'notice on permitted materials'.

The most important thing is that you should accustom yourself to using your chosen calculator and feel comfortable with it. Specifically, calculators must:

- be hand-held, compact and portable
- be quiet in operation
- have no external wires
- be non-programmable
- not be capable of receiving, storing or displaying user-supplied non-numerical data.

If you are aiming to carry out serious statistical analysis (which is beyond the level of this course) you will probably want to use some statistical software package such as Minitab or SPSS. It is **not** necessary for this course to have such software available, but if you do have access to it you could profit from using it.

1.7 How to use the subject guide

1.7.1 Using the subject guide and the textbook

This subject guide has been structured such that it is tailored to the specific requirements of the examinable material. It is 'written to the course', unlike textbooks which typically include additional material which will not be examinable. Therefore, the subject guide should act as your principal resource and you should refer to the specific sections in Newbold et al. as indicated.

A key benefit of the textbook is that it contains a wealth of further examples and exercises which can be used to check and consolidate your understanding. As previously mentioned, solutions to selected even-numbered exercises are provided in the textbook enabling you to check your answers.

1.7.2 Online study resources

In addition to the subject guide and the Essential reading, it is vitally important that you take advantage of the study resources that are available online for this course, including the VLE and the Online Library.

You can access the VLE, Online library and your University of London email account via the Student Portal: http://my.londoninternational.ac.uk

You should have received your login details for the Student Portal with your official offer, which was emailed to the address that you gave on your application form. You have probably already logged in to the Student Portal in order to register. As soon as you registered, you will automatically have been granted access to the VLE, Online Library and your fully functioning University of London email account.

If you have forgotten these login details, please click on the 'Forgotten your password' link on the login page.

The Student Portal forms an important part of your study experience with the University of London and should therefore be accessed regularly.

1.7.3 Virtual learning environment

The VLE, which complements this subject guide, has been designed to enhance your learning experience, providing additional support and a sense of community. In addition to making printed materials more accessible, the VLE provides an open space for you to discuss interests and to seek support from other students, working collaboratively to solve problems and discuss subject material. In a few cases, such discussions are driven and moderated by an academic who offers a form of feedback on all discussions. In other cases, video material, such as audio-visual tutorials, are available. These will typically focus on taking you through difficult concepts in the subject guide. For quantitative courses, such as Mathematics and Statistics, fully worked-through solutions of practice examination questions are available. For some qualitative courses, academic interviews and debates will provide you with advice on approaching the subject and examination questions, and will show you how to build an argument effectively.

Past examination papers and *Examiners' commentaries* from the past three years are available to download which provide advice on how each examination question might best be answered. Self-testing activities allow you to test your knowledge and recall of the academic content of various courses and where possible sessions from previous years' Study Weekends have been recorded. Finally, a section of the VLE has been dedicated to providing you with expert advice on preparing for examinations and developing digital literacy skills.

Unless otherwise stated, all websites in this subject guide were accessed in January 2014. We cannot guarantee, however, that they will stay current and you may need to perform an internet search to find the relevant pages.

1.7.4 Making use of the Online Library

The Online Library contains a huge array of journal articles and other resources to help you read widely and extensively.

To access the majority of resources via the Online Library you will either need to use your University of London Student Portal login details, or you will be required to register and use an Athens login: http://tinyurl.com/ollathens

The easiest way to locate relevant content and journal articles in the Online Library is to use the **Summon** search engine.

If you are having trouble finding an article listed on the reading list, try removing any punctuation from the title, such as single quotation marks, question marks and colons.

For further advice, please see the online help pages:
www.external.shl.lon.ac.uk/summon/about.php

1.7.5 Structure of the subject guide

Statistics is fundamentally a *cumulative* discipline – that is, **the following 11 chapters are not a series of self-contained units, rather they build on each other sequentially**. As such, you are strongly advised to follow the subject guide in chapter order, understanding the material in the early chapters before embarking on later material. There is little point in rushing past material which you have only partially understood in order to reach the later chapters.

1.7.6 Time management

About one-third of your private study time should be spent reading and the other two-thirds doing problems. (Note the emphasis on practising problems!) We normally recommend that if you are intending to study **ST104a Statistics 1** over the course of one academic year, then you would need to devote a **minimum** of seven hours per week to your studies.

To help your time management, we have converted the chapters and topics of this course into **approximate** weeks to devote to each subject if you were, for example, spending 20 weeks on **ST104a Statistics 1**. 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 2 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 2	–	2 weeks
Chapter 3	–	1 week
Chapter 4	–	2 weeks
Chapter 5	–	1 week
Chapter 6	–	1 week
Chapter 7	–	1 week
Chapter 8	–	1 week
Chapter 9	–	1 week
Chapter 10	–	2 weeks
Chapter 11	–	1 week
Chapter 12	–	2 weeks
Revision(!)	–	5 weeks

1.7.7 Recommendations for working through the chapters

The following procedure is recommended for each chapter:

1. Carefully read the aims of the chapter and the learning outcomes.
2. Read the introduction, noting carefully the sections of Newbold et al. to which you are referred.
3. Now work through each section of the chapter making sure you can understand the examples given. In parallel, watch the accompanying video tutorials for each section.
4. Review the intended learning outcomes carefully, almost as a checklist. Do you think you have achieved your targets?
5. Attempt the learning activities which appear near the end of each chapter.
6. Attempt the chapter's self-test quizzes on the VLE.
7. Attempt the sample examination questions given at the end of the chapter. You can review the video solutions, but do so **after** attempting the questions yourself!
8. When you have finished the material in the chapter, try the suggested questions from Newbold et al. You can treat these as additional activities. This time, though, you will have to think a little about which part of the new material you have learnt is appropriate to each question.
9. If you have problems at this point, go back to the subject guide and work through the area you find difficult again. Don't worry – you will improve your understanding to the point where you can work confidently through the problems.
10. Once you have completed your work on the whole subject guide, you will be ready for examination revision. Start with work on the Sample examination paper which you will find at the back of the subject guide.

The last few steps are most important. It is easy to think that you have 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 (refer to the 'Time management' section above). Note that we have given worked examples and activities to cover each substantive topic in the subject guide. The Newbold et al. examples are added for further consolidation of the whole chapter topic and also to help you work out exactly what the questions are about! One of the problems students sometimes have in an examination is that they waste time trying to understand to which part of the syllabus particular questions relate. These final questions, together with the further explanations on the VLE, aim to help with this before you tackle the sample examination questions at the end of each chapter.

Try to be disciplined about this: don't look up the answers until you have done your best. Promise? 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 – honestly!

So to conclude, perseverance with problem solving is your passport to a strong examination performance. Attempting (ideally successfully!) all the cited exercises is of paramount importance.

1.8 Examination advice

Important: the information and advice given in the following section are based on the examination structure used at the time this subject guide was written. Please note that subject guides may be used for several years. Because of this, we strongly advise you to check both the current Regulations for relevant information about the examination, and the VLE where you should be advised of any forthcoming changes. You should also carefully check the rubric/instructions on the paper you actually sit and follow those instructions.

This half course is assessed by a two-hour, unseen, written examination. No books may be taken into the examination, but you will be provided with the necessary extracts from the *New Cambridge Statistical Tables* and a formula sheet (which can be found in past examination papers on the VLE). These will be supplied as part of the examination question paper for **ST104a Statistics 1** rather than being provided separately. A calculator may be used when answering questions on this paper and it must comply in all respects with the specification given in the Regulations.

Section A is a series of short compulsory questions worth 50 per cent of the total marks. In Section B, you should attempt two out of a choice of three longer questions, each of which is worth 25 per cent of the total marks. As Section A will seek to assess a broad cross section of the syllabus, we strongly advise you to study the whole syllabus. A sample examination paper is provided at the end of this subject guide.

Students should be aware that graph paper will be supplied as part of the examination question paper for this half course, so they will not need to request it from the invigilator. Students will need to detach this graph paper from the back of the question paper and tie any sheets that they use into their examination answer booklets. Students should therefore practise using this type of graph paper as part of their preparation for the examination.

Examiners will expect this graph paper to be used where indicated and they will be

looking for greater precision in the answers to these questions than in recent years.

Strategies for success, your academic and study skills handbook, also provides additional guidance on examination technique.

Remember, it is important to check the VLE for:

- up-to-date information on examination and assessment arrangements for this course
- where available, past examination papers and *Examiners' commentaries* for the course which give advice on how each question might best be answered.

Chapter 2
Mathematical revision – simple algebra and coordinate geometry

2.1 Aims

This chapter outlines the essential mathematical building blocks which you will need to work with in this course. Most of them will be revision to you but some new material is introduced. Particular aims are:

- to be familiar with the basic rules of arithmetic operations
- to learn further notation concerned with absolute value and 'greater than' ($>$) and 'less than' ($<$) signs
- to use summation signs to calculate simple statistical measures
- to learn how to plot a straight line graph and identify its slope and intercept.

2.2 Learning outcomes

After completing this chapter, and having completed the Recommended reading and activities, you should be able to:

- manipulate arithmetic and algebraic expressions using the simple rules
- recall and use common signs: square, square root, 'greater than', 'less than' and absolute value
- demonstrate use of the summation sign and explain the use of the 'i', or index, of x
- draw the straight line for a linear function.

2.3 Recommended reading

- Anthony, M. and N. Biggs *Mathematics for Economics and Finance*. (Cambridge: Cambridge University Press, 1996) [ISBN 9780521559133] 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. If you

are studying for **MT105a Mathematics 1**, that course will support your studies in **ST104a Statistics 1**. However, you will not need all of the material in **MT105a Mathematics 1** here.

In addition there is essential 'watching' of this chapter's accompanying video tutorials accessible via the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

2.4 Introduction

This opening chapter introduces some basic concepts and mathematical tools on which the rest of the course is built. Before proceeding to the rest of the subject guide, it is essential that you have a solid understanding of these fundamental concepts and tools.

You should be confident users of the basic mathematical operations: addition, subtraction, multiplication and division, and be able to use these operations on a basic electronic calculator. The content of Chapter 1 is expected to be a 'refresher' of the elementary algebraic and arithmetic rules from schooldays. Some material featured in this chapter may be new to you, such as summation signs and graphs of linear functions. If so, you should master these new ideas before progressing.

Finally, remember that although it is unlikely that an examination question would test you on the topics in this chapter alone, the material covered here may well be an important part of the answer!

2.5 Arithmetic operations

We begin with elementary arithmetic operations which will be used when working with figures in **ST104a Statistics 1**. Students often let themselves down by understanding the statistical concepts, but fail to manage a problem because they are unable to deal with the required arithmetic. Although this is not primarily an arithmetic paper, many calculations will be used, so it is vital to ensure that you are comfortable with the examples and activities presented here.

The acronym to remember is 'BODMAS', which tells us the correct order (that is, the priority) in which mathematical operations are performed:

- **B**rackets
- **O**rder (i.e. powers, square roots, etc.)
- **D**ivision
- **M**ultiplication
- **A**ddition
- **S**ubtraction.

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 $a \times b = a \cdot b$
- the **quotient** of a and b means a divided by b, i.e. a/b.

Example 2.1

What is $(35 \div 7 + 2) - (4^2 - 8 \times 3)$?

'BODMAS' tells us to work out brackets first. Here there are two sets of brackets, so let us do them one at a time:

- First bracket: $35 \div 7 + 2$.
- Do division first: $35 \div 7 + 2 = 5 + 2$.
- Then perform the addition: $5 + 2 = 7$.
- Second bracket: $4^2 - 8 \times 3$.
- Do order first: $4^2 - 8 \times 3 = 16 - 8 \times 3$.
- Next do multiplication: $16 - 8 \times 3 = 16 - 24$.
- Then perform the subtraction: $16 - 24 = -8$.

Now the problem has been simplified we complete the calculation with the final subtraction: $7 - (-8) = 7 + 8 = 15$. Note the two negatives become positive!

2.6 Squares and square roots

The *power* is the number of times a quantity is to be multiplied by itself. For example, $3^4 = 3 \times 3 \times 3 \times 3 = 81$. Any number raised to the power 2 is called 'squared', hence x^2 is 'x squared', which is simply $x \times x$.

Remember that squared values, such as x^2, are always non-negative. This is important, for example, when we compute the quantity s^2 in Chapter 4 or r^2 in Chapter 12 which involve squared terms, so a negative answer should ring alarm bells telling us a mistake has been made!

It might be helpful to think of the **square root** of x (denoted $\sqrt{x}$) as the reverse of the square, such that $\sqrt{x} \times \sqrt{x} = x$. Note positive real numbers have two square roots: $\pm\sqrt{81} = \pm 9$, although the positive square root will always be used in **ST104a Statistics 1**. In practice, the main problems you will encounter involve taking square roots of numbers with decimal places. Be careful that you understand that 0.9 is the square root of 0.81 and that 0.3 is the square root of 0.09 (and **not** 0.9!). Of course, in the examination you can perform such calculations on your calculator, but it always helps to have an idea of what the answer should be as a feasibility check of your answer!

2.7 Fractions and percentages

A **fraction** is part of a whole and can be expressed as either:

- common fractions: for example 1/2 or 3/8, or
- decimal fractions: for example 0.5 or 0.375.

In the common fraction, the top number is the *numerator* and the bottom number is the *denominator*. In practice, decimal fractions are more commonly used.

When multiplying fractions together, just multiply all the numerators together to obtain the new numerator, and do the same with the denominators. For example:

$$\frac{4}{9} \times \frac{1}{3} \times \frac{2}{5} = \frac{4 \times 1 \times 2}{9 \times 3 \times 5} = \frac{8}{135}.$$

Percentages give an alternative way of representing fractions by relating a particular quantity to the whole in parts per hundred. For example, 60% is 60 parts per 100, which, as a common fraction, is simply 60/100.

2.8 Some further notation

2.8.1 Absolute value

One useful sign in statistics is | | which denotes the **absolute value**. This is the numerical value of a real number regardless of its sign (positive or negative). The absolute value of x, sometimes referred to as the *modulus* of x, or 'mod x', is $|x|$. So $|7.1| = |-7.1| = 7.1$.

Statisticians sometimes want to indicate that they only want to use the positive value of a number. For example, let the distance between town X and town Y be 5 miles. Suppose someone walks from X to Y – a distance of 5 miles. A mathematician would write this as +5 miles. Later, after shopping, the person returns to X and the mathematician would record him as walking −5 miles (taking into account the direction of travel). Hence this way the mathematician can show the person ended up where he started. We, however, may be more interested in the fact that the person has had some exercise that day! So we need notation to indicate this. The absolute value enables us to take only the positive values of our *variables*. The distance, d, from Y to X may well be expressed mathematically as -5 miles, but you will probably be interested in the absolute amount, so $|-d| = d$.

2.8.2 Inequalities

An **inequality** is a mathematical statement that one quantity is greater or less than another:

- $x > y$ means 'x is greater than y'

- $x \geq y$ means 'x is greater than or equal to y'
- $x < y$ means 'x is less than y'
- $x \leq y$ means 'x is less than or equal to y'
- $x \approx y$ means 'x is approximately equal to y'.

2.9 Summation operator, $\sum$

The **summation operator**, $\sum$, is likely to be new to many of you. It is widely used in statistics and you will come across it frequently in **ST104a Statistics 1**, so make sure you are comfortable using it before proceeding further!

Statistics is all about data analysis, so to use statistical methods we need data. Individual observations are typically represented using a subscript notation. For example, the heights of n people[1] would be called $x_1, x_2, \ldots, x_n$, where the subscript denotes the order in which the heights are observed (x_1 represents the height of the first person, etc.). Hence x_i represents the height of the ith individual and, in order to list them all, the subscript i must take all integer values from 1 to n. So the whole *set* of observations is $\{x_i : i = 1, \ldots, n\}$ which can be read as 'a set of observations x_i such that i goes from 1 to n'.

Summation opeartor, $\sum$

The *sum* of a set of n observations, that is $x_1 + x_2 + \cdots + x_n$, may be written as:

$$\sum_{i=1}^{i=n} x_i \tag{2.1}$$

by introducing the summation operator, $\sum$ (the Greek capital letter sigma), which can be read as 'the sum of'. Therefore $\sum_{i=1}^{i=n} x_i$ is read 'sigma x_i, for i equals 1 to n'.

So the summation is said to be over i, where i is the **index of summation** and the range of i, in (2.1), is from 1 to n. So the lower bound of the range is the value of i written underneath $\sum$, and the upper bound is written above it. Note the lower bound can be any integer (positive, negative or zero), such that the summation is over all values of the index of summation in step increments of size one from the lower bound to the upper bound.

As stated above, $\sum$ appears frequently in statistics. For example, in Chapter 4 you will meet *descriptive statistics* including the arithmetic mean of observations which is defined as:

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

[1] Throughout this course, n will denote a *sample size*.

As seen in this example, rather than write out $\sum_{i=1}^{i=n} x_i$ in full, when all the x_is are summed we sometimes write short-cuts, such as $\sum_{i=1}^{n} x_i$, or (when the range of summation is obvious) just $\sum x_i$.

Note that the resulting sum does not involve i in any form. Hence the sum is unaffected by the choice of letter used for the index of summation, so, for example:

$$\sum_{i=1}^{n} x_i = \sum_{j=1}^{n} x_j = \sum_{k=1}^{n} x_k.$$

Sometimes the way that x_i depends on i is known. For example, if $x_i = i$, we have:

$$\sum_{i=1}^{3} x_i = \sum_{i=1}^{3} i = 1 + 2 + 3 = 6.$$

However, do **not** always assume $x_i = i$!

Example 2.2 If $\{x_i : i = 1, \ldots, n\}$ is a set of observations, we might observe $x_1 = 4$, $x_2 = 5$, $x_3 = 1$, $x_4 = -2$ and $x_5 = 9$. Then:

$$\begin{aligned} \sum_{i=1}^{4} x_i^2 &= 4^2 + 5^2 + 1^2 + (-2)^2 = 46 \\ \sum_{i=4}^{5} x_i(x_i - 2) &= \sum_{i=4}^{5} \left(x_i^2 - 2x_i\right) = ((-2)^2 - 2 \times -2) + (9^2 - 2 \times 9) = 71 \end{aligned}$$

remembering to use BODMAS in the second example.

2.10 Graphs

In Chapter 4 you will spend some time learning how to present material in graphical form, and also in the representation of the *normal distribution* in Chapter 6. You should make sure you have understood the following material. If you are taking **MT105a Mathematics 1**, you will need and use these ideas. If you are not, you are encouraged to read up on and practise examples in Anthony and Biggs.

When a variable y depends on another variable x, we can represent the relationship mathematically using *functions*. In general we write this as $y = f(x)$, where f is the rule which allows us to determine the value of y when we input the value of x. **Graphs** are diagrammatic representations of such relationships, using coordinates and axes. 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 begin with the x-axis and y-axis as shown in Figure 2.1.

We then plot all points of the form $(x,\ f(x))$. Therefore, at x units from the origin (the point where the axes cross), we plot a point whose height above the x-axis (that is, whose y-coordinate) is $f(x)$, as shown in Figure 2.2.

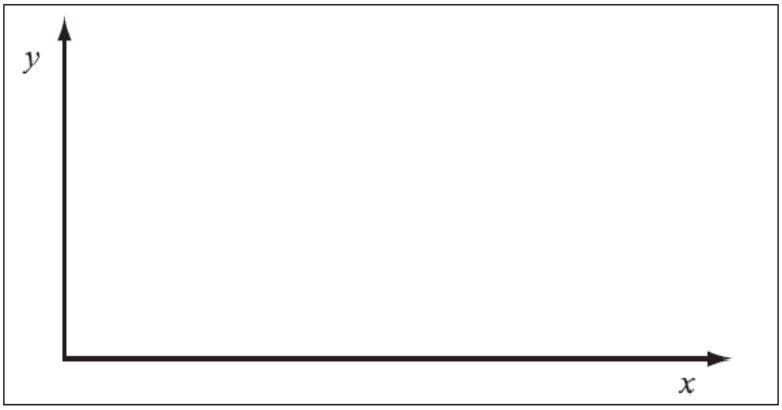


Figure 2.1: Graph axes.

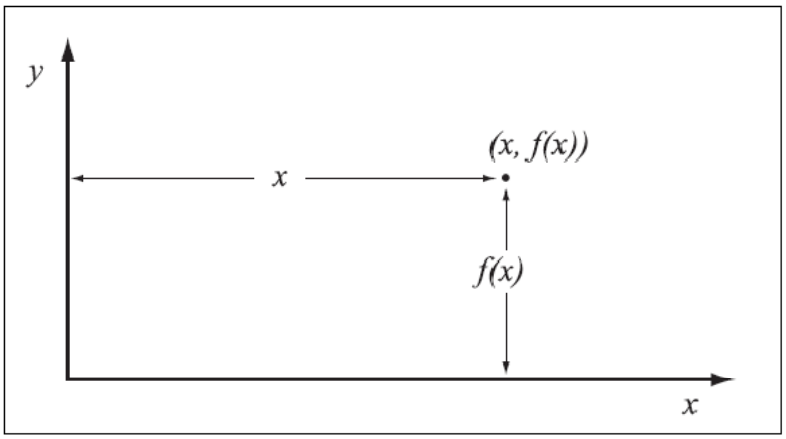


Figure 2.2: Example of a plotted coordinate.

Joining all points together of the form $(x,\ f(x))$ results in a curve (or sometimes a straight line), which is called the graph of $f(x)$. A typical curve might look like that shown in Figure 2.3.

However, 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 in practice and more sophisticated techniques are needed. There are two function types which you need to know about for this course:

- **linear functions** (i.e. the graph of a straight line, see below), and
- **normal functions** (which we shall meet frequently in later chapters).

2.11 The graph of a linear function

Linear functions are those of the form $f(x) = mx + c$ and their graphs are straight lines which are characterised by a gradient (or slope), m, and a y-intercept (where $x = 0$) at the point $(0,\ c)$.

A sketch of the function $y = 2x + 3$ is provided in Figure 2.4, and the function $y = -x + 2$ is shown in Figure 2.5.

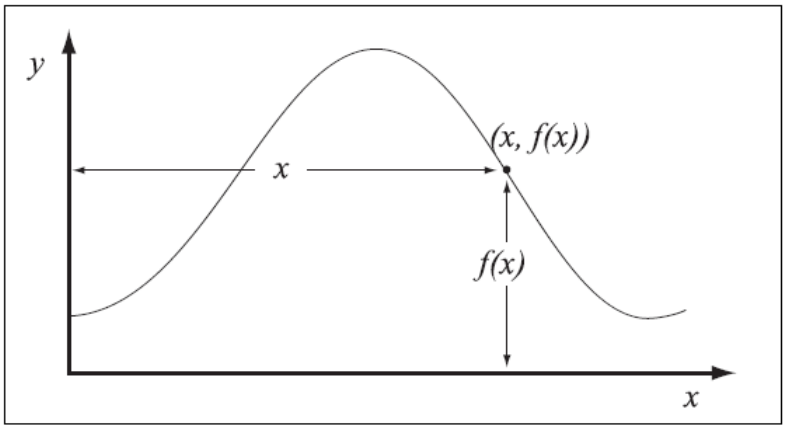


Figure 2.3: The graph of a generic function, $y = f(x)$.

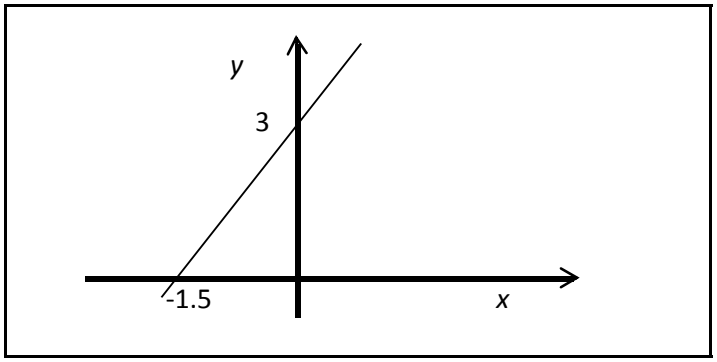


Figure 2.4: A sketch of the linear function $y = 2x + 3$.

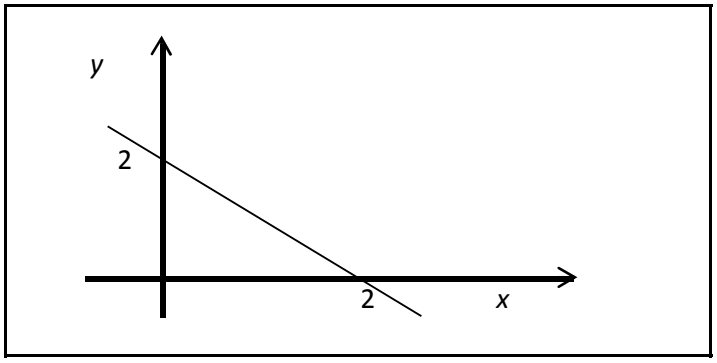


Figure 2.5: A sketch of the linear function $y = -x + 2$.

2.12 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 learning activities below, understood what you have done and, if necessary, worked through additional examples in the basic textbooks.

2.13 Key terms and concepts

- Absolute value
- Fractions
- Linear function
- Power
- Summation operator
- BODMAS
- Inequalities
- Percentages
- Square root

2.14 Learning activities

1. Work out the following:
   (a) $(2 + 4) \times (3 + 7)$
   (b) $1/3$ of $12 - 4 \div 2$
   (c) $(1 + 4)/5 \times (100 - 98)$.

   If you find these difficult at all, go to Anthony and Biggs, or your old school textbook and work on some more examples before you do anything else.

2. Work out the following (use a calculator where necessary):
   (a) $\sqrt{16}$
   (b) $(0.07)^2$
   (c) $\sqrt{0.49}$.

3. (a) What is 98% of 200?
   (b) Give 17/25 as a percentage.
   (c) What is 25% of 98/144?

4. Give the absolute values for:
   (a) $|-8|$
   (b) $|15 - 9|$.

5. (a) For which of 2, 3, 7 and 9 is $x > 3$?
   (b) For which of 2, 3, 7 and 9 is $x < 3$?
   (c) For which of 2, 3, 7 and 9 is $x \leq 3$?
   (d) For which of 2, 3, 7 and 9 is $x^2 \geq 49$?

6. Given $x_1 = 3$, $x_2 = 1$, $x_3 = 4$, $x_4 = 6$ and $x_5 = 8$, find:
   (a) $\sum_{i=1}^{5} x_i$
   (b) $\sum_{i=3}^{4} x_i^2$.

   Given also that $p_1 = 1/4$, $p_2 = 1/8$, $p_3 = 1/8$, $p_4 = 1/3$ and $p_5 = 1/6$, find:
   (c) $\sum_{i=1}^{5} p_i x_i$
   (d) $\sum_{i=3}^{5} p_i x_i^2$.

   If you find these difficult, go back to an elementary textbook and do some more work on this. It is most important that you deal with this before you embark on the topic of descriptive statistics, such as means, in Chapter 4.

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 12.

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

In addition, attempt the 'Test your understanding' self-test quizzes available on the VLE.

2.15 A reminder of your learning outcomes

After completing this chapter, and having completed the Recommended reading and activities, you should be able to:

- manipulate arithmetic and algebraic expressions using the simple rules
- recall and use common signs: square, square root, 'greater than', 'less than' and absolute value
- demonstrate use of the summation sign and explain the use of the 'i', or index, of x
- draw the straight line for a linear function.

2.16 Sample examination questions

1. Suppose $x_1 = 4$, $x_2 = 1$ and $x_3 = 2$. For these figures, give:

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

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

   (a)

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

   (b)

$$\frac{1}{n}\sum_{i=1}^{4} x_i^2.$$

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

Chapter 3
The nature of statistics

3.1 Aims

This chapter gives a general introduction to some of the statistical ideas which you will be learning about in this course. It should also enable you to link different parts of the syllabus and see their relevance to each other and also to the other courses you are studying. You should aim to:

- have an overview of the ideas to be introduced in later chapters
- connect the ideas of sampling to real-life problems
- be able to decide to take topics in a different sequence from the subject guide where this is more appropriate for your overall study plan.

3.2 Learning outcomes

After completing this chapter, and having completed the Essential reading and activities, 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
- explain the principle of random sampling
- identify what the parameter(s) of a function is (are)

and you should then 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 in Chapters 10 and 11 of the subject guide.

3.3 Essential reading

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060] Chapter 1 and Section 6.1.

3.4 Further reading

- Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics for Business and Economics*. (Chichester: John Wiley & Sons, 1990) fourth edition [ISBN 9780471615170] Chapter 1.

If you are taking **SC1021 Principles of sociology**, read Chapter 2 of that subject guide too and enrich your knowledge of both subjects! If you are not studying for **SC1021 Principles of sociology** you might like to find this chapter online and read it through.

3.5 Introduction

This chapter sets the framework for the subject guide. Read it carefully, because the ideas introduced are fundamental to this subject, and, although 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 2 before you begin work on later chapters.**

3.6 Coverage of the course

ST104a Statistics 1 is divided into six main sections:

1. **Data presentation** (Chapter 4).
   - This covers: Ways of summarising data; measures of location and dispersion; graphical presentation.
2. **Probability** (Chapter 5).
   - This covers: Probability theory; probability trees.
3. **The normal distribution and ideas of sampling** (Chapter 6).
   - This covers: Definitions of random variables; expectations and variances; normal distribution; central limit theorem.
4. **Decision analysis** (Chapters 7, 8 and 9).
   - This covers: Estimation and confidence intervals for means, proportions and differences; Student's t distribution; hypothesis testing; Type I and Type II errors; contingency tables.
5. **Sampling design** (Chapter 10).
   - This covers: Key features of different data sampling techniques.

6. **Modelling for decision making** (Chapters 11 and 12).
    - This covers: Measures of correlation; spurious correlation; model selection; model testing.

In the presence of a real-world problem, the aim is to develop a quantitative framework for decision-making under uncertainty. The first step is to conduct an experiment and collect data on which to base the decisions (Chapter 10). The second step is to explore the data (Chapter 4) and the third step is to assign good probability models (Chapters 5 and 6). The final step is the task of statistical inference (or else decision analysis) and this uses the material of Chapters 7, 8 and 9, and possibly also Chapters 11 and 12.

3.7 Terminology

Before progressing to Chapter 4, it is sensible to spend time familiarising yourself with some important terminology and ideas which you will meet later on in the course (and also in **ST104b Statistics 2**, for those studying that as well).

In brief, statistics is a mathematical science concerned with analysing and interpreting *data*. Data collection is not a zero-cost exercise – costs involve both time and money and, in the real world, time and budget resources are finite! (If we did perform a total enumeration of the population, then this is called a **census**.) Who might collect data? In principle anyone – governments, corporations, individuals etc.

As is so often the case in statistics, some words have technical meanings that overlap with their common everyday use but are not the same. It is important that you are aware of these so that you can read and use them correctly.

3.7.1 Population

Population is one such word. Most people think of the word as being about the number of people living in a country. Statisticians, however, use it more generally to refer to the collection of items[1] which we would like to study. So, for example:

- if we were studying children aged two to five years old we would refer to them as the study 'population' and not consider older or younger children, or their parents
- if we were making an analysis of toy retailers in a country, we might look at the 'population' of toy shops.

Often, it is not feasible to collect full information on the entire population due to the size of many populations[2] (and the time and financial constraints described above). Hence, we collect a **sample** drawn from the population (see below).

It can sometimes be difficult to decide which population should be sampled. For instance, if we wished to sample n listeners to a radio station specialising in music,

[1] Note 'items' need not be restricted to people – researchers may wish to study a variety of different populations: fish, countries, dictators, businesses, alumni, criminals, politicians, ...

[2] Indeed, the population size may well be finite, but unknown. For example, how many fish live in the sea?

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 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 – that is, our **survey population** may differ from our **target population** (of course, in an ideal world the two would be the same).

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 persons. Think about this carefully; it is particularly important to those of you studying sociology or marketing.

3.7.2 Bias

Besides the common-sense meaning of **bias**, in statistics there is a more technical meaning.[3] Intuitively, bias is something we should aim to minimise (if not eliminate entirely). A full discussion of bias will be given later in the subject guide, but for now remember that the term **systematic bias** refers to an estimation methodology which results in systematic errors, i.e. consistently over- or under-estimating some population *parameter* (discussed shortly). It would be difficult to justify a methodology that was biased, hence sample survey methodology places great emphasis on unbiased surveys and unbiased estimators.

3.7.3 Parameters

At this point it is useful to think a little about **parameters**. We can define a parameter (or set of parameters) as a measure (or set of measures) which completely describes a function.

For example, in the linear equation $y = mx + c$, the parameters[4] are m and c which completely determine the position of a straight line in a graph (where m denotes the slope, and c the y-intercept). Hence as m, c, or both vary, then we obtain different lines.

Consider another application: you are asked to draw a circle. How do you do this – what things do you need to know in order to draw it? A little thought makes us realise that, of course, you need to know the *centre* of the circle. You also need to know how big it is – we need to know its *radius*. If we only knew the centre we could literally draw millions of different, but concentric, circles; if we knew the radius, but not the centre, then we could again draw millions of circles each of the same size, but with different

[3] It is covered in more technical detail in **ST104b Statistics 2** when exploring the properties of *estimators*.

[4] Mathematical textbooks will typically use m and c as the parameters of a line; however, in statistics we typically use α and β for the intercept and slope, respectively. The linear function is then written as $y = \alpha + \beta x$. Of interest later on will be the *estimation* of these parameters when their values are unknown – this is achieved (for a line) in **regression** (Chapter 12).

positions in space. There is, however, only *one* circle which has a given centre *and* radius. This very interesting idea of finding a minimum number of properties (attributes) to define a shape was first described by the Greek mathematical philosophers who gave them their names.

In Chapter 6 you will meet the *normal distribution*. One of the reasons statisticians like to use the normal distribution is that, despite its complicated functional form, it is completely determined by two parameters: its **mean**, μ, and **variance**, σ^2.[5] If we know these two parameter values, we can draw a unique normal curve. As you will discover, this is extremely useful for working out probabilities and confidence intervals, and for testing hypotheses (part of *statistical inference* – covered in Chapters 7 and 8).

3.7.4 Sampling

As discussed above, a census is not usually a feasible option; however, we may wish to know the value of some attribute of a particular population (specifically, the value of a population parameter). In the absence of population data, we resort to drawing a **sample** from the population of interest, and use the sample observations (our data) to *estimate* a certain parameter when its true value is unknown (due to the absence of population-level data).[6]

Of course, although sampling avoids the costs associated with a census, it is possible to make an argument against the practice. One could argue that since some members of the population will be excluded from the sample, then the process is inherently undemocratic. However, although the omission of some population members will certainly give rise to **sampling error**, sampling might well be more accurate, since more time can be spent verifying the sample information collected. The key aim of sampling is to select a *representative sample* from the population. Chapter 10 will explore various sampling techniques in depth. A brief summary of the main themes is outlined below.

Random sampling

In order to collect a random sample, the researcher needs a **sampling frame** – a list of all population members. For example, if you want to investigate UK secondary schools, the sampling frame would be a list of all the secondary schools in the UK. The key advantages of random sampling are:

- it avoids systematic bias, and
- it allows an assessment of the size of the sampling error.

Access to a sampling frame allows a *known, non-zero probability* of selection to be attached to each population unit. This is the main feature of a random sample, hence the collection of random sampling techniques is known as *probability sampling*. A **simple random sample** is a special case where each population unit has an *equal*,

[5] Concepts formally introduced in Chapter 4.

[6] This is what we term *statistical inference*, that is inferring 'things' (such as parameter values) about a population based on sample data.

known, non-zero probability of selection. 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)
- **cluster** (where precision may not be better, but time and money may be saved on interview surveys involving travel to the site of the interview).

Randomisation (the process by which a random sample is selected) is popular since it avoids the biases caused by the prejudices of the person taking the sample. Of course, just by chance, a random sample may have the appearance of extreme bias – for example, drawing from a population of men and women produces a sample containing men only. Such difficulties are usually resolved with some suitable form of restricted randomisation. One might for instance, in this example, use a *stratified* random sample (by gender) and select half the sample from all men and the other half from all women.

Random sampling is carried out, for many populations, by choosing at random *without replacement*[7] a subset of the very large number of population units, N.[8] If the sample is a sufficiently small proportion of the population (n/N is sufficiently small), then there is no appreciable difference between the inferences possible using sampling with replacement and sampling without replacement.

3.8 Summary

This chapter has laid out the coverage of **ST104a Statistics 1** detailing the six main sections of the syllabus. Many frequently cited terms in statistics have been introduced providing an introductory overview of the issue of sampling – specifically, probability sampling techniques. These will be explored in much greater depth in Chapter 10. Also, the concept of parameters was introduced which is extremely important since statistical inference (covered in Chapters 7, 8, 9 and 12) is concerned with the estimation and testing of unknown parameters.

3.9 Key terms and concepts

- Census
- Parameter
- Sample
- Sampling frame
- Stratified sampling
- Cluster sampling
- Population
- Sampling error
- Simple random sample
- Systematic bias

[7]Sampling without replacement means that once a population unit is selected it is permanently removed from the remaining population, which ensures no population unit is repeatedly observed.

[8]That is, N denotes the population size.

3.10 Learning activities

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

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

3. What parameters could you find for a circle?

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

In addition, attempt the 'Test your understanding' self-test quizzes available on the VLE.

3.11 Further exercises

If you are taking **SC1021 Principles of sociology**, this would be a good time to work through the activities in Chapter 2 of that subject guide.

3.12 A reminder of your learning outcomes

After completing this chapter, and having completed the Essential reading and activities, 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
- explain the principle of random sampling
- identify what the parameter(s) of a function is (are).

3.13 Self assessment

This chapter is an introduction to ideas you will meet again later. Check that you understand the content of the syllabus.

If you want to study the ideas of sampling and causation in greater detail now before studying the theoretical statistical concepts in the next chapters, you could go straight to Chapters 10 and 11, before returning to Chapter 4.

Chapter 4
Data presentation

4.1 Aims

This chapter contains two separate but related themes, both to do with the understanding of data. These are:

- to find graphical representations for data, which allow one to see their most important characteristics
- to calculate simple numbers, such as the mean or interquartile range, which will summarise those characteristics.

4.2 Learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- calculate the following: arithmetic mean, median, mode, standard deviation, variance, quartiles, range and interquartile range
- explain the use and limitations of the above quantities
- draw and interpret: histograms, stem-and-leaf diagrams, boxplots and cumulative frequency distributions
- incorporate 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.

4.3 Essential reading

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060] Sections 1.3, 1.5 and 2.1–2.3.

In addition there is essential 'watching' of this chapter's accompanying video tutorials accessible via the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

4.4 Further reading

- Aczel, A.D. *Complete Business Statistics*. (London: McGraw-Hill Higher Education, 2009) seventh edition [ISBN 9780071287531] Chapter 1.
- Anderson, D.R., D.J. Sweeney, T.A. Williams, J. Freeman and E. Shoesmith *Statistics for Business and Economics*. (South-Western Cengage Learning, 2010) eleventh edition [ISBN 9780324783247] Chapters 1–3.
- Lind, D.A., W.G. Marchal and S.A. Wathen *Statistical Techniques in Business and Economics*. (Boston: McGraw-Hill Higher Education, 2009) fourteenth edition [ISBN 978007110004] Chapters 3 and 4.
- Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics for Business and Economics*. (Chichester: John Wiley & Sons, 1990) fourth edition [ISBN 9780471615170] Chapter 2.

4.5 Introduction

Both themes considered in this chapter (graphical representations for data and calculating values which summarise characteristics of data) could be applied to population[1] data, but in most cases (namely here) they are applied to a sample. The notation would change a bit if a population was being represented. Most graphical representations are very tedious to construct in practice without the aid of a computer. However, you will understand much more if you try a few by hand (as is commonly asked in examinations). You should also be aware that spreadsheets do not always use correct terminology when discussing and labelling graphs. It is important, once again, to go over this material slowly and make sure you have mastered the basic statistical definitions introduced *before* you proceed to more theoretical ideas. Make sure that, once you have completed the chapter and its activities, you make time to practise the recommended questions from Newbold et al.

4.6 Types of variable

Many of the questions for which people use statistics to help them understand and make decisions involve types of variables which can be measured. Obvious examples include height, weight, temperature, lifespan, rate of inflation and so on. When we are dealing with such a variable – for which there is a generally recognised method of determining its value – we say that it is a **measurable variable**. The numbers which we then obtain come ready-equipped with an order relation, i.e. we can always tell if two measurements are equal (to the available accuracy) or if one is greater or less than the other.

Data[2] are obtained on any desired **variable**. For most of this course, we will be dealing with variables which can be partitioned into two types.

[1] Refer back to the 'Terminology' in Section 3.7.

[2] Note that the word 'data' is plural, but is very often used as if it was singular. You will probably see both forms written in textbooks.

Types of variable

1. **Discrete** data: things you can *count*. Examples: number of passengers on a flight and telephone calls received each day in a call centre. Observed values for these will be $0, 1, 2, \ldots$ (i.e. non-negative integers).
2. **Continuous** data: things you can *measure*. Examples: height, weight and time, all of which can be measured to several decimal places.

Of course, before we do any sort of data analysis, we need to collect data. Chapter 10 will discuss a range of different techniques which can be employed to obtain a sample. For now, we just consider some examples of situations where data might be collected:

- pre-election opinion poll asks 1,000 people about their voting intentions
- market research survey asks homemakers how many hours of television they watch per week
- census[3] interviewer asks each householder how many of their children are receiving full-time education.

4.6.1 Categorical variables

A polling organisation might be asked to determine whether, say, the political preferences of voters were in some way linked to their food preferences: for example, do supporters of Party X tend to be vegetarians? Other market research organisations might be employed to determine whether or not users were satisfied with the service which they obtained from a commercial organisation (a restaurant, say) or a department of local or central government (housing departments being one important instance).

This means that we are concerned, from time to time, with **categorical variables** in addition to measurable variables. So we can count the frequencies with which an item belongs to particular categories. Examples include:

(a) The number of vegetarians who support Party X.

(b) The total number of vegetarians (in a sample).

(c) The number of Party X supporters who are vegetarians.

(d) The total number of Party X supporters (in a sample).

(e) The number of diners at a restaurant who were dissatisfied/indifferent/satisfied with the service.

In cases (b) and (d) we are doing simple **counts**, within a sample, of a single category, while in cases (a) and (c) we are looking at some kind of cross-tabulation between variables in two categories: dietary v. political preferences in (a), and political v. dietary preferences in (c) – they are not the same!

[3]Recall that a census is the total enumeration of a population, hence this would not be a sample.

There is no obvious and generally recognised way of putting political or dietary preferences in order (in the way that we can certainly say that $2.28 < 2.32$). It is similarly impossible to **rank** (as the technical term has it) many other categories of interest: in combatting discrimination against people, for instance, organisations might want to look at the effects of gender, religion, nationality, sexual orientation, disability, or whatever, but the whole point of combatting discrimination is that different 'varieties' within each category cannot be ranked.

In case (e), by contrast, there is a clear ranking: the restaurant would be pleased if there were lots of people who expressed themselves satisfied rather than dissatisfied. Such considerations lead us to distinguish between two main types of variable, the second of which is itself subdivided.

Classification of variables

- **Measurable variables**, where there is a generally recognised method of measuring the value of the variable of interest.
- **Categorical variables**, where no such method exists (or, often enough, is even possible), but among which:
  - some examples of categorical variables can be put in some sensible order (case (e)), and hence are called **ordinal** (categorical) variables
  - some examples of categorical variables cannot be put in any sensible order, but are only known by their names, and hence are called **nominal** (categorical) variables.

4.7 Data presentation

Datasets consist of potentially vast amounts of data. Hedge funds, for example, have access to very large databases of historical price information on a range of financial assets, such as so-called 'tick data' – very high-frequency intra-day data. Of course, the human brain cannot easily make sense of such large quantities of numbers when presented with them on screen. However, the human brain can cope with graphical representations of data. By producing various plots, we can instantly 'eyeball' to get a bird's-eye view of the dataset. So, at a glance, we can quickly get a feel for the data and determine whether there are any interesting features, relationships, etc. which could then be examined in greater depth. (For those studying **ST104b Statistics 2**, in modelling we often make **distributional assumptions**, and a suitable variable plot allows us to easily check the feasibility of a particular distribution.) To summarise, plots are a great medium for *communicating* the salient features of a dataset to a wide audience.

The main representations we use in **ST104a Statistics 1** are **histograms**, **stem-and-leaf diagrams** and **boxplots**.[4] We also use **scatter plots** for *two*

[4]There are many other representations available from software packages, in particular **pie charts** and standard **bar charts** which are appropriate when dealing with *categorical* data, although these will not

measurable variables (covered in Chapter 12).

4.7.1 Presentational traps

Before we see our first graphical representation you should be aware when reading articles in newspapers, magazines and even within academic journals, that it is easy to mislead the reader by careless or poorly-defined diagrams. Hence presenting data effectively with diagrams requires careful planning.

- A good diagram:
  - provides a clear summary of the data
  - is a fair and honest representation
  - highlights underlying patterns
  - allows the extraction of a lot of information quickly.
- A bad diagram:
  - confuses the viewer
  - misleads (either accidentally or intentionally).

Advertisers and politicians are notorious for 'spinning' data for their own objectives!

4.7.2 Dot plot

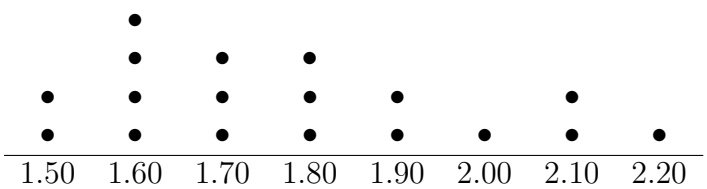
Although a dot plot would not be asked in an examination question, the simplicity of such a diagram makes it an ideal starting point to think about the concept of a **distribution**. For small datasets, this type of plot is very effective for seeing the data's underlying distribution. We use the following procedure:

1. Obtain the **range** of the dataset (values spanned by the data), and draw a horizontal line to accommodate this range.
2. Place dots (hence the name 'dot plot'!) corresponding to the values above the line, resulting in the empirical distribution.

Example 4.1

Hourly wage rates (in £) for clerical assistants (historic data!):

2.20	1.50	1.80	1.60	2.10	1.80	1.60	1.70	1.50
1.60	1.90	1.70	1.60	2.10	1.70	1.80	1.90	2.00



be considered further in this course.

Instantly, some interesting features emerge from the dot plot which are not immediately obvious from the raw data. For example, most clerical assistants earn less than £2 and nobody (in the sample) earns more than £2.20.

4.7.3 Histogram

Histograms are excellent diagrams to use when we want to depict the frequency distribution of (discrete or continuous) variables. To construct histograms, data are first organised into a table which arranges the data into *class intervals* (also called *bins*) – subdivisions of the total range of values which the variable takes. To each class interval, the corresponding *frequency* is determined, that is the number of observations of the variable which falls in each class interval.

Example 4.2 Weekly production output[5] of a factory over a 50-week period. (You can choose what the manufactured good is!)

360	383	368	365	371	354	360	359	358	354
385	381	393	392	395	410	437	420	456	441
458	453	451	444	445	382	379	372	363	350
362	365	365	372	364	396	396	398	402	406
450	469	467	460	454	438	358	362	404	459

In principle the bins do not have to be of equal width, but for simplicity we use equal bin widths whenever possible (the reason why will become clear below). A word of warning: recall our objective is to represent the distribution of the data. As such, too many bins will dilute the distribution, while too few will concentrate it (using technical jargon, will *degenerate* the distribution). Either way, the pattern of the distribution will be lost – defeating the purpose of the histogram. As a guide, six or seven bins should be sufficient, but remember to exercise common sense!

Class interval	Interval width	Frequency	Frequency density	Cumulative frequency
[300, 360)	60	6	0.100	6
[360, 380)	20	14	0.700	20
[380, 400)	20	10	0.500	30
[400, 420)	20	4	0.200	34
[420, 460)	40	13	0.325	47
[460, 500)	40	3	0.075	50

The table above includes two additional columns: (i.) 'Frequency density' – obtained by calculating 'frequency divided by interval width' (for example, $6/60 = 0.100$), and (ii.) 'Cumulative frequency' – obtained by simply determining the running total of the class frequencies (for example, $6 + 14 = 20$). Note the final column is not required for a histogram per se, although the computation of cumulative frequencies may be useful when determining medians and quartiles (to be discussed later in this chapter).

To construct the histogram, adjacent bars are drawn over the respective class intervals such that the **area of each bar is proportional to the interval frequency**. This explains why equal bin widths are desirable since this reduces the problem to making the **heights proportional to the interval frequency**. However, you may be told to use a particular number of bins or bin widths, such that the bins will not all be of equal width. In such cases, you will need to compute the frequency density as outlined above. The histogram for the above example is shown in Figure 4.1.

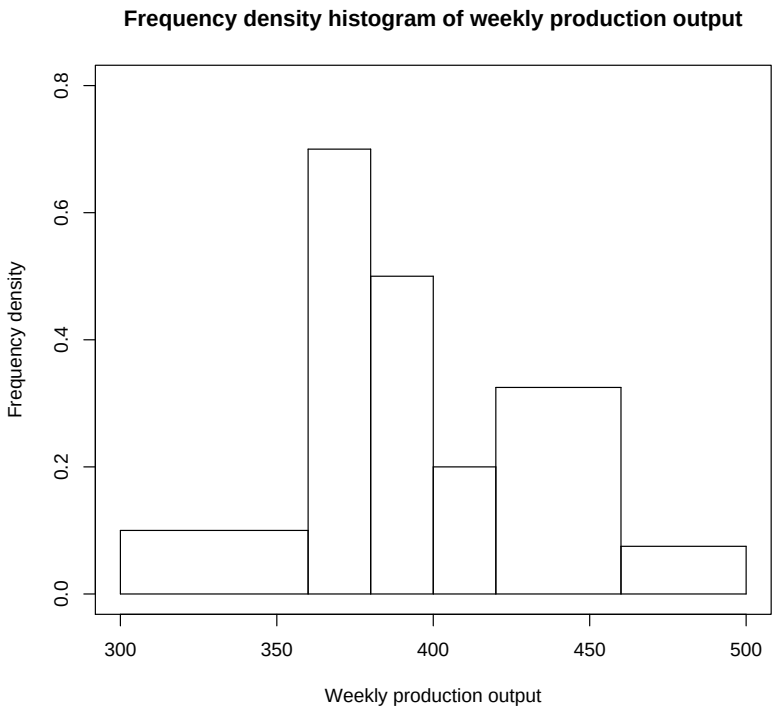


Figure 4.1: Frequency density histogram of weekly production output.

Key points to note:

- All bars are centred on the midpoints of each class interval.
- **Informative** labels on the histogram, i.e. title and axis labels. The idea is the reader can work out what the graphic is showing!
- Because area represents frequency, it follows that the dimension of bar heights is number per unit class interval, hence the y-axis should be labelled 'Frequency density' rather than 'Frequency'.
- **Keen students only:** experiment with different numbers of bins and see how sensitive the histogram shape is to this choice!

[5]This is a discrete variable since the output will take integer values, i.e. something which we can *count*.

4.7.4 Stem-and-leaf diagram

A stem-and-leaf diagram requires the raw data. As the name suggests, it is formed using a 'stem' and corresponding 'leaves'. Choice of the stem involves determining a major component of a typical data item, for example the '10s' unit, or if data are of the form $1.4, 1.8, 2.1, 2.9, \ldots$, then the integer part would be appropriate. The remainder of the data value plays the role of the 'leaf'. Applied to the weekly production dataset, we obtain the stem-and-leaf diagram shown below in Example 4.3. Note the following points:

- These stems are equivalent to using the (discrete) class intervals:

$$[350,\ 359],\ [360,\ 369],\ \ [370,\ 379],\ \ldots,\ [460,\ 469].$$

- Leaves are vertically aligned, hence rotating the stem-and-leaf diagram 90 degrees anti-clockwise reproduces the shape of the data's distribution, just as would be revealed with a histogram.
- The leaves are placed in order of magnitude within the stems – therefore it is a good idea to sort the raw data into ascending order first of all.
- Unlike the histogram, the actual data values are preserved. This is advantageous if we want to calculate various (*descriptive* or *summary*) statistics later on.

Example 4.3 Continuing with Example 4.2, the stem-and-leaf diagram is:

Stem-and-leaf diagram of weekly production

Stem (Tens)	Leaves (Units)
35	044889
36	0022345558
37	1229
38	1235
39	235668
40	246
41	0
42	0
43	78
44	145
45	0134689
46	079

Note the *informative* title and labels for the stems and leaves.

So far we have considered how to summarise a dataset graphically. This methodology is appropriate to get a *visual* feel for the distribution of the dataset. In practice, we would also like to summarise things *numerically* – and so we shall. There are two key properties of a dataset that will be of particular interest.

Key properties of a dataset

- **Measures of location** – a central point about which the data tends (also known as measures of central tendency).
- **Measures of spread** – a measure of the variability of the data, i.e. how spread out it is about the central point (also known as measures of dispersion).

4.8 Measures of location

Mean, **median** and **mode** are the three principal measures of location. In general, these will not all give the same numerical value for a given dataset/distribution.[6] These three measures (and, later, measures of spread) will now be introduced using the following simple sample dataset:

$$32 \quad 28 \quad 67 \quad 39 \quad 19 \quad 48 \quad 32 \quad 44 \quad 37 \quad 24. \tag{4.1}$$

4.8.1 Mean

The preferred measure of location/central tendency, which is simply the 'average' of the data. It will be frequently applied in various statistical inference techniques in later chapters.

Sample mean

Using the summation operator, $\sum$, which, remember, is just a form of 'notational shorthand', we define the sample mean[7] as:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Example 4.4 For the dataset in (4.1) above:

$$\bar{x} = \frac{32 + 28 + \cdots + 24}{10} = \frac{370}{10} = 37.$$

Of course, it is possible to encounter datasets in frequency form, that is each data value is given with the corresponding frequency of observations for that value, f_k, for

[6] 'In general' $\neq$ always! An interesting example is the normal distribution (introduced in Chapter 6) which is symmetric about the mean (and so mean = median) and achieves a maximum at this point, i.e. mean = median = mode.

[7] $\bar{x}$ will be used to denote an observed *sample* mean, while μ will denote its population counterpart, that is the *population* mean.

$k = 1, \ldots, K$, where there are K different variable values. In such a situation, use the formula:

$$\bar{x} = \frac{\sum_{k=1}^{K} f_k x_k}{\sum_{k=1}^{K} f_k}. \tag{4.2}$$

Note that this preserves the idea of 'adding up all the observations and dividing by the total number of observations'. This is an example of a **weighted mean**, where the weights are the *relative frequencies*.

If the data are given in grouped-frequency form, for example as shown in the table in Example 4.2, then the individual data values are unknown[8] – all we know is the *interval* in which each observation lies. The sensible solution is to use the midpoint of the interval as a 'representative proxy' for each observation recorded as belonging within that class interval. Hence you still use the grouped-frequency mean formula (4.2), but each x_i value will be substituted with the appropriate class interval midpoint.

Example 4.5 Using the weekly production data in Example 4.2, the interval midpoints are: 330, 370, 390, 410, 440 and 480, respectively. These will act as the data values for the respective classes. The mean is then calculated by:

$$\bar{x} = \frac{\sum_{k=1}^{K} f_k x_k}{\sum_{k=1}^{K} f_k} = \frac{(6 \times 330) + \cdots + (3 \times 480)}{6 + \cdots + 3} = 397.2.$$

Compared to the true mean of the raw data (which is 399.72), we see that using the midpoints as proxies gives a mean very close to the true sample mean value. Note the mean is **not** rounded up or down since it is an arithmetic result.

Note a drawback with the mean is its **sensitivity to outliers** (extreme observations). For example, suppose we record the net worth of 10 randomly chosen people. If Warren Buffett, say, was included, his substantial net worth would drag the mean upwards considerably! By increasing the sample size n, the effect of his inclusion, although diluted, would still be 'significant', assuming we were not just sampling from the population of billionaires!

4.8.2 Median

The median is the middle value of the *ordered* dataset, where observations are arranged in (ascending) order. By definition, 50 per cent of the observations are greater than the median, and 50 per cent are less than the median.

[8] Of course, we do have the raw data for the weekly production example and so could work out the exact sample mean, but here suppose we did not have access to the raw data, instead we were just given the table of class frequencies in Example 4.2.

Median

Arrange the n numbers in ascending order,[9] say $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$,[10] then:

- If n is odd, there is an explicit middle value, so the median, m, is $x_{((n+1)/2)}$.
- If n is even, there is no explicit middle value, so take the average of the values either side of the 'midpoint', hence $m = \left(x_{(n/2)} + x_{(n/2+1)}\right)/2$.

Example 4.6 For the dataset in (4.1), the ordered observations are:

$$19 \quad 24 \quad 28 \quad 32 \quad 32 \quad 37 \quad 39 \quad 44 \quad 48 \quad 67.$$

$n = 10$, i.e. even, so we compute the average of the fifth and sixth ordered observations, that is:

$$m = \frac{1}{2}\left(x_{(5)} + x_{(6)}\right) = \frac{1}{2}(32 + 37) = 34.5.$$

If we only had data in grouped-frequency form (as in Example 4.2), then we can make use of the cumulative frequencies. Since $n = 50$, the median is the 25.5th ordered observation which must be in the [380, 400) interval because once we exhaust the ordered data up to the [360, 380) interval we have only covered 20 observations, while once the [380, 400) interval is exhausted we have accounted for 30 observations, meaning the median lies in this interval. Assuming the raw data are not accessible, we could use the midpoint (i.e. 390) as denoting the median. Alternatively we could use an *interpolation* method which uses the following 'general' formula for grouped data, once you have identified the class which includes the median (such as [380, 400) above):

$$\text{Endpoint of previous class} + \frac{\text{Class width} \times \#\text{ of remaining observations}}{\text{Class frequency}}.$$

Example 4.7 Returning to the weekly production example, the median would be:

$$380 + \frac{20 \times (25.5 - 20)}{10} = 391.$$

For comparison, using the raw data, $x_{(25)} = 392$ and $x_{(26)} = 393$ (quickly obtained from the stem-and-leaf diagram), giving a median of 392.5.

Although the median is advantageous in not being influenced by outliers (Warren Buffett's net worth would be $x_{(n)}$ and so would not affect the median), in practice it is of limited use in formal statistical inference.

For symmetric data, the mean and median are always equal. Hence this is a good way to verify whether a dataset is symmetric. Asymmetrical distributions are skewed, where

[9] If you have constructed a stem-and-leaf diagram, you would have already done this!

[10] These are known as 'order statistics', for example $x_{(1)}$ is the first order statistic, i.e. the smallest observed value, and $x_{(n)}$ is the largest observed value.

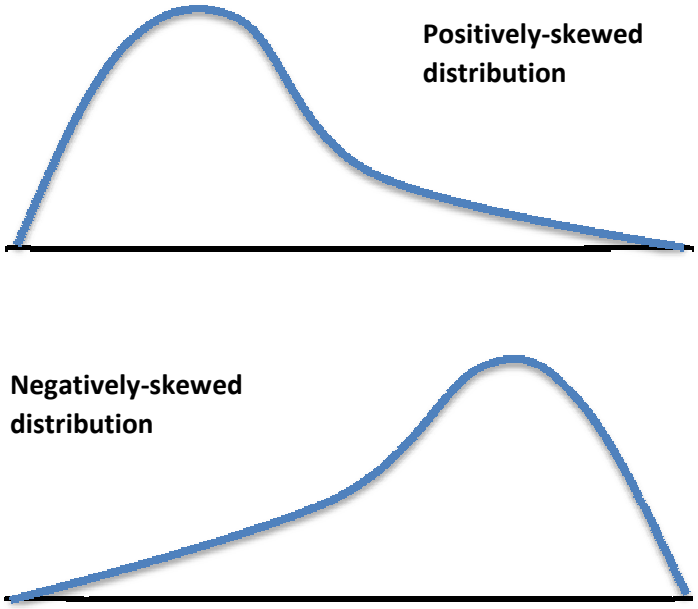


Figure 4.2: Different types of skewed distributions.

skewness measures the departure from symmetry. Although you will not be expected to compute the coefficient of skewness (its numerical value), you need to be familiar with the two types of skewness.

Skewness

- Mean $>$ median indicates a **positively-skewed** distribution (also, referred to as 'right-skewed').
- Mean $<$ median indicates a **negatively-skewed** distribution (also, referred to as 'left-skewed').

Graphically, skewness can be determined by identifying where the long 'tail' of the distribution lies. If the long tail is heading toward $+\infty$ (*positive* infinity) on the x-axis (i.e. on the right-hand side), then this indicates a positively-skewed (right-skewed) distribution. Similarly, if heading toward $-\infty$ (*negative* infinity) on the x-axis (i.e. on the left-hand side) then this indicates a negatively-skewed (left-skewed) distribution, as illustrated in Figure 4.2.

Example 4.8 The hourly wage rates used in Example 4.1 are skewed to the right, due to the influence of the 'extreme' values $2.0, 2.1, 2.1$ and 2.2 – the effect of these (akin to Warren Buffett's effect mentioned above) is to 'drag' or 'pull' the mean upwards, hence mean $>$ median.

Example 4.9 For the weekly production dataset in Example 4.2, we have calculated the mean and median to be 399.72 and 392.50, respectively. Since the

mean is greater than the median, the data form a positively-skewed distribution, as confirmed by the histogram in Figure 4.1.

4.8.3 Mode

Our final measure of location is the mode.

Mode

By definition, the mode is the most frequently occurring value in a dataset.

It is perfectly possible to encounter a multimodal distribution where several data values are tied in terms of their frequency of occurrence.

Example 4.10 The modal value of the dataset in (4.1) is 32, since it occurs twice while the other values only occur once each.

Example 4.11 For the weekly production dataset, looking at the stem-and-leaf diagram in Example 4.3, we can quickly see that 365 is the modal value (the three consecutive 5s opposite the second stem stand out). If just given grouped data, then instead of reporting a modal value we can determine the *modal class*, which is $[360, 380)$ with 14 observations. (The fact that this includes 365 here is a coincidence – the modal class and modal value are not the same thing.)

4.9 Measures of spread

The dispersion (or spread) of a dataset is extremely relevant when drawing conclusions from it. Thus it is important to have a useful measure of this property, and several candidates exist – these are reviewed below. As expected, there are advantages and disadvantages to each.

4.9.1 Range

Our first measure of spread is the range.

Range

The range is simply the largest value minus the smallest value, that is:

$$\text{Range } = x_{(n)} - x_{(1)}.$$

Example 4.12 For the dataset in (4.1), the range is:

$$x_{(n)} - x_{(1)} = 67 - 19 = 48.$$

Clearly, the range is very sensitive to extreme observations since (when they occur) they are going to be the smallest and/or largest observations, hence this measure is of limited appeal. If we were confident that no outliers were present (or decided to filter out any outliers), then the range would better represent the true spread of the data.

More formally, we could consider the idea of the **interquartile range (IQR)** instead – that is, the upper (third) quartile, Q_3, minus the lower (first) quartile, Q_1. The upper quartile has 75 per cent of observations below it (and 25 per cent above it) while the lower quartile has 25 per cent below (and 75 per cent above). Unsurprisingly the median, given our earlier definition, is the middle (second) quartile. By discarding the top 25 per cent and bottom 25 per cent of observations, respectively, we restrict attention solely to the central 50 per cent of observations.

Interquartile range

$$The IQR is defined as: \text{IQR} = Q_3 - Q_1$$

where Q_1 and Q_3 are the first (lower) and third (upper) quartiles, respectively.

Example 4.13 Continuing with the dataset in (4.1), computation of these quartiles can be problematic since, for example, for the lower quartile we require the value such that 2.5 observations are below it and 7.5 values are above it. A suggested remedy[11] (motivated by the median calculation when n is even) is to use:

$$Q_1 = \frac{1}{2}(x_{(2)} + x_{(3)}) = \frac{1}{2}(24 + 28) = 26.$$

Similarly:

$$Q_3 = \frac{1}{2}(x_{(7)} + x_{(8)}) = \frac{1}{2}(39 + 44) = 41.5.$$

Hence the IQR is $41.5 - 26 = 15.5$. Contrast this with the range of 48.

4.9.2 Boxplot

At this point, it is useful to introduce another graphical method – the boxplot.[12] In a boxplot, the middle horizontal line is the median and the upper and lower ends of the box are the upper and lower quartiles, respectively. The 'whiskers' are drawn from the quartiles to the observations furthest from the median, but not by more than one-and-a-half times the IQR (i.e. excluding outliers). The whiskers are terminated by horizontal lines. Any extreme points beyond the whiskers are plotted individually. An example of a (generic) boxplot is shown in Figure 4.3.

If you are presented with a boxplot, then it is easy to obtain all of the following: median, quartiles, IQR, range and skewness. Recall skewness (departure from

[11] There are many different methodologies for computing quartiles, and conventions vary from country to country. Any rational, and justified, approach is perfectly acceptable in **ST104a Statistics 1**. For example, interpolation methods, as demonstrated previously for the case of the median, are valid.

[12] Some people call such plots box-and-whisker plots. No prizes for guessing why!

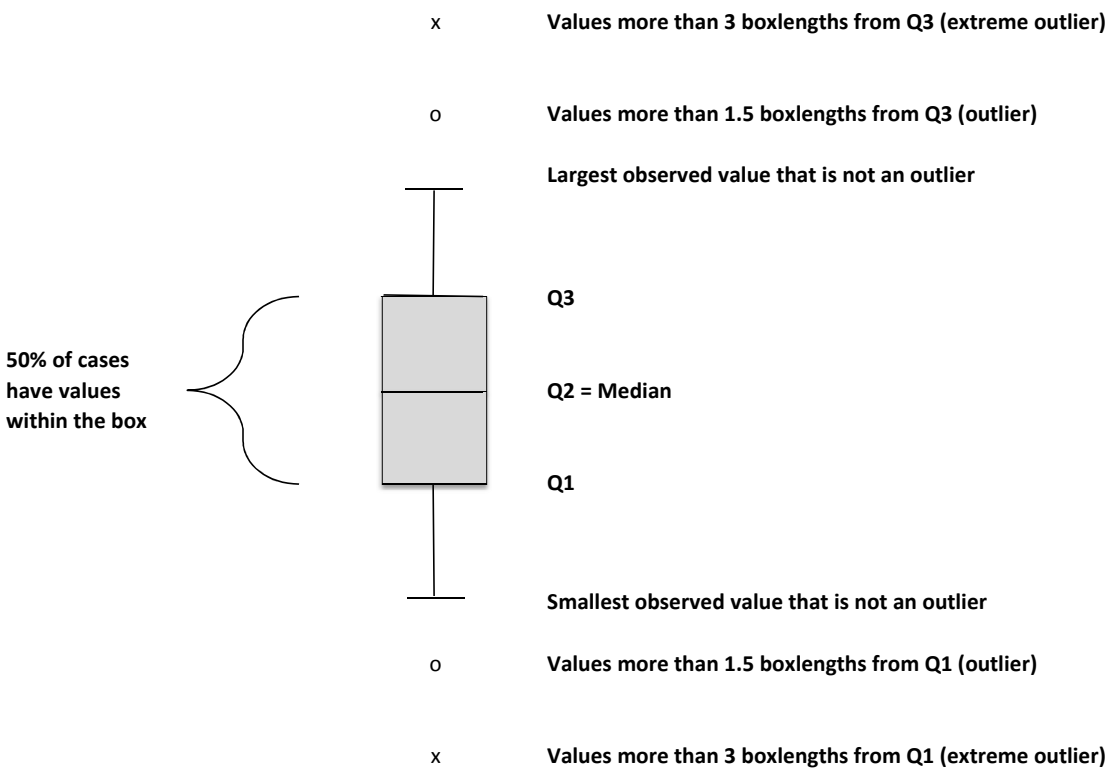


Figure 4.3: An example of a boxplot (not to scale).

symmetry) is characterised by a long tail, attributable to outliers, which are readily apparent from a boxplot.

> **Example 4.14** From the boxplot shown in Figure 4.4, it can be seen that the median, Q_2, is around 74, Q_1 is about 63 and Q_3 is approximately 77. The numerous outliers provide a useful indicator that this is a negatively-skewed distribution as the long tail covers lower values of the variable. Note also that $Q_3 - Q_2 < Q_2 - Q_1$.

4.9.3 Variance and standard deviation

These are much better and more useful statistics for representing the spread of a dataset. You need to be familiar with their definitions and methods of calculation for a sample of data values $x_1, x_2, \ldots, x_n$.

Begin by computing the so-called 'corrected sum of squares', S_{xx}, the sum of the squared deviations of each data value from the (sample) mean, where:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2. \tag{4.3}$$

Recall from earlier $\bar{x} = \sum_{i=1}^{n} x_i/n$. The proof for why the expressions in (4.3) are equivalent is beyond the scope of this course.

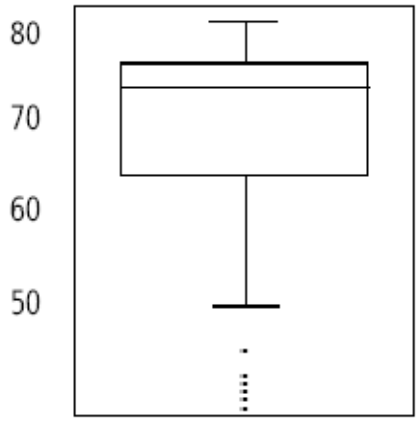


Figure 4.4: A boxplot showing a negatively-skewed distribution.

Sample variance

The **sample variance** is defined as:

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

Note the divisor used to compute s^2 is $n-1$, not n. Do not worry about why,[13] just remember to divide by $n-1$ when computing a *sample* variance.[14] To obtain the **sample standard deviation**, s, we just take the (positive) square root of the sample variance, s^2.

Sample standard deviation

The **sample standard deviation** is:

$$s = \sqrt{s^2} = \sqrt{\frac{S_{xx}}{n-1}}.$$

Example 4.15 Using the dataset in (4.1), $\bar{x} = 37$, so:

$$S_{xx} = (32-37)^2 + (28-37)^2 + \cdots + (24-37)^2 = 25 + 81 + \cdots + 169 = 1698.$$

Hence $s = \sqrt{1698/(10-1)} = 13.74$.

Note that, given $\sum x_i^2 = 15388$, we could have calculated S_{xx} using the other

[13] This is covered in **ST104b Statistics 2.**

[14] In contrast for *population* data, the *population* variance is $\sigma^2 = \sum_{i=1}^{N}(x_i - \mu)^2/N$, i.e. we use the N divisor here. N denotes *population* size while n denotes *sample* size. Also, note the use of μ (*population* mean) instead of $\bar{x}$ (*sample* mean).

expression:

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 15388 - 10 \times 37^2 = 1698.$$

So this alternative method is much quicker to calculate S_{xx}.

Sample variance for grouped-frequency data

For grouped-frequency data with K classes, to compute the sample variance we would use the formula:

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

Recall that the last bracketed squared term is simply the mean formula for grouped data. Note that for grouped-frequency data we can ignore the 'divide by $n-1$' rule, since we would expect n to be very large in such cases, such that $n - 1 \approx n$ and so dividing by n or $n-1$ makes negligible difference in practice.

Example 4.16

Let us now consider an extended example bringing together many of the issues considered in this chapter. We will draw a histogram, here with equal class widths, and introduce the cumulative frequency diagram.

A stockbroker is interested in the level of trading activity on a particular stock exchange. He has collected the following data, which are weekly average volumes (in millions), over a 29-week period. This is an example of *time series* data. Note that this variable is treated as if it was discrete, but because the numbers are so large the variable can be treated as continuous.

172.5	154.6	163.5
161.9	151.6	172.6
172.3	132.4	168.3
181.3	144.0	155.3
169.1	133.6	143.4
155.0	149.0	140.6
148.6	135.8	125.1
159.8	139.9	171.3
161.6	164.4	167.0
153.8	175.6	

So to construct a histogram we first decide on the choice of class intervals, which is a subjective decision. The objective is to convey information in a useful way. In this case the data lie between (roughly) 120 and 190 million shares/week, so class intervals of width 10 million will give seven classes. With almost 30 observations this

choice is probably adequate; more observations might support more classes (a class interval of 5 million, say); fewer observations would, perhaps, need a larger interval of 20 million.

The class intervals are thus defined like this:

$$120 \leq \text{ Volume } < 130, \quad 130 \leq \text{ Volume } < 140, \quad \text{etc.}$$

or alternatively [120, 130), [130, 140), etc. We now proceed to determine the frequency density (and cumulative frequencies, for later).

Class interval	Interval width	Frequency	Frequency density	Cumulative frequency
[120, 130)	10	1	0.1	1
[130, 140)	10	4	0.4	5
[140, 150)	10	5	0.5	10
[150, 160)	10	6	0.6	16
[160, 170)	10	7	0.7	23
[170, 180)	10	5	0.5	28
[180, 190)	10	1	0.1	29

Hence the frequency density histogram is as shown in Figure 4.5.

A development from the histogram is the *cumulative frequency diagram*. Recall that a cumulative frequency gives us the total frequency of observations that fall below the upper endpoint of a class interval. Therefore it is the upper endpoint that we use in the diagram. Figure 4.6 displays this for the trading data example.

A variation on this is the *cumulative relative frequency diagram* which accumulates the **relative frequencies** (typically expressed as percentages), rather than the frequencies, where the relative frequency of the kth interval is calculated by:

$$\text{Relative frequency}_k = f_k/n.$$

Class interval	Frequency	Relative frequency (%)	% Cumulative relative frequency
[120, 130)	1	3.45	3.45
[130, 140)	4	13.79	17.24
[140, 150)	5	17.24	34.48
[150, 160)	6	20.69	55.17
[160, 170)	7	24.14	79.31
[170, 180)	5	17.24	96.55
[180, 190)	1	3.45	100.00

So, for example, in the above table since $n = 29$ the relative frequency for the first interval is $1/29 = 0.0345$, or 3.45%. Plotting the cumulative relative frequencies yields the cumulative relative frequency diagram, shown in Figure 4.7.

Note all the diagrams use the grouped frequencies, rather then the original raw data. For comparison, Figure 4.8 uses the raw *ungrouped* data, providing a more 'precise' cumulative relative frequency diagram.

Finally, we use the grouped data to compute particular descriptive statistics – specifically the mean, variance and standard deviation. There are $K = 7$ classes, so we perform the appropriate intermediate calculations by groups.

Class interval	Midpoint, x_k	Frequency, f_k	$f_k x_k$	$f_k x_k^2$
[120, 130)	125	1	125	15625
[130, 140)	135	4	540	72900
[140, 150)	145	5	725	105125
[150, 160)	155	6	930	144150
[160, 170)	165	7	1155	190575
[170, 180)	175	5	875	153125
[180, 190)	185	1	185	34225
Total, $\sum$		29	4535	715725

Hence the grouped mean is:

$$\bar{x} = \frac{\sum_{k=1}^{7} f_k x_k}{\sum_{k=1}^{7} f_k} = \frac{4535}{29} = 156.4$$

and the grouped variance is:

$$s^2 = \frac{\sum_{k=1}^{7} f_k x_k^2}{\sum_{k=1}^{7} f_k} - \left(\frac{\sum_{k=1}^{7} f_k x_k}{\sum_{k=1}^{7} f_k} \right)^2 = \frac{715725}{29} - (156.4)^2 = 219.2$$

giving a standard deviation of $\sqrt{219.2} = 14.8$. For comparison, the ungrouped mean, variance and standard deviation are 156.0, 217.0 and 14.7, respectively (compute these yourself to verify!). Note the units for the mean and standard deviation are 'millions of shares/week', while the units for the variance are the square of those for the standard deviation, i.e. '(millions of shares/week)^2', so this is an obvious reason why we often work with the standard deviation, rather than the variance, due to the original and meaningful units.

4.10 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.

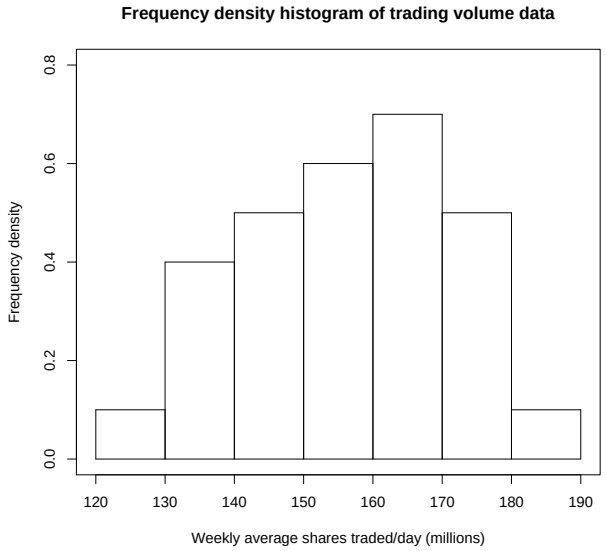


Figure 4.5: Frequency density histogram of trading volume data.

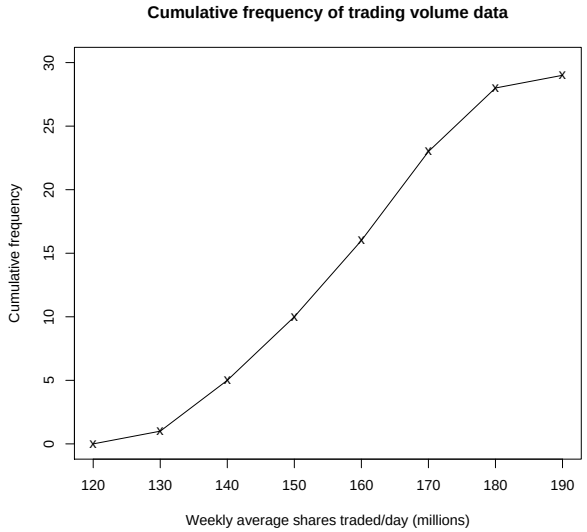


Figure 4.6: Cumulative frequency of trading volume data.

4.11 Key terms and concepts

- Boxplot
- Categorical variable
- Continuous variable
- Cumulative frequency
- Discrete variable
- Distribution
- Dot plot
- Histogram
- Interquartile range
- Mean
- Measurable variable
- Median
- Mode
- Nominal
- Ordinal
- Outliers
- Range
- Relative frequency
- Skewness
- Standard deviation
- Stem-and-leaf diagram
- Variance

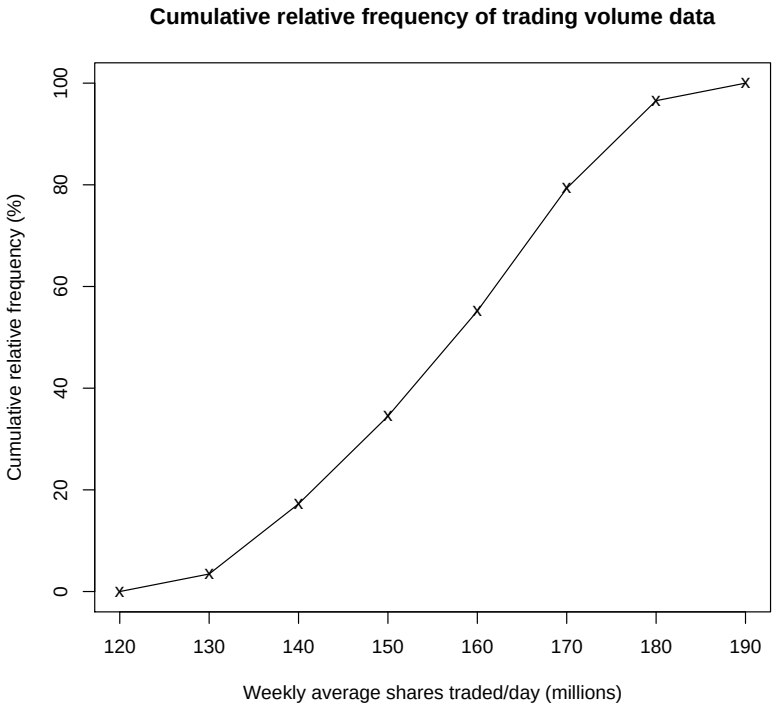


Figure 4.7: Cumulative relative frequency of trading volume data.

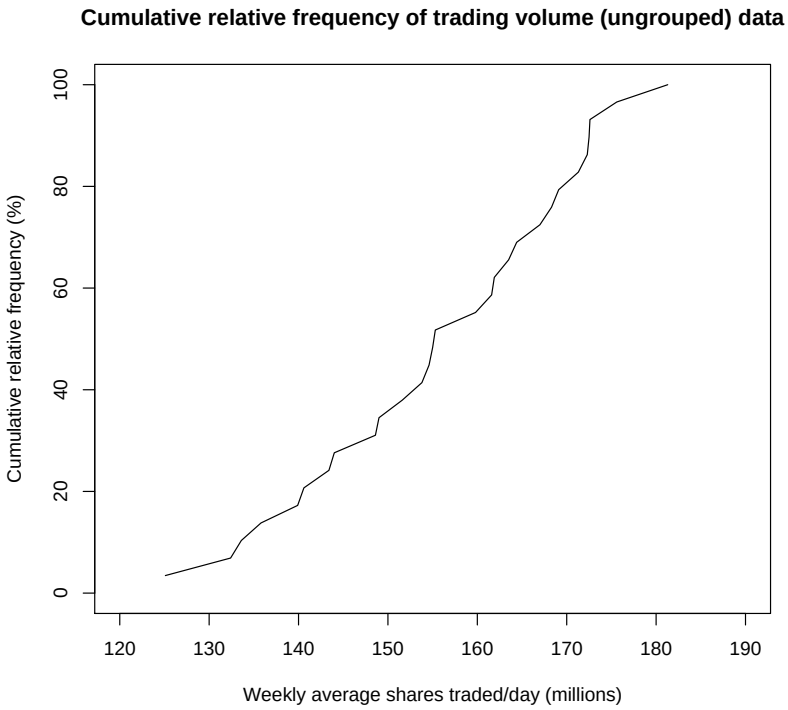


Figure 4.8: Cumulative relative frequency of trading volume (ungrouped) data.

4.12 Learning activities

1. Identify and describe the variables in each of the following examples:
   (a) Voting intentions in a poll of 1,000 people.
   (b) The number of hours homemakers watch television per week.
   (c) The number of children per household receiving full-time education.

2. Find the mean of the number of hours of television watched per week by 10 homemakers:

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

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

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	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 two values) in this raw data. However, rounding to the nearest 10 gives values at 110.)

4. Work out s^2 for a sample of nine 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.

5. Think about why and when you would use each of the following:
   (a) histogram
   (b) stem-and-leaf diagram.

   When would you **not** do so?

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

7. Calculate the range, variance, standard deviation and interquartile range of the spreadsheet prices shown below. Check against the answers given after the data.

Name	Price	Price − Mean	(Price − Mean)^2
Kuma K-Spread II	52	−82.54	6812.60
Logistix	64	−70.54	4975.67
Sage Planner	65	−69.54	4835.60
SuperCalc 3.21	82	−52.54	2760.29
VP Planner Plus	105	−29.54	875.52
SuperCalc v4	107	−27.54	758.37
Multiplan IV	110	−24.54	602.14
Borland QUATTRO	115	19.54	381.75
Legend Twin Level 3	155	20.46	418.67
SuperCalc v5	195	60.46	3655.60
Plan Perfect	195	60.46	3655.60
Microsoft Excel	239	104.46	10912.21
Lotus 1-2-3	265	130.46	17020.21
Total		0.00	57661.23
Variance = 4435.48	Std. Dev. = 66.60	IQR = 113	

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

In addition, attempt the 'Test your understanding' self-test quizzes available on the VLE.

4.13 Further exercises

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060]. Relevant exercises from Sections 1.3, 1.5 and 2.1–2.3.

4.14 A reminder of your learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- calculate the following: arithmetic mean, median, mode, standard deviation, variance, quartiles, range and interquartile range
- explain the use and limitations of the above quantities

- draw and interpret: histograms, stem-and-leaf diagrams, boxplots and cumulative frequency distributions
- incorporate 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.

4.15 Sample examination questions

1. The data below show the number of daily phone calls received by an office supplies company over a period of 25 working days.

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.

   (b) Find the first and third quartiles of the data.

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

   (d) Comment on your figures.

2. Say whether the following statement is **true** or **false** and briefly give your reason(s). 'The mean of a dataset is always greater than the median.'

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

   (a) the median

   (b) the mean.

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.

5. If $n = 4$, $x_1 = 1$, $x_2 = 4$, $x_3 = 5$ and $x_4 = 6$, find:

$$\frac{1}{3}\sum_{i=2}^{4} x_i.$$

   Why might you use this figure to estimate the mean?

6. State whether the following statement is **true** or **false** and briefly give your reason(s). 'Three quarters of the observations in a dataset are less than the lower quartile.'

7. If $x_1 = 4$, $x_2 = 2$, $x_3 = 2$, $x_4 = 5$ and $x_5 = 6$, calculate:
   (a) the mode
   (b) the mean.

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

Chapter 5
Probability

5.1 Aims

This chapter introduces the fundamental concepts of probability. In other courses, particularly in **ST104b Statistics 2** and **MN3032 Management science methods** you may make full use of probability in both theory and in decision trees, and highlight the ways in which such information can be used.

We will look at probability at quite a superficial level in this course. Even so, you may find that, although the concepts of probability introduced are simple, their application in particular circumstances may be very difficult. Try not to lose heart. Aim to:

- be capable of dealing with the basic concepts and follow through on examples and activities
- relate the new ideas of probability to the examples given
- appreciate its uses and applications.

5.2 Learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- apply the ideas and notation involved in set theory to simple examples
- recall the basic axioms of probability and apply them
- distinguish between the ideas of conditional probability and independence
- draw and use appropriate Venn diagrams
- draw and use appropriate probability trees.

5.3 Essential reading

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060] Sections 3.1–3.3 and 3.5.

In addition there is essential 'watching' of this chapter's accompanying video tutorials accessible via the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

5.4 Further reading

- Aczel, A.D. *Complete Business Statistics*. (London: McGraw-Hill Higher Education, 2009) seventh edition [ISBN 9780071287531] Sections 2.1–2.6.
- Anderson, D.R., D.J. Sweeney, T.A. Williams, J. Freeman and E. Shoesmith *Statistics for Business and Economics*. (South-Western Cengage Learning, 2010) eleventh edition [ISBN 9780324783247] Sections 4.1–4.4.
- Lind, D.A., W.G. Marchal and S.A. Wathen *Statistical Techniques in Business and Economics*. (Boston: McGraw-Hill Higher Education, 2009) fourteenth edition [ISBN 978007110004] The first half of Chapter 5.
- Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics for Business and Economics*. (Chichester: John Wiley & Sons, 1990) fourth edition [ISBN 9780471615170] Sections 3.1–3.3.

5.5 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 making decisions.

5.6 The concept of probability

Probability forms the bedrock on which statistical methods are based. This chapter will be devoted to understanding this important concept, its properties and applications.

One can view probability as a quantifiable measure of one's degree of belief in a particular **event** or **set**[1] of interest. To motivate the use of the terms 'event' and 'set', we begin by introducing the concept of an **experiment**. An experiment can take many forms, but to prevent us from becoming over-excited, let us consider two mundane examples:

i. the toss of a (fair) coin

ii. the roll of a (fair) die.

Sample space

We define the **sample space**, S, as the set of all possible outcomes for an experiment.

[1] These terms can be used interchangeably.

Example 5.1 Thus for our two examples we have:

i. Coin toss: $S = \{H, T\}$, where H and T denote 'heads' and 'tails', respectively, and are called the **elements** or **members** of the sample space.

ii. Die score: $S = \{1, 2, 3, 4, 5, 6\}$.

So the coin toss sample space has two elementary outcomes, H and T, while the score on a die has six elementary outcomes. These individual elementary outcomes are themselves events, but we may wish to consider slightly more exciting events of interest[2] – for example, for the die score, we may be interested in the event of obtaining an even score, or a score greater than 4, etc. Hence we proceed to define an event.

Event

An event is a collection of elementary outcomes from the sample space S of an experiment, and therefore it is a **subset of** S.

Typically we can denote events by letters for notational efficiency. For example, $A =$ 'an even score', and $B =$ 'a score greater than 4'. Hence $A = \{2, 4, 6\}$ and $B = \{5, 6\}$.

The universal convention is that we define **probability** to lie on a scale from 0 to 1 inclusive.[3] Hence the probability of any event A, say, is denoted $P(A)$ and is a real number somewhere on the unit interval, i.e. $P(A) \in [0, 1]$, where '$\in$' means 'is a member of'. Note the following:

- If A is an impossible event, then $P(A) = 0$.
- If A is a certain event, then $P(A) = 1$.
- For events A and B, if $P(A) > P(B)$, then A is more likely to happen than B.

We thus have a probability scale from 0 to 1 on which we are able to *rank* events, as evident from the $P(A) > P(B)$ result above. However, we need to consider how best to *quantify* these probabilities. Let us begin with the experiments where each elementary outcome is **equally likely**, hence our (fair) coin toss and (fair) die score fulfil this criterion (conveniently).

Determining event probabilites for equally likely elementary outcomes

For an experiment with equally likely elementary outcomes, let N be the total number of equally likely elementary outcomes, and let n be the number of these elementary outcomes that are favourable to our event of interest, A. Then:

$$P(A) = \frac{n}{N}.$$

[2] Admittedly, for a single coin toss, we cannot go beyond a choice of heads or tails.
[3] Multiplying by 100 yields a probability as a percentage.

Example 5.2

i. So returning to our coin toss example, if A is the event 'heads', then $N = 2$ (H and T), $n = 1$ (H), so for a *fair*[4] coin, $P(A) = 1/2 = 0.5$.[5]

ii. For the die score, if A is the event 'an even score', then $N = 6$ (1, 2, 3, 4, 5 and 6), $n = 3$ (2, 4 and 6), so for a *fair* die, $P(A) = 3/6 = 1/2 = 0.5$.

Finally, if B is the event 'a score greater than 4', then $N = 6$ (as before), $n = 2$ (5 and 6), hence $P(B) = 2/6 = 1/3$.

5.7 Relative frequency

So far we have only considered equally likely experimental outcomes. Clearly, to apply probabilistic concepts more widely, we require a more general interpretation – the **relative frequency** interpretation.

Relative frequency approach to probability

Suppose the event A associated with some experiment either does or does not occur. Also, suppose we conduct this experiment *independently*[6] F times. Suppose that, following these repeated experiments, A occurs f times. Hence the 'frequentist' approach to probability would regard:

$$P(A) = \frac{f}{F}$$

as $F \to \infty$.

Example 5.3 For a coin toss with event $A = \{H\}$, if the coin is fair we would *expect* that repeatedly tossing the coin F times would result in *approximately* $f = F/2$ heads, hence $P(A) = (F/2)/F = 1/2$. Of course, this approach is not confined to fair coins!

Intuitively, this is an appealing interpretation and is extremely useful when we come to its use in statistical inference later on. However, do be aware that we are not advocating that you perform all these experiments! Rather they are imaginary experiments, but the concept gives a meaning to numerical probability.[7]

[4]Remember we are assuming equally likely elementary outcomes here, so a fair coin is required. If we had a *biased* coin, then this approach would fail to accurately quantify probabilities.

[5]Probabilities can be reported as proper fractions or in decimal form. If in decimal form (which is preferred), report answers to a maximum of four decimal places in the examination.

[6]'Independent' is an important term, discussed later.

[7]There is another 'school' of probability thought, known as the 'Bayesian' school. We will touch on this only briefly via Bayes' formula later in this chapter. In short, the Bayesian view is that probability is a degree of belief in an event's occurrence *based on the observer's knowledge of the situation.*

5.8 'Randomness'

Statistical inference is concerned with the drawing of conclusions from data that are subject to *randomness*, perhaps due to the sampling procedure, perhaps due to observational errors, perhaps for some other reason. Let us stop and think why, when we repeat an experiment under apparently identical conditions, we get different results. The answer is that although the conditions may be as identical as we are able to control them to be, there will inevitably be a large number of uncontrollable (and frequently unknown) variables that we do not measure and which have a cumulative effect on the result of the sample or experiment. For example, weather conditions may affect the outcomes of biological or other 'field' experiments.

The cumulative effect, therefore, is to cause variations in our results. It is this variation that we term *randomness* and, although we never fully know the true generating mechanism for our data, we can take the random component into account via the concept of probability, which is, of course, why probability plays such an important role in data analysis.

5.9 Properties of probability

We begin this section by presenting three simple, self-evident truths known as **axioms** which list the basic properties we require of event probabilities.

Axioms of probability

1. For any event A, $0 \leq P(A) \leq 1$.
2. For the sample space S, $P(S) = 1$.
3. If $\{A_i\}$, $i = 1, \ldots, n$, are *mutually exclusive* events, then the probability of their 'union' is the sum of their respective probabilities, that is:

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

The first two axioms should not be surprising. The third may appear a little more daunting. Events are labelled **mutually exclusive** when they cannot both simultaneously occur.

> **Example 5.4** When rolling a die once, the event $A =$ 'get an even score' and the event $B =$ 'get an odd score' are mutually exclusive.

Extending this, a collection of events is *pairwise* mutually exclusive if no two events can happen simultaneously. For instance, the three events A, B and C are pairwise mutually exclusive if A and B cannot happen together *and* B and C cannot happen together *and* A and C cannot happen together. Another way of putting this is that a collection of events is pairwise mutually exclusive if *at most one* of them *can* happen.

Related to this is the concept of a collection of events being **collectively exhaustive**. This means *at least one* of them *must* happen, i.e. all possible experimental outcomes are included among the collection of events.

5.9.1 Notational vocabulary

Axiom 3 above introduced a new symbol. For the remainder of this chapter, various symbols connecting sets will be used as a form of notational shorthand. It is important to be familiar with these symbols, hence two 'translations' are provided – one for children and one for grown-ups.[8]

Symbol	'Child' version	'Adult' version	Example
$\cup$	or	union	$A \cup B =$ 'A union B'
$\cap$	and	intersect	$A \cap B =$ 'A intersect B'
^{c}	not	complement of	$A^c =$ 'complement of A'
$\mid$	given	conditional on	$A \mid B =$ 'A conditional on B'

Also, do make sure that you distinguish between a *set* and the *probability of a set*. This distinction is important. A *set*, remember, is a collection of elementary outcomes from S, whereas a probability (from Axiom 1) is a number on the unit interval, $[0, 1]$. For example, $A =$ 'an even die score', while $P(A) = 0.5$, for a fair die.

5.9.2 Venn diagrams

The previous coin and die examples have been rather simple (for illustrative purposes). Hence it is highly likely (in fact with a probability of 1!) that you will encounter more challenging sets and sample spaces. Fear not, there is a helpful geometric technique which can often be used – we represent the sample space elements in a **Venn diagram**.

Imagine we roll a die twice and record the total score. Hence our sample space will be:

$$S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}.$$

Suppose we are interested in the following three events:

- $A =$ an even total, that is $A = \{2, 4, 6, 8, 10, 12\}$.
- $B =$ a total strictly less than 8, that is $B = \{2, 3, 4, 5, 6, 7\}$.
- $C =$ a total greater than 4, but less than 10, that is $C = \{5, 6, 7, 8, 9\}$.

Having defined these events, it is therefore possible to insert every element in the sample space S into a Venn diagram, as shown in Figure 5.1.

The box represents S, so every possible outcome of the experiment (the total score when a die is rolled twice) appears within the box. Three (overlapping) circles are drawn representing the events A, B and C. Each element of S is then inserted into the appropriate area. For example, the area where the three circles all intersect represents the event $A \cap B \cap C$ into which we place the element '6', since this is the only member of S which satisfies all three events A, B and C.

[8] I shall leave you to decide which one of these mutually exclusive and exhaustive sets applies to you.

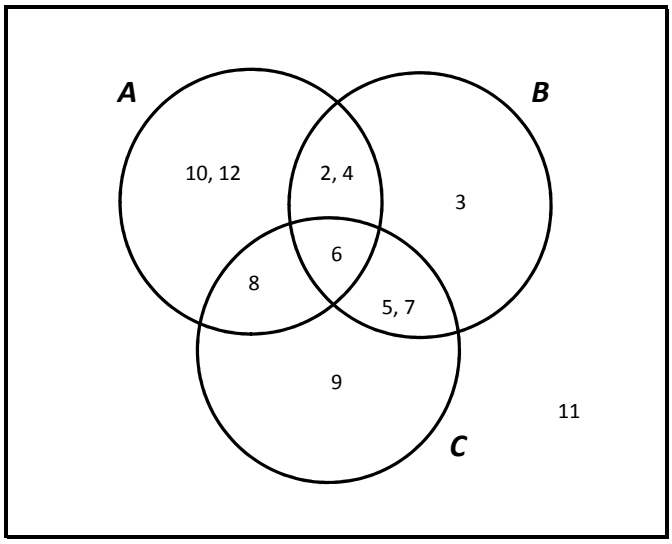


Figure 5.1: Venn diagram for pre-defined sets A, B and C recording the total score when a die is rolled twice.

Example 5.5 Using Figure 5.1, we can determine the following sets:

- $A \cap B = \{2, 4, 6\}$
- $A \cap B \cap C = \{6\}$
- $A \cap B \cap C^c = \{2, 4\}$
- $(A \cup C)^c \cap B = \{3\}$
- $A \cap C = \{6, 8\}$
- $(A \cup B \cup C)^c = \{11\}$
- $A^c \cap B = \{3, 5, 7\}$
- $A \,|\, C = \{6, 8\}$.

5.9.3 The additive law

We now introduce our first probability 'law' – the **additive law**.

The additive law

Let A and B be any two events. The additive law states that:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

So $P(A \cup B)$ is the probability that *at least* one of A and B occurs, and $P(A \cap B)$ is the probability that *both* A and B occur.

Example 5.6 We can think about this using a Venn diagram. The total area of the Venn diagram in Figure 5.2 is assumed to be 1, so area represents probability. Event A is composed of all points in the left-hand circle, and event B is composed of all points in the right-hand circle. Hence:

$$P(A) = \text{area } x + \text{area } z \qquad P(B) = \text{area } y + \text{area } z$$
$$P(A \cap B) = \text{area } z \qquad P(A \cup B) = \text{area } x + \text{area } y + \text{area } z.$$

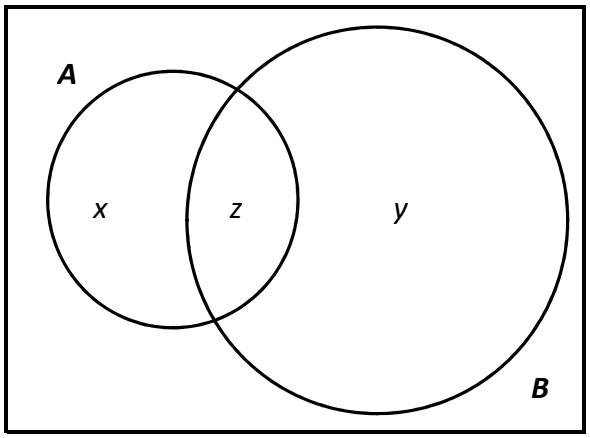


Figure 5.2: Venn diagram illustrating the additive law.

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= (\text{area } x + \text{area } z) + (\text{area } y + \text{area } z) - (\text{area } z) \\ &= \text{area } x + \text{area } y + \text{area } z. \end{aligned}$$

Hence, to compute $P(A \cup B)$ we need to subtract $P(A \cap B)$ otherwise that region would have been counted twice.

Example 5.7 Consider an industrial situation in which a machine component can be defective in two ways such that:

- $P(\text{defective in first way}) = P(D_1) = 0.01$
- $P(\text{defective in second way}) = P(D_2) = 0.05$
- $P(\text{defective in both ways}) = P(D_1 \cap D_2) = 0.001.$

Then it follows that the probability that the component is defective is:

$$P(D_1 \cup D_2) = P(D_1) + P(D_2) - P(D_1 \cap D_2) = 0.01 + 0.05 - 0.001 = 0.059.$$

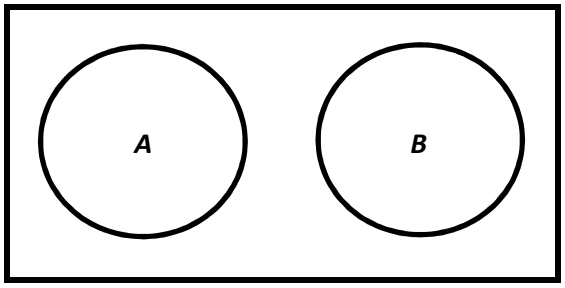


Figure 5.3: Venn diagram illustrating two mutually exclusive events.

Additive law – special case 1

If A and B are *mutually exclusive* events, i.e. they cannot occur simultaneously, then $P(A \cap B) = 0$. Hence:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - 0 \\ &= P(A) + P(B). \end{aligned}$$

Such events can be depicted by two non-overlapping sets in a Venn diagram as shown in Figure 5.3. Now revisit Axiom 3, to see this result generalised for n mutually exclusive events.

Additive law – special case 2

The probability of an event A **not** happening, i.e. the complement, A^c, is:

$$P(A^c) = 1 - P(A).$$

5.9.4 The multiplicative law

This rule is concerned with the probability of two events happening at the same time – specifically when the two events have the special property of **independence**. An informal definition of independence is that two events are said to be independent if one event has no influence on the other.

The multiplicative law (for independent events)

Formally, events A and B are independent if the probability of their intersect is the product of their individual probabilities, that is:

$$P(A \cap B) = P(A) \cdot P(B).$$

Example 5.8 Consider rolling two fair dice. The score on one die has no influence on the score on the other. Hence the respective scores are independent events. Hence:

$$P(\text{two sixes}) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}.$$

Note the multiplicative (or product) law does not hold for dependent events, which is the subject of **conditional probability**, discussed shortly. Also, take a moment to ensure that you are comfortable with the terms 'mutually exclusive' and 'independent'. These are **not** the same thing, so do not get these terms confused!

5.10 Conditional probability and Bayes' formula

We have just introduced the concept of independent events – one event has no influence on another. Clearly, there are going to be many situations where independence does not in fact hold, i.e. the occurrence of one event has a 'knock-on' effect on the probability of another event occurring.

Example 5.9 For a single roll of a fair die, let A be 'roll a 6', and B be 'an even number'. The following probabilities are obvious:

- $P(A) = 1/6$
- $P(B) = 1/2$
- $P(A \mid B) = 1/3$.

So, we see that the probability of a 6 changes from 1/6 to 1/3 once we are given the information that 'an even number' has occurred. Similarly, the probability of an even number changes from 1/2 to 1, conditional on a 6 occurring, i.e. $P(B \mid A) = 1$.

Example 5.10 In order to understand and develop formulae for conditional probability, consider the following simple example, representing the classification by gender and subject (where A, B, C and D are defined below) of 144 college students.

Subject	Female	Male	Total
A: Mathematics	4	14	18
B: Economics	17	41	58
C: Science	4	25	29
D: Arts	28	11	39
Total	53	91	144

Let F = 'Female' and M = 'Male' (obviously!), then $P(A) = 18/144$, $P(F) = 53/144$ and $P(A \cap F) = 4/144$. Note that $P(A \cap F) \neq P(A) \cdot P(F)$, hence A and F are not independent events.

From the table we have the following probabilities:

- $P(A \mid F) = 4/53 \neq P(A)$
- $P(F \mid A) = 4/18 \neq P(F)$.

The correct relationship of these *conditional* probabilities to the original *unconditional* probabilities is:

$$P(A \mid F) = \frac{4/144}{53/144} = \frac{4}{53} = \frac{P(A \cap F)}{P(F)}.$$

Similarly:

$$P(F \mid A) = \frac{4/144}{18/144} = \frac{4}{18} = \frac{P(A \cap F)}{P(A)}. \qquad \text{(Note } P(A \cap F) = P(F \cap A).\text{)}$$

Activity 5.1 Check that some of the other conditional probabilities obtained from the table in Example 5.10 satisfy this formula, such as $P(C \mid F) = 4/53$.

Note also another important relationship involving conditional probability is the 'total probability formula' (discussed in greater depth shortly). This expresses an unconditional probability in terms of other, conditional probabilities.

Example 5.11 Continuing with Example 5.10, we have:

$$\begin{aligned} P(A) = \frac{18}{144} &= \left(\frac{4}{53} \times \frac{53}{144}\right) + \left(\frac{14}{91} \times \frac{91}{144}\right) \\ &= P(A \mid F)P(F) + P(A \mid M)P(M). \end{aligned}$$

5.10.1 Bayes' formula

We formally define conditional probabilities.

Conditional probability

For any two events A and B, we define conditional probabilities as:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \text{and} \qquad P(B \mid A) = \frac{P(A \cap B)}{P(A)}. \tag{5.1}$$

In words: 'the probability of one event, given a second event, is equal to the probability of both, divided by the probability of the second (conditioning) event.'

This is the simplest form of Bayes' formula, and this can be expressed in other ways. Rearranging (5.1), we obtain:

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A),$$

from which we can derive Bayes' formula.

Bayes' formula

The simplest form of Bayes' formula is:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

5.10.2 Total probability formula

The simplest case of the total probability formula involves calculating the probability of an event A from information about its two conditional probabilities with respect to some other event B and its complement, B^c, together with knowledge of $P(B)$. Note that:

- B and B^c are mutually exclusive, and
- B and B^c are collectively exhaustive.

Fulfilment of these criteria (being mutually exclusive and collectively exhaustive) allows us to view B and B^c as a **partition** of the sample space.

The (simplest form of the) total probability formula

The total probability formula is:

$$P(A) = P(A \mid B) \cdot P(B) + P(A \mid B^c) \cdot P(B^c).$$

In words: 'the probability of an event is equal to its conditional probability on a second event times the probability of the second event, plus its probability conditional on the second event not occurring times the probability of that non-occurrence.'

There is a more general form of the total probability formula. Let $B_1, B_2, \ldots, B_n$ partition the sample space S into n pairwise mutually exclusive (at most one of them can happen) and collectively exhaustive (at least one of them must happen) events. For example, for $n = 4$, see Figure 5.4.

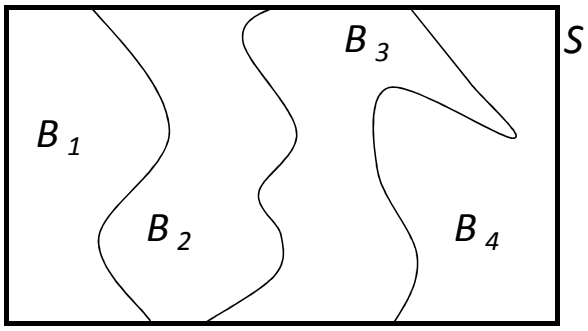


Figure 5.4: An example of a partitioned sample space.

Figure 5.5 superimposes an event A.

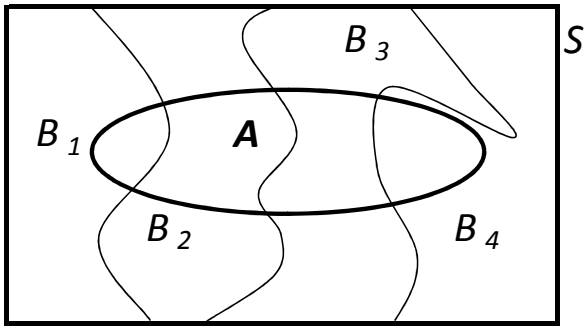


Figure 5.5: The event A within a partitioned sample space.

Extending the simple form of the total probability formula, we obtain:

$$\begin{aligned} P(A) &= \sum_{i=1}^{n} P(A \mid B_i)P(B_i) \\ &= P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_n)P(B_n). \end{aligned}$$

Recall that $P(B \mid A) = P(A \cap B)/P(A) = P(A \mid B)P(B)/P(A)$, so assuming we have the partition B and B^c, then:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B) \cdot P(B) + P(A \mid B^c) \cdot P(B^c)}.$$

A more general partition gives us a more complete form of Bayes' formula.

General form of Bayes' formula

For a general partition of the sample space S into $B_1, B_2, \ldots, B_n$, and for some event A, then:

$$P(B_k \mid A) = \frac{P(A \mid B_k)P(B_k)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)}.$$

Activity 5.2 In an audit Bill analyses 60% of the audit items and George analyses 40%. Bill's error rate is 5% and George's error rate is 3%. Suppose an item is sampled at random.

(a) What is the probability that it is in error (audited incorrectly)?

(b) If the chosen item is incorrect what is the probability that Bill is to blame?

Activity 5.3 Two fair coins are tossed. You are told that 'at least one is a head'. What is the probability that both are heads?

5.10.3 Independent events (revisited)

The terms 'dependent' and 'independent' reflect the fact that the probability of an event is changed when another event is known to occur only if there is some dependence between the events. If there is such a dependence, then $P(A \mid B) \neq P(A)$.

It follows from this that two events, A and B, are independent if and only if:

$$P(A \mid B) = P(A).$$

Recall from the multiplicative law in Section 5.9, that under independence $P(A \cap B) = P(A) \cdot P(B)$. Substituting this into our conditional probability formula gives the required result:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A)$$

provided $P(B) > 0$. Hence if A and B are independent, knowledge of B, i.e. '$\mid B$', is of no value to us in determining the probability of A occurring.

5.11 Probability trees

Probability problems can often be represented in several ways. We have already seen how Venn diagrams provide a convenient way to visualise the members of various combinations of events.

Here, we introduce the **probability tree**, also referred to as a **tree diagram**. This is best explained by way of an example.

Example 5.12 The London Special Electronics company is investigating a fault in its manufacturing plant. The tests done so far prove that the fault is at one of three locations: A, B or C. Taking all the tests and other evidence into account, the company assesses the chances of the fault being at each site as:

Suspect site	A	B	C
Probability of it being site of fault	0.5	0.2	0.3

The next test they will do is expected to improve the identification of the correct site, but (like most tests) it is not entirely accurate.

- If the fault is at A, then there is a 70 per cent chance of a correct identification; i.e. given that A is the site of the problem then the probability that the test says that it is at A is 0.7, and in this case the probabilities of either possible error are equal.
- If the fault is at B, then there is a 60 per cent chance of a correct identification, and in this case too the probabilities of either possible error are equal.
- If the fault is at C, then there is an 80 per cent chance of a correct identification, and in this case too the probabilities of either possible error are equal.

Draw a probability tree for this problem, and use it to answer the following:

(a) What is the probability that the new test will (rightly or wrongly) identify A as the site of the fault?

(b) If the new test does identify A as the site of the fault, find:

i. the company's revised probability that C is the site of the fault

ii. the company's revised probability that B is not the site of the fault.

Let A, B and C stand for the events: 'fault is at A', 'fault is at B' and 'fault is at C', respectively. Also, let a, b and c stand for the events: '*the test says* the fault is at A', '*the test says* the fault is at B' and '*the test says* the fault is at C', respectively. The probability tree is shown in Figure 5.6.

(a) The probability $P(a)$ is the sum of the three values against 'branches' which include the event a, that is:

$$0.35 + 0.04 + 0.03 = 0.42.$$

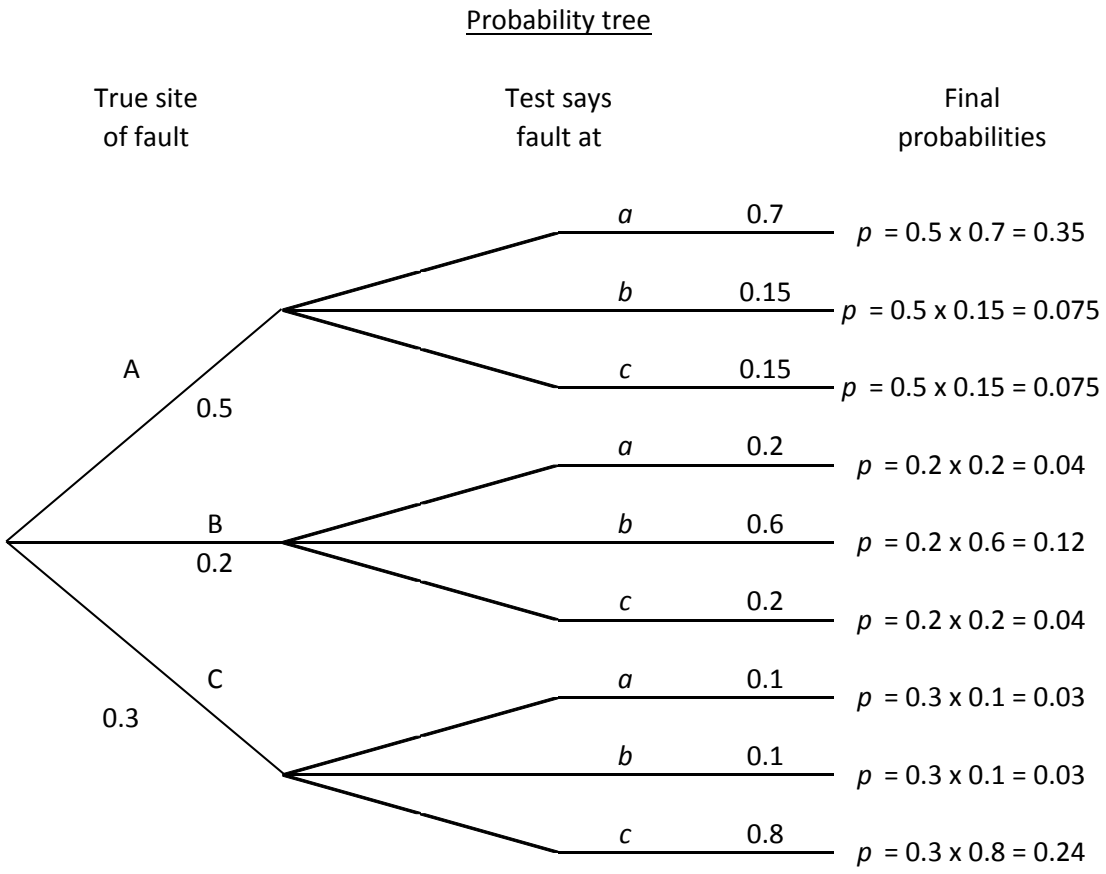


Figure 5.6: Probability tree for Example 5.12.

(b) i. The conditional probability $P(C \mid a)$ is the value for the $C \cap a$ branch divided by $P(a)$, that is:

$$\frac{0.03}{0.35 + 0.04 + 0.03} = \frac{1}{14} = 0.071.$$

ii. The conditional probability $P(B^c \mid a)$ is the sum of the values for the $A \cap a$ and $C \cap a$ branches divided by $P(a)$, that is:

$$\frac{0.35 + 0.03}{0.42} = 0.905.$$

Alternatively:

$$P(B^c \mid a) = 1 - \frac{\text{value of } (B \cap a) \text{ branch}}{P(a)} = 1 - \frac{0.04}{0.42} = 0.905.$$

5.12 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.

5.13 Key terms and concepts

- Additive law
- Bayes' formula
- Conditional
- Equally likely
- Exhaustive
- Independence
- Multiplicative law
- Partition
- Probability tree
- Sample space
- Subset
- Venn diagram
- Axioms
- Collectively exhaustive
- Element
- Event
- Experiment
- Member
- Mutually exclusive
- Probability
- Relative frequency
- Set
- Tree diagram

5.14 Learning activities

1. When throwing a die, we have:

$$S = \{1, 2, 3, 4, 5, 6\}, \qquad E = \{3, 4\}, \qquad \text{and} \qquad F = \{4, 5, 6\}.$$

Determine:

(a) F^c

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

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

(d) $E^c \cap F$.

2. Consider the following information.

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

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?

3. Draw the appropriate Venn diagram to show each of the following in connection with Question 1:
   (a) $E \cup F = \{3, 4, 5, 6\}$
   (b) $E \cap F = \{4\}$
   (c) $E^c = \{1, 2, 5, 6\}$.

4. There are three sites a company may move to: A, B and C. We are told that $P(A)$ (the probability of a move to A) is $1/2$, and $P(B) = 1/3$. What is $P(C)$?

5. Two events A and B are independent with $P(A) = 1/3$ and $P(B) = 1/4$. What is $P(A \cap B)$?

6. A company gets 60% of its supplies from manufacturer A, and 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) Show that the sum of the probabilities of all outcomes is 1.
   (c) The way the tree is used depends on the information required. For example, show that the tree can be used to show that the probability of receiving a bad part is $0.028 + 0.018 = 0.046$.

7. 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 such 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.

   (Use set theory and the laws of probability, or a probability tree.)

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

In addition, attempt the 'Test your understanding' self-test quizzes available on the VLE.

5.15 Further exercises

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060]. Relevant exercises from Sections 3.1–3.3 and 3.5.

5.16 A reminder of your learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- apply the ideas and notation involved in set theory to simple examples
- recall the basic axioms of probability and apply them
- distinguish between the ideas of conditional probability and independence
- draw and use appropriate Venn diagrams
- draw and use appropriate probability trees.

5.17 Sample examination questions

1. Say whether the following statement is **true** or **false** and briefly give your reason(s). 'If two events are independent, then they must be mutually exclusive.'

2. If X can take values of 1, 2 and 4 with $P(X = 1) = 0.3$, $P(X = 2) = 0.5$ and $P(X = 4) = 0.2$, what are:
   (a) $P(X^2 < 4)$
   (b) $P(X > 2 \,|\, X \text{ is an even number})$?

3. Write down and illustrate the use in probability of:
   (a) the addition law
   (b) the multiplicative rule.

4. A student can enter a course either as a beginner (73% of all students) or as a transferring student (27% of all students). It is found that 62% of beginners eventually graduate, and that 78% of transferring students eventually graduate.
   (a) Find:
      i. the probability that a randomly chosen student is a beginner who will eventually graduate
      ii. the probability that a randomly chosen student will eventually graduate
      iii. the probability that a randomly chosen student is either a beginner or will eventually graduate, or both.
   (b) Are the events 'Eventually graduates' and 'Enters as a transferring student' statistically independent?
   (c) If a student eventually graduates, what is the probability that the student entered as a transferring student?

(d) 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?

5. A coffee machine may be defective because it dispenses the wrong amount of coffee, event C, and/or it dispenses the wrong amount of sugar, event S. The probabilities of these defects are:

$$P(C) = 0.05, \qquad P(S) = 0.04 \qquad \text{and} \qquad P(C \cap S) = 0.01.$$

What proportion of cups of coffee has:

(a) at least one defect

(b) no defects?

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

Chapter 6
The normal distribution and ideas of sampling

6.1 Aims

Now is the time where the ideas of measurement (introduced in Chapter 4) and probability (Chapter 5) come together to form the ideas of sampling. Your aims should be to:

- understand the concept of a random variable and its distribution
- work with the normal distribution
- see the connection between the mathematical ideas of sampling introduced here and their uses (introduced in Chapter 3, and covered in detail in Chapter 10).

6.2 Learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- summarise the meaning of $\mathrm{E}(X)$ and $\mathrm{Var}(X)$
- compute areas under the curve for a normal distribution
- state and apply the central limit theorem
- explain the relationship between sample size and the standard error of the sample mean.

6.3 Essential reading

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060] Sections 4.1, 4.3, 5.1–5.3, 6.1 and 6.2.

In addition there is essential 'watching' of this chapter's accompanying video tutorials accessible via the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

6.4 Further reading

- Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics for Business and Economics*. (Chichester: John Wiley & Sons, 1990) fourth edition [ISBN 9780471615170] Chapters 4 and 6.

6.5 Introduction

If all governmental political and economic decisions were based on figures from entire populations, think how costly and time-consuming that would be. You could argue that one would never in fact tackle the problem itself. If we needed to know exactly how many homeless families there are in a particular country before implementing policy, or the exact number of companies which employ people on a part-time basis, and for how many hours, then work might scarcely get off the ground! Think, for example, of the population of your country and the cost of a complete census, even if only limited questions are asked about residents.

Fortunately for everyone, statisticians have come up with a way of *estimating* such figures using samples from the population. They can also tell us how accurate these estimates are likely to be for a particular sample size and this is why survey sampling, rather than taking a full census, is the preferred tool of governments, social researchers and market research companies! The rest of this chapter develops the ideas you need in order to use random sampling successfully.

6.6 The random variable

Formally, a **random variable** is a 'mapping' of the elementary outcomes in the sample space to real numbers. This allows us to attach probabilities to the experimental outcomes. Hence the concept of a random variable is that of a measurement which takes a particular value for each possible trial (experiment).

Frequently, this will be a numerical value. For example, suppose we sample five people and measure their heights, hence 'height' is the random variable and the five (observed) values of this random variable are the realised measurements for the heights of these five people.

As a further example, suppose a fair die is thrown four times and we observe two 6s, a 3 and a 1. The random variable is the 'score on the die', and for these four trials it takes the values 6, 6, 3 and 1. (In this case, since we do not know the true order in which the values occurred, we could also say that the results were 1, 6, 3 and 6 or 1, 3, 6 and 6, etc.)

An example of an experiment with non-numerical outcomes would be a coin toss, for which recall $S = \{H, T\}$. We can use a random variable, X, to convert the sample space elements to real numbers:

$$X = \begin{cases} 1 & : \text{ if heads} \\ 0 & : \text{ if tails.} \end{cases}$$

The value of any of the aforementioned measurable variables will typically vary from sample to sample, hence the name 'random variable'.

So each experimental random variable has a collection of possible outcomes, and a numerical value associated with each outcome. We have already encountered the term 'sample space' which is the set of all possible (numerical) values of the random variable.

Random variables come in two 'types': **discrete** and **continuous**. Discrete random variables are those associated with *count data*, such as the score on a die, while continuous random variables refer to *measurable data*, such as height, weight and time.

Example 6.1 Examples of discrete random variables include:

Experiment	Random variable	Sample space
Die is thrown	Value on top face	$\{1, 2, 3, 4, 5, 6\}$
Coin is tossed five times	Number of heads	$\{0, 1, 2, 3, 4, 5\}$
Twenty people sampled	Number with blue eyes	$\{0, 1, 2, \ldots, 19, 20\}$
Machine operates for a day	Number of breakdowns	$\{0, 1, 2, \ldots\}$

Example 6.2 Possible examples of continuous random variables include:

- In economics, measuring values of inputs or outputs, workforce productivity or consumption.
- In sociology, measuring the proportion of people in a population with a particular preference.
- In engineering, measuring the electrical resistance of materials.
- In physics, measuring temperature, electrical conductivity of materials.

Note the repeated appearance of 'measuring' in the above examples. Hence continuous variables deal with measured data, while discrete variables deal with count data.

We typically use a **capital** letter to denote the random variable. For example, $X =$ 'score of a die'. The letter X is often adopted, but it is perfectly legitimate to use any other letter: Y, Z, etc. In contrast, a **lower case** letter denotes a particular *value* of the random variable. For example, if the die results in a 3, then this is $x = 3$.

A natural question to ask is 'what is the probability of any of these values?'. That is, we are interested in the **probability distribution** of the random variable. Special probability distributions are the subject matter of **ST104b Statistics 2**; however, there is one very special distribution which will concern us – the **normal distribution**. Before the normal distribution is introduced though, it is necessary to introduce two important measures.

6.7 Population mean and variance

'Mean' and 'variance' were introduced in Chapter 4, although importantly these referred to the **sample** versions, i.e. $\bar{x}$ and s^2, which were calculated based on **sample data**. Here, our attention is on the **population (theoretical)** mean and variance. These play a key role in sampling theory which will occupy us for the next few chapters. In order to work with these measures, we first need to introduce the **expectation operator, 'E'**.

6.7.1 Population mean

Certain important properties of distributions arise if we consider probability-weighted averages of random variables, and of functions of random variables.[1] For example, we might want to know the **average** of a random variable.

It would be foolish to simply take the average of all the values taken by the random variable, as this would mean that very unlikely values (those with small probabilities of occurrence) would receive the same weighting as very likely values (those with large probabilities of occurrence). The obvious approach is to use the **probability-weighted average** of the sample space values.

Expected value of a discrete random variable

If $x_1, x_2, \ldots, x_N$ are the possible values of the random variable X, with corresponding probabilities $p_1, p_2, \ldots, p_N$, then:

$$\mathrm{E}(X) = \mu_X = \mu = \sum_{i=1}^{N} p_i x_i = p_1 x_1 + p_2 x_2 + \cdots + p_N x_N.$$

Note this **population mean** can be written as $\mathrm{E}(X)$ (in words 'the expectation of the random variable X'), or μ_X (in words 'the (population) mean of X'). Also, note the distinction between the *sample* mean, $\bar{x}$, (introduced in Chapter 4) based on observed sample values, and the *population* mean, μ_X, based on the theoretical probability distribution.

Example 6.3 If the 'random variable' X happens to be a constant, then $x_1 = x_2 = \cdots = x_N = k$, and it is always the case that $p_1 = p_2 = \cdots = p_N = 1$, so trivially $\mathrm{E}(X) = k$.

Example 6.4 Let X represent the value shown when a fair die is thrown once. Its **probability distribution** (that is, how the sample space probability of 1 is distributed across the values the random variable takes) is:

$X = x$	1	2	3	4	5	6	Total
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6	1

Note that the sum of the individual probabilities is 1 (a consequence of the second

[1] A function, $f(X)$, of a random variable X is, of course, a new random variable, say $Y = f(X)$.

probability axiom). So the (population) mean is calculated to be:

$$\mathrm{E}(X) = \sum_{i=1}^{6} p_i x_i = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \cdots + \frac{1}{6} \cdot 6 = 3.5.$$

Example 6.5 Suppose the random variable X takes the values 3, 6 and 8 with probabilities 0.2, 0.7 and 0.1, respectively. Hence its probability distribution is:

$X = x$	3	6	8	Total
$P(X = x)$	0.2	0.7	0.1	1

The population mean is calculated to be:

$$\mathrm{E}(X) = \sum_{i=1}^{3} p_i x_i = (0.2 \times 3) + (0.7 \times 6) + (0.1 \times 8) = 5.6.$$

6.7.2 Population variance

The concept of a probability-weighted average (or expected value) can be extended to functions of the random variable. For example, if X takes the values $x_1, x_2, \ldots, x_N$ with corresponding probabilities $p_1, p_2, \ldots, p_N$, then:

$$\mathrm{E}\left(\frac{1}{X}\right) = \sum_{i=1}^{N} p_i \frac{1}{x_i} \qquad \text{provided } x_i \neq 0$$

$$\mathrm{E}(X^2) = \sum_{i=1}^{N} p_i x_i^2.$$

One very important average associated with a distribution is the expected value of the square of the deviation[2] of the random variable from its mean, μ, known as the population variance, σ^2.

Variance of a discrete random variable

If $x_1, x_2, \ldots, x_N$ are the possible values of the random variable X, with corresponding probabilities $p_1, p_2, \ldots, p_N$, then:

$$\sigma^2 = \mathrm{E}((X - \mu)^2) = \sum_{i=1}^{N} p_i (x_i - \mu)^2.$$

This can be seen to be a measure – not the only one, but the most widely used by far – of the spread of the distribution and is known as the **(population) variance** of the random variable.

[2]Which roughly means 'distance with sign'.

The (positive) square root of the variance is known as the **standard deviation** and, given the variance is typically denoted by σ^2, the standard deviation is denoted by σ.

Example 6.6 Returning to the fair die example, we now compute the variance of X as follows, noting that $\mu = 3.5$.

$X = x$	1	2	3	4	5	6	Total
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6	1
$(x - \mu)^2$	25/4	9/4	1/4	1/4	9/4	25/4	
$(x - \mu)^2 \cdot P(X = x)$	25/24	9/24	1/24	1/24	9/24	25/24	$70/24 = 2.92$

Hence $\sigma^2 = \mathrm{E}((X - \mu)^2) = 2.92$ and $\sigma = \sqrt{2.92} = 1.71$.

The tabular format in Example 6.6 has some advantages. Specifically:

- It helps to have (and to calculate) the 'Total' column since, for example, if a probability, $P(X = x)$, has been miscalculated or miscopied then the row total will not be 1 (recall the second probability axiom). This would therefore highlight an error so, with a little work, could be identified.
- It is often useful to do a group of calculations as fractions over the same denominator (as in the final row of the table in Example 6.6), rather than to cancel or to work with them as decimals, because important patterns can be more obvious, and calculations easier.

It can be shown[3] that the population variance can be expressed in an alternative form:

$$\mathrm{Var}(X) = \mathrm{E}((X - \mu)^2) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 .$$

In words, 'the (population) variance is equal to the mean of the square minus the square of the mean'.

Example 6.7 Using the same single die example from Example 6.6:

$X = x$	1	2	3	4	5	6	Total
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6	1
$x \cdot P(X = x)$	1/6	2/6	3/6	4/6	5/6	6/6	$21/6 = 3.5 = \mu$
$x^2 \cdot P(X = x)$	1/6	4/6	9/6	16/6	25/6	36/6	91/6

Hence $\mu = \mathrm{E}(X) = 3.5$, $\mathrm{E}(X^2) = 91/6$, so the variance is $91/6 - (3.5)^2 = 2.92$, as before. However, this method is usually easier.

[3] This means the result is important, but the derivation of the result is not required for this course.

6.8 The normal distribution

The **normal distribution** is the most important distribution in statistical theory and it is essential to much statistical theory and reasoning. It is, in a sense, the 'parent' distribution of all the *sampling distributions* which we shall meet later.

In order to get some feel for the normal distribution, let us consider the exercise of constructing a histogram of people's heights (assumed to be normally distributed). Suppose we start with 100 people and construct a histogram using sufficient class intervals such that the histogram gives a good representation of the data's distribution. This will be a fairly 'ragged' diagram, but useful nonetheless.

Now suppose we increase our sample size to 500 and construct an appropriate histogram for these observed heights, but using more class intervals now that we have more data. This diagram will be smoother than the first, peaked in the centre, and roughly symmetric about the centre. The normal distribution is emerging! If we continue this exercise to sample sizes of 5,000 or even 50,000, then we will eventually arrive at a very smooth bell-shaped curve similar to that shown in Figure 6.1. Hence we can view the normal distribution as the smooth limit of the basic histogram as the sample size becomes very large.

Such a diagram represents the distribution of the population. It is conventional to adjust the vertical scale so that the total area under the curve is 1, and so it is easy to view the area under the curve as probability (recall the second probability axiom). The mathematical form for this curve is well-known and can be used to compute areas, and thus probabilities – in due course we shall make use of statistical tables for this purpose.

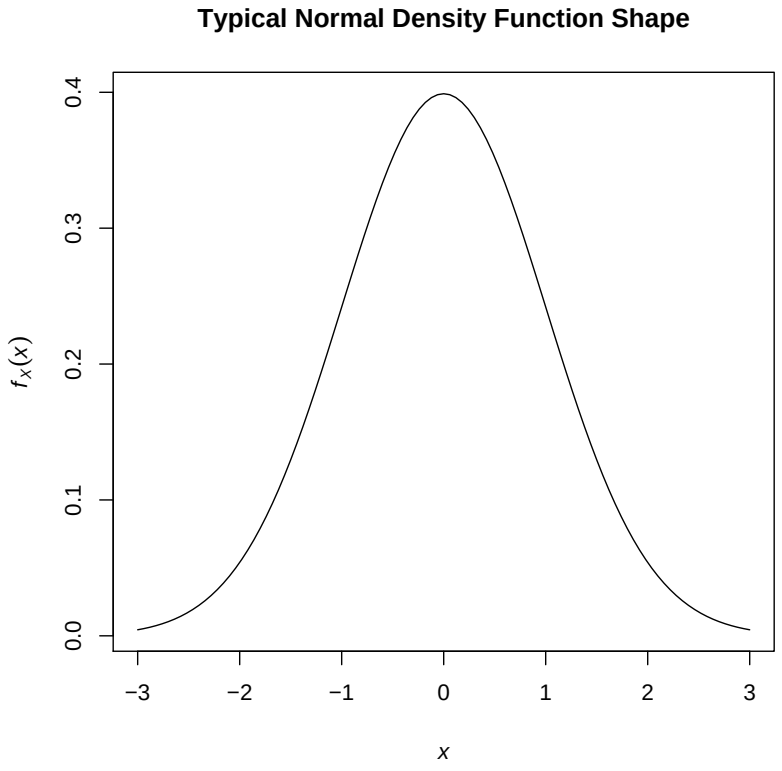


Figure 6.1: Density function of the (standard) normal distribution.

6.8.1 Relevance of the normal distribution

The normal distribution is relevant to the application of statistics for many reasons such as:

- Many naturally occurring phenomena can be *modelled* as following a normal distribution. Examples include heights of people, diameters of bolts, weights of pigs, etc.[4]

- A very important point is that averages of sampled variables (discussed later), indeed *any functions* of sampled variables, also have probability distributions. It can be demonstrated, theoretically and empirically, that, provided the sample size is reasonably large, the distribution of the sample mean, $\bar{X}$, will be (approximately) normal regardless of the distribution of the original variable. This is known as the **central limit theorem (CLT)** which we will return to later.

- The normal distribution is often used as the distribution of the error term in standard statistical and econometric models such as linear regression. This assumption can be, and should be, checked. This is considered further in **ST104b Statistics 2**.

6.8.2 Consequences of the central limit theorem

The consequences of the CLT are twofold:

- A number of statistical methods that we use have a **robustness** property, i.e. it does not matter for their validity just what the true population distribution of the variable being sampled is.

- We are justified in assuming normality for statistics which are sample means or linear transformations of them.

The CLT was introduced above 'provided the sample size is reasonably large'. In practice, 30 or more observations are usually sufficient (and can be used as a rule-of-thumb), although the distribution of $\bar{X}$ may be normal for n much less than 30. This depends on the distribution of the original (population) variable. If this population distribution is in fact normal, then all sample means computed from it will be normal. However, if the population distribution is very non-normal, then a sample size of (at least) 30 would be needed to justify normality.

[4]Note the use of the word '*modelled*'. This is due to the 'distributional assumption' of normality. Since a normal random variable X is defined over the entire real line, i.e. $-\infty < x < \infty$, we know a person cannot have a negative height, even though the normal distribution has positive, non-zero probability over negative values. Also, nobody is of infinite height (the world's tallest man ever, Robert Wadlow, was 272 cms), so clearly there is a finite upper bound to height, rather than ∞. Therefore height does **not** follow a true normal distribution, but it is a good enough *approximation* for modelling purposes.

6.8.3 Characteristics of the normal distribution

The equation which describes the normal distribution takes the general form:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(x-\mu)^2/2\sigma^2}.$$

The shape of this function is the bell-shaped curve, such as in Figure 6.1. Don't panic, you will **not** be working with this function directly! That said, do be aware that it involves **two parameters**: the mean, μ, and the variance, σ^2.[5]

- Since the normal distribution is symmetric about μ, the distribution is centred at μ. As a consequence of this symmetry, the mean is equal to the median. Also, since the distribution peaks at μ, it is also equal to the mode. In principle, the mean can take any real value, i.e. $-\infty < \mu < \infty$.
- The variance is σ^2, hence the larger σ^2, the larger the spread of the distribution. Note that variances cannot be negative, hence $\sigma^2 > 0$.

If the random variable X has a normal distribution with parameters μ and σ^2, we denote this as $X \sim N(\mu, \sigma^2)$.[6] Given the infinitely many possible values for μ and σ^2, and given that a normal distribution is *uniquely defined* by these two parameters, there is an infinite number of normal distributions due to the infinite combinations of values for μ and σ^2.

The most important normal distribution is the special case when $\mu = 0$ and $\sigma^2 = 1$. We call this the **standard normal distribution**, denoted by Z, i.e. $Z \sim N(0, 1)$. Tabulated probabilities that appear in statistical tables are for the standard normal distribution.

6.8.4 Standard normal tables

We now discuss the determination of normal probabilities using standard statistical tables. (Extracts from the) *New Cambridge Statistical Tables* will be provided in the examination. Here we focus on Table 4.

Standard normal probabilities

Table 4 of the *New Cambridge Statistical Tables* lists 'lower-tail' probabilities, which can be represented as:

$$P(Z \leq z) = \Phi(z), \qquad z \geq 0$$

using the conventional Z notation for a standard normal variable.[7]

Note the *cumulative* probability[8] for the Z distribution, $P(Z \leq z)$, is often denoted $\Phi(z)$. We now consider some examples of working out probabilities from $Z \sim N(0, 1)$.

[5] 'Parameters' were introduced in Chapter 3.

[6] Read '$\sim$' as 'is distributed as'.

[7] Although Z is the conventional letter used to denote a standard normal variable, Table 4 uses (somewhat annoyingly) 'x' to denote 'z'.

[8] A cumulative probability is the probability of being **less than or equal to** some particular value.

Example 6.8 If $Z \sim N(0, 1)$, what is $P(Z > 1.2)$?

When computing probabilities, it is useful to draw a quick sketch to visualise the specific area of probability that we are after.

So for $P(Z > 1.2)$ we require the upper-tail probability shaded in red in Figure 6.2. Since Table 4 gives us lower-tail probabilities, if we look up the value 1.2 in the table we will get $P(Z \leq 1.2) = 0.8849$. The total area under a normal curve is 1, so the required probability is simply $1 - \Phi(1.2) = 1 - 0.8849 = 0.1151$.

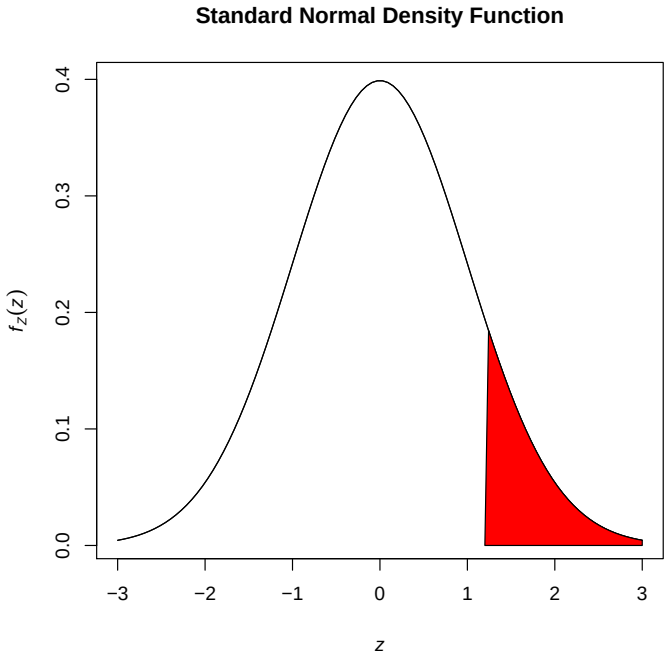


Figure 6.2: Standard normal distribution with shaded area depicting $P(Z > 1.2)$.

Example 6.9 If $Z \sim N(0, 1)$, what is $P(-1.24 < Z < 1.86)$?

Again, begin by producing a sketch.

The probability we require is the sum of the blue and red areas in Figure 6.3. Using Table 4, which note only covers $z \geq 0$, we proceed as follows.

The red area is given by:

$$\begin{aligned} P(0 \leq Z \leq 1.86) &= P(Z \leq 1.86) - P(Z \leq 0) \\ &= \Phi(1.86) - \Phi(0) \\ &= 0.9686 - 0.5 \\ &= 0.4686. \end{aligned}$$

The blue area is given by:

$$\begin{aligned} P(-1.24 \leq Z \leq 0) &= P(Z \leq 0) - P(Z \leq -1.24) \\ &= \Phi(0) - \Phi(-1.24) \\ &= \Phi(0) - (1 - \Phi(1.24)) \\ &= 0.5 - (1 - 0.8925) \\ &= 0.3925. \end{aligned}$$

Note by symmetry of Z about $\mu = 0$, $P(Z \leq -1.24) = P(Z \geq 1.24) = 1 - \Phi(1.24)$. So although Table 4 does not give probabilities for negative z values, we can exploit the symmetry of the (standard) normal distribution.

Hence $P(-1.24 < Z < 1.86) = 0.4686 + 0.3925 = 0.8611$.

Alternatively:

$$\begin{aligned} P(-1.24 < Z < 1.86) &= \Phi(1.86) - \Phi(-1.24) \\ &= \Phi(1.86) - [1 - \Phi(1.24)] \\ &= 0.9686 - (1 - 0.8925) \\ &= 0.8611. \end{aligned}$$

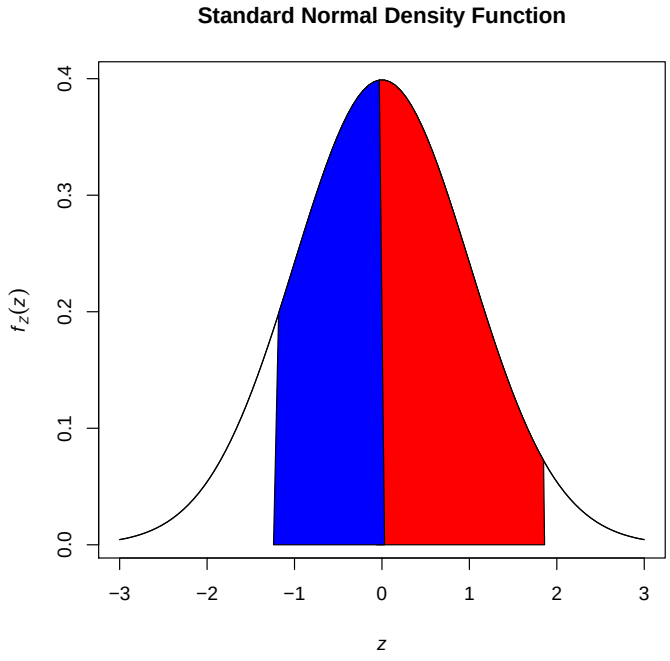


Figure 6.3: Standard normal distribution depicting $P(-1.24 < Z < 1.86)$ as the shaded areas.

6.8.5 The general normal distribution

We have already discussed that there exists an infinite number of different normal distributions due to the infinite pairs of parameter values since $-\infty < \mu < \infty$ and

$\sigma^2 > 0$. The good news is that Table 4 of the *New Cambridge Statistical Tables* can be used to determine probabilities for *any* normal random variable X, such that $X \sim N(\mu, \sigma^2)$.

To do so, we need a little bit of magic – **standardisation**. This is a special transformation which converts $X \sim N(\mu, \sigma^2)$ into $Z \sim N(0, 1)$.

The transformation formula for standardisation

If $X \sim N(\mu, \sigma^2)$, then the transformation:

$$Z = \frac{X - \mu}{\sigma}$$

creates a standard normal random variable, i.e. $Z \sim N(0, 1)$. So to standardise X we subtract the mean and divide by the standard deviation.

To see why,[9] first note that any linear transformation of a normal random variable is also normal – hence as X is normal, so too is Z, since the standardisation transformation is linear in X. It remains to show that standardisation results in a random variable with zero mean and unit variance.

Since $X \sim N(\mu, \sigma^2)$, we have:

$$\mathrm{E}(Z) = \mathrm{E}\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma}\mathrm{E}(X - \mu) = \frac{1}{\sigma}(\mathrm{E}(X) - \mu) = \frac{1}{\sigma}(\mu - \mu) = 0. \qquad (6.1)$$

This result exploits the fact that σ is a constant, hence can be taken outside the expectation operator. Turning to the variance, we have:

$$\mathrm{Var}(X) = \mathrm{Var}\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma^2}\mathrm{Var}(X - \mu) = \frac{1}{\sigma^2}\mathrm{Var}(X) = \frac{1}{\sigma^2}\sigma^2 = 1. \qquad (6.2)$$

This result uses the fact that we must square a constant when taking it outside the 'Var' operator.

Example 6.10

Suppose $X \sim N(5, 4)$. What is $P(5.8 < X < 7.0)$?

[9] This bit of theory, that is (6.1) and (6.2), is purely for illustrative purposes (and for the interested student!), you will **not** have to reproduce this in the examination.

$$
\begin{aligned}
P(5.8 < X < 7.0) &= P\left(\frac{5.8-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{7.0-\mu}{\sigma}\right) \\
&= P\left(\frac{5.8-5}{\sqrt{4}} < \frac{X-5}{\sqrt{4}} < \frac{7.0-5}{\sqrt{4}}\right) \\
&= P(0.4 < Z < 1) \\
&= P(Z \leq 1) - P(Z \leq 0.4) \\
&= 0.8413 - 0.6554 \qquad \text{(from Table 4)} \\
&= 0.1859.
\end{aligned}
$$

6.9 Sampling distributions

In Chapter 3, 'Statistics' was introduced as a discipline for data analysis. In this chapter we have encountered the normal probability distribution. Probability distributions typically have associated **parameters**, such as the (theoretical) mean, μ, and the (theoretical) variance, σ^2, for the normal distribution. By convention, Greek letters are used to denote **population parameters**, whose values in practice are typically *unknown*.

The next few chapters of this course will be the subject of **statistical inference**, whereby we *infer* unknown population parameters based on **sample** data. As an example, suppose we wanted to investigate the height of the UK population. As previously discussed, it is reasonable to assume that height is a normally-distributed random variable with some mean μ and some variance σ^2. What are the *exact* values of these parameters? To know these values precisely would require data on the heights of the entire UK population – all 60-plus million people!

Population sizes, denoted by N, are typically very large and clearly no-one has the time, money or patience to undertake such a marathon data collection exercise. Instead we opt to collect a **sample** (some subset of the population) of size n.[10] Having collected our sample, we then **estimate** the *unknown* population parameters based on the *known* (observed) sample data. Specifically, we **estimate population quantities based on their respective sample counterparts**.

A 'statistic' (singular noun) is just some known function calculated from data. A *sample statistic* is calculated from *sample data*. At this point, be aware of the following distinction between an estimator and an estimate.

[10] If $n = N$, and we sample without replacement, then we have obtained a complete enumeration of the population – a census.

Estimator versus estimate

An **estimator** is a statistic (which is a random variable) describing how to obtain a **(point) estimate** (which is a real number) of a population parameter.

Example 6.11 The sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$ is the estimator for the population mean, μ. If we had drawn at random from the population the sample data $4, 8, 2, 6$, then the (point) estimate for μ would be $\bar{x} = 20/4 = 5$. Notice the notation – the estimator is written as 'capital' $\bar{X}$ and is a **random variable**, while the estimate is written as 'lower case' $\bar{x}$ as it is computed for a specific sample, hence is **fixed (constant)** for that particular sample. Had the random sample been instead $2, 9, 1, 4$ (from the same population), then the estimator for μ would still be $\bar{X} = \sum_{i=1}^{n} X_i/n$, but the estimate would now be $\bar{x} = 16/4 = 4$.

6

So we see that sample statistics vary from sample to sample due to the random nature of the sample. Hence estimators are random variables with corresponding probability distributions, namely **sampling distributions**.

Sampling distribution

A sampling distribution is the probability distribution of an estimator.

Before we proceed further, let us take a moment to review some population quantities and their respective sample counterparts.

Population quantity	Sample counterpart
Probability distribution	Histogram
(Population) mean, μ	(Sample) mean, $\bar{x}$
(Population) variance, σ^2	(Sample) variance, s^2
(Population) standard deviation, σ	(Sample) standard deviation, s
(Population) proportion, π	(Sample) proportion, p

The *precision* (or quality) of point estimates such as $\bar{x}$, s^2 and p will depend on the sample size n, and in principle on the population size N, if finite. In practice if N is large relative to n, then we can use approximations which are more than adequate for practical purposes, but would only be completely accurate if the population truly was infinite. In what follows we assume N is large enough to be treated as infinite.

6.10 Sampling distribution of $\bar{X}$

Let us consider a large population (big enough to be treated as if it was infinite) which is normally distributed, i.e. $N(\mu, \sigma^2)$. Suppose we take an initial random sample of size n and obtain the sample mean, $\bar{x}_1$. Next we take a second, independent sample (from

the same initial population) and compute $\bar{x}_2$. Continue taking new, independent samples, resulting in a series of sample means: $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots$. These values will, of course, vary from sample to sample, hence if we constructed a histogram of these values we would see the *empirical* sampling distribution of $\bar{X}$. Fortunately, however, we are often able to determine the exact, *theoretical* form of sampling distributions without having to resort to such empirical 'simulations'.

Sampling distribution of $\bar{X}$ for normal populations

When taking a random sample of size n from a $N(\mu, \sigma^2)$ population, it can be shown [11] that:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

The central limit theorem concerns the sampling distribution of $\bar{X}$ when sampling from any non-normal distribution (with some exceptions).

Central limit theorem

When taking a random sample of size n from a non-normal population with finite mean μ and finite variance σ^2, then:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

approximately, as $n \to \infty$.

So the difference between sampling from normal and non-normal populations is that $\bar{X}$ is *exactly* normally distributed in the former case, but only *approximately* so in the latter case. The approximation is reasonable for n at least 30, as a rule-of-thumb. Although because this is an asymptotic approximation (i.e. as $n \to \infty$), the bigger n is, the better the normal approximation.

We can use standardisation to compute probabilities involving $\bar{X}$, but we must remember to divide by $\sigma/\sqrt{n}$ rather than σ, since the variance of $\bar{X}$ is σ^2/n. $\sigma/\sqrt{n}$ is known as the **standard error** of $\bar{X}$.

Standard error

The standard deviation of an estimator is called its standard error.

For example, $\bar{X} \sim N(\mu, \sigma^2/n)$ (either exactly, or approximately by the CLT), so $\mathrm{Var}(\bar{X}) = \sigma^2/n$, hence the standard deviation of $\bar{X}$, i.e. the standard error, is:

$$\sqrt{\mathrm{Var}(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

Note the standard error decreases as n increases.

[11] Which means the actual proof is beyond the scope of this course.

Example 6.12 A random sample of 16 values is taken from a normal distribution with $\mu = 25$ and $\sigma = 3$. What is $P(\bar{X} > 26)$?

Since the population is normally distributed, the sampling distribution of $\bar{X}$ is exactly $N(\mu, \sigma^2/n) = N(25, 9/16)$. If Z is the standard normal random variable, then:

$$P(\bar{X} > 26) = P\left(Z > \frac{26 - 25}{3/\sqrt{16}}\right) = P(Z > 1.33) = 1 - \Phi(1.33) = 1 - 0.9082 = 0.0918$$

using Table 4 of the *New Cambridge Statistical Tables*.

6.11 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 7, 8 and 9, and work on ideas of statistical **estimation** and **inference**. Do not worry if you found some later sections of this chapter difficult to understand. Work at the learning activities and the sample examination questions below.

6.12 Key terms and concepts

- Central limit theorem
- Normal distribution
- Population variance
- Random variable
- Standard error
- Statistical inference
- Expectation operator
- Population mean
- Probability distribution
- Sampling distribution
- Standardisation
- Statistical tables

6.13 Learning activities

1. Check the following using Table 4 of the *New Cambridge Statistical Tables*.

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

   (a) $P(Z \geq 1) = 1 - \Phi(1) = 0.1587$

   (b) $P(Z \leq 1) = \Phi(1) = 1 - 0.1587 = 0.8413$.

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 deviations of the mean.

   (c) 99% of normal random variables fall within 3 standard deviations of the mean.

   Draw the areas concerned on a normal distribution.

3. The following six observations give the time taken, in seconds, to complete a 100-metre sprint by all six individuals competing in the race – hence this is population data.

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

   (a) Find the population mean, μ, and the population standard deviation, σ, of the sprint times.

   (b) Calculate the sample mean for each possible sample of:

      i. two individuals

      ii. three individuals

      iii. four individuals.

   (c) Work out the mean for each set of sample means (it must come to μ!) and compare the standard deviations of the sample means about μ.

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

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

In addition, attempt the 'Test your understanding' self-test quizzes available on the VLE.

6.14 Further exercises

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060]. Relevant exercises from Sections 4.1, 4.3, 5.1–5.3, 6.1 and 6.2.

6.15 A reminder of your learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- summarise the meaning of $\mathrm{E}(X)$ and $\mathrm{Var}(X)$
- compute areas under the curve for a normal distribution
- state and apply the central limit theorem

- explain the relationship between sample size and the standard error of the sample mean.

6.16 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?

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?

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 55 with a standard deviation of 6.

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

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

Chapter 7
Estimation

7.1 Aims

This chapter develops the ideas of sampling introduced in Chapter 6 and looks at the actual estimation of population parameters using both normal and Student's t distributions. Your aims are to:

- estimate values for common parameters such as means and proportions
- decide which is the appropriate distribution to use – normal or Student's t.

7.2 Learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- calculate sample means, standard deviations and proportions, and demonstrate their use as estimators
- construct a confidence interval for a sample mean, a sample proportion, the difference between two sample means and the difference between two sample proportions
- know when to use Student's t distribution.

You do **not** need to:

- demonstrate the central limit theorem
- know about the concepts relating to a 'good' estimator.

7.3 Essential reading

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060] Sections 7.1–7.3 and 8.1–8.4.

In addition there is essential 'watching' of this chapter's accompanying video tutorials accessible via the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

7.4 Further reading

- Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics for Business and Economics*. (Chichester: John Wiley & Sons, 1990) fourth edition [ISBN 9780471615170] Chapter 7 and Sections 8.1–8.3 and 8.5.

7.5 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 (known as 'elements'). 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 (for example, 'How many items does an individual purchase per year, on average?') or a proportion ('What proportion of items manufactured has 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** (due to time and financial constraints), so instead one collects information about a **sample** drawn from the population and infers the required information about the population. In **ST104a Statistics 1**, we will look at the most commonly-used estimators. If you take **ST104b Statistics 2**, you will learn what the properties of a 'good' estimator are and look at other measures to be estimated apart from means and proportions.

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

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

7.6 Principle of confidence intervals

A point estimate is our 'best guess' of an unknown population parameter based on sample data. Due to the random nature of sample data, we do not expect to estimate the parameter exactly (unless we are very lucky). Hence there is some uncertainty in our estimation so it would be advisable to communicate the level of uncertainty (or imprecision) in conjunction with the point estimate.

Standard errors (recall the standard error is the square root of the variance, i.e. the standard deviation, of an estimator) act as measures of estimation (im)precision and these are used in the construction of **confidence intervals** – the subject of this chapter. Informally, you can think of a confidence interval as representing our 'best guess plus or minus a bit', where the magnitude of this 'bit' is dependent on the level of precision. More formally, an $x\%$ confidence interval **covers** the unknown parameter with $x\%$ probability **over repeated samples**.[1] This method of expressing the accuracy of an estimate is easily understood and requires no statistical sophistication for interpretation.

So, it is important to distinguish *point estimation* (using sample data to obtain a numerical estimate of an unknown population parameter) from a *confidence interval* (an interval estimate of the parameter whose width indicates how reliable the point estimate is). At this point, you may wish to reread Section 6.9 to ensure that you are clear about the distinction between *statistics* (such as point and interval estimators) and *parameters* (population characteristics).

Clearly, a very wide confidence interval would show that our estimate was not very reliable, as we could not be sure that its value was close to the true parameter value, whereas a narrow confidence interval would correspond to a more reliable estimate. The degree of confidence that we have in our confidence interval can be expressed numerically; an ideal situation is a narrow interval with a high **coverage probability**. With these points in mind we now show how such intervals can be computed from data in some basic situations.

7.7 General formulae for normally-distributed statistics

Chapter 6 introduced the sampling distribution of $\bar{X}$. We will now draw on this sampling theory.

7.7.1 Standard error known

Let $\hat{\theta}$ be an unbiased[2] estimator for θ such that its sampling distribution is:

$$N(\theta,\, \mathrm{Var}(\hat{\theta})) = N(\theta,\, (\mathrm{S.E.}(\hat{\theta}))^2)$$

recalling that the standard error of an estimator, denoted here as $\mathrm{S.E.}(\hat{\theta})$, is the square root of its variance, denoted here as $\mathrm{Var}(\hat{\theta})$. Assume this standard error is **known**. We require a confidence interval defined by a pair of values such that the probability of the interval covering θ, the *coverage probability*, is high.

Since $\hat{\theta}$ is normally distributed it follows that, upon standardising $\hat{\theta}$, we have:

$$Z = \frac{\hat{\theta} - \theta}{\mathrm{S.E.}(\hat{\theta})} \sim N(0,\, 1).$$

[1] A useful phrase in examinations!

[2] An unbiased estimator means the point estimate is correct *on average*, over repeated samples. This concept is covered in greater depth in **ST104b Statistics 2.**

Hence, assuming we desire a 95% coverage probability, we have:

$$P\left(-1.96 < \frac{\hat{\theta} - \theta}{\text{S.E.}(\hat{\theta})} < 1.96\right) = 0.95$$

using Table 4 of the *New Cambridge Statistical Tables*. Note we can also write 1.96 as $z_{0.025}$, that is the z-value which cuts off 2.5% probability in the upper tail of the standard normal distribution.

Since $\text{S.E.}(\hat{\theta}) > 0$ (a standard error must be strictly positive):

$$\begin{aligned} 0.95 &= P\left(-1.96 < \frac{\hat{\theta} - \theta}{\text{S.E.}(\hat{\theta})} < 1.96\right) \\ &= P\left(-1.96 \times \text{S.E.}(\hat{\theta}) < \hat{\theta} - \theta < 1.96 \times \text{S.E.}(\hat{\theta})\right) \\ &= P\left(-1.96 \times \text{S.E.}(\hat{\theta}) < \theta - \hat{\theta} < 1.96 \times \text{S.E.}(\hat{\theta})\right) \\ &= P\left(\hat{\theta} - 1.96 \times \text{S.E.}(\hat{\theta}) < \theta < \hat{\theta} + 1.96 \times \text{S.E.}(\hat{\theta})\right) \end{aligned}$$

where we multiply by -1 to go from the second to the third line, and therefore the inequality sign is reversed.

Endpoints for a 95% confidence interval (standard error known)

A 95% confidence interval for θ has **endpoints** $\hat{\theta} \pm 1.96 \times \text{S.E.}(\hat{\theta})$. Hence the reported confidence interval would be:

$$(\hat{\theta} - 1.96 \times \text{S.E.}(\hat{\theta}),\ \hat{\theta} + 1.96 \times \text{S.E.}(\hat{\theta})).$$

This is a simple, but very important, result. As we shall see, it can be applied to give confidence intervals in many different situations such as the estimation of a mean, proportion, difference in means and difference in proportions.

The above derivation was for a 95% confidence interval, i.e. with a 95% coverage probability, which is a generally accepted confidence requirement. Of course, it is possible to have different levels of confidence, say 90% or 99%. Fortunately, we can use the same argument as above; however, a different multiplier coefficient drawn from the standard normal distribution is required (i.e. not $z_{0.025} = 1.96$). Using Table 5 of the *New Cambridge Statistical Tables*, it is easy to obtain such coefficients. For convenience, key values are given below, where z_α denotes the z-value which cuts off $100\alpha\%$ probability in the upper tail of the standard normal distribution.

For 90% confidence, use the multiplier	$z_{0.05} = 1.6449$
For 95% confidence, use the multiplier	$z_{0.025} = 1.9600$
For 99% confidence, use the multiplier	$z_{0.005} = 2.5758$.

Check these values from Table 5 in the *New Cambridge Statistical Tables*, and make sure you know how to work out other ones, for example 80% or 98% confidence.[3]

[3] These would be $z_{0.1} = 1.2816$ and $z_{0.01} = 2.3263$, respectively.

Unfortunately, the method used so far in this section is limited by the assumption that the standard error of $\hat{\theta}$ is known. This means, in effect, that we need to know the true population variance and, in most practical cases, the assumption that we know either the population variance or the sampling standard error will not be justified.

In such cases it will be necessary to **estimate the standard error** from the data. This requires a modification both of the approach and, in particular, of the general formula for the endpoints of a confidence interval.

7.7.2 Standard error unknown

Here we consider cases where the standard error of $\hat{\theta}$ is estimated, hereby denoted E.S.E.$(\hat{\theta})$ which stands for the estimated standard error of $\hat{\theta}$. It may be tempting to think that:

$$\frac{\hat{\theta} - \theta}{\text{E.S.E.}(\hat{\theta})} \sim N(0,\, 1)$$

i.e. it is standard normal, **but** because of the additional sampling variability of the estimated standard error, this new, transformed function of the data will have a more dispersed distribution than the standard normal – the Student's t distribution on ν degrees of freedom (discussed below). Hence:

$$\frac{\hat{\theta} - \theta}{\text{E.S.E.}(\hat{\theta})} \sim t_\nu.$$

7.7.3 Student's t distribution

'Student' was the pen name of William S. Gosset (1876–1937) who is credited with developing this distribution.[4] You should be familiar with its generic shape, which resembles the standard normal (i.e. bell-shaped and symmetric about 0) but with 'fatter' tails. We get different versions of the t distribution for different **degrees of freedom**,[5] ν. Graphical examples of the t distribution for various degrees of freedom are given in Figure 7.1, but note that as $\nu \to \infty$ (in words, 'as the degrees of freedom tend to infinity'), we approach the standard normal distribution – that is, $t_\nu \to N(0,\, 1)$ as $\nu \to \infty$ (in words, 'the Student's t distribution tends to the standard normal distribution as the degrees of freedom tend to infinity').

For our purposes, we will use this distribution whenever we are performing inference for population means when population variances are *unknown*, and hence estimated from the data. The correct degrees of freedom will depend on the degrees of freedom used to estimate the variance (more on this later).

Assuming a 95% coverage probability, using Table 10 of the *New Cambridge Statistical*

[4] Curious students may wonder why Gosset was not vain enough to call it Gosset's t distribution – after all, who wouldn't want to achieve such statistical immortality? In fact the reason for the pen name was due to the refusal of Guinness (his employer when he discovered this distribution) to allow its researchers to publish papers fearing leaks of trade secrets, forcing Gosset to publish anonymously. Poor Gosset.

[5] We will not discuss 'degrees of freedom' in any great depth in **ST104a Statistics 1.**

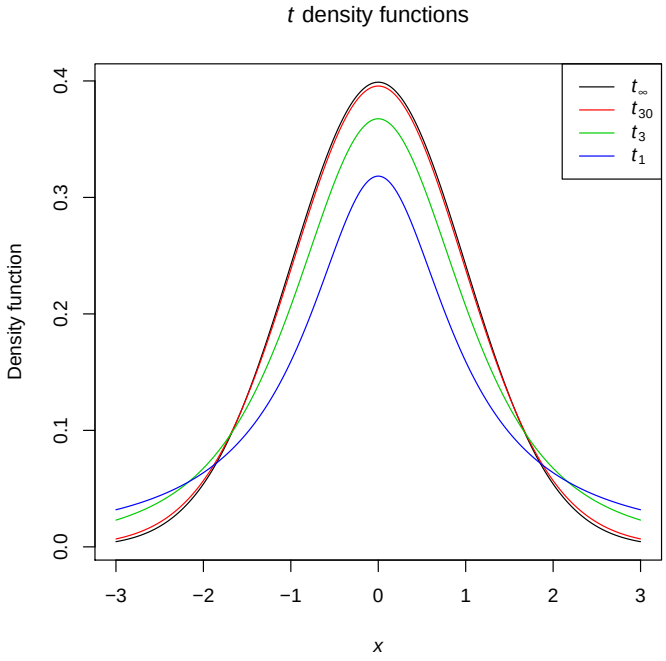


Figure 7.1: Student's t distribution – various degrees of freedom.

7

Tables,[6] for a given ν we can find $t_{0.025,\nu}$ such that:

$$P\left(-t_{0.025,\nu} < \frac{\hat{\theta}-\theta}{\text{E.S.E.}(\hat{\theta})} < t_{0.025,\nu}\right) = 0.95$$

where $t_{0.025,\nu}$ cuts off 2.5% probability in the upper tail of the t distribution with ν degrees of freedom. On rearranging the inequality within the brackets we get:

$$P\left(\hat{\theta} - t_{0.025,\nu}\cdot\text{E.S.E.}(\hat{\theta}) < \theta < \hat{\theta} + t_{0.025,\nu}\cdot\text{E.S.E.}(\hat{\theta})\right) = 0.95.$$

Endpoints for a 95% confidence interval (standard error unknown)

A 95% confidence interval for θ has **endpoints** $\hat{\theta} \pm t_{0.025,\nu}\cdot\text{E.S.E.}(\hat{\theta})$, leading to a reported confidence interval of the form:

$$(\hat{\theta} - t_{0.025,\nu}\cdot\text{E.S.E.}(\hat{\theta}),\ \hat{\theta} + t_{0.025,\nu}\cdot\text{E.S.E.}(\hat{\theta}))$$

where $t_{0.025,\nu}$ is the t-value which cuts off 2.5% probability in the upper tail of the t distribution with ν degrees of freedom, obtained from either Table 9 or Table 10 in the *New Cambridge Statistical Tables*.

This general result can be used in a large number of situations to compute confidence intervals.

In order to use these results we need to know how to calculate or estimate, respectively, the standard errors for the various cases (using the sample data). We now proceed to demonstrate these techniques for some important scenarios.

[6] Make sure you are comfortable using this table.

Important note: In the following applications, whenever the t distribution is used and we have *large* sample size(s) (hence large degrees of freedom), it is acceptable to use standard normal values as approximations due to the tendency of the t distribution to the standard normal as the degrees of freedom approach infinity. What constitutes a 'large' sample size is rather subjective; however, as a rule-of-thumb treat anything over 30 as 'large' in ST104a Statistics 1.

7.8 Confidence interval for a single mean (σ known)

Given observed sample values $x_1, x_2, \ldots, x_n$, the point estimate for μ is $\bar{x} = \sum_{i=1}^{n} x_i/n$.

Assuming the (population) variance, σ^2, is known, the standard error of $\bar{x}$ is $\sigma/\sqrt{n}$ (*exactly* if the population is normal; *approximately*, for large enough n, when sampling from non-normal populations by the central limit theorem).

Confidence interval endpoints for a single mean (σ known)

In such instances, we use the standard normal distribution when constructing a confidence interval with endpoints:

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \qquad \Longrightarrow \qquad \left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

where $z_{\alpha/2}$ is the z-value which cuts off $\alpha/2$ probability in the upper tail of the standard normal distribution to ensure a $100(1-\alpha)\%$ confidence interval. For example, for $\alpha = 0.05$, we have a $100(1-0.05)\% = 95\%$ confidence interval, and we require the z-value which cuts off $\alpha/2 = 0.025$, i.e. 2.5% probability in the upper tail of the standard normal distribution, that is 1.96, which can be obtained from Table 5 in the *New Cambridge Statistical Tables*.

Example 7.1 Measurements of the diameter of a random sample of 200 ball bearings produced by a machine gave a sample mean of $\bar{x} = 0.824$. The population standard deviation, σ, is 0.042. Find a 95% confidence interval and a 99% confidence interval for the true mean value of the diameter of the ball bearings.

We are told that $\sigma = 0.042$. So a 95% confidence interval, where $\alpha = 0.05$, has endpoints:

$$\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}} = 0.824 \pm 1.96 \times \frac{0.042}{\sqrt{200}} = 0.824 \pm 0.006$$

where $z_{0.025} = 1.96$ is the z-value which cuts off $\alpha/2 = 0.025$ probability in the upper tail of the standard normal distribution. In other words, the interval is $(0.818, \ 0.830)$ which **covers** the true mean with a probability of 95%.

To compute a 99% confidence interval (where $\alpha = 0.01$), since σ is known we require $z_{\alpha/2} = z_{0.005}$, that is the z-value which cuts off 0.5% probability in the upper tail of the standard normal distribution. Using Table 5 of the *New Cambridge Statistical*

Tables, the z-value is 2.5758. So a 99% confidence interval has endpoints:

$$\bar{x} \pm 2.576 \times \frac{\sigma}{\sqrt{n}} = 0.824 \pm 2.5758 \times \frac{0.042}{\sqrt{200}} = 0.824 \pm 0.008.$$

In other words, the interval is (0.816, 0.832). Note the higher level of confidence has resulted in a wider confidence interval.

7.9 Confidence interval for a single mean (σ unknown)

In practice it is unusual for σ^2 to be known – why would we know σ^2, but not μ? However, we can estimate σ^2 with the sample variance s^2,[7] using the estimator:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

This is because S^2 is an unbiased estimator[8] of σ^2. The degrees of freedom associated with this estimator are $n-1$, where n is the sample size.

Confidence interval endpoints for a single mean (σ unknown)

In such instances, we use the t distribution when constructing a $100(1-\alpha)\%$ confidence interval with endpoints:

$$\bar{x} \pm t_{\alpha/2,\,n-1} \cdot \frac{s}{\sqrt{n}} \qquad \Longrightarrow \qquad \left(\bar{x} - t_{\alpha/2,\,n-1} \cdot \frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2,\,n-1} \cdot \frac{s}{\sqrt{n}}\right)$$

where $t_{\alpha/2,\,n-1}$ is the t-value which cuts off $100(\alpha/2)\%$ probability in the upper tail of the t distribution with $n-1$ degrees of freedom, obtained from either Table 9 or Table 10 in the *New Cambridge Statistical Tables*.

Example 7.2

A company producing designer label jeans carries out a sampling exercise in order to estimate the average price that retailers are charging for the jeans. A random sample of 12 retailers gives the following sample of prices (in £):

34.50, 36.00, 43.50, 29.50, 44.50, 47.50, 45.50, 53.00, 40.00, 38.00, 33.00, 54.50.

We seek a 95% confidence interval for the mean retailer's price of the jeans.

Clearly, $n = 12$ and it is not hard to check $\bar{x} = 41.625$ and $s = 7.84$.

So the estimated standard error of the sample mean is:

$$\frac{s}{\sqrt{n}} = \frac{7.84}{\sqrt{12}} = 2.2632 \qquad \text{on } n-1 = 11 \text{ degrees of freedom.}$$

[7] An example of an unknown population parameter being estimated by its sample counterpart.

[8] Recall an unbiased estimator means the point estimate is correct *on average*, over repeated samples. This concept is covered in greater depth in **ST104b Statistics 2.**

Hence a 95% confidence interval for μ has endpoints $41.625 \pm 2.201 \times 2.2632$, i.e. the confidence interval is:

(£36.64, £46.61).

Two important points:

- Make sure you see where 2.201 comes from in Table 10 – it is $t_{0.025,\,11}$, i.e. the t-value above which lies 2.5% probability for a Student's t distribution with 11 degrees of freedom.
- Make sure you report confidence intervals in the form (£36.64, £46.61); that is you **must** compute the actual endpoints and report these as an *interval*, as that is what a confidence *interval* is! Note the lower endpoint should be given first.

7.10 Confidence interval for a single proportion

We often want to estimate a (population) proportion, for example in surveys of people's social, political or consumer attitudes.[9] Take an opinion poll in which we investigate voters' support for a particular political party. We can ask the question 'Do you support party X?', to which the n respondents each answer either 'yes' or 'no'. If r people reply 'yes', then the sample proportion who support the party is r/n. This is our point estimate of the true proportion of support in the population, π.

Sampling distribution of the sample proportion estimator, P

The sampling distribution of the sample proportion, P, (estimator of the population proportion, π)[10] is:

$$P \sim N\left(\pi,\, \frac{\pi(1-\pi)}{n}\right) \qquad \text{(approximately)}. \tag{7.1}$$

Hence $\text{Var}(P) = \pi(1-\pi)/n$, so the standard error is $\text{S.E.}(P) = \sqrt{\pi(1-\pi)/n}$. Unfortunately, this depends on π, precisely what we are trying to estimate, hence the true standard error is unknown, so must itself be estimated. As π is unknown, the best we can do is replace it with our point estimate for it, $p = r/n$, hence the estimated standard error is:

$$\text{E.S.E.}(p) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{r/n\,(1-r/n)}{n}}.$$

[9] In Chapter 4, categorical variables were introduced. Political party affiliation is an example of a categorical variable.

[10] This result is a consequence of the central limit theorem applied to the proportion of successes for a 'binomial' distribution. Full details can be found in **ST104b Statistics 2**, although such details are beyond the scope of this course.

Confidence interval endpoints for a single proportion

A $100(1-\alpha)\%$ confidence interval for a single proportion has endpoints:

$$p \pm z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \qquad \Longrightarrow \qquad \left(p - z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}},\, p + z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \right)$$

where $z_{\alpha/2}$ is the z-value which cuts off $100(\alpha/2)\%$ probability in the upper tail of the standard normal distribution, obtained from Table 5 in the *New Cambridge Statistical Tables*.

Note that although we are estimating a variance, **for proportions we do not use Student's t distribution** for two reasons:

- The standard error has not been estimated by a corrected sum of squares calculation, unlike $S^2 = S_{xx}/(n-1)$.
- The sample size n has to be large for the normal approximation to hold, and so the standard normal distribution is appropriate in this case.

7

Example 7.3 A survey is conducted by a bank to estimate the proportion of its customers who would be interested in using a proposed new mobile telephone banking service. If we denote the population proportion of customers who are interested in this proposal by π, and it is found that 68 out of 150 customers are in favour, then we would estimate π by $p = 68/150 = 0.453$. Hence a 95% confidence interval for π has endpoints:

$$0.453 \pm 1.96 \times \sqrt{\frac{0.453(1-0.453)}{150}} \qquad \Longrightarrow \qquad (0.37,\, 0.53).$$

Note, for a 95% confidence interval, $\alpha = 0.05$ and so we use $z_{0.025} = 1.96$ in the computation of the confidence interval.

7.11 Sample size determination

The question 'How big a sample do I need to take?' is a common one when sampling. The answer to this depends on the **quality of inference** that the researcher requires from the data. In the estimation context, this can be expressed in terms of the accuracy of estimation.

If the researcher requires that there should be a 95% chance that the estimation error should be no larger than e units (we refer to e as the **tolerance on the sampling error**), then this is equivalent to having a 95% confidence interval of width $2e$. Note here e represents the **half-width** of the confidence interval since the point estimate is, by construction, at the centre of the confidence interval.

For known variance, a 95% confidence interval is based on the error of estimation

being no greater than 1.96 standard errors. Hence we can conveniently use the relation:

$$1.96 \times \text{S.E.}(\hat{\theta}) \leq e$$

and solve this simple equation for the required sample size (note the standard error is a function of n).

Sample size determination for a single mean

To estimate μ to within e units with $100(1-\alpha)\%$ confidence, we require a sample of size:

$$n \geq \frac{{z_{\alpha/2}}^2 \sigma^2}{e^2} \tag{7.2}$$

where $z_{\alpha/2}$ is the z-value which cuts off $100(\alpha/2)\%$ probability in the upper tail of the standard normal distribution, obtained from Table 5 in the *New Cambridge Statistical Tables*.

Sample size determination for a single proportion

To estimate π to within e units with $100(1-\alpha)\%$ confidence, we require a sample of size:

$$n \geq \frac{{z_{\alpha/2}}^2 p(1-p)}{e^2} \tag{7.3}$$

where $z_{\alpha/2}$ is the z-value which cuts off $100(\alpha/2)\%$ probability in the upper tail of the standard normal distribution, obtained from Table 5 in the *New Cambridge Statistical Tables*.

Example 7.4 A simple random sample of 50 households is taken from a population of 1,000 households in an area of a town. The sample mean and standard deviation of weekly expenditure on alcoholic beverages are £18 and £4, respectively. How many *more* observations are needed to estimate μ to within 1 unit with 99% confidence?

Here, $e = 1$, and we can assume $\sigma = 4$ since the initial sample with $n = 50$ is 'large' and hence we would expect $s \approx \sigma$. For 99% confidence, we use $z_{0.005} = 2.5758$. Hence, using (7.2), we have:

$$n \geq \frac{(2.5758)^2 \times 4^2}{1^2} = 106.16.$$

Remembering that n must be an integer, the smallest n satisfying this is 107.[11] So 57 *more* observations are needed.

Example 7.5

The reaction time of a patient to a certain stimulus is known to have a standard deviation of 0.05 seconds. How large a sample of measurements must a psychologist take in order to be 95% confident and 99% confident, respectively, that the error in the estimate of the mean reaction time will not exceed 0.01 seconds?

[11] Note that we round *up*, otherwise had we rounded down it would lead to less precision.

For 95% confidence, we use $z_{0.025} = 1.96$ using Table 5 of the *New Cambridge Statistical Tables*. So, using (7.2), n is to be chosen such that:

$$n \geq \frac{(1.96)^2(0.05)^2}{(0.01)^2}.$$

Hence, we find that $n \geq 96.04$. Since n must be an integer, 97 observations are required to achieve an error of 0.01 or less with 95% probability.

For 99% confidence, we use $z_{0.005} = 2.5758$. So, using (7.2), n is to be chosen such that:

$$n \geq \frac{(2.5758)^2(0.05)^2}{(0.01)^2}.$$

Hence, we find that $n \geq 165.87$. Since n must be an integer, 166 observations are required to achieve an error of 0.01 or less with 99% probability.

Note that a higher level of confidence requires a larger sample size as more information (sample data) is needed to achieve a higher level of confidence for a given tolerance, e.

Example 7.6 A pilot study estimates a proportion to be 0.4. If we wish to be 90% confident of estimating the true population proportion with an error no greater than 0.03, how large a sample is needed?

Here $e = 0.03$, and we have an initial estimate of $p = 0.4$. For 90% confidence, we use $z_{0.05} = 1.645$ using Table 5 of the *New Cambridge Statistical Tables*. Hence, using (7.3), we have:

$$n \geq \frac{(1.645)^2 0.4(1-0.4)}{(0.03)^2} = 721.61.$$

So, rounding up, we require a sample size of 722.

7.12 Difference between two population proportions

The correct approach to the comparison of two population proportions, π_1 and π_2, is, obviously, via the *difference* between the population proportions. The sample proportions P_1 and P_2 are, for large sample sizes n_1 and n_2, respectively, (approximately) normally distributed as, from (7.1), we have:

$$P_1 \sim N\left(\pi_1, \frac{\pi_1(1-\pi_1)}{n_1}\right) \qquad \text{and} \qquad P_2 \sim N\left(\pi_2, \frac{\pi_2(1-\pi_2)}{n_2}\right).$$

When random samples are drawn from two independent populations, then these distributions are statistically independent, and it can be shown[12] that their difference is also normally distributed such that:

$$P_1 - P_2 \sim N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right). \tag{7.4}$$

[12] This is code for 'beyond the scope of the course'!

Clearly, $\text{Var}(P_1 - P_2)$, and hence $\text{S.E.}(P_1 - P_2) = \sqrt{\text{Var}(P_1 - P_2)}$, depends on the *unknown* parameters π_1 and π_2. So we must resort to the estimated standard error:

$$\text{E.S.E.}(P_1 - P_2) = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}.$$

Confidence interval endpoints for the difference between two proportions

With point estimates for π_1 and π_2 of $p_1 = r_1/n_1$ and $p_2 = r_2/n_2$, respectively, a $100(1-\alpha)\%$ confidence interval for the difference between two proportions has endpoints:

$$(p_1 - p_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \tag{7.5}$$

where $z_{\alpha/2}$ cuts off $100(\alpha/2)\%$ probability in the upper tail of the standard normal distribution, obtained from Table 5 in the *New Cambridge Statistical Tables*.

Example 7.7 We use (7.5) to calculate 95% and 90% confidence intervals for the difference between the population proportions of the general public who are aware of a particular commercial product before and after an advertising campaign. Two surveys were conducted and the results of the two random samples were:

	Sample size	Number aware
Before campaign	150	68
After campaign	120	65

To avoid working with negative differences whenever possible (to keep notation simpler by working with positive values), let $p_1 = r_1/n_1 = 65/120 = 0.5417$ and $p_2 = r_2/n_2 = 68/150 = 0.4533$. Hence our point estimate for the difference is:

$$\text{'After'} - \text{'Before'} = p_1 - p_2 = 0.5417 - 0.4533 = 0.0884.$$

So a 95% confidence interval for the difference in population proportions, $\pi_1 - \pi_2$, has endpoints:

$$0.0884 \pm 1.96 \times \sqrt{\frac{0.4533(1-0.4533)}{150} + \frac{0.5417(1-0.5417)}{120}}$$

which can be expressed as $(-0.031, 0.208)$.

A 90% confidence interval has endpoints:

$$0.0884 \pm 1.645 \times \sqrt{\frac{0.4533(1-0.4533)}{150} + \frac{0.5417(1-0.5417)}{120}}$$

which can be expressed as $(-0.012, 0.189)$.

Note that both confidence intervals *include zero* (since they both have a negative lower bound and positive upper bound). This suggests there is no significant difference in public awareness. This idea has close parallels with hypothesis testing, introduced in Chapter 8.

7.13 Difference between two population means

In this section we are primarily interested in the difference between two population means, i.e. $\mu_1 - \mu_2$. There are four cases to be considered, depending on whether variances are known or unknown, in which case we can assume they are either equal or unequal, and the case of paired datasets.

7.13.1 Unpaired samples – variances known

Suppose we have random samples of size n_1 and n_2 from two normal populations, $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. The sampling distributions of $\bar{X}_1$ and $\bar{X}_2$ are hence:

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{and} \quad \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

A natural estimator for $\mu_1 - \mu_2$ is $\bar{X}_1 - \bar{X}_2$. Random samples drawn from two independent populations are statistically independent, hence $\bar{X}_1$ and $\bar{X}_2$ are independent. The sampling distribution of their difference is:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right). \tag{7.6}$$

Confidence interval endpoints for the difference between two means

If the population variances σ_1^2 and σ_2^2 are known, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ has endpoints:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{7.7}$$

where $z_{\alpha/2}$ cuts off $100(\alpha/2)\%$ probability in the upper tail of the standard normal distribution, obtained from Table 5 in the *New Cambridge Statistical Tables*.

7.13.2 Unpaired samples – variances unknown and unequal

We have the same set-up as above, with the same sampling distribution for $\bar{X}_1 - \bar{X}_2$ in (7.6), but now the population variances σ_1^2 and σ_2^2 are unknown. Assuming *large* sample sizes, say greater than 30, we can replace these unknown parameters with the respective sample variances s_1^2 and s_2^2 and continue to use standard normal values. The justification is that since the sample sizes are large, we would expect accurate estimates of the population variances, such that $s_1^2 \approx \sigma_1^2$ and $s_2^2 \approx \sigma_2^2$.

Confidence interval endpoints for the difference between two means

If the population variances σ_1^2 and σ_2^2 are unknown, provided sample sizes n_1 and n_2 are large (greater than 30), a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ has endpoints:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $z_{\alpha/2}$ cuts off $100(\alpha/2)\%$ probability in the upper tail of the standard normal distribution, obtained from Table 5 in the *New Cambridge Statistical Tables*.

7.13.3 Unpaired samples – variances unknown and equal

In some circumstances we may be able to justify the assumption that the two populations being sampled are of equal variability. In which case suppose we have random samples of size n_1 and n_2 from two normal populations, $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, i.e. the populations have a **common variance**, σ^2, which is unknown.[13] The sampling distributions of $\bar{X}_1$ and $\bar{X}_2$ are hence:

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right) \quad \text{and} \quad \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right).$$

A natural estimator for $\mu_1 - \mu_2$ is still $\bar{X}_1 - \bar{X}_2$. Since we have unpaired (independent) samples, $\bar{X}_1$ and $\bar{X}_2$ are statistically independent, hence the sampling distribution of their difference is:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).$$

The problem is that this common variance σ^2 is unknown, so needs to be estimated. Should we use S_1^2 or S_2^2 as an estimator for σ^2? Answer: use both, by **pooling** the two sample variances, since both contain useful information about σ^2.

Pooled variance estimator

The pooled variance estimator, where S_1^2 and S_2^2 are sample variances from samples of size n_1 and n_2, respectively, is:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{7.8}$$

on $n_1 + n_2 - 2$ degrees of freedom, where the subscript 'p' denotes 'pooled'.

Hence S_p^2 is the weighted average of sample variances S_1^2 and S_2^2, where the weights are $(n_1 - 1)/(n_1 + n_2 - 2)$ and $(n_2 - 1)/(n_1 + n_2 - 2)$, respectively.

So if $n_1 = n_2$, then we give the sample variances equal weight. Intuitively this should make sense. As the sample size increases, a sample variance provides a more accurate

[13] If the population variances were known, then we would know their true values. Hence no assumptions would be necessary and we could use (7.7).

estimate of σ^2. Hence if $n_1 \neq n_2$, the sample variance calculated from the larger sample is more reliable, so is given greater weight in the pooled variance estimator. Of course, if $n_1 = n_2$, then the variances are equally reliable, hence are given equal weight.

Confidence interval endpoints for the difference between two means

If the population variances σ_1^2 and σ_2^2 are unknown but assumed equal, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ has endpoints:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,\, n_1+n_2-2} \cdot \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{7.9}$$

where s_p^2 is the estimate from the pooled variance estimator (7.8), and where $t_{\alpha/2,\, n_1+n_2-2}$ is the t-value which cuts off $100(\alpha/2)\%$ probability in the upper tail of the Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom, obtained from either Table 9 or Table 10 in the *New Cambridge Statistical Tables*.[14]

An obvious problem is how to decide whether to assume the unknown variances are equal or unequal. Well, firstly it would be wise to check the question to see if it explicitly states one way or the other. If it is not clear, then consider the following points:

7

- If $\sigma_1^2 = \sigma_2^2$, then we would expect approximately equal sample variances, i.e. $s_1^2 \approx s_2^2$, since both sample variances would be estimating the same (common) variance. If the sample variances are very different, then this would suggest $\sigma_1^2 \neq \sigma_2^2$.[15]
- If we are sampling from two 'similar' populations (companies in the same industry, for example) then an assumption of equal variability in these 'similar' populations would be reasonable.

Example 7.8 Two companies supplying a similar service are compared for their reaction times (in days) to complaints. Random samples of recent complaints to these companies gave the following:

	Sample size	Sample mean	Sample std. dev.
Company A	12	8.5	3.6
Company B	10	4.8	2.1

We compute a 95% confidence interval for the true difference in reaction times and use this interval to decide if one company is faster than the other.

Because the markets are 'similar', it is reasonable to assume (though it is only an assumption!) that the two population variances are equal. Under this assumption, using (7.8), we have:

$$s_p^2 = \frac{(12-1) \times (3.6)^2 + (10-1) \times (2.1)^2}{12 + 10 - 2} = 9.1125$$

[14] Recall for sufficiently 'large' degrees of freedom, a standard normal approximation can be used, provided a justification is given.

[15] A formal hypothesis test of H_0: $\sigma_1^2 = \sigma_2^2$ could be performed. We will not consider that in this course, although it is covered in **ST104b Statistics 2.**

on $n_1 + n_2 - 2 = 12 + 10 - 2 = 20$ degrees of freedom. So the estimated standard error of the difference in sample means is:

$$\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{9.1125 \left(\frac{1}{12} + \frac{1}{10} \right)} = \sqrt{9.1125 \times 0.1833} = 1.2925.$$

Hence a 95% confidence interval for $\mu_A - \mu_B$, using (7.9), is:

$$(8.5 - 4.8) \pm 2.086 \times 1.2925 \quad \Longrightarrow \quad (1.01,\ 6.39)$$

where $t_{0.025,\,20} = 2.086$ (we have estimated the common variance, so use the Student's t distribution, here with 20 degrees of freedom) using Table 10 of the *New Cambridge Statistical Tables*.

As zero does *not* lie in this interval, we conclude that the data suggests that Company B is faster than Company A in reacting to complaints. (Note the variable is reaction time, so a *lower* sample mean suggests a *faster* true average reaction time.)

7.13.4 Paired (dependent) samples

Paired-sample methods are used in special cases when the two samples are not statistically independent. For our purposes, such paired data are likely to involve observations on the *same* individuals in two different states – specifically 'before' and 'after' some intervening event. A paired-sample experimental design is advantageous since it allows researchers to determine whether or not significant changes have occurred as a result of the intervening event free of bias from other factors, since these have been controlled for by observing the same individuals.[16]

A necessary, but not sufficient, indicator for the presence of paired sample data is that $n_1 = n_2$, in order to have 'pairs' of data values. Common sense needs to be exercised to determine whether or not we have paired data. An example of such a dataset would be observations of the *same* individuals at two different points in time, typically 'before' and 'after' some event, such as statistical 'IQ' before and after taking **ST104a Statistics 1**.[17]

This scenario is easy to analyse as the paired data can simply be reduced to a 'one sample' analysis by working with **differenced data**. That is, suppose two samples generated sample values $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$, respectively (note the same number of observations, n, in each sample). Compute the differences, that is:

$$d_1 = x_1 - y_1, \quad d_2 = x_2 - y_2, \quad \ldots, \quad d_n = x_n - y_n.$$

It follows that $\bar{x}_d = \bar{x} - \bar{y}$, so that by using the differences to compute a confidence interval for μ_d, then we get the required confidence interval for $\mu_X - \mu_Y$. The technique therefore follows that in Section 7.9.

[16] Experimental studies are discussed in Chapter 11.

[17] One hopes, of course, that this would result in a mean increase, or at the very least no decrease, in statistical 'IQ'!

Example 7.9

The table below shows the before and after weights (in pounds) of 8 adults after a diet. Provide a 95% confidence interval for the loss of weight due to the diet. Are you convinced that the diet reduces weight?

Before	After	Before	After
127	122	150	144
130	120	147	138
114	116	167	155
139	132	153	152

The differences (calculated as 'Before − After') are:

$$5 \quad 10 \quad -2 \quad 7 \quad 6 \quad 9 \quad 12 \quad 1.$$

Hence $n = 8$, $\bar{x}_d = 6$ and $s_d = 4.66$ on $n - 1 = 7$ degrees of freedom. The estimated standard error of the sample mean is $s/\sqrt{n} = 4.66/\sqrt{8} = 1.65$. Using the t distribution on 7 degrees of freedom, for 95% confidence we use $t_{0.025,\,7} = 2.365$.

So a 95% confidence interval for the mean difference in weight before and after the diet is $\bar{x}_d \pm t_{0.025,\,n-1} \cdot s/\sqrt{n}$, that is:

$$6 \pm 2.365 \times 1.65 \qquad \Longrightarrow \qquad (2.1,\ 9.9).$$

Since zero is not included in this confidence interval, we conclude that the diet *does* appear to reduce weight, i.e. the average weight loss appears to be positive.

7

7.14 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 standard normal distribution.

Remember:

- We use the standard 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 the variance or standard deviation from a sample, we will need to use Student's t, whatever the size of the sample.
- If the sample size is large, then the standard normal distribution approximates Student's t.
- We use the standard normal distribution whenever we are dealing with proportions.

7.15 Key terms and concepts

- Confidence interval
- Degrees of freedom
- Endpoints
- Paired sample
- Standard error
- Tolerance
- Coverage probability
- Differences
- Half-width
- Pooled variance
- Student's t

7.16 Learning activities

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

2. Look again at Example 7.1, but this time read the description 'Measurements of the diameter of a random sample of 200 ball bearings produced by a machine gave a sample mean $\bar{x} = 0.824$. The sample standard deviation was $s = 0.042$. Find a 95% confidence interval and a 99% confidence interval for the true mean value of the diameter of ball bearings.'

   (Note: Although this time you have been told that you only have a sample estimate for the standard deviation, you can justify using the standard 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 Example 7.1.)

3. Open your *New Cambridge Statistical Tables* at Table 10. Note that different probability tails are given for $\nu = 1,\ 2$ etc. (ν is the same as $n - 1$ in this case). Now consider a 95% confidence interval for μ when $n = 21$ (i.e. for 20 degrees of freedom). You can see the t value is $t_{0.025,\,20} = 2.086$. However, when ν is very large, the t value is 1.96 – exactly the same as for the standard normal distribution. Although you can see that t values are given for quite large degrees of freedom, we generally assume that the standard normal distribution can be used instead of Student's t if the degrees of freedom are greater than 30 (some textbooks say 40, others say 50).

4. A random sample of 200 students is observed. 30 of them say they are 'really enjoying' Statistics.

   (a) Calculate the proportion of students in this sample saying they are 'really enjoying' Statistics and then provide a 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 are 'really enjoying' Statistics.

   (b) Provide a 95% confidence interval for this value. Think about why the two confidence intervals are different.

   (c) Provide a 95% confidence interval for the difference between the two proportions.

5. A business requires an expensive check on the value of stock in its warehouse. In order to do this, a random sample of 50 items is observed and valued. The average value of these is computed to be £320.41 with a (sample) standard deviation of £40.60. It is known that there are 9,875 items in total stock.
   (a) Estimate the total value of the stock to the nearest £10,000.
   (b) Calculate a 95% confidence interval for the mean value of all items and hence determine a 95% confidence interval for the total value of the stock.

6. The reaction times, in seconds, for eight police officers were found to be:

   0.28 0.23 0.21 0.26 0.29 0.21 0.25 0.22.

   Determine a 90% confidence interval for the mean reaction time of all police officers.

7. A random sample of 100 voters contained 60 Labour supporters. Give a 95% confidence interval for the proportion of Labour voters in the population.

8. A sample of 954 adults in early 1987 found that 23% of them held shares.
   (a) Given a UK adult population of 41 million, and assuming a proper random sample was taken, find a 95% confidence interval for the number of shareholders in the UK.

   A 'similar' survey the previous year had found a total of 7 million shareholders.
   (b) Assuming 'similar' means the same sample size, find a 95% confidence interval for the increase in shareholders between the two years.

9. Two advertising companies each give quotations for nine different campaigns. Their quotations (in £000s) are shown in the following table. Calculate a 95% confidence interval for the true difference between mean quotations. Can you deduce from this confidence interval if one company is more expensive than the other?

Company	1	2	3	4	5	6	7	8	9
A	39	24	36	42	45	30	38	32	39
B	46	26	32	39	51	34	37	41	44

10. A college contains 12,000 students. A random sample of 400 students are interviewed and it is found that 240 use the refectory. Use these data to calculate:
    (a) a 95% confidence interval
    (b) a 99% confidence interval

    for the total number of students who use the refectory.

    The college catering officer claims that the refectory is used by at least 9,000 students and that the survey has yielded a low figure due to sampling variability. Is this claim reasonable?

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

In addition, attempt the 'Test your understanding' self-test quizzes available on the VLE.

7.17 Further exercises

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060]. Relevant exercises from Sections 7.1–7.3 and 8.1–8.4.

7.18 A reminder of your learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- calculate sample means, standard deviations and proportions, and understand their use as estimators
- construct a confidence interval for a sample mean, a sample proportion, the difference between two sample means and the difference between two sample proportions
- know when to use Student's t distribution.

You do **not** need to:

- demonstrate the central limit theorem
- know about the concepts relating to a 'good' estimator.

7.19 Sample examination questions

1. Would you say the following statement is **true** or **false**? Give brief reasons. 'When calculated from the same dataset, a 91% confidence interval is wider than a 96% confidence interval.'

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.

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 π. Give the formula for pooled proportions when comparing the two samples. (Note that you need to check this on your examination formula sheet for **ST104a Statistics 1**.)

   (b) An independent assessment is made of the services provided by two holiday companies. Of a random 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 between the proportions of satisfied customers between the two companies. Basing your conclusion on this confidence interval, do you believe that one company gives more satisfaction than the other?

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

Chapter 8
Hypothesis testing

8.1 Aims

In Chapters 6 and 7 you were introduced to the idea of the probability that a parameter could lie within a range of values, and in particular the confidence interval (generally 90%, 95% or 99% confidence) 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 concept of testing a hypothesis. The arithmetic and underlying ideas you need are similar to those you met in the last chapter.

You should aim to consolidate your familiarity with ideas of randomness and work on the different aspects of hypothesis testing as presented here.

8.2 Learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- set up the null and alternative hypotheses for a problem and state whether the latter is one- or two-sided, hence leading to a one- or two-tailed test
- define and apply the terminology of statistical testing
- perform statistical tests on means and proportions
- construct and explain a simple chart showing the kinds of errors that can be made in hypothesis testing.

8.3 Essential reading

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060] Sections 9.1–9.4 and 10.1–10.3.

In addition there is essential 'watching' of this chapter's accompanying video tutorials accessible via the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

8.4 Further reading

- Aczel, A.D. *Complete Business Statistics*. (London: McGraw-Hill Higher Education, 2009) seventh edition [ISBN 9780071287531] Chapters 7, 8 and 14 (taking care to omit topics not in the learning outcomes).
- Anderson, D.R., D.J. Sweeney, T.A. Williams, J. Freeman and E. Shoesmith *Statistics for Business and Economics*. (South-Western Cengage Learning, 2010) eleventh edition [ISBN 9780324783247] Chapter 9 and Sections 10.2, 12.2 and 14.9.
- Lind, D.A., W.G. Marchal and S.A. Wathen *Statistical Techniques in Business and Economics*. (Boston: McGraw-Hill Higher Education, 2009) fourteenth edition [ISBN 978007110004] Chapters 9, 10, 11 and the latter half of 16.
- Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics for Business and Economics*. (Chichester: John Wiley & Sons, 1990) fourth edition [ISBN 9780471615170] Chapter 9.

8.5 Introduction

In this chapter we introduce a branch of statistical inference known as **hypothesis testing**. Typically we choose between two statements about the value of a parameter based on evidence obtained from sample data which, due to sample data being subject to randomness, can be regarded as realisations of some random variable of interest.

Our objective is to choose between these two conflicting statements about the **population**, where these statements are known as **hypotheses**. By convention these are denoted by H_0 and H_1. *Qualitative* examples of such hypotheses are:

Null hypothesis, H_0	Alternative hypothesis, H_1
A modified process **does not** produce a higher yield than the standard process.	A modified process **does** produce a higher yield than the standard process.
True economic output **has not** increased over a year.	True economic output **has** increased over a year.
A person **is not** gifted with Extra Sensory Perception.	A person **is** gifted with Extra Sensory Perception.
The average level of lead in the blood of people in a particular environment **is as high** as 3.0.	The average level of lead in the blood of people in a particular environment **is lower than** 3.0.

From these examples we see that we use H_0 to represent 'no difference', 'no improvement', 'no effect' etc., and this is known as the **null hypothesis**, while H_1 is the **alternative hypothesis**.

Many statistical procedures can be represented as statements about the values of population parameters such as the mean, μ, or variance, σ^2.[1] The first step in any hypothesis testing problem is to 'translate' the real problem into its technical analogue.

[1] There is also a branch of statistics which deals with so-called 'non-parametric tests', although we will not consider such tests in this course.

In **ST104a Statistics 1**, the hypotheses will only contain parameters. For example, the previous hypotheses can all be 'translated' into technical forms similar to:

'We observe $x_1, x_2, \ldots, x_n$ from $N(\mu, \sigma^2)$ and we wish to test:

$$\mathrm{H}_0 : \mu = \mu_0 \qquad \text{vs.} \qquad \mathrm{H}_1 : \mu > \mu_0$$

where μ_0 is some specified value such as 0 or 3.'

Performing the 'translation' correctly can itself be complicated, hence this requires careful thought. Specifically, it is the form of H_1 which needs consideration. The null hypothesis, H_0, will always denote the parameter value with equality (=),[2] such as:

$$\mathrm{H}_0 : \mu = \mu_0.$$

In contrast the alternative hypothesis, H_1, will take one of three forms, i.e. using $\neq$, $<$, or $>$, that is:

$$\mathrm{H}_1 : \mu \neq \mu_0 \qquad \text{or} \qquad \mathrm{H}_1 : \mu < \mu_0 \qquad \text{or} \qquad \mathrm{H}_1 : \mu > \mu_0.$$

Note that **only one** of these forms will be used per test. To determine which form to use will require careful consideration of the wording in the question.

The form $\mathrm{H}_1 : \mu \neq \mu_0$ is an example of a two-tailed (two-sided) test and we use this form with questions worded such as 'test the hypothesis that μ is zero'. Here, there is no indication of the value of μ if it is not zero – do we assume $\mu > 0$ or $\mu < 0$ in such cases? We cannot be sure, so we take the 'conservative' option of a two-sided test, i.e. $\mu \neq 0$.

In contrast, had the question been phrased as 'test whether or not μ is greater than zero', then unambiguously we would opt for $\mathrm{H}_1 : \mu > \mu_0$. This is an example of an upper-tailed (one-sided) test. Unsurprisingly, 'test whether or not μ is less than zero' leads to $\mathrm{H}_1 : \mu < \mu_0$ which is a lower-tailed (also one-sided) test.

Later, when testing for *differences* between two population means or proportions, you need to look out for **comparative phrases** indicating if one population value should exceed the other (for example, testing whether 'A' is taller/smarter/faster than 'B'). Practising problems will make you proficient in correctly specifying your hypotheses.

8.6 Statistical tests

Consider the following two situations:

1. Suppose that it has been suggested that a particular environment is such that the average lead content of blood is as high as 3 units. In order to assess the truth of this, 50 samples are analysed and it is found that the average of these is 2.8 and that the sample standard deviation is 0.20.[3] The standard error is, therefore, 0.028

[2] Such an H_0 is called a *simple* null hypothesis. It is possible to have a *composite* null hypothesis, e.g. $\mathrm{H}_0 : \mu \geq 0$ which allows for more than one parameter value, however, in **ST104a Statistics 1** we will only focus on simple forms of H_0.

[3] Even when the true mean of the blood lead level is 3, we do not expect the sample mean to be exactly 3. But we **do** expect the sample mean to be 'close' to 3, while being willing to tolerate a certain amount of deviation from that value, due to chance. The important question here is 'Is this deviation from 3 due to chance alone, or is it too large to be explained purely by chance?'. The deviation of 2.8 from 3 is far too large, given the standard error.

so that the sample mean is more than 7 standard errors below the hypothetical mean – highly unlikely *if* the hypothesis is true! The data are extremely inconsistent with the hypothetical mean of 3, and so we reject this value on the evidence of the data.

2. A coin is tossed 100 times in order to decide whether or not it is a fair coin. We observe 98 heads and 2 tails. Common sense tells us that this is strong evidence that the coin is biased towards heads. More formally, we reject the hypothesis that the coin is fair as the chance of obtaining 98 heads with a fair coin is extremely low; the observation, 98 heads, is inconsistent with the coin being fair (since we would have *expected* (approximately) 50 heads).

In these examples it would be perfectly reasonable to make common sense judgements since the data were so extreme, but these 'common sense' judgements are using exactly the same underlying logic as do the more sophisticated tests involving the formal calculation of a test statistic. By fully understanding this logic, we are able to make optimal decisions in less extreme and more complicated cases. A further advantage of the formal approach is that it enables us to express numerically the degree of confidence in our inference procedure; this can be extremely important in communicating our conclusions via reports, presentations, etc.

Choosing between competing hypotheses requires us to conduct a statistical test. The test procedures themselves involve straightforward calculations using the sample data. At this point it is worth emphasising the following: **always assume the null hypothesis, H_0, is true**. That is, H_0 is our 'working hypothesis' which we hold to be true until we obtain *significant* evidence against it.

A **test statistic** is the formal mechanism used to evaluate the support given to H_0 by sample data. Different test statistics are used for testing different sorts of hypotheses. For example, tests of a single mean, single proportion, differences in means or proportions all require different test statistics.

Since H_0 refers to a population parameter, we use a suitable estimator to estimate it based on the sample data. For example, $\bar{X}$ would be used when testing μ. Test statistics are constructed from the sampling distribution of the estimator (introduced in Chapter 6). Hence test statistics are themselves random variables, as they are functions of estimators (which are random variables). It follows, therefore, that test statistics also follow probability distributions.

Since we assume H_0 to be true, the test statistic's distribution is conditional on H_0. Using the given sample data, we evaluate the test statistic to obtain a **test statistic value** (analogous to obtaining a point estimate using an estimator). If we find this test statistic value is sufficiently extreme – as determined by the test statistic distribution – then we will have sufficiently strong evidence to justify rejecting H_0.

Of course, we need to specify what constitutes a 'sufficiently extreme' test statistic value. In fact we define a **critical region** which spans test statistic values which are considered sufficiently extreme. It follows that our **decision rule** is to reject H_0 if the test statistic value falls in this critical region. If it does not, then we fail to reject H_0, which is retained as our working hypothesis. For any test, the critical region is defined by one or two **critical values**, depending on whether we are performing a one-tailed or two-tailed test, respectively. These critical values are obtained from the test statistic's

distribution under H_0.

8.7 Types of error

In any hypothesis test there are two types of inferential decision error that could be committed. Clearly, we would like to reduce the probabilities of these errors as much as possible. These two types of error are called **Type I error** and **Type II error**.

Type I and Type II errors

- **Type I error**: Rejecting H_0 when it is true. This can be thought of as a 'false positive'. Denote the probability of this type of error by α.
- **Type II error**: Failing to reject H_0 when it is false. This can be thought of as a 'false negative'. Denote the probability of this type of error by β.

Both errors are undesirable and, depending on the context of the hypothesis test, it could be argued that either one is worse than the other.[4] However, on balance, a Type I error is usually considered to be more problematic. The possible **decision space** can be presented as:

	Decision	
	Not reject H_0	Reject H_0
H_0 true	Correct decision	Type I error
H_1 true	Type II error	Correct decision

For example, if H_0 was being 'innocent' and H_1 was being 'guilty', a Type I error would be finding an innocent person guilty (bad for him/her), while a Type II error would be finding a guilty person innocent (bad for the victim/society, but admittedly good for him/her!).

The complement of a Type II error, that is $1 - \beta$, is called the **power** of the test – the probability that the test will reject a false null hypothesis. You will not be expected to calculate either β or $(1 - \beta)$ in this course. However, you should understand the implications of the different types of error and be able to complete the chart in the second learning activity at the end of this chapter.

> **Example 8.1** In the example about lead levels discussed earlier, we would make a Type I error if we decided the lead level was less than 3 when it was actually as high as 3, and we would make a Type II error if we believed the level to be 3 when it was actually lower than 3. These errors are taken into consideration in the formulation of test procedures – the standard tests presented below are all devised to minimise the probabilities of making such errors.

[4] Which is better – a medical test incorrectly concludes a healthy person has a terminal illness, or incorrectly concludes that a terminally ill person is perfectly healthy?

8.8 Tests for normal populations

Let us suppose that, having obtained sample data from a normal distribution, we use estimator $\hat{\theta}$ to estimate an unknown parameter θ. Assume $\hat{\theta}$ has the sampling distribution:

$$\hat{\theta} \sim N(\theta, V)$$

where $V = \text{Var}(\hat{\theta})$ is assumed to be known. Hence the standard error, S.E.$(\hat{\theta})$, is also known. The hypothesis test will proceed using the following steps.

1. **Define the hypotheses.** In this case suppose we test:

   $$\text{H}_0 : \theta = \theta_0 \qquad \text{vs.} \qquad \text{H}_1 : \theta > \theta_0.$$

   Hence large values of $\hat{\theta}$ would imply H_1, while small values would tend to favour H_0. Note in this case the alternative hypothesis is one-sided and this has an effect on the format of the test, i.e. here we perform an **upper-tailed test**.

2. **State the test statistic and its distribution, then compute its value.** Regardless of the type of parameter we are testing, we choose a test statistic whose distribution, conditional on H_0, has a standard tabulated distribution such as the standard normal, Student's t, or chi-squared distributions (the chi-squared distribution is introduced in Chapter 9). In this case, due to the known standard error, standardisation of $\hat{\theta}$ readily yields:

   $$\frac{\hat{\theta} - \theta_0}{\text{S.E.}(\hat{\theta})} \sim N(0, 1).$$

   Next compute the test statistic value using the sample data, i.e. $\hat{\theta}$ will be our point estimate of θ.

3. **Define the critical region for a given significance level, α.**[5] Obtain critical value(s) from statistical tables to define the critical region. For this example, let $\alpha = 0.05$, where α is the Type I error probability. Since we are conducting an upper-tailed test, we require $z_\alpha = z_{0.05} = 1.645$, i.e. the value, for the standard normal distribution, above which lies $\alpha = 0.05 = 5\%$ probability.

4. **Decide whether or not to reject H_0.** Now we decide whether or not to reject H_0 (our working hypothesis which, up to now, is assumed to be true). If the test statistic value lies in the critical region then we will reject H_0, otherwise we do not reject H_0. A rejected H_0 means the test is **statistically significant** at the specific significance level used. In the examination you will be required to test at two significance level sequentially – more on this later.

5. **Draw conclusions.** It is always important to draw conclusions in the context of the variables of the original problem. It is not sufficient to conclude 'the test is significant at the 5% significance level'. This is just a technical step which guides us to make a better decision about the original *real-world problem*, and final conclusions should be drawn in terms of that problem.

[5] See the next section for a discussion of significance levels.

An important aspect of this test is the simple form of the test statistic, namely:

$$\frac{\text{Point estimator} - \text{Hypothesised value}}{\text{(Estimated) standard error}}.$$

Although this is not the format of all test statistics, it does cover a wide class of situations, listed next.

- Test of a single population mean, $H_0 : \mu = \mu_0$.
- Test of a single population proportion, $H_0 : \pi = \pi_0$.
- Test of the equality of two means, $H_0 : \mu_1 = \mu_2$.
- Test of the equality of two proportions, $H_0 : \pi_1 = \pi_2$.

8.9 Significance levels

Step 3 of the previous test procedure referred to a **significance level,** α. We control for the probability of committing a Type I error by setting the probability of this to be α. Hence the significance level, or **size**, of a test is just the probability of committing a Type I error. If we perform a test at the 5% significance level, say, then we are actually basing our decision on a procedure which gives us a 5% chance of making a Type I error (rejecting H_0 when H_0 is true).

In the development of the general normal test above, we used a significance level of 5%. This is one of the most commonly used significance levels, although 1% and 10% are also sometimes used. The value of quoting the significance level used is that it gives an acceptable method of expressing the user's confidence in the decision that was made. A test that is significant at the 1% level expresses a strong conviction that H_0 is untrue, whereas a test that is significant only at the 10% level expresses merely some doubt on H_0.

The significance levels – 10%, 5% and 1%, corresponding to α values of 0.10, 0.05 and 0.01, respectively – provide a means of deciding which values of the test statistic are extreme. Consider, for example, a one-sided, upper-tailed test of:

$$H_0 : \mu = 2 \qquad \text{vs.} \qquad H_1 : \mu > 2.$$

We calculate the test statistic value and if it falls deep enough into the right-hand tail (the critical region) of the test statistic's distribution then we reject H_0. Where do we draw the line which signals the start of this critical region? That is, what critical value should we use? An accepted convention is to use a critical value that 'cuts off' a tail of size α. So, if the test statistic follows a standard normal distribution, for an upper-tailed test, using Table 5 of the *New Cambridge Statistical Tables*, we denote this cut-off point as z_α which for:

$$\begin{aligned} \alpha = 0.10 &\implies z_{0.10} = 1.2816 \\ \alpha = 0.05 &\implies z_{0.05} = 1.6449 \\ \alpha = 0.01 &\implies z_{0.01} = 2.3263 \end{aligned}$$

such that if the test statistic value exceeds the critical value, then the 'test is significant at the $100\alpha\%$ significance level'.

Clearly, significance at the 1% level ($2.3263 <$ test statistic value) implies significance at the 5% level (since $1.6449 < 2.3263$) which, in turn, implies significance at the 10% level (since $1.2816 < 1.6449 < 2.3263$). When a test is significant we should *always* state the *smallest* level of significance as possible as this provides the best measure of the compatibility (or rather, *incompatibility*) of the data with the null hypothesis.

If the test statistic value falls between 1.2816 and 1.6449, then the test is significant at the 10% level, but not at the 5% level. This type of result arises from rather inconclusive data and so it would be better to sample more data in the hope that a firmer conclusion can be drawn one way or the other. Such an outcome is saying that the null hypothesis is to be treated with suspicion, but that we cannot confidently reject it. These points and ideas will be reinforced in the subsequent examples.

8.9.1 Order of conducting tests

The best practice for answering hypothesis test questions in the examination is to test at **two significance levels**. However, the **order** in which you apply the significance levels is taken into account by the Examiners. Noting the discussion above, a test which is significant at the 5% significance level, say, **must** also be significant at the 10% significance level, so you would be penalised if you used 10% as your second significance level having just rejected at the 5% significance level. In such a situation it would be more appropriate to test at the 1% significance level next.

Unless an examination question states otherwise, a sensible strategy to follow is to initially test at the 5% significance level and then test either at the 1% or 10% significance level, depending on whether or not you have rejected at the 5% significance level. A 'decision tree' depiction of the procedure to follow is presented in Figure 8.1.

As indicated in Figure 8.1, it is possible to state whether a test result is 'highly significant', 'moderately significant', 'weakly significant' or 'not significant', and therefore we can convey a measure of the 'strength' of any statistical significance.

8.10 One- and two-tailed tests

The tests that have been described so far have been **one-tailed** (specifically, upper-tailed). This is because the alternative hypotheses used were one-sided. Recall the discussion in Section 8.5 which identified three different forms of H_1. For example, if testing a population mean $H_0 : \mu = \mu_0$, these are:

$$\text{i. } H_1 : \mu \neq \mu_0 \qquad \text{or} \qquad \text{ii. } H_1 : \mu < \mu_0 \qquad \text{or} \qquad \text{iii. } H_1 : \mu > \mu_0.$$

If H_1 is of type (i.), then both tails of the test statistic's distribution will form the critical region, i.e. a **two-tailed test**. The critical region will still be of total area α, so it is logical to split this α evenly between the two tails. So if the test statistic follows a standard normal distribution, for a two-tailed test, using Table 5 of the *New Cambridge*

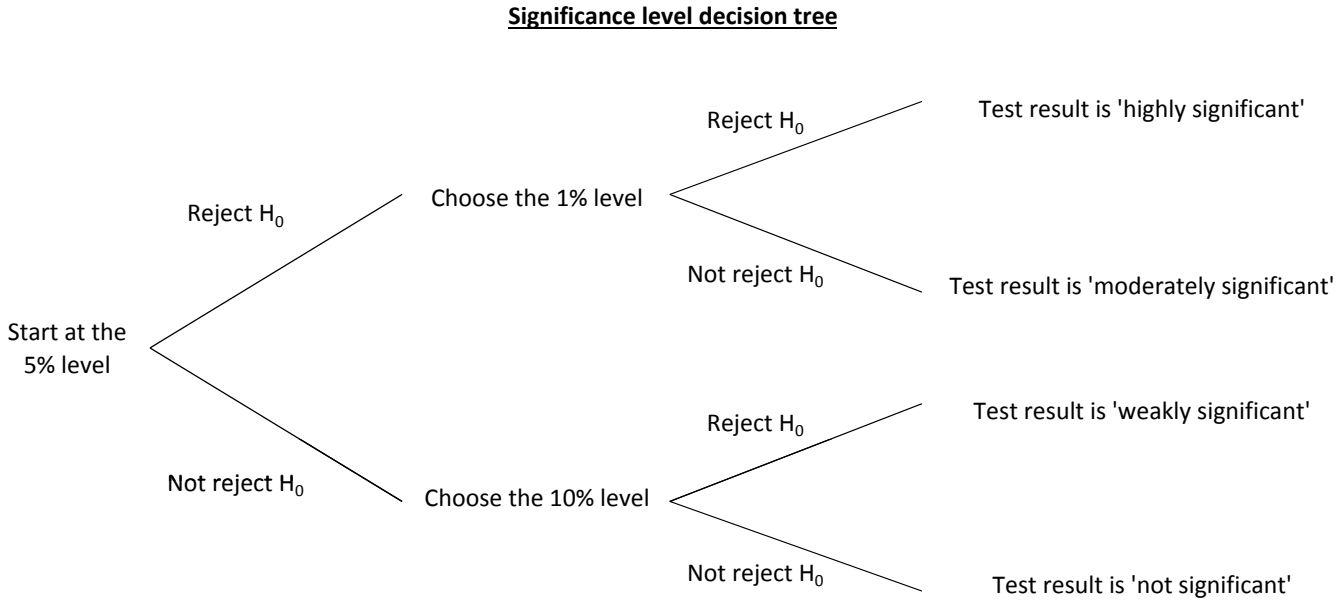


Figure 8.1: Significance level decision tree.

Statistical Tables and noting symmetry of the standard normal distribution about zero:

$$\begin{aligned}
\alpha = 0.10 &\implies z_{\alpha/2} = z_{0.05} = 1.6449, \text{ giving critical values of } \pm 1.6449 \\
\alpha = 0.05 &\implies z_{\alpha/2} = z_{0.025} = 1.9600, \text{ giving critical values of } \pm 1.9600 \\
\alpha = 0.01 &\implies z_{\alpha/2} = z_{0.005} = 2.5758, \text{ giving critical values of } \pm 2.5758
\end{aligned}$$

such that if the test statistic value exceeds either critical value (i.e. greater than the upper critical value, or less than the lower critical value), then the 'test is significant at the $100\alpha\%$ significance level'.

If H_1 is of type (ii.), then this is a **lower-tailed test**. Hence the critical region is in the left-hand tail of the test statistic's distribution. So if the test statistic follows a standard normal distribution, for a lower-tailed test, using Table 5 of the *New Cambridge Statistical Tables* and noting symmetry of the standard normal distribution about zero:

$$\begin{aligned}
\alpha = 0.10 &\implies z_{0.90} = -1.2816 \\
\alpha = 0.05 &\implies z_{0.95} = -1.6449 \\
\alpha = 0.01 &\implies z_{0.99} = -2.3263
\end{aligned}$$

such that if the test statistic value exceeds the critical value (i.e. falls below the critical value), then the 'test is significant at the $100\alpha\%$ significance level'.

If H_1 is of type (iii.), then we have an **upper-tailed test**, which has already been discussed.

8.11 *P*-values

Before proceeding to examples of hypothesis test situations, we will discuss an alternative approach to hypothesis testing based on the ***p*-value**.

So far we have spoken about defining a critical region and rejecting the null hypothesis, H_0, whenever the test statistic value falls in this critical region. An alternative approach is to compute the probability of obtaining the observed test statistic value *or a more extreme value* conditional on H_0 being true, using the probability distribution of the test statistic. This probability is the p-value.

So, given we reject H_0 if the test statistic value is sufficiently extreme, we would also reject H_0 if we obtained a sufficiently small p-value. To summarise, in the p-value world of testing, we use the following simple decision rule.

Hypothesis testing – p-value decision rule

- Reject H_0 if p-value $< \alpha$, where α is the significance level.
- Do not reject H_0 if p-value $\geq \alpha$.

It remains to show how to calculate the p-value. Again, this will depend on the form of H_1 and the test statistic distribution.

Calculating p-values

Let the test statistic value be x, and the test statistic (a random variable as it is constructed from an estimator) be X. Assuming a test statistic distribution which is *symmetric about zero* for the *two-tailed test*, in general these can be summarised as follows.

Form of H_1	p-value calculation
$H_1 : \theta \neq \theta_0$	$2 \times P(X \geq \lvert x \rvert)$
$H_1 : \theta < \theta_0$	$P(X \leq x)$
$H_1 : \theta > \theta_0$	$P(X \geq x)$

In each case use the *New Cambridge Statistical Tables* corresponding to the distribution of test statistic X.

Example 8.2 If $X \sim N(0, 1)$, and we obtain the test statistic value $x = -1.82$, then for a lower-tailed test the p-value would be:

$$P(X \leq -1.82) = 1 - \Phi(1.82) = 1 - 0.9656 = 0.0344$$

using Table 4 of the *New Cambridge Statistical Tables*. Hence this is significant at the 5% significance level ($0.0344 < 0.05$), but not at the 1% significance level ($0.0344 > 0.01$), indicating a moderately significant result.

The next few sections detail various hypothesis testing situations analogous to the confidence interval situations presented in Chapter 7.

8.12 Hypothesis test for a single mean (σ known)

We first consider the test of a single population mean when the population standard deviation, σ, is known. To test $\mathrm{H}_0 : \mu = \mu_0$ when sampling from $N(\mu, \sigma^2)$ we use the following test statistic.

z-test of hypothesis for a single mean (σ known)

In this case, the test statistic is:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1). \tag{8.1}$$

Hence critical values are obtained from the standard normal distribution, i.e. using Table 5 of the *New Cambridge Statistical Tables*.

Example 8.3 The mean lifetime of 100 components in a sample is 1,570 hours and their standard deviation is known to be 120 hours. μ 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 perform a two-tailed test since we are testing whether or not μ is 1,600.[6]

Hence we test:

$$\mathrm{H}_0 : \mu = 1{,}600 \qquad \text{vs.} \qquad \mathrm{H}_1 : \mu \neq 1{,}600.$$

Since σ is known, we use (8.1) to calculate the test statistic value, which is:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{1{,}570 - 1{,}600}{120/\sqrt{100}} = -2.5.$$

We first test at the 5% significance level. Since this is a two-tailed test, the critical values are ± 1.96. Since the test statistic value is in the critical region ($-2.5 < -1.96$), we reject H_0. If we now refer to the significance level decision tree (Figure 8.1), we now test at the 1% level with critical values of ± 2.5758. Since $-2.5758 < -2.5$ we are (just) unable to reject H_0 and conclude that the test result is 'moderately significant'. Hence at the 5% significance level there is evidence to suggest that the mean lifetime of components, μ, is not equal to 1,600.

If we wanted to compute the p-value, then from Section 8.11, for this two-tailed test the p-value would be:

$$p\text{-value} = 2 \times P(Z > |-2.5|) = 2 \times P(Z \geq 2.5) = 2 \times 0.0062 = 0.0124.$$

Since $0.01 < 0.0124 < 0.05$, we see that using the p-value testing method we still conclude (as we must) that the test is significant at the 5% significance level, but not at the 1% significance level.

8.13 Hypothesis test for a single mean (σ unknown)

Here the only difference with the previous example is that σ^2 is unknown. In which case this is estimated with the sample variance estimator S^2. To test $\mathrm{H}_0 : \mu = \mu_0$ when sampling from $N(\mu, \sigma^2)$ we use the following test statistic.

t test of hypothesis for a single mean (σ unknown)

In this case, the test statistic is:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}. \tag{8.2}$$

Hence critical values are obtained from the Student's t distribution with $n-1$ degrees of freedom, i.e. using Table 10 of the *New Cambridge Statistical Tables*.

Example 8.4 A study on the impact of comprehensive planning on financial performance reported that the average annual return on investment for American banks was 10.2%,[7] and suggested that banks who exercised comprehensive planning would do better than this. A random sample of 26 such banks gave the following returns on investment. Do these data support the conjecture?

10.00, 11.90, 9.90, 10.09, 10.31, 9.96, 10.34, 10.30, 10.50,
10.23, 10.72, 11.54, 10.81, 10.15, 9.04, 11.55, 10.81, 8.69,
10.74, 10.31, 10.76, 10.92, 11.26, 11.21, 10.20, 10.76.

We test:

$$\mathrm{H}_0 : \mu = 10.2 \qquad \text{vs.} \qquad \mathrm{H}_1 : \mu > 10.2.$$

Note that the alternative hypothesis is one-sided as this is the region of interest (we hypothesise that banks exercising comprehensive planning perform *better* than 10.2%). This implies that an upper-tailed test is needed.

The summary statistics are $n = 26$, $\bar{x} = 10.5$ and $s = 0.714$. Hence, using (8.2), the test statistic value is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{10.5 - 10.2}{0.714/\sqrt{26}} = 2.14.$$

We compare this test statistic value with the upper tail of the Student's t distribution with $26 - 1 = 25$ degrees of freedom.

If we choose $\alpha = 0.05$,[8] then Table 10 of the *New Cambridge Statistical Tables* can be used to find the (upper tail) critical value of 1.708. For $\alpha = 0.01$, the critical value is 2.485. Since $1.708 < 2.14 < 2.485$, we conclude that the test is significant at the 5% level, but not at the 1% level, indicating that the result is moderately significant.

[6]Common sense might lead us to perform a lower-tailed test since 1,570 < 1,600, suggesting that if μ is not 1,600, then it is likely to be less than 1,600. However, since the question is phrased as a two-tailed test, a justification for performing a lower-tailed test would be required, should you decide to opt for a lower-tailed test. Indeed, in principle the alternative hypothesis should be determined *before* data are collected, to avoid the data biasing our choice of alternative hypothesis.

Using the p-value approach, the p-value is calculated to be:

$$p\text{-value} = P(t_{25} > 2.14) \approx 0.02$$

using Table 9 of the *New Cambridge Statistical Tables*, with 24 as the nearest available degrees of freedom. We see that $0.01 < 0.021 < 0.05$ so, of course, we reach the same conclusion. Note that we can view the p-value as the smallest α value such that we would reject the null hypothesis.[9]

We conclude that there is moderate evidence against the null hypothesis and that comprehensive planning does improve the average annual return on investment.

8.14 Hypothesis test for a single proportion

We now consider the hypothesis test for a single proportion. Recall from Section 7.10 that the standard error for a proportion is $\text{S.E.}(\pi) = \sqrt{\pi(1-\pi)/n}$. When testing $\text{H}_0 : \pi = \pi_0$, **under** $\mathbf{H}_0$, the standard error is $\sqrt{\pi_0(1-\pi_0)/n}$ leading to the following test statistic, achieved by standardising the sample proprotion estimator, P, given in (7.1).

z test of hypothesis for a single proportion

In this case, the test statistic is:

$$Z \cong \frac{P - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \sim N(0,\, 1) \qquad \text{(approximately)}. \tag{8.3}$$

Hence critical values are obtained from the standard normal distribution, i.e. using Table 5 of the *New Cambridge Statistical Tables*.

Example 8.5 To illustrate this, let us reconsider Example 7.3 in which a survey was conducted by a bank to estimate the proportion of its customers who would be interested in using a proposed new mobile telephone banking service. If we denote the population proportion of customers who are interested in this proposal by π, and it is found that 68 out of 150 customers are in favour, then we would estimate π with $p = 68/150 = 0.453$. Let us suppose that other surveys have shown that 40% of the public are interested in mobile telephone banking and it is proposed to test whether or not the above survey agrees with this figure, i.e. we conduct a two-tailed test. Hence:

$$\text{H}_0 : \pi = 0.4 \qquad \text{vs.} \qquad \text{H}_1 : \pi \neq 0.4.$$

[7] Before the financial crisis!

[8] Recall that if the question does not specify a significance level, then $\alpha = 0.05$, representing a 5% significance level, is a good starting point.

[9] You are not expected to calculate p-values from the Student's t distribution in the examination, although the Examiners will not penalise those who use this approach provided the p-value is compared against two appropriate significance levels.

The test statistic value, using (8.3), is:

$$z = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.453 - 0.4}{\sqrt{0.4 \times 0.6/150}} = 1.325.$$

Let $\alpha = 0.05$, then we compare this test statistic value with the critical values ± 1.96 (using Table 5 of the *New Cambridge Statistical Tables*), so it can be seen that the test is not significant at the 5% significance level as we do not reject H_0 since $1.325 < 1.96$. The significance level decision tree (see Figure 8.1) now requires us to test at the 10% significance level, which gives critical values of ± 1.6449, so again we do not reject H_0 since $1.325 < 1.6449$, i.e. the test result is not statistically significant.

We conclude that these data are consistent with the null hypothesis and that the level of interest shown among the bank's customers for the proposed new mobile telephone banking service may well be 40%.

8.15 Difference between two population proportions

As with confidence intervals, the correct approach to the comparison of two population proportions, π_1 and π_2, is to consider the *difference* between them, i.e. $\pi_1 - \pi_2$. When testing for equal proportions (i.e. a zero difference), the null hypothesis is therefore $H_0 : \pi_1 - \pi_2 = 0$, or equivalently $H_0 : \pi_1 = \pi_2$. We derive the test statistic by standardising the sampling distribution of the difference in two independent sample proportions, $P_1 - P_2$, given in (7.4), leading to the proposed test statistic:

$$Z = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{\pi_1(1-\pi_1)/n_1 + \pi_2(1-\pi_2)/n_2}} \sim N(0,\, 1).$$

However, when evaluating this test statistic, which values do we use for π_1 and π_2? In the test of a single proportion, we had $H_0 : \pi = \pi_0$, where π_0 is the tested value. When comparing two proportions, under H_0 no value is given for π_1 and π_2, only that they are equal, that is $\pi_1 = \pi_2 = \pi$, where π is the **common proportion** whose value, of course, is still unknown! Hence we need to estimate π from the sample data using the **pooled proportion estimator**.

Pooled proportion estimator

If R_1 and R_2 represent the number of 'favourable' responses from two independent samples with sample sizes of n_1 and n_2, respectively, then the pooled proportion estimator is:

$$P = \frac{R_1 + R_2}{n_1 + n_2}. \tag{8.4}$$

This leads to the following revised test statistic.

z test for the difference between two proportions

In this case, the test statistic is:

$$Z \cong \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{P(1-P)\,(1/n_1 + 1/n_2)}} \sim N(0,\, 1) \qquad \text{(approximately)}. \qquad (8.5)$$

Hence critical values are obtained from the standard normal distribution, i.e. using Table 5 of the *New Cambridge Statistical Tables*.

Example 8.6 To illustrate this, let us reconsider Example 7.7 to test for a difference between the population proportions of the general public who are aware of a particular commercial product before and after an advertising campaign. Two surveys were conducted and the results of the two random samples were:

	Sample size	Number aware
Before campaign	150	68
After campaign	120	65

If π_1 and π_2 are the true population proportions before and after the campaign, respectively, then we wish to test:

$$\mathrm{H}_0 : \pi_1 = \pi_2 \qquad \text{vs.} \qquad \mathrm{H}_1 : \pi_1 < \pi_2.$$

Note that we use a one-sided alternative on the assumption that the campaign would not decrease awareness![10]

After checking that this way round the value we get is positive (avoiding negative differences wherever possible to keep the notation as 'light' as possible), we estimate the difference in the population proportions as:

$$\text{'After'} - \text{'Before'} = p_2 - p_1 = \frac{65}{120} - \frac{68}{150} = 0.0884.$$

On the assumption that H_0 is true, we estimate the common proportion, π, using (8.4), as:

$$\frac{68 + 65}{150 + 120} = 0.4926.$$

So our test statistic value, using (8.5), is:

$$z = \frac{0.0884}{\sqrt{0.4926(1 - 0.4926)\,(1/150 + 1/120)}} = 1.44.$$

For a z test at the 5% significance level, we compare this test statistic value with the upper-tail critical value of 1.645, for this upper-tailed test. Since $1.44 < 1.6449$, we fail to reject H_0 and conclude that the test is not significant at the 5% significance level. But it is significant at the 10% significance level (because $1.2816 < 1.44$), and so we maintain belief in the null hypothesis with some reservation as the data have thrown some doubt on it. It is possible that the campaign has increased awareness but, on the basis of these samples, we would not be fully convinced – the test is at

best 'weakly significant'. If we had to take an important decision, we would almost certainly want more information before doing so!

Incidentally, the p-value for this problem is simply $P(Z > 1.44) = 0.0749$, using Table 4 of the *New Cambridge Statistical Tables*. So, clearly, we would be unable to reject H_0 for any significance level $\alpha < 0.0749$, agreeing with our earlier findings.

8.16 Difference between two population means

In this section we are primarily interested in the difference between two population means, $\mu_1 - \mu_2$. As in Chapter 7, there are four cases to consider.

8.16.1 Unpaired samples – variances known

Suppose we have two random samples of size n_1 and n_2, respectively, drawn from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, where σ_1^2 and σ_2^2 are known. Testing for the equality of means gives the null hypothesis $H_0 : \mu_1 = \mu_2$ or, in terms of their difference, $H_0 : \mu_1 - \mu_2 = 0$. Standardising the sampling distribution of $\bar{X}_1 - \bar{X}_2$, shown in (7.6), gives the test statistic.

z test for the difference between two means (variances known)

In this case, the test statistic is:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0, 1). \tag{8.6}$$

Hence critical values are obtained from the standard normal distribution, i.e. using Table 5 of the *New Cambridge Statistical Tables*.

Note if testing for the equality of means, then $\mu_1 - \mu_2 = 0$ under H_0. Hence, in (8.6), we set the term $(\mu_1 - \mu_2) = 0$.

8.16.2 Unpaired samples – variances unknown and unequal

We have the same set-up as above, with the same sampling distribution for $\bar{X}_1 - \bar{X}_2$, but now the population variances σ_1^2 and σ_2^2 are unknown. Assuming *large* sample sizes we can replace these unknown parameters with the sample variance estimators S_1^2 and S_2^2 to obtain the test statistic.

[10] An example of the importance of using common sense in determining the alternative hypothesis!

z test for the difference between two means (variances unknown)

If the population variances σ_1^2 and σ_2^2 are unknown, provided sample sizes n_1 and n_2 are large (greater than 30), then:

$$Z \cong \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim N(0,\, 1) \quad \text{(approximately, for large samples).} \tag{8.7}$$

Hence critical values are obtained from the standard normal distribution, i.e. using Table 5 of the *New Cambridge Statistical Tables*.

Note if testing for the equality of means, then $\mu_1 - \mu_2 = 0$ under H_0. Hence, in (8.7), we set the term $(\mu_1 - \mu_2) = 0$.

8.16.3 Unpaired samples – variances unknown and equal

Although still unknown, if we assume the population variances are equal to some common variance, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then we only have one (common) unknown variance to estimate. As with confidence intervals, we utilise the pooled variance estimator, given in (7.8).

t test for the difference between two means (variances unknown)

If the population variances σ_1^2 and σ_2^2 are unknown but assumed equal, then:

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_p^2\,(1/n_1 + 1/n_2)}} \sim t_{n_1+n_2-2} \tag{8.8}$$

where S_p^2 is the pooled variance estimator, given in (7.8).
Hence critical values are obtained from the Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom, i.e. using Table 10 of the *New Cambridge Statistical Tables*.[11]

Note if testing for the equality of means, then $\mu_1 - \mu_2 = 0$ under H_0. Hence, in (8.7), we set the term $(\mu_1 - \mu_2) = 0$.

Look back at Section 7.13 for hints about how to decide whether or not $\sigma_1^2 = \sigma_2^2$.

Example 8.7 To illustrate this, let us reconsider Example 7.8 where two companies supplying a similar service are compared for their reaction times (in days) to complaints. Random samples of recent complaints to these companies gave the following statistics:

	Sample size	Sample mean	Sample std. dev.
Company A	12	8.5	3.6
Company B	10	4.8	2.1

[11] Recall for sufficiently 'large' degrees of freedom, a standard normal approximation can be used, provided a justification is given.

We want to test for a true difference in reaction times, that is:

$$\mathrm{H}_0 : \mu_A = \mu_B \qquad \text{vs.} \qquad \mathrm{H}_1 : \mu_A \neq \mu_B.$$

Because the markets are 'similar', it is reasonable to assume (though it is only an assumption!) that the two population variances are equal. Under this assumption, using (7.8), we have:

$$s_p^2 = \frac{(12-1)\times(3.6)^2 + (10-1)\times(2.1)^2}{12+10-2} = 9.1125.$$

Using (8.8), the test statistic value is therefore:

$$t = \frac{8.5 - 4.8}{\sqrt{9.1125\,(1/12 + 1/10)}} = 2.87.$$

There are $12 + 10 - 2 = 20$ degrees of freedom, hence we obtain critical values from the t_{20} distribution using Table 10 of the *New Cambridge Statistical Tables*. For a two-tailed test at the 5% significance level, we obtain critical values of ± 2.086, hence the test is significant at this level since $2.086 < 2.87$. Moving to the 1% significance level, as per the significance level decision tree in Figure 8.1, the critical values are ± 2.845, so again we (just) reject H_0 since $2.845 < 2.87$. On this basis, we have a highly significant result when we reject H_0, and we conclude that the mean reaction times are different. Indeed, it appears that Company B reacts faster than Company A.

8

8.16.4 Paired (dependent) samples

Recall from Section 7.13, that for paired (dependent) samples we work with differenced data to reduce matters to a one sample analysis. As before, we compute the differenced data as:

$$d_1 = x_1 - y_1, \quad d_2 = x_2 - y_2, \quad \ldots, \quad d_n = x_n - y_n$$

reducing the two-sample problem to a one-sample problem. The test is then analogous to the hypothesis test of a single mean with σ unknown.

t test for the difference in means in paired samples

Using the sample mean and sample standard deviation of differenced data, then:

$$T = \frac{\bar{X}_d - \mu_d}{S_d/\sqrt{n}} \sim t_{n-1}. \tag{8.9}$$

Hence critical values are obtained from the Student's t distribution with $n-1$ degrees of freedom, i.e. using Table 10 of the *New Cambridge Statistical Tables*.

Example 8.8 The table below shows the before and after weights (in pounds) of 8 adults after a diet. Test whether there is evidence that the diet is effective.

Before	After	Before	After
127	122	150	144
130	120	147	138
114	116	167	155
139	132	153	152

We want to test:

$$\mathrm{H}_0 : \mu_1 = \mu_2 \qquad \text{vs.} \qquad \mathrm{H}_1 : \mu_1 > \mu_2$$

which is equivalent to:

$$\mathrm{H}_0 : \mu_d = 0 \qquad \text{vs.} \qquad \mathrm{H}_1 : \mu_d > 0$$

where we choose a one-tailed test because we are looking for a reduction (if there is any change, we would expect weight *loss* from a diet!) and we define $\mu_d = \mu_1 - \mu_2$ since we anticipate that this way round the values will (more likely) be positive.

The differences (calculated as 'Before − After') are:

$$5 \quad 10 \quad -2 \quad 7 \quad 6 \quad 9 \quad 12 \quad 1.$$

Hence $n = 8$, $\bar{x}_d = 6$ and $s_d = 4.66$ on $n - 1 = 7$ degrees of freedom. Using (8.9), the test statistic value is:

$$t = \frac{6 - 0}{4.66/\sqrt{8}} = 3.64.$$

At the 5% significance level, the upper-tail critical value is 1.895 (using Table 10 of the *New Cambridge Statistical Tables*), hence we reject H_0 since $1.895 < 3.64$. Using Figure 8.1, we proceed to test at the 1% significance level, which uses a critical value of 2.998. Therefore we, again, reject H_0 since $2.998 < 3.64$ and we conclude that the test is highly significant. These data strongly suggest that the diet reduces weight.

8.17 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 6 and 7 before you move on to look at the chi-squared distribution in Chapter 9.

8.18 Key terms and concepts

- Alternative hypothesis
- Critical region
- Lower-tailed test
- P-value
- Significance level
- Common proportion
- Critical value
- Null hypothesis
- Power
- Test statistic (value)

- Two-tailed test
- Type II error
- Type I error
- Upper-tailed test

8.19 Learning activities

1. Think about each of the following statements. Then give the null and alternative hypotheses and say whether they will need one- or two-tailed tests.

   (a) The general mean level of family income in a population is known to be 10,000 'ulam' per year. You take a random sample in an urban area U and find the mean family income is 6,000 ulam per 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 the 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?

2. Complete the following chart:

Real situation	**Result from your test**	
	Not reject H_0	Reject H_0
H_0 true	Correct decision Probability $(1-\alpha)$ called the **confidence level** of the test	
H_1 true	Type II error	Probability $(1-\beta)$ called the **power** of the test

3. The manufacturer of a patient 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 π.) Is your test one- or two-tailed?

4. A sample of seven is taken at random from a large batch of (nominally 12 volt) batteries. These are tested and their true voltages are shown below:

   12.9 11.6 13.5 13.9 12.1 11.9 13.0.

   (a) Test if the mean voltage of the whole batch is 12 volts.

   (b) Test if the mean batch voltage is less than 12 volts.

5. Explain what you understand by the statement: 'The test is significant at the 5% significance level'. How would you interpret a test that was significant at the 10% significance level, but not at the 5% significance level?

   In a particular city it is known, from past surveys, that 25% of homemakers regularly use a washing powder named 'Snolite'. After an advertising campaign, a survey of 300 randomly selected homemakers showed that 100 had recently purchased 'Snolite'. Is there evidence that the campaign had been successful?

6. If you live in California, the decision to purchase earthquake insurance is a critical one. An article in the *Annals of the Association of American Geographers* (June 1992) investigated many factors that California residents consider when purchasing earthquake insurance. The survey revealed that only 133 of 337 randomly selected residences in Los Angeles County were protected by earthquake insurance.

   (a) What are the appropriate null and alternative hypotheses to test the research hypothesis that less than 40% of the residents of Los Angeles County were protected by earthquake insurance?

   (b) Do the data provide sufficient evidence to support the research hypothesis? (Use $\alpha = 0.10$.)

   (c) Calculate and interpret the p-value for the test.

7. A random sample of 250 households in a particular community was taken and in 50 of these the lead level in the water supply was found to be above an acceptable level. A sample was also taken from a second community, which adds anti-corrosives to its water supply and, of these, only 16 out of 320 households were found to have high levels of lead level. Is this conclusive evidence that the addition of anti-corrosives reduces lead levels?

8. Two different methods of determination of the percentage fat content in meat are available. Both methods are used on portions of the same meat sample. Is there any evidence to suggest that one method gives a higher reading than the other?

Meat Sample	Method		Meat Sample	Method	
	I	II		I	II
1	23.1	22.7	9	38.4	38.1
2	23.2	23.6	10	23.5	23.8
3	26.5	27.1	11	22.2	22.5
4	26.6	27.4	12	24.7	24.4
5	27.1	27.4	13	45.1	43.5
6	48.3	46.8	14	27.6	27.0
7	40.5	40.4	15	25.0	24.9
8	25.0	24.9	16	36.7	35.2

9. The data in the following table show the numbers of daily parking offences in two areas of a city. The day identifications are unknown and the recordings were not necessarily made on the same days. Is there evidence that the areas experience different numbers of offences?

Area A	Area B	Area A	Area B
38	32	33	28
38	38	27	32
29	22	32	34
45	30	32	24
42	34	34	

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

In addition, attempt the 'Test your understanding' self-test quizzes available on the VLE.

8.20 Further exercises

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060]. Relevant exercises from Sections 9.1–9.4 and 10.1–10.3.

8

8.21 A reminder of your learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- set up the null and alternative hypotheses for a problem and state whether the latter is one- or two-sided, hence leading to a one- or two-tailed test
- define and apply the terminology of statistical testing
- perform statistical tests on means and proportions
- construct and explain a simple chart showing the kinds of errors that can be made in hypothesis testing.

8.22 Sample examination questions

1. Say whether the following statement is **true** or **false** and briefly give your reasons. 'The power of a test is the probability that the correct hypothesis is chosen.'

2. Explain your attitude to a null hypothesis if a test of the hypothesis is significant:
   (a) at the 1% significance level
   (b) at the 10% significance level, but not at the 5% significance level.

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) $+1.95$?

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 of the proportion and make a test. What test would you use?

5. Look at Question 3 of the Sample examination questions in Chapter 7. 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 decide whether or not to reject the null hypothesis.

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

Chapter 9
Contingency tables and the chi-squared test

9.1 Aims

This chapter introduces you to a further application of hypothesis testing using a new distribution – the chi-squared (χ^2) distribution. This distribution enables you to work with categorical variables. Your aims are to:

- learn a new application
- link this with different applications in the business and social statistical fields.

9.2 Learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- set up the null and alternative hypotheses appropriate for a contingency table
- compute the degrees of freedom, expected frequencies and appropriate critical values of chi-squared for a contingency table
- summarise 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 a one-row or one-column contingency table as above.

9.3 Essential reading

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060] Sections 14.1 and 14.3.

In addition there is essential 'watching' of this chapter's accompanying video tutorials accessible via the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

9.4 Further reading

- Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics for Business and Economics*. (Chichester: John Wiley & Sons, 1990) fourth edition [ISBN 9780471615170] Chapter 17.

9.5 Introduction

In Chapter 8 we focused on testing the value of a particular population parameter of interest, such as a mean, μ, or a proportion, π. Being able to perform such statistical tests is of particular use when making policy decisions. Here we shall look at two additional testing procedures, one which deals with testing for association between two *categorical* variables (introduced in Chapter 4), and a second which considers the *shape* of the distribution from which the sample data was drawn. Both tests can easily be applied to business management and social science fields.

9.6 Correlation and association

If we are asking whether two measurable variables (see Section 4.6) are related to each other, the technical way of describing the case when they are is to say that the two variables are **correlated**. It is possible to measure the strength of a correlation based on sample data, and also to study to what extent changes in a variable 'explain' changes in another – a topic known as **regression**. These topics will be discussed in Chapter 12.

But we also often wish to answer questions such as 'Do vegetarians tend to support Party X?' or 'Do Party X supporters tend to be vegetarians?'. If either question has the affirmative answer, we say that there is **association** between the two categories (or 'attributes' or 'factors'). We will need frequency data in order to answer such questions, but, in this course, we will **not** be concerned with the strength of any such association, rather we restrict ourselves to testing for its existence and commenting on the nature of any association which we discover.

Finally, note that if one variable is measurable and the other categorical, we would still be testing for the existence of association instead of computing correlation. In which case we would 'demote' the measurable variable to categorical status by creating levels. For example, age could be converted to a categorical variable by creating age groups.[1]

9.7 Tests for association

This type of test, tests the null hypothesis that two factors (or attributes) are **not** associated, against the alternative hypothesis that they **are** associated. Each data unit we sample has one level (or 'type' or 'variety') of each factor.

[1]Note age group would be an ordinal variable, since we could rank the age groups in order, from youngest to oldest.

Example 9.1 Suppose that we are sampling people, and that one factor of interest is hair colour (black, blonde, brown etc.) while another factor of interest is eye colour (blue, brown, green etc.). In this example, each sampled person has one level of each factor. We wish to test whether or not these factors are associated. Hence:

H_0 : There is no association between hair colour and eye colour[2]

H_1 : There is association between hair colour and eye colour.

So, under H_0, the distribution of eye colour is the same for blondes as it is for brunettes etc., whereas if H_1 is true it may be attributable to blonde-haired people having a (significantly) higher proportion of blue eyes, say.

The association might also depend on the sex of the person, and that would be a third factor which was associated with (i.e. *interacted with*) both of the others.[3] The main way of analysing these questions is by using a **contingency table**, discussed in the next section.

9.7.1 Contingency tables

In a **contingency table**, also known as a **cross-tabulation**, the data are in the form of frequencies (counts), where the observations are organised in cross-tabulated categories. We sample a certain number of units (people perhaps) and classify them according to the (two) factors of interest.

Example 9.2 In three areas of a city, a record has been kept of the numbers of burglaries, robberies and car thefts that take place in a year. The total number of offences was 150, and they were divided into the various categories as shown in the following **contingency table**:

Area	Burglary	Robbery	Car theft	Total
A	30	19	6	55
B	12	23	14	49
C	8	18	20	46
Total	50	60	40	150

The cell frequencies are known as **observed frequencies** and show how the data are spread across the different combinations of factor levels. The first step in any analysis is to complete the row and column totals (as already done in this table).

[2]When conducting tests for association, the null hypothesis can be expressed either as 'There is no association between categorical variables X and Y', or that 'Categorical variables X and Y are independent'. The corresponding alternative hypothesis would then replace 'no association' or 'independent' with 'association' or 'not independent (dependent)', respectively.

[3]We will not consider interactions in testing in **ST104a Statistics 1**.

9.7.2 Expected frequencies

We proceed by computing a corresponding set of **expected frequencies**, conditional on the null hypothesis of no association between the factors, i.e. that the factors are independent.

Now suppose that you are only given the row and column totals for the frequencies. If the factors were assumed to be independent, consider how you would calculate the expected frequencies. Recall from Chapter 5 that if A and B are two independent events, then $P(A \cap B) = P(A) \cdot P(B)$. We now apply this idea.

Example 9.3 For the data in Example 9.2, if a record was selected at random from the 150 records:

- P(a crime being a burglary) $= 50/150$
- P(a crime being in area A) $= 55/150$.

Hence, under H_0, we have:

$$P(\text{a crime being a burglary in area A}) = \frac{50}{150} \cdot \frac{55}{150}$$

and so the **expected number** of burglaries in area A is:

$$150 \times \frac{50}{150} \cdot \frac{55}{150}.$$

So the expected frequency is obtained by multiplying the product of the 'marginal' probabilities by n, the total number of observations. This can be generalised as follows.

Expected frequencies in contingency tables

The expected frequency, E_{ij}, for the cell in row i and column j of a contingency table with r rows and c columns, is:

$$E_{ij} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{total number of observations}}$$

where $i = 1, \ldots, r$ and $j = 1, \ldots, c$.

Example 9.4 The completed expected frequency table for the data in Example 9.2 is (rounding to two decimal places – which is recommended in the examination):

Area	Burglary	Robbery	Car theft	Total
A	18.33	22.00	14.67	55
B	16.33	19.60	13.07	49
C	15.33	18.40	12.27	46
Total	50	60	40	150

Make sure you can replicate these expected frequencies using your own calculator.

9.7.3 Test statistic

To motivate our choice of test statistic, if H_0 is true then we would expect to observe *small* differences between the observed and expected frequencies, while *large* differences would suggest that H_1 is true. This is because the expected frequencies have been calculated conditional on the null hypothesis of independence hence, if H_0 is actually true, then what we actually observe (the observed frequencies) should be (approximately) equal to what we expect to observe (the expected frequencies).

χ^2 test of association

Let the contingency table have r rows and c columns, then formally the test statistic[4] used for tests of association is:

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}. \tag{9.1}$$

Critical values are found using Table 8 of the *New Cambridge Statistical Tables*.

Notice the 'double summation' here just means summing over all rows and columns. This test statistic follows an (approximate) chi-squared distribution with $(r-1)(c-1)$ degrees of freedom, where r and c denote the number of rows and columns, respectively, in the contingency table.

9.7.4 The χ^2 distribution

A **chi-squared variable** is only defined over positive values. The precise shape of the distribution is dependent on the degrees of freedom, which is a parameter of the distribution. Figure 9.1 illustrates the chi-squared distribution for a selection of degrees of freedom.

Although the shape of the distribution does change quite significantly for different degrees of freedom, note the distribution is always positively skewed.

9.7.5 Degrees of freedom

The general expression for the number of degrees of freedom in tests for association is:

$$\text{(Number of cells)} - \text{(Number of times data is used to calculate } E_{ij}\text{s)}\,.$$

For an $r \times c$ contingency table, we begin with rc cells. We lose one degree of freedom for needing to use the total number of observations to compute the expected frequencies. However, we also use the row and column totals in these calculations, but we only need

[4]This test statistic was first used by the statistician Karl Pearson, hence it is often referred to as 'Pearson's chi-squared statistic'. We will not worry in this course about how this test statistic is derived.

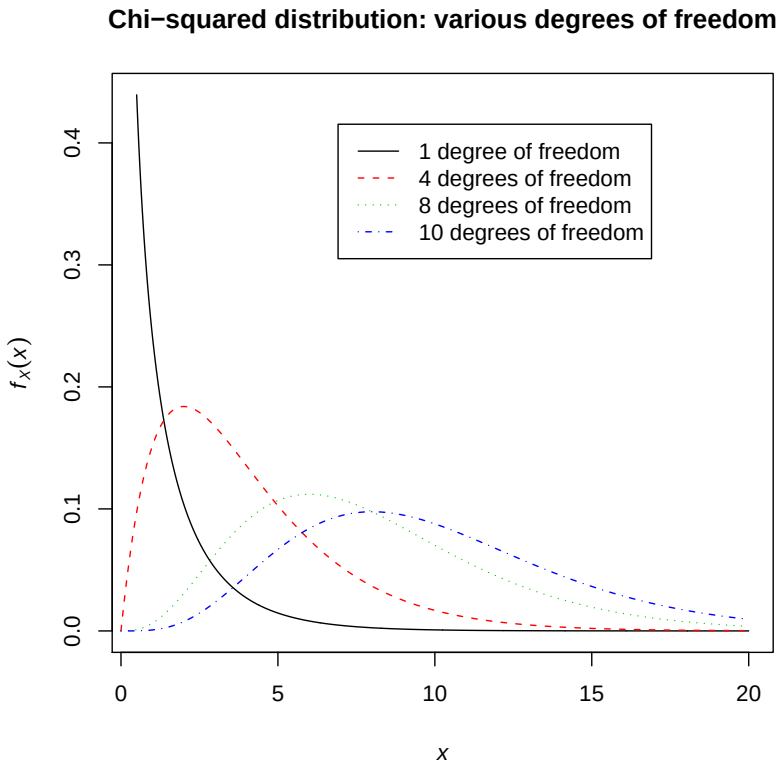


Figure 9.1: Examples of the chi-squared distribution with a selection of degrees of freedom.

$r-1$ row totals and $c-1$ column totals, as the final one in each case can be deduced using the total number of observations. Hence we only lose $r-1$ degrees of freedom for the row totals, similarly we only lose $c-1$ degrees of freedom for the column totals.

Hence the overall degrees of freedom are:

$$\nu = rc - (r-1) - (c-1) - 1 = (r-1)(c-1).$$

9.7.6 Performing the test

As usual, we choose a significance level, α, at which to conduct the test. But are we performing a one-tailed test or a two-tailed test? To determine this, we need to consider what sort of test statistic value would be considered extreme under H_0. As seen in Figure 9.1, the chi-squared distribution only takes positive values. The squared term in the numerator of the test statistic in (9.1) ensures the test statistic value will be positive (the E_{ij}s in the denominator are clearly positive too). If H_0 is true, then observed and expected frequencies should be quite similar, since the expected frequencies are computed conditional on the null hypothesis of independence. This means that $|O_{ij} - E_{ij}|$ should be quite small for all cells. In contrast if H_0 is not true, then we would expect comparatively large values for $|O_{ij} - E_{ij}|$ due to large differences between the two sets of frequencies. Therefore, upon squaring $|O_{ij} - E_{ij}|$, sufficiently *large* test statistic values suggest rejection of H_0. Hence χ^2 tests of association are always **upper-tailed tests**.

Example 9.5 Using the data in Example 9.2, we proceed with the hypothesis test. Note it is advisable to present your calculations as an extended contingency table as shown below, where the three rows in each cell correspond to the observed frequencies, the expected frequencies and the test statistic contributor.

		Burglary	Robbery	Car theft	Total
	O	30	19	6	55
A	E	18.33	22.00	14.67	55
	$(O-E)^2/E$	7.48	0.41	5.15	
	O	12	23	14	49
B	E	16.33	19.60	13.07	49
	$(O-E)^2/E$	1.13	0.59	0.06	
	O	8	18	20	46
C	E	15.33	18.40	12.27	46
	$(O-E)^2/E$	3.48	0.01	4.82	
Total		50	60	40	150

Using (9.1), we obtain a test statistic value of:

$$\sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 7.48 + 0.41 + \cdots + 4.82 = 23.13.$$

Since $r = c = 3$, we have $(r-1)(c-1) = (3-1)(3-1) = 4$ degrees of freedom. For $\alpha = 0.05$, using Table 8 of the *New Cambridge Statistical Tables*, we obtain an upper-tail critical value of 9.488. Hence we reject H_0 since $9.488 < 23.13$. Moving to the 1% significance level, the critical value is now 13.28 so, again, we reject H_0 since $13.28 < 23.13$. Therefore the test is highly significant and we conclude that there is an association between the factors.

Looking again at the contingency table, comparing observed and expected frequencies, the interpretation of this association becomes clear – burglary is the main problem in area A whereas car theft is a problem in area C. (We can deduce this by looking at the cells which large test statistic contributors, which are a consequence of large differences between observed and expected frequencies.)

Incidentally, the p-value for this upper-tailed test is (using a computer) $P(23.13 \leq \chi^2_4) = 0.000119$, emphasising the extreme significance of this test statistic value.

The conclusions in Example 9.5 are fairly obvious, given the small dimensions of the contingency table. However, for data involving more factors, and more factor levels, this type of analysis can be very insightful. Cells that make a large contribution to the test statistic value (i.e. which have large values of $(O-E)^2/E$) should be studied carefully when determining the nature of an association. This is because, in cases where H_0 has been rejected, rather than simply conclude that there is an association between two categorical variables, it is helpful to describe the nature of the association.

9.7.7 Extending to an appropriate test of proportions

If we had a 2×2 contingency table, then there are two ways of testing an equivalent hypothesis – (i.) a χ^2 test as outlined above, and (ii.) a test of the difference in proportions as detailed in Chapter 8. See the fourth Learning activity at the end of this chapter for an example to practise.

9.8 Goodness-of-fit tests

In addition to tests of association, the chi-squared distribution is often used more generally in so-called 'goodness-of-fit' tests. We may, for example, wish to answer hypotheses such as 'Is it reasonable to assume the data follow a particular distribution?'. This justifies the name 'goodness-of-fit' tests, since we are testing whether or not a particular probability distribution provides an adequate *fit* to the observed data. The null hypothesis will assert that a *specific* hypothetical population distribution **is** the true one. The alternative hypothesis is that this *specific* distribution **is not** the true one.

'Goodness-of-fit' tests are covered in detail in **ST104b Statistics 2** for a variety of probability distributions. However, there is a special case which we shall consider in **ST104a Statistics 1** when you are only dealing with one row or one column. This is when we wish to test that the sample data are drawn from a (discrete) uniform distribution, i.e. that each characteristic is equally likely.

> **Example 9.6** Is a given die fair? If the die is fair, then the values of the faces (1, 2, 3, 4, 5 and 6) are all equally likely so we have the following hypotheses:
>
> H_0 : Score is uniformly distributed vs. H_1 : Score is not uniformly distributed.

9

9.8.1 Observed and expected frequencies

As with tests of association, the goodness-of-fit test involves both observed and expected frequencies. In all goodness-of-fit tests, the sample data must be expressed in the form of **observed frequencies** associated with certain classifications of the data. Assume k classifications, hence observed frequencies can be denoted by O_i, $i = 1, \ldots, k$.

> **Example 9.7** Extending Example 9.6, for a die the obvious classifications would be the six faces. If the die is thrown n times, then our observed frequency data would be the number of times each face appeared. Here $k = 6$.

Recall that in hypothesis testing we always assume that the null hypothesis, H_0, is true. In order to conduct a goodness-of-fit test, **expected frequencies** are computed conditional on the probability distribution expressed in H_0. The test statistic will then involve a comparison of the observed and expected frequencies. In broad terms, if H_0 is true, then we would expect *small* differences between these two sets of frequencies, while *large* differences would indicate support for H_1.

We now consider how to compute expected frequencies for discrete uniform probability distributions – the only goodness-of-fit distributions considered in **ST104a Statistics 1**.

Expected frequencies in goodness-of-fit tests

For discrete uniform probability distributions, expected frequencies are computed as:

$$E_i = n \times \frac{1}{k}, \qquad i = 1, 2, \ldots, k$$

where n denotes the sample size and $1/k$ is the uniform (equal, same) probability for each characteristic.

Expected frequencies should not be rounded, just as we do not round sample means, say. Note that the final expected frequency (for the kth category) can easily be computed using the formula:

$$E_k = n - \sum_{i=1}^{k-1} E_i.$$

This is because we have a constraint that the sum of the observed and expected frequencies must be equal,[5] that is:

$$\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i$$

which results in a loss of one degree of freedom (discussed below).

9.8.2 The goodness-of-fit test

Having computed expected frequencies, we now need to formally test H_0 and so we need a test statistic.

Goodness-of-fit test statistic

For a discrete uniform distribution with k categories, observed frequenices O_i and expected frequencies E_i, the test statistic is:

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1} \qquad \text{approximately under } \mathrm{H}_0. \tag{9.2}$$

Note that this test statistic does **not** have a true χ^2_{k-1} distribution under H_0, rather it is only an approximating distribution. An important point to note is that this approximation is only good enough provided **all the expected frequencies are at least 5**. (In cases where one (or more) expected frequencies are less than 5, we **merge** categories with neighbouring ones until the condition is satisfied – this point is covered in greater depth in **ST104b Statistics 2**.)

[5] Recall the motivation for computing expected frequencies in the first place – assuming H_0 is true, we want to know how a sample of size n is expected to be distributed across the k categories.

As seen in (9.2), there are $k-1$ degrees of freedom when testing a discrete uniform distribution. k is the number of categories (after merging), and we lose one degree of freedom due to the constraint that:

$$\sum_i O_i = \sum_i E_i.$$

As with the test of association, goodness-of-fit tests are **upper-tailed** tests as, under H_0, we would expect to see *small* differences between the observed and expected frequencies, as the expected frequencies are computed conditional on H_0. Hence *large* test statistic values are considered extreme under H_0, since these arise due to large differences between the observed and expected frequencies.

Example 9.8 A confectionery company is trying out different wrappers for a chocolate bar – its original, A, and two new ones, B and C. It puts the bars on display in a supermarket and looks to see how many of each wrapper type have been sold in the first hour, with the following results.

Wrapper type	A	B	C	Total
Observed frequencies	8	10	15	33

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

To answer this we need to test:

H_0 : There is no difference in preference for the wrapper types[6]

H_1 : There is a difference in preference for the wrapper types.

How do we work out the expected frequencies? Well, for **equal** preferences, with three choices each ($k=3$), the expected frequencies will be:

$$E_i = 33 \times \frac{1}{3} = 11, \qquad i = 1, 2, 3.$$

Applying (9.2), our test statistic value is:

$$\sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i} = \frac{(8-11)^2}{11} + \frac{(10-11)^2}{11} + \frac{(15-11)^2}{11} = 2.364.$$

The degrees of freedom will be $k-1 = 3-1 = 2$. At the 5% significance level, the upper-tail critical value is 5.991, using Table 8 of the *New Cambridge Statistical Tables*, so we do not reject H_0 since $2.364 < 5.991$. If we now consider the 10% significance level[7] the critical value is 4.605, so again we do not reject H_0 since $2.364 < 4.605$. Hence the test is not significant. So it looks as if there are no strict preferences for a particular wrapper type based on the choices observed in the supermarket during the first hour.

[6]This is the same as testing for the suitability of the discrete uniform distribution, i.e. that each

9.9 Summary

You should regard this chapter as an opportunity to revise your work on hypothesis testing in Chapter 8 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.

9.10 Key terms and concepts

- Association
- Chi-squared distribution
- Cross-tabulation
- Expected frequency
- Observed frequency
- Categorical variable
- Contingency table
- Discrete uniform distribution
- Goodness-of-fit test

9.11 Learning activities

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 and very dissatisfied. Levels of housing type 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 a χ^2 test.

   (Remember that you will reject the null hypothesis if your calculated test statistic value is greater than the critical value obtained from Table 8 of the *New Cambridge Statistical Tables*.)

2. Draw the χ^2 curve and put in the rejection region for 5% and 1% significance levels with 6 degrees of freedom. Make sure you understand which calculated values of χ^2 will lead you to reject your H_0.

3. 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.

wrapper is equally popular.

[7]Remember if the test is not significant at the 5% significance level, we should then test at the 10% significance level. See Figure 8.1.

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

(a) Does the population proportion of skilled workers vary with the area? (Be careful to ensure you know what you are doing!)

(b) Test:

$$H_0 : \pi_D = \pi_{others} \quad \text{vs.} \quad H_1 : \pi_D < \pi_{others}.$$

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.

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.

	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 H_0 and H_1 in each case.)

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

Season	**Live births**
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 93 (Spring), 80 (Summer), 100 (Autumn) and 92 (Winter).

(Hint: You would expect, if the births are uniformly distributed over the year, that the number of births would be proportionate to the number of days per season. So

work out your expected frequencies by taking the number of days per season divided by the number of days in the year and multiplying by the total number of births over the year.)

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

In addition, attempt the 'Test your understanding' self-test quizzes available on the VLE.

9.12 Further exercises

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060]. Relevant exercises from Sections 14.1 and 14.3.

9.13 A reminder of your learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- set up the null and alternative hypotheses appropriate for a contingency table
- compute the degrees of freedom, expected frequencies and appropriate critical values of chi-squared for a contingency table
- summarise 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 a one-row or one-column contingency table as above.

9.14 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 an association between advertising and sales of a product. You have 4 levels of advertising (A, B, C and D) and 3 levels of sales (low, medium and high).

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

2. The table below shows a contingency table for a sample of 1,104 randomly selected adults from three types of environment (City, Town and Rural) who have been 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

3. Two surveys have collected information on adult and teenage cigarette use. The results of the first survey are given in the first table below (sample size 4,000) and the results of the second survey, carried out two years later on a new sample of 1,000 households, are underneath it.

   (a) Without doing any further calculations, comment on any association between rows and columns for the first survey.

   (b) Calculate the chi-squared statistic for the second survey and test for association between rows and columns.

   (c) Write a short report explaining what the first survey table shows about the nature of the association, if any, between adult and teenage cigarette use in a household, and (by comparing the two tables) discuss whether or not the extent of any association changed over the years.

Survey 1		**Adult cigarette use**				**Total**
		Yes		No		
Teenage	Yes	198	(8.4%)	170	(10.4%)	368
cigarette use	No	2,164	(91.6%)	1,468	(89.6%)	3,632
$\chi^2 = 4.61$	Total	2,362	(100.0%)	1,638	(100.0%)	4,000

Survey 2		**Adult cigarette use**				**Total**
		Yes		No		
Teenage	Yes	48	(8.1%)	45	(11.0%)	93
cigarette use	No	544	(91.9%)	363	(89.0%)	907
	Total	592	(100.0%)	408	(100.0%)	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 uniformly between the four seasons? Outline the steps in your work.

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

Chapter 10
Sampling design

10.1 Aims

If you are using this subject guide in the order in which it is presented, you can now take a break from learning any more new statistical ideas for a while. Now is your chance to consolidate these ideas by looking at their applications in business and the social sciences. You should aim to consider the following types of question.

- What are your choices when you design a survey?
- How reliable are results of surveys reported in the media?
- Are random samples the best we can do in every circumstance?

10.2 Learning outcomes

After completing this chapter, and having completed the Essential reading and activities, 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 methods of contact that may be used in a survey and the related implications:
  - interviewer
  - postal
  - other self-administered
  - telephone.

10.3 Essential reading

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060] Section 6.1 and Chapter 17.

In addition there is essential 'watching' of this chapter's accompanying video tutorials accessible via the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

10.4 Further reading

- Office for National Statistics *Social Trends*. (Basingstoke: Palgrave Macmillan, 2009) [ISBN 9780230220508].

Those taking **SC1021 Principles of sociology** should reread Chapter 2 in that subject guide and look at Chapter 3 to give context to their studies for **ST104a Statistics 1**.

10.5 Introduction

This chapter describes the main stages of a survey and the sources of error at each stage. Remember that a first treatment and background to this subject matter is given earlier in this subject guide in Chapter 3. It makes sense to reread that now before continuing. 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 subject guide unless you understand the sources and limitations of the data you are using. This is important to academics and policymakers alike! The material in this chapter and Chapter 12 is a useful preparation or companion for **SC1021 Principles of sociology**.

You should find Newbold et al. Chapter 17 useful background. Note, however, that while the ideas presented are important conceptually, the computations included in Chapter 17 of Newbold et al. will **not** be required for **ST104a Statistics 1**.

10.6 Motivation for sampling

In Chapter 3, we introduced the term **target population** as representing the collection of units (people, objects, etc.) in which we are interested. In the absence of time and budgetary constraints we conduct a **census**, that is a total enumeration of the population. Examples include the population census in the UK and in other countries around the world. Its advantage is that there is no **sampling error** because all population units are observed, and so there is no estimation of population parameters – we can, in principle, determine the true values exactly.

Of course, in the real world, we do need to take into account time and budgetary constraints. Due to the large size, N, of most populations, an obvious disadvantage with

a census is cost. So a census is not feasible in practice because there is only a limited amount of information which is economical to collect. Also, **non-sampling errors** may occur. For example, if we have to resort to using cheaper (hence less reliable) interviewers, then they may erroneously record data, misunderstand a respondent, etc. So we select a **sample**, that is a certain number of population members are selected and studied. The selected members are known as **elementary sampling units**.

Sample surveys (hereafter 'surveys') are how new data are collected on a population and tend to be based on samples rather than censuses. However, in practice, surveys are usually conducted under circumstances which cannot be fully controlled. Since the sample size, n, can be very large, the design and implementation of surveys requires considerable planning and teamwork.

Example 10.1 Examples of sample surveys include:

- **Demography**: study of births, deaths, families, etc.
- **Government**: study of crime, education, employment, housing, health, etc.
- **Market research**: study of consumer preferences, attitudes, etc.
- **Political science**: study of voting intention.

Selected respondents may be contacted in a variety of methods such as face-to-face interviews, telephone, mail or email questionnaires. Sampling errors will occur (since not all population units are observed, so we resort to estimating population parameter values, which means there will be some uncertainty attached to our point estimates), but because of the smaller numbers involved ($n \ll N$)[1] resources can be used to ensure high quality interviews or to check completed questionnaires. Therefore non-sampling errors should be less and consequently researchers can ask more questions.

So a census is difficult to administer, time-consuming, expensive and does not guarantee completely accurate results due to non-sampling errors. In contrast, a sample is easier, faster and cheaper, although it introduces sampling errors while non-sampling errors can still occur. Unsurprisingly, we would like sampling (and non-sampling) errors to be small.

Before proceeding, it is worth spending a moment thinking about the classification of data into **primary data** and **secondary data**:

- Primary data are *new* data collected by researchers for a particular purpose.
- Secondary data refer to existing data which have already been collected by others or for another purpose.

Of course, if secondary data exist, then there would be no need to collect new data! In **ST104a Statistics 1**, the focus will be on the collection of primary data using various sampling techniques.

[1] The symbol $\ll$ means 'much less than'.

10.7 Types of sample

We seek a sample which is **representative** of the whole population to ensure our estimated parameter values are good approximations of the true parameter values. A particular difficulty is how to select a sample which yields a 'fair' representation. Fortunately, there are many different sampling techniques (reviewed below) from which to choose. As expected, these all have relative advantages and disadvantages, hence each will be suitable in particular situations. Knowledge of these techniques is frequently examined in this course, so it is strongly advised that you become very familiar with this material.

Of course, consideration also needs to be given to the choice of sample size, n. In practice we face a trade-off – n should be large enough to give a fair representation of the population, yet be sufficiently small to be practical and cost-effective. Section 6.11 considered this problem in the context of setting a tolerance on the maximum permitted sampling error. Qualitative factors to take into consideration when determining n are study objectives, budget, time constraints, type of data analysis planned and non-sampling errors.

Types of sample

We proceed by partitioning sampling techniques into two groups:

- Non-probability (non-random) sampling
- Probability (random) sampling.

10.7.1 Non-probability sampling

Non-probability samples are characterised by the following properties:

- Some population units have no chance (zero probability) of being selected.
- Units which can be selected have an *unknown* (non-zero) probability of selection.
- Sampling errors cannot be quantified.
- Non-probability samples are used in the absence of a sampling frame.[2]

We shall consider three types of non-probability sampling:

- **Convenience sampling** – individuals are selected from the members of the population which happen to be in the vicinity. For example, those people in a shopping mall at the time of the survey, say a school day. This technique is convenient, hence the name! Of course, such an approach is unlikely to yield a representative sample. (In the mall example, children and working professionals are unlikely to be in the mall at the time.)

[2]Sampling frames (lists) were discussed at length in Chapter 3.

- **Judgement sampling** – this uses expert opinion to judge which individuals to choose. An example would be a research firm selecting people to participate in a focus group[3] by telephoning people and identifying who matches the *target profile* through a series of questions developed by an expert.

- **Quota sampling** – this attempts to obtain a representative sample by specifying quota controls on certain specified characteristics, such as age, gender, social class and any other variables relevant to the investigation being undertaken. The (approximate) distribution of such characteristics in the population is required in order to replicate it in the sample.

Of these, quota sampling is the most important type which we will consider now in greater depth.

Quota sampling

Quota sampling is useful in the absence of a sampling frame, since non-probability sampling techniques do not require a sampling frame. Another reason for conducting a quota sample, instead of a random sample, might be speed. We may be in a hurry and not want to spend time organising interviewers for a random sample – it is much quicker to set target numbers (quotas) to interview.

As with any form of sampling, the objective is to obtain a sample which is representative of the population. Quota sampling is no exception and to do so the interviewer:

- seeks out units which satisfy some control characteristics (known as **quota controls**), such as age, gender and social class

- requires the *distribution* of these characteristics in the population in order to replicate it in the sample.

Quota sampling is cheap, but it may be systematically biased by the choice of interviewee made by the interviewer, and their willingness to reply. For instance, interviewers might avoid choosing anyone who looks threatening, or too busy, or too strange! Quota sampling also does not allow us to measure the sampling error – a consequence of it being a non-probability sampling technique.

Each of the three techniques above (convenience, judgement and quota) has an appeal; however, in all cases we have no real guarantee that we have achieved an adequately representative sample – we might do by chance, of course, but this would be highly unlikely! For example, were the women we interviewed only those working in local offices? Were the young adults all students?

Basically, since 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 (confidence intervals and hypothesis tests) do not apply. Specifically, **standard errors** (a key ingredient in inferential procedures) are not measurable. However, in the absence of a sampling frame then we will have to resort to non-probability sampling methods.

[3]Focus groups are used for qualitative research by asking a group of individuals about their opinions toward a new product or advertising campaign, consumer preferences, etc.

That said, non-random samples are also frequently used by market research organisations or companies when speed (if not accuracy) is important. They are rarely used by governments.

Example 10.2 You would likely use a quota sample (the main non-probability sample considered in **ST104a Statistics 1**) in the following situations:

- **When speed is important.** Clearly, an interviewer with a target to reach a certain number (quota) of people on a given day is likely to be quicker than one which requires a *specific* person or household to be contacted (as determined by a random sample). Typical quota controls for the interviewer to meet are:
  - age
  - gender
  - socio-economic group, or social class.

  Note 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).
- **No available sampling frame covering the target population.** If you think obtaining a list is likely to be very complicated, then a *sensible* targeting of the population by taking a quota sample might be helpful. 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 for data protection reasons, say. One of the things you could do in these circumstances is to carry out a quota sample 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 take a random sample of routes, interviewing the relevant driver as he or she completes their journey.
- **When you need to reduce cost.** Clearly, time-saving is an important element in cost-saving.
- **When accuracy is not important.** You may not need to have an answer to your question to the high – and *known* – level of accuracy that is possible using a random sample; rather you merely require an **idea** about a subject. Perhaps you only need to know if people, on the whole, *like* your new flavour of ice cream in order to judge whether or not there is likely to be sufficient consumer demand to justify full-scale production and distribution. In this case, asking a representative group of people (quota) would be perfectly adequate for your needs.

Although there may be several reasons to justify the use of a quota sample, you should be aware of the problem caused by the **omission of non-respondents**. Because you only count the individuals who reply (unlike random sampling where your estimate has

to allow for bias through non-response), the omission of non-respondents[4] can lead to serious errors as 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.

One possible remedy to this problem is to introduce more detailed quota controls. For example, we might ask for age, gender, employment status, marital status and/or the particular age of the respondent's children. However, 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 this is not done, the temptation to cheat, and therefore make results completely invalid, will be strong.

10.7.2 Probability sampling

Probability sampling means that every population unit has a **known (not necessarily equal), non-zero probability of being selected** in the sample. In all cases selection is performed through some form of *randomisation*. For example, a pseudo-random number generator can be used to generate a sequence of 'random' numbers.

Relative to non-probability methods, probability sampling can be expensive and time-consuming, and also requires a sampling frame. We aim to minimise both the (random) sampling error and the systematic sampling bias. Since the probability of selection is known, standard errors can be computed which allows confidence intervals to be determined and hypothesis tests to be performed.

We shall consider five types of probability sampling:

- **Simple random sampling (SRS)** – a special case where each population unit has a known, *equal* and non-zero probability of selection. Of the various probability samples, its simplicity is desirable and it produces unbiased estimates, although more accurate methods exist (i.e. those with smaller standard errors).

- **Systematic random sampling** – a 1-in-x systematic random sample is obtained by randomly selecting one of the first x units in the sampling frame and then selecting every subsequent xth unit. Though easy to implement, it is important to consider how the sampling frame is compiled. For example, the data may exhibit a *periodic* property such as sales in a shop. All Monday sales are likely to be similar, as are all Saturday sales, but Saturdays are traditionally much busier shopping days. So sales from a 1-in-7 systematic sample will have a large variation *between* days, but a small variation *within* days. This would lead us to underestimate the true variation in sales. A 1-in-8 systematic sample would be better. Can you say why?[5]

[4] Non-response is discussed later in this chapter.

[5] A 1-in-7 systematic sample would lead to only one day of the week being represented in the sample, such as Mondays. A 1-in-8 systematic sample would have different days of the week. For example, if the first day chosen was a Monday, the next day selected would be 8 days later, i.e. Tuesday of the following week, followed by a Wednesday etc.

- **Stratified random sampling** – this sampling technique achieves a higher level of accuracy (smaller standard errors) by exploiting natural groupings within the population. Such groupings are called *strata*[6] and these are characterised as having population units which are similar *within* strata, but are different *between* strata. More formally, elements within a stratum are *homogeneous* while the strata are collectively *heterogeneous*.

  A simple random sample is taken from each stratum, thus ensuring a representative overall sample since a broad cross-section of the population will have been selected – provided suitable strata were created. Hence great care needs to be taken when choosing the appropriate **stratification factors**, which should be relevant to the purpose of the sample survey such as age or gender.

  Imagine we were investigating student satisfaction levels at a university. If we took a simple random sample of students we could, by chance, select just (or mainly) first-year students whose opinions may very well differ from those in other year groups. So in this case a sensible stratification factor would be 'year of study'. By taking a simple random sample from each stratum you ensure that you will not end up with an extreme (non-representative) sample and also avoid the possibility of one particular group not being represented at all in the sample. Of course, in order to perform stratified random sampling, we would need to be able to allocate each population unit to a stratum. For students this should be straightforward, since the university's database of student names would no doubt include year of study too.

- **Cluster sampling** – is used to reduce costs (time and money). Here the population is divided into *clusters*, ideally such that each cluster is as variable as the overall population (i.e. heterogeneity *within* clusters and homogeneity *between* clusters). Next, some of the clusters are selected by SRS. This is where the economy savings are expected (in face-to-face interviews), since typically the clusters may be constructed on a geographical basis allowing interviewers to restrict themselves to visiting households in certain areas rather than having to travel long distances to meet the respondents required by a simple random sample who potentially cover a large area. A **one-stage cluster sample** would mean that *every* unit in the chosen clusters is surveyed.

  It may be that the cluster sizes are large meaning that it is not feasible to survey every unit within a cluster. A **two-stage cluster sample** involves taking a simple random sample from the clusters which were selected by SRS during the first stage. When a subsample is taken from a selected cluster, this is known as a **multistage design**.

  Individual respondents will be identified in a random manner – but crucially *within* an area. You will reduce costs (the interviewer will be able to complete a higher number of interviews in a given time, using less petrol and reducing shoe leather costs), but will probably have to sacrifice a degree of accuracy, that is the method is less efficient in general compared to, say, stratified random sampling. However, cluster sampling is very useful when a sampling frame is not immediately available for the entire population. For example, individual universities will have databases of their own students, but a national student database does not exist (admittedly it

[6] 'Strata' is the plural form, 'stratum' is the singular form.

may be possible to create a national student database, but it would take a lot of effort, time and money).

Clustering is clearly useful in an interviewer-administered survey. It is less important as a design feature for telephone or postal interviews, unless you are particularly 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 (hence standard error) of your sample estimate will be greater.

A further reason, in addition to cost savings, 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.

- **Multistage sampling** – refers to the case when sample selection occurs at two or more successive stages (the two-stage cluster sample above is an example). Multistage sampling is frequently used in large surveys. During the first stage, large compound units are sampled (*primary units*). During the second stage, smaller units (*secondary units*) are sampled from the primary units. From here, additional sampling stages of this type may be performed, as required, until we finally sample the basic units.

  As we have already seen, this technique is often used in cluster sampling as we initially sample main clusters, then clusters within clusters, etc. We can also use **multistage sampling with a mixture of techniques**. A large government survey will likely incorporate elements of both stratification and clustering at different stages. A typical multistage sample in the UK might involve the following:

  - Divide the areas of the country into strata by industrial region.
  - Sample clusters (local areas) from **each** industrial region.
  - From each local area choose some areas for which you have lists (such as electoral registers) and take a simple random sample from the chosen lists (clusters).

Note that from a technical perspective, stratified sampling can be thought of as an extreme form of two-stage cluster sampling where at the first stage *all* clusters in the population are selected. In addition, one-stage cluster sampling is at the opposite end of this spectrum. We can summarise as follows.

Similarities and differences between stratified and cluster sampling

- Stratified sampling: *all* strata chosen, *some* units selected in each stratum.
- One-stage cluster sampling: *some* clusters chosen, *all* units selected in each sampled cluster.
- Two-stage cluster sampling: *some* clusters chosen, *some* units selected in each sampled cluster.

Example 10.3 You have been asked to make a sample survey of each of the following. Would you use random or quota sampling? Explain.

i. Airline pilots, for their company, about their use of holiday entitlement in order to bring in a new work scheme.

In this case 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 collect 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.

ii. 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.

The situation for the tourist survey is different from that for airline pilots. There will be not just 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.

iii. Household expenditure for government assessment of the effect of different types of taxes.

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.

10.8 Types of error

We can distinguish between two types of error in sampling design[7] as follows:

- **Sampling error**: This occurs as a result of us selecting a sample, rather than performing a census (where a total enumeration of the population is undertaken). It is attributable to random variation due to the sampling scheme used. For probability sampling, we can estimate the statistical properties of the sampling

[7] These should **not** be confused with Type I and Type II errors, which only concern hypothesis testing and are covered in Section 8.7.

error, i.e. we can compute (estimated) standard errors which allow confidence intervals to be determined and hypothesis tests to be performed.

- **Non-sampling error**: This occurs as a result of the (inevitable) failures of the sampling scheme. In practice it is very difficult to quantify this sort of error, typically through separate investigation. We distinguish between two sorts of non-sampling error:
  - **Selection bias** – this may be due to (i.) the sampling frame not being equal to the target population, (ii.) the sampling frame not being strictly adhered to, or (iii.) non-response bias.
  - **Response bias** – the actual measurements might be wrong due to, for example, ambiguous question wording, misunderstanding of a word in a questionnaire by less-educated people, or sensitivity of information which is sought. *Interviewer bias* is another aspect of this, where the interaction between the interviewer and interviewee influences the response given in some way, either intentionally or unintentionally, such as through leading questions, the dislike of a particular social group by the interviewer, the interviewer's manner or lack of training, or perhaps 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.

10.9 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:

- to find the standard error which can be attached to different kinds of questions and hence to underpin the sampling design chosen
- 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 subsample 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.

10.10 Non-response and response bias

Bias caused by non-response and response is worth a special entry. It can cause problems at every stage of a survey, both random and quota, and however administered.

The first problem can be in the sampling frame. Is an obvious group missing?

Example 10.4 The following are examples of coverage bias in the sampling frame.

- If the sampling frame is of householders, those who have just moved in will be missing.
- If the sampling frame 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** (data not provided by a unit that we wish to sample) is one of the major problems of sample surveys as the non-respondents, in general, cannot be treated like the rest of the population. As such, 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.

Classifications of non-response

We can classify non-response as follows:

- **Item non-response** occurs when a sampled member fails to respond to a question in the questionnaire.
- **Unit non-response** occurs when no information is collected from a sample member.

Non-response may be due to any of the following factors:

- **not-at-home** due to work commitments, or on holiday
- **refusals** due to subject matter, or sponsorship of the survey
- **incapacity to respond** due to illness, or language difficulties
- **not found** due to vacant houses, incorrect addresses, moved on
- **lost schedules** due to information being lost or destroyed after it had been collected.

How should we deal with non-response? Well, note that increasing the sample size will **not** solve the problem – the only outcome would be that we have more data on the types of individuals who are *willing* to respond! Instead, we might look at improving our survey procedures such as data collection and interviewer training. Non-respondents could be followed up using call-backs, or an alternative contact method to the original survey in an attempt to subsample the non-respondents. A *proxy interview* (where a unit from your sample is substituted with an available unit) may be another possibility. (Note that non-response also occurs in quota sampling but is not generally recorded – see the earlier discussion.) However, an obvious remedy is to provide an **incentive** (for example, cash or entry into a prize draw) to complete the survey – this exploits the notion that human behaviour can be influenced in response to the right incentives!

Response error is very problematic because 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 good example from the UK 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 HM Revenue & Customs!

Sources of response error

The main sources of response error include:

- **Role of the interviewer** due to the characteristics and/or opinions of the interviewer, asking leading questions and the incorrect recording of responses.
- **Role of the respondent** who may lack knowledge, forget information or be reluctant to give the correct answer due to the sensitivity of the subject matter.

Control of response errors typically involves improving the recruitment, training and supervision of interviewers, reinterviewing, consistency checks and increasing the number of interviewers.

In relation to all these problems, pilot work is very important. It may also be possible to carry out a check on the interviewers and methods used after the survey (post-enumeration surveys).

10.11 Method of contact

A further point you should think about when assessing how to carry out a survey is the **method of contact**. The most common methods of contact are face-to-face interviews, telephone interviews and online/postal/mail (so-called 'self-completion') interviews. 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 target population is on the telephone (and how good the telephone system is).
- Self-completion questionnaires can have a low response rate.

We now explore some[8] of the advantages and disadvantages of various contact methods.

- **Face-to-face interviews:**
  - Advantages: Good for personal questions; allow for probing issues in greater depth; permit difficult concepts to be explained; can show samples (such as new product designs).
  - Disadvantages: (Very) expensive; not always easy to obtain detailed information on the spot.

[8]This is not necessarily an exhaustive list. Can you add any more?

- **Telephone interviews:**
  - Advantages: Easy to achieve a large number of interviews; easy to check on quality of interviewers (through a central switchboard perhaps).
  - Disadvantages: Not everyone has a telephone so the sample can be biased; cannot usually show samples; although telephone directories exist for *landline* numbers, what about mobile telephone numbers? Also, young people are more likely to use mobile telephones rather than landline telephones, so are more likely to be excluded.

- **Self-completion interview:**
  - Advantages: Most people can be contacted this way (there will be little non-response due to not-at-home reasons); allow time for people to look up details such as income, tax returns etc.
  - Disadvantages: High non-response rate – it requires effort to complete the questionnaire; answers to some questions may influence answers to earlier questions since the whole questionnaire is revealed to the respondent – this is important where the order of a questionnaire matters; you have no control over who answers the questionnaire.

Example 10.5 Examples of occasions when you might use a particular method are as follows.

- **Face-to-face interviewer – 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 telephone, or are unwilling to talk to strangers by telephone.

- **Telephone interviewer – a survey of businessmen and their attitudes to a new item of office equipment.**

  All of them will have a telephone, and also the questions should be simple to ask. Here, booking time for a telephone interview at work (once it has been agreed with the management) should be much more effective than waiting for a form to be filled in, or sending interviewers to disrupt office routine.

- **Postal/mail questionnaire – 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, since they may be hoping for a pay rise! Also, the recipients 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 details), while the respondent household is required to fill in a self-completion diary (showing expenditure, which could not be obtained by an interview alone).

Similarly, telephone interviews may be combined with a mail-shot subsampled individual survey, in the case of offices and businesses 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 in advance, or as they are telephoned.

Remember also that email surveys are already widespread and becoming increasingly popular, although they are only appropriate when the population to be studied uses them heavily and is likely to reply to your questions, such as employees in your office. An obvious advantage is that this method is very cheap to administer.

10.12 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 subject 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.

10.13 Key terms and concepts

- Census
- Convenience sampling
- Incentive
- Multistage design
- Non-response
- Pilot survey
- Primary data
- Representative
- Quota sampling
- Sample survey
- Sampling frame
- Selection bias
- Stratified random sampling
- Target population
- Cluster sampling
- Judgement sampling
- Method of contact
- Non-probability sampling
- Non-sampling error
- Post-enumeration survey
- Probability sampling
- Response error
- Sample
- Sampling error
- Secondary data
- Simple random sampling
- Systematic random sampling

10.14 Learning activities

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.

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 per week. Two controls should be easy for your interviewer to identify. How can you help them with the third control?

3. Find out about one of the government surveys carried out in your own country.

   This should help you understand the problems involved in designing a useful survey and help you with illustrations for your examination 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!)

4. Your company has five sites and a mixture of managerial, clerical 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 rail season tickets.

   You, as a statistician, have been asked to design 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 the strata will help you.

   Hint: Do not be afraid to use two sets of strata in these circumstances!

5. You have been asked to design a random sample in order to study the way schoolchildren learn in your country. Explain the clusters you might choose, and why.

6. Your textbook will have a clear and detailed example of a multistage 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 mind.

7. What form of contact might you use for your questionnaire in the following circumstances?
   (a) A random sample of schoolchildren 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 television programmes.

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 audiotape. How might you stratify your sample? Explain.

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

In addition, attempt the 'Test your understanding' self-test quizzes available on the VLE.

10.15 Further exercises

Try the following example exercises in the subject guide for course **SC1021 Principles of sociology**: 2.2, 2.9 and 2.12.

10.16 A reminder of your learning outcomes

After completing this chapter, and having completed the Essential reading and activities, 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 methods of contact that may be used in a survey and the related implications:
  - interviewer
  - postal
  - other self-administered
  - telephone.

10.17 Sample examination questions

1. (a) Define a 'quota' sample.
   (b) What are the main reasons for using such a sample, and what are the alternatives?
   (c) What are the main sources of error in a quota sample, and how would you deal with them?

2. Given the data from Chapter 7, Sample examination question 3 part (b) and Chapter 8, Sample examination 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.

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.

4. Discuss the statistical problems you might expect to have in each of the following situations:
   (a) Conducting a census of the population.
   (b) Setting up a study of single-parent families.
   (c) Establishing future demand for post-compulsory education.

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

Chapter 11
Some ideas underlying causation – the use of control groups and time order

11.1 Aims

As with Chapter 10, this chapter does not introduce any new statistical calculations. However, it addresses the key ideas of causation in non-experimental science. It also prepares you for the final ideas – namely regression and correlation – you need to know in order to complete **ST104a Statistics 1**.

11.2 Learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- distinguish between an experiment in the natural sciences and the observations possible in social science, business or management studies
- determine sensible controls for a given experiment or trial
- set up a panel study
- explain the merits and limitations of a longitudinal/panel survey compared with a cross-sectional survey.

11.3 Essential reading

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060] Section 13.2.

In addition there is essential 'watching' of this chapter's accompanying video tutorials accessible via the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

11.4 Further reading

- Shipman, M. *The Limitations of Social Research*. (London: Longman, 1997) fourth edition [ISBN 9780582311039] Part 2.

Those taking **SC1021 Principles of sociology** should reread Section 2.4 and Chapter 3 with care!

11.5 Introduction

So far, we have looked at ways of collecting and describing data. Chapters 3 and 10 introduced you to the basic ideas of sampling from populations. The main ways of describing and presenting data were covered in Chapter 4. Chapters 6 to 8 dealt with the ways we can assess the relevance or significance of these figures, while Chapter 9 introduced the idea of 'association' between categorical variables. Chapter 12 will complete the process, so far as **ST104a Statistics 1** is concerned, by covering correlation and regression.

Before we look at the ideas of correlation and regression more formally, it is important to take stock of the limitations of assuming **causation** when working in 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, then 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'[1]), the assumption is made that the treatment has *caused* the change between the first and second set of observations.

If we are dealing with such a controlled experiment, the meaning of our statistical measures and work is very clear. However, in social science, business and management fields, things are rarely that simple. We are generally faced with figures that show changes in variables but the treatment given is not easy to assess.

> **Example 11.1** Consider determining the effectiveness of an advertising campaign for your new chocolate bars with improved wrappers. Your company measures the sales of the old chocolate bars, product A, in two areas, X and Y, before introducing the new chocolate bars, product B, and running a four-month advertising campaign.
>
> Imagine your marketing manager's joy when sales of product B are much higher than for product A in area X. Clearly, the changes have worked! But, oh dear, product B is achieving a lower level of sales than product A in area Y. How can this be? Well, on closer investigation, we find that our main rival has withdrawn their product from area X and concentrated their sales focus on area Y (where they have realised that a larger number of their target population lives). So your success with product B in area X is not necessarily related to your advertising campaign, rather it may be entirely due to your rival's actions. Clearly, whatever measures you use, there is a problem with the measurement of your results. 'Other things' are no longer equal, since the behaviour of the rival changed while you were conducting your experiment.

[1] Or '*ceteris paribus*', in Latin.

So it is important to **control for confounding factors**, that is, factors which are correlated with the observed variables (such as the rival's actions in Example 11.1). Failure to properly control for such factors may lead us to treat a false positive as a genuine causal relationship.

11.6 Observational studies and designed experiments

In medical, 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 economic and other social science applications of statistics, one usually just observes a sample of the available population, without controlling for any of the factors that may influence the measures observed. Such 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 better to have a designed experiment, whenever possible. Of course, the social sciences do not really lend themselves to designed experiments. For example, we do not have multiple versions of an economy to which we can apply different mixtures of fiscal and monetary policies. Instead, we have one economy to which we apply a particular mix and observe what happens.

In **ST104a Statistics 1**, we will only look at *regression* and *correlation* as ways of studying the connections between variables, but analysis of variance (covered in **ST104b 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 (or a placebo) when a new treatment is available. After the trial is complete, if the new treatment is shown to be worse than the established one, there may 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 others.

When we come to study regression in Chapter 12, it will be important to remember its limitations. It is easy to be too optimistic about its use. 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 positive 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.

So how do statisticians try to measure causal relationships in the social sciences? They use two main weapons:

- the control group

- time order.

We now proceed to consider both of these.

11.7 Use of the control group

We are often limited in the social sciences because we are unable to perform experiments either 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 possibility of being allowed to experiment, as you are afraid that fluoride might have harmful side effects. However, you know that fluoride occurs naturally in some communities. What can you do, as a statistician?

11.7.1 Observational study

In an **observational study** data are collected on units (not necessarily people) **without any intervention**. Researchers do their best not to influence the observations in any way. A sample survey is a good example of such a study, where data are collected in the form of questionnaire responses. As discussed in Chapter 10, every effort is made to ensure response bias is minimised (if not completely eliminated).

> **Example 11.2** To assess the likely effect on tooth decay of adding fluoride to the water supply, you can look at the data for your non-fluoridated 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 that may differ between the two. If you can **match** in this way (i.e. find two communities that share the same characteristics and *only* differ in the fluoride concentration of their water supply), your results may have some significance.

So, to credibly establish a *causal* link in an observational study, all other relevant factors need to be adequately **controlled**, such that any change between observation periods can be explained by only one variable.

11.7.2 Experimental study

In an **experiment**, an **intervention** or **treatment** is administered to some or all of the experimental units (usually people). Allocation of the treatment (or perhaps a

combination of treatments) is determined by using a form of **randomisation**.[2] Some time after the treatments are given, the **outcomes** are recorded by the researchers. Data analysis involves comparing outcomes across all treatments.

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 need to be taken 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**).

In order to avoid bias in assessing the effects of the treatments, **blinding** and **double blinding** are recommended. In blind experimentation the experimental units are unaware of whether they receive a treatment or a *placebo*, while in a double-blind set-up the people administering the treatment and placebo also do not know which is which. The sample size in each treatment group must be large enough to ensure that medically (or socially) important differences can be detected. Sample size determination (as previously discussed for observational studies in Section 7.11) to ensure adequate *power* is a routine part of experimental design.

On the whole, such 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.

11.8 Time order

Another way we might attempt to disentangle causal relationships is to look at the *order* in which events occurred. Clearly, if we eat more, we gain weight, other things being equal. For example, if we do not exercise more to offset the extra consumption of calories!

This underpins work on *time series* data which you will meet if you study **EC2020 Elements of econometrics**. For now, you should know a little about **longitudinal** and **panel** surveys, where the same individuals are resurveyed over time.

11.8.1 Longitudinal surveys

Policy makers use these surveys over a long period of time 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.

[2]For example, treatment A is administered if a fair coin toss comes up 'heads', otherwise treatment B is administered.

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, say.

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 cooperate, he or she may have altered their perceptions (the participants may have become too involved in the study).

Despite the disadvantages, such studies are widely regarded as being the best way of studying change over time.

11.8.2 Panel surveys

In business and management research, we generally use slightly less ambitious surveys called **panels**. They also involve contacting the same individuals 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 to 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. For now, make sure you understand how a longitudinal or panel study is set up.

11.9 Case study: Smoking and lung cancer

In order to clarify the issues raised in this chapter, we will consider a case study. It describes how control groups were used in order to see what was connected with lung cancer. You might be interested 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
- being in hospital at roughly the same time.

Using the measures that you will meet in Chapter 12, it was found that the main difference between the two groups was in smoking behaviour – those with lung cancer were more likely to be heavy smokers than the other group. Although this study was carried out some time ago, similar studies have confirmed the results.

But what did the study tell us? Consider again Section 11.5. Descriptive statistics alone **cannot** distinguish between the following ideas.

- Smoking causes lung cancer:

$$\text{smoking} \Rightarrow \text{lung cancer}$$

- Smoking is a symptom of developing lung cancer:

$$\text{lung cancer} \Rightarrow \text{smoking}$$

- Your personality factor leads to smoking and lung cancer:

$$\text{personality factor X} \Rightarrow \text{smoking} + \text{lung cancer}.$$

It is important that you understand this. Although animal experimentation (ethics aside) may help to resolve the conundrum to some extent, we are really at a 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 UK. Their smoking histories were collected and subsequent death rates from lung cancer checked. The results confirmed the initial findings.

However, although these results make the causal connection 'heavy smoking $\longrightarrow$ lung cancer' much more likely, it is still possible that the heavy smokers may have some other characteristic (for example, genetic) 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.

11.10 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 prepares the way for the material in Chapter 12.

11.11 Key terms and concepts

- Blinding
- Conditioning
- Causation
- Confounding factor

- Control group
- Experiment
- Longitudinal survey
- Outcome
- Randomisation
- Treatment
- Drop out
- Intervention
- Observational study
- Panel survey
- Time order

11.12 Learning activities

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 per hour.

   What control factors might you use in order to examine the likely effect on road accidents of a change in your country?

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 television programmes your company is making for under-10s in the after school/before homework time slot.

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

In addition, attempt the 'Test your understanding' self-test quizzes available on the VLE.

11.13 Further exercises

For those taking **SC1021 Principles of sociology**, try Activity 2.6 again in the subject guide.

11.14 A reminder of your learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- distinguish between an experiment in the natural sciences and the observations possible in social science, business or management studies
- determine sensible controls for a given experiment or trial
- set up a panel study
- explain the merits and limitations of a longitudinal/panel survey compared with a cross-sectional survey.

11.15 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.

2. Write notes on the following:
   (a) blind trials
   (b) control groups
   (c) measuring causation.

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

Chapter 12
Correlation and regression

12.1 Aims

This final chapter of the subject guide takes us back to the ideas of the formula for a straight line, which you worked on back in Chapter 2. Here, we are going to use basic mathematics to understand the relationship between two variables.

If you are taking, or likely to take, **ST104b Statistics 2**, and later **EC2020 Elements of econometrics**, you should ensure that you cover all the examples here very carefully as they prepare you for some work on those courses.

If you do not intend to take your statistics any further, you need, at the very least as a social scientist, to be able to assess whether relationships shown by data are what they appear!

Be careful not to confuse the techniques used here – on measurable variables – with those you worked with in Chapter 9 on categorical variables!

12.2 Learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- draw and label a scatter diagram
- calculate r
- explain the meaning of a particular value and the general limitations of r and r^2 as measures
- calculate a and b for the line of best fit in a scatter diagram
- explain the relationship between b and r
- summarise the problems caused by extrapolation.

12.3 Essential reading

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060] Sections 1.6, 2.4, 11.1–11.3 and 14.6.

In addition there is essential 'watching' of this chapter's accompanying video tutorials accessible via the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

12.4 Further reading

- Aczel, A.D. *Complete Business Statistics*. (London: McGraw-Hill Higher Education, 2009) seventh edition [ISBN 9780071287531] Chapters 10 and 11.
- Anderson, D.R., D.J. Sweeney, T.A. Williams, J. Freeman and E. Shoesmith *Statistics for Business and Economics*. (South-Western Cengage Learning, 2010) eleventh edition [ISBN 9780324783247] Chapter 14 and Section 15.1.
- Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics for Business and Economics*. (Chichester: John Wiley & Sons, 1990) fourth edition [ISBN 9780471615170] Chapters 12–14.

12.5 Introduction

In Chapter 9, you were introduced to idea of testing for association between different attributes of a variable using the chi-squared distribution. We did this by looking at the number of individuals falling into a category, or experiencing a particular contingency.

Correlation and **regression** 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 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 – so-called multivariate analysis. *Factor analysis*, *discriminant analysis*, *principal component analysis* and some kinds of *cluster analysis* (not to be confused with cluster sampling!) all use related ideas and techniques.[1]

When we use these terms we are concerned with using models for **prediction** and **decision making**. So, how do we model the relationship between two variables? We are going to look at:

- correlation – which measures the **strength** of a *linear* relationship
- regression – which is a way of **representing** that *linear* relationship.

It is important you understand what these two terms have in common, but also the differences between them.

12

12.6 Scatter diagrams

Let us assume that we have some data in paired form: (x_i, y_i), $i = 1, 2, \ldots, n$.

[1]Though not covered in **ST104a Statistics 1**, these do appear in **MN3141 Principles of marketing**. Regression techniques are also emphasised in **EC2020 Elements of econometrics**.

Example 12.1 An example of paired data is the following which represents the number of people unemployed and the corresponding monthly reported crime figures for twelve areas of a city.

Unemployed, x	2,614	1,160	1,055	1,199	2,157	2,305
Number of offences, y	6,200	4,610	5,336	5,411	5,808	6,004

Unemployed, x	1,687	1,287	1,869	2,283	1,162	1,201
Number of offences, y	5,420	5,588	5,719	6,336	5,103	5,268

This dataset will be used repeatedly throughout this chapter.

When dealing with paired data, the first action is to construct a **scatter plot** (or **scatter diagram**) of the data, and visually inspect it for any apparent relationship between the two variables.[2] Figure 12.1 shows such a scatter plot for these data.

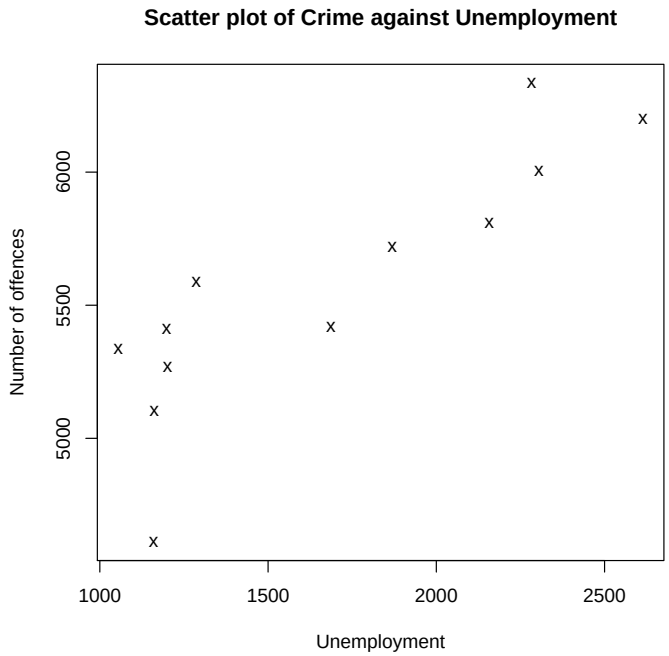


Figure 12.1: Scatter plot of 'Crime' against 'Unemployment'.

Figure 12.1 gives an impression of a *positive, linear* relationship, i.e. it can be seen that x (unemployment) and y (number of offences) increase together, roughly in a straight line, but subject to a certain amount of scatter. So, the relationship between x and y is not *exactly* linear – the points do not lie exactly on a straight line.

Data showing a general 'upward shape' like this are said to be **positively correlated**, and we shall see how to quantify this correlation. Other possible scatter patterns are shown in Figure 12.2.

[2]Notice that, in contrast to categorical variables considered in Chapter 9, for the remainder of this course we will deal with measurable variables only.

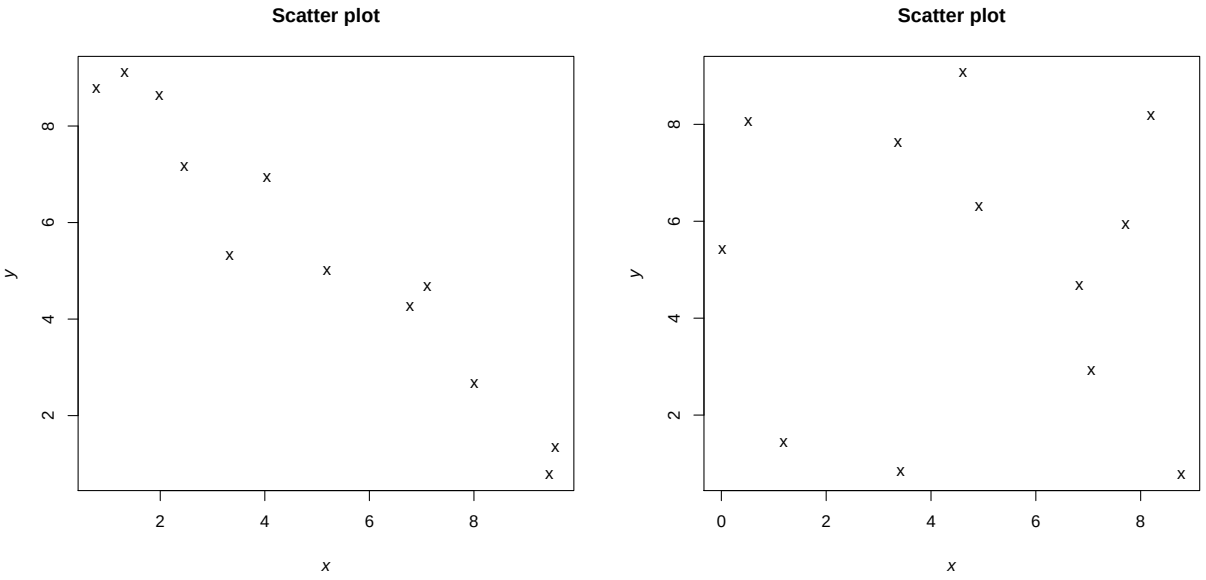


Figure 12.2: Scatter plots – negatively correlated variables (left) and uncorrelated variables (right).

The left-hand plot shows data that have a **negative correlation**, i.e. y decreases as x increases, and vice versa. The right-hand plot shows **uncorrelated** data, i.e. no clear relationship between x and y. Note that **correlation** assesses the strength of the *linear* relationship between two variables. Hence uncorrelated data, in general, just means an absence of linearity. It is perfectly possible that uncorrelated data are related, just not linearly – for example, x and y may exhibit a quadratic relationship (considered later).

Example 12.2 Below is a list of variables, along with their expected correlation.

Variables	Expected correlation
Height and weight	Positive
Rainfall and sunshine hours	Negative
Ice cream sales and sun cream sales	Positive
Hours of study and examination mark	Positive
Car's petrol consumption and goals scored	Zero

12.7 Causal and non-causal relationships

When two variables are correlated, an interesting question that arises is whether the correlation indicates a **causal** relationship. In Example 12.2, it is natural to assume that hours of study and examination marks are positively correlated, due to more study resulting in better examination performance, i.e. a higher mark. In this case, the relationship is causal.

For the ice cream and sun cream sales, however, the relationship is not causal. It is not the selling of ice cream that causes sun cream sales to rise, nor vice versa. Rather both sales respond to warm weather and so these sales are seasonal, with both rising and

falling together in response to other variables such as temperature and sunshine hours.

It should be clear from this that care needs to be taken in interpreting correlated relationships. Remember correlation does not imply causality!

Example 12.3 Let us consider a few more examples. In each case we observe strong correlations.

i. 'Average salary of school teachers' and 'Consumption of alcohol measured by country'.

ii. 'Stork population in Bavaria' and 'Human birth rate'.

iii. 'Size of student population' and 'Number of juvenile offenders by local area in a country'.

Would you seriously think there was a *causal* connection in these cases? Let us look at them in a little more detail.

i. It would be frivolous to deduce 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.

ii. This 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 (to raise their children in pleasant surroundings) which encourage storks?

iii. 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.

12.8 Correlation coefficient

The strength of a linear relationship between two random variables is given by the **correlation coefficient**. For random variables X and Y, the *population* correlation coefficient, denoted ρ, is defined as:

$$\rho = \frac{\mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y)))}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

and some texts may refer to this as the 'product-moment correlation coefficient'. Technically, we say that ρ can only be determined if we have perfect knowledge of the 'bivariate density function' of X and Y.[3] In practice, it is more likely that we will wish to *estimate* ρ, using the **sample correlation coefficient**, from a set of sample observations of X and Y, i.e. using sample paired data (x_i, y_i), $i = 1, 2, \ldots, n$.

[3] A bivariate density function shows the dependence of two variables (here X and Y) and is covered in **ST104b Statistics 2**.

Sample correlation coefficient

The sample correlation coefficient, denoted r, is:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}} \tag{12.1}$$

where:

$$\text{corrected sum of squares of } x\text{-data} \quad : \quad S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$\text{corrected sum of squares of } y\text{-data} \quad : \quad S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

$$\text{corrected sum of cross-products} \quad : \quad S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.$$

It is quite common in examination questions to be given certain summary statistics (for example, $\sum_i x_i$, $\sum_i x_i^2$, $\sum_i y_i$, $\sum_i y_i^2$ and $\sum_i x_i y_i$) to save you time from computing all quantities directly using the raw data. Hence it may be easier for you to remember the expressions for S_{xx}, S_{yy} and S_{xy} (respectively the 'corrected sum of squares' for x and y, and corresponding cross-products), and how they combine to give r.

Example 12.4 For the dataset in Example 12.1, we have $n = 12$, $\bar{x} = 1665$ and $\bar{y} = 5567$, and so:

$$S_{xx} = 3431759, \quad S_{yy} = 2584497 \quad \text{and} \quad S_{xy} = 2563066.$$

Using (12.1), we have:

$$r = \frac{2563066}{\sqrt{3431759 \times 2584497}} = 0.861.$$

This is a strong, positive correlation.[4] We also note that the value of r agrees with the scatter plot shown in Figure 12.1.

The sample correlation coefficient measures how closely the points in the scatter plot lie to a straight line, and the sign of r tells us the direction of this line, i.e. upward-sloping or downward-sloping, for positive and negative r, respectively. It does *not* tells us the gradient of this line – this is what we will determine in regression.

12

[4] If you were following **ST104b Statistics 2** and were to test the statistical significance of this r, using the kind of techniques and ideas in Chapter 8, you would find that the result is highly significant at the 1% significance level. Note that you are **not** expected to test for the significance of r in **ST104a Statistics 1**.

Properties of the correlation coefficient

Both types of correlation coefficient (ρ and r) have the following properties. They:

- are **independent** of the **scale** of measurement
- are **independent** of the **origin** of measurement
- are **symmetric**; that is, the correlation of X and Y is the same as the correlation of Y and X
- can only take values between ± 1, i.e. $-1 \leq \rho \leq 1$ and $-1 \leq r \leq 1$, alternatively $|\rho| \leq 1$ and $|r| \leq 1$ – correlation coefficients always have an absolute value less than or equal to 1.

Having defined the correlation coefficient, it is important to remember the following when interpreting r (or ρ):

- An r (ρ) near the top of this range (i.e. near 1) indicates a strong, positive linear relationship between X and Y.
- An r (ρ) near the bottom of this range (i.e. near -1) indicates a strong, negative linear relationship between X and Y.
- Zero correlation indicates that X and Y are not linearly related, i.e. the variables are uncorrelated.

Example 12.5 To reiterate, correlation assesses the strength of *linear* relationships between variables only. $\rho = 0$ does **not** imply that X and Y are independent, since the variables could have a non-linear relationship. For example, if:

$$y = x(1 - x), \qquad \text{for } 0 \leq x \leq 1$$

then the correlation is zero (as they are not linearly related), but, clearly, there is a well-defined relationship between the two variables, so they are certainly not independent. Figure 12.3 demonstrates this point for simulated sample data, where we see a clear relationship between x and y, but it is clearly not a linear relationship.[5] Data of this kind would have a (sample) correlation near zero (here, $r = 0.148$), which emphasises the linear nature of r.

12.8.1 Spearman rank correlation

Just as we saw in Chapter 4 that we can use the median and interquartile range as measures of location and dispersion, respectively, instead of the mean and standard deviation (or variance), it is possible to calculate the **Spearman rank correlation**, r_s, instead of r. Section 14.6 in Newbold et al. covers this in more detail.

[5] In fact, these data are scattered around a parabola with (approximate) equation $y = 2(x-15)(85-x)$.

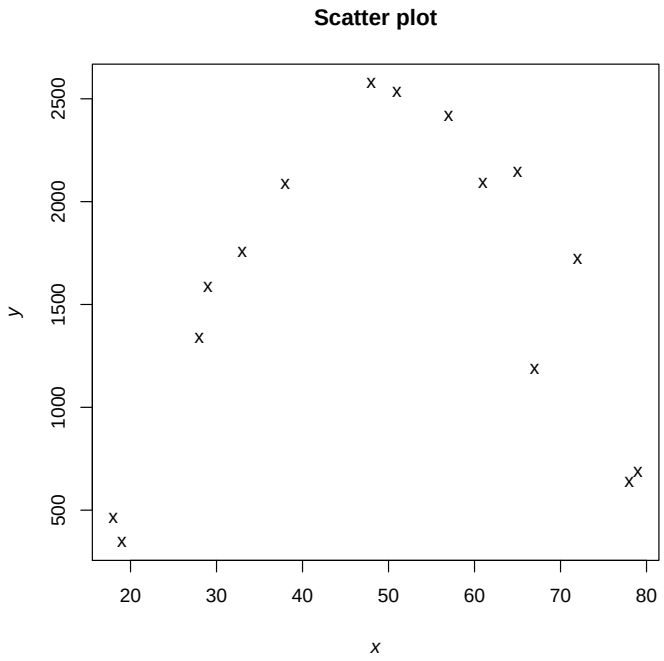


Figure 12.3: Scatter plot showing a non-linear relationship.

To compute r_s we **rank** the x_i and y_i values in ascending order and use the r formula above using the ranks of the x_is and y_is, rather than their actual values. Of course, it may be that we *only* have the rankings, in which case we would have to use this method.

Spearman rank correlation

If there are no tied rankings of x_i and y_i, the Spearman rank correlation is:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{12.2}$$

where the d_is are the differences in the ranks between each x_i and y_i.

As with other order statistics, such as the median and quartiles, it is helpful to use the Spearman rank correlation if you are worried about the effect of extreme observations (outliers) in your sample. The limits for r_s are the same as for r, that is $|r_s| \leq 1$.

Example 12.6 An aptitude test has been designed to examine a prospective salesman's ability to sell. Ten current staff sit the test. Instead of putting achieved scores in the computer, a research assistant ranks the individuals in ascending order in terms of the test as well as productivity. The data are:

Staff member	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 of sales ability?

We first compute the differences in the ranks:

Staff member	A	B	C	D	E	F	G	H	I	J
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 therefore:

$$\sum_{i=1}^{10} d_i^2 = 46.$$

Hence, using (12.2), we have:

$$r_s = 1 - \frac{6 \times 46}{10((10)^2 - 1)} = 0.72$$

which is quite strong, indicating that the test is a reasonably good predictor.

12.9 Regression

In **ST104a Statistics 1**, we can only hope to review the fundamentals of what is a very large (and important) topic in statistical analysis – several textbooks and courses focus exclusively on regression analysis. Here, we shall concentrate on the **simple linear (bivariate) regression** model. This will allow calculations to be performed on a hand held calculator, unlike multiple regression (with multiple variables) which typically requires the help of statistical computer packages due to the complexity and sheer number of calculations involved.

In the simple model, we have two variables Y and X:

- Y is the **dependent (or response) variable** – that which we are trying to explain
- X is the **independent (or explanatory) variable** – the factor we think influences Y.

Multiple regression is just a natural extension of this set-up, but with more than one explanatory variable.[6]

There can be a number of reasons for wanting to establish a mathematical relationship between a response variable and an explanatory variable, for example:

- To find and interpret unknown parameters in a known relationship.
- To understand the reason for such a relationship – is it causal?
- To predict or forecast Y for specific values of the explanatory variable.

[6]Multiple linear regression is covered in **ST104b Statistics 2**.

12.9.1 The simple linear regression model

We assume that there is a true (population) linear relationship between a response variable y and an explanatory variable x of the *approximate* form:

$$y = \alpha + \beta x$$

where α and β are fixed, but unknown, population parameters. Our objective is to estimate α and β using (paired) sample data (x_i, y_i), $i = 1, \ldots, n$.

Note the use of the word 'approximate'. Particularly in the social sciences, we would not expect a perfect linear relationship between the two variables. Hence we modify this basic model to:

$$y = \alpha + \beta x + \epsilon$$

where ϵ is some random perturbation from the initial 'approximate' line. In other words, each y observation *almost* lies on the line, but 'jumps' off the line according to the **random variable** ϵ. This perturbation is often referred to as the 'error term'.

For each pair of observations (x_i, y_i), $i = 1, \ldots, n$, we can write this as:

$$y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, \ldots n.$$

The random deviations $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ corresponding to the n data points are assumed to be **independent and identically normally distributed**, with **zero mean** and **constant (but unknown) variance,** σ^2. That is:

$$\epsilon_i \sim N(0, \sigma^2), \qquad i = 1, \ldots, n.$$

This completes the model specification.

Specification of the simple linear regression model

To summarise, we list the assumptions of the simple linear regression model:

- A linear relationship between the variables of the form $y = \alpha + \beta x + \epsilon$.
- The existence of three model parameters: the linear equation parameters α and β, and the error term variance, σ^2.
- $\text{Var}(\epsilon_i) = \sigma^2$ for all $i = 1, \ldots, n$, i.e. it does not depend on the explanatory variable.
- The ϵ_is are independent and $N(0, \sigma^2)$ distributed for all $i = 1, \ldots, n$.

You may feel that some of these assumptions are particularly strong and restrictive. For example, why should the error term variance be constant across all observations? Indeed, your scepticism serves you well! In a more comprehensive discussion of regression, such as in **EC2020 Elements of econometrics**, model assumptions would be properly tested to assess their validity. Given the limited scope of regression in **ST104a Statistics 1**, sadly we are too time-constrained to consider such tests in detail. However, do be aware that with any form of modelling, a thorough critique of model assumptions is essential.[7] Analysis based on false assumptions leads to invalid results – a bad thing.

[7] The validity of certain model assumptions is considered in **ST104b Statistics 2**.

12.9.2 Parameter estimation

As mentioned above, our principal objective is to estimate α and β, that is the y-intercept and slope of the true line. To fit a line to some data, we need a criterion for establishing which straight line is in some sense 'best'. The criterion used is to minimise the sum of the squared distances between the observed values of y_i and the values predicted by the model. (This 'least squares' estimation technique is explored in depth in **ST104b Statistics 2**.)

The estimated least squares regression line is written as:

$$\hat{y} = a + bx$$

where a and b denote our estimates for α and β, respectively. The *derivation* of the formulae for a and b is **not** required,[8] although you do need to know how to calculate point estimates for α and β.

Simple linear regression line estimates

We estimate α and β with a and b, where:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}} \quad (12.3)$$

$$a = \bar{y} - b\bar{x}. \quad (12.4)$$

Note you need to compute b first, since this is needed to compute a.

12.9.3 Prediction

Having estimated the regression line, an important application of it is for **prediction**. That is, for a given value of the explanatory variable, we can use it in the estimated regression line to obtain a prediction for y.

Predicting y for a given value of x

For a given value of the explanatory variable, x_0, the predicted value of the dependent variable, $\hat{y}$, is:

$$\hat{y} = a + bx_0.$$

Remember to attach the appropriate *units* to the prediction (i.e. the units of measurement of the original y data). Also, for that matter, ensure the value you use for x_0 is correct – if the original x data is in 000s, then a prediction of y when the explanatory variable is 10,000, say, would mean $x_0 = 10$, and not 10,000!

[8] The derivation is presented in **ST104b Statistics 2**.

Example 12.7 A study was made by a retailer to determine the relationship between weekly advertising expenditure and sales (in thousands of pounds). Find the equation of a regression line to predict weekly sales from advertising. Predict weekly sales when advertising costs are £35,000. The data are:

Advertising costs (in £000s)	40	20	25	20	30	50
Sales (in £000s)	385	400	395	365	475	440

Advertising costs (in £000s)	40	20	50	40	25	50
Sales (in £000s)	490	420	560	525	480	510

Summary statistics, representing sales as y and advertising costs as x, are:

$$\sum x = 410,\ \sum x^2 = 15650,\ \sum y = 5445,\ \sum y^2 = 2512925,\ \sum xy = 191325.$$

So, using (12.3) and (12.4), the parameter estimates are:

$$\begin{aligned} b &= \frac{191325 - (12 \times (410/12) \times (5445/12))}{15650 - (12 \times (410/12)^2)} = 3.221 \\ a &= \frac{5445}{12} - 3.221 \times \frac{410}{12} = 343.7. \end{aligned}$$

Hence the estimated regression line is:

$$\hat{y} = 343.7 + 3.221x.$$

The predicted sales for £35,000 worth of advertising is:

$$\hat{y} = 343.7 + 3.221 \times 35 = 456.4, \text{ which is £456,400.}$$

Note that since the advertising costs were given in £000s, we used $x_0 = 35$, and then converted the predicted sales into pounds.

12.9.4 Points to watch about linear regression

Non-linear relationships

We have only seen here how to use a straight line to model the best fit. So, we could be missing some quite important non-linear relationships, particularly if we were working in the natural sciences.

Which is the dependent variable?

Note it is essential to correctly establish which is the dependent variable, y. In Example 12.7, you would have a different line if you had taken advertising costs as y and sales as x! So remember to exercise your common sense – we would expect sales to react to advertising campaigns rather than vice versa.

Extrapolation

In Example 12.7, we used our estimated line of best fit to predict the value of y for a given value of x, i.e. advertising expenditure of £35,000. Such prediction is only reliable if you are dealing with figures which lie *within* the dataset. If you use the estimated regression line for prediction using x_0 which lies *outside* the range of the available x data, then this is known as **extrapolation**, for which any predictions should be viewed with extreme caution.

For Example 12.7, it may not be immediately obvious that the relationship between advertising expenditure and sales could change, but a moment's thought should convince you that, were you to quadruple advertising expenditure, you would be unlikely to get a nearly 13-fold increase in sales! Basic economics would suggest diminishing marginal returns to advertising expenditure.

Sometimes it is very easy to see that the relationship must change. For instance, consider Example 12.8, which shows an anthropologist's figures on years of education of a mother and the number of children she has, based on a Pacific island.

Example 12.8 Figures from an anthropologist show a negative linear relationship between the number of years of education, x, of a mother and the number of live births she has, y. The estimated regression line is:

$$\hat{y} = 8 - 0.6x$$

based on figures of women with between 5 and 8 years of education who had 0 to 8 live births. This looks sensible. We predict $\hat{y} = 8 - 0.6(5) = 5$ live births for those with 5 years of education, and $\hat{y} = 8 - 0.6(10) = 2$ live 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 so had 15 years of education. She clearly cannot have $\hat{y} = 8 - 0.6(15) = -1$ children! And, if someone missed school entirely, is she really likely to have $\hat{y} = 8 - 0.6(0) = 8$ children? We have no way of knowing. The relationship shown by the existing figures will probably not hold beyond the x data range of 5 to 8 years of education.

12.10 Points to note about correlation and regression

As already mentioned, the Examiners frequently give you the following summary statistics:

$$\sum x_i, \quad \sum x_i^2, \quad \sum y_i, \quad \sum y_i^2, \quad \text{and} \quad \sum x_i y_i$$

in order to save you computation time. If you do not know how to take advantage of these, 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 highly unlikely to get any credit. This part of the syllabus leads directly into **EC2020 Elements of econometrics**, so it is important that you understand it if you are taking that course.

What is the relationship between correlation and regression?

As we have seen, all the calculations we do for correlation and regression involve similar summation measures, so they must be connected in some way. This is indeed the case.

For example, a large r^2 (the square of the sample correlation coefficient) 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 small r^2 means that the points are scattered (resulting in a large standard error of b).

While you are **not** expected to determine confidence intervals nor perform hypothesis tests for the parameters α, β and ρ (these are all part of **ST104b Statistics 2**), it is important that you are aware of these ideas. On the other hand, be sure you are clear that a large $|r|$ does not necessarily mean that the slope of the regression line is steep – a strong correlation can exist if the points are all close to a line with a moderate gradient.

Multiple correlation and regression

The ideas we 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. These techniques are discussed a little more in **ST104b Statistics 2**, and their use and limitations are part of the **MN3141 Principles of marketing** 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.

12.11 Summary

This brings us to the end of our introduction to dealing with the estimation of two variables (bivariate statistics). This chapter has been chiefly concerned with two important features when modelling the (linear) relationship between two variables: (i.) correlation – measuring the strength of a (linear) relationship, and (ii.) regression – a way of representing that linear relationship.

12.12 Key terms and concepts

- Causality
- Correlation coefficient
- Error term
- Independent (explanatory) variable
- Prediction
- Scatter diagram (plot)
- Spearman rank correlation
- Correlation
- Dependent (response) variable
- Extrapolation
- Non-linear relationship
- Regression
- Simple linear regression
- Uncorrelated

12.13 Learning activities

1. Sketch a scatter diagram for each of the following situations:
   (a) $r = 0$
   (b) r is very strong and positive
   (c) r is very strong and negative
   (d) r is quite strong, but negative.

2. Think of an example where you feel the correlation is clearly spurious (that is, there is correlation, but no causal connection) and explain how it might arise.

   Now think of a 'clear' correlation and the circumstances in which you might accept causality.

3. The following figures give examination and project results (in percentages) for eight students.
   (a) Find the Spearman's rank correlation coefficient, r_s, for the data.
   (b) Compare r_s with the sample correlation coefficient, r.

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

4. Work out b and a in Example 12.7 using advertising costs 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 Example 12.7!

5. Try to think of a likely linear relationship between x and y which would probably work over some of the data, but then break down like that in the anthropologist case in Example 12.8.

   This should make sure you understand the difference between interpolation (which statisticians do all the time) and extrapolation (which they should not).

6. Sketch a scatter diagram with the line of best fit for the following a, b and r.
   (a) $a = 2$, $b = 1/2$ and $r = 0.9$
   (b) $a = -3$, $b = -2$ and $r = -0.3$.

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

In addition, attempt the 'Test your understanding' self-test quizzes available on the VLE.

12.14 Further exercises

- Newbold, P., W.L. Carlson and B.M. Thorne *Statistics for Business and Economics*. (London: Prentice-Hall, 2012) eighth edition [ISBN 9780273767060]. Relevant exercises from Sections 1.6, 2.4, 11.1–11.3 and 14.6.

12.15 A reminder of your learning outcomes

After completing this chapter, and having completed the Essential reading and activities, you should be able to:

- draw and label a scatter diagram
- calculate r
- explain the meaning of a particular value and the general limitations of r and r^2 as measures
- calculate a and b for the line of best fit in a scatter diagram
- explain the relationship between b and r
- summarise the problems caused by extrapolation.

12.16 Sample examination questions

1. State whether the following statements are **true** or **false** and explain why.
   (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, 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 steep slope b.'

2. The following table shows the number of computers in thousands, x, produced by a company each month and the corresponding monthly costs in £000s, y, for running its computer maintenance department.

Number of computers (in thousands), x	Maintenance costs (in £000s), 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:

$$\sum_{i=1}^{10} x_i = 75.5, \qquad \sum_{i=1}^{10} y_i = 1080, \qquad \sum_{i=1}^{10} x_i y_i = 8184.9,$$

$$\sum_{i=1}^{10} x_i^2 = 573.33, \qquad \sum_{i=1}^{10} y_i^2 = 116988.$$

(a) Draw the scatter diagram.

(b) Calculate the sample correlation coefficient for computers and maintenance costs.

(c) Compute the best-fitting straight line for y and x.

(d) Comment on your results. How would you check on the strength of any relationship you have found?

Solutions to these questions can be found on the VLE in the **ST104a Statistics 1** area at http://my.londoninternational.ac.uk

Appendix A
Sample examination paper

Important note: This Sample examination paper reflects the examination and assessment arrangements for this course in the academic year 2014–15. The format and structure of the examination may have changed since the publication of this subject guide. You can find the most recent examination papers on the VLE where all changes to the format of the examination are posted.

Time allowed: 2 hours.

Candidates should answer **THREE** of the following **FOUR** questions: **QUESTION 1** of Section A (50 marks) and **TWO** questions from Section B (25 marks each). **Candidates are strongly advised to divide their time accordingly**.

A list of formulae and extracts from statistical tables are provided after the final question on this paper.*

Graph paper is provided at the end of this question paper. If used, it must be detached and fastened securely inside the answer book.

A calculator may be used when answering questions on this paper and it must comply in all respects with the specification given with your Admission Notice. The make and type of machine must be clearly stated on the front cover of the answer book.

* Note these are not included here.

Section A

Answer all **eight** parts of Question 1 (50 marks in total).

1. (a) Define each of the following briefly:

   i. Interviewer bias.

   ii. Cluster sampling.

   iii. Quota sampling.

(6 marks)

(b) Which of the following is the odd one out? Explain briefly.

Variance, mean, standard deviation, range.

(2 marks)

(c) In an examination, the scores of students who attend type A schools are normally distributed about a mean of 55 with a standard deviation of 6. The scores of students who attend type B schools are normally distributed about a mean of 60 with a standard deviation of 5. Which type of school would have a higher proportion of students with marks above 70? Explain your answer.

(5 marks)

(d) What does it mean for two variables to be uncorrelated? How is it possible for two variables to be strongly related, but still uncorrelated?

(3 marks)

(e) The owner of a fish market finds that the mean weight for a catfish is 3.2 pounds with a standard deviation of 0.8 pounds. Assume that the weights of the catfish are normally distributed.

i. What is the probability that the weight of a randomly selected catfish is greater than 4.8 pounds?

ii. Above what weight (in pounds) do 89.8% of the weights occur?

iii. You buy a sample of 25 catfish. What is the probability that the mean is less than 3 pounds?

(10 marks)

(f) Let $x_1 = 5$, $x_2 = 1$, $x_3 = 3$ and $x_4 = 1$. Find:

$$\text{i. } \sum_{i=3}^{4} x_i^3 \qquad \text{ii. } \sum_{i=2}^{4} (x_i - 3)^2 \qquad \text{iii. } \sum_{i=1}^{3} 4x_i.$$

(6 marks)

(g) Three fair coins are thrown (and throws are independent of each other).

i. Find the probability that exactly one is a head.

ii. You are told that at least one is a head. What is the probability that all are heads in this case?

(4 marks)

(h) State whether the following are possible or not. Give a brief explanation. (*Note that no marks will be awarded for a simple possible/not possible reply.*)

i. The correlation coefficient for the relationship between hours of revision and examination mark is -2.3.

ii. Quota sampling allows us to quantify bias and standard errors.

iii. It is possible to have a chi-squared value of -3.

iv. If the probability that it will rain tomorrow is $1/5$ and the probability that you will not wear a raincoat is given, then the probability that it rains and you find you have no raincoat is $1/4$.

(8 marks)

(i) A company has two machines which produce cupboards. 75% are produced by the new machine and the remaining 25% by the older machine. In addition, the new machine produces 8% defective cupboards. The old machine produces 23% defective cupboards.

i. What is the probability that a randomly chosen cupboard produced by the company is defective?

ii. Given that a randomly chosen cupboard has been chosen and found to be defective, what is the probability it was produced by the new machine?

(6 marks)

Section B

Answer **TWO** questions from Section B (25 marks each).

2. (a) Over a number of years, the demand for daily newspapers in a locality had been 20% for the Sun, 15% for the Star, 30% for the Times, 20% for the Examiner and 15% for the Independent. To determine whether demand has changed, the manager randomly records the papers purchased by 300 readers. The results are given below:

Paper	Sun	Star	Times	Examiner	Independent
No. of readers	65	40	75	90	30

Use the χ^2 test to determine whether demand for these papers has changed.

(10 marks)

(b) A survey is carried out to determine whether or not there is any association between age group and being in favour of working from home on two days per week. The raw results of the survey are as follows:

Age group	18–24	25–35	36–50	Over 50
In favour	20	62	40	40
Against	32	50	46	80

i. Test the claim that there is no association between age group and being in favour at the 5% significance level.

ii. If the null hypothesis in (i.) is rejected, identify the nature of the association.

(15 marks)

3. (a) Fill in the missing entries (??) in the following table:

H_0	??
H_1	$\mu > 50$
The decision rule (for $\alpha = 0.05$) is	??
Sample data:	$n = 36$, $\bar{x} = 56.5$, $s = 4.5$
The test statistic value is	??
Conclusion	??

(8 marks)

(b) Health advisory services state that 50% of heart attack patients die before reaching hospital – the chances of surviving a heart attack are dramatically increased by receiving medical help as quickly as possible. A survey of 240 heart attack victims discovered that men sought help more urgently than women.

	Number	Sample statistics of those reaching hospital after the initial attack (in hours)
Men	134	$\bar{x} = 3.40$, $s = 1.42$
Women	106	$\bar{x} = 8.24$, $s = 4.24$

i. Calculate a 95% confidence interval for the difference between the mean times to reach hospital for men and women.

ii. Is it possible to infer that there is a difference between the mean times taken to reach hospital from the data given above? Explain your answer.

(12 marks)

(c) According to the Quarterly National Household Survey, the mean weekly family income is estimated to be £1,250 with a standard deviation of £120. Assuming incomes are normally distributed, calculate the percentage of households whose weekly income is between £750 and £1,500.

(5 marks)

4. (a) The following data give the price in euros of a ten-minute call at 11:00 on a weekday for a local call. The prices refer to August each year. Normal tariffs without special rates are used.

	1997	1998	1999	2000	2001	2002	2003	2004
Spain	0.20	0.32	0.32	0.28	0.28	0.28	0.28	0.28
Ireland	0.58	0.58	0.49	0.51	0.51	0.51	0.51	0.49

i. Plot, on the same diagram, approximate lines for the price of a local call in Spain and Ireland from 1997 to 2004. Give a brief verbal comparison of the trends in prices in the two countries.

ii. Fit a least squares line for prices in Ireland over this period, using $x = 0$ for 1997, $x = 1$ for 1998, etc.

N.B. summary statistics for these data are:

$$\sum_i x_i = 28,\ \sum_i y_i = 4.18,\ \sum_i x_i^2 = 140,\ \sum_i y_i^2 = 2.1934,\ \sum_i x_i y_i = 14.17.$$

iii. Calculate the sample correlation coefficient.

iv. How appropriate do you think it is to fit a linear trend to these data?

(15 marks)

(b) You run a market research company and have been asked to research the likely demand for extra sports facilities in your city following a decision by the government to promote exercise for general health.

Outline your survey design, making it clear whether it is random or not, and giving appropriate design factors. You have been given generous funding and are expected to look at all aspects of the question and all groups of the population.

(10 marks)

[END OF PAPER]

Appendix B
Sample examination paper – Examiners' commentary

Section A

1. (a) **Reading for this question**

This question asks for brief definitions of various aspects of sample surveys. Candidates should ensure they are familiar with all aspects of sampling, as explained in Chapter 10. A good answer would not only describe each term, but also provide a potential remedy (in the case of interviewer bias) or an advantage/disadvantage (in the case of the sampling methods).

Approaching the question

i. Interviewer bias – opinion or prejudice on the part of the interviewer which occurs during the interview process, thus affecting outcomes, or it may be a consistent reaction to a particular type of interviewer on the part of the interviewee (for example, interviewers with an upper-class accent may make the interviewees all respond in a similar way). The problem is best avoided with well-trained and well-qualified interviewers to conduct interviews objectively.

ii. Cluster sampling – a form of *probability sampling* useful for interview surveys to reduce costs. Achieved by partitioning the target population into clusters, then sampling the desired number of clusters using simple random sampling. One-stage cluster sampling would then involve surveying every member of the chosen clusters. Costs (of travel) are only saved if the survey is conducted face-to-face (in contrast with mail or telephone surveys).

iii. Quota sampling – a form of *non-probability sampling* when no sampling frame is available. Commonly used for obtaining fast results, when you need to reduce costs and when accuracy is not too important. Bias can be caused by the choices made by interviewers when they fill their quota control allocations.

(b) **Reading for this question**

A sound knowledge of descriptive statistics is required here. Chapter 4 introduces measures of location and spread, so it should be easy to allocate each term to one of these categories.

Approaching the question

The odd one out is the mean (which is a measure of location), while all the other terms are measures of spread.

(c) **Reading for this question**

Examination questions involving the normal distribution (covered in Chapter 6) often require standardisation. Note the question does not ask for a probability, rather just which type of school has a higher proportion of students with marks above 70. Hence it is not strictly necessary to compute probabilities here, instead z values would be sufficient.

Approaching the question

Type A schools have a z score of $z_A = (70 - 55)/6 = 2.5$, while Type B schools have a z score of $z_B = (70 - 60)/5 = 2.0$. Since the proportion *greater* than 70 in Type A schools will be less than in Type B schools, due to the higher z score, then Type B schools have the higher proportion. This could be accompanied by a sketch of a (standard) normal distribution to illustrate the point.

Alternatively, candidates could work out the actual proportion of students achieving above 70: for Type A schools this is 0.00621, while for Type B schools this is 0.02275. Hence Type B schools have the higher proportion.

(d) **Reading for this question**

Once again, definitions are needed to answer the question. In (d) candidates need to understand the concept of correlation, which is the strength of a linear relationship between two variables. Chapter 12 gives details.

Approaching the question

'Uncorrelated' means no *linear* relationship between two variables and a correlation coefficient of zero. Two variables can be strongly related, yet still be uncorrelated, if they have a *non-linear* relationship.

(e) **Reading for this question**

Like part (c), this question also concerns the normal distribution (Chapter 6), although here actual probabilities are required which means candidates will need to use the *New Cambridge Statistical Tables*. Note that in (ii.) a form of reverse standardisation is required, and in (iii.) it is necessary to work with the sample mean, $\bar{X}$, rather than X.

Approaching the question

i. $P(X > 4.8) = P(Z > (4.8 - 3.2)/0.8) = P(Z > 2) = 0.02275$. Note that this is the same z value which appeared in (d)!

ii. $P(Z > z) = 0.898$. From Table 4, $z = -1.27$, hence:

$$P(Z > -1.27) = P\left(X > \frac{x - 3.2}{0.8}\right).$$

Therefore $(x - 3.2)/0.8 = -1.27$, hence $x = 2.184$.

iii. The sample mean $\bar{X} \sim N(\mu, \sigma^2/n)$, hence for $n = 25$, $\bar{X} \sim N(3.2, 0.0256)$. So:

$$P(\bar{X} < 3) = P\left(Z < \frac{3 - 3.2}{\sqrt{0.0256}}\right) = P(Z < -1.25) = 0.1056.$$

(f) **Reading for this question**

This question examines candidates' competence with the summation operator, as covered in Chapter 2. It is important to ensure the correct values are used for the index of summation, i. No marks would be awarded if i values where used instead of x_i values.

Approaching the question

i. $\sum_{i=3}^{4} x_i^3 = 3^3 + 1^3 = 28.$

ii. $\sum_{i=2}^{4} (x_i - 3)^2 = (1-3)^2 + (3-3)^2 + (1-3)^2 = 4 + 0 + 4 = 8.$

iii. $\sum_{i=1}^{3} 4x_i = (4 \times 5) + (4 \times 1) + (4 \times 3) = 36.$

(g) **Reading for this question**

Part (g) deals with probability (Chapter 5). There is a helpful hint in the question, i.e. that the three coins are independent, which means the multiplicative law can be used. Note, too, that the coins are fair, which means the probabilities of heads and tails are both equal to 0.5.

Approaching the question

i. Let H_i and T_i denote heads and tails for coin i, respectively, $i = 1, 2, 3$. For exactly one head we consider:

$$P(H_1 \cap T_2 \cap T_3) + P(T_1 \cap H_2 \cap T_3) + P(T_1 \cap T_2 \cap H_3) = 3 \times \left(\frac{1}{2}\right)^3 = \frac{3}{8}.$$

ii. This is a conditional probability problem. Let A denote 'at least one head' and B denote 'all heads'. First note that $P(B) = (1/2)^3 = 1/8$. Now since A^c is 'all tails', we have:

$$P(A) = 1 - P(A^c) = 1 - \left(\frac{1}{2}\right)^3 = \frac{7}{8}.$$

So we require $P(B \mid A) = P(A \cap B)/P(B)$, but $A \cap B = B$, so:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B)}{P(A)} = \frac{1/8}{7/8} = \frac{1}{7}.$$

(h) **Reading for this question**

Each of these statements requires candidates to go back to first principles, typically involving various definitions. Candidates are reminded of the need to give a brief explanation in their answer, rather than just answering 'True' or 'False' which would receive no marks.

Approaching the question

i. False. Correlation coefficient must be in the interval $[-1, 1]$.

ii. False. Quota sampling is a non-probability form of sampling meaning bias and standard errors cannot be measured.

iii. False. Chi-squared values cannot be negative.

iv. False. If $P(\text{rain}) = 1/5$ and $P(\text{rain} \cap \text{coat}) = 1/4$ where rain and taking coat are independent, then this implies $P(\text{coat}) = 5/4 > 1$, therefore this is impossible since probabilities cannot exceed 1.

(i) **Reading for this question**

This question involves probability calculations (see Chapter 5). In the first part, the total probability formula is used, and Bayes' formula is needed for the second part.

Approaching the question

i. Let D = defective. Hence:

$$P(D) = P(D \,|\, \text{New})P(\text{New}) + P(D \,|\, \text{Old})P(\text{Old}).$$

So, $P(D) = (0.08 \times 0.75) + (0.23 \times 0.25) = 0.1175$.

ii. Using Bayes' formula:

$$\begin{aligned} P(\text{New} \,|\, \text{D}) &= \frac{P(\text{D} \,|\, \text{New})P(\text{New})}{P(\text{D} \,|\, \text{New})P(\text{New}) + P(\text{D} \,|\, \text{Old})P(\text{Old})} \\ &= \frac{0.08 \times 0.75}{(0.08 \times 0.75) + (0.23 \times 0.25)} \\ &= 0.5106. \end{aligned}$$

Section B

2. (a) **Reading for this question**

This question requires candidates to perform a goodness-of-fit test using the historical demand values. Chapter 9 discusses such tests. Note the importance of testing at two significance levels, starting with 5%.

Approaching the question

First, state the hypotheses:

H_0: Demand for these papers has not changed

H_1: Demand for these papers has changed.

Next, compute the expected frequencies given historic demand:

- The Sun $= 0.2 \times 300 = 60$.
- The Star $= 0.15 \times 300 = 45$.
- The Times $= 0.3 \times 300 = 90$.
- The Examiner $= 0.2 \times 300 = 60$.

- The Independent $= 0.15 \times 300 = 45$ (or $300 - 255$).

Under H_0, the test statistic is:

$$\chi^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$$

and there are $5 - 1 = 4$ degrees of freedom. The test statistic value is 23.47.

For $\alpha = 0.05$, the critical value is 9.488, hence we reject H_0 (since $9.488 < 23.47$) and conclude demand has changed.

Setting $\alpha = 0.01$, the critical value is now 13.28, hence we again reject H_0 (since $13.28 < 23.147$) and can say that the test result is highly significant.

(b) **Reading for this question**

As with part (a), this question can be solved using a chi-squared test statistic, so again refer to Chapter 9. The difference here is that the expected frequencies will have to be computed using a common proportion.

Approaching the question

i. We wish to test:

H_0: There is no association between age group and being in favour of working from home.

H_1: There is association between age group and being in favour of working from home.

The following table shows the required calculations.

		18–24	25–35	36–50	Over 50	Total
In favour	O	20	62	40	40	162
	E	22.77	49.04	37.65	52.54	162
	$(O-E)^2/E$	0.34	3.43	0.15	2.99	
Against	O	32	50	46	80	208
	E	29.33	62.96	48.35	67.46	208
	$(O-E)^2/E$	0.26	2.67	0.11	2.33	
Total		52	112	86	120	370

The test statistic value is $0.34 + 3.43 + \cdots + 2.33 = 12.28$, on $(r-1)(c-1) = (2-1)(4-1) = 3$ degrees of freedom.

At the 5% significance level, the critical value is 7.815 (obtained from Table 8 of the *New Cambridge Statistical Tables*). Since $7.815 < 12.28$, we reject H_0 at the 5% significance level. Moving to the 1% significance level, the critical value is now 11.34. Since $11.34 < 12.28$, we again reject H_0.

We conclude that the test is highly significant as there is strong evidence of an association between age group and being in favour of working from home.

ii. In order to identify the nature of the association, we seek the large contributors to the (highly significant) test statistic value. These are the entries of 3.43, 2.99, 2.67 and 2.33. Comparing the observed and expected frequencies, it is clear that the 25–35 age group tends to be in favour of

working from home, while the over 50 age group tends to be against working from home.

3. (a) **Reading for this question**

The purpose of completing the missing table entries is for candidates to demonstrate their ability to recognise the core components of a hypothesis test of a single population mean. Such tests are covered in Chapter 8.

Approaching the question

Table entries:

H_0	$\mu = 50$
H_1	$\mu > 50$
The decision rule (for $\alpha = 0.05$) is	Reject H_0 if the test statistic value is > 1.6896 i.e. this is a one-tailed t-test (note s is given, not σ) with $n - 1 = 35$ degrees of freedom
Sample data	Sample size, $n = 36$; $\bar{x} = 56.5$, $s = 4.5$
The test statistic value is	$t = (\bar{x} - \mu)/(s/\sqrt{n}) = (56.5 - 50)/(4.5/\sqrt{36}) = 8.67$
Conclusion	Reject H_0, hence sufficient evidence to suggest $\mu > 50$

(b) **Reading for this question**

Here a 95% confidence interval for the difference in two means is required. Chapter 7 covers all the types of confidence intervals which may be asked in this course. Note since the sample standard deviations for men and women are very different (and we have large sample sizes), it would be inappropriate to use a pooled variance here, so we estimate the true variances with the sample variances.

Approaching the question

First, calculate the difference between the sample means, which is $8.24 - 3.40 = 4.84$. The confidence interval endpoints are given by:

$$(\bar{x}_2 - \bar{x_1}) \pm z \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

so, substituting in the data values:

$$4.84 \pm 1.96 \cdot \sqrt{\frac{1.42^2}{134} + \frac{4.24^2}{106}}.$$

Finally, evaluate the endpoints and report the confidence interval as:

$$(3.998,\ 5.682).$$

Since the confidence interval does not include zero, it is likely that there is a difference between the times taken to reach hospital for men and women, given the sample data. A formal hypothesis test need not be carried out (and no marks would be awarded if actually provided), since we are 95% confident that the true *absolute* average difference in population means is in the interval (3.998, 5.682).

(c) **Reading for this question**

This is a straightforward application of the standardisation of a normal random variable. See Chapter 6 for further examples.

Approaching the question

Let X be weekly income, hence $X \sim N(1250, (120)^2)$ and we require $P(750 \leq X \leq 1500)$. We have:

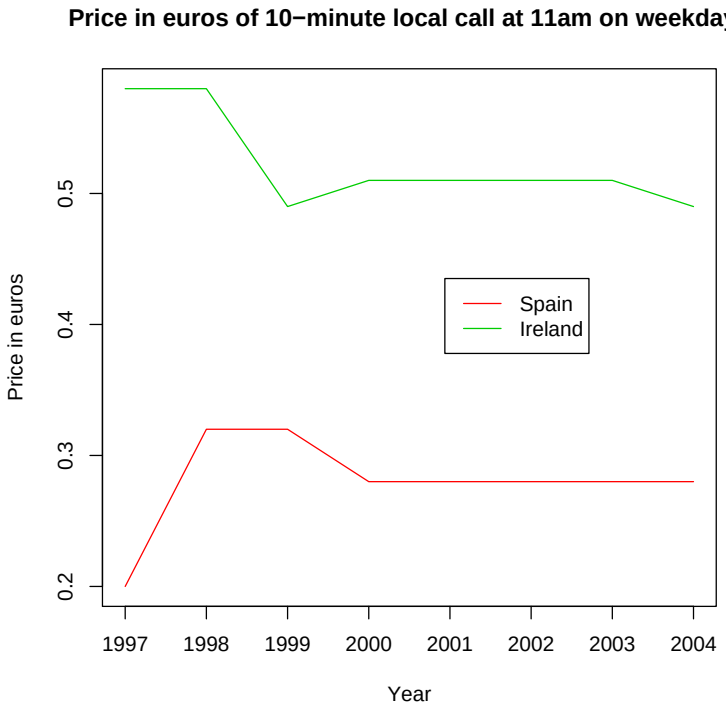
$$\begin{aligned} P(750 \leq X \leq 1500) &= P\left(\frac{750-1250}{120} \leq Z \leq \frac{1500-1250}{120}\right) \\ &= P(-4.167 \leq Z \leq 2.083) \\ &= \Phi(2.08) - \Phi(-4.17) \\ &= \Phi(2.08) - (1 - \Phi(4.17)) \\ &= 0.9814 - (1 - 1) \\ &= 0.9814. \end{aligned}$$

4. (a) **Reading for this question**

Part (a) examines the topics of correlation and regression, which can be found in Chapter 12. The first task asks candidates to construct a diagram of these time series data. So remember to adequately label the diagram. The regression part requires the estimation of the parameters of the simple linear regression model. Finally, candidates need to interpret the value of r in relation to the suitability of a linear regression model.

Approaching the question

i. The diagram should look like the following. Note the informative title, key indicating each country and axis labels.



We see that Ireland's cost of local phone calls has steadily decreased since 1997, although it is clearly more expensive than Spain throughout the time period for which data are provided.

ii. Calculation of least squares regression line: $\hat{y} = a + bx$, where:

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x^2 - n\bar{x}^2} = -0.01095238$$

$$a = \bar{y} - b\bar{x} = 0.5608333.$$

Hence the estimated regression line is, rounding appropriately:

$$\hat{y} = 0.561 - 0.011x.$$

iii. The sample correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}} = -0.73.$$

iv. $r = -0.73$ suggests a fairly strong, negative correlation between time and price of local calls in Ireland. This implies that it is indeed appropriate to fit a linear regression line to these data (and is consistent with the negative trend).

(b) **Reading for this question**

This is a typical examination question which asks candidates to devise a sample survey. Such questions draw on all the material in Chapter 10. Note the importance of specifying whether or not the survey is random.

Approaching the question

The following are suggested points to make in an answer. These are not necessarily exhaustive. The Examiners will give credit to any sensible (and justified) ideas.

- Generous funding is specified, so a random survey design should be used, and non-probability methods should be rejected.
- Candidates should specify that each unit in a random sample has a known (although not necessarily equal) probability of being selected.
- Mention of sampling frame/list, for example electoral register, schools (for children), national list of addresses, etc.
- Recommend using a pilot survey – for example to assess clarity of questions.
- Question asks to look at all groups of the population, hence candidates should define stratification and give examples of sensible strata, for example socio-economic group, age, etc. with a justification for each.

- For large-scale random surveys, cluster/multistage sampling could also be considered.
- Discussion of method of contact, for example face-to-face interview, telephone survey, postal/online questionnaires with discussion of advantages and disadvantages.
- Potential questions to include in the survey relevant to researching 'demand for extra sports facilities'.

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 International Programmes.
If you have any comments about this guide, either general or specific (including corrections, non-availability of Essential readings, etc.), please take the time to complete and return this form.

Title of this subject guide: .....

.....

Name .....

Address .....

.....

Email .....

Student number .....

For which qualification are you studying? .....

Comments

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Please continue on additional sheets if necessary.

Date: .....

Please send your completed form (or a photocopy of it) to:
Publishing Manager, Publications Office, University of London International Programmes,
Stewart House, 32 Russell Square, London WC1B 5DN, UK.