Introducing STATISTICS

2nd EDITION

GRAHAM UPTON
IAN COOK

REVISED FOR NEW A LEVEL SPECIFICATION

OXFORD

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide in

Oxford New York

Auckland Bangkok Buenos Aires Cape Town Chennai Dar es Salaam Delhi Hong Kong Istanbul Karachi Kolkata Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi São Paolo Shanghai Taipei Tokyo Toronto

Oxford is a registered trade mark of Oxford University Press in the UK and in certain other countries

First published 1998

Reprinted (with corrections) 1999

Second edition 2001

British Library Cataloguing in Publication Data

Data available

ISBN 0 19 914 801 5

10 9 8 7 6 5 4

Typeset by Tech Set Ltd, Gateshead, Tyne and Wear
Printed and bound in Great Britain by Bell & Bain Ltd.

Contents

Preface to the Second Edition

Early Statistics syllabuses presented the subject as a branch of mathematics. The emphasis was on formulae, with, for example, questions on artificial continuous distributions being an excuse to practise integration.

In reality, Statistics is about the interpretation of data and the modelling of the processes that have given rise to those data. Modern syllabuses (or 'specifications' as they are now sometimes called) increasingly reflect the need to do more than simply summarise data. The first edition of *Introducing Statistics* emphasised the need to draw inferences and in this second edition we have added further inferential material.

Amongst the additions to Chapter 1 are several examples of the graphical comparison of similar data sets. This chapter includes five new sections and ends with a discussion of the (largely unwanted) characteristics to be expected in real data. Chapter 2 has been augmented by sections on the use of coded values, Bayes' theorem is included in Chapter 4, and the method for determining the distribution of a simple function of a random variable is now included in Chapter 9. In Chapter 14 there is a new section dealing with properties of regression line estimators and, later, a subsection on nonsense correlation.

We have also taken the opportunity to introduce some questions on sampling for Chapter 3. These questions are somewhat open-ended, as were a number of existing questions for which we did not give answers in the first edition. In this edition we now provide possible answers to these questions, though no answers to such questions should be regarded as prescriptive.

This book covers all the material in the statistics modules of the EDEXCEL, AQA, OCR, WJEC and NICCEA syllabuses for A-level Mathematics and in the corresponding syllabuses set by Cambridge International Examinations. Although it is still possible for a student to study an A-level Mathematics syllabus that contains no Statistics, for some boards Statistics can amount to half the syllabus. Since there are substantial differences between the syllabuses, this book will also cover much of the material in A-level Further Mathematics, and AS-level Statistics. Students studying A-level Statistics, or requiring a book for a first course at university level should use the companion volume *Understanding Statistics*.

GJGU
ITC
University of Essex
Colchester
June 2000

Preface to the First Edition

Long long ago, when the authors were at school, the subject Statistics was almost unheard of. Few universities had specialist Statistics teachers and there was little if any Statistics taught in schools — those were certainly not the 'good old days' so far as Statistics was concerned. Since that time there has been a continuing expansion in the teaching of the subject as its relevance to everyday life, the conduct of research, and government has become increasingly appreciated.

This book concentrates on the fundamentals of the subject. To determine its contents we took careful note of the statistics sections of the current single subject A-level syllabuses. This book covers the union of those syllabuses. Of course, syllabuses seem ever changing, but this is not a problem for us since all of them must begin in very much the same way in dealing with these fundamentals. This book will therefore be suitable for a wide audience.

The detailed list of contents is given in the next few pages. A summary is as follows: Chapters 1 to 3 describe the basic mechanics of collecting, displaying and summarising data; Chapters 4 to 12 develop the common probability-based models (binomial, Poisson, normal) and use them to draw conclusions about the properties of large populations on the basis of information from small samples; Chapter 13 introduces a simple method for checking model validity; Chapter 14 provides an introduction to the study of relationships between variables.

A comparison of this book (*IS*, for short) with our other volume *Understanding Statistics* (*US*) will immediately reveal that *IS* is shorter (about two-thirds the size) and that there is nothing in *IS* that does not appear in *US*. The economy in size has not been accomplished by curtailing explanations, nor by reducing the large numbers of worked examples and set questions on the topics treated. Instead, the reduction has been effected by omitting the more advanced or esoteric sections to leave only the essentials. Our hope is that a student using *IS* will be enthused by the material and, like Oliver Twist, will ask for more! In this case 'more' will be readily to hand in the shape of *US*, with its corresponding format.

In Statistics, as in the rest of Mathematics, the best way to learn is by doing questions. For that reason we have included hundreds of questions, with answers, in the book. We are very grateful to the examining boards listed below for permission to reproduce their questions. The source of each question is indicated by the corresponding initials at the end of the question. The numerical answers given in the book are, of course, our responsibility and any errors (we hope none!) are due to us and not to the examining boards. Where only part of a question has been used this is indicated by (P) after the attribution.

Associated Examining Board [AEB]
Northern Examination and Assessment Board [NEAB], formerly the Joint Matriculation Board [JMB]
Oxford and Cambridge Schools Examination Board [O&C], which also gave permission to reproduce questions from the examinations for the Mathematics in Education and Industry Project [MEI] and the School Mathematics Project [SMP]
University of Cambridge Local Examinations Syndicate [UCLES], which gave permission to reproduce questions from the University of Oxford Delegacy for Local Examinations (UODLE).
London Examinations, a division of Edexcel Foundation, formerly the University of London Examinations and Assessment Council [ULEAC] and the University of London School Examinations Board [ULSEB]
Welsh Joint Education Committee [WJEC]

All that remains is to wish you, the reader, pleasure in the use of this book. We hope that you too may be bitten by the Statistics bug.

GJGU
ITC
University of Essex
Colchester
October 1997

Glossary of notation

Inevitably some letters are used to denote different quantities in different contexts. However, this should cause no confusion since the context will make clear which definition is appropriate.

∞	Infinity.
$\sim$	'has distribution'.
$\approx$	'approximately equals'.
$\bar{}$	'mean'.
$\hat{}$	'estimate'.
$'$	'complementary' (of events).
$\mid$	'conditional', so $A \mid B$ means 'A occurs given that B occurs'.
!	Factorial: $r! = r(r-1)(r-2)\cdots 1$.
$\cap$	The intersection of two events.
$\sum$	Summation: $\sum_{i=1}^{n} x_i = x_1 + \cdots + x_n$.
$\cup$	The union of two events.
a	Intercept of (estimated) regression line.
α (alpha)	Intercept of population regression line.
b	Slope of (estimated) regression line.
β (beta)	Slope of population regression line.
$\mathrm{B}(n, p)$	The binomial distribution with parameters n and p.
c	A critical value.
cdf	The cumulative distribution function.
χ^2_ν	Chi-squared distribution with ν degrees of freedom.
d_i	A difference between ranks.
e	2.718 281 828 ....
$\mathrm{E}(X)$	Expected value or expectation of X.
E	An event.
E'	The complementary event to E.
E_i	An expected frequency.
f_j	The frequency with which the value x_j occurs.
$\mathrm{f}(x)$	The probability density function of X.
$\mathrm{F}(x)$	The cumulative distribution function of X.
$\mathbf{H}_0$, $\mathbf{H}_1$	The null and alternative hypotheses.
λ (lambda)	The parameter of a Poisson distribution.
m	The median.
μ (mu)	The population mean.
n	The number of observations.
$n(E)$	The number of outcomes in the event E.
$\binom{n}{r}$	The number of distinct combinations of r objects chosen from n.
${}^nP_r$	The number of distinct permutations of r objects chosen from n.
$\mathrm{N}(\mu, \sigma^2)$	Normal distribution, mean μ, variance σ^2.
ν (nu)	Degrees of freedom of t or χ^2 distributions
O_i	An observed frequency.
p, P	Probability.
P_x	$\mathrm{P}(X = x)$
$\mathrm{P}(z)$	Used in place of $\Phi(z)$ by some tables.
pdf	Probability density function.
ϕ (phi)	Probability density function for $\mathrm{N}(0, 1)$.
Φ (capital phi)	Cumulative distribution function for $\mathrm{N}(0, 1)$.
π (pi)	3.141 592 653 59 ....
q	$1 - p$, often the probability of a 'failure'.
Q_1, Q_2, Q_3	The lower quartile, median and upper quartile.
$\mathrm{Q}(z)$	Used by some tables to mean $1 - \Phi(z)$.
ρ (rho)	Population correlation coefficient.
r	The number of successes.
r	Product–moment correlation coefficient.
r_s	Spearman's rank correlation coefficient.
s	The square root of s^2.
s^2	$(= \sigma^2_{n-1})$ An unbiased estimate of the population variance.
s_p^2	The pooled estimate of the population variance.
S	The sample space.
S^2	The random variable corresponding to s^2.
S_{xx}, S_{xy}, S_{yy}	$\sum(x_i - \bar{x})^2$, $\sum(x_i - \bar{x})(y_i - \bar{y})$, $\sum(y_i - \bar{y})^2$
σ (sigma),	The population standard deviation.
σ^2	The population variance.
σ_n	The sample standard deviation.
σ_n^2	The sample variance.
σ^2_{n-1}	$(= s^2)$ An unbiased estimate of the population variance.
t_ν	A t-distribution with ν degrees of freedom.
T	A random variable having a t-distribution.
$\mathrm{Var}(X)$	The variance of X.
$\bar{x}$, $\bar{X}$	The sample mean (value or random variable).
x_i	An observed value.
$X, Y, \ldots$	Random variables.
X^2	The goodness-of-fit statistic.
X_c^2	The Yates-corrected version of X^2.
Z	A random variable having a $\mathrm{N}(0,1)$ distribution.
z	A test statistic: an observation on Z.

1 Summary diagrams and tables

She may look at it because it has pictures

Florence Nightingale, on a book of statistics that she had sent to Queen Victoria

One picture is worth ten thousand words

Frederick R Barnard

1.1 The purpose of Statistics

In most countries the biggest employer of statisticians is the government, which collects numerical information about all aspects of life. The information collected in the form of human statistics (such as the numbers out of work), of financial statistics (such as the rate of inflation), and on other aspects of life is regularly reported in newspapers and on the news. In addition to these **population** statistics, **sample** statistics are also reported. For example, market research agencies (e.g. Gallup) also collect numerical information which can dominate the news in advance of a general election.

As an example, a single issue of *The Times* contained the following:

- Drink–drive statistics (*Source:* The Government).
- Mothers' feelings about going back to work (sample statistics – *Source:* Gallup).
- Numbers of visitors to Britain subdivided by nationality (sample statistics – *Source:* International Passenger Survey).
- A breakdown of British Rail assets (*Source:* British Rail annual report).
- Pages of statistics on stocks and shares.
- Much more interesting pages of sports statistics!
- World weather statistics (*Source:* The Meteorological Office).

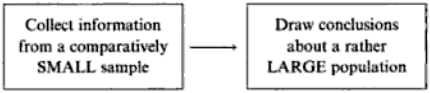
In the modern world we are inundated with statistics; the subject Statistics is concerned with trying to make sense of all this numerical information.

Project

Choose a single issue of a 'quality' newspaper and search for reports that include statistics. Try to decide what type of organisation collected the reported statistics.
Counting just one for all the sports reports, one for all the financial reports and one for all the weather information, how many different reports can you find in a single issue of the paper?
How many different organisations appear to have collected the statistics?

A good example of the purpose of Statistics is provided by the opinion poll. A poll is taken of a few thousand people. From the information that these

people provide, remarkably accurate conclusions are drawn that refer to the entire population, which is many times greater in number.



1.2 Variables and observations

The term **variable** refers to the description of the quantity being measured, and the term **observed value** or **observation** is used for the result of the measurement. Some examples are:

Variable	*Observed value*
Weight of a person	80 kg
Speed of a car	70 mph
Number of letters in a letter box	23
Colour of a postage stamp	Tyrian plum

If the value of a variable is the result of a random observation or experiment (e.g. the roll of a die), then the variable is called a **random variable**.

1.3 Types of data

The word 'data' is the plural of 'datum', which means a piece of information – so **data** are pieces of information. There are three common types of data: qualitative, discrete and continuous.

Qualitative data (also referred to as **nominal** data or **categorical** data) consist of descriptions using *names*. For example:

'Head' or 'Tail'
'Black' or 'White'
'Bungalow', 'House' or 'Castle'

Discrete data consist of numerical values in cases where we can make a list of the possible values. Often the list is very short:

1, 2, 3, 4, 5 and 6.

Sometimes the list will be infinitely long, as for example, the list:

0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, ...

Continuous data consist of numerical values in cases where it is not possible to make a list of the outcomes. Examples are measurements of physical quantities such as weight, height and time.

Note

- The distinction between discrete and continuous data is often blurred by the limitations of our measuring instruments. For example, when measuring people, we may record their heights to the nearest centimetre, in which case observations of a continuous quantity are being recorded using discrete values.

1.4 Tally charts and frequency distributions

Given below are the scores made in their final round by the 30 leading golfers in the Scottish Open golf championship:

62, 65, 63, 65, 70, 68, 65, 67, 67, 69, 70, 70, 70, 67, 68,
68, 66, 69, 74, 67, 68, 69, 69, 69, 71, 71, 72, 69, 72, 68

The data are discrete since the scores are all whole numbers. It is easy to see that most scores are around 70 or a little less. But it is not so easy to see which score was most common. A simple summary is provided by a **tally chart**.

The tally chart is constructed on a single 'pass' through the data. For each score a vertical stroke is entered on the appropriate row, with a diagonal stroke being used to complete each group of five strokes. This is much easier than going through the data counting the number of occurrences of a 62 and then repeating this for each individual score.

Notes

- Counting the tallies is made easy by using the 'five-bar gates'.
- If the tallies are equally spaced then the chart provides a useful graphical representation of the data.

Tally chart of the scores made in their final round by the 30 leading golfers in the Scottish Open

Score	Tallies
62	|
63	|
64	
65	𝍸
66	|
67	𝍸
68	𝍸
69	𝍸 |
70	𝍸
71	||
72	||
73	
74	|

The tally count for each outcome is called the **frequency** of that outcome. For example, the frequency of the outcome 65 was 3. The set of outcomes with their corresponding frequencies is called a **frequency distribution**, which can be displayed in a **frequency table**, as illustrated below:

Final round score	62	63	64	65	66	67	68	69	70	71	72	73	74
Number of golfers	1	1	0	3	1	4	5	6	4	2	2	0	1

1.5 Stem-and-leaf diagrams

Tally charts become uncomfortably long if the range of possible values is very large, as with these individual scores from a low-scoring Sunday league cricket match:

22, 58, 12, 17, 4, 7, 26, 10, 13, 1, 39, 0, 1, 10, 6, 0, 11, 14, 1, 0

A convenient alternative is the **stem-and-leaf diagram** (also called a **stemplot**), in which the stem represents the most significant digit (i.e. the 'tens') and the leaves are the less significant digits (the 'units'). The following stem-and-leaf chart has been created following the order of the data:

0	4, 7, 1, 0, 1, 6, 0, 1, 0
1	2, 7, 0, 3, 0, 1, 4
2	2, 6
3	9
4	
5	8
tens	*units*

If the original stem-and-leaf diagram had been created on rough paper then a tidied version could have the leaves neatly ordered as shown below:

0	0, 0, 0, 1, 1, 1, 4, 6, 7
1	0, 0, 1, 2, 3, 4, 7
2	2, 6
3	9
4	
5	8
tens	*units*

These charts are sometimes presented with 'split' stems (for finer detail). This is illustrated below, with the units between 0 and 4 (inclusive) separated from the units between 5 and 9 (inclusive):

0	0, 0, 0, 1, 1, 1, 4
0	6, 7
1	0, 0, 1, 2, 3, 4
1	7
2	2
2	6
3	
3	9
4	
4	
5	
5	8
tens	*units*

It is now particularly easy to see that most players scored less than 15 and that the highest score of 58 was a long way clear of the rest.

Stem-and-leaf diagrams retain the original data information, but present it in a compact and more easily understandable way: this is the hallmark of an efficient data summary.

Stem-and-leaf diagrams can be used both with discrete data and with continuous data (treating the latter as though it were discrete). They are much easier to understand when the stem involves a power of ten, but other units may be employed if the stem would otherwise be too long or too short. It is often wise to provide an explanation (a **Key**) with the diagram.

Example 1

The internal phone numbers of a random selection of individuals from a large organisation are given below.
Summarise these numbers using a stem-and-leaf diagram.

3315, 3301, 2205, 2865, 2608, 2886, 2527, 3144, 2154, 2645, 3703, 2610, 2768, 3699, 2345, 2160, 2603, 2054, 2302, 2997, 3794, 3053, 3001, 2247, 3402, 2744, 3040, 2459, 3699, 3008, 3062, 2887, 2215, 2213, 3310, 2508, 2530, 2987, 3699, 3298, 2021, 3323, 2329, 2845, 2247, 3196, 3412, 2021

A quick glance at the data reveals that all the numbers begin with either a 2 or a 3, implying that they all lie between 2000 and 3999 (inclusive). Taking a stem with units of 100 would lead to a large diagram: instead,

therefore, we work with units of 200, so that the leaves range between 0 and 199, inclusive. In this case the number 3315 is represented as a stem of '3200' and a leaf of '115' (so that $3200 + 115 = 3315$). The resulting diagram (with unordered leaves) is as follows:

2000	154, 160, 54, 21, 21
2200	5, 145, 102, 47, 15, 13, 129, 47
2400	127, 59, 108, 130
2600	8, 45, 10, 168, 3, 144
2800	65, 86, 197, 87, 187, 45
3000	144, 53, 1, 40, 8, 62, 196
3200	115, 101, 110, 98, 123
3400	2, 12
3600	103, 99, 194, 99, 99

Key:
3200 | 115 = 3315

The stem could also be labelled '20', '22', etc, with the caption 'hundreds'.

Example 2

The masses (in g) of a random sample of 20 sweets were as follows:

1.13, 0.72, 0.91, 1.44, 1.03, 1.39, 0.88, 0.99, 0.73, 0.91,
0.98, 1.21, 0.79, 1.14, 1.19, 1.08, 0.94, 1.06, 1.11, 1.01

Summarise these results using a stem-and-leaf diagram.

A quick scan reveals that the masses are all in the region of 1 g, so that an appropriate choice would be multiples of 0.1 for the stem and multiples of 0.01 for the leaves.

0.7	2, 3, 9
0.8	8
0.9	1, 9, 1, 8, 4
1.0	3, 8, 6, 1
1.1	3, 4, 9, 1
1.2	1
1.3	9
1.4	4

Key:
0.7 | 2 = 0.72

Example 3

The ages of the patients in one wing of a hospital were as follows:

Males 24, 56, 71, 88, 55, 73, 32, 59, 66, 60, 90, 42, 77
Females 40, 59, 93, 77, 86, 82, 60, 35, 76, 82, 84, 37, 61

Summarise the data using a back-to-back stem-and-leaf diagram.

In this case we have a central stem with leaves on either side depending on the gender of the patient. After ordering, we obtain

Males		*Females*
4	20	
2	30	5, 7
2	40	0
9, 6, 5	50	9
6, 0	60	0, 1
7, 3, 1	70	6, 7
8	80	2, 2, 4, 6
0	90	3

Key: 2 | 80 = 82 = 80 | 2

The older patients are predominantly female.

Exercises 1a

1 The numbers of absentees in a class over a period of 24 days were:

0, 3, 1, 2, 1, 0, 4, 0, 1, 1, 2, 3,
1, 0, 0, 2, 4, 6, 4, 2, 1, 0, 1, 1

By first drawing up a tally chart obtain a frequency table.

2 A bridge player keeps a note of the numbers of aces that she receives in successive deals. The numbers are:

0, 2, 3, 0, 0, 2, 1, 1, 0,
2, 3, 0, 1, 1, 2, 1, 0, 0

Draw up a tally chart and hence obtain a frequency table.

3 The numbers of eggs laid each day by 8 hens over a period of 21 days were:

6, 7, 8, 6, 5, 8, 6, 8, 6, 5, 6,
4, 7, 6, 8, 7, 5, 7, 6, 7, 5

Draw up a tally chart and hence obtain a frequency table.

4 For each potato plant, a gardener counts the numbers of potatoes whose mass exceeds 100 g. The results are:

8, 5, 7, 10, 8, 6, 5, 6, 4, 8,
10, 9, 8, 7, 3, 10, 11, 6, 9, 8

Obtain a frequency table.

5 A choirmaster keeps a record of the numbers turning up for choir practice. The numbers were:

25, 28, 32, 31, 31, 34, 28, 31, 29,
28, 32, 32, 30, 29, 29, 31, 28, 28

Obtain a frequency table.

6 The numbers of matches in a box were counted for a sample of 25 boxes. The results were:

51, 52, 48, 53, 47, 48, 50, 51, 50,
46, 52, 53, 51, 48, 49, 52, 50,
48, 47, 53, 54, 51, 49, 47, 51

Obtain a frequency table.

7 The marks obtained in a mathematics test marked out of 50 were:

35, 42, 31, 27, 48, 50, 24, 27,
21, 37, 41, 34, 12, 18, 27

Construct a stem-and-leaf diagram to represent the data.

8 A baker kept a count of the number of doughnuts sold each day.
The numbers were:

35, 47, 34, 46, 62, 41, 35, 47, 51,
56, 73, 38, 41, 44, 51, 45, 74

Construct a stem-and-leaf diagram to show the data.

9 The total scores in a series of basketball matches were:

215, 224, 182, 200, 229, 219,
209, 217, 195, 162, 210, 213,
204, 208, 197, 192, 187, 213

Construct a stem-and-leaf diagram to represent the above data.

10 The masses (in g) of a random collection of 16 pebbles are as follows:

17.4, 32.1, 24.4, 37.6, 51.0, 41.4,
19.9, 36.2, 41.3, 50.2, 37.7, 28.4,
26.3, 22.2, 33.5, 42.4

Summarise these data using a stem-and-leaf diagram.

11 An experiment to discover the movement of antibiotics in a certain variety of broad bean plants was made using 10 cut shoots and 10 rooted plants. The specimens were all immersed in a solution containing a certain chemical. After 18 hours the amounts of the chemical in a section of the lowest leaf of each specimen were as shown below.

Cut shoots	55	61	57	60	52
Rooted plants	53	50	43	46	35
Cut shoots	65	48	58	68	63
Rooted plants	48	39	44	56	51

Display these data using a back-to-back stem-and-leaf diagram. State your conclusions concerning any difference between the two sets of results.

12 Twenty patients were treated with a drug that relieves arthritis. However, eight of the patients suffered an adverse reaction. These patients were aged 44, 51, 64, 33, 39, 37, 41 and 72. The patients that suffered no reaction were aged 53, 29, 53, 67, 54, 57, 51, 68, 38, 44, 63 and 53. Display these data using a back-to-back stem-and-leaf diagram. Does the adverse reaction appear to be age-related?

1.6 Bar charts

The lengths of the rows of a tally chart or of a stem-and-leaf diagram provide an instant picture of the data. This picture is neatened by using bars whose lengths are proportional to the numbers of observations of each outcome (i.e. to the frequencies). In the resulting diagram, known as a **bar chart**, the bars may be either **horizontal** (like the tally chart) or **vertical**.

Notes

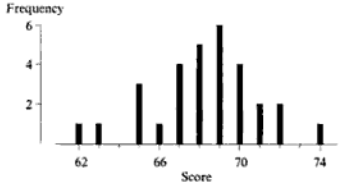
- Bar charts are easier to read if the width of the bars is different from the width of the gaps between the bars.
- It is not necessary to show the origin on the graph.
- A bar chart in which the bars are simply lines is called a **line graph**.

Example 4

Illustrate the golf scores of Section 1.4 (p.3) using a bar chart.

With lots of different values we use narrow bars centred on the values 62, 63, etc. The origin does not appear!

Vertical bar chart of the scores made in their final round by the 30 leading golfers in the Scottish Open



Example 5

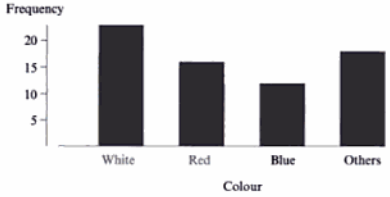
A car salesman is interested in the colour preferences of his customers. For one type of car his records are as follows.

Blue	White	Red	Others
12	23	16	18

Represent these figures using a vertical bar chart.

With just four categories narrow bars would look silly! We therefore use wide bars separated by narrower gaps. The categories are not numerical so they could be arranged in any order. A sensible order is to arrange the single colours in descending order of observed frequency, ending with the 'Others' category.

Bar chart of sales of cars of different colours



Practical

Roll an ordinary six-sided die 24 times, recording each outcome as it occurs (e.g. 3, 6, 2, 2, ...). Summarise the data using a tally chart and write down your frequency distribution. Compare your distribution with that of a neighbour. There may be large differences due to random variation! Combine the two sets of results and illustrate them with a vertical bar chart. Does it look as though your dice were fair?

Calculator practice

Graphical calculators can produce crude bar graphs. These are good enough to provide an idea of the data, but fail to indicate that discrete x-values are involved. Produce a diagram for the golf data in Section 1.4 on your calculator and compare it with our diagram.

1.7 Multiple bar charts

When data occur naturally in groups and the aim is to contrast the variations within different groups, a **multiple bar chart** may be used. This consists of groups of two or more adjacent bars separated from the next group by a gap having, ideally, a different width to the bars themselves.

The diagram may be horizontal or vertical with the values either specified on the diagram or indicated using a standard axis.

If there are two groups of data which refer to discrete variables then it is often effective to show the data on two separate bar charts, so as to highlight the differences in the distributions of the values. In this case care should be taken to use the same scales for each chart.

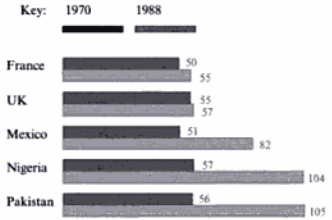
Example 6

The following data, taken from the *Monthly Bulletin of Statistics* published by the United Nations, show the 1970 and 1988 estimated populations (in millions) for five countries.
Illustrate the data using a multiple bar chart.

	France	Mexico	Nigeria	Pakistan	UK
1970	50	51	57	56	55
1988	55	82	104	105	57

The data show the differing rates of population growth of the two European countries and the three non-European countries and provide a graphic (literally!) illustration of a world problem. To increase visibility the countries are re-ordered in terms of their 1988 populations.

Populations of five countries in 1970 and in 1988 (figures are in millions)



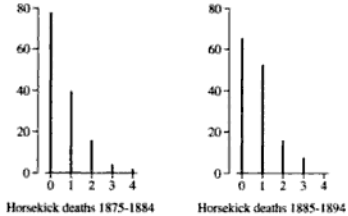
Example 7

In the late 19th century the Prussian army kept meticulous records of the numbers of their cavalrymen, in the 14 cavalry corps, who died as a consequence of a horse kick! The figures for 1875–1884 and 1885–1894 are given in the table

	Number of deaths				
	0	1	2	3	4
1875–1884	78	40	16	4	2
1885–1894	66	51	16	7	0

Since the data are discrete, bar charts are appropriate. The interest is in the possibility of a change in the distribution of deaths over time. The separate bar charts have the same vertical scales to permit comparison. There does not appear to be convincing evidence of any change.

Same-scale bar charts of the numbers of deaths of Prussian cavalrymen in two ten-year periods



1.8 Compound bars for proportions

In a compound bar chart the length of a complete bar signifies 100% of the population. The bar is subdivided into sections that show the relative sizes of components of the populations. By comparing the sizes of the subdivisions of two parallel compound bars, differences can be seen between the compositions of the separate populations. The populations need not be populations of living creatures – they could be, for example, the populations of nails in two builders' trucks!

Example 8

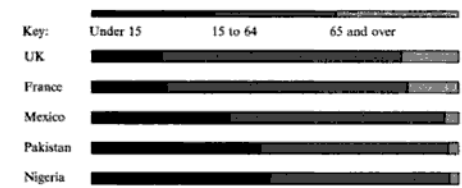
One consequence of the dramatic growth in population of the 'third world' countries is that a high proportion of the population of these countries is young and there are few old people. The United Nations publication *World Population Prospects* gives the following figures for 1990 populations:

	France	Mexico	Nigeria	Pakistan	UK
% under 15	20.2	37.2	48.4	45.7	18.9
% 15 to 64	66.0	59.0	49.2	51.6	65.6
% 65 and over	13.8	3.8	2.4	2.7	15.5

Illustrate these figures in an appropriate diagram.

The data are conveniently presented in percentage form and, since comparisons are intended, composite bar charts are appropriate. It is difficult to know in what order to present the countries: we have used increasing order of the youngest age group, since this appears on the left of the diagram.

Compound bars showing, for five countries, the proportions of the population in three age ranges



1.9 Pie charts

Pie charts are the circular equivalent of compound bar charts. The areas of the portions of the pie are in proportion to the quantities being represented. Occasionally you may see pies of different sizes; these indicate different population sizes. When drawn correctly the areas (and not the radii) will be in proportion to the differing population sizes.

Example 9

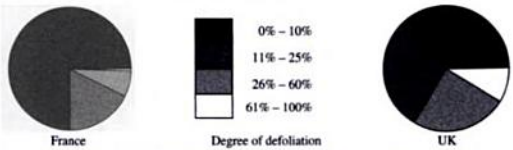
The European Community *Forest Health Report 1989* classifies trees by the extent of their defoliation (i.e. by their loss of leaves). Trees that are in good health have defoliation levels of between 0% and 10%. The following data show the proportions of conifers with various amounts of defoliation in France and the UK.

Illustrate the data using pie charts.

	Extent of defoliation			
	0%–10%	11%–25%	26%–60%	61%–100%
France	0.750	0.176	0.068	0.006
UK	0.358	0.303	0.250	0.089

The separate pies have the same size (since we are not concerned with the *quantities* of conifers in the two countries). The pie segments are shaded to assist with their visibility. The shading scale is chosen so that the colour is darker where there are more leaves (least defoliated).

It can be seen that a sizeable proportion of the UK conifers are heavily defoliated, whereas about three-quarters of the French conifers are in good health (0%–10% defoliation). However, the comparison is not quite fair since other information in the report shows that French conifers are rather younger than those in the UK.



Pie charts comparing the amounts of defoliation of conifers in France and in the UK in 1989

Exercises 1b

1 The numbers of absentees in a class over a period of 24 days were:

0, 3, 1, 2, 1, 0, 4, 0, 1, 1, 2, 3,
1, 0, 0, 2, 4, 6, 4, 2, 1, 0, 1, 1

Construct a bar chart for the above data.

2 A bridge player keeps a note of the numbers of aces that she receives in successive deals. The numbers are:

0, 2, 3, 0, 0, 2, 1, 1, 0,
2, 3, 0, 1, 1, 2, 1, 0, 0

Construct a bar chart for the above data.

3 The numbers of eggs laid each day by 8 hens over a period of 21 days were:

6, 7, 8, 6, 5, 8, 6, 8, 6, 5, 6,
4, 7, 6, 8, 7, 5, 7, 6, 7, 5

Construct a bar chart for the above data.

4 For each potato plant, a gardener counts the numbers of potatoes whose mass exceeds 100 g. The results are:

8, 5, 7, 10, 8, 6, 5, 6, 4, 8,
10, 9, 8, 7, 3, 10, 11, 6, 9, 8

Construct a bar chart for the above data.

5 The numbers of goals scored in the first three divisions of the Football League on 4 February 1995 were:

6, 5, 3, 3, 5, 3, 1, 2, 4, 2, 2, 5,
1, 2, 2, 3, 8, 2, 4, 5, 3, 3, 0, 2,
5, 0, 1, 0, 3, 0, 1, 2, 7, 1, 2

Construct a bar chart for the above data.

6 The shoe sizes of the members of a football team are:

10, 10, 8, 11, 10, 9, 9, 10, 11, 9, 10

Represent the data on (i) a bar chart,
(ii) a compound bar chart,
(iii) a pie chart.

7 A school recorded the numbers of candidates achieving the various possible grades in their A-level subjects.

E: 21; D: 47; C: 69; B: 72; A: 53

Represent the data on (i) a bar chart,
(ii) a compound bar chart, (iii) a pie chart.

8 The proportions of males in the audiences of various sporting fixtures are as follows:
Football match 85%, Rugby match 70%,
Tennis match 45%, Badminton match 40%,
Gymnastics 35%.
Represent these findings using a compound bar chart.

9 Random samples of individuals aged 20–60 are interviewed in five regions of the country. The percentages of males and of females who are found to be in full-time employment are given in the following table.

	Male	Female
South-East	84	61
East Anglia	78	57
West Midlands	70	58
South-West	65	40
Scotland	63	36

Illustrate the data using a multiple bar chart.

10 A school recorded the numbers of candidates achieving the various possible grades in their A-level subjects.

For boys the figures were:

E: 14; D: 29; C: 42; B: 42; A: 21

For girls the figures were:

E: 7; D: 18; C: 27; B: 30; A: 32

Illustrate these results:
(i) using a compound bar chart,
(ii) using two pie charts,
(iii) using a multiple bar chart.

11 The Registrar General's *Annual Reports* reveals the following figures concerning the marital status of men who married in the years 1872, 1931 and 1965:

	1872	1931	1965
Bachelor	86.3	91.7	88.5
Widower	13.7	7.6	4.9
Divorced	*	0.7	6.6

The figures are percentages, with * indicating a figure of less than 0.1%. Display the figures using:
(i) pie charts,
(ii) a multiple bar chart,
(iii) a compound bar chart.

1.10 Grouped frequency tables

The following data are the masses (in g) of 30 brown pebbles chosen at random from those on one area of a shingle beach:

3.4, 12.3, 7.5, 8.2, 8.6, 15.4, 6.9, 7.0, 2.9, 5.0,
13.5, 8.4, 9.9, 11.8, 4.6, 7.7, 3.8, 7.7, 8.6, 14.6,
4.3, 7.9, 9.1, 11.9, 17.4, 6.3, 8.7, 10.1, 5.1, 10.2

A bar chart of these data would look like a very old comb that had had an unfortunate accident! It is obviously sensible to work with ranges of values, which we call **classes**, rather than with the individual values. As a start we summarise the data (perhaps using a tally chart to help with the counting) in order to form a **grouped frequency table**:

Range of masses (g)	1.95–3.95	3.95–5.95	5.95–7.95	7.95–9.95	9.95–11.95	11.95–13.95	13.95–15.95	15.95–17.95
Frequency	3	4	7	7	4	2	2	1

Notes

- Inspecting the recorded data it appears that the measurements were made correct to the nearest 0.1 g. Thus pebbles with masses recorded as lying in the range 2 g–3.9 g have true masses lying in the range 1.95 g–3.95 g.
- The values 1.95, 3.95, ..., 15.95 are the **lower class boundaries (l.c.b.)** of their classes, while the values 3.95, 5.95, ..., 17.95 are the **upper class boundaries (u.c.b.)**. Therefore:
  - u.c.b. of one class = l.c.b. of the next class
  - class width = (u.c.b. − l.c.b.)

  In the example, each of the eight classes has width 2.
- Published tables frequently use the rounded figures in the grouped frequency table, and may give only the class mid-point or just one of the class boundaries (usually the l.c.b.). For example the pebble data might be reported thus:

Range of masses (nearest 0.1 g)	Frequency
2–	3
4–	4
6–	7
8–	7
10–	4
12–	2
14–	2
16–	1

  Great care and some ingenuity is often needed to deduce the true class boundaries – this is, however, typical of published data!
- Many quantities that we measure are not really continuous, but are best treated as such. The following data consists of the advertised prices (in £, in 1992) of second-hand Ford Sierras all less than 3 years old.

  8195, 4995, 9995, 9995, 8995, 8695, 5995, 5495, 7495, 7895, 7295, 8995, 8695, 8495, 7495, 8995, 4995, 7495, 4795, 4995, 4995, 8895, 5495, 6495, 5795, 5695, 5195, 5995, 7995, 7350, 12395, 4995, 9495, 6495

  These prices would be much easier to read with a 5 added! Although price in £ is not a continuous quantity (since all the prices are in whole numbers of pounds), the possible prices are so close together that it is sensible to treat it as such.

Price range (£)	4000–4999	5000–5999	6000–6999	7000–7999	8000–8999
Frequency	6	7	2	7	8

Price range (£)	9000–9999	10000–10999	11000–11999	12000–12999
Frequency	3	0	0	1

1.11 Difficulties with grouped frequencies

- **The value zero** For example, suppose the duration of phone calls are measured to the nearest minute. Then a call of duration '2 minutes' actually lasted for between 1.5 and 2.5 minutes – a range of one minute. Similarly, a call of duration '3 minutes' refers to a range of one minute. The

same is true for every recorded phone call length *except* '0 minutes' which refers to calls of between 0 and 0.5 minutes in duration. The treatment of zero here (for a *continuous* variable) should be contrasted with that below.

- **Grouped discrete data** Suppose a test is marked out of 100 and it is decided to use the classes 0–24, 25–49, 50–74 and 75–100. Natural intermediate class boundaries are 24.5, 49.5 and 74.5. These boundaries lie 0.5 outside the stated ranges of the classes. In order to be consistent, this suggests using −0.5 and 100.5 as the two remaining boundaries, even though negative marks, and marks in excess of 100, are not feasible. The treatment of zero in this note differs from that in the previous note because the quantity being measured here is discrete and not continuous.
- **Age** Unlike almost every other variable, age is reported in *truncated* form. A person who claims to be 'aged 14' is actually aged at least 14.0, but has not yet reached 15.0.

Adolphe Quetelet (1796–1874) was a dominant force in Belgian science for 50 years. His job was as astronomer and meteorologist at the Royal Observatory in Brussels, but his fame was due to his work as a statistician and sociologist! He was one of the founders of the Royal Statistical Society (of London). He spent much time constructing tables and diagrams to show relationships between variables. He was interested in the concept of an 'average man' in the same way as today we talk of the 'average family'.

1.12 Histograms

Bar charts are not appropriate for data with grouped frequencies for ranges of values, instead we use a **histogram** which is a diagram in which rectangles are used to represent frequencies. A histogram differs from the bar chart in that the rectangles may have different widths, but the key feature is that, for each rectangle:

area is proportional to **class frequency**

When all the class widths are equal, histograms are easy to construct, since then not only is area $\propto$ frequency, but also height $\propto$ frequency.

Note

- Some computer packages attempt to make histograms three-dimensional. Avoid these if you can, since the effect is likely to be misleading.

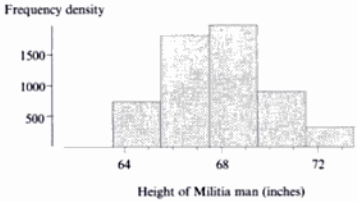
Example 10

Use a histogram to display the following data, which refer to the heights of 5732 Scottish militia men. The data were reported in the *Edinburgh Medical and Surgical Journal* of 1817 and were analysed by Adolphe Quetelet (see above).

Height (ins)	64–65	66–67	68–69	70–71	72–73
Frequency	722	1815	1981	897	317

It appears that the heights were recorded to the nearest inch, so the class boundaries are 63.5, 65.5, 67.5, 69.5, 71.5 and 73.5. These define the locations of the sides of the rectangles while the heights are proportional to 722, 1815, etc.

Histogram of the heights of 5732 Scottish militia men in 1817



Notes

- The y-axis has been labelled *frequency density* rather than frequency because it is *area* which is proportional to frequency.
- Because the classes all have the same width, the vertical scale (frequency density) could be labelled 'Frequency per 2 inch height range'. However, these units have been omitted because they would confuse rather than inform anyone looking at the diagram!

Example 11

As glaciers retreat they leave behind rocks known as 'erratics' because they are of a different type to the normal rocks found in the area. In Scotland many of these erratics have been used by farmers in the walls of their fields. One measure of the size of these erratics is the cross-sectional area, measured in cm^2, visible on the outside of a wall. The areas of 30 erratics are given below.

Provide an appropriate display of these data.

216, 420, 240, 100, 247, 128, 540, 594, 160, 286, 216, 448, 380, 509, 90, 156, 135, 225, 304, 144, 152, 143, 135, 266, 286, 154, 154, 386, 378, 160

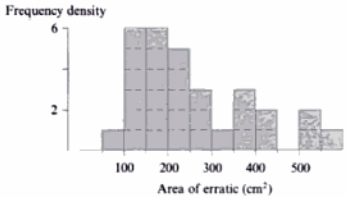
A quick inspection of the data reveals that the values range from 90 to 594. Since many values are possible for cross-sectional area, it will be necessary to group the data and to portray it using a histogram. A good impression of the distribution of a set of values can usually be obtained by using between 5 and 15 classes. This suggests using classes of width 50, with 'natural' boundaries at 50, 100 and so on.

We use a tally chart to help with the counting and obtain:

Range of areas (cm^2)	50–99	100–149	150–199	200–249
Frequency	1	6	6	5
Range of areas (cm^2)	250–299	300–349	350–399	400–449
Frequency	3	1	3	2
Range of areas (cm^2)	450–499	500–549	550–599	
Frequency	0	2	1	

The resulting histogram shows that there is a long 'tail' of large values but no corresponding tail of small values: the distribution is said to be **skewed to the right** or **positively skewed**. This commonly happens when (as here) we are dealing with physical quantities that have no obvious upper bound, but cannot be negative.

Histogram of cross-sectional area of erratics, using classes of equal width



Notes

- The class boundaries are really at 49.5, 99.5, etc, and the histogram is plotted using these values. However, the axis is labelled (accurately) using less 'awkward' values. Of course, you might not have noticed!
- The internal 'boxes' serve to emphasise the relation between frequency and area and would not normally be shown.

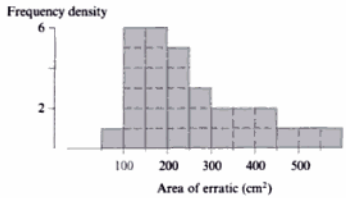
The histogram shows that the typical erratic had a cross-sectional area of around 200 cm^2, and that some were much larger. With a bigger sample we would expect more or less steadily decreasing frequencies as the values of area increase. In order to eliminate the 'jagged' nature of this diagram, we could use wider categories for the larger values of cross-sectional area:

Range (cm^2)	50–99	100–149	150–199	200–249
Frequency	1	6	6	5

Range (cm^2)	250–299	300–449	450–599
Frequency	3	6	3

The histogram corresponding to the revised table has a reasonably smooth outline, with a more-or-less steady decrease from the peak. Effectively all that has happened is that a few of the 'boxes' have tumbled off local peaks into neighbouring troughs!

Smoothed histogram of cross-sectional area of erratics, using classes of unequal width



Notes

- The total area of the histogram is unaltered.
- The units on the y-axis are simply to enable the viewer to get an accurate impression of the relative heights of different parts of the histogram.

Example 12

The table on the following page summarises the results of a 1992 assessment of the knowledge of the mathematics content of the National Curriculum by 7-year-olds.

Results were reported for 105 Local Education Authorities, with the figures in the table being the percentages of pupils who succeeded in attaining level 2 or better.
Illustrate these data in an appropriate fashion.

% reaching level 2	50–59	60–63	64–65	66–67	68–69	70–71
Number of LEAs	4	4	5	8	7	18

% reaching level 2	72–73	74–75	76–77	78–79	80–83
Number of LEAs	17	11	21	5	5

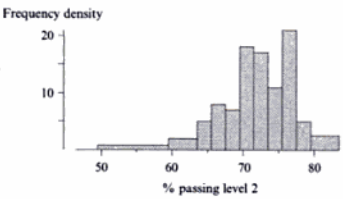
Assuming that the reported figures have been rounded to the nearest percentage point the class boundaries are 49.5, 59.5, 63.5, 65.5, ..., 79.5 and 83.5. Most classes have a width of 2 percentage points, but the classes at either end are wider. Taking 2 percentage points as being the 'standard' width, and recalling that it is *area* that is proportional to frequency, the height of the rectangle representing the final class frequency must be $\frac{5}{2}$, since this class is twice as wide as the standard class. Similarly, the heights of the first two classes will be $\frac{4}{5}$ and $\frac{4}{2}$, since their widths are respectively 5 times and 2 times the standard width.

We can set out these calculations in a table as follows.

Class	Class width w	Frequency f	Frequency density $\frac{f}{w}$
50–59	10	4	0.40
60–63	4	4	1.00
64–65	2	5	2.50
66–67	2	8	4.00
68–69	2	7	3.50
70–71	2	18	9.00
72–73	2	17	8.50
74–75	2	11	5.50
76–77	2	21	10.50
78–79	2	5	2.50
80–83	4	5	1.25

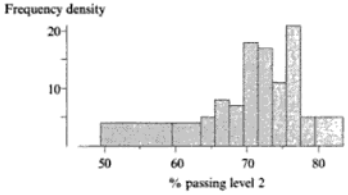
The heights of the sections of the histogram are in proportion to the frequency densities given in the final column of the table.

Histogram showing numbers of LEAs achieving various percentage success rates in Mathematics in the National Curriculum for 7-year-olds



The resulting correct histogram with its thin tails should be contrasted with the incorrect fat-tailed histogram in which no allowance has been made for the extra width of the end intervals.

An incorrect histogram in which, for the end classes, it is height rather than area that has been made proportional to frequency



Note

- The most convenient scale for the *y*-axis is usually in terms of frequencies for the narrowest class width. In this example the indicated scale of 10 and 20 would correspond to frequencies of 10 and 20 in a 2% range of success rates.

Example 13

The data in the table summarise the results of an experiment in the 'seeding' of rainstorms. Of 52 clouds, 26 were randomly selected and 'seeded' by dropping silver nitrate crystals. Estimates were made of the amount of water released by each cloud and these are given in the table.

Unseeded	1203, 830, 372, 346, 321, 244, 163, 148, 95, 87, 81, 69, 47, 41, 37, 29, 29, 26, 26, 24, 22, 17, 12, 5, 5, 1
Seeded	2745, 1698, 1656, 978, 703, 489, 430, 334, 303, 275, 275, 255, 242, 201, 199, 130, 119, 118, 115, 92, 41, 33, 31, 18, 8, 4

The raw data values are extremely skewed (lots of small values and a few very large values) so it makes sense to work with logarithms of the values, instead of the values themselves. The resulting values (using logarithms to base 10 and rounding to two decimal places) are as follows.

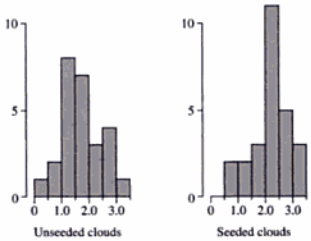
Unseeded	3.08, 2.92, 2.57, 2.54, 2.51, 2.39, 2.21, 2.17, 1.98, 1.94, 1.91, 1.84, 1.67, 1.61, 1.57, 1.46, 1.46, 1.41, 1.41, 1.38, 1.34, 1.23, 1.08, 0.70, 0.70, 0.00
Seeded	3.44, 3.23, 3.22, 2.99, 2.85, 2.69, 2.63, 2.52, 2.48, 2.44, 2.44, 2.41, 2.38, 2.30, 2.30, 2.11, 2.08, 2.07, 2.06, 1.96, 1.61, 1.52, 1.49, 1.26, 0.90, 0.60

Because the two sets are related it makes sense to use the same type of diagram (here a histogram) and, crucially, the same scales. It is usually better to use separate diagrams because it is difficult to distinguish two histograms on a single diagram.

The first step is to form group frequencies.

	0–	0.5–	1.0–	1.5–	2.0–	2.5–	3.0–
Unseeded	1	2	8	7	3	4	1
Seeded	0	2	2	3	11	5	3

Same-scale histograms contrasting the distributions of rainfall amounts (using logarithms of the raw data) for seeded and unseeded clouds



The shift towards larger values is clearly seen in the histogram of the data for the seeded clouds – it does appear that seeding clouds initiates greater rainfall.

Practical

How long can you get a 10p piece to spin on a flat surface?
Using a watch from which seconds can be read accurately, note the lengths of times of four spins. Your personal value will be the length of the longest spin.
Collect the personal values for the whole class and represent the data using a histogram.
Is the histogram roughly symmetrical, or is it skewed?
Was your personal value typical, or was it unusually short or long?

Calculator practice

Some 'bar charts' produced by graphical calculators are really histograms! These charts are only suitable for cases where the class widths are all the same. Use your calculator to reproduce the histogram of heights of militia men.

Exercises 1c

1 A market gardener plants 20 potatoes and weighs the potatoes obtained from each plant. The results, in g, are as follows:

853, 759, 891, 923, 755, 885, 821, 911, 789, 854, 861, 915, 784, 853, 891, 942, 758, 867, 896, 835

Construct a frequency table with class boundaries at 750, 800, . . . , 950.
Show the results in a histogram.

2 The lengths of 20 cucumbers were measured with the following results, in cm:

29.3, 30.5, 34.0, 31.7, 27.8, 29.4, 32.6, 33.4, 29.8, 29.8, 35.4, 36.3, 26.4, 38.8, 37.5, 34.5, 28.6, 31.9, 27.6, 32.0

Construct a frequency table with class boundaries at 25.0, 27.0, . . . , and draw a histogram for the data.

3 A consumers' association tests the lives of car batteries of a particular brand, with the following results, in completed months:

45, 49, 55, 61, 47, 55, 63, 68, 58,
51, 40, 46, 50, 51, 57, 58, 49, 44,
65, 62, 53, 58, 43, 37, 48

Represent the data by a histogram.

4 The mileages travelled by delegates at a conference were as follows:

38, 47, 22, 15, 71, 54, 43, 22, 79, 65,
43, 33, 23, 12, 58, 63, 52, 32, 43, 48,
21, 25, 27, 48, 55, 10, 23, 37, 47, 51

Represent the data by a histogram.

5 The masses (in g to the nearest g) of a random collection of offcuts taken from the floor of a carpenter's shop are summarised below:

0–19	20–39	40–59	60–99
4	17	12	6

Display the data using a histogram.

6 The marks gained in an examination are summarised below.

0–29	30–49	50–69	70–99
4	12	37	14

Represent the data using a histogram.

7 In 1993 the age distribution of the population of the UK (in thousands) was:

Total	Under 1	1–4	5–14
58 191	759	3129	7417

15–24	25–34	35–44	45–59
7723	9295	7787	10 070

60–64	65–74	75–84	Over 85
2839	5169	3020	982

Source: *Population Trends, No 78, Winter 1994*

Choosing a sensible upper limit (which you should state) for the top age category, construct a histogram showing the above data.

8 In 1991 the distribution of the age of a mother at the live birth of a child in the UK was (in thousands):

All	Under 20	20–24	25–29
699.2	52.4	173.4	248.7

30–34	35–39	Over 40
161.3	53.6	9.8

Source: *Population Trends, No 78, Winter 1994*

Making suitable assumptions, which you should state, construct a histogram showing the above data.

9 Quetelet (see earlier biography) analysed the following data, which give the heights (x mm) of potential French conscripts. Those with heights less than 157 cm were excused from military service.

Height range	Frequency
$1435 \leq x < 1570$	28 620
$1570 \leq x < 1597$	11 580
$1597 \leq x < 1624$	13 990
$1624 \leq x < 1651$	14 410
$1651 \leq x < 1678$	11 410
$1678 \leq x < 1705$	8780
$1705 \leq x < 1732$	5530
$1732 \leq x < 1759$	3190
$1759 \leq x < 1840$	2490

Plot these data on a histogram.

10 A survey of cars in a car park reveals the following data on the ages of cars:

<2 yrs	2–4 yrs	5–8 yrs	9–12 yrs
35	51	83	35

Represent the data by a histogram.

11 The lengths (in minutes, to the nearest minute) of the phone calls made between two teenagers are summarised in the table below.

0–4	5–9	10–14	15–19	20–29
2	7	15	18	5

Illustrate these data using a histogram.

1.13 Frequency polygons

The idea of the histogram is to give a visual impression of which values are likely to occur and which values are less likely. The 'chunky' outline of a histogram is not 'a thing of beauty' and an alternative exists *whenever the frequencies have been grouped in classes that are all of equal width*. The **frequency polygon** is constructed as follows. For each class, locate the point with x-coordinate equal to the mid-point of the class and with y-coordinate corresponding to the class frequency. Successive points are then joined to form the polygon. In order to obtain a closed figure, extra classes with zero frequencies are added at either end of the frequency distribution.

Notes

- As with the histogram it is *area* that is proportional to frequency. This is very difficult to achieve with unequal class widths. With equal class widths the above method results in a frequency polygon having the same area as the corresponding histogram.
- Since the frequency polygon is only used with classes of equal width, class frequencies provide a convenient scale for the y-axis.

Example 14

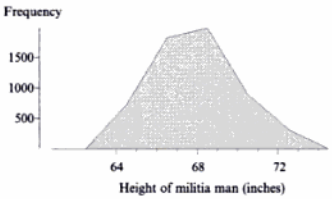
Illustrate the data on the heights of Scottish militia men (Example 10) using a frequency polygon.

Height (ins)	64–65	66–67	68–69	70–71	72–73
Frequency	722	1815	1981	897	317

After adding extra classes, having the same widths but zero frequencies, the data are now summarised in the following table. The addition of the end classes enables us to complete the frequency polygon.

Class mid-point (ins)	62.5	64.5	66.5	68.5	70.5	72.5	74.5
Frequency	0	722	1815	1981	897	317	0

Frequency polygon showing the heights of 5732 Scottish militia men in 1817



Calculator practice

Using the statistical draw mode, a graphical calculator produces a frequency polygon very easily – though the instructions may refer to 'a line graph'. Use your calculator to reproduce the previous polygon.

1.14 Cumulative frequency diagrams

An alternative form of diagram provides answers to questions such as 'What proportion of the data have values less than x?'. In such a diagram, cumulative frequency on the y-axis is plotted against observed value on the x-axis. The result is a graph in which, as the x-coordinate increases, the y-coordinate cannot decrease.

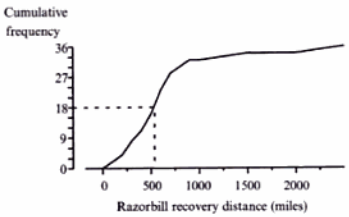
With grouped data the first step is to produce a table of cumulative frequencies. These are then plotted against the corresponding upper class boundaries (u.c.b.). The successive points may be connected either by straight-line joins (in which case the diagram is called a **cumulative frequency polygon**) or by a curve (in which case the diagram is called an **ogive**).

Example 15

In studying bird migration a standard technique is to put coloured rings around the legs of the young birds at their breeding colony. The source of a bird subsequently seen wearing coloured rings can therefore be deduced. The following data, which refer to recoveries of razorbills, consist of the distances (measured in hundreds of miles) between the recovery point and the breeding colony. Illustrate these data using a cumulative frequency polygon and estimate the distance exceeded by 50% of the birds.

Distance (miles) (x)	Frequency	Cumulative frequency
$x < 100$	2	2
$100 \leqslant x < 200$	2	4
$200 \leqslant x < 300$	4	8
$300 \leqslant x < 400$	3	11
$400 \leqslant x < 500$	5	16
$500 \leqslant x < 600$	7	23
$600 \leqslant x < 700$	5	28
$700 \leqslant x < 800$	2	30
$800 \leqslant x < 900$	2	32
$900 \leqslant x < 1000$	0	32
$1000 \leqslant x < 1500$	2	34
$1500 \leqslant x < 2000$	0	34
$2000 \leqslant x < 2500$	2	36

Cumulative frequency polygon of the distances travelled by razorbills between their breeding colony and their recovery point.



The cumulative frequency polygon shows that 50% of the razorbills had travelled more than 520 miles.

Note

- If the recording inaccuracy (e.g. 'to the nearest mile') is small by comparison with the range of the data (2500 miles), there is no need to be over-particular about the end-points. The difference between a value plotted at $x = 99.5$ and $x = 100$ will not be visible!

Calculator practice

Write a routine for cumulating frequencies and use the line graph facility to draw a cumulative frequency diagram. Test it with the razorbill data.

Step diagrams

A cumulative frequency diagram for ungrouped data is sometimes referred to as a **step polygon** or **step diagram** because of its appearance.

Example 16

In a compilation of Sherlock Holmes stories, the 13 stories that comprise *The Return of Sherlock Holmes* have the following numbers of pages:

13.7, 15.5, 16.4, 12.8, 20.8, 13.7, 11.2, 13.7, 11.7, 15.0, 14.1, 14.8, 17.1

The lengths are given to the nearest tenth of a page.
Illustrate these data using a step diagram.

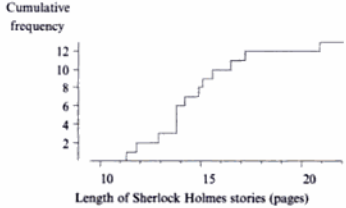
Treating the values as being exact, we use them as the boundaries in a cumulative frequency table. We first need to order the values:

11.2, 11.7, 12.8, 13.7, 13.7, 13.7, 14.1, 14.8, 15.0, 15.5, 16.4, 17.1, 20.8

The resulting table is therefore:

Story length, x	Cumulative frequency
$x < 11.2$	0
$11.2 \leqslant x < 11.7$	1
$11.7 \leqslant x < 12.8$	2
$12.8 \leqslant x < 13.7$	3
$13.7 \leqslant x < 14.1$	6
$14.1 \leqslant x < 14.8$	7
⋮	⋮
$20.8 \leqslant x$	13

Notice that the cumulative frequencies 'jump' at each of the observed values. It is this that gives rise to the vertical strokes in the diagram. The horizontal strokes represent the ranges given in the table.



Step polygon of lengths of Sherlock Holmes stories

Exercises 1d

1 Students were asked to estimate the length (in mm) of a line. Their responses are summarised in the following table.

10–19	20–29	30–39	40–49	50–59	60–69
1	3	10	16	6	1

Represent the data using (i) a frequency polygon, (ii) a cumulative frequency polygon.

2 The lifetimes (in hours to the nearest hour) of bulbs in an advertising hoarding were recorded and are summarised in the following table.

650–699	700–749	750–799	800–849	850–899
1	7	18	9	2

Represent the data using (i) a frequency polygon, (ii) a cumulative frequency polygon.

3 The numbers of eggs laid each day by 8 hens over a period of 21 days were:

6, 7, 8, 6, 5, 8, 6, 8, 6, 5, 6,
4, 7, 6, 8, 7, 5, 7, 6, 7, 5

Display these results using a step diagram.

4 For each potato plant, a gardener counts the numbers of potatoes whose mass exceeds 100 g. The results are:

8, 5, 7, 10, 8, 6, 5, 6, 4, 8,
10, 9, 8, 7, 3, 10, 11, 6, 9, 8

Display these results using a step diagram.

5 The numbers of goals scored in the first three divisions of the Football League on 4 February 1995 were:

6, 5, 3, 3, 5, 3, 1, 2, 4, 2, 2, 5, 1, 2, 2, 3, 8, 2,
4, 5, 3, 3, 0, 2, 5, 0, 1, 0, 3, 0, 1, 2, 7, 1, 2

Display these results using a step diagram.

6 A survey of cars in a car park reveals the following data on the ages of cars:

< 2 yrs	2–4 yrs	5–8 yrs	9–12 yrs
35	51	83	35

Draw a cumulative frequency polygon.

7 In 1993 the age-distribution of the population of the UK (in thousands) was:

Total	Under 1	1–4	5–14
58 191	759	3129	7417
15–24	25–34	35–44	45–59
7723	9295	7787	10 070
60–64	65–74	75–84	Over 85
2839	5169	3020	982

Source: *Population Trends, No 78, Winter 1994*
Draw a cumulative frequency polygon.

8 In 1992 the distribution of the age of a mother at the live birth of a child in the UK was (in thousands):

All	Under 20	20–24	25–29
699.2	52.4	173.4	248.7
30–34	35–39	Over 40	
161.3	53.6	9.8	

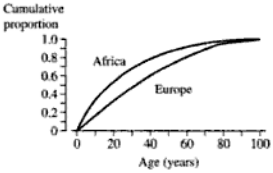
Source: *Population Trends, No 78, Winter 1994*
Display the data using a cumulative frequency polygon.

1.15 Cumulative proportion diagrams

The vertical scale of a cumulative frequency diagram is frequency, with a top value corresponding to n, the number of data items. If we change the scale, by dividing all the entries by n, then we have an axis that records cumulative proportions on a scale from 0 to 1.

This sort of diagram is used when we are interested in proportions rather than numbers. It is particularly effective when comparing two data sets of different sizes. As an example, the diagram contrasts the age distributions of the populations of an African and a European nation.

Cumulative proportion diagram contrasting the age distributions of African and European nations

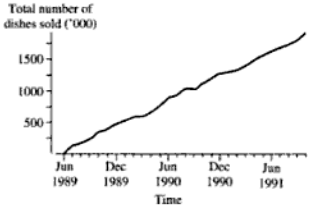


The curve for the African nation rises steeply, reflecting the large numbers of young people: about a quarter of the population are aged under 10, and nearly half are aged under 20. The corresponding proportions for a European nation are about 14% and 30%, respectively.

1.16 Time series

Time-series graphs are probably the type of diagram most frequently encountered in newspapers. They are also possibly the most straightforward: time is plotted on the x-axis and the quantity of interest is plotted on the y-axis.

Cumulative sales of satellite dishes in the UK since July 1989



The figure shows the growth in total of satellite dishes in the United Kingdom between July 1989 and October 1991. The source is a report published in *The Times* in November 1991. The data probably consist of estimates rather than direct counts, since otherwise one would have to conclude that around 20 000 satellite dishes were returned to the shops in September 1990!

The straight-line joins between the successive values are useful here since they enable us to estimate the total sales at intermediate points in time. The implication of the (almost) relentless upward progress of the graph is that monthly sales of satellite dishes remained steady at an average of around 70 000 a month.

Note

- Beware of advertisements showing time series that rise rapidly!
  1. The starting point for the graph may have been chosen to give a false impression. Compare the effects of displaying the time series 83, 91, 87, 82, 74, 82, 91, 89 (with no obvious change on average), to the display of the last four items only (74, 82, 91, 89) where there appears to be a strong increase.
  2. The vertical scale may be exaggerated – check where 0 would occur.

1.17 Scatter diagrams

Often it is of interest to collect data on two variables (x and y, say) simultaneously. The interest arises because we may be able to explain the variation in one variable in terms of the variation in the other variable.

We examine the possible relation between the variables by plotting each pair of values as a point on a diagram using Cartesian coordinates. The resulting diagram is called a scatter diagram (or **scatter plot**). An individual data item would appear on the diagram as a dot (a **data point**) or a cross. Other simple symbols may be used; for example, if data come from several sources then the symbol used could be different for each source.

A time-series plot is really a scatter diagram in which the pairs of values arise in order and the line joining the pairs indicates that order.

Note

- We return to scatter diagrams in Chapter 14.

Example 17

The following data relates soil erosion (in kg/day) to daily average wind velocity (in km/h^{-1}) in a region in the sandy plains of Rajasthan in India. Plot these data in an appropriate scatter diagram.

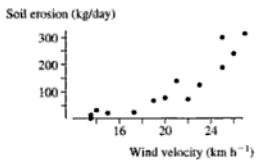
Wind velocity	13.5	13.5	14	15	17.5
Soil erosion	0	10	31	20	20

Wind velocity	19	20	21	22	23
Soil erosion	66	76	137	71	122

Wind velocity	25	25	26	27
Soil erosion	188	300	239	315

We being with a straightforward diagram in which the x-coordinate indicates the daily average wind velocity and the y-coordinate shows the resulting estimated solid erosion.

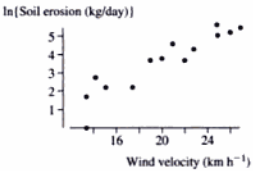
Scatter diagram showing the relation between wind velocity and soil erosion in a sandy Indian plain



The statistician often needs to try many diagrams before finding the one that makes the most useful display of the data. Our first effort suggests that soil erosion increases dramatically as the wind velocity picks up, and that the relation between the two variables is not linear. One possibility is that the relationship is exponential: this can be examined by plotting the (natural) logarithm of the soil erosion against wind speed.

The revised diagram does appear to show a linear relation between ln(soil erosion) and wind velocity.

Revised scatter diagram showing the relation between wind velocity and soil erosion in a sandy Indian plain using a logarithmic scale for the y-axis



Note

- Notice that the x-axis does not begin at zero, but at about 12. Using a 'broken' axis like this enables us to 'fill' the diagram with the data. If we had insisted on including zero then all the data would have been in a small part of the right-hand side of the diagram. Effective diagrams often require one or both of the axes to be broken in this way, though the result must not be misleading to the viewer.

Calculator practice

Graphical calculators are particularly effective for drawing scatter diagrams of pairs of values. Reproduce the original scatter diagram and then investigate other ways of transforming the data so as to get scatter diagrams that appear to be more linear in form.

Computer project

Investigate how to produce scatter diagrams using a spreadsheet. As with graphical calculators it is easy also to experiment with transformations of one or both of x and y with the aim of producing an approximate straight line on the diagram.

Exercises 1e

1 The total amount of snow cover over Europe and Asia during October for the years 1970 to 1979 is given (in millions of square kilometres) in the table.

1970	1971	1972	1973	1974
6.5	12.0	14.9	10.0	10.7
1975	1976	1977	1978	1979
7.9	21.9	12.5	14.5	9.2

Display this information on a suitable diagram.

2 The acidity of milk (the pH value, y) depends upon the temperature at which it is stored (x °C).
Some experimental results are shown below.

x	4	24	38	40
y	6.85	6.63	6.62	6.57
x	40	60	70	78
y	6.52	6.38	6.32	6.34

Display this information on a suitable diagram.

3 The average numbers of deaths per 1000 population in Norway during the period 1750–1850 are summarised below.

1750	1770	1790	1810	1830	1850
25.5	23.6	22.9	26.8	19.7	17.2

Display this information on a suitable diagram.

4 A doctor records the number of patients that he sees in the second week of four particular months in the year. The numbers are as follows:

Month	Jan.	Apr.	Jul.	Oct.
No. seen ('94)	255	235	176	219
No. seen ('95)	215	207	139	243

Represent the data using a suitable diagram.

5 The total value of goods (in thousands of £) produced by a manufacturer each quarter, together with the value of goods exported (in thousands of £) are given below.

1993	1st qtr	2nd qtr	3rd qtr	4th qtr
prodn.	238	316	297	286
exports	57	89	94	82

1994	1st qtr	2nd qtr	3rd qtr	4th qtr
prodn.	211	297	241	270
exports	63	82	108	103

1995	1st qtr	2nd qtr	3rd qtr	4th qtr
prodn.	224	289	285	228
exports	76	97	114	91

Represent the data using a suitable diagram.

1.18 Choosing which display to use

The previous sections have presented a rather bewildering variety of types of diagram. The following table is intended to help you select the appropriate diagram for your data by identifying the diagrams most commonly associated with different types of data.

Data type		*Type of display*
Discrete data	Few different values	Tally chart for counting, bar chart for display
	Many different values	Stem-and-leaf diagram, or histogram, or box-whisker diagram (Section 2.13)
Qualitative data	To show frequencies	Bar chart
	To show proportions	Pie chart
Grouped data	Unequal intervals	Histogram
	Equal intervals	Histogram, or frequency polygon
Cumulative data	Continuous variable	Cumulative frequency diagram, or cumulative proportion diagram
	Discrete variable	Step diagram

Two samples	Same categories	Multiple bar chart, or compound bar chart, or side-by-side pie charts
	Discrete	Back-to-back stem-and-leaf, or side-by-side bar charts, or multiple bar chart
	Continuous	Cumulative frequency diagram, or cumulative proportion diagram
Two variables	One variable is time Data in pairs	Time series Scatter diagram

When reporting data, either in a table or as individual values, try to organise the units so that the figures presented are simple integers. (This is the idea that underlies the stem-and-leaf diagram.) Here are some examples.

Context	*A bad choice*	*A good choice*
Masses of flour bags	Report in kg: 1.004, 1.032, 1.040, 1.011	Report in g the excess over 1 kg: 4, 32, 40, 11
Years	2000, 2005, 2006, 2002	0, 5, 6, 2
Average exam marks	With irrelevant accuracy: 53.377, 62.401, 15.822	To nearest integer: 53, 62, 16

1.19 Dirty data

It is almost certain that a large set of data will contain errors! This may seem rather pessimistic but remember that "To err is human ...". Here are some examples of the way that errors arise.

- **Mistype** The correct value was obtained, but it was written down or typed incorrectly. The most common errors are digits interchanged (1183 instead of 1138) and a digit double-typed (993 instead of 93).
- **Mistaken answer** An interviewee misunderstands a question and gives the wrong answer. For example, an individual earning £24 000 a year replies £24 000 when asked to state his monthly earnings.
- **Mistaken measurement** There is an innate preference for 'nice' round numbers. Suppose that a large collection of pebbles with masses in the range 20 g to 60 g are weighed. The values reported are likely to show pronounced peaks at 30 g, 40 g, 50 g as well as lesser peaks at 35 g, 45 g, etc.
- **Mistaken rounding** Values are frequently rounded. This causes a problem with 'halves'. For example, how do you round 3.465 to two decimal places, or round 4.5 to the 'nearest' integer? A bad rule is to always round in the same direction, since this will bias totals and averages. A better rule is to always round to an even digit – thus 3.465 becomes 3.46 and 4.5 becomes 4.

- **Misreporting** In one famous example of misreporting, the specific gravity of beer in barrels was being determined. Barrels with a specific gravity greater than a had an acceptable alcohol quantity and were rolled downhill. However, barrels with a specific gravity less than a had to be rolled uphill for further treatment. When the data were examined it was found that there were remarkably few barrels with specific gravity just below a! The measurer had taken a 'favourable view' of his measuring instrument.
- **Biased sampling** Chapter 3 provides some examples of ways in which the sample may misrepresent the population as a consequence of a poor sampling procedure.

The first job of a statistician, with any data set, is to examine it using pictures and summary statistics (Chapter 2) in a search for mistaken data items. Even when there are no errors in the data there are often unusual values. These are called **outliers** (see pp 53, 60) and their presence can cause problems for the subsequent data analysis.

Chapter summary

- The principal purpose of Statistics is the drawing of conclusions about large populations (human or otherwise) from comparatively small amounts of data.
- Diagrams are an effective way of conveying information.
- If only a small number of discrete values are possible, then the best approach is often to use a **tally chart**, followed by a summary in a **frequency table** and representation using a **bar chart**.
- If a large number of discrete values are possible, then the best approach is often to use a **stem-and-leaf diagram**, followed by a summary in a **grouped frequency table** and representation using a **histogram.**
- **Pie charts** and **compound bar charts** are useful when the features of interest are the relative sizes of the frequencies in alternative categories.
- There are many ways of portraying data. Whatever method is used, try to make it self-explanatory for the reader (and, if possible, interesting!).
- When data are collected on two variables simultaneously, representation using either a **scatter diagram** or a **time-series** graph may be appropriate.

Exercises 1f (Miscellaneous)

1 The total scores given in *The Times* for Welsh and Scottish Rugby matches on 4 February 1995 were:

44, 21, 23, 26, 24, 39, 56, 22, 28, 25, 63, 83, 42, 39, 24, 23, 38, 61, 44, 19, 31, 24, 24, 60, 45, 48, 39, 34, 50, 46, 53, 43, 43

Construct a stem-and-leaf diagram to display the above data.

2 The daughter of a market gardener plants twenty sunflower plants in her garden. When they are full grown, she measures them and records their heights in metres as follows:

1.60, 1.72, 2.23, 2.12, 1.70, 1.93, 1.69, 2.11, 1.99, 2.08, 2.11, 1.79, 2.01, 1.88, 1.93, 2.22, 1.92, 2.44, 1.87, 1.76

Summarise these data using a stem-and-leaf (!) diagram.

3 In 1665, a total of 97 308 people died in London (compared to just 9967 births). The principal cause of death was the plague, which accounted for 68 596 of the deaths. Show this information on a pie chart.

Of the deaths not due to the plague, the principal causes (according to the *Annual Bill of Mortality for London*, and using its spelling) were these:

Aged	1545
Ague and Feaver	5257
Chrisomes and Infants	1258
Consumption and Tissick	4808
Convulsion and Mother	2036
Dropsie and Timpany	1478
Griping in the Guts	1288
Spotted feaver and Purples	1929
Surfet	1251
Teeth and Worms	2614

Illustrate this information on a bar chart.

4 According to *Punch*, in the first half of 1869 there were 106 traffic fatalities in London: "Eight persons were killed by horses, 3 by carriages, 6 by omnibuses, 15 by cabs, 33 by vans or waggons, and 40 by carts. One person was killed by a dray." Display this information using a pie chart.

5 The following table shows data about road traffic in each of the years from 1984 to 1993. Row A shows estimates of the total distance (in billions of kilometres) travelled by cars and taxis in each of the years. Row B shows estimates of the total distance (in billions of kilometres) travelled by all motor vehicles in each of the years.

Year	1984	1985	1986	1987	1988
A: cars, taxis ($\times 10^9$ km)	244.0	250.5	264.4	284.6	305.4
B: all motor vehicles ($\times 10^9$ km)	303.1	309.7	325.3	350.5	375.7

Year	1989	1990	1991	1992	1993
A: cars, taxis ($\times 10^9$ km)	331.3	335.9	335.2	336.4	336.8
B: all motor vehicles ($\times 10^9$ km)	406.9	410.8	411.6	410.4	410.2

(i) State briefly two general features of the data in Row A.

(ii) Describe a suitable way of presenting all the data in the table in a single diagram.

(iii) Comment on each of the following statements, justifying your answer.

(a) "In 1993, of the kilometres travelled by all motor vehicles, less than 1 in 5 was travelled by motor vehicles other than cars and taxis."

(b) "The figures prove that the number of motor vehicles increased each year between 1984 and 1991." [UCLES]

6 The following table shows the average number of persons per household in the United Kingdom in each of seven consecutive years.

	1985	1986	1987	1988	1989	1990	1991
Males	1.258	1.236	1.223	1.229	1.217	1.193	1.169
Females	1.339	1.317	1.310	1.288	1.292	1.281	1.253
All	2.596	2.554	2.533	2.516	2.509	2.475	2.422

© G Dennis (ed). *Annual Abstract of Statistics 1993*, Central Statistical Office

(i) Give a reason why, for some years, the value in the third row is not the exact sum of the values in the first and second rows.

(ii) State briefly two distinct features of the data.

(iii) State why it is not possible to deduce from the above data that the population of the UK decreased between 1985 and 1991. [UCLES]

7 During a particular month a family spends £52.27 on meat, £23.10 on fruit and vegetables, £19.72 on drink, £12.41 on toiletries, £102.68 on groceries and £9.82 on miscellaneous items. These data are to be represented by a pie chart of radius 5 cm.

(*a*) Calculate, to the nearest degree, the angle corresponding to each of the above classifications. (DO NOT DRAW THE PIE CHART.)

The following month the family spends 20% more in total.

(*b*) Find the radius of a comparable pie chart to represent the data on this occasion. [ULEAC]

8 Telephone calls arriving at a switchboard are answered by the telephonist. The following table shows the time, to the nearest second, recorded as being taken by the telephonist to answer the calls received during one day.

Time to answer (to nearest second)	Number of calls
10–19	20
20–24	20
25–29	15
30	14
31–34	16
35–39	10
40–59	10

Represent these data by a histogram.
Give a reason to justify the use of a histogram to represent these data. [ULEAC]

9 The following table shows the time to the nearest minute, spent reading during a particular day by a group of school children

Time	Number of children
10–19	8
20–24	15
25–29	25
30–39	18
40–49	12
50–64	7
65–89	5

(*a*) Represent these data by a histogram.
(*b*) Comment on the shape of the distribution. [ULEAC]

2 Summary statistics

Lies, damned lies and statistics

Benjamin Disraeli

2.1 The purpose of summary statistics

Simple! The purpose of summary statistics is to replace a huge indigestible mass of numbers (the **data**) by just one or two numbers that, together, convey most of the essential information.

Well, perhaps it is not so simple, since this is a pretty stiff challenge! No single summary statistic can tell us all about a set of data. Different statistics emphasise different aspects of the data and it will not always be evident which aspect is more important. An example of the difficulties is provided by an interchange many years ago, in the Houses of Parliament. The MPs were debating the need for road signs in Wales to give directions in both Welsh and English. The discussion went something like this:

MP A: Since less than 10% of the population of Wales speak Welsh it is unnecessary to include directions in Welsh.

This seems like a pretty convincing statistic! But wait:

MP B: Over 90% of the area of Wales is inhabited by a population whose principal language is Welsh – directions in Welsh are essential.

It is easy to see why Disraeli was rather hard on Statistics! Both the above statements were essentially correct at the time (though the percentages are invented by the present authors): but they led to opposite inferences. Clearly we have to be careful to choose our summary statistics to be appropriate.

For **univariate** data (i.e. data concerned with a single quantity) there are two main types of summary statistic: **measures of location** and **measures of spread**. Measures of location answer the question 'What sort of size values are we talking about?'. Measures of spread answer the question 'How much do the values vary?'. Both are discussed in this chapter.

The main purpose of Statistics is to draw conclusions about a (usually large) **population** from a (usually small) **sample** of **observed values**: the **observations**. In this chapter we study various ways of providing numerical summaries of the observations.

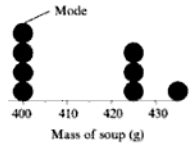
2.2 The mode

The **mode** of a set of discrete data is the single value that occurs most frequently. This is the simplest of the measures of location, but is of limited use. If there are two such outcomes that occur with equal frequency then there is no unique mode and the data are described as being **bimodal**; if there are three or more such outcomes then the data are called **multimodal**. The associated adjective is 'modal', so we are sometimes asked to find the **modal value**.

Example 1

At the supermarket I buy 8 tins of soup. According to the information on the tins, four have mass 400 g, three have mass 425 g and one has mass 435 g. Find the mode.

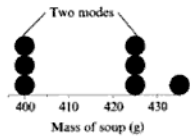
The mode is 400 g because 400 g is the most common value.



Example 2

After unpacking the shopping, I feel hungry and have soup for lunch. I choose one of the 400 g cans bought in the previous example. What is an appropriate description of the frequency distribution of the remaining 7 masses?

There are now two modes, one at 400 g and one at 425 g: an appropriate description is 'bimodal'.



Calculator practice

When used to plot a bar chart or frequency polygon, graphical calculators may also indicate the value of the mode and the associated frequency.
Use a calculator to determine the mode of the soup tin data.
What happens if the 435 g tin is replaced by a 425 g tin?

Modal class

For continuous data (or for grouped discrete data) the mode exists only as an idea! When measured with sufficient accuracy, all observations on a continuous variable will be different: even if John and Jim both *claim* to be 1.8 metres tall, we can be certain that their heights will not be *exactly* the same. However, if we plot a histogram of a sample of men's heights, we will usually find that it has a peak in the middle: the class corresponding to this peak is called the **modal class**.

For **qualitative data** (in which items are described by the qualities they possess) we again refer not to a mode, but to a modal class. In the next example 'hair colour' is a qualitative variable.

Example 3

The hair colours and heights of a class of male university students are summarised below.
Determine the modal classes.

Hair colours			
Black	Brown	Blond	Red
63	53	9	3

Heights (m), x			
$1.4 \leqslant x < 1.7$	$1.7 \leqslant x < 1.9$	$1.9 \leqslant x < 2.0$	$2.0 \leqslant x < 2.2$
19	56	42	11

For hair-colour the modal class is 'Black'. For height the modal class is '1.9–2.0 m' and not '1.7–1.9 m'. To see why, imagine drawing the histogram: the heights of the middle two rectangles would be in the proportion 28 to 42 (because the width of the second class is twice the width of the third class).

2.3 The median

The word 'median' is just a fancy name for 'middle'! After all the observations have been collected, they can be arranged in a row in order of magnitude, with the smallest on the left and the largest on the right (or vice versa). Values in the middle of the ordered row will therefore be intermediate in size and should give a good idea of the general size of the data. For example, suppose the observed values are 13, 34, 19, 22 and 16. Arranged in order of magnitude these become:

13, 16, **19**, 22, 34

The middle value, 19, is called the **median**.

When there are an even number of observations there are two middle values. By convention the median is then taken as their average. For the soup tins in Example 1, the values were:

400, 400, 400, **400**, **425**, 425, 425, 435

The value of the median is taken to be

$$\tfrac{1}{2}(400 + 425) = 412.5$$

In general, with n observed values arranged in order of size, the median is calculated as follows:

If n is odd and equal to $(2k + 1)$, say, then the median is the $(k + 1)$th ordered value.

If n is even and equal to $2k$, say, then the median is the average of the kth and the $(k + 1)$th ordered values.

Note

- A useful preliminary is to summarise the data using a stem-and-leaf diagram, since this immediately provides an ordering of the data.

Example 4

A chemistry professor has an accurate weighing machine and two sons, Charles and James, who are keen on playing conkers. One day, Charles and James collect some new conkers. On their return home, following a dispute over who has the best conkers, they use their father's balance to determine the weights of the conkers (in g). Their results are as follows:

Charles: 31.4, 44.4, 39.5, 58.7, 63.6, 51.5, 60.0
James: 60.1, 34.7, 42.8, 38.6, 51.6, 55.1, 47.0, 59.2

Which boy's collection of conkers has the higher median weight?

We first arrange each set of values in ascending order, and then highlight the central value(s):

Charles: 31.4, 39.5, 44.4, **51.5**, 58.7, 60.0, 63.6
James: 34.7, 38.6, 42.8, **47.0**, **51.6**, 55.1, 59.2, 60.1

The median for James is the average of 47.0 and 51.6, which is 49.3. This is less than the median for Charles, which is 51.5, so Charles's collection has the higher median weight.

Calculator practice

Some calculators will report the median value. You may need to delve deeply into the calculator manual in order to find the correct sequence of keystrokes. You should check that the calculator reports the correct value for the median both in the case of an even number of data items and in the case of an odd number. Use the data above as a check. (Remember that the calculator depends on its built-in instructions – these are not always correct!)

2.4 The mean

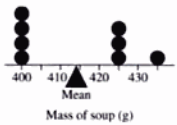
This measure of location is often called the **average**, and can be used with both discrete and continuous data. The mean is equal to the sum of all the observed values divided by the total number of observations. Unlike the value of the mode, the value of the mean will usually not be equal to any one of the individual observed values. Thus the mean mass (in g) of the tins of soup is:

$$\frac{(400+400+400+400+425+425+425+435)}{8} = 413.75$$

It is time to introduce some algebra! Suppose that the data set consists of n observed values, denoted by $x_1, x_2, \ldots, x_n$. Then the **sample mean**, which is usually denoted by $\bar{x}$, is given by

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n} \qquad (2.1)$$

One way of thinking about the mean is as the **centre of mass** when the observations are 'balanced' on the x-axis.



The sample mean viewed as a centre of mass

Calculator practice

Many calculators are capable of calculating the mean of a set of data using an appropriate sequence of key strokes.
Determine how to do this using your own calculator.

Exercises 2a

1 A school records the numbers of candidates achieving the various possible grades in their A-level subjects.

E: 21; D: 47; C: 69; B: 72; A: 53

Find the modal class.

2 A survey of cars in a car park reveals the following data on the ages of cars:

< 2yrs	2–4 yrs	5–8 yrs	9–12 yrs
35	51	83	35

Determine the modal class.

3 In 1993 the age distribution of the population of the UK (in thousands) was:

Total	Under 1	1–4	5–14
58 191	759	3129	7417
15–24	25–34	35–44	45–59
7723	9295	7787	10 070
60–64	65–74	75–84	Over 85
2839	5169	3020	982

Source: *Population Trends, No 78, Winter 1994*

Determine the modal class.

4 Most people have more than the average number of legs!
Explain.

5 Eight athletes run 100 m. The times taken (in s) are:

10.34, 10.68, 10.81, 11.02,
11.35, 11.71, 11.82, 11.95

Find the average time taken.

6 A bridge player keeps a note of the numbers of aces that she receives in successive deals. The numbers are:

0, 2, 3, 0, 0, 2, 1, 1, 0,
2, 3, 0, 1, 1, 2, 1, 0, 0

Find (i) the mode, (ii) the mean, of the numbers of aces received.

7 A gardener classifies a potato having a mass over 100 g as being 'large'. The gardener grows a number of potato plants and, for each plant, he counts the number of large potatoes, obtaining the following results:

8, 5, 7, 10, 8, 6, 5, 6, 4, 8,
10, 9, 8, 7, 3, 10, 11, 6, 9, 8

Find (i) the mode, (ii) the mean, of the numbers of large potatoes.

8 The heights, in m, of 12 walnut seedlings, after twenty years' growth, were:

4.3, 5.2, 4.1, 3.5, 5.2, 4.8,
5.3, 4.8, 3.7, 4.1, 4.5, 5.0

Find the mean height.

9 A computer is programmed to generate 8 random numbers between -1 and $+1$.
The numbers generated are:

0.269, -0.679, 0.507, -0.663,
0.325, -0.960, 0.741, 0.484

Find the mean.

10 A student's bank balance at the end of each month was recorded in £. A negative quantity denotes an overdraft.
The figures were as follows:

341.32,	97.53,	-57.44,	255.93,
5.89,	-83.33,	152.81,	-23.11
-105.73,	-204.50,	-150.46,	-85.39

Find her mean bank balance at the end of each month.

11 The weights, in kg, of the Cambridge Boat Race crew in 1995 were:

90.7, 89.4, 93.4, 92.1, 82.6, 92.5, 94.4, 89.8

The weights of the Oxford crew were:

86.9, 90.3, 94.8, 97.5, 89.6, 89.8, 91.9, 89.1

Find the mean weight of each crew and verify that the Oxford crew is heavier than the Cambridge crew by an average of 0.63 kg per man.

12 The mean of the following numbers is 20:

20, 18, c, 24, 23, 13

Find the value of c.

13 The numbers of goals scored in the first three divisions of the Football League Championship on 4 February 1995 were:

6, 5, 3, 3, 5, 3, 1, 2, 4, 2, 2, 5, 1, 2, 2, 3, 8, 2,
4, 5, 3, 3, 0, 2, 5, 0, 1, 0, 3, 0, 1, 2, 7, 1, 2

Find (i) the mode, (ii) the mean, of the number of goals scored.

14 The heights of the Cambridge Boat Race crew in the 1995 race were:

6 ft 3 in, 6 ft 5 in, 6 ft 3 in, 6 ft 4 in,
6 ft 2 in, 6 ft 6 in, 6 ft 4 in, 6 ft 2 in

The heights of the Oxford crew were:

6 ft 3 in, 6 ft 1 in, 6 ft 5 in, 6 ft 4 in,
6 ft 5 in, 6 ft 3 in, 6 ft 3 in, 6 ft 2 in

Find the difference in their median heights.

15 The numbers of matches in a box were counted for a sample of 25 boxes. The results were:

51, 52, 48, 53, 47, 48, 50, 51, 50,
46, 52, 53, 51, 48, 49, 52, 50, 48,
47, 53, 54, 51, 49, 47, 51

By constructing a tally chart, or otherwise, find the median number of matches in a box.

16 A record is kept of the number of patients attending each day at a medical practice. The numbers are:

45, 41, 37, 48, 44, 29, 32, 43, 41, 37, 38,
31, 43, 39, 35, 31, 42, 40, 35, 42, 35

Construct a stem-and-leaf diagram and hence find the median number of patients attending per day.

17 A baker keeps count of the number of doughnuts sold each day for three weeks. The numbers are:

35, 47, 34, 46, 55, 82, 41, 35, 47,
51, 56, 75, 38, 41, 44, 51, 45, 74

By constructing a steam-and-leaf diagram, or otherwise, find the median number of doughnuts sold per day.

18 The marks obtained in a mathematics test marked out of 50 were:

35, 42, 31, 27, 48, 50, 24, 27,
21, 37, 41, 34, 12, 18, 27

Find:
(i) the mean mark,
(ii) the median mark.

19 A choirmaster keeps a record of the numbers turning up for choir practice:

25, 28, 32, 31, 31, 34, 28, 31, 29,
28, 32, 32, 30, 29, 29, 31, 28, 28

(i) Find the mean number attending.
(ii) Determine the median number.

20 The shoe sizes of the members of a football team are:

10, 10, 8, 11, 10, 9, 9, 10, 11, 9, 10

Find:
(i) the mean shoe size,
(ii) the median shoe size,
(iii) the modal shoe size.

2.5 Advantages and disadvantages of the mode, mean and median

Advantages

- If a mode exists it is certain to have a value that was actually observed.
- The median can be calculated in some cases where the mean or mode cannot. For example, suppose 99 homing pigeons fly from A to B. The median time of flight can be calculated as soon as the 50th pigeon has arrived – we don't need to wait for the last exhausted traveller (who may never arrive!).

Disadvantages

- The mode may not be unique (because two or more values may be equally frequent).
- The mean may be significantly affected by the inclusion of a mistaken observation (e.g. a tin of soup misreported as having mass 4000 g) or of an unusual observation (e.g. the salary of the boss of a factory included with those of the factory workers).
- The statistical properties of the mode and the median are difficult to determine.

In practice much more use is made of the mean than of either of the other two measures of location.

Practical

How many four-legged pets does the typical family have?
Use a tally chart to record the combined number of dogs, cats, hamsters, etc, for each member of your class.
Determine the mean, median, and mode of these values. Which was easiest to calculate?
An organisation wishes to estimate the total number of four-legged pets in your area.
Which of your three statistics is likely to be most useful to them?

2.6 Sigma (Σ) notation

Expressions such as $(x_1 + x_2 + \cdots + x_n)$ are tedious to write. We want to write 'Sum of the x-values', but this is not very mathematical – it doesn't contain Greek letters! Instead, therefore, we write

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n \tag{2.2}$$

The Σ sign is the Greek equivalent of S and is pronounced 'sigma'. Confusingly, we shall also meet shortly the Greek equivalent of s which is also pronounced 'sigma', but looks quite different (σ) and has a very different statistical interpretation.

Notes

- In the shorthand formula the letter i is simply an index. Any letter could be used, but it must replace i everywhere it appears. For example:

$$\sum_{j=1}^{4} y_j = \sum_{i=1}^{4} y_i = \sum_{r=1}^{4} y_r = y_1 + y_2 + y_3 + y_4$$

- Changing the value of n results in a change in the terms being summed. For example:

$$\sum_{j=1}^{3} y_j = y_1 + y_2 + y_3 \quad \text{but} \quad \sum_{j=1}^{2} y_j = y_1 + y_2$$

Applications of sigma notation

Here are some further examples of the use of the Σ sign:

$$\sum_{r=1}^{3} r = 1 + 2 + 3 = 6$$

$$\sum_{s=2}^{4} s^2 = 2^2 + 3^2 + 4^2 = 29$$

$$\sum_{j=1}^{2} (2j + 5) = \{(2 \times 1) + 5\} + \{(2 \times 2) + 5\} = 16$$

$$\sum_{k=2}^{3} (k^2 + 6k) = \{2^2 + (6 \times 2)\} + \{3^2 + (6 \times 3)\} = 43$$

There are four particularly useful results that involve manipulation of the Σ sign:

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \tag{2.3}$$

$$\sum_{i=1}^{n} cx_i = c \sum_{i=1}^{n} x_i \tag{2.4}$$

$$\sum_{i=1}^{n} c = nc \tag{2.5}$$

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{m} x_i + \sum_{i=m+1}^{n} x_i \tag{2.6}$$

In the above c is a constant and m is an integer such that $1 \leqslant m < n$. Result (2.5) in particular should be noted. It follows immediately from result (2.4)

by putting all the x-values equal to 1. All four results are easily proved by writing out the various summations in full.

Notes

- Often the limits of the summation are obvious, in which case they may be dropped from the formula. For example, for the mean of n observations $x_1, x_2, \ldots, x_n$ we could write

$$\bar{x} = \frac{\Sigma x_i}{n}$$

- In ordinary text we write $\sum_{i=1}^{n} x_i$ instead of

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

- As a shorthand, when the formulae are thick on the ground, the suffix may also be omitted:

$$\bar{x} = \frac{\Sigma x}{n} \qquad (2.7)$$

Exercises 2b

1 It is given that $x_1 = 2$, $x_2 = 3$, $x_3 = 5$, $x_4 = 1$, $x_5 = 3$, $x_6 = 2$, $x_7 = 0$, $x_8 = 2$
Verify that:

(i) $\sum_{i=1}^{8}(x_i + 2) = \sum_{i=1}^{8} x_i + 16$

(ii) $\sum_{i=1}^{8} x_i = \sum_{i=1}^{4} x_i + \sum_{i=5}^{8} x_i$

(iii) $\sum_{i=1}^{8}(3x_i) = 3\sum_{i=1}^{8} x_i$

(iv) $\sum_{j=1}^{8} x_j = \sum_{i=1}^{4} x_i + \sum_{k=5}^{8} x_k$

2 It is given that $x_1 = 2$, $x_2 = -3$, $x_3 = 0$, $x_4 = -1$, $y_1 = 3$, $y_2 = -2$, $y_3 = 10$, $y_4 = 2$
Verify that:

(i) $\sum_{i=2}^{4}(x_i + y_i) = \sum_{i=2}^{4} x_i + \sum_{i=2}^{4} y_i$

(ii) $\sum_{i=1}^{4}(x_i y_i) = 10$

(iii) $\left(\sum_{i=1}^{4} x_i\right)\left(\sum_{i=1}^{4} y_i\right) = -26$

3 Find:

(i) $\sum_{j=1}^{8} j$

(ii) $\sum_{j=1}^{8} j^2$

(iii) $\sum_{j=1}^{8} (j-2)^2$

4 A set of data for 10 observations has $\Sigma x = 365$. Find the mean.

5 The summarised data for a set of observations is $n = 60$, $\Sigma y = 74\,344$.
Find the mean value of y.

6 The results of 30 experiments to find the value of the acceleration due to gravity are summarised by $\Sigma g = 294.34$.
Find the mean value.

7 Eight numbers have a mean of 16.
Given that the first seven numbers have a total of 130, determine the value of the eighth number.

8 A set of 25 observations was found to have a mean of 15.2. It was subsequently found that one item of data had been wrongly recorded as 23 instead of 28.
Find the revised value of the mean.

2.7 The mean of a frequency distribution

We have seen in Chapter 1 that data are often represented by a frequency distribution. For example, for the soup tins (see Section 2.4) we have:

Reported mass (g) x	400	425	435
Observed frequency f	4	3	1

The sum of the frequencies $(4 + 3 + 1)$ is equal to n, the total number of observations. The sum of the three products 4×400, 3×425 and 1×435 is equal to the sum of the eight observations, and so the mean mass is $\frac{1600 + 1275 + 435}{4 + 3 + 1} = 413.75$ g, as before. All that we have done is to collect together equal values of x.

So an alternative general formula for the sample mean is

$$\bar{x} = \frac{\sum_{j=1}^{m} f_j x_j}{\sum_{j=1}^{m} f_j} \tag{2.8}$$

where here the summation is over the m different values of x that were recorded. In the example, $m = 3$, $x_1 = 400$, $x_2 = 425$, $x_3 = 435$, $f_1 = 4$, $f_2 = 3$ and $f_3 = 1$.

Now $\sum_{j=1}^{m} f_j$ equals n, the total number of observations, so a simpler form for the previous formula is:

$$\bar{x} = \frac{1}{n} \Sigma f_j x_j$$

which we may write (more casually!) as $\frac{\Sigma fx}{n}$.

Calculator practice

Most calculators with statistical functions have some special key sequence for dealing with the input of grouped frequencies. Investigate how this can be done with your calculator and test the procedure using the soup tin data.

2.8 The mean of grouped data

The formula for the mean of a frequency distribution can also be used to provide an estimate of the sample mean of a set of grouped data:

$$\bar{x} = \frac{\Sigma f_j x_j}{n}$$

In this case x_j is the **class mid-point** for the jth of m classes, f_j is the frequency for this class and $n = \sum_{j=1}^{m} f_j$. This is only an estimate of the actual sample mean since we do not know the individual sample values.

Notes

- The estimate is often referred to as the **grouped mean**.
- The difference in the values of the grouped mean and the true sample mean will usually be very small.
- It is usually much quicker to group a set of data and calculate the grouped mean than to calculate the sample mean directly (unless a computer is taking the strain!).
- Sometimes we only have grouped data available!

Example 5

The following data summarise the distances travelled by a fleet of 190 buses before experiencing a major breakdown.

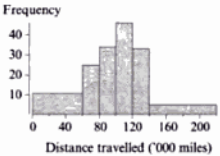
Distance ('000 miles) (d)	$d \leqslant 60$	$60 < d \leqslant 80$	$80 < d \leqslant 100$
Mid-point (x)	30	70	90
Frequency (f)	32	25	34

Distance ('000 miles) (d)	$100 < d \leqslant 120$	$120 < d \leqslant 140$	$140 < d \leqslant 220$
Mid-point (x)	110	130	180
Frequency (f)	46	33	20

Calculate the grouped mean of these data.

It is a good idea to draw a rough sketch of the data in order to 'get a feel' for the data. A glance at the (not so rough!) sketch suggests that the distribution has centre of mass at about 100 thousand miles. If our calculated answer is very different from this then it should be checked for a possible error.

Consider the 32 buses that travelled less than 60 000 miles before breaking down. Each breakdown occurred somewhere between 0 miles and 60 000 miles so a sensible estimate of the average distance travelled by these buses would be 30 000 miles. Hence an estimate of the total distance travelled by those 32 buses would be $32 \times 30\,000 = 960\,000$ miles. Repeating for each of the classes, our overall estimate of the total mileage is $\Sigma f_i x_i = 18\,720\,000$, and hence the grouped mean, $\dfrac{18\,720\,000}{32}$, is about 98 500 miles.



Histogram of the distances travelled by a fleet of buses before breakdown

Calculator practice

Most statistical calculators can be used to calculate the mean of grouped data.
Test your calculator using the data given above.

Computer project

It is easy to program a spreadsheet to calculate the mean of a set of grouped data.
Test your program using the previous data.

Exercises 2c

1 A marine biologist is studying the population of limpets on a rocky coast. The numbers of a rare type of limpet that are found in 1 m square sections of the undercliff are summarised in the table below.

No. of limpets	0	1	2	3	4
No. of squares	73	19	5	2	1

Calculate the mean number of limpets per square metre of undercliff.

2 A proofreader reads through a 250-page manuscript. The numbers of mistakes found on each page are summarised in the table below.

No. of mistakes	0	1	2	3	4
No. of pages	61	109	53	23	4

Determine the mean number of errors found per page.

3 Construct a frequency distribution for the following data:

5, 7, 5, 3, 1, 4, 5, 4, 3, 2, 1, 3, 4, 5, 7, 6
8, 4, 3, 1, 5, 3, 5, 7, 3, 2, 4, 2, 6, 5, 2, 2

Find the mean and the median.

4 Construct a frequency distribution for the following data:

20, 30, 35, 25, 20, 30, 35, 25,
20, 30, 35, 40, 30, 35, 35, 25,
20, 40, 20, 25, 25, 30, 20, 20

Find the mean and the median.

5 A shop sells light bulbs. Mr Watt, the proprietor, makes a note one week of the wattage of the bulbs that he sells. At the end of the week he has noted the following:

100, 100, 100, 60, 100, 40, 150,
60, 100, 100, 100, 60, 40, 150,
100, 100, 100, 60, 100, 60

Construct a frequency distribution for the data and obtain (i) the mean, (ii) the median.

6 A sales representative records his daily mileage (in completed miles) for a period of 4 weeks:

153, 127, 142, 82, 91, 125, 113,
105, 93, 105, 88, 122, 96, 145,
136, 115, 107, 125, 98, 94

Group the data using class intervals of width 20, giving classes of 80–99, 100–119, etc.
Find:
(i) the grouped mean,
(ii) the modal class.

7 A garage notes the mileages of cars brought in for a 15 000-mile service. The data is summarised in the following table.

Mileage ('000 miles)	14–	15–	16–	17–
No. of cars	8	15	13	9

Assuming that the upper limit of the final class is 17 999, find (i) an estimate for the mean, (ii) the modal class.

8 Each day, x, the number of diners in a restaurant was recorded and the following grouped frequency table was obtained.

x	16–20	21–25	26–30	31–35	36–40
No. of days	67	74	38	39	42

Using the above grouped data find:
(i) an estimate of the mean,
(ii) the modal class.

9 A die is rolled twenty times with the following results.

Outcome	1	2	3	4	5	6
Frequency	2	4	a	7	2	b

Given that the mean is 3.6, obtain the values of a and b.

10 Subsidies for loft insulation are offered to households whose net income is less than £25 000 per annum. Applicants for these subsidies classified their incomes as follows.

Annual income	No. of applicants
–£4999	3
£5000–£9999	17
£10 000–£14 999	31
£15 000–£19 999	28
£20 000–£24 999	16

Determine the value of the grouped mean.

2.9 Using coded values to simplify calculations

Consider the problem of finding the mean of the following values:

3001, 3003, 3005, 3005, 3007, 3007, 3007, 3009

We could calculate:

$$\frac{1}{8}\{3001 + 3003 + (2 \times 3005) + (3 \times 3007) + 3009\} = 3005.5$$

but this needs a calculator and lots of button pressing. It is much easier to calculate:

$$3000 + \frac{1}{8}\{1 + 3 + (2 \times 5) + (3 \times 7) + 9\} = 3005.5$$

As a second example, consider the problem of finding the mean of:

0.000 01, 0.000 03, 0.000 05, 0.000 05,
0.000 07, 0.000 07, 0.000 07, 0.000 09

We could calculate:

$$\frac{1}{8}\{0.000\,01 + 0.000\,03 + (2 \times 0.000\,05) + (3 \times 0.000\,07) + 0.000\,09\} = 0.000\,055$$

but it is much easier to calculate:

$$0.000\,01 \times \frac{1}{8}\{1 + 3 + (2 \times 5) + (3 \times 7) + 9\} = 0.000\,055$$

Both the examples above have used **coded** data. Algebraically, we replaced the observations $x_1, x_2, \ldots$ by the coded values $y_1, y_2, \ldots$ In the first example, $y_i = x_i - 3000$ and in the second example, $y_i = 100\,000x_i$. The first example used a shift of location and the second a change of scale.

These two ideas may be combined. Suppose we want to find the mean of the following data:

10 500, 11 500, 12 500, 12 500, 13 500, 13 500, 13 500, 14 500

Writing $y = \dfrac{x - 10\,000}{500}$ we once again get the values 1, 3, 5, 5, 7, 7, 7 and 9 which have mean 5.5. Since $x = 10\,000 + 500y$, the mean of the x-values is:

$$10\,000 + (500 \times 5.5) = 12\,750$$

A general shift of location and change of scale is represented algebraically by the (linear) coding:

$$y = \frac{x - a}{b}$$

For convenience b is taken to be positive. When rewritten this expression gives:

$$x = a + by$$

and the mean, $\bar{x}$, is related to the mean, $\bar{y}$, of the coded values by

$$\bar{x} = a + b\bar{y}$$

In the first example $a = 3000$, $b = 1$; in the second $a = 0$, $b = \dfrac{1}{10\,000}$ and in the third $a = 10\,000$, $b = 500$.

Note

- Working with coded values usually saves time. Since calculators and computers place limits on the numbers of significant figures that they can handle, using coded values can also improve accuracy.

Example 6

The jackets on display in the window of a men's outfitters have the following prices (in £):

49.95, 79.95, 79.95, 99.95, 139.95

Use a coding method to determine the average jacket price.

Let x be the displayed price. A useful coding is $y = x + 0.05$. The prices then become 50, 80, 80, 100, 140. The sum of the 5 y-values is 450, so $\bar{y} = 90$. Thus $\bar{x} = \bar{y} - 0.05 = 90 - 0.05 = 89.95$.
The average price is £89.95.

Example 7

A bus inspector notes the numbers of passengers on buses travelling on a certain route. He records the following values:

31, 45, 40, 38, 39, 42, 36, 38, 44, 39, 32, 32, 38

Using the coding $y = x - 30$, determine the mean of these data.

Taking the observed values to be x, the y-values are:

1, 15, 10, 8, 9, 12, 6, 8, 14, 9, 2, 2, 8

These are simple numbers that won't strain our powers of mental arithmetic! Their total is 104 and $n = 13$, so that $\bar{y} = \frac{104}{13} = 8$ and hence $\bar{x} = \bar{y} + 30 = 38$.

The mean number of passengers is 38.

Example 8

A manufacturer wished to test the accuracy of the '2000 ohm' resistors being produced by a machine. A random sample of 100 resistors was selected and their actual resistances were determined (correct to the nearest ohm). The results are shown in the table.

Resistance	Frequency
1995	1
1996	3
1997	5
1998	9
1999	19
2000	21
2001	16
2002	15
2003	4
2004	4
2005	2
2006	1

Determine the mean resistance of these resistors.

The values are clustered around the nominal value of 2000. A sensible coding is therefore provided by $y = x - 2000$, where x is the recorded resistance.

x	y	f	fy	Total
1995	−5	1	−5	
1996	−4	3	−12	
1997	−3	5	−15	
1998	−2	9	−18	
1999	−1	19	−19	−69
2000	0	21	0	
2001	1	16	16	
2002	2	15	30	
2003	3	4	12	
2004	4	4	16	
2005	5	2	10	
2006	6	1	6	90
Total		100		21

Notice the way that the negative values are summed separately from the positive values. The overall total is $90 - 69 = 21$ and so $\bar{y} = \frac{21}{100} = 0.21$. Since $\bar{x} = \bar{y} + 2000$, the mean resistance is 2000.21 ohms.

An alternative coding, that would avoid negative numbers, would be $y = x - 1995$.

Calculator practice

Compare the speed and accuracy of calculating the mean of the two sets of data given above using (i) the actual values and (ii) the coded values. You should find that using the coded values you work both more quickly and more accurately.

Exercises 2d

1 Given that the numbers 3, 5, 6, 14 and 12 have mean 8, write down the mean of each of the following sets of numbers:
(i) 1003, 1005, 1006, 1014, 1012
(ii) 2.03, 2.05, 2.06, 2.14, 2.12
(iii) 1030, 1050, 1060, 1140, 1120

2 Find the mean, median and mode of the following observations:

1.000 000 002, 1.000 000 005,
1.000 000 006, 1.000 000 003,
1.000 000 009, 1.000 000 006,
1.000 000 005, 1.000 000 006,
1.000 000 003

3 The valuations (£x) of a collection of 12 antiques are reported as being:

600, 680, 1000, 750, 600, 850,
1000, 880, 1000, 650, 600, 1000

Use the coding $y = \frac{1}{10}(x - 600)$ to find the mean value of y and hence determine the mean valuation.

4 A choirmaster keeps a record of the numbers turning up for choir practice.

25, 28, 32, 31, 31, 34, 28, 31, 29,
28, 32, 32, 30, 29, 29, 31, 28, 28

Using a coding with each number reduced by 20, find the mean number turning up for choir practice.

5 The prices (£x) of pairs of shoes in the window display of a shoe shop are given below.

34.95, 44.95, 49.95, 69.95,
54.95, 64.95, 64.95, 54.95

(i) Using the coding $y = \frac{1}{5}(x + 0.05)$, determine the mean price.
(ii) Verify that the same result is obtained using the coding $y = x - 54.95$

6 The prices (in £) of various Indian dishes in a supermarket are given below.

1.99, 3.99, 2.99, 2.99, 1.99, 2.49, 1.99, 2.49

Using an appropriate coding, determine the mean price of these dishes.

7 The scores obtained by the leading 50 competitors in the first round of the 1995 US Masters are summarised below:

Score	66	67	68	69	70	71	72	73	74
Frequency	3	2	2	7	7	9	7	9	4

Using a suitable coding, find the mean of these scores.

8 The numbers of matches in a box were counted for a sample of 25 boxes. The results were:

51, 52, 48, 53, 47, 48, 50, 51, 50,
46, 52, 53, 51, 48, 49, 52, 50, 48,
47, 53, 54, 51, 49, 47, 51

Use a coding in which 40 is subtracted from each number to find the mean number of matches in a box.

Use a coding in which 50 is subtracted from each number (giving some negative values) to find the mean number of matches in a box.

Verify that your two answers are the same.

9 Records are kept for 18 days of the midday barometric pressure, in millibars.

1022, 1016, 1032, 1008, 998, 985,
993, 1004, 1009, 1011, 1015, 1020,
1007, 1001, 995, 993, 975, 972

Using a suitable coding, find the mean midday barometric pressure.

10 The gap, x mm, in a sample of spark plugs was measured with the following results:

0.81, 0.83, 0.81, 0.81, 0.82, 0.80, 0.81,
0.83, 0.84, 0.81, 0.82, 0.84, 0.80

Use the coding $y = 100x - 80$ to find the mean gap.

11 A garage notes the mileages of cars brought in for a 15 000-mile service. The data is summarised in the following table.

Mileage ('000 miles)	14–	15–	16–	17–
No. of cars	8	15	13	9

Taking the groups as having mid-points 14 500, ..., 17 500, and using the coding $y = \frac{1}{1000}(x - 14\,500)$, where x is the mileage, find the grouped mean.

12 Each day, x, the number of diners in a restaurant was recorded and the following grouped frequency table was obtained.

x	16–20	21–	26–	31–	36–40
No. of days	67	74	38	39	42

Using the coding $y = \frac{1}{5}(x - 18)$, where x is the number of diners, estimate the mean number of diners per day.

13 A set of data is summarised by $n = 8$, $\Sigma(x - 5) = 7.2$
Find the mean of x.

14 The annual salaries (x £'000) of the employees of a company are summarised in the following table.

Salary	Frequency
$5 \leqslant x < 10$	35
$10 \leqslant x < 15$	42
$15 \leqslant x < 20$	58
$20 \leqslant x < 30$	14
$30 \leqslant x < 50$	3
$50 \leqslant x < 100$	1

Use the coding $y = \frac{1}{5}(x - 7.5)$ to find the grouped mean salary.

15 A set of data is summarised by $\sum_1^{12}(y + 0.5) = 0.234$
Find the mean of y.

16 A set of ten observations is such that $\Sigma(2x + 3) = 427$
Find the mean of x.

17 Given that $n = 8$ and $\Sigma\{2(z + 3)\} = 752$, find the mean of z.

2.10 The median of grouped data

We begin by forming a cumulative frequency distribution. The median can then be estimated using linear interpolation (shown in this section) or, less accurately, by reading off a value from a cumulative frequency diagram (shown in the next section). In either case, with grouped data and n observations, it is customary to use $\frac{n}{2}$ in the calculations rather than $\frac{n+1}{2}$ (though the resulting difference is most unlikely to affect one's view of the data!).

Example 9

Continuing with Example 5, we report the bus data using cumulative frequencies and upper class boundaries:

Distance ('000 miles)	460	480	4100	4120	4140	4220
Cumulative frequency	32	57	91	137	170	190

Estimate the distance exceeded by half the buses.

There are 190 buses and $\frac{190}{2} = 95$. Since 95 falls between 91 and 137, the median distance falls between 100 and 120 thousand miles. A total of $(137 - 91) = 46$ buses fall in this class. The median is estimated as:

$$100 + \frac{(95 - 91)}{(137 - 91)} \times (120 - 100) = 101.74 \text{ (to 2 d.p.)}$$

so the median is about 101 700 miles.

Example 10

During 1983, motorists in Adelaide in South Australia were subject to random tests for alcohol consumption. Measurements of blood alcohol content (BAC) were made in units of mg of alcohol per 100 ml of blood.

BAC: Upper class boundary	15	25	35	45	65
Cumulative frequency	397	785	1083	1298	1580

BAC: Upper class boundary	95	125	155	205	400
Cumulative frequency	1793	1903	1951	1989	2003

Estimate the median BAC value.

One half of 2003 is 1001.5, which lies between 785 and 1083. The median therefore lies between 25 and 35. The estimated value is given by:

$$25 + \frac{(1001.5 - 785)}{(1083 - 785)} \times (35 - 25) = 32.27 \text{ (to 2 d.p.)}$$

The median blood alcohol content is estimated as being about 32 mg per 100 ml of blood.

Calculator practice

Investigate whether your calculator is able to determine the median of grouped data. If there is not a preset sequence of key strokes available, then you may wish to write a short program to calculate the quantity.

2.11 Quartiles, deciles and percentiles

The median is a value that subdivides the ordered data into two halves. Further subdivision is also possible: the **quartiles** subdivide the data into quarters, the **deciles** provide a subdivision into tenths, and the **percentiles** provide a subdivision into hundredths. There are three quartiles: the **lower quartile**, Q_1, the median (Q_2), and the **upper quartile**, Q_3. The percentiles are simply called the 1st percentile, the 2nd percentile, and so on. The median is the 5th decile and the 50th percentile. A study of the values of the deciles or quartiles gives us an idea of the spread of the data, but an 'idea' is all we get and there is no need for great precision.

Grouped data

With grouped data, life is straightforward! In general, the rth percentile is the '$\left(\frac{rn}{100}\right)$th' observation. The median is therefore the '$\left(\frac{n}{2}\right)$th' observation as in the previous section), while the quartiles are the '$\left(\frac{n}{4}\right)$th' and '$\left(\frac{3n}{4}\right)$th' observations. We have used inverted commas as a reminder that interpolation will usually be needed (though it would be inappropriate to report the value obtained to any great accuracy).

Example 11

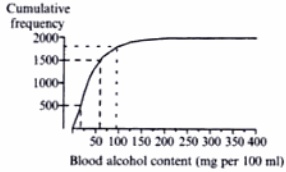
Determine the lower and upper quartiles and the 9th decile of the Adelaide motorists' data in Example 10:

BAC: Upper class boundary	15	25	35	45	65
Cumulative frequency	397	785	1083	1298	1580

BAC: Upper class boundary	95	125	155	205	400
Cumulative frequency	1793	1903	1951	1989	2003

Since the data are grouped we can use linear interpolation within the groups or we can attempt to read the figures off the cumulative frequency diagram. We first attempt to use the diagram.

Cumulative frequency diagram of blood alcohol content of Adelaide motorists in 1983



The lower and upper quartiles correspond to cumulative frequencies of $2003 \times 0.25 = 501$ and $2003 \times 0.75 = 1502$. Reading off the diagram (with considerable difficulty!) we find that these correspond to about 20 and 60. The 9th decile (the 90th percentile) corresponds to a cumulative frequency of $2003 \times 0.90 = 1803$ and, from the diagram, has a value of about 100.

We now use interpolation. For the lower quartile we have the estimate:

$$15 + \frac{(501 - 397)}{(785 - 397)} \times (25 - 15) = 18$$

and, for the upper quartile we have:

$$45 + \frac{(1502 - 1298)}{(1580 - 1298)} \times (65 - 45) = 59$$

For the 9th decile, the same approach gives:

$$95 + \frac{(1803 - 1793)}{(1903 - 1793)} \times (125 - 95) = 98$$

Ungrouped data

In Section 2.3 (p.35) the definition given for the median of ungrouped data was quite complicated! It does not look much better when expressed in another way: the median of ungrouped data is the '$\left(\frac{n}{2} + \frac{1}{2}\right)$th' observation, where the inverted commas serve as a reminder that interpolation may be needed.

Noting that $\frac{n}{2} = \frac{50n}{100}$, we now generalise this result and define the rth percentile of ungrouped data to be the '$\left(\frac{rn}{100} + \frac{1}{2}\right)$th' observation.

With this definition the lower and upper quartiles are, respectively, the '$\left(\frac{n}{4} + \frac{1}{2}\right)$th' and '$\left(\frac{3n}{4} + \frac{1}{2}\right)$th' ordered observations.

Note

- *There are no universally agreed formulae for any of these quantities (except for the median).* However, since quartiles and percentiles are of limited use, this is not really a source of worry! There is no virtue in reporting values for the quartiles to great accuracy: they should be reported using at most one more decimal place than that given in the original data.

Example 12

The numbers of words in the first 18 sentences of Chapter 1 of *A Tale of Two Cities* by Charles Dickens are as follows:

118, 39, 27, 13, 49, 35, 51, 29, 68, 54, 58, 42, 16, 221, 80, 25, 41, 33

whilst the numbers of words in the first 17 sentences of Chapter 1 of *Not a Penny More, Not a Penny Less* by Jeffrey Archer are as follows:

8, 10, 15, 13, 32, 25, 14, 16, 32, 25, 5, 34, 36, 19, 20, 37, 19

Determine the median, quartiles and first decile for each data set.

Rearranging the Dickens data in order of magnitude we get:

Since $\frac{18}{2}+\frac{1}{2}=9.5$ the median is the average of the 9th and 10th ordered observations, namely $\frac{1}{2}(41+42)=41.5$.

Since $\frac{18}{4}+\frac{1}{2}=5$, the lower quartile is the 5th observation, namely 29.

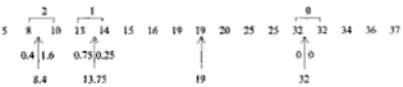
Since $\frac{3\times18}{4}+\frac{1}{2}=14$, the upper quartile is 58.

For the first decile, we need the '$\left(\frac{10\times18}{100}+\frac{1}{2}\right)$th' observation. Since $\frac{10\times18}{100}+\frac{1}{2}=2.3$, we need to interpolate between the 2nd and 3rd ordered observations, and the required value is:

$$16+\{0.3\times(25-16)\}=18.7$$

The first decile is about 19.

Rearranging the Archer data in order we get:



Since $\frac{17}{2}+\frac{1}{2}=9$ the median is the 9th largest observation, namely 19.

Since $\frac{17}{4}+\frac{1}{2}=4.75$, we must interpolate between the 4th and 5th ordered observations, getting $13+\{0.75\times(14-13)\}=13.75$. The lower quartile is about 14.

Since $\frac{3\times17}{4}+\frac{1}{2}=13.25$, we must interpolate between the 13th and 14th ordered observations. Since both of these are 32, the interpolated value will also be 32, which is therefore the value of the upper quartile.

For the first decile we need to calculate the value of $\frac{10\times17}{100}+\frac{1}{2}$, which is 2.2. Interpolating between the second and the third of the ordered observations we get $8+\{0.2\times(10-8)\}=8.4$. The first decile is about 8.

The difference in writing styles is evident!

Computer project

Write a computer program to calculate quartiles, deciles and percentiles. If your computer has graphical capabilities then the program could be extended to display the cumulative frequency diagram along with indications of the locations of the quartiles. A well-written program should automatically scale the axes so that the diagram fills the screen.

Exercises 2e

1 The gap, x mm, in a sample of spark plugs was measured with the following results:

0.81, 0.83, 0.81, 0.81, 0.82, 0.80, 0.81,
0.83, 0.84, 0.81, 0.82, 0.84, 0.80

Find the lower and upper quartiles for this data set.

2 The numbers of matches in a box were counted for a sample of 25 boxes. The results were:

51, 52, 48, 53, 47, 48, 50, 51, 50,
46, 52, 53, 51, 48, 49, 52, 50, 48,
47, 53, 54, 51, 49, 47, 51

Find the second and eighth deciles for this set of data.

3 Records are kept for 18 days of the midday barometric pressure, in millibars.

1022, 1016, 1032, 1008, 998, 985,
993, 1004, 1009, 1011, 1015, 1020,
1007, 1001, 995, 993, 975, 972

Find the values of the lower and upper quartiles.

4 Find the lower and upper quartiles and the 9th decile for the following data:

20, 30, 35, 25, 20, 30, 35, 25,
20, 30, 35, 40, 30, 35, 35, 25,
20, 40, 20, 25, 25, 30, 20, 20

5 Find the lower and upper quartiles and the 15th percentile for the following data:

5, 7, 5, 3, 1, 4, 5, 4, 3, 2, 1, 3, 4, 5, 7, 6,
8, 4, 3, 1, 5, 3, 5, 7, 3, 2, 4, 2, 6, 5, 2, 2

6 A baker keeps a count of the number of doughnuts sold each day for three weeks. The numbers are:

35, 47, 34, 46, 55, 82, 41, 35, 47,
51, 56, 75, 38, 41, 44, 51, 45, 74

By constructing a stem-and-leaf diagram, or otherwise, find the lower and upper quartiles of the number of doughnuts sold per day. Find also the 4th decile.

7 A garage notes the mileages of cars brought in for a 15 000-mile service. The data are summarised in the following table.

Mileage ('000 miles)	14–	15–	16–	17–
No. of cars	8	15	13	9

Find the lower and upper quartiles and the 5th and 20th percentiles.

8 Each day, x, the number of diners in a restaurant was recorded and the following grouped frequency table was obtained.

x	16–20	21–	26–	31–	36–40
No. of days	67	74	38	39	42

Treating x as though it is a continuous variable with class boundaries at 15.5, 20.5, 25.5, 35.5 and 40.5, find the lower and upper quartiles and the 2nd and 8th deciles.

9 In an investigation of delays at a roadworks, the times spent, by a sample of commuters, waiting to pass through the roadworks were recorded to the nearest minute. Shown below is part of a cumulative frequency table resulting from the investigation.

Upper class boundary	2.5	4.5	7.5	8.5	9.5
Cumulative number of commuters	0	6	21	48	97

Upper class boundary	10.5	12.5	15.5	20.5
Cumulative number of commuters	149	178	191	200

(a) For how many of the commuters was the time recorded as 11 minutes or 12 minutes?

(b) Estimate (i) the lower quartile, (ii) the 81st percentile, of these waiting times.

[ULEAC]

10

Volume (in litres) of petrol	Number of sales
5 or less	6
10 or less	20
15 or less	85
20 or less	148
25 or less	172
30 or less	184
35 or less	194
40 or less	200

The table gives an analysis of a random sample of 200 sales of unleaded petrol at a petrol station.

(*a*) Using scales of 2 cm to 5 litres on the horizontal axis and 2 cm to 20 sales on the vertical axis, draw a cumulative frequency curve for the data.

(*b*) Use your curve to estimate the median volume of unleaded petrol sales.

Unleaded petrol is sold for 52.3p per litre. Use your curve to estimate

(*c*) the 40th percentile of the value of unleaded petrol sales,

(*d*) the percentage of sales above £12.

[ULSEB]

2.12 Range, interquartile range and midrange

The **range** of a set of numerical data is the difference between the highest and lowest values. It is the simplest possible measure of spread. It cannot be used with grouped data and it ignores the distribution of intermediate values. A single very large or very small value would give a misleading impression of the spread of the data. This happens with the Dickens data where the range $(221 - 13 = 208)$ gives a distorted impression because of the single unusually long sentence.

More useful, because it concentrates on the middle portion of the distribution, is the **interquartile range (IQR)** which is the difference between the upper and lower quartiles. The **semi-interquartile range** is sometimes quoted: it is half the interquartile range.

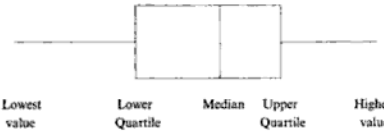
For the Dickens data of Example 9, the interquartile range is $Q_3 - Q_1 = 58 - 29 = 29$, and the semi-interquartile range is 14.5. The less variable Archer data has a semi-interquartile range equal to 9.1 (to 1 d.p.).

The highest and lowest values obviously bracket the remaining data so a simple alternative to the mean or median as a measure of the location of the data is provided by the **midrange** $= \frac{1}{2}$(highest value + lowest value). However, although simple, this measure is rarely used because (unlike the mean and the median) it is sensitive to outliers (see Section 2.13).

2.13 Box–whisker diagrams

Box–whisker diagrams present a simple picture of the data based on the values of the quartiles. They are also known as **boxplots**. The general form of a box–whisker diagram is shown in the diagram.

The general form of a box-whisker diagram



Box–whisker diagrams provide a particularly convenient way of comparing two distributions.

Note

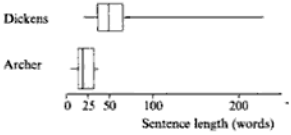
- There is no agreed rule for determining the thickness of the box. When comparing samples a sensible procedure would be to make the box areas proportional to the sample sizes. The thicknesses of the boxes are therefore in proportion to the ratios of the respective sample sizes divided by the corresponding IQR.

Example 13

Use box–whisker diagrams to compare the sentence lengths of Dickens and Archer for the data of Example 12.

Following the previous note, we give the boxes areas in the ratio 18 to 17. The interquartile ranges were 29 and 18.25, so we use thicknesses in proportion to $\frac{18}{29} : \frac{17}{18.25}$. Note that this is an optional extra!

Box–whisker diagrams comparing the sentence lengths of two authors



The difference in the distributions of the sentence lengths is very apparent.

Refined boxplots

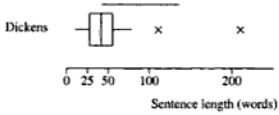
A useful refinement of the simple box–whisker diagram highlights any unusually extreme data values (which are known as **outliers** and should be examined for possible transcription or other errors).

In these **refined boxplots** all values lying outside specified limits (typically $Q_1 - 1.5(Q_3 - Q_1)$ and $Q_3 + 1.5(Q_3 - Q_1)$) are indicated individually using crosses.

With these limits the upper whisker stretches from Q_3 to the largest data value that is smaller than $Q_3 + 1.5(Q_3 - Q_1)$, while the lower whisker stretches from Q_1 to the smallest value that is greater than $Q_1 - 1.5(Q_3 - Q_1)$.

Example 14

Draw a refined boxplot for the sentence lengths of Dickens given in Example 12 (see also Example 13).



For the sentences from *A Tale of Two Cities* we had $Q_3 = 58$ and $Q_1 = 29$. Thus $Q_3 + 1.5(Q_3 - Q_1) = 58 + 1.5 \times (58 - 29) = 101.5$. The values 118 and 221 exceed 101.5 and are therefore outliers. The largest data value that is smaller than 101.5 is 80, so the upper whisker extends from 58 to 80.

To examine the lower values we first calculate $Q_1 - 1.5(Q_3 - Q_1) = 29 - 1.5(58 - 29) = -14.5$. There are no sentences shorter than this (!) and thus the bottom whisker extends from 29 to the smallest observed value, 13.

Project

Choose two of your own favourite authors and repeat the Dickens/Archer experiment. Try to choose authors whose styles you think may be different. Choose descriptive passages rather than passages of dialogue, but don't choose them because they seem to have particularly long (or short) sentences, or you will bias the results! Construct box–whisker diagrams for each author. Do there seem to be differences in the two distributions of sentence length?

Exercises 2f

1 The following data are the numbers of deaths of army officers caused by horse kicks, for the Prussian Army during the period 1875 to 1894. In order of size the numbers are:

3, 4, 5, 5, 6, 6, 7, 8, 9, 9, 10,
11, 11, 11, 12, 14, 15, 15, 17, 18

Find the range, interquartile range and midrange.
Illustrate the data using a box-whisker diagram.

2 One year the numbers of academic staff (including part-time staff) in the various departments of the University of Essex (a small, friendly university) were as follows:

19.0, 15.7, 25.3, 28.0, 15.0, 10.0,
12.0, 10.3, 22.0, 24.8, 13.8, 25.9,
23.0, 21.3, 12.0, 11.0, 23.0

Find the range, interquartile range and midrange.
Illustrate the data using a box–whisker diagram.

3 The record times (in hours) for marathon sessions of various games, as reported in the 1986 *Guinness Book of Records*, are as follows:

Backgammon 151, Bridge 180, Chess 200,
Darts 133, Draughts 108, Monopoly 660,
Pool 300, Scrabble 153, Snooker 301,
Table tennis 148, Tiddlywinks 300

Illustrate the data using a refined boxplot.

4 One year the numbers of undergraduates in the various departments of that friendly University of Essex were as shown below:

173, 166, 255, 225, 107, 146, 199, 107,
348, 327, 236, 390, 424, 252, 125, 161, 343

Illustrate the data using a refined boxplot.

5 The systolic blood pressures of 12 smokers and 12 non-smokers are as follows (in the standard units):

Smokers:
122, 146, 120, 114, 124, 126,
118, 128, 130, 134, 116, 130

Non-smokers:
114, 134, 114, 116, 138, 110,
112, 116, 132, 126, 108, 116

Contrast these two sets of data using side-by-side refined boxplots.

6 One year the number of overseas postgraduate students in the departments of a certain university (you know where!) were as given below:

13, 8, 34, 24, 26, 22, 14, 44,
0, 104, 19, 26, 9, 7, 41, 57, 6

Illustrate these data using a refined boxplot.

7 The numbers of students on degree schemes involving mathematics at a really excellent university are as follows:

22, 11, 5, 20, 15, 13, 3, 2, 5, 1, 2

Determine the range and interquartile range and illustrate the data using a refined boxplot.

8 The times (in s) taken for a group of experienced rats to run through a maze are to be compared with the times for a group of inexperienced rats. The data are:

Experienced rats:
121, 137, 130, 128, 132, 127, 129,
131, 135, 130, 126, 120, 118, 125

Inexperienced rats:
135, 142, 145, 156, 149, 134, 139,
126, 147, 152, 153, 145, 144

(i) Summarise the two data sets using stem-and-leaf diagrams.
(ii) Find the median and the upper and lower quartiles for each group of rats.
(iii) Plot the two sets of data on a single graph using boxplots.
Comment on the results.

9 A random sample of size 500 was selected from the persons listed in a residential telephone directory of a county in Wales. The number of letters in each surname was counted and the distribution of the name-lengths is given in the table below.

Name-length	3	4	5	6	7	8	9	10	11
Frequency	4	31	103	124	111	63	38	19	7

(a) (i) Represent this distribution graphically.
(ii) Find the median and the interquartile range of the distribution.

(b) The 500 persons in the sample are unlikely to be representative of the population of the county.
Name one group of people which is likely to be under-represented.

(c) State, with a reason, whether the sample of surnames is likely to be representative of the surnames of the population of this county in Wales. [NEAB]

10 A random sample of 51 people were asked to record the number of miles they travelled by car in a given week. The distances, to the nearest mile, are shown below.

67	76	85	42	93	48	93	46	52
72	77	53	41	48	86	78	56	80
70	70	66	62	54	85	60	58	43
58	74	44	52	74	52	82	78	47
66	50	67	87	78	86	94	63	72
63	44	47	57	68	81			

(a) Construct a stem and leaf diagram to represent these data.
(b) Find the median and the quartiles of this distribution.
(c) Draw a box plot to represent these data.
(d) Give one advantage of using
(i) a stem and leaf diagram,
(ii) a box plot, to illustrate data such as that given above. [ULEAC]

11 The following table, extracted from *Welsh Social Trends*, shows the distribution of the number of persons per household in Wales in 1981.

Number of persons in household	1	2	3	4	5	6+
Percentage of households	21	31	18	18	8	4

(i) By plotting the cumulative percentage step polygon, or otherwise, determine the median and the semi-interquartile range of the number of persons per household.

(ii) State, giving your reason, whether the mean number of persons per household is greater than, equal to, or less than the median number. [WJEC]

2.14 Deviations from the mean

Suppose we wish to summarise the following data:

0, 99, 99, 100, 100, 100, 100, 100, 101, 101, 200

This set of data has mean, median and mode equal to 100, lower quartile equal to 99, upper quartile equal to 101 and range equal to 200. The same is true for this second set of data:

0, 0, 99, 99, 100, 100, 100, 101, 101, 200, 200

However, this second set of data has four extreme observations, compared with only two in the first set. This extra variability can be quantified by calculating the differences between the observations and their mean:

Set 1: −100, −1, −1, 0, 0, 0, 0, 0, 1, 1, 100
Set 2: −100, −100, −1, −1, 0, 0, 0, 1, 1, 100, 100

In each case the differences sum to zero. This always happens since, for a set of n observations $x_1, \ldots, x_n$ with sample mean $\bar{x}$, given by $n\bar{x} = \sum x_i$:

$$\begin{aligned}\sum_{i=1}^{n} (x_i - \bar{x}) &= (x_1 - \bar{x}) + \cdots + (x_n - \bar{x}) \\ &= (x_1 + \cdots + x_n) - (\bar{x} + \cdots + \bar{x}) \\ &= \sum_{i=1}^{n} x_i - n\bar{x} \\ &= n\bar{x} - n\bar{x} \\ &= 0\end{aligned}$$

2.15 The variance

A natural measure of spread is provided by the sum of the squares of the deviations from the mean:

$$\Sigma(x_i - \bar{x})^2 = (x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2$$

The more variation there is in the x-values, the larger will be the value of $\Sigma(x_i - \bar{x})^2$. However, the sum might be large simply because of the number of x-values, and some sort of average value is needed.

Dividing by n would seem natural, but (unfortunately!) there is a strong case for dividing instead by $(n - 1)$.

Using the divisor n

This is appropriate in two cases:

1. If the values $x_1, \ldots, x_n$ represent an entire population.
2. If the values $x_1, \ldots, x_n$ represent a sample from a population and we are interested in *the variation within the sample itself*.

In both cases the n observed values are all that interest us and the natural average squared deviation, denoted by σ_n^2, is given by

$$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2.9}$$

The quantity σ_n^2 should be read as 'sigma n squared'.

If $x_1, \ldots, x_n$ represent a sample of data then σ_n^2 is called the **sample variance**, while if $x_1, \ldots, x_n$ represent the entire population then σ_n^2 is called the **population variance**.

An example of a case where the x-values refer to the entire population is where $x_1, \ldots, x_n$ represent the heights of *all* the children in a particular class in a school. If, for some reason, we are *only* interested in this class then σ_n^2 is appropriate.

Using the divisor $(n - 1)$

This is appropriate in the following case:

> The values $x_1, \ldots, x_n$ represent a sample from a population and we are interested in estimating *the variation in the population*. The sample is important only because it gives information about the larger population.

For example, we might collect information about the heights of the children in a class so as to gain an impression of the distribution of the heights of children in corresponding classes nationwide.

In this case, a slight adjustment is made to the formula by dividing by $(n-1)$ instead of by n. The revised quantity is sometimes denoted by σ_{n-1}^2, sometimes by $\hat{\sigma}^2$, but more commonly by s^2:

$$s^2 = \sigma_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.10}$$

We will call this quantity the **unbiased estimate of the population variance** (and will use the s^2 notation).

Notes

- Since s^2 and σ_n^2 are positive multiples of a sum of squares:
  –they cannot have negative values,
  –they have units which are squares of the units of x.
- If s^2 or σ_n^2 is equal to zero then each of the x-values must be equal to the mean, $\bar{x}$, and therefore also equal to each other.
- Except when both are zero, $s^2 > \sigma_n^2$.
- Practising statisticians seldom use the divisor n because they are interested in drawing inferences about a population from a sample. The important questions are those about the unseen population rather than the particular sample observed.

Special note

- There is considerable variation from book to book, from exam board to exam board and from one set of statistical tables to another, concerning the names and symbols to be used for the two forms of variance formula introduced above. The quantity with divisor n, which we denote by σ_n^2, is denoted by $\hat{s}^2$ in one set of tables and by s^2 in another, both of which call it the 'sample variance'. Another set of tables uses S^2 for the same quantity and calls it the 'unadjusted variance'. A majority of tables use s^2 as we do, to denote the quantity with divisor $n-1$. There is also a general agreement that s^2 should be referred to as '**the unbiased estimate of the population variance**', although it too is often called the 'sample variance'.
  You should consult the formula sheet, tables, or syllabus for your exam board to be sure which formula you will be expected to use.

2.16 Calculating the variance

If $\bar{x}$ is an integer then the values of $(x_1 - \bar{x})^2, \ldots, (x_n - \bar{x})^2$ will be quite easy to calculate. However, $\bar{x}$ will usually be an awkward decimal and it is much easier to use the result:

$$\Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} \tag{2.11}$$

Hence:

$$\sigma_n^2 = \frac{\Sigma x_i^2}{n} - \frac{(\Sigma x_i)^2}{n^2}$$

and:

$$s^2 = \sigma_{n-1}^2 = \frac{1}{n-1}\left\{\Sigma x_i^2 - \frac{1}{n}(\Sigma x_i)^2\right\} = \frac{n}{n-1}\,\sigma_n^2$$

Notes

- In all important cases the quantities Σx_i^2 and $(\Sigma x_i)^2$ are *not* equal, since:

  $$\Sigma x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2$$

  whereas:

  $$(\Sigma x_i)^2 = (x_1 + x_2 + \cdots + x_n)^2$$
  $$= (x_1^2 + x_2^2 + \cdots + x_n^2) + 2(x_1x_2 + x_1x_3 + \cdots + x_{n-1}x_n)$$

- The proof of the result in Equation (2.11) requires some messy algebra:

$$\begin{aligned}\Sigma(x_i-\bar{x})^2 &= \Sigma(x_i^2-2x_i\bar{x}+\bar{x}^2)\\ &= \Sigma x_i^2-2\bar{x}\Sigma x_i+n\bar{x}^2\\ &= \Sigma x_i^2-2\left(\frac{\Sigma x_i}{n}\right)\Sigma x_i+n\left(\frac{\Sigma x_i}{n}\right)^2\\ &= \Sigma x_i^2-\frac{1}{n}(\Sigma x_i)^2\end{aligned}$$

- Another way of writing $\Sigma(x_i-\bar{x})^2$ is as $\Sigma x_i^2-n\bar{x}^2$, but for numerical calculations it is usually more accurate to calculate $\frac{1}{n}(\Sigma x_i)^2$ than to calculate $n\bar{x}^2$. The latter form is more useful in algebraic manipulations.

2.17 The sample standard deviation

We define the sample standard deviation, σ_n, as being the square root of the sample variance, σ_n^2.

$$\sigma_n=\sqrt{\frac{\Sigma x_i^2}{n}-\frac{(\Sigma x_i)^2}{n^2}} \tag{2.12}$$

Notes

- The units of the standard deviation are the same as the units of x – i.e. if x is a number of apples then so is σ_n.
- In some books you may come across s, the square root of s^2, being described as the 'sample standard deviation'.
- The words 'standard deviation' are often abbreviated to s.d.

Calculator practice

Calculators with statistical functions will calculate one or both of σ_n and $s(=\sigma_{n-1})$. You should check which statistic(s) your calculator provides. Usually the values of n, Σx and Σx^2 will have been calculated and stored in accessible memories in the process. You should be aware of where these quantities are stored and how they can be accessed.

Example 15

The nine planets of the solar system have approximate equatorial diameters (in thousands of km) as follows:

4.9, 12.1, 12.8, 6.8, 142.8, 120.0, 52.4, 49.5, 2.5

Determine the standard deviation of these diameters.

We begin by calculating $\Sigma x_i=(4.9+\cdots+2.5)=403.8$ and $\Sigma x_i^2=(4.9^2+\cdots+2.5^2)=40\,374.6$. The mean diameter is $\frac{1}{9}\times 403.8=44.87$ thousand kilometres. Assuming that we are interested in these nine planets for their own sake rather than for what they may imply about planets elsewhere in the universe, we now calculate:

$$\sigma_n = \sqrt{\frac{\Sigma x_i^2}{n} - \frac{(\Sigma x_i)^2}{n^2}}$$

$$= \sqrt{\frac{40\,374.6}{9} - \frac{403.8^2}{9^2}}$$

$$= \sqrt{4486.066\,667 - 2013.017\,778}$$

$$= \sqrt{2473.048\,889}$$

$$= 49.73 \text{ (to 2 d.p.)}$$

The standard deviation of the equatorial diameters is about 50 thousand kilometres.

Notes

- The working is carried through to considerable accuracy to guard against **round-off errors** and against loss of significance, since the calculations often involve determining the relatively small difference between two large numbers. With inaccurate calculations

$$\frac{\Sigma x_i^2}{n} - \frac{(\Sigma x_i)^2}{n^2}$$

may *appear* to be negative due to loss of significant figures!
- When reporting results a reasonably accurate value (49.73) should be easily available, but the description (50) of the result should be as simple as possible. In most situations there will be little interest in the difference between 49.73 and 50!

Example 16

An office manager wishes to get an idea of the number of phone calls received by the office during a typical day. A week is chosen at random and the numbers of calls on each day of the (5-day) week are recorded. They are as follows:

15, 23, 19, 31, 22

Determine (i) the sample mean, (ii) the sample standard deviation, (iii) s^2, the unbiased estimate of the population variance.

(i) For these data $\Sigma x = 110$, $\Sigma x^2 = 2560$. The mean is $\frac{110}{5} = 22$.
(ii) The calculation of the sample standard deviation takes a little longer:

$$\sigma_n = \sqrt{\frac{\Sigma x_i^2}{n} - \frac{(\Sigma x_i)^2}{n^2}}$$

$$= \sqrt{\frac{2560}{5} - \frac{110^2}{25}}$$

$$= \sqrt{28}$$

$$= 5.29 \text{ (to 2 d.p.)}$$

The sample standard deviation is about 5.

(iii) This requires the $(n-1)$ divisor:

$$s^2 = \frac{1}{n-1}\left\{\Sigma x_i^2 - \frac{1}{n}(\Sigma x_i)^2\right\}$$

$$= \frac{1}{4}\left(2560 - \frac{110^2}{5}\right)$$

$$= \frac{140}{4}$$

$$= 35$$

The unbiased estimate of the population variance, s^2, is equal to 35.

Approximate properties of the standard deviation

Providing the sample size is reasonably large and the data are not too skewed (i.e. there is not a long 'tail' of very large or very small values) it is possible to make the following approximate statements which are based on theory covered later in Chapter 10:

- About two-thirds of the individual observations will lie within one standard deviation of the sample mean.
- For most data sets about 95% of the individual observations will lie within two standard deviations of the sample mean. Observations that are more than two standard deviations from the sample mean may be regarded as **outliers** (see also Section 2.13).
- Almost all the data will lie within three standard deviations of the sample mean.
- A useful check that your calculations have not gone hopelessly wrong is provided by noting that the standard deviation will usually be between a third and a sixth of the range.

These are *very approximate* statements which enable us to check our calculations. Because they are approximate we need not worry whether we are using s or σ_n (hurrah!).

As an example of their use, suppose that the observed data consists of values ranging between 0 and 30. We expect a mean of about 15 (since this is half-way between 0 and 30) and a standard deviation of between 5 and 10. If our calculations find a standard deviation of 4 then this should not worry us, but if we calculate a value of 40, then we will certainly have made a mistake.

The statements also enable us to draw inferences about the population from which the data has been sampled. As stated at the beginning of Chapter 1, this is the principal purpose of Statistics.

Example 17

Use the approximate properties of the standard deviation to make statements concerning the likely numbers of daily phone calls received by the office featured in Example 16.

Here the sample is very small, so we cannot place too much reliance on our approximations.

The range of values observed was $31 - 15 = 16$, so we anticipate a mean of about $\frac{1}{2}(31 + 15) = 23$ and a standard deviation of between $\frac{1}{3} \times 16 = 5.3$ and $\frac{1}{6} \times 16 = 2.6$. The calculated mean and standard deviation were 22 and 5.29. It is reasonable to assume that we have not made a mistake!

The office manager can conclude that on two-thirds of days the office will receive between $22 - 6 = 16$ and $22 + 6 = 28$ calls (there is no point in using great precision since these are only very crude approximations).

Assuming that the week sampled was typical, the office is unlikely ever to receive fewer than $22 - (3 \times 6) = 4$ calls, or more than $22 + (3 \times 6) = 40$ calls.

Exercises 2g

1 A card player notes the number of hearts that she receives during a sample of 5 random deals. The numbers are 3, 2, 4, 4, 1.
Find the sample mean and the sample standard deviation.

2 The numbers of television licences bought at a particular Post Office on a sample of 5 randomly chosen weekdays were 15, 9, 23, 12, 17.
Find the mean and standard deviation of this sample.

3 The numbers of potatoes in a sample of 2 kg bags were 12, 15, 10, 12, 11, 13, 9, 14.
Find the mean and an unbiased estimate of the population variance.

4 During his entire life, Mr I Walton, a most unlucky angler, caught just six fish. Their masses, in kg, were 1.35, 0.87, 1.61, 1.24, 0.95, 1.87.
Find the mean and variance of this population.

5 A random sample of seven runner beans have lengths (in cm, to the nearest cm) given as 28, 31, 24, 33, 28, 32, 30.
Find the value of s.

6 In an experiment, a cupful of cold water is poured into a kettle and the time taken for the water to boil is noted. The experiment was conducted six times giving the following results (in seconds):

125, 134, 118, 143, 128, 131.

Find the value of s^2.

7 The midday temperature (in °C) was noted at an Antarctic weather station on every day of a particular week of the year. The results were -25, -18, -41, -34, -25, -33, -27. Treating these results as a population, find their mean and standard deviation.

8 A random sample has values summarised by

$$n = 8,\ \Sigma x = 671,\ \Sigma x^2 = 60\,304.$$

Find the mean and the value of s^2.

9 Twenty observations of t are summarised by

$$\Sigma t = 23.16,\ \Sigma t^2 = 35.4931.$$

Find the mean and the value of s^2.

10 A population is summarised by

$$\sum_{j=1}^{52} y_j = 3.751,\ \sum_{j=1}^{52} y_j^2 = 0.535\,691$$

Find the mean and the value of s.

11 A random sample is summarised by $n = 13$, $\Sigma u = -27.3$, $\Sigma u^2 = 84.77$.
Find the mean and sample standard deviation of u.

12 A random sample has $n = 11$, $s = 1.4$ and $\Sigma x^2 = 50$.
Find $\bar{x}$.

13 A random sample of 15 observations has sample mean 11.2 and sample variance 13.4. One observation of 21.2 is judged to be unreliable.
Find the sample mean and the sample variance of the remaining 14 observations.

2.18 Variance and standard deviation for frequency distributions

When data have been summarised in the form:

'the value x_j occurs with frequency f_j'

the formulae for the variance need rewriting. With m distinct values of x, the formula for the sample variance, σ_n^2, becomes:

$$\sigma_n^2 = \frac{1}{n}\left\{\sum_{j=1}^{m} f_j x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{m} f_j x_j\right)^2\right\} \tag{2.13}$$

and the formula for the unbiased estimate of the population variance becomes:

$$s^2 = \frac{1}{n-1}\left\{\sum_{j=1}^{m} f_j x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{m} f_j x_j\right)^2\right\} \tag{2.14}$$

where n is the total of the individual frequencies:

$$n = \sum_{j=1}^{m} f_j$$

As before, the sample standard deviation is simply the square root of the sample variance.

The same revised formulae are used when working with grouped data. In this case the x-values are the mid-points of the class intervals and the f-values are the class frequencies. The value obtained will usually be a slight underestimate of the true sample variance (or of the true value of s^2).

Example 18

Determine the variance of the marks obtained by 99 students which are summarised in the following grouped frequency table:

Mark range	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89
Mid-point (x)	14.5	24.5	34.5	44.5	54.5	64.5	74.5	84.5
Frequency (f)	8	18	25	22	16	6	3	1

We start by calculating:
$\Sigma f_i x_i = (8 \times 14.5) + \cdots + (1 \times 84.5) = 3965.5$ and
$\Sigma f_i x_i^2 = (8 \times 14.5^2) + \cdots + (1 \times 84.5^2) = 182\,084.75$. Thus:

$$\sigma_n^2 = \frac{1}{99}\left(182\,084.75 - \frac{3965.5^2}{99}\right) \approx 234.79$$

and $\sigma_n = 15.32$ (to 2 d.p.).

A quick check suggests that the calculations are correct since the range of the mid-points is 70 and 15.32 lies comfortably inside the predicted range of $\frac{70}{6} = 11.7$ to $\frac{70}{3} = 23.3$.

The mean is $\frac{1}{99} \times 3965.5 = 40.06$ (to 2 d.p.). Suppose at this stage that the original frequency table is mislaid! Applying the approximate rules we deduce that about two-thirds of the data are in the interval between $(40 - 15) = 25$ and $(40 + 15) = 55$, whilst almost all the data are in the interval -6 (!) to 86.

Calculator practice

If your calculator is described as 'statistical', then it can probably be used to calculate the mean, standard deviation and variance of grouped data. Find out the correct sequence of buttons to press!

Computer project

Computers love numbers! An advantage of a spreadsheet is that you can see what is happening: if you enter the wrong number it is likely to become obvious as the computer performs the calculations. Also, of course, it is easy to correct an error.
Write a program to calculate the mean and variance for the data of the previous example. Revise the data by reducing all the marks by 20.
What happens to the mean and variance?
What happens if you now double the previous marks?

2.19 Variance calculations using coded values

Earlier we introduced the general coding $y_i = \dfrac{x_i - a}{b}$, with $b > 0$. When rearranged, this gives

$$x_i = a + by_i$$

This coding resulted in the mean, $\bar{y}$, of the coded values being related to the original mean, $\bar{x}$, by

$$\bar{x} = a + b\bar{y}$$

so that $$x_i - \bar{x} = (a + by_i) - (a + b\bar{y}) = b(y_i - \bar{y})$$

Thus

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = b^2 \sum_{i=1}^{n} (y_i - \bar{y})^2$$

If we denote the sample variance of the x-values by σ_x^2, and the sample variance of the y-values by σ_y^2, then, on dividing the previous equation through by n on both sides, we get

$$\sigma_x^2 = b^2 \sigma_y^2$$

Notes

- Essentially the same formulae apply to grouped data. In this case the original x-values are the class mid-points.
- Writing s_x^2 and s_y^2 for the unbiased estimates of the population variances of the x-values and the y-values, the previous coding leads to the comparable result that

$$s_x^2 = b^2 s_y^2$$

Example 19

The protein content of milk depends upon a cow's diet. The following observations are the percentages of protein in the milk produced by 25 cows fed on a diet of barley:

3.73, 3.33, 3.25, 3.11, 3.53, 3.73, 3.42, 3.57, 3.13, 3.27, 3.60, 3.26, 3.40, 3.24, 3.63, 3.15, 3.00, 3.28, 3.84, 3.57, 3.35, 3.24, 3.66, 3.50, 3.47

Calculate the mean, variance and standard deviation of these data.

Calculation of the mean and variance of these data is simplified by using the coding $y = 100(x - 3)$, for which $a = 3$ and $b = \frac{1}{100}$. This results in the y-values:

73, 33, 25, 11, 53, 73, 42, 57, 13, 27, 60, 26, 40,
24, 63, 15, 0, 28, 84, 57, 35, 24, 66, 50, 47

For these y-values we have $\Sigma y_i = 1026$ so that $\bar{y} = \frac{1026}{25} = 41.04$. Hence $\bar{x} = 3 + \frac{1}{100}\bar{y} = 3.4104$. The mean is 3.41 (to 2 d.p.).

We also have $\Sigma y_i^2 = 53\,814$ so that:

$$\Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - \frac{1}{n}(\Sigma y_i)^2 = 53\,814 - \frac{1026^2}{25} = 11\,706.96$$

and hence $\sigma_y^2 = \frac{1}{25} \times 11\,706.96 = 468.28$. Hence:

$$\sigma_x^2 = \left(\frac{1}{100}\right)^2 \times 468.28 = 0.046\,828$$

The variance is therefore 0.047 (to 3 d.p.)

The standard deviation is $\sqrt{0.046\,828} = 0.216$ (to 3 d.p.).

Exercises 2h

1 The numbers of absentees in a class over a sample period of 24 days were:

0, 3, 1, 2, 1, 0, 4, 0, 1, 1, 2, 3,
1, 0, 0, 2, 4, 6, 4, 2, 1, 0, 1, 1

Find:

(i) the mean number of absentees,
(ii) the modal number of absentees,
(iii) the sample variance of the number of absentees.

2 The numbers of eggs laid each day by 8 hens over a period of 21 days were:

6, 7, 8, 6, 5, 8, 6, 8, 6, 5, 6,
4, 7, 6, 8, 7, 5, 7, 6, 7, 5

Find:

(i) the modal number of eggs laid per day,
(ii) the mean number of eggs laid per day,
(iii) the sample standard deviation of the number of eggs laid per day.

3 The shoe sizes of the members of a football team are:

10, 10, 8, 11, 10, 9, 9, 10, 11, 9, 10

Determine the variance of this population.

4 A choirmaster keeps a record of the numbers turning up for choir practice on a sample of 18 randomly chosen days.

The numbers are:

25, 28, 32, 31, 31, 34, 28, 31, 29,
28, 32, 32, 30, 29, 29, 31, 28, 28

Using the coding $y = x - 30$, where x is the observed number, determine the value of s_x^2. Confirm your answer by direct calculation without using coding.

5 A biased six-sided die is tossed 60 times giving the following results:

Side of die	1	2	3	4	5	6
Frequency	6	15	2	4	16	17

Without using a calculator (except for the final division), calculate the sample mean and the unbiased estimate of the population variance, showing your working clearly.

6 A driver keeps records of his average mileage per gallon, recording his findings to the nearest integer. His first 25 results are summarised below.

mpg	34	35	36	37	38	39
Frequency	2	4	10	6	2	1

Using the transformation $x = m - 34$, where m is the mpg, and without using a calculator (except for the final division), calculate the sample mean and the unbiased estimate of the population variance, showing your working clearly.

7 A random sample of values of x is:

20, 30, 35, 25, 20, 30, 35, 25, 20, 30, 35, 40,
30, 35, 35, 25, 20, 40, 20, 25, 25, 30, 20, 20

Using the coding $y = \dfrac{x - 35}{5}$, determine the unbiased estimate of the variance of the population.

8 The prices of a set of books, in £, are as follows:

12.95, 12.95, 12.95, 9.95, 16.95,
16.95, 16.95, 16.95, 14.95, 14.95

Use a suitable coding to determine the sample mean and the sample variance of these prices.

9 The gap, x mm, in a sample of spark plugs was measured with the following results:

0.81, 0.83, 0.81, 0.82, 0.80, 0.81, 0.81
0.83, 0.84, 0.81, 0.82, 0.84, 0.80

Use the coding $y = 100x - 80$ to find the unbiased estimate of the population variance of spark plug gaps.

10 A garage notes the mileages of cars brought in for a 15 000-mile service. The data is summarised in the following table.

Mileage ('000 miles)	14–	15–	16–	17–
No. of cars	8	15	13	9

Taking the groups as having mid-points 14 500, ..., 17 500, and using the coding $y = \dfrac{x - 14\,500}{1000}$, where x is the mileage, find the unbiased estimate of the population variance for these grouped data.

11 Each day, x, the number of diners in a restaurant was recorded and the following grouped frequency table was obtained.

x	16–20	21–25	26–30	31–35	36–40
No. of days	67	74	38	39	42

Using the coding $y = \dfrac{x - 18}{5}$, find the value of s_x.
Confirm your answer by direct calculation without using coding.

12 Records are kept for 18 days of the midday barometric pressure, in millibars.

1022, 1016, 1032, 1008, 998, 985,
993, 1004, 1009, 1011, 1015, 1020,
1007, 1001, 995, 993, 975, 972

Using a suitable coding, find the value of s.

13 The total scores in a series of basketball matches were:

215, 224, 182, 200, 229, 219,
209, 217, 195, 162, 210, 213,
204, 208, 197, 192, 187, 213

Using a suitable coding, find the sample mean and the sample variance.

14 The heights of a random sample of 100 Christmas trees were measured with the following results, where h is the height of a tree in metres.

$0.5 < h \leqslant 1.0$	$1.0 < h \leqslant 1.5$	$1.5 < h \leqslant 2.0$
8	23	48

$2.0 < h \leqslant 2.5$	$2.5 < h \leqslant 3.0$
16	5

Find the grouped mean and the value of s for this set of grouped data.

15 A shopkeeper analyses his sales, in order to determine how much each customer spends. The amount spent is denoted by £c. The results are summarised below.

$0 < c \leqslant 10$	$10 < c \leqslant 15$	$15 < c \leqslant 20$
128	223	148

$20 < c \leqslant 30$	$30 < c \leqslant 50$
56	15

Find the grouped mean amount spent per customer.
Find also the value of s for these grouped data.

16 A machine tests the distance w, measured in thousands of km, that car tyres travel before the tyre wear reaches a critical amount. For a random sample of tyres, the results are summarised as follows.

$0 < w \leqslant 25$	$25 < w \leqslant 30$	$30 < w \leqslant 35$
12	23	48

$35 < w \leqslant 45$	$45 < w \leqslant 60$
15	3

Find the grouped mean for these data.
Find also the unbiased estimate of the population variance based on these grouped data.

17 To test their ability to perform tasks accurately, a class of chemistry students are asked to put precisely one kg of flour into a beaker. The class teacher then chooses six students at random and uses an extremely accurate balance (that records weights in milligrams) to determine the actual amounts of flour.

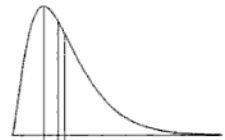
The results are:

1 000 007, 1 000 006, 999 992, 1 000 015, 999 998, 1 000 000

Obtain the sample mean and the value of s, giving your answers in milligrams, correct to two decimal places.

2.20 Symmetric and skewed data

If a population is approximately **symmetric** then in a sample of reasonable size the mean and median will have similar values. Typically their values will also be close to that of the mode of the population (if there is one!). A population that is not symmetric is said to be **skewed**. A distribution with a long 'tail' of high values is said to be **positively skewed**, in which case the mean is usually greater than the mode or the median. If there is a long tail of low values then the mean is likely to be the lowest of the three location measures and the distribution is said to be **negatively skewed**.



A positively skewed distribution

Various measures of skewness exist. One, known as **Pearson's coefficient of skewness**, is given by:

$$\frac{\text{mean} - \text{mode}}{\text{standard deviation}}$$

If the mode is not known, or if there is more than one, or if there is insufficient data for it to be reliably calculated, an alternative is:

$$\frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

An alternative to Pearson's coefficient is the **quartile coefficient of skewness**:

$$\frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

This coefficient takes values between -1 (when $Q_2 = Q_3$) and 1 (when $Q_2 = Q_1$). If the median (Q_2) lies midway between the two quartiles then this coefficient has value 0. It is positive if $(Q_3 - Q_2) > (Q_2 - Q_1)$ and negative if $(Q_3 - Q_2) < (Q_2 - Q_1)$.

Note

- It is feasible for one coefficient to have a negative value and another to have a positive value! Nowadays, professional statisticians rarely use any of these coefficients: they are included here only for syllabus reasons!

Example 20

The distribution of essay marks (out of 20) in a group of 80 students is as follows:

Mark, x	9	10	11	12	13	14	15	16	17	18
Frequency, f	1	1	4	11	17	19	15	8	3	1

Determine the values of the various measures of skewness.

The quartile coefficient is easily calculated. The quartiles are given by $Q_1 = 13$, $Q_2 = 14$, and $Q_3 = 15$ so that $(Q_3 - 2Q_2 + Q_1) = 0$ and the quartile coefficient is therefore 0.

Pearson's coefficient is, perhaps, more sensitive – but also requires more calculation. The sample mean is 13.8 and the sample standard deviation is 1.68, so, using the mean and the mode (14) we get -0.12, while when using the median we get -0.36.

There is some indication of negative skewness, but the three different formulae give (typically!) rather different values.

Exercises 2i

1 Calculate Pearson's coefficient of skewness:
(i) for a set of data having mean 15.0, mode 12.0 and standard deviation 3.1,
(ii) for a set of data having mean 100, mode 112 and standard deviation 20,
(iii) for a set of data having mean -0.9, median -1.1 and variance 0.9.

2 Calculate a measure of skewness for a set of data having 25th, 50th and 75th percentiles equal to, respectively, 14, 31 and 73.

3 The numbers of games of squash played in a given week by a random sample of university students were as follows:

No. of games	0	1	2	3	7
No. of students	42	11	2	1	1

Determine Pearson's coefficient of skewness using the mode.

4 Heights of Sitka spruce trees in a plantation, x m, are summarised below.

$x < 1.5$	$1.5 \leqslant x < 2$	$2 \leqslant x < 2.5$
11	58	74

$2.5 \leqslant x < 3$	$3 \leqslant x < 4$
41	6

Determine the quartile coefficient of skewness.

5 The cost (£x) of the purchases by 30 randomly chosen customers in a supermarket are summarised in the following table.

$x < 20$	$20 \leqslant x < 50$	$50 \leqslant x < 80$
3	19	5

$80 \leqslant x < 100$	$100 \leqslant x < 150$
2	1

Determine the quartile coefficient of skewness.

6 A random sample of 100 adults were asked to state which of the numbers 0, 1, 2, 5, 10, 50, 100 was the best approximation to the number of times that they had been to church in the previous year. Their replies are summarised below:

No. of times	0	1	2	5	10	50	100
No. of replies	72	13	7	2	3	2	1

Determine Pearson's coefficient of skewness using the mode.

7 The numbers of letters that are delivered to a particular house are recorded for 30 consecutive days (excluding Sundays).
On 4 days no letters are delivered, on 12 days 1 letter is delivered, on 6 days 2 letters are delivered, on 5 days 3 letters are delivered. On the remaining days more than 3 letters are delivered. Calculate a coefficient of skewness.

8 After an oil spill, local beaches are checked for oiled birds. To get an idea of the nature of the problem, the beaches are divided into 100 m stretches and the numbers of oiled birds are recorded separately for each stretch. Fifty of the records are summarised below.

0, 1, 5, 2, 19, 47, 21, 8, 7, 4, 0, 1, 1, 0, 0, 0,
0, 0, 0, 1, 3, 15, 11, 4, 3, 7, 2, 2, 0, 0, 0, 0,
0, 1, 0, 0, 1, 0, 0, 1, 4, 6, 6, 0, 1, 2, 2, 0, 0, 0

(i) Determine the mean, median, standard deviation, lower quartile and upper quartile.
(ii) Determine Pearson's coefficient of skewness using the mode.
(iii) Determine Pearson's coefficient of skewness using the median.
(iv) Determine the quartile coefficient of skewness.

9 A supermarket stocks 184 different types of wine. The prices (£x) are summarised in the following table.

$x < 2$	$2 \leqslant x < 3$	$3 \leqslant x < 4$
7	59	58
$4 \leqslant x < 5$	$5 \leqslant x < 8$	$x > 8$
37	15	8

Determine the value of a coefficient of skewness.

2.21 The weighted mean and index numbers

A company gives all its employees a £1000 pay rise, and wants to know the consequent rise *in percentage terms* in its annual wage bill. To find the answer we need to know either all the original wages, or, both the number of employees and the original total wage bill. Suppose that we have the following information:

Staff type	Old wage (£'000)	New wage (£'000)	Number of employees
Junior	15	16	24
Middle	20	21	8
Senior	25	26	8

We begin by calculating the original total wage bill (in £'000) for the 40 employees. This was:

$$(24 \times 15) + (8 \times 20) + (8 \times 25) = 720$$

so the old mean wage was $\frac{720}{40} = 18$ thousand pounds. Note that the average wage was not a simple average of the three possible wages, but was **weighted** by the numbers of employees involved. Without any effort we have calculated a **weighted mean**! If we have a number of values (x) and associated weights (w), then the weighted average is:

$$\frac{\Sigma w_i x_i}{\Sigma w_i}$$

which is simply the usual formula $\frac{\Sigma f_i x_i}{\Sigma f_i}$ in disguise.

The new wage bill is:

$$(24 \times 16) + (8 \times 21) + (8 \times 26) = 760$$

so the overall percentage increase is:

$$\frac{(760 - 720)}{720} \times 100 = 5.56\%$$

For every £100 that the firm used to pay, it will now pay £105.56.

Suppose we wish to contrast this increase with that for another department that uses the same pay scales but has 21 Junior, 3 Middle and 1 Senior member of staff. The old total bill for this department was £400 000, while the new bill is £425 000. This is an increase of 6.25%: for every £100 that the firm used to pay it will now pay £106.25.

Reporting the changes by reference to the convenient yardstick of £100 makes for easy comparisons. The £ sign is redundant so far as the comparisons are concerned so that a summary using **index numbers** would be as shown in the following table:

	First department	Second department
Index for original wage	100	100
Revised index	105.56	106.25

Exercises 2j

1 A box of fruit contains ten apples, thirty bananas and forty lemons. If the average weight of the apples is 140 g, the average weight of the bananas is 130 g and the average weight of the lemons is 115 g, determine the average weight of the fruit in the box.

2 The mark awarded to a Mathematics student is a weighted average of the student's coursework and her exam mark, with the weights being 4 to 1 in favour of the exam mark.
If the student gets 53% in her exam and 62% in her coursework, determine her overall mark.

3 A sports shop stocks four grades of badminton racket. These sell at £15, £25, £30 and £50. Given that the shop has, respectively, 20, 12, 8 and 2 of the four types, determine the mean cost of the badminton rackets in the shop.

4 A particular type of postage stamp occurs in two versions. According to the stamp catalogue one version is worth 20p while the other is worth 60p. If the average value of this type of stamp is 23p and if there are 5 million of the cheaper version in circulation, determine how many there are of the dearer version.

5 (i) If the price of eggs rises by 10% and the old price index for eggs was 120, determine the new price index.
(ii) If the price of bacon rises by 8% and the old index was 116, determine the new index.
(iii) An 'all-day breakfast' consists of eggs costing 40p, bacon costing 35p, and other items costing 62p. The eggs and bacon are now subject to the price rises noted in parts (i) and (ii), while the other breakfast ingredients remain at their old price.
If the old price index for the breakfast was 100, determine the new price index.

6 The Jones family give regular parties. For an average party they buy 6 bottles of wine and 2 kg of a special cheese. The Smiths are trying to keep up with the Joneses, so they buy the same sorts of wine and cheese. However, for their parties, they buy an average of 7 bottles of wine and 1.8 kg of cheese.

In 1990 wine cost an average of £3.10 a bottle and cheese cost an average of £5.25 per kilogram. In 1995 both these prices had risen by £1.

Denoting the cost of a 1990 party by the index number 100, find, for each family, the 1995 index number.

7 The April 1995 version of the *Monthly Digest of Statistics* records the changes in the numbers of various types of road transport using indices.

Year 19–	Cars & taxis	Buses & coaches	Vans	Motor-cycles
87	147	126	131	108
88	157	134	145	97
89	171	140	160	96
90	173	142	161	90
91	173	149	168	87
92	173	142	166	73
93	173	142	164	68

(i) Convert these to indices with 100 corresponding to the 1987 value.
(ii) Plot these results on a time-series graph.
(iii) Which form of transport has shown the greatest percentage increase between 1987 and 1993?

Chapter summary

- **Sigma notation:**

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

$$\Sigma c = nc$$

$$\Sigma(x + y) = \Sigma x + \Sigma y$$

- **Measures of location:**
  - The **mode** is the single value that occurs most frequently (if there is one).
  - The **mean** is the 'average value', denoted by $\bar{x}$.
    - For individual values, $x_1, \ldots, x_n$:

$$\bar{x} = \frac{\Sigma x_i}{n}$$

    - When value x_i occurs with frequency f_i:

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

    - For grouped data, x_i is the mid-point of class i.
  - The **median** (Q_2) is the middle value of ordered values.
    - With $(2k + 1)$ observations the median is the $(k + 1)$th.
    - With $2k$ observations the median is the average of the kth and the $(k + 1)$th.
    - With grouped data (n observations) the median is calculated as the value of the '$\left(\frac{n}{2}\right)$th' observation, using linear interpolation.
  - **Quartiles** (Q_1, Q_2 and Q_3) and **deciles** divide the ordered data into, respectively, quarters and tenths.

- **Measures of spread:**
  - The **range** is the difference between the largest and smallest observations.
  - The **interquartile range** (IQR) is the difference between the upper and lower quartiles.
  - The **variance** has units that are the squares of the units of x.
    - If the observations comprise the entire population, or they represent a sample from a population and we are interested in *the variation within the sample itself*, then the variance is:

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

$$= \frac{1}{n}\left\{\Sigma x_i^2 - \frac{1}{n}(\Sigma x_i)^2\right\}$$

    If $x_1, \ldots, x_n$ constitutes the entire population then σ_n^2 is the **population variance**; otherwise it is the **sample variance**.

(*continued*)

- If the observations constitute a sample from a population, then the quantity referred to as the **unbiased estimate of the population variance**, is:

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

$$= \frac{1}{n-1}\left\{\Sigma x_i^2 - \frac{1}{n}(\Sigma x_i)^2\right\}$$

- If the data are summarised with x_j occurring with frequency f_j, where x_j may denote a class mid-point, then the sample variance is:

$$\sigma_n^2 = \frac{1}{n}\left\{\Sigma f_j x_j^2 - \frac{1}{n}(\Sigma f_j x_j)^2\right\}$$

and the corresponding value of s^2 is given by:

$$s^2 = \frac{1}{n-1}\left\{\Sigma f_j x_j^2 - \frac{1}{n}(\Sigma f_j x_j)^2\right\}$$

where $n = \Sigma f_j$.

- The **standard deviation** (s.d.) is the square root of the variance and has the same units as x.

- **Using coded values**

Using the coding $y = \dfrac{x-a}{b}$, with $b > 0$,

$$\bar{x} = a + b\bar{y}, \qquad \sigma_x^2 = b^2\sigma_y^2, \; s_x^2 = b^2 s_y^2$$

- An **outlier** is an extreme value. There is no set definition. Values that might be considered to be outliers include (i) values that are more than $1.5 \times$ IQR above the upper quartile (or below the lower quartile), (ii) values that are more than 2 standard deviations above or below the sample mean.

- **Boxplots (Box–whisker diagrams)**: indicate the least and greatest values together with the quartiles and the median.

- **Skewness:**
  - **Pearson's coefficient** equals

$$\frac{(\text{mean} - \text{mode})}{\text{s.d.}} \quad \text{or} \quad \frac{3(\text{mean} - \text{median})}{\text{s.d.}}$$

  - The **quartile coefficient** equals

$$\frac{(Q_3 - 2Q_2 + Q_1)}{(Q_3 - Q_1)}$$

- The **weighted mean** of values $x_1, x_2, \ldots$, with weights $w_1, w_2, \ldots$, is given by:

$$\frac{\Sigma w_i x_i}{\Sigma w_i}$$

- The **index number** for some value (e.g. sales figures) is the ratio of this value to a reference value (usually a corresponding value from a previous time point). The ratio is usually multiplied by 100.

Exercises 2k (Miscellaneous)

1 Shirt sizes are given in multiples of $\frac{1}{2}$. The following data refer to the shirt sizes of a random sample of 250 adult males.

Size	14	$14\frac{1}{2}$	15	$15\frac{1}{2}$	16	> 16
Frequency	19	41	43	53	38	56

For these data calculate, where possible, the values of the mean, median, mode, range and variance.

2 For England and Wales, the percentages of households of various sizes, in 1993, were as follows:

1 person	27
2 people	35
3 people	16
4 people	15
5 people	5
6 or more people	2

Source: *Social Trends, 25*, 1995.

Find the modal class.
Represent the data by a suitable diagram.

3 The total scores in a series of basketball matches were:

215, 224, 182, 200, 229, 219,
209, 217, 195, 162, 210, 213,
204, 208, 197, 192, 187, 213

Use a stem-and-leaf diagram to find the median total score.

4 A market gardener sowed 20 sunflower seeds in each of 100 specially prepared seed trays. The number of seeds, n, that germinated in each of the trays was recorded. The values of n and their frequencies are summarised below.

No. germinating	20	19	18	17	16	15	< 15
No. of trays	53	25	12	6	3	1	0

(*a*) Exhibit the distribution of n using a line graph.
(*b*) Calculate the mean and the mode of the distribution of n.
(*c*) Calculate the overall proportion of seeds that germinated.

5 The numbers of households, in England, receiving local authority home help or home care services were tabulated, in thousands, against the age of the oldest client as follows:

Under 18	3.5
18–64	47.0
65–74	83.8
75–84	207.9
85 and over	143.6

Source: *Social Trends, 25*, 1995

Find the modal class.
Represent the data by a suitable diagram.

6 The midnight temperature is recorded, in °C. The figures for 25 Dec to 6 Jan are:

−3, −2, −5, 1, 3, 2, 2,
0, −4, −7, −8, −4, 2

Find:
(i) the mean temperature,
(ii) the median temperature,
(iii) the variance of the temperatures,
(iv) the standard deviation of the temperatures.

7 A region is divided into a lattice of 100 one-metre square quadrats. The number of different plant species is determined for each quadrat, giving the results summarised below.

No. of species	4	5	6	7	8	9
No. of quadrats	1	2	5	9	8	15

No. of species	10	11	12	13	14	15
No. of quadrats	12	8	10	15	10	5

(i) For this set of data determine the mean, median, standard deviation, lower quartile and upper quartile.
(ii) Explain why Pearson's coefficient of skewness using the mode cannot be calculated. Calculate it using the median.
(iii) Determine the quartile coefficient of skewness.

8 The cumulative distribution of the ages (in years) of the employees of a company is given in the following table.

Age	< 15	< 20	< 30	< 40
Cum. Freq.	0	17	39	69

Age	< 50	< 60	< 65	< 100
Cum. Freq.	87	92	98	98

Find:

(i) the median age and the upper and lower quartiles,

(ii) the grouped mean and standard deviation for this population.

9 A grouped frequency distribution of the ages of 358 employees in a factory is shown in Table 1.

Age last birthday	16–20	21–25	26–30
Number of employees	36	56	58

Age last birthday	31–35	36–40	41–45
Number of employees	52	46	38

Age last birthday	46–50	51–60	61–
Number of employees	36	36	0

Table 1

Estimate, to the nearest month, the mean and the standard deviation of the ages of these employees.

Graphically, or otherwise, estimate

(*a*) the median and the interquartile range of the ages, each to the nearest month,

(*b*) the percentage, to one decimal place, of the employees who are over 27 years old and under 55 years old. [ULSEB]

10 (i) At a university, a random sample of 100 students was taken and each student recorded his/her intake of milk (in ml) during a given day. The results are summarized in Table 1.

Milk intake	<25	25–	50–	100–	150–
No. of students	1	3	20	48	11

Milk intake	200–	300–	500–	700–	800–
No. of students	11	4	1	1	0

Table 1

(*a*) Draw a histogram to illustrate these data.

(*b*) Estimate the mean milk intake, explaining the limitations of your calculation.

(*c*) Draw a cumulative frequency curve to fit these data. From your curve, estimate, to the nearest 5 ml, the median intake of milk on that day for all students at this university.

(*d*) State, with reasons, whether you consider the mean or the median to be the more appropriate measure of the milk intake of students at the university on that day.

(ii) A measuring rule was used to measure the length of a rod of stated length 1 m. On 8 successive occasions the following results, in millimetres, were obtained.

999 1000 999 1002
1001 1000 1002 1001

Calculate unbiased estimates of the mean and, to 2 significant figures, the variance of the errors occurring when this rule is used for measuring a 1 m length. [ULSEB]

11 The table given below shows a grouped frequency distribution of the recorded heights, measured to the nearest centimetre, of 50 girls.

Height (cm)	102–105	106–107	108–109
No. of girls	14	16	10

Height (cm)	110–111	112–115
No. of girls	8	2

Find estimates of

(i) the median of the heights,

(ii) the upper quartile of the heights,

(iii) the proportion of the girls whose heights exceed 108.8 cm. [WJEC]

12 Summarised below are the values of the orders (to the nearest £) taken by a sales representative for a wholesale firm during a particular year.

Value of order (£)	Number of orders
less than 10	3
10–19	9
20–29	15
30–39	27
40–49	29
50–59	34
60–69	19
70–99	10
100 or more	4

(*continued*)

(*a*) Using interpolation, estimate the median and the semi-interquartile range for these data.

(*b*) Explain why the median and the semi-interquartile range might be more appropriate summary measures for these data than the mean and standard deviation. [ULEAC]

13 The table below shows the age (at last birthday) at which women married in 1986 in England and Wales.

Age (in yrs)	16–20	21–24	25–29	30–34
Women (in tens of thousands)	6	12	8	3

Age (in yrs)	35–44	45–54	55–99
Women (in tens of thousands)	3	1	1

Draw a histogram and a cumulative frequency diagram to illustrate these data.
Hence estimate

(i) the number of women who were aged 40 or over when they married,

(ii) the median age of marriage for women. [O&C]

14 Measurements of the time intervals between successive arrivals of telephone calls at an office exchange were taken.
The first 100 time intervals were recorded and the following grouped frequency distribution was obtained.

Time interval (x mins)	Frequency
$0 < x \leqslant 0.5$	39
$0.5 < x \leqslant 1.0$	23
$1.0 < x \leqslant 2.0$	23
$2.0 < x \leqslant 3.0$	9
$3.0 < x \leqslant 6.0$	6

(i) Draw a histogram to illustrate this distribution.

(ii) Calculate, showing your working, estimates for the mean and the standard deviation of the distribution.

(iii) Explain briefly which aspects of the data are measured by the mean and the standard deviation. [JMB]

15 On September 1st 1992 the grouped frequency distribution of the ages (in completed years) of 1000 pupils aged under 16 in a comprehensive school was as given in the following table.

Age (in completed years)	11	12	13	14	15
Frequency	165	184	216	231	204

(i) Calculate, to three significant figures, estimates for the mean and standard deviation of the ages of these pupils on September 1st 1992.

(ii) Draw a cumulative frequency polygon and estimate, to three significant figures, the median age of the pupils on September 1st 1992.

(iii) Given in addition that there were 222 pupils aged 16 or over, estimate, to three significant figures, the median age of all the pupils in the school on September 1st 1992. [NEAB]

16 The weekly consumption of cheese in ounces has been estimated for 50 participants in a nutrition study. The figures are given below.

3.89	3.80	4.01	3.84	3.91
4.16	3.98	3.87	3.97	4.04
3.96	4.12	4.05	4.03	4.02
4.07	3.90	4.03	3.94	3.91
3.97	3.91	3.98	4.05	4.03
4.16	4.09	4.13	4.07	4.00
4.06	3.97	4.07	3.90	3.91
4.02	4.20	4.11	3.99	4.02
4.01	4.01	4.05	4.18	3.99
4.21	3.77	3.96	3.84	3.83

(*a*) Construct a stem and leaf diagram of these data.

(*b*) Find the quartiles.

(*c*) Represent the data by a box and whisker plot.

(*d*) Using classes of common width and taking the first class to be 3.75–3.79, form a grouped frequency distribution from the data and represent this grouped distribution by a suitable histogram.

(*e*) Give a brief summary of the main features of the distribution of consumption of cheese by the 50 participants. [O&C]

17 Auditem Ltd., an accounting firm, recorded the time, x minutes, to the nearest minute, taken to audit each account. The values of x below are those recorded for a random sample of accounts they have audited.

37	33	24	36	31	31	24	51
31	47	40	40	55	42	30	34
41	36	42	46	34	38	33	42
56	37	39	36	31	30	45	50
43	41	46	41	30	51	36	21
32	34	62	43	46	34	34	56
32	62	30					

(*a*) For these data,
(i) construct a stem and leaf diagram,
(ii) find the median and quartiles,
(iii) draw a box plot.

(*b*) Write down which of the mode, the median and the mean you would prefer to use as a representative value for these data. Justify your choice. [ULEAC]

18 Give **one** advantage and **one** disadvantage of grouping data in a frequency table.

The table shows the trunk diameters, in centimetres, of a random sample of 200 larch trees.

Diameter (cm)	15–	20–	25–	30–	35–	40–50
Frequency	22	42	70	38	16	12

Plot a cumulative frequency curve of these data.

By use of this curve, or otherwise, estimate the median and the interquartile range of the trunk diameters of larch trees.

A random sample of 200 spruce trees yields the following information concerning their trunk diameters, in centimetres.

Minimum	Lower quartile	Median	Upper quartile	Maximum
13	27	32	35	42

Use this data summary to draw a second cumulative frequency curve on your graph.

Comment on any similarities or differences between the trunk diameters of larch and spruce trees. [AEB 93]

19 The table shows the distribution of ages of school pupils in the United Kingdom in 1984.

Age in completed years	*Number of pupils* (1000)
2 to 4	887
5 to 10	4140
11	825
12 to 14	2631
15 to 16	1183
17 to 18	210

(*a*) What is the age range represented by the entry 11 in the table? Explain what is meant by the number 2631 in the table. How many pupils are represented in this table?

(*b*) Calculate an estimate of the mean age of pupils.

(*c*) For each of the classes calculate its frequency density and on graph paper draw a histogram of the data. Comment briefly on the distribution of ages of pupils. [UODLE]

20 (a) Data are often presented in graphical form rather than in their raw state.
Give
(i) **one** reason for using graphical presentation,
(ii) **one** disadvantage of graphical presentation.

Explain briefly the difference **in use** between a *bar diagram* and a *histogram*.

(b) Electric fuses, nominally rated at 30 A, are tested by passing a gradually increasing current through them and recording the current, x amperes, at which they blow. The results of this test on a sample of 125 such fuses are shown in the following table.

Current (xA)	Number of fuses
$25 \leqslant x < 28$	6
$28 \leqslant x < 29$	12
$29 \leqslant x < 30$	27
$30 \leqslant x < 31$	30
$31 \leqslant x < 32$	18
$32 \leqslant x < 33$	14
$33 \leqslant x < 34$	9
$34 \leqslant x < 35$	4
$35 \leqslant x < 40$	5

(*continued*)

Draw a histogram to represent these data.

For this sample calculate

(i) the median current,

(ii) the mean current,

(iii) the standard deviation of current.

A measure of the *skewness* (or asymmetry) of a distribution is given by

$$\frac{3(\text{mean}-\text{median})}{\text{standard deviation}}.$$

Calculate the value of this measure of skewness for the above data. Explain briefly how this skewness is apparent in the shape of your diagram. [JMB]

21 In an attempt to devise an aptitude test for applicants seeking work on a factory's assembly line, it was proposed to use a simple construction puzzle. As an initial step in the evaluation of this proposal, the times taken to complete the puzzle by a random sample of 95 assembly line employees were observed with the following results.

Time to complete puzzle (seconds)	Number of employees
10–	5
20–	11
30–	16
40–	19
45–	14
50–	12
60–	9
70–	6
80–100	3

Draw a cumulative frequency diagram to represent these data. Hence, or otherwise, estimate the median and the interquartile range.

Calculate estimates of the mean and the standard deviation of this sample.

It is decided to grade the applicants on the basis of their times taken, as good, average or poor.

Method *A* states that the percentages of applicants in these grades are to be approximately 15%, 70% and 15% respectively. Estimate the grade limits.

Method *B* grades applicants as

good, if the time taken is less than (mean−standard deviation),

poor, if the time taken is more than (mean+standard deviation),

average, otherwise.

Compare methods *A* and *B* with respect to the percentages in each grade, and comment. [JMB]

22 (a) Give an example of data for which the most appropriate measure of location might reasonably be

(i) the mode,

(ii) the median,

(iii) the mean.

(b) As part of a work study investigation for the Royal Mail, a daily record was made for each of six days of the number of letters, x, delivered to each of the 175 private houses on a particular postal route. The table below summarises the results for the 1050 possible deliveries.

Number of letters delivered daily	Percentage of deliveries
0	13.2
1	26.7
2	18.9
3	15.8
4	10.5
5	5.2
6	2.6
7	1.1
510	6.0

Construct a suitable pictorial representation of these data.

Calculate the median and the interquartile range for the number of letters delivered daily.

For these data give **two** reasons why the interquartile range is a more appropriate measure of dispersion than the standard deviation. [NEAB(P)]

23 The following table shows a grouped frequency distribution of the gross annual earnings of 110 employees at a certain factory in 1988.

(*continued*)

Gross Earnings	Number of employees
up to £5000	14
above £5000 and up to £6000	25
above £6000 and up to £7000	30
above £7000 and up to £10 000	25
above £10 000 and up to £15 000	13
above £15 000	3

(*a*) Estimate graphically, or by calculation,
(i) the median and the semi-interquartile range of the gross earnings of these 110 employees, giving **each** answer correct to the nearest £,
(ii) the percentage of the employees whose gross earnings exceeded £9000, giving your answer to the nearest whole number.

(*b*) Explain why it is not possible, from the above data, to obtain reliable estimates of the mean and the standard deviation of the gross earnings.

(*c*) More detailed information on the gross earnings of the 110 employees in 1988 showed that the mean was £7470 and the standard deviation was £3550. During the year, the employees' union were negotiating for the mean gross earnings to be increased in 1989 to £8000. Find the constant percentage increase across the board that would meet this claim; give your answer correct to the nearest whole number.

If this percentage was granted find the value of the standard deviation of the employees' gross earnings in 1989.

Suppose, instead, that the employers decided to give each employee an extra £500 in 1989. In this case, find the mean and the standard deviation of the 110 employees' gross earnings in 1989. [WJEC]

24 Over a period of four years a bank keeps a weekly record of the number of cheques with errors that are presented for payment. The results for 200 accounting weeks are as follows.

Number of cheques with errors (x)	Number of weeks (f)
0	5
1	22
2	46
3	38
4	31
5	23
6	16
7	11
8	6
9	2

($\Sigma fx = 706$; $\Sigma fx^2 = 3280$)

Construct a suitable pictorial representation of these data.

State the modal value and calculate the median, mean and standard deviation of the number of cheques with errors in a week.

Some textbooks measure the *skewness* (or asymmetry) of a distribution by

$$\frac{3(\text{mean} - \text{median})}{\text{standard deviation}},$$

and others measure it by

$$\frac{(\text{mean} - \text{mode})}{\text{standard deviation}}.$$

Calculate and compare the values of these two measures of skewness for the above data.

State how this skewness is reflected in the shape of your graph. [AEB 90]

25 A teacher is introducing the concept of weighted averages and index numbers. She uses as her example the cost of a typical breakfast for one person. Her typical breakfast consists of egg, bacon, bread, butter and tea. She has the following information available.

Item	*Quantity*	*Cost* 1986	*Cost* 1990
Bacon	1 lb	£0.88	£1.60
Butter	$\frac{1}{2}$ lb	£0.42	£1.00
Egg	1 dozen	£0.36	£1.20
Bread	1 loaf	£0.35	£0.65
Tea	$\frac{1}{4}$ lb	£0.44	£1.08

(continued)

(*a*) One student suggested that a reasonable estimate for the cost of a typical breakfast in 1986 would be obtained by adding together all the costs in that column and dividing the total by 5. His teacher was not impressed. Put forward **two** criticisms of this proposed method.

(*b*) The teacher went on to say that the sort of quantities consumed in this typical breakfast were:

$\frac{1}{10}$lb of bacon, $\frac{1}{30}$lb of butter, 1 egg, 2 slices of bread ($=\frac{1}{10}$ of a loaf), $\frac{1}{100}$lb of tea.

Calculate a realistic cost for such a meal in 1986.

(*c*) By 1990 the costs of the items had risen to that shown in the 1990 column. Assuming the same consumption as in 1986, calculate an index number for the cost of a typical breakfast in 1990 using 1986 as a base. [UODLE]

26 The table below shows part of the calculation of a simple unweighted index of wage costs for a firm that employs semi-skilled workers, skilled workers, clerical staff and supervisory staff. Copy the table, filling in the missing values indicated by dots.

	DATA Weekly wage rates (£)		
	1980	1984	1988
Semi-skilled	52	72	82
Skilled	79	93	110
Clerical	58	71	76
Supervisory	88	126	160

	CALCULATION Wage rates relative to 1980 (1980 = 100)		
	1980	1984	1988
Semi-skilled	100	138.46	157.69
Skilled	100	117.72	•
Clerical	100	•	•
Supervisory	100	•	•
Index (overall average index of wage costs (1980 = 100)	100	130.44	•

What drawback does this index have in indicating the firm's wage costs in these three years?

In 1980, the firm employed 30 semi-skilled workers, 40 skilled workers, 20 clerical staff and 10 supervisory staff. Calculate the firm's total weekly wage bill for these workers in 1980. Calculate also these totals for 1984 and 1988, assuming the numbers of workers remained the same. Use these totals to construct an index, based on 1980 = 100, for 1984 and 1988.

Do you consider this new index to be an improvement on the original unweighted index? Why?

What further information would you wish to have in judging whether the new index is satisfactory? Supposing this information were available, describe briefly how you would use it in index construction. Discuss whether there would still be any drawbacks in your procedure. [O&C]

27 The tables below show meteorological data from the same site for the years 1994 and 1995. For each month, the tables show the mean maximum temperature (i.e. the mean of the daily maximum temperatures during the month), the mean minimum temperature, the highest and lowest temperatures recorded and the total rainfall recorded.

	Mean Max °C	Mean Min °C	Highest Temp °C	Lowest Temp °C	Rainfall mm
Jan	7.6	2.1	12.0	−1.8	53.6
Feb	5.7	−0.1	11.8	−4.8	47.2
Mar	10.8	4.0	16.0	0.0	35.4
Apr	11.5	4.0	21.4	−0.9	53.2
May	14.3	6.5	20.5	−0.3	58.2
Jun	19.4	9.5	26.3	4.2	20.6
Jul	24.2	12.5	30.4	7.3	7.2
Aug	21.3	11.5	29.8	4.6	33.0
Sep	16.1	9.5	20.5	5.9	54.4
Oct	13.0	5.7	17.6	−0.3	60.8
Nov	11.3	6.3	16.8	−1.6	25.8
Dec	8.4	1.8	13.2	−7.9	52.2
Year	13.6	6.1	30.4	−7.9	501.6

Summary of weather for 1994

(continued)

	Mean Max °C	Mean Min °C	Highest Temp °C	Lowest Temp °C	Rainfall mm
Jan	8.8	4.4	13.5	−7.1	106.3
Feb	9.7	3.8	12.4	−1.9	59.5
Mar	10.3	5.3	16.1	−6.8	56.0
Apr	14.6	5.7	19.7	−4.3	14.2
May	17.1	8.1	26.0	−1.2	32.8
Jun	19.0	9.5	30.8	5.2	11.8
Jul	20.5	11.1	31.2	6.7	24.6
Aug	22.8	11.1	32.8	5.0	8.4
Sep	18.1	9.1	22.0	0.8	121.3
Oct	14.1	7.1	24.0	−0.9	24.5
Nov	11.1	5.1	14.0	−4.8	45.2
Dec	9.3	4.1	12.1	−10.6	89.0
Year	14.9	5.4	q	−10.6	r

Summary of weather for 1995

(i) Describe how the data in the two shaded columns for 1995 may be illustrated in one diagram.
(ii) Determine the values to be inserted in the 1995 table at the places marked **q** and **r**.
(iii) Can the following deductions be made from the tables? Justify your answers.

(a) At least 3 days in 1995 were hotter than any in 1994.
(b) There were more days of frost (i.e. temperatures below 0°C) in 1995 than in 1994. [UCLES]

28 The following table shows data taken from the *Annual Abstract of Statistics, 1993*. The first row shows the numbers, in thousands, of colour television licences in force in the United Kingdom on the 31st March in each of ten successive years. The second row shows the numbers, in thousands, of all licences in force (i.e. those for colour and those for black and white).

Year	1983	1984	1985	1986	1987
Colour licences ('000)	14 699	15 370	15 819	16 025	16 539
All licences ('000)	18 494	18 632	18 716	18 705	18 953

Year	1988	1989	1990	1991	1992
Colour licences ('000)	17 134	17 469	17 964	18 111	18 426
All licences ('000)	19 354	19 396	19 645	19 546	19 631

(i) State briefly what general trends are shown by the figures in the table.
(ii) Describe how the information in the table could be illustrated in one diagram.
(iii) Comment fully on the accuracy of all numerical aspects of the following newspaper report.

"The number of black and white licences decreased by over 68% during the 10-year period. This is a mean annual reduction of 6.8% and by 1992 there were only about 1200 black and white licences in force."

[UCLES]

3 Data collection

It is a capital mistake to theorize before one has data

The Memoirs of Sherlock Holmes, Sir Arthur Conan Doyle

3.1 Data collection by observation

This has been the standard method of data collection for millennia. The famous theories, such as Newton's Theory of Gravitation and Einstein's Theory of Relativity, all have their roots in numerical data collected by careful observation. On a more mundane level, decisions concerning local traffic flow (e.g. 'Would it help to replace the crossroads by a roundabout?') are based on observations of flow made by video cameras or teams of observers.

The collection of data of a scientific nature (e.g. physical, chemical, biological data) relies almost exclusively on observation. However, in the last two centuries there has been increasing interest in the social sciences (e.g. sociology, politics, economics) for which other methods of data collection are relevant.

3.2 The purpose of sampling

Some might say that without sampling there would be no Statistics! This is because:

by careful study of a relatively small amount of data
(the *sample*)
we draw conclusions about a very much larger set of data
(the *population*),
without actually studying the whole of the larger set.

For the sample to be useful:

- *it must not be biased*
  If we are interested in the distribution of the shoe sizes of army officers, then our sample should not be restricted to tall officers.
- *the sample must be taken from the correct population*
  If we are interested in the characteristics of army officers we should not be studying submariners!

3.3 Methods for sampling a population

The simple random sample

Most sampling methods endeavour to give every member of the population the same probability of being included in the sample. If each member of the sample is selected by the equivalent of drawing lots, then the sample selected is described as being a **simple random sample**.

One procedure for drawing lots is the following:

1. Make a list of all N members of the population.
2. Assign each member of the population a different number.
3. For each member of the population place a correspondingly numbered ball in a bag.

4 Draw n balls from the bag, without replacement. The balls should be chosen at random.
5 The numbers on the balls identify the chosen members of the population.

This is like the procedure used in deciding the draw for the Cup competition in football. The drawing of the balls from the bag is sometimes televised.

An automated version would use the computer to simulate the drawing of the balls from the bag. The principles of simulation are discussed in the following sections of this chapter.

The principal difficulty with the above procedure is the first step: the creation of a list of all N members of the population. This list is known as the **sampling frame**. In many cases there will be no such central list. For example, suppose it was desired to test the effect of a new cattle feed on a random sample of British cows. Each individual farm may have a list of its own cows (Daisy, Buttercup, ...), but the Government keeps no central list.

For the country as a whole there is not even a 100% accurate list of people (because of births, deaths, immigration and emigration).

Because of the straightforward nature of the simple random sample, most analyses (and almost all exam questions) assume that this kind of sample has been used to obtain the data. The necessary adjustments that may be required when dealing with other methods of sampling are well beyond the scope of this book. However, the nature of these other methods of sampling needs to be discussed.

A variant of the simple random sampling procedure is **sampling with replacement**, which is also known as **unrestricted sampling**. The procedure is identical except that, at step 4, having drawn the ball from the bag (at random) and noted its number, the ball is replaced in the bag. This variant is rarely used but might be appropriate if failure to replace the 'ball' had unwelcome consequences. For example, fish sampled from a river might die if not replaced.

Cluster sampling

Even if there were a 100% accurate list of the population, simple random sampling of the entire British population would almost certainly not be performed because of the expense. It is easy to imagine the groans emitted by the pollsters on drawing a ball from the bag corresponding to an inhabitant of Land's End, or the Shetland Isles. The intrepid interviewer would be a much travelled individual!

To avoid this problem, populations that are geographically scattered are usually divided into conveniently sized regions. A possible procedure is then:

1 Choose a region at random.
2 Choose individuals at random from that region.

The consequences of this procedure are that instead of a random scatter of selected individuals there are scattered **clusters** of individuals. The selection probabilities for the various regions are not equal, but are adjusted to be in proportion to the number of individuals that the regions contain. If the ith region contains N_i individuals, then the chance that it is selected is chosen to be $\frac{N_i}{N}$, where $N = \Sigma N_i$.

The size of the chosen region is usually sufficiently small that a single interviewer can perform all the interviews in that region without incurring huge travel costs. In practice, because of the sparse population and the difficulties of travel in the highlands and islands of Scotland, studies of the British population are usually confined to the region south of the Caledonian Canal.

Stratified sampling

Most populations contain identifiable **strata**, which are distinctive non-overlapping subsets of the population. For example, for human populations, useful strata might be 'males' and 'females', or 'receiving education', 'working' and 'retired', or combinations such as 'retired female'. From census data we might know the proportions of the population falling into these different categories. With stratified sampling, we ensure that these proportions are reproduced by the sample. Suppose, for example, that the age distribution of the adult population in a particular district is as given in the table below.

Aged under 40	Aged between 40 and 60	Aged over 60
38%	40%	22%

A simple random sample of 200 adults would be unlikely to reproduce these figures *exactly*. If we were very unfortunate, over half the individuals in the sample might be aged under 40. If the sample were concerned with people's taste in music, then, by chance, the simple random sample might provide a misleading view of the population.

A **stratified sample** is made up of separate simple random samples for each of the strata. In the present case, we would choose a simple random sample of 76 adults aged under 40, a separate simple random sample of 80 adults aged between 40 and 60, and a separate simple random sample of 44 adults aged over 60.

Stratified samples exactly reproduce the characteristics of the strata and this almost always increases the accuracy of subsequent estimates of population parameters. Their slight disadvantage is that they are a little more difficult to organise.

Systematic sampling

Both cluster sampling and stratified sampling subdivide the population into components. In both cases the final stage consists of selecting a random sample from a portion of the population. One possible method of doing the final selection is by simple random sampling. An alternative is to use **systematic sampling**, which is described below for the case of a sample of n individuals to be drawn from a listed population of N individuals.

1 Choose one individual at random.
2 Choose every kth individual thereafter, returning to the beginning of the list when the end is reached. The value of k is not crucial, but should be chosen beforehand. A popular choice is a convenient value close to $\frac{N}{n}$.

The use of this wide spacing guards against the list consisting of clusters of similar individuals.

For example, suppose we wish to choose six individuals from a list of 250. A convenient value for k might be 40. Suppose that the first individual selected is number 138. The remainder would be numbers 178, 218, 8, 48 and 88.

If the list has been ordered by some relevant characteristic (e.g. age, or years of service), then, with $k \approx \frac{N}{n}$, this procedure produces a spread of values for the characteristic – a type of informal stratification.

Quota sampling

This is the method often used for street interviews. The interviewer is given a series of targets. For example, he or she might be instructed to interview equal numbers of men and women, of whom one-quarter should be aged over 60 and one-third should be in low-paid jobs. The instructions would be more detailed than these, with the idea being that each interviewer will select a representative cross-section of the population. It is easy to see that an interviewer might have some difficulty in completing his or her **quota** – as night falls the search for an elderly red-bearded giant might still be on!

The results of quota sampling must always be viewed with a little suspicion, since the interviewees are not chosen at random. However, assuming that the quotas have been determined sensibly, the results of a quota sample will be more reliable than those from an **opportunity sample** in which the interviewer gathers information from anyone who will oblige. The population sampled is therefore the population of obliging people – rather different, perhaps, from the general population!

Self-selection

However bad quota sampling may be, it is wonderful by comparison with self-selection! The latter is exemplified by radio or television 'phone-ins' where listeners or viewers record their 'vote'. The views of the apathetic majority are seriously under-represented (though maybe they don't have any to represent!).

A national survey

To illustrate the methods of sampling discussed in the previous section, we now give a brief outline of the method of selection for the households included in the BHPS (*British Household Panel Study*), which is an annual study conducted by staff at the University of Essex.

The *sampling frame* for the BHPS was the Postcode Address File, a (computer-based) master list of all the 1.5 million postcodes in Britain. Each individual postcode (e.g. CO5 8JU) is a member of a so-called Postcode Sector (e.g. CO5 8). There are around 9000 postcode sectors, each of which identifies a *cluster* of about 2500 households.

A simplified version of the stages involved in the selection of the households is as follows.

1 *Selection of sectors*

Sector selection was accomplished by using *systematic* sampling of a cleverly reordered list. The reordering process was as follows:

(a) The 9000 postcode sectors were subdivided into 18 geographical regions.

(b) Within each region the postcode sectors were arranged in an ordered list. The first sector in the list was the postcode sector having the highest proportion of professional heads of households, with the last in the list being that sector with the lowest proportion. These proportions were determined using data from the 1981 Census.

(c) Each regional group was now split into a 'high' half and a 'low' half, to form two subgroups.

(d) Each subgroup was reordered by using descending order of percentage of pensioners as a criterion and was again split into two.

(e) The reordering process was repeated once more, using another social characteristic, to give a total of 144 subgroups, each separately ordered.

The systematic selection of sectors from the reordered list will result in the selected sectors being spread across the country and across the characteristics used for the reordering. The effect is similar to that of *stratification*.

2 *Selection of households*
Simple! These were a *simple random sample* of about 35 households chosen from each selected sector. In all, about 8000 households were selected.

It can be seen that a large survey is likely to combine several different types of sampling procedure. A school survey would obviously be nothing like as complicated, but it is still true that care will be needed to avoid bias.

3.4 Random numbers

Suppose that we have a box containing ten balls labelled 0–9 and otherwise identical. A ball is drawn at random from the box, its number is noted and the ball is then returned to the box. A sequence of draws of this nature will result in a sequence of digits, occurring in an unpredictable fashion, but nevertheless having the property that, in a sequence of $10N$ digits, the expected number of occurrences for each digit would be N. An example of the start of such a sequence is shown below:

85113	47660	38795	86932	04334
60952	44952	45981	54876	87666
44303	61914	54504	18774	29845
50836	38781	50084	98521	78069
19190	50125	54011	39418	12020

Note that these **random numbers** may be read as individual digits (8, 5, 1, 1, ...) or as pairs of digits (85, 11, 34, 76, ...) or in whatever is a convenient manner. In each case they are in effect observations from a discrete uniform distribution.

The numbers may also be interpreted as observations from a *continuous* uniform distribution by the simple expedient of introducing some decimal points:

.85113 .47660 .38795 .86932 .04334

In effect these are random observations from a uniform distribution with range 0 to 1, truncated to an accuracy of 5 decimal places.

Pseudo-random numbers

Suppose we require a sample of 1000 random digits. A thousand draws of balls from a box would be feasible but very tedious. Instead therefore, we make use of computer-generated **pseudo-random numbers**. These are numbers which are generated by a mathematical formula. They have the following properties:

- someone who did not know where they came from would be unable to deduce that they had not been generated using a ten-sided die, but
- the computer could generate exactly the same sequence time after time if this was required.

In practice the description 'pseudo-' is usually dropped and these numbers are also described as **random numbers**.

Tables of random numbers

Although it is easy to produce pseudo-random numbers using the computer, and many calculators also produce numbers of this type, it is often convenient to be able to refer to a printed list of random numbers. Many books of tables and text-books (including this one!) therefore have such a list (see Appendix, p. 445).

To use such a list, it is sufficient that the starting point, and the direction (up, down, left, right, diagonal, etc.) should be decided upon before the numbers in the table can be seen. This is to guard against biased selection of an 'interesting' number with which to start.

Computer project

Many calculators claim to generate a sequence of random numbers. These are usually pseudo-random numbers that are approximately uniformly distributed in the interval (0,1). Similarly, virtually all computer languages have a readily available function (often called RAN or RAND) to generate uniform numbers in this interval.
Investigate the commands required by your calculator/computer.
Do the numbers really seem unpredictable?

3.5 Methods of data collection by questionnaire (or survey)

The most common method for collecting social science data is by means of a **questionnaire** which consists of a series of questions concerning the facts of someone's life or their opinions on some subjects. The recipient of a questionnaire is usually referred to as the **respondent**.
There are three principal methods of collecting the data using a questionnaire:

1 Face-to-face interview
2 By post
3 By phone

The face-to-face interview

In this case the interviewer and the respondent communicate directly, either in a street interview (in which the interviewer selects passers-by for interview) or in an interview in the respondent's home.

Advantages

- **Complex structure** The structure of the questionnaire (e.g. 'If answer is "Yes" then go to question 23c') can be relatively complicated, since only the interviewer needs to understand it.
- **Consistency** If the interviewer does the writing, then the questionnaire will be completed in a consistent fashion.
- **Help** If the respondent has difficulty understanding a question, then the interviewer is available to give an explanation.
- **Response rate** The **response rate** is defined as the number of interviews completed divided by the number attempted. Assuming that the interviewer is friendly, this is likely to be quite high (say 70%).

Disadvantages

- **Expense** The procedure uses up a lot of time for each interviewer. There may also be costs associated with the travelling between respondents.
- **Bias** Although the questionnaire is completed in a consistent fashion, this consistency may contain bias (e.g. the interviewer consistently misinterprets an answer, or gives misleading guidance).
- **Lack of anonymity** A respondent may refuse to answer questions because of being embarrassed by the presence of the interviewer.

The 'postal' questionnaire

Here we mean any questionnaire that is given out for self-completion and return by (anonymous) respondents. An example would be a questionnaire about school food consisting of questions on two sides of paper, to be returned by 'posting' in a box at the end of a lunchtime.

The principal advantage of this method of gathering information is:

- **Economy** Since no interviewer is required, it is a cheap method of collecting data.

However, set against this advantage is a major disadvantage:

- **Non-response** The response rate (measured as the proportion of questionnaires that are returned) can be very low indeed (e.g. 10%) and is rarely greater than 50%. This low level of response is a problem because the replies received are unlikely to be representative of those of the population as a whole. People who take the trouble to fill in and return a questionnaire are not typical (it is well known that 'apathy reigns O.K.')! If the response rate is very low then the replies may be seriously misleading.

The telephone interview

Telephone interviews are occasionally used by market research organisations as a cheaper alternative to face-to-face interviews in the case of short questionnaires. A major problem with a telephone interview (apart from the would-be respondent putting the phone down!) is that it is difficult to relate the information obtained to the population as a whole, because the people interviewed will not be representative.

In Britain in the 1990s, only about 85% of households possess a phone, while about 25% of domestic subscribers are ex-directory. Consequently only about two-thirds of British households are listed in the telephone directory.

When this problem is compounded with that of non-response it is easy to see that the reliability of telephone surveys is rather doubtful. If you read in a newspaper that a telephone survey has unearthed some interesting new facts about British society, then our advice would be to take this information with a pinch of salt.

3.6 Questionnaire design

To ask someone a series of questions might seem to be a ridiculously simple task, but this is certainly not the case. It is easy accidentally to create unanswerable questions, while small changes to the wording can make a difference to the answer obtained. Even the order of questions needs careful thought.

Some poor questions

1 Do you think that boys or girls have the better dress sense or is it simply the influence of their parents?
Unanswerable! This 'question' is at least two questions and is unlikely to be understood by anyone (including its author!).

2 Does your family watch a lot of television?
Unanswerable! Some family members may be TV addicts, whereas others scarcely ever watch. Also 'a lot' is not a well-defined quantity.

3 Do you think that Statistics is:
(a) a very interesting subject,
(b) an interesting subject,
(c) quite an interesting subject?
A biased set of choices.

4 Are you alive?
Not worth asking! Avoid questions that will be answered the same way by everyone (or almost everyone).

5 I am going to ask you about the Monarchy. Bertrand Russell once said ... [something long and rambling taking several minutes to read]. Do you agree?
Avoid long questions – the respondent will forget what the question is about.

6 You are against the death penalty, aren't you?
This is a leading question – the respondent is being pressurised into saying 'Yes'. The defence counsel would object!

7 What do you think of the verisimilitude of this simulacrum?
Avoid unfamiliar words.

8 Are you aged
(a) over 30,
(b) under 21,
(c) under 18?
When giving a range of alternatives make sure that they are non-overlapping and include all possibilities.

9 When they are not playing at home, Arsenal are not a good side at scoring goals. Do you agree or disagree?
Avoid double (or multiple) negatives – some respondents will misunderstand the question.

10 Please don't be embarrassed by this question: do you pick your nose?
But for the preamble, many respondents would have answered the question without worry. Don't invite respondents not to respond!

11 Where were you on March 7th?
Unless this question is asked soon afterwards, it is unlikely to get a response! Questions about the distant past are likely to require the respondent to guess.

12 Are you a communist?
Since communists are rather out of fashion at present, some supporters of communism are unlikely to own up. Respondents tend to give 'socially acceptable' answers.

Some good questions

The best questions are probably those that have been used in surveys conducted by market research or other organisations that specialise in asking questions. From their experience they will know which questions work well. A large public library may be able to help with this.

- Books on survey methods may contain example questionnaires.
- The 'quality' newspapers may report questions asked in national surveys by an organisation such as Gallup.
- The survey organisations themselves may publish questionnaire details.

Most good questions are *short* and *simple*.
The same applies to questionnaires!

The order of questions

Two general rules are:

- Start with easy questions.
  This encourages the respondent to participate.
- Ask general questions (e.g. 'How satisfied are you with school lunches?') first, and specific questions (e.g. 'What do you dislike most about school lunches?') afterwards.
  This is to avoid the 'satisfaction' question being influenced by the subsequent 'dislike' question.

Some questions occur naturally before others. For example, if one were investigating a respondent's history, it would be natural to begin with questions about childhood before questions about middle-age.

Question order and bias

The order in which questions are asked can influence a respondent's reply. Contrast:

1 Do you intend to be an organ donor?
2 Did you know that dozens of people die each year because there are not enough organ donors?

with:

1 Did you know that dozens of people die each year because there are not enough organ donors?
2 Do you intend to be an organ donor?

Filtered questions

Many questionnaires have what might be described as 'miss-out sections' (flagged by statements such as 'If NO then go to Q24'). Thus a question such as:

How much money did you earn last week?

should not precede:

Were you employed last week?

since, if the answer to the second question is 'No', then the first question should not be asked (it should be **filtered** out).

Open and closed questions

An **open question** is one in which there are no suggested answers:

> What is your opinion of the Prime Minister?

The advantage of this type of question is that the respondent can choose precisely how to answer. The disadvantage is that every respondent may answer in a different way, making it difficult to summarise the data obtained.

A **closed question** is one in which there is a prescribed set of alternative answers:

> How do you think the present Government compares with others that we have had?
> Is it (i) above average, (ii) average, (iii) below average?

With a closed question the respondent may find difficulty because none of the alternatives offered is found to be suitable. However, this problem will not arise if all possibilities are covered (as in this example).

The order of answers for closed questions

We noted earlier that question order can affect the responses obtained. The same is true of the alternative answers provided for closed questions.

- There is a bias towards the left-hand answer in 'postal' questionnaires. *Because the respondent reads from left to right and may get bored before reaching the right-hand answers.*
- There is a bias towards the right-hand answer in face-to-face interviews. *Because this is the last answer read out and is therefore the one that the respondent remembers most easily.*
- If there is a sequence of similar questions the respondent is likely to develop a 'habit' and answer each the same way. *So it is a good idea to vary the questions – this also makes the questionnaire more interesting.*

The pilot study

Before using a questionnaire it is essential to make sure that it 'works'. Are there any ambiguous questions? Are there closed questions that cause trouble because a possibility has been overlooked? Are there any questions that you have forgotten to ask? The **pilot study** uses the entire questionnaire with a small number of people who need not be chosen in any scientific way. The aim is simply to find and overcome any difficulties *before* using the real questionnaire.

3.7 Primary and secondary data

Primary data is data collected by, or for, the researcher conducting the current statistical analysis. The process of collecting such data has been the theme of this chapter. Frequently, however, the researcher needs further information and often this information is freely available in government publications, such as *The Monthly Digest of Statistics* or *The Annual Abstract of Statistics*. Data of this type, originally collected for some other purpose, is described as **secondary data**.

Exercises 3a

1 (a) 'Nine out of ten cats prefer Catto'. Comment.
(b) 'Out of a random sample of ten cats, nine preferred Catto.' Comment.
(c) 'Out of a random sample of ten cats, nine preferred Catto to their regular brand.' Comment.

2 A town council is considering building a swimming pool for its residents. Comment on each of the following possible questions for a questionnaire to be issued to all the local inhabitants.
(a) Do you think people would use a new swimming pool?
(b) Do you think the council should build a swimming pool?
(c) Put the following projects in your order of preference: (i) a new library, (ii) a new swimming pool (iii) a new multi-storey car park, (iv) a new leisure centre.

3 A researcher wishes to find out the favourite television programmes of newly married couples. The researcher wishes to conduct face-to-face interviews and must therefore locate potential interviewees. Briefly state the advantages and disadvantages of each of the following methods.
(a) Advertise in a national newspaper for newly-weds stating the wish to conduct face-to-face interviews about favourite television programmes.
(b) Advertise in a national newspaper for newly-weds stating the wish to conduct face-to-face interviews.
(c) Consult the 'Marriages' column of a national newspaper.
(d) Consult the 'Marriages' section of a local newspaper.
(e) Choose a random sample of local newspapers (from a national list), and then choose random samples from the 'Marriages' section of each paper.

4 In Greyfriars Academy, classes 3, 4 and 5 have 35, 42 and 28 students, respectively. John Smith is a member of class 4. One student is to be chosen from the three classes to represent the school in a general knowledge competition. Three methods of selection are suggested:
(i) Make a list of all the students and choose one at random.
(ii) Choose a class at random and choose a student at random from that class.
(iii) Choose a class at random and discard it. Choose a class at random from the remaining two. Choose a student at random from this class.
Determine the probability that John Smith is chosen under each scheme. Which of the schemes are fair?

5 A researcher wishes to obtain information on the types of cars using a multi-storey car park. His procedure is to choose a random day of the week and, after his arrival at the car park entrance at 10 am, to note the type of every 10th car that enters the car park. He goes home at 6 pm to tabulate his results. Comment on this sampling scheme.

6 A television station wishes to sample the opinions of the electors in a small town. The names of the N electors are listed on an electoral roll. Each of the following procedures aims to provide a sample of n electors.
(i) Use the electoral roll. Generate n random integers in the range $(1, \ldots, N)$ and interview the corresponding electors.
(ii) Use the electoral roll. Suppose that the ratio of N to n is approximately equal to the integer r. Generate an integer in the range $(1, \ldots, r)$. Denoting this as k, interview the electors numbered $k, k+r, k+2r, \ldots$.
(iii) Use the electoral roll. Generate $3n$ random integers in the range $(1, \ldots, N)$. Send each of these a letter (assuming a response rate of 1 in 3).
Comment on the merits of each procedure.

7 The local Cheapsell supermarket is interested in getting the views of its shoppers on the proposed new layout of its shelves. Comment on the merits of each of the following possible sampling procedures.
(i) Hand a leaflet out to every shopper who enters the shop between 1 pm and 2 pm on a randomly chosen Monday. Completed leaflets are to be deposited in a bin in the foyer.
(ii) Choose a day of the week at random and offer a leaflet, for immediate completion, to every shopper who enters the shop.

(iii) Suppose the supermarket is open 11 hours a day, for six days a week. During the first week interview exactly 10 shoppers between 9 am and 10 am on Monday, exactly 10 between 10 am and 11 am on Tuesday, ..., exactly 10 between 2 pm and 3 pm on Saturday. In the next week continue by interviewing 10 people between 3 pm and 4 pm on Monday, and so on, advancing by an hour a day and continuing for a total of five weeks.

8 A television show addresses issues of national interest. A key element of the show is the 'Let the people decide' feature. Viewers are given alternative numbers to phone depending on whether they are in favour of, or against, the proposition of the week. State, with reasons, whether or not this provides a good test of public opinion.

9 A philosopher has a great many books, which he keeps in two rather dusty rooms (one brown and the other grey). In the brown room are two large bookcases with 300 books in each bookcase. In the grey room are three smaller bookcases, two containing 200 books and the third containing just 100 books. One of his books, entitled 'The Meaning of Life', is on the smallest bookcase. The philosopher is on his way upstairs and wants to choose a book to read in bed. Determine the probability that 'The Meaning of Life' is chosen if

(i) he chooses a book by selecting a card at random from his card index (which has one card for each book),

(ii) he chooses a bookcase at random, and then a book at random from that bookcase,

(iii) he chooses a room at random, then a bookcase at random from the bookcases in that room, then a book at random from that bookcase.

Chapter summary

- **Sampling frame**: The list of population members.
- **Simple random sampling**: Selection of individuals directly from the sampling frame with equal probability of selection for each individual.
- **Cluster sampling**: Selection from randomly chosen groups of neighbouring individuals.
- **Stratified sampling**: Division of the sampling frame into non-overlapping subsets called **strata**, with proportionate simple random sampling from each subset.
- **Systematic sampling**: Sampling at regular intervals from an ordered sampling frame.
- **Quota sampling**: Non-random sampling of targeted types of individuals.
- **Pseudo-random numbers**: Numbers, created using a mathematical formula, that appear indistinguishable from genuinely random numbers.

4 Probability

Probable impossibilities are to be preferred to improbable possibilities

Aristotle

4.1 Relative frequency

Suppose we roll a die and are interested in the outcome '6'. To get some idea of how likely the outcome is, we roll the die repeatedly. Here are the first 30 rolls:

2 4 4 1 2 3 2 4 3 1 4 5 6 4 3
2 3 6 2 4 3 4 2 2 5 4 6 5 3 3

After 10 rolls we have had no '6's and might think that getting a '6' is impossible! However, as the number of rolls increases so '6's begin to appear: after 30 rolls, we have had 3 '6's – a **relative frequency** of $\frac{3}{30} = 0.1$.

What will happen as we increase the number of rolls? The answer is that the number of '6's will increase, but the proportion of '6's (the relative frequency) will stabilise. The limiting value of this relative frequency is called the **probability**. So, if all six sides of the die are equally likely (which is the case for a *fair die*), then the limit of the relative frequency will be $\frac{1}{6}$ and we will say that the probability of a '6' is $\frac{1}{6}$.

4.2 Preliminary definitions

- A **statistical experiment** is one in which there are a number of possible outcomes and we have no way of predicting which outcome will actually occur. Sometimes the experiment may have already taken place, but we remain ignorant of the outcome.
- The **sample space**, often denoted by S, is the set of all possible outcomes of the experiment.
  - The use of the word *sample* in the definition of S is an unfortunate historical accident – it does *not* refer to a sample of observations.
- An **event** is any set of possible outcomes of the experiment (thus an event is a subset of S). When rolling a die we might be interested in events such as 'getting an even number' or 'getting a number greater than 3'.
- A **simple event** is an event consisting of a single outcome. When rolling a die the simple events are '1', '2', etc.

Example 1

Many board games require the rolling of an ordinary six-sided die. The possible outcomes are 1, 2, 3, 4, 5, 6. Before the die is rolled we cannot predict the outcome – so this is an example of a statistical experiment.

In our new notation the six values are the six possible simple events and the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. An example of a simple event is 'the outcome is a 6'. Examples of events are 'the outcome is an even number' and 'the outcome is a number less than 4'.

4.3 The probability scale

Assigned to the event E is a number, known as the probability of the event E, which takes a value in the range 0 to 1 (inclusive). The number is denoted by $P(E)$. In addition to satisfying:

$$0 \leqslant P(E) \leqslant 1$$

the value of $P(E)$ is chosen so that:

If E is impossible, then $P(E) = 0$
If E is certain to occur, then $P(E) = 1$

Intermediate values of $P(E)$ have natural interpretations:

$P(E) = 0.5 \longrightarrow E$ is as likely to occur as not to occur
$P(E) = 0.001 \longrightarrow E$ is very unlikely
$P(E) = 0.999 \longrightarrow E$ is highly likely

Example 2

Suppose we toss an ordinary coin. Define the events A and B to be:

A: The coin comes down heads.
B: The coin explodes in a flash of green light.

We can reasonably assume that $P(A) = \frac{1}{2}$ and that $P(B) = 0$.

4.4 Probability with equally likely outcomes

Suppose that the sample space, S, consists of $n(S)$ possible outcomes, and suppose that each is *equally likely*. Suppose that the number of outcomes in the event E is $n(E)$. Then $P(E)$, the probability that the event E occurs, is given by the equation:

$$P(E) = \frac{n(E)}{n(S)} \tag{4.1}$$

This clearly satisfies the requirement that $0 \leqslant P(E) \leqslant 1$.

Example 3

A fair die is tossed. The event A is defined as 'the number obtained is a multiple of 3'.
Determine $P(A)$.

Here, the sample space S consists of the outcomes $\{1, 2, 3, 4, 5, 6\}$, so that $n(S) = 6$. The outcomes corresponding to A are $\{3, 6\}$, so $n(A) = 2$. Thus $P(A) = \frac{n(A)}{n(S)} = \frac{2}{6} = \frac{1}{3}$.

Example 4

Two fair coins are tossed. The event A is defined to be 'exactly one head is tossed'.
Determine $P(A)$.

Consider the coin that is tossed first. This coin is equally likely to give a head (H) or a tail (T). Suppose it gives a head. The second coin is now tossed. This coin is also equally likely to give a head or a tail, so, if the first coin was a head there are two equally likely sequences: HH and HT. On the other hand, if the first coin gave a tail then the equally likely sequences are TH and TT. Since the first coin was equally likely to give a head as a tail, the four outcomes HH, HT, TH, TT, which make up the sample space, S, are equally likely. The event A corresponds to the outcomes HT, TH. Thus $n(A) = 2$, $n(S) = 4$ and so $P(A) = \frac{n(A)}{n(S)} = \frac{2}{4} = \frac{1}{2}$.

Exercises 4a

1 An unbiased die is thrown.
Find the probability that:
(i) the score is even,
(ii) the score is at least two,
(iii) the score is at most two,
(iv) the score is divisible by 3.

2 A box contains 4 red balls, 6 green balls and 5 yellow balls. A ball is drawn at random.
Find the probability that:
(i) the ball is green,
(ii) the ball is red,
(iii) the ball is not yellow.

3 A card is drawn at random from a pack.
Find the probability that:
(i) the card is a Spade,
(ii) the card is an Ace,
(iii) the card is the Ace of Spades,
(iv) the card is a 'court card' (King, Queen or Jack).

4 A computer produces a 4-digit random number in the range 0000 to 9999 inclusive, in such a way that all such numbers are equally likely.
Find the probability that:
(i) the number is at least 1000,
(ii) the number lies between 1000 and 5000 inclusive,
(iii) the number is 4321,
(iv) the number ends in 0,
(v) the number begins and ends with 1.

5 A disc carries the numbers 1 and 2 on its faces. It is thrown with a fair die. The score is the sum of the two numbers that show.
Find the probability that:
(i) the score is at least 4,
(ii) the score is at most 6.

6 A bag contains 30 balls. The balls are numbered 1, 2, 3, ..., 30. A ball is drawn at random.
Find the probability that the number on the ball:
(i) is divisible by 3,
(ii) is not divisible by 3,
(iii) is divisible by 4,
(iv) is a prime (2, 3, 5, ...),
(v) differs from 10 by less than 5,
(vi) differs from 25 by more than 6.

7 Two unbiased dice, one red and one green, are thrown and the separate scores are observed. Represent the result as (r, g), where r and g are the scores on the red and green dice respectively.
Give a reason why there are 36 of these simple events.
Hence find the probability that:
(i) a double six is obtained,
(ii) a double (any score) is obtained,
(iii) the sum of the two scores is 4,
(iv) the sum of the two scores is 5,
(v) the score on the red die is 3 more than the score on the green die,
(vi) both scores are divisible by 3.

8 I have 14 coins in my purse. There are two 1p coins, three 2p coins, four 5p coins and five 10p coins. I choose a coin at random.
Find the probability that:
(i) it is a 2p coin,
(ii) it is worth at least 5p,
(iii) it is worth less than 3p,
(iv) it is worth at least 1p,
(v) it is worth at least 20p.

4.5 The complementary event, E'

An event E either occurs or it does not! We cannot have events 'half-occurring'. Each of the possible equally likely outcomes therefore corresponds to the event occurring or to the event not occurring. If $n(E)$ is the number of outcomes for which E occurs and $n(S)$ is the size of the sample space, then $n(S) - n(E)$ is the number of outcomes corresponding to the event 'E does not occur', which is called the **complementary event**, and is denoted by E'. Thus:

$$P(E') = \frac{n(S) - n(E)}{n(S)} = 1 - \frac{n(E)}{n(S)} = 1 - P(E)$$

This result:

$$P(E') = 1 - P(E) \tag{4.2}$$

or its equivalent:

$$P(E) = 1 - P(E')$$

often enables us to simplify calculations.

Notes

- The complementary event is sometimes denoted by $\bar{E}$, $C(E)$ or E^c.
- $(E')' = E$, since if E' does not occur then E occurs and vice versa.

Example 5

We toss a red die and a blue die. Both dice are fair.
We wish to find $P(A)$, where A is the event 'the total of the numbers shown by the two dice exceeds 3'.

We begin by finding $n(S)$, the number of possible outcomes in the sample space. There are six equally likely outcomes for the red die.
Whichever of these outcomes arises, there will also be six equally likely outcomes for the blue die. In all, therefore, there are thirty-six equally likely outcomes: $n(S) = 36$. We can see this easily on a diagram which can also be used to show the possible totals of the two dice:

		Red die					
		1	2	3	4	5	6
Blue die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

The complementary event, A', is the event that 'the total of the two dice does not exceed 3'. Now whereas there are lots of outcomes for which A occurs, there are very few for which A' occurs and it is easy to count them: the (red die, blue die) possibilities are $\{(1,1), (1,2), (2,1)\}$. Hence $n(A') = 3$. The 33 remaining outcomes in the diagram correspond to the event A.

Now, since all the outcomes are equally likely:

$$P(A') = \frac{n(A')}{n(S)} = \frac{3}{36} = \frac{1}{12}$$

But $P(A) = 1 - P(A')$, so that $P(A) = \frac{11}{12}$.

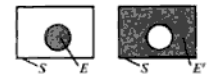
Note

- The fact that the dice were coloured does not affect $P(A)$. It simply makes it easier to describe what is happening. All that is required is some method of distinguishing the dice, and this is *always* possible, even if the dice are described as being identical! We could refer to the dice as being rolled one after the other, or being rolled by different people, or being rolled at the same time from different starting points in the die shaker.

John Venn (1834–1923) was a Cambridge lecturer whose speciality was logic. His major work, *The Logic of Chance*, was published in 1866. It is best known today for the introduction of the diagrams that now bear his name. Venn had a general interest in all branches of Statistics and a letter that he wrote in 1887 to the editor of the influential journal *Nature* stimulated an explosion of interest in the mathematical theory of Statistics.

4.6 Venn diagrams

A Venn diagram is a simple representation of the sample space, that is often helpful in seeing 'what is going on'. Usually the sample space is represented by a rectangle, with individual regions within the rectangle representing events.
It is often helpful to imagine that the actual areas of the various regions in a Venn diagram are in proportion to the corresponding probabilities. However, there is no need to spend a long time drawing these diagrams – their use is simply as a reminder of what is happening.

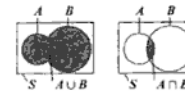


Venn diagrams illustrating the events E and E'

4.7 Unions and intersections of events

Suppose A and B are two events associated with a particular statistical experiment. We now consider the events denoted by $A \cup B$ and $A \cap B$, which are defined as follows:

$A \cup B$	'A **or** B'	At least one of A and B occurs.
$A \cap B$	'A **and** B'	Both A and B occur.



Venn diagrams illustrating the events $A \cup B$ and $A \cap B$

Notes

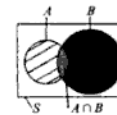
- $A \cup B$ includes the possibility that both A *and* B occur.
- In set notation:
  $A \cup B$ is called the **union** of A and B,
  $A \cap B$ is called the **intersection** of A and B.

The number of outcomes in A is $n(A)$ and the number of outcomes in B is $n(B)$. Also a total of $n(A \cap B)$ outcomes is in both A and B. The outcomes in $A \cup B$ include all those in A and all those in B but no others. However, if we simply add together $n(A)$ and $n(B)$ we will overstate the number in $A \cup B$ because we will have counted those in $A \cap B$ twice.
Hence:

$$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

Dividing throughout by $n(S)$ we get:

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B) \qquad (4.3)$$



Outcomes in $A \cap B$ are counted twice

Notes

- In ordinary English, the phrase '*A or B*' usually means one of *A* and *B*, but *not* both. However, in probability questions '*A* or *B*' does not rule out the possibility of both occurring. The ambiguity disappears if set notation is used.
- From Equation (4.3) (or from the Venn diagram) it can be seen that $P(A \cup B)$ must be at least as great as the greater of $P(A)$ and $P(B)$.
- Similarly, $P(A \cap B)$ cannot be greater than the lesser of $P(A)$ and $P(B)$.

Example 6

Each month a mail-order firm awards a 'Star Prize' to a randomly chosen shopper. The firm uses the following procedure. It first chooses eight shoppers at random. The names of these eight shoppers are put into a hat. A guest celebrity then draws the name of the lucky winner of the 'Star Prize' from the hat and the other seven shoppers are awarded consolation prizes.

One month the first of the eight shoppers was a male living in the south of England. The complete list of those chosen was:

	Shopper number							
	1	2	3	4	5	6	7	8
Male (M) or Female (F)	M	F	F	F	M	F	F	F
North (N) or South (S)	S	S	N	S	N	S	S	N

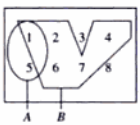
The events *A* and *B* are defined by:

A: The winner of the 'Star Prize' is male.
B: The winner of the 'Star Prize' lives in the south.

(i) Define, in words, the events $A \cap B$ and $A \cup B$.
(ii) Determine the probabilities of these events.

(i) The event $A \cap B$ is the event: 'the winner of the "Star Prize" is a male living in the south'.
The event $A \cup B$ is the event: 'the winner of the "Star Prize" is either a male, or lives in the south (or both)'.

(ii) The situation is illustrated in the Venn diagram, with the eight simple events, which are all equally likely, making up the sample space, *S*, being identified by numbers. Note that not all sets are egg-shaped! It can be seen that only the first of the eight simple events corresponds to $A \cap B$.



The following table provides a comprehensive list of the various events:

Event (E)	Simple events in E	$n(E)$	$P(E)$
Sample space, S	1, 2, 3, 4, 5, 6, 7, 8	8	1
A	1, 5	2	$\frac{2}{8} = \frac{1}{4}$
B	1, 2, 4, 6, 7	5	$\frac{5}{8}$
$A \cap B$	1	1	$\frac{1}{8}$
$A \cup B$	1, 2, 4, 5, 6, 7	6	$\frac{6}{8} = \frac{3}{4}$

As a check note that:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \tfrac{2}{8} + \tfrac{5}{8} - \tfrac{1}{8} = \tfrac{6}{8}$$

The probabilities of the two events are $\frac{1}{8}$ and $\frac{3}{4}$.

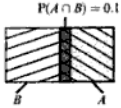
Example 7

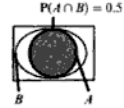
For the sample space S it is known that $P(A) = 0.5$, $P(B) = 0.6$.
Determine the minimum and maximum possible values of $P(A \cap B)$.
Illustrate each case using a Venn diagram.

Substituting into Equation (4.3) we have:

$$P(A \cup B) = 0.5 + 0.6 - P(A \cap B) = 1.1 - P(A \cap B)$$

Since $P(A \cup B)$ cannot exceed 1, the minimum value of $P(A \cap B)$ is 0.1.
When $P(A \cap B) = 0.1$, $P(A \cup B)$ takes its maximum value = 1 and $A \cup B$ is the whole of S. The smaller of $P(A)$ and $P(B)$ is $P(A)$, so the maximum value for $P(A \cap B)$ is 0.5, in which case $A \cap B$ is the whole of A.





Example 8

Interviews with 18 people revealed that 5 of the 8 women and 8 of the 10 men preferred drinking coffee to tea.
Determine the probability that the first person interviewed was either a woman or someone who preferred coffee to tea.

In the absence of any information to the contrary we begin by assuming that each of the 18 people was equally likely to have been the first to be interviewed. If we were to guess who was first interviewed, there would be a probability of $\frac{1}{18}$ that we would guess correctly.

We define the events W: 'the person was a woman' and C: 'the person preferred coffee to tea'. We *assume* that the interviewing was done at random, so that each of the 18 people has probability $\frac{1}{18}$ of being the first to be interviewed. The question as stated is slightly ambiguous (questions often are!) – we assume it requires calculation of $P(W \cup C)$.

Now $P(W) = \frac{8}{18}$, $P(C) = \frac{13}{18}$, and
$P(W \text{ and } C) = P(\text{the person was a woman who preferred coffee to tea}) = P(W \cap C) = \frac{5}{18}$.

Hence:

$$P(W \text{ or } C) = \frac{8}{18} + \frac{13}{18} - \frac{5}{18} = \frac{16}{18} = \frac{8}{9}$$

so the probability that the first person interviewed was either a woman or someone who preferred coffee to tea is $\frac{8}{9}$.

	Prefers coffee	Prefers tea	Total
Women	5	3	8
Men	8	2	10
Total	13	5	18

An easy way of seeing the answer in this case is by totalling the underlined numbers in the table, and expressing the total as a proportion of the overall total (18).

4.8 Mutually exclusive events

Events $A, B, \ldots, M$, are said to be **mutually exclusive** if the occurrence of one of them implies that none of the others can occur. If D and E are two mutually exclusive events then $P(D \cap E) = 0$.

Note

- All simple events are mutually exclusive.

The addition rule

If the events A and B are mutually exclusive, then Equation (4.3) simplifies, since $P(A \cap B) = 0$, to give:

$$P(A \cup B) = P(A) + P(B) \tag{4.4}$$

which is known as the **addition rule**.

Note

- The addition rule *only* applies to mutually exclusive events.

Example 9

An Irish rugby club contains 40 players, of whom 7 are called O'Brien, 6 are called O'Connell, 4 are called O'Hara, 8 are called O'Neill and there are 15 others. The 40 players draw lots to decide who should be captain of the first team. Determine the probability that the captain of the first team is:
(i) called either O'Brien or O'Connell,
(ii) is not called either O'Hara or O'Neill.

The sample space consists of the 40 players, each of whom is equally likely to be selected as captain. Denote the event that 'the captain is an O'Brien' by the symbol B, with C, H and N denoting the other events. The events B, C, H and N are mutually exclusive, since a player cannot have two surnames.

(i) $P(B \text{ or } C) = P(B \cup C) = P(B) + P(C) = \frac{7}{40} + \frac{6}{40} = \frac{13}{40}$

The probability that the captain is called O'Brien or O'Connell is $\frac{13}{40}$.

(ii) $P(\text{Neither } H \text{ nor } N) = 1 - P(H \text{ or } N) = 1 - \{P(H) + P(N)\}$
$= 1 - \{\frac{4}{40} + \frac{8}{40}\} = 1 - \frac{12}{40} = \frac{28}{40} = \frac{7}{10}$

The probability that the captain is not called O'Hara or O'Neill is $\frac{7}{10}$.

The question can also be answered by making a table showing the possibilities:

	O'Brien	O'Connell	O'Hara	O'Neill	Others
Number	7	6	4	8	15
Satisfy part (i)	*	*			
Satisfy part (ii)	*	*			*

From the table we see immediately that there are 13 players satisfying the requirements of (i) and 28 satisfying (ii), so the probabilities are $\frac{13}{40}$ and $\frac{28}{40}$, respectively.

4.9 Exhaustive events

Two events are said to be **exhaustive** if it is certain that at least one of them occurs. For example, when rolling a die it is certain that at least one of the events A: 'the number obtained is either 1, 2, 3 or 5' and B: 'the number obtained is even' will occur. In this example, if a 2 is obtained then both A and B occur. If the events A and B are exhaustive then:

$$P(A \cup B) = 1 \qquad (4.5)$$

Notes

- Any event A and its complement, A', are both exhaustive and mutually exclusive:

  $$P(A \cup A') = 1, \qquad P(A \cap A') = 0$$

- The events $A, B, \ldots, N$ are said to be exhaustive if it is certain that at least one of them occurs:

  $$P(A \text{ or } B \text{ or } \ldots \text{ or } N) = P(A \cup B \cup \ldots \cup N) = 1$$

  Thus the simple events that make up the sample space, S, are mutually exclusive and exhaustive.

Exercises 4b

1 A fair die is thrown. Events A, B, C, D are defined as follows:

A: The score is even.
B: The score is divisible by 3.
C: The score is not more than 2.
D: The score exceeds 3.

Verify that:
$P(A) + P(B) = P(A \cup B) + P(A \cap B)$
Find:
(i) $P(A')$, (ii) $P(B')$, (iii) $P(C')$, (iv) $P(D')$
(v) Identify two pairs of events that are mutually exclusive, and verify the addition rule in each case.
(vi) Identify three events that are exhaustive.
(vii) Find $P(A \cup B \cup C)$.
(viii) Find $P(C \cap D)$.

2 Two fair dice, one red and one green, are thrown and the separate scores are observed. The outcome is denoted by (r, g), where r and g are the scores on the red and green dice respectively.
Represent these outcomes on a 6×6 grid, with r-axis horizontal and g-axis vertical. Events A, B, C are defined as follows:

A: The score on the red die exceeds the score on the green die.
B: The total score is six or more.
C: The score on the red die does not exceed 4.

(i) Identify on your diagram the sets corresponding to A, B, C.
(ii) Verify that:
$P(A) + P(B) = P(A \cup B) + P(A \cap B)$
(iii) Verify that:
$P(A) + P(C) = P(A \cup C) + P(A \cap C)$
(iv) Identify a pair of events that are exhaustive.
(v) Find $P(A')$, $P(B')$, $P(C')$.
(vi) Find $P(A' \cup B)$, $P(A \cap B')$, $P(B \cup C)$, $P(B' \cap C')$, $P(B' \cup C')$.

3 A man tosses two fair dice. One is numbered 1 to 6 in the usual way. The other is numbered 1, 3, 5, 7, 9, 11. The events A to E are defined as follows:

A: Both dice show odd numbers.
B: The number shown by the normal die is the greater.
C: The total of the two numbers shown is greater than 10.
D: The total is less than or equal to 4.
E: The total is odd.

(a) Determine the probability of each event.
(b) State which pairs of events (if any) are exclusive and which (if any) are exhaustive.

4 For the sample space S it is given that:
$P(A) = 0.5$, $P(A \cup B) = 0.6$,
$P(A \cap B) = 0.2$.
Find:
(i) $P(B)$,
(ii) $P(A' \cap B)$,
(iii) $P(A \cap B')$,
(iv) $P(A' \cap B')$.

5 For the sample space S it is given that:
$P(A' \cap B) = \frac{3}{7}$, $P(A \cap B') = \frac{2}{7}$, $P(A' \cap B') = \frac{1}{7}$.
Find:
(i) $P(A \cap B)$,
(ii) $P(A)$,
(iii) $P(B)$.

6 For the sample space S it is given that:
$P(B \cap C) = 0$, $P(A \cap B) = \frac{1}{20}$, $P(A \cap C) = \frac{2}{5}$,
$P(A) = \frac{3}{5}$, $P(B) = \frac{3}{20}$, $P(C) = \frac{13}{20}$.
Sketch a corresponding Venn diagram and indicate $A \cap B$ and $A \cap C$.
Find:
(i) $P(A' \cap B)$,
(ii) $P(A' \cap C)$,
(iii) $P(A \cup B)$,
(iv) $P(B \cup C)$,
(v) $P(A' \cap B' \cap C')$.

7 A card is drawn at random from a normal pack of 52 cards. Events A, B, C, D are defined as follows:

A: The card drawn is either a Queen or a Heart.
B: The card drawn is a black King.
C: The card drawn is either an Ace or a King or a Queen or a Jack.
D: The card drawn is a Spade.

Find:
(i) $P(A)$, (ii) $P(B)$, (iii) $P(C)$, (iv) $P(D)$,
(v) $P(A \cap D)$, (vi) $P(A \cup D)$, (vii) $P(A \cup B)$,
(viii) $P(C \cap D)$, (ix) $P(C \cup D)$,
(x) $P(B \cap D)$, (xi) $P(B \cup D)$.

8 A survey of 1000 people revealed the following voting intentions.

	Women	Men	Total
Con	153	130	283
Lab	220	194	414
LibDem	157	146	303
Total	530	470	1000

A person is chosen at random from the sample. Find the probability that the person chosen:
(i) intends to vote Conservative,
(ii) is a woman intending to vote Labour,
(iii) is either a woman or intends to vote Conservative,
(iv) is neither a man nor intends to vote LibDem,
(v) is a man and intends to vote either LibDem or Labour.

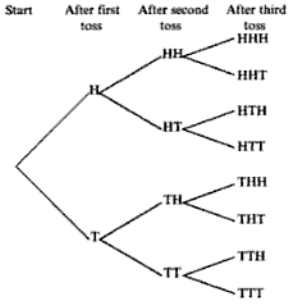
4.10 Probability trees

Probability trees are diagrams that help us to see what is happening!
Consider the following problem. A fair coin is tossed three times. Determine P(exactly two heads are obtained).

Each time we toss the coin the number of distinguishable outcomes increases:

After first toss	Either H or T
After second toss	The sequence of outcomes must be HH, HT, TH or TT
After third toss	Either HHH, HHT, HTH, HTT, THH, THT, THH or TTT

The same possibilities are represented more simply (and we are less likely to miss out one of the possibilities!) in a tree diagram in which the final column lists the entire sample space.

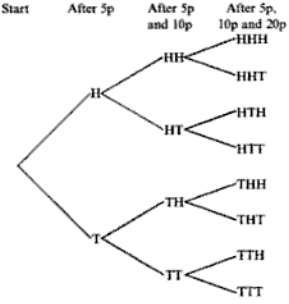


Each of the eight sequences is equally likely to occur. Three sequences include exactly two heads (HHT, HTH, THH) and so the probability of obtaining exactly two heads is $\frac{3}{8}$.

Consider the new problem.

A man tosses a 5p coin, a 10p coin and a 20p coin.
Determine P(exactly two heads are obtained).

Essentially the same tree diagram does the trick:

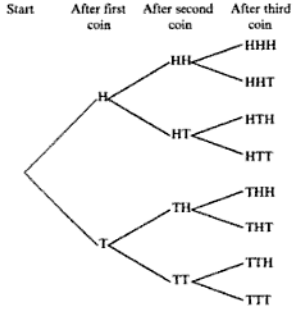


The probability is again equal to $\frac{3}{8}$.

Consider one final problem.

A woman tosses three coins.
Determine P(exactly two heads are obtained).

Once again we use the same tree diagram:



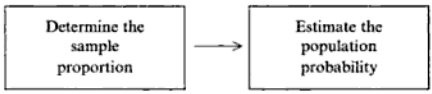
This time the tree has been labelled 'After first coin', 'After second coin' and 'After third coin'. We can think of the three coins as being tossed one after the other, so as to identify which coin is which. More mischievously, we can imagine having written the words 'First coin', 'Second coin' and 'Third coin' on the coins before tossing them. The required probability is again $\frac{1}{8}$.

Although all three problems refer to coin tosses, they describe different physical situations that are all equivalent in terms of their probability structure. This is an example of a general principle: most probability problems can be translated into problems concerning either the tossing of (possibly bent) coins, or the drawing of coloured balls from boxes! The setters of probability problems do their best to disguise this fact!

4.11 Sample proportions and probability

So far the probability to be associated with an event has been expressed in terms of the numbers of simple events in a sample space in which all the possible outcomes are equally likely. An alternative view of probability is a consequence of the general idea that a sample of observations gives information about the population from which it is derived. The bigger the sample, the more reliable is the information.

We have to adapt this approach when the outcomes in the sample space are no longer equally likely. For example, if we are interested in the probability that a bent penny comes down heads, then an obvious approach is to toss the penny a number of times (our sample) and see what proportion of the time a head is obtained:



As the sample size increases, so the observed sample proportion of occasions on which the event E occurs will vary. However, the variations will generally decrease in magnitude, and we expect that the observed sample proportion will approach a value that we will take to be the probability of E and will denote by $P(E)$.

Consider the following two situations:

Experiment	Event
A fair die is tossed.	*A*: A 6 is obtained.
A car is chosen at random.	*B*: The car is white.

For event A it seems reasonable that if we were to roll a fair die a huge number of times then 'obviously' the event A would occur on about one-sixth of occasions: $P(A) = \frac{1}{6}$. There is no need to do any real sampling – we need only think about it!

For event B, however, there is no alternative to real sampling. To have any idea of the value of $P(B)$, we need to examine a large sample of cars to find out what (roughly) is the proportion of cars that are white.

Project

So, what is the probability of the event, B, that a randomly chosen car is white? To answer this, you need to go to a convenient nearby road and count cars, keeping a tally of the number that are white. To see how the sample proportion stabilises as the sample size increases, complete the following table for your results:

Number of cars n	*Number of white cars* w	*Sample proportion* $p = \frac{w}{n}$
2		
5		
10		
50		
100		
200		
500		

You may wish to stop before seeing 500 cars, if the road is not a busy one!

Your best estimate of $P(B)$ is simply your final value for p. Compare the value that you get with the values obtained by others in your class (who, hopefully, all observed different sets of cars). Decide on a class estimate for $P(B)$.

Computer project

Computers are a good source of so-called 'random numbers'. For now, all we need to know about these numbers is that, if the random-number generator is set to produce numbers between 0 and 1, and is working correctly, then exactly 10% of the random numbers will have values less than 0.1. In probability terms, if E is the event 'number less than 0.1', then, theoretically, $P(E) = 0.1$.

Write a computer program to produce a table of the form shown for the car project above. Since the computer is doing the counting the table can go on a little longer – a final sample size of 10 000 should suffice!

Because of random variation, you should not expect always to see exactly 1000 'successes', but the theory discussed later in Chapter 10 suggests that you are likely to obtain between 940 and 1060 'successes', corresponding to an estimate of $P(E)$ in the range 0.094 to 0.106.

Calculator practice

Many calculators also have an inbuilt random-number generator which generates random numbers between 0 and 1.

If your calculator is programmable, then you could write a short program to simulate the rolling of a six-sided fair die. A random number between 0 and $\frac{1}{6}$ would correspond to 1, a number between $\frac{1}{6}$ and $\frac{2}{6}$ would correspond to 2, and so on.

Use such a program to simulate 6000 rolls of a die and to count the numbers of 1's, 2's and so on. If the random-number generator is fair you should nearly always get between 900 and 1100 of each of the six outcomes.

4.12 Unequally likely possibilities

The results so far have been obtained while considering equally likely simple events. However, this restriction is artificial and Equations (4.2) to (4.5) hold equally well for unequally likely events.

Example 10

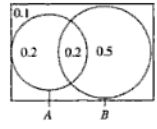
The events A and B are such that $P(A) = 0.4$, $P(B') = 0.3$ and $P(A \cap B) = 0.2$.
Determine (i) $P(A \cup B)$, (ii) $P(A' \cap B')$.

(i) Since $P(B') = 0.3$, $P(B) = 1 - 0.3 = 0.7$

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

$$P(A \cup B) = 0.4 + 0.7 - 0.2 = 0.9$$

(ii) By inspection of a Venn diagram we can see that $P(A' \cap B') = 1 - P(A \cup B)$. Thus $P(A' \cap B') = 0.1$.



4.13 Physical independence

The coin-tossing examples of Sections 4.4 (p. 93) and 4.10 (p. 101) are examples of situations in which the separate components (e.g. toss of one coin and toss of another coin) are physically independent events. By **physical independence** we mean that the outcome of one component (e.g. the first toss) can have no possible influence on the outcome of any other component (e.g. the second toss).

The multiplication rule

If A and B are two events *relating to physically independent situations* then:

$$P(A \cap B) = P(A) \times P(B) \tag{4.6}$$

More generally, if $A, B, \ldots, N$ *all relate to physically independent situations* (for example, N separate tosses of a coin) then:

$$P(A \cap B \cap \cdots \cap N) = P(A) \times P(B) \times \cdots \times P(N) \tag{4.7}$$

This very useful result is known as the **multiplication rule**.

Example 11

A bent penny has probability 0.8 of coming down heads when it is tossed. The penny is tossed six times.

What is the probability that it shows heads on every occasion?

The six tosses are physically independent – there is no way that the outcome of one of the tosses can affect the outcomes of the other tosses. Therefore:

$$\begin{aligned} \text{P(6 heads)} &= \text{P('Head on first toss' } \textit{and} \text{ 'Head on second toss' } \cdots \textit{and} \text{ 'Head on sixth toss')} \\ &= \text{P('Head on first toss')} \times \text{P('Head on second toss')} \times \cdots \times \text{P('Head on sixth toss')} \\ &= 0.8 \times 0.8 \times \cdots \times 0.8 \\ &= 0.8^6 \\ &= 0.262 \text{ (to 3 d.p.)} \end{aligned}$$

The probability of getting 6 heads with the bent penny is just over a quarter.

Example 12

A computer system consists of a keyboard, a monitor and the computer itself. The three parts are manufactured separately. From past experience it is known that, on delivery, the probability that the monitor works correctly is 0.99, the probability that the keyboard works correctly is 0.98 and the probability that the computer works correctly is 0.95. What is the probability that:

(i) the entire system works correctly,

(ii) exactly two of the components work correctly?

Define the events M, K and C as follows:

M: The monitor works correctly.
K: The keyboard works correctly.
C: The computer works correctly.

In part (i) we want $\text{P}(M \textit{ and } K \textit{ and } C) = \text{P}(M \cap K \cap C)$. Since the parts are manufactured separately the three events refer to physically independent manufacturing processes and therefore:

$$\begin{aligned} \text{P}(M \cap K \cap C) = \text{P}(M) \times \text{P}(K) \times \text{P}(C) &= 0.99 \times 0.98 \times 0.95 \\ &= 0.922 \text{ (to 3 d.p.)} \end{aligned}$$

To answer part (ii) we have to examine a number of possibilities. The situation of interest is one in which just one of the three components is working incorrectly (or not working at all!). This may be the monitor or it may be the keyboard or it may be the computer. Writing it all out in words would be dreadfully tedious, so we use the union/intersection notation. The event of interest is:

$$E = E_1 \cup E_2 \cup E_3$$

where:

$$E_1 = (M' \cap K \cap C), \qquad E_2 = (M \cap K' \cap C), \qquad E_3 = (M \cap K \cap C')$$

Here, for example, M' is the complement of the event M, in other words the event: The monitor does not work correctly.

The events E_1, E_2 and E_3 are mutually exclusive, so, using the addition rule:

$$P(E) = P(E_1) + P(E_2) + P(E_3)$$

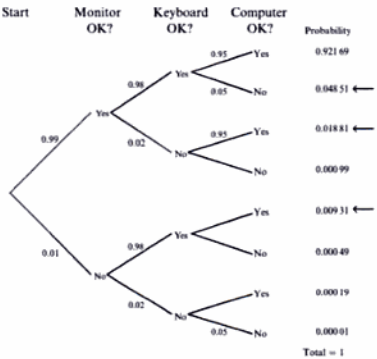
Because of physical independence:

$$P(E_1) = P(M' \cap K \cap C) = P(M') \times P(K) \times P(C)$$

and, since $P(M') = 1 - P(M)$, we finally get:

$$\begin{aligned} P(E) &= \{1 - P(M)\} \times P(K) \times P(C) + P(M) \times \{1 - P(K)\} \times P(C) \\ &\quad + P(M) \times P(K) \times \{1 - P(C)\} \\ &= (0.01 \times 0.98 \times 0.95) + (0.99 \times 0.02 \times 0.95) \\ &\quad + (0.99 \times 0.98 \times 0.05) \\ &= 0.009\,31 + 0.018\,81 + 0.048\,51 \\ &= 0.076\,63 \\ &= 0.077 \text{ (to 3 d.p.)} \end{aligned}$$

If the above solution seems rather daunting, then a probability tree will be very welcome:



The final column gives the products of the probabilities of the corresponding branches.

Exercises 4c

1 Two fair dice, each having faces numbered 1, 1, 2, 2, 2, 3 are thrown.
Draw up a probability tree.
Hence find the probability that:
(i) the total score is 4,
(ii) the total score is less than 4,
(iii) the total score exceeds 4,
(iv) at least one die shows 2.

2 Two fair dice, each having faces numbered 1, 1, 1, 1, 2, 2 are thrown.
Draw up a probability tree for the scores.
Find the probability that:
(i) the total score is 2,
(ii) the total score is 3,
(iii) the total score is 4.
A third similar die is thrown.
Add this to your probability tree and hence find the probability that:
(iv) the total score is 4,
(v) the total score is 5.

3 A man travels to work each day by train for three days. Each day the probability that the train is late is 0.1.
Find the probability that his train to work is late on at most two occasions.

4 The probability that a biased coin comes down heads is 0.4. It is tossed three times.
Find the probability of:
(i) exactly two heads,
(ii) at least two heads.

5 Family A has two daughters and one son.
Family B has three daughters and one son.
Family C has two daughters and two sons. One child is chosen at random from each family.
Draw up a probability tree.
Find the probability that:
(i) 3 girls are chosen,
(ii) at least 2 girls are chosen,
(iii) no girls are chosen,
(iv) a girl is chosen from A and the other two are of opposite sex to one another.

6 A child is allowed a lucky dip from each of three boxes. One box contains 10 chocolates and 15 mints, one box contains 8 apples and 4 oranges, and the third box contains 7 (plastic) dinosaurs and 3 (plastic) turtles. Events A, B, C are defined as follows:
A: The child gets a chocolate and a dinosaur.
B: The child gets a mint or a turtle (or both).
C: The child gets an apple.
Find (i) $P(A)$, (ii) $P(B)$, (iii) $P(A \cap C)$, (iv) $P(B \cup C)$, (v) $P(A \cap B)$, (vi) $P(A \cup B)$.

7 A woman travels to work by car. There are three roundabouts on the road. The probability that she is delayed at the first roundabout is 0.3. The corresponding figures for the second and third roundabouts are 0.5 and 0.7 respectively.
Find the probability that:
(i) she is delayed at only one roundabout,
(ii) she is delayed at 2 or more roundabouts.

8 Two chess grand masters, Xerxes and Yorick, play a tournament of 3 games. Past experience of games between these two players suggests that the results of successive games are independent of one another and that, for each game:
P(Xerxes wins) $= \frac{1}{4}$
P(Yorick wins) $= \frac{1}{5}$
P(draw) $= \frac{11}{20}$
Determine the probabilities of each of the following events:
A: Xerxes wins all three games.
B: Exactly two games are drawn.
C: Yorick wins at least one game.
D: Xerxes wins more games than Yorick wins.

4.14 Orderings

Consider the following problem.

> Four markers are arranged in a line. The markers are labelled A, B, C and D. Assuming that all possible arrangements are equally likely, determine the probability that the markers are in the order ABCD.

A systematic (alphabetical) list of the possible arrangements is:

ABCD	ABDC	ACBD	ACDB	ADBC	ADCB
BACD	BADC	BCAD	BCDA	BDAC	BDCA
CABD	CADB	CBAD	CBDA	CDAB	CDBA
DABC	DACB	DBAC	DBCA	DCAB	DCBA

In all there are 24 possible orderings of the markers. Since each ordering is equally likely, the required probability is $\frac{1}{24}$.

The problem with this sort of approach is that frequently the number of elementary events in the sample space is so large that we may miss a few out! What is needed is a formula that allows us to count the possibilities *without* actually making a list. This formula can be deduced from study of the table of possibilities given above. There are 4 possibilities for the first marker. Suppose that this is A (the possible orderings are those in the first row of the table). There are then 3 possible candidates for the second marker (B, C or D). Suppose that B is second. Then there are 2 possibilities for third place (C or D), with whichever is left being in last place. We see that there are 24 possible orderings because $4 \times 3 \times 2 \times 1 = 24$.

In general, therefore, if there were n objects, the number of possible orderings would be:

$$n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

This is tedious to write out, so we use the notation:

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

The quantity $n!$ is read as 'n **factorial**'.

Notes

- $(n+1)! = (n+1) \times n!$
- For convenience, $0!$ is defined to be equal to 1.

Calculator practice

Check out the values of 0!, 1!, 2! and so on on your calculator.
What happens when you try to calculate 99!?
Why does this happen?
What is the largest value of n for which your calculator can calculate n!?
You could also try to calculate 4.5! More expensive calculators will give a value of about 52.3, whereas cheaper or older calculators will refuse to give an answer. If your calculator does give an answer, then you might like to plot the values of, say, 3.9!, 4!, 4.1!, ..., 5!, 5.1! on graph paper. Do you get a smooth curve?

Example 13

A supermarket uses a code to identify each product that it stocks. The code consists of an ordering (without repetition) of the letters A–E, followed by an ordering (without repetition) of the numbers 1–6. How many different codes can be formed?

The number of orderings of 5 objects is $5! = 5 \times 4! = 5 \times 24 = 120$. The number of orderings of 6 objects is $6! = 6 \times 5! = 720$. Since each ordering of the letters can be associated with any one of the 720 orderings of the numbers, there are a total of $120 \times 720 = 86\,400$ different codes.

Orderings of similar objects

Consider the following problem.

> Four markers are arranged in a line. The markers are labelled A, B, A and B. Determine the number of distinguishable orderings.

The only change from the previous situation is that the marker labelled C is now labelled A, while D has become B. Making the appropriate adjustments to the previous table, we get:

ABAB	ABBA	AABB	AABB	ABBA	ABAB
BAAB	BABA	BAAB	BABA	BBAA	BBAA
AABB	AABB	ABAB	ABBA	ABAB	ABBA
BABA	BAAB	BBAA	BBAA	BAAB	BABA

There is now a lot of repetition! The only distinguishable arrangements are (in alphabetical order) AABB, ABAB, ABBA, BAAB, BABA and BBAA. The reduction comes about because A followed by C and C followed by A now give an identical result (A followed by A). This halves the number of distinguishable orderings. A similar halving results from the replacement of D by B.

The general rule is as follows:

If there are n objects, consisting of a of one type, b of a second type, and so on, then the number of *distinct* arrangements of the objects in a line is:

$$\frac{n!}{a!b!\ldots}$$

Example 14

The four letters of the word COOK are arranged in a line.

(i) How many distinct arrangements are there?

(ii) If an arrangement is chosen at random, what is the probability that the two Os are consecutive?

(i) There are 4 letters, consisting of 1 C, 2 Os and 1 K. The number of arrangements is therefore:

$$\frac{4!}{1!1!2!} = \frac{24}{1 \times 1 \times 2} = 12$$

There are 12 possible arrangements of the letters in the word COOK.

(ii) We require the two Os to be consecutive. Imagine that they are glued together. We then have only three items to arrange in order: C, OO and K. The number of possible orderings is $3! = 6$. Thus 6 of the 12 possible arrangements of the letters in the word COOK involve a double O: the required probability is $\frac{6}{12} = \frac{1}{2}$.

In this question the number of orderings is small so that we could write them all out. Life is not always that easy, however, as the next example demonstrates.

Example 15

Five chairs are arranged in a line. Five boys are to be seated on the chairs. If Alfred and Bruce sit next to each other then a fight is sure to start.

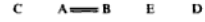
(i) How many possible arrangements are there if there are no restrictions on the seating arrangements?

(ii) If the boys are assigned seats at random, what is the probability that Alfred and Bruce are not sitting next to one another?

(i) There are $5! = 120$ possible arrangements, all equally likely.

(ii) The easy way to answer this is to consider the complementary event 'a fight starts'! Imagine that Alfred and Bruce are 'glued' together in the order AB. There are now 4 'objects' (boys or doubleboys) to be arranged in order.

There are then $4! = 24$ possible arrangements of the objects. There are a further 24 possible arrangements with Albert and Bruce 'glued' in the order BA. In all, therefore, there are 48 unsatisfactory arrangements and therefore $120 - 48 = 72$ satisfactory arrangements. The probability that Albert and Bruce are not sitting next to each other is therefore $\frac{72}{120} = \frac{3}{5}$.



One arrangement with A, B 'glued'

Arrangements of n objects in a circle are more restrictive because there are n possible 'starting points' for the circle. Denoting the directions North, South, East and West by the letters N, S, E and W, the familiar clockwise ordering NESW could also be represented as ESWN, SWNE or WNES, depending upon one's starting point. The general rule for objects arranged in a circle is as follows.

The number of arrangements of n objects arranged in a circle is equal to the corresponding number of arrangements on a line, divided by n.

Note

- If the circle can be 'turned over', so that clockwise and anticlockwise arrangements are indistinguishable, the number of arrangements is equal to the corresponding number of arrangements on a line, divided by $2n$ rather than n.

Example 16

If the five chairs of the previous example are now arranged in a circle, what is the probability that Albert and Bruce are not sitting next to each other?

The number of equally likely distinct arrangements is now $\frac{120}{5} = 24$. The number of AB arrangements is now $\frac{24}{4} = 6$, and the number of BA arrangements is also 6, so the total number of unsatisfactory arrangements is 12. The probability that Albert and Bruce do not sit next to each other is therefore $\frac{1}{2}$, somewhat smaller than before.

One arrangement with A, B 'glued'

Note

- A simpler solution to the last example is as follows. Obviously Albert is sitting somewhere. This leaves four seats. Bruce is equally likely to be in any of these. Two of the four are not next to Albert, so the required probability is $\frac{2}{4} = \frac{1}{2}$.

  This type of argument can also be used for Example 15 (ii), but care is needed since the cases where Albert sits at an end of the line (and therefore has just one neighbour) are different from the other cases where he has two neighbours.

Example 17

The five letters of the word UPTON are arranged in a line. How many different arrangements are possible?

The letters are arranged in a random order. Find the probability that
(i) the two vowels are next to each other,
(ii) the two vowels are not next to each other,
(iii) either the letters NOT appear next to each other and in that order or the letters UP appear next to each other and in that order (or both).

All the letters are different, so the number of different arrangements is $5! = 120$.

(i) We can glue the two vowels together in 2! orders. We now have 4 objects, (for example, OU, P, T and N) that can be ordered in 4! ways. There are therefore $2! \times 4!$ orderings and the probability of the two vowels being next to each other is therefore

$$\frac{2! \times 4!}{5!} = \frac{2}{5}$$

(ii) This is the complementary event to the event in part (i) and the probability is $1 - \frac{2}{5} = \frac{3}{5}$

(iii) There are two events of interest:

A The letters NOT appear next to each other

B The letters UP appear next to each other

We are being asked to find $P(A \cup B)$. One way of doing this is to find each of $P(A)$, $P(B)$ and $P(A \cap B)$.

If the letters NOT appear next to each other and in that order then we have three objects to arrange in order (U, P and NOT). There are 3! possible arrangements.

Similarly, if the letters UP appear next to each other and in that order then there are four objects to arrange in order (UP, T, O and N) with 4! possible arrangements.

Finally, if both the letters NOT appear next to each other and the letters UP appear next to each other then there are just two objects to arrange in order (UP and TON) with 2! possible arrangements.

The probability that either the letters NOT appear next to each other and in that order or the letters UP appear next to each other and in that order (or both) is therefore given by

$$\frac{3!}{5!} + \frac{4!}{5!} - \frac{2!}{5!} = \frac{6 + 24 - 2}{120} = \frac{28}{120} = \frac{7}{30}$$

Exercises 4d

1 Six children, Alice, Brenda, Caroline, David, Edward and Frank, stand in a line.
How many different orders are possible?
They stand in random order.
Find the probability that:
(i) the three girls are next to each other,
(ii) Brenda and Frank are next to each other,
(iii) Caroline and David are not next to each other.

2 A hand of cards consists of all 13 Hearts from an ordinary pack. In how many different orders can they be arranged?
The cards are arranged in random order.
Find the probability that:
(i) the Ace is first and the King is last,
(ii) the Ace and King, in either order, are the first two cards,
(iii) either the Ace is first or the King is last or both,
(iv) the Ace is somewhere in front of the King.

3 I empty out my purse. There are four 1p coins, three 2p coins, two 5p coins and one 10p coin. Assume that coins of the same value are indistinguishable from each other. In how many different ways can the 10 coins be arranged in a line?
In how many of these ways are the three 2p coins all next to each other?
The coins are arranged in a line in random order.
Find the probability that:
(i) the two 5p coins are not next to each other,
(ii) the 10p coin has a 5p coin next to it on each side.

4 Thirteen counters, 4 red, 4 green, 3 blue and 2 yellow, are arranged in order in a line. The counters are identical except for their colour.
Find the number of distinguishable orderings.
The counters are arranged in random order.
Find the probability that:
(i) the 4 green counters are all next to each other,
(ii) all counters of the same colour are next to each other.

5 Six novels, labelled A, B, C, D, E, F, have to be arranged in order of merit for a literary prize.
Find the total number of different ways in which this can be done.
Suppose that the novels are arranged in random order. Find the probability that:
(i) F is first,
(ii) A is last,
(iii) C is first and D is second,
(iv) D comes immediately after C,
(v) either B or E (or both) appear in the first two places.

6 The six children, Alice, Brenda, Caroline, David, Edward and Frank, now stand in a circle. Distinguishing between clockwise order and anticlockwise order, find the number of different possible orders.
Find the probability that:
(i) the three girls are next to each other,
(ii) Brenda and Frank are next to each other,
(iii) Caroline and David are not next to each other.

7 Find the number of different arrangements of the six letters in the word ELEVEN in which
(i) all three letters E are consecutive,
(ii) the first letter is E and the last letter is N.
[UCLES]

4.15 Permutations and combinations

Consider the following problem.

> A pack of 52 playing cards (all different) is shuffled. Determine the probability that the top card in the pack is the Ace of Spades, the next is the Ace of Hearts and the next is the Ace of Diamonds.

Now any one of the 52 cards could have been at the top of the pack. This leaves 51 cards, any one of which might have been next. Similarly, there are 50 possibilities for the third card. There are therefore a total of $52 \times 51 \times 50 = 132\,600$ possibilities for the first three cards in order. Only one of these corresponds to the event described, so the probability of that event is $\frac{1}{132\,600}$.

The number of *ordered* arrangements of r objects chosen from a collection of n objects, is denoted by ${}^nP_r$ (read as **'n p r'** or **'n perm r'**) and each ordering is called a **permutation** of the selected objects.

The value of ${}^nP_r$ is given by:

$$ {}^nP_r = n \times (n-1) \times \cdots \times (n-r+1) \tag{4.8} $$

Note that there are r terms in the expression on the right of this equation. An alternative expression, using factorials, is:

$$ {}^nP_r = \frac{n!}{(n-r)!} $$

Using Equation (4.8) with $n = 52$ and $r = 3$, we get ${}^{52}P_3 = 52 \times 51 \times 50 = 132\,600$, as before.

Consider now the slightly different problem.

> A pack of 52 playing cards (all different) is shuffled. Determine the probability that the top three cards in the pack are the Ace of Spades, the Ace of Hearts and the Ace of Diamonds.

This problem differs from the previous one in that *the order in which the cards appear is irrelevant*. There are $3! = 6$ possible orders for three cards, so the number of *distinguishable* groups of three cards, chosen from 52, is the number of ordered possibilities (132 600) divided by 6 giving the answer 22 100. The probability that the first three cards are the three aces is therefore $\frac{1}{22\,100}$.

The number of *unordered* arrangements of r objects selected from a collection of n objects, is denoted by either ${}^nC_r$ or $\binom{n}{r}$. In this book we use the second form which is that used in all modern advanced statistical texts. In either case the formula is read as either **'n c r'** or **'n choose r'**. Each collection of selected objects is a **combination**.

The general formula for $\binom{n}{r}$ is:

$$ \binom{n}{r} = \frac{{}^nP_r}{r!} = \frac{n \times (n-1) \times \cdots \times (n-r+1)}{r \times (r-1) \times \cdots \times 1} = \frac{n!}{(n-r)!r!} \tag{4.9} $$

It should be noted that the fraction:

$$ \frac{n \times (n-1) \times \cdots \times (n-r+1)}{r \times (r-1) \times \cdots \times 1} $$

has r terms in both the numerator and the denominator.

Using Equation (4.9) with $n = 52$ and $r = 3$, we get:

$$ \binom{52}{3} = \frac{52 \times 51 \times 50}{3 \times 2 \times 1} $$
$$ = 22\,100 $$

Notes

- $\binom{n}{r} = \binom{n}{n-r}$; $\binom{n}{0} = \binom{n}{n} = 1$
- Some calculators have buttons for calculating permutations and combinations.
- Combinations occur naturally in the context of the binomial expansion, since:

$$ (a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r} $$

Example 18

A woman is planting rose bushes. She has eight different bushes, each with a different colour flower, and she will plant five of the bushes in her back garden.
How many different choices does she have?

Order matters here, so the number of possible arrangements is:

$${}^8P_5 = \frac{8!}{5!} = 8 \times 7 \times 6 = 336$$

Example 19

A pack of cards is shuffled and a 'hand' of 13 randomly chosen cards is dealt to one card player.
How many possible hands can that player receive?

In this case the order in which the player receives the cards is irrelevant. The number of possible hands is therefore:

$$\binom{52}{13} = \frac{52 \times 51 \times 50 \times \cdots \times 41 \times 40}{13 \times 12 \times 11 \times \cdots \times 2 \times 1}$$

$$\approx 6.35 \times 10^{11}$$

There are about 635 thousand million possible hands!

Example 20

At the beginning of a game show a contestant is allowed a five-second glimpse of a table on which is placed a fluffy toy and four other objects (all different). At the end, the contestant is asked to name as many of the objects as possible.
(i) How many different combinations of objects might be named?
(ii) What proportion include the fluffy toy?

(i) The contestant may name 0, 1, 2, 3, 4 or 5 of the objects. The total number of combinations is therefore:

$$\binom{5}{0} + \binom{5}{1} + \binom{5}{2} + \binom{5}{3} + \binom{5}{4} + \binom{5}{5}$$

$$= 1 + 5 + 10 + 10 + 5 + 1 = 32$$

(ii) Given that the fluffy toy is named, the contestant may name up to four of the remaining objects. The total number of combinations including the fluffy toy is therefore:

$$\binom{4}{0} + \binom{4}{1} + \binom{4}{2} + \binom{4}{3} + \binom{4}{4} = 1 + 4 + 6 + 4 + 1 = 16$$

The proportion of the combinations that include the fluffy toy is therefore $\frac{16}{32} = \frac{1}{2}$.

Note

- An alternative approach to part (i) is to argue that each of the five objects can either be 'chosen' or 'not chosen'. There are therefore 2 possibilities for each of 5 objects, so the total number of combinations is $2^5 = 32$. In part (ii) the number of possibilities is reduced to $2^4 = 16$, and so the required proportion is $\frac{16}{32} = \frac{1}{2}$.

Exercises 4e

1 A delegation of 3 students is to be chosen from a class of 15.
In how many ways can this be done?
The class consists of 10 girls and 5 boys.
(i) If two of the delegates are to be girls and the other is to be a boy, in how many ways can this be done?
(ii) If the delegation is to include at least one boy and at least one girl, in how many ways can this be done?

2 How many different hands of 13 cards, drawn from an ordinary pack, are there that contain 6 Spades, 4 Hearts, 2 Diamonds and 1 Club? How many hands contain 6 from one suit, 4 from another, 2 from another and one from the fourth suit?

3 In the state of Utopia, the alphabet contains 25 letters. A car registration number consists of two **different** letters of the alphabet followed by an integer n such that $1004 \leqslant n \leqslant 4999$. Find the number of possible car registration numbers. [UCLES(P)]

4 A nursery school teacher has 4 apples, 3 oranges, and 2 bananas to share among 9 children, with each child receiving one fruit. Find the number of different ways in which this can be done. [UCLES]

5 A code consists of blocks of ten digits, four of which are zeros and six of which are ones; e.g. 1011011100. Calculate the number of such blocks in which the first and last digits are the same as each other. [UCLES]

6 A computer terminal displaying text can generate 16 different colours numbered 1 to 16. Any one of colours 1 to 8 may be used as 'background colour' on the screen, and any one of colours 1 to 16 may be used as 'text colour'; however, selecting the same colour for background and text renders the text invisible so this combination is not used. Find the number of different usable combinations of background colour and text colour. [UCLES]

7 Find the number of ways in which 4 questions can be chosen from the 7 questions in an examination paper, assuming that the order in which the questions are chosen is not relevant. [UCLES]

8 Prizes are to be awarded to four different members of a group of eight people. Find the number of ways in which the prizes can be awarded
(i) if there is a 1st prize, a 2nd prize, a 3rd prize and a 4th prize,
(ii) if there are two 1st prizes and two 2nd prizes. [UCLES]

9 Twelve horses run in a race. The published results list the horses finishing first, second and third. Assuming there are no dead-heats, find the number of different possible published results. [UCLES]

10 A party of 12 people is to make a journey in 3 cars, with 4 people in each car. Each car is driven by its owner. Find the number of ways in which the remaining 9 people may be allocated to the cars. (The arrangement of people within a particular car is not relevant.) [UCLES]

11 The digits of the number 314152 are rearranged so that the resulting number is odd. Find the number of ways in which this can be done. [UCLES]

12 A school is asked to send a delegation of six pupils selected from six badminton players, six tennis players and five squash players. No pupil plays more than one game. The delegation is to consist of at least one, and not more than three, players drawn from each game. Giving full details of your working, find the number of ways in which the delegation can be selected. [UCLES(P)]

4.16 Sampling with replacement

This is easy! The situation is one of physical independence and we can use the addition and multiplication rules and probability trees. Here is a typical problem.

> A pack of cards consists of the Queens of Spades, Hearts, Diamonds and Clubs together with the Ace, King and Jack of Spades. The pack is shuffled and a card is chosen at random. After its identity has been noted, the card is replaced in the pack, which is again shuffled. This is repeated on two further occasions.
> Determine the probability that a Queen is chosen on only one occasion.

On each occasion the probability that a Queen is chosen is $\frac{4}{7}$. Using Q to denote a Queen and R to denote one of the other cards, the possibilities that include exactly one Queen are RRQ, RQR and QRR. For each of these possibilities, the probability is the product of $\frac{3}{7}$, $\frac{3}{7}$ and $\frac{4}{7}$, so the overall probability is:

$$3 \times \left(\frac{3}{7}\right)^2 \times \frac{4}{7} = \frac{108}{343}$$

which is about 0.315 (to 3 d.p.).

4.17 Sampling without replacement

Consider the following problem.

> A pack of cards consists of the Queens of Spades, Hearts, Diamonds and Clubs together with the Ace, King and Jack of Spades. The pack is shuffled and three cards are chosen at random.
> Determine the probability that just one of the three cards is a Queen.

This is similar to the previous problem, but in this case the cards must be different, whereas in the previous case the same card might have been selected on more than one occasion.

In our new problem the order of selection is again unimportant and we are therefore concerned with combinations rather than permutations. The number of distinct combinations of three cards chosen from seven cards is:

$$\binom{7}{3} = \frac{7 \times 6 \times 5}{3 \times 2 \times 1} = 35$$

These are listed systematically in the following table using the shorthand of A, K and J for the Ace, King and Jack and with S, H, D and C representing the four queens.

When making lists it is important to work systematically (or we will get hopelessly lost!). In this case we work alphabetically:

ACD	ACH	<u>ACJ</u>	<u>ACK</u>	ACS	ADH	<u>ADJ</u>
<u>ADK</u>	ADS	<u>AHJ</u>	<u>AHK</u>	AHS	AJK	<u>AJS</u>
<u>AKS</u>	CDH	CDJ	CDK	CDS	CHJ	CHK
CHS	<u>CJK</u>	CJS	CKS	DHJ	DHK	DHS
<u>DJK</u>	DJS	DKS	<u>HJK</u>	HJS	HKS	<u>JKS</u>

The 12 outcomes corresponding to the event of interest are underlined.

For each of the $\binom{4}{1} = 4$ possible selections of a Queen there are $\binom{3}{2} = 3$ possible selections of two other cards from the three available. The total number of possibilities is the product $\binom{4}{1} \times \binom{3}{2} = 4 \times 3 = 12$. The probability of the event of interest is:

$$\frac{\binom{4}{1} \times \binom{3}{2}}{\binom{7}{3}} = \frac{12}{35}$$

This problem is a simple example of a general type illustrated by the following.

A box contains a total of N balls. The balls are of k different types. There are N_1 balls of type 1, N_2 of type 2, and so on $\left(\sum_{i=1}^{k} N_i = N\right)$. A random sample of n balls is taken from the box *without replacement*. What is the probability that the sample contains exactly n_1 balls of type 1, n_2 of type 2, and so on $\left(\sum_{i=1}^{k} n_i = n\right)$? The order of selection is unimportant.

In this case an outcome consists of an unordered collection of n balls. The total number of outcomes in the sample space is the number of ways in which a random sample of n balls can be selected from a group of N balls. This is just the number of ways of choosing n from N, which is $\binom{N}{n}$.

The number of ways of choosing n_1 balls from the N_1 balls of type 1, is $\binom{N_1}{n_1}$. Whichever of these selections occurs there are also $\binom{N_2}{n_2}$ selections of balls of type 2, and so on. The total number of selections corresponding to the required event (i.e. the total number of outcomes) is therefore:

$$\binom{N_1}{n_1} \times \binom{N_2}{n_2} \times \cdots \times \binom{N_k}{n_k}$$

The probability of simultaneously choosing n_1 from N_1, n_2 from N_2, and so on, is therefore:

$$\frac{\binom{N_1}{n_1} \times \binom{N_2}{n_2} \times \cdots \times \binom{N_k}{n_k}}{\binom{N}{n}}$$

Note

- The amount of thought required for this sort of problem can be minimised as follows! Write down in a row the numbers of each of the different types in the population (i.e. $N_1, N_2, \ldots, N_k$, which sum to N). In a row below these write down the corresponding numbers that are required for the sample (including zeros). These are the numbers $n_1, n_2, \ldots, n_k$ which sum to n. With suitably placed brackets we have the required numerator while $\binom{N}{n}$ provides the denominator.

Example 21

A committee of five is chosen by drawing lots from a group of eight men and four women.
Determine the probability that the committee contains exactly three men.

Since nobody can be chosen more than once, selection is without replacement. An outcome consists of an unordered group of three people. We now suspend thought and simply identify the values of the parts of N and n. We have $N = 12$, $n = 5$, $N_1 = 8$, $N_2 = 4$, $n_1 = 3$ and $n_2 = 2$. Hence:

$$\frac{\binom{N_1}{n_1} \times \binom{N_2}{n_2}}{\binom{N}{n}} = \frac{\binom{8}{3} \times \binom{4}{2}}{\binom{12}{5}}$$

$$= \frac{8 \times 7 \times 6}{3 \times 2 \times 1} \times \frac{4 \times 3}{2 \times 1} \times \frac{5 \times 4 \times 3 \times 2 \times 1}{12 \times 11 \times 10 \times 9 \times 8}$$

$$= 56 \times 6 \times \frac{1}{792}$$

$$= \frac{14}{33}$$

The probability that the committee contains exactly three men is $\frac{14}{33}$, which is 0.424 to 3 decimal places.

Example 22

A notorious gang of outlaws contains five gunfighters called Smith, four called Jones and one called Cassidy. In a gunfight, three of the gang are killed. Assuming that each gunfighter had the same probability of being killed, what is the probability that the three killed in the gunfight all had different names?

This time the outcomes are unordered groups of three outlaws. There are three types of outlaw: Smith, Jones and Cassidy. The numbers of these are 5, 4 and 1 (total 10), while the numbers required are 1, 1 and 1 (total 3). Hence the required probability is:

$$\frac{\binom{5}{1} \times \binom{4}{1} \times \binom{1}{1}}{\binom{10}{3}} = \frac{5 \times 4 \times 1}{120}$$

$$= \frac{1}{6}$$

The probability that the three ex-gunfighters had different names is $\frac{1}{6}$.

Example 23

Three letters are chosen at random (without replacement) from the word STATISTICS. What is the probability that:
(i) they are all the same,
(ii) they are all consonants,
(iii) they are all different,
(iv) exactly two are the same?

The sample space consists of all possible unordered selections of 3 letters from the 10 letters S, T, A, T, I, S, T, I, C and S. The number of outcomes is the number of ways of choosing 3 letters from 10 letters (ignoring the repetition of the letters), and is therefore $\binom{10}{3} = 120$.

(i) One of these selections is SSS and another is TTT. These are the only outcomes that consist of three letters all the same and so the required probability is $\frac{2}{120} = \frac{1}{60}$.

(ii) There are seven consonants and three vowels in STATISTICS. The number of ways of choosing three consonants and no vowels is $\binom{7}{3} \times \binom{3}{0} = 35$. Hence the probability of this event is $\frac{35}{120} = \frac{7}{24}$.

(iii) Part (i) was simple because it was easy to spot that there were only two possible outcomes. Part (ii) was easy because the letters were split into two types. But this part is more difficult because there are 5 types of letter (S, T, A, I and C) to consider and – worse still – we need to consider these three at a time.

Since $\binom{5}{3} = 10$, there are 10 different types of outcomes to consider. These are (ignoring order) STI, STA, STC, SIA, SIC, SAC, ITA, ITC, IAC and TAC. The number of ways of obtaining, in some order, an outcome of type STI, is the number of ways of obtaining an S, times the number of ways of obtaining a T, times the number of ways of obtaining an I. This is:

$$\binom{3}{1} \times \binom{3}{1} \times \binom{2}{1} = 18$$

The table below shows the numbers of possibilities for all ten types of outcome.

Outcome type	STI	STA	STC	SIA	SIC	SAC	TIA	TIC	TAC	IAC	Total
Number of possibilities	18	9	9	6	6	3	6	6	3	2	68

The probability that the three chosen letters are all different is $\frac{68}{120} = \frac{17}{30}$.

(iv) This question can be answered by enumeration, though care is needed to make sure that all the possibilities have been noted. The method proceeds as before. Thus, for an outcome of type SSI the number of possibilities is:

$$\binom{3}{2} \times \binom{2}{1} = 6$$

However, it is simpler to recognise that we have already done the hard work! The three letters will either be all the same, all different or will have two letters the same and one different. We have calculated that there are 2 combinations in which the letters are all the same and 68 in which they are all different. By subtraction, therefore, the number of combinations in which one letter occurs exactly twice is 50 and the probability required is $\frac{50}{120}$, i.e. $\frac{5}{12}$.

Exercises 4f

1 There are ten bottles arranged in a random order on a shelf. Five are green, three are blue and two are yellow. Two bottles are knocked off the shelf. Determine the probability that:
(i) both bottles are green,
(ii) both bottles are the same colour,
(iii) the bottles are of different colours.

2 A class of 100 students comprises a group of 40 people called 'idiots' and a group of 60 called 'complete idiots'. A sample of three students is selected at random from the class. Determine the probability that the sample contains more 'complete idiots' than 'idiots'.

3 A bag of fruit contains 5 apples, 8 oranges and 3 pears. Three fruit are chosen, at random and without replacement, from the bag.
Find the probability that:
(i) no apples are chosen,
(ii) all the chosen fruit are different,
(iii) exactly one apple is chosen,
(iv) exactly two apples are chosen,
(v) three apples are chosen,
(vi) two apples and one orange are chosen.

4 A man is taking 12 shirts with him on a flight. He takes 4 formal shirts and 8 casual shirts, of which 3 are long-sleeved and 5 are short-sleeved. He splits his shirts randomly between his two cases, putting 6 shirts in each case. One of his cases is lost.
Find the probability that he has lost:
(i) exactly three formal shirts,
(ii) more than two formal shirts,
(iii) all his long-sleeved casual shirts.

5 A committee consists of 5 people: Anne, Bridget, Charles, Diana and Edward. Two members are to be chosen at random to be Chair and Vice-chair.
In how many different ways can these offices be filled?
Find the probability that:
(i) both the members chosen are men,
(ii) both are women,
(iii) the Chair is a woman and the Vice-chair is a man,
(iv) the Chair is a man and the Vice-chair is a woman,
(v) the two are of opposite sex.

6 Manjula has the following coins in her purse: eight 1p coins, three 2p coins, four 5p coins, two 10p coins and four 20p coins. In the dark she drops three coins.
Find the probability that:
(i) each of the coins lost is worth 5p or more,
(ii) the total value of the three coins is 3p,
(iii) the total value of the three coins is less than 7p,
(iv) all three coins have the same value.

7 A club committee consists of 2 married couples, 3 single women and 5 single men. Four members are to be chosen at random from the 12 members of the committee to form a delegation to represent the club at a conference. Find the probabilities that the delegation will consist of
(i) 4 single men,
(ii) 2 men and 2 women. [JMB(P)]

8 A bag contains 5 red balls, 3 blue balls and 2 white balls. Four balls are drawn at random without replacement from the bag. Calculate the probability that the four balls drawn contain at least one of each colour. [WJEC]

9 A choir has 7 sopranos, 6 altos, 3 tenors and 4 basses. At a particular rehearsal, three members of the choir are chosen at random to make the tea.
(i) Find the probability that all three tenors are chosen.
(ii) Find the probability that exactly one bass is chosen. [UCLES(P)]

4.18 Conditional probability

The probability that we associate with the occurrence of an event is always likely to be influenced by the information that we have available. Suppose, for example, that I see a man lying motionless on the grass in a nearby park and am interested in the probability of the event 'the man is dead'. In the absence of other information a reasonable guess might be that the probability

is one in a million. However, if I have just heard a shot ring out, and a suspicious-looking man with a smoking revolver is standing nearby then the probability would be rather higher!

We write:

$$P(B|A)$$

to mean the probability that the event B occurs (or has occurred) given the information that the event A occurs (or has occurred).

The quantity $B|A$ is read as 'B **given** A' and $P(B|A)$ is described as a **conditional probability** since it refers to the probability that B occurs (or has occurred) *conditional* on the event that A occurs (or has occurred).

Example 24

A statistician has two coins, one of which is fair, while the other is double-headed. She chooses one coin at random and tosses it. The events A_1, A_2 and B are defined as follows:

A_1: The fair coin is chosen.
A_2: The double-headed coin is chosen.
B: A head is obtained.

Determine the values of $P(B|A_1)$ and $P(B|A_2)$.

If the fair coin is tossed then the probability of a head is $\frac{1}{2}$: $P(B|A_1) = \frac{1}{2}$.
If the double-headed coin is tossed then the probability is 1: $P(B|A_2) = 1$.

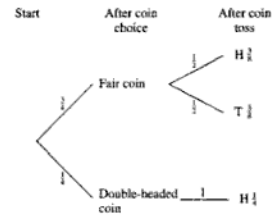
Shortly we will relate $P(B|A)$ to the unconditional probabilities of the events A, B, $A \cap B$, but first we look at two examples that involve equally likely simple events.

Example 25

A box contains four coins. Three coins are fair, but the fourth coin is double-headed. A coin is chosen at random and tossed.

(i) Determine the probability that a head is obtained.
(ii) Given that a head is obtained, determine the conditional probability that it was the double-headed coin that was tossed.

We can see the various possibilities quite easily using a probability tree.



(i) There are three alternative 'outcomes' (heads with a fair coin, tails with a fair coin, heads with the double-headed coin). Two correspond to getting a head, hence $P(\text{head}) = \frac{3}{8} + \frac{1}{4} = \frac{5}{8}$.

(ii) The contribution to P(head) from the double-headed coin is $\frac{1}{4}$ and hence the required conditional probability is

$$\frac{\frac{1}{4}}{\frac{5}{8}} = \frac{2}{5}$$

Example 26

An electronic display is equally likely to show any of the digits $1, \ldots, 8, 9$. Determine the probability that it shows a prime number (i.e. one of 2, 3, 5 and 7):

(i) given no knowledge about the number,

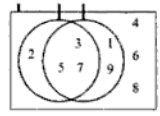
(ii) given the information that the number is odd.

Let B be the event 'a prime number' and A be the event 'an odd number'. Thus $A \cap B$ is the event 'an odd prime number'.

(i) Since there are nine possible outcomes, $n(S) = 9$. Since there are four outcomes corresponding to the event of interest, $n(B) = 4$. Since the outcomes are all equally likely, $\mathrm{P}(B) = \dfrac{n(B)}{n(S)} = \frac{4}{9}$.

(ii) Given the information that the number is odd, we know that it must be one of the $n(A)$ numbers 1, 3, 5, 7 and 9. Initially, each of these outcomes was equally likely. The knowledge that one of them has occurred does not make their chances of occurrence unequal. Of these five possible outcomes, three (3, 5 and 7) are prime. These outcomes are the simple events corresponding to the event $A \cap B$. Thus,

$$\mathrm{P}(B|A) = \frac{n(A \cap B)}{n(A)} = \frac{3}{5}.$$



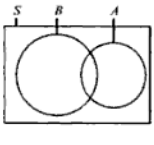
The previous examples illustrated, in two particular cases, the result that, for equally likely simple events:

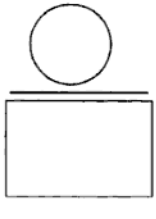
$$\mathrm{P}(B|A) = \frac{n(A \cap B)}{n(A)}$$

If we divide both the numerator and the denominator of the right-hand side of this equation by $n(S)$, we obtain:

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)} \tag{4.10}$$

This result is always true (provided A is a possible event!) and is not confined to equally likely events. We can illustrate the result using Venn diagrams.



$\mathrm{P}(B) =$ 

$\mathrm{P}(B|A) =$ 

Knowing that A has occurred means that we can ignore all of the sample space except for that part occupied by the event A. The part of A in which B also occurs is the part denoted by $A \cap B$, and Equation (4.10) is seen to be a simple statement about proportions.

Rearranging the previous equation, we get:

$$P(A \cap B) = P(A) \times P(B|A) \tag{4.11}$$

Reversing the roles of A and B:

$$P(B \cap A) = P(B) \times P(A|B)$$

Since $A \cap B$ and $B \cap A$ are descriptions of the same event, the intersection of A and B, we have:

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

and hence:

$$P(A \cap B) = P(A) \times P(B|A) = P(B) \times P(A|B) \tag{4.12}$$

The generalised multiplication rule

For three events, repeated application of Equation (4.12) gives:

$$P(A \cap B \cap C) = P(A) \times P(B|A) \times P(C|A \cap B) \tag{4.13}$$

from which the extension to larger numbers of events is clear.

Since $(A \cap B \cap C)$ is the same as, for example, $(B \cap C \cap A)$, another of many equivalent expressions for $P(A \cap B \cap C)$ is:

$$P(A \cap B \cap C) = P(B) \times P(C|B) \times P(A|B \cap C)$$

4.19 Statistical independence

Two events A and B are said to be **statistically independent** if knowledge that one occurs does *not* alter the probability that the other occurs. Formally, if A and B are two statistically independent events with non-zero probabilities, then:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \cap B) = P(A) \times P(B)$

Notes

- Any one of the above three equations is enough to guarantee independence of A and B (assuming that both have non-zero probability of occurrence).
- Physically independent events are always statistically independent.
- The words 'statistically' and 'physically' are often omitted and events are simply referred to as being 'independent'.
- Exclusive events with positive probability cannot be independent.

Example 27

Two events A and B are such that $P(A) = 0.5$, $P(B) = 0.4$ and $P(A|B) = 0.3$.

(i) State whether the events are independent.

(ii) Find the value of $P(A \cap B)$.

(i) The events A and B are *not* independent since $P(A) \neq P(A|B)$

(ii) $P(A \cap B) = P(B) \times P(A|B) = 0.4 \times 0.3 = 0.12$

Example 28

Two events A and B are such that $P(A) = 0.7$, $P(B) = 0.4$ and $P(A|B) = 0.3$.
Determine the probability that neither A nor B occurs.

It is not obvious how to answer this! One way is to 'doodle', by writing down the probabilities of things we do know! So, from Equation (4.12):

$$P(A \cap B) = P(B) \times P(A|B) = 0.4 \times 0.3 = 0.12$$

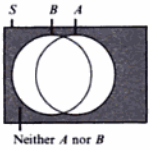
From Equation (4.3) we can now obtain $P(A \cup B)$:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.7 + 0.4 - 0.12 = 0.98$$

But, looking at a Venn diagram:

$$P(\text{neither } A \text{ nor } B) = 1 - P(A \cup B)$$

The required probability is therefore $1 - 0.98 = 0.02$



An alternative approach, more algebraic in nature, begins by organising the information in a table of probabilities of the joint events $A \cap B$, $A \cap B'$, $A' \cap B$, $A' \cap B'$, with the required value, $P(A' \cap B')$ being set equal to x.

	B	B'	Total
A			0.7
A'		x	0.3
Total	0.4	0.6	1.0

which gives:

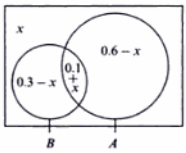
	B	B'	Total
A	$0.1 + x$	$0.6 - x$	0.7
A'	$0.3 - x$	x	0.3
Total	0.4	0.6	1.0

In obtaining the second table we have used, for example, the fact that:

$$P(B) = P(B \cap A) + P(B \cap A')$$

We also know that $P(A|B) = 0.3$. Hence $\dfrac{0.1 + x}{0.4} = 0.3$. Multiplying both sides by 0.4, we get $0.1 + x = 0.3 \times 0.4$, so that $x = 0.12 - 0.1 = 0.02$, as before.

The same approach could be adopted using the Venn diagram shown.



Example 29

The following table gives information on two aspects of the habitats of some tropical lizards for a sample of 207 habitats.

		Perch diameter (cm) 410	>10	Total
Perch height (m)	> 1.5	64	22	86
	41.5	86	35	121
Total		150	57	207

Suppose that one of the 207 habitat locations in the sample is chosen at random. Determine, correct to 2 decimal places, the probability that:
(i) the perch diameter is greater than 10 cm,
(ii) the perch diameter is greater than 10 cm, given the information that the perch height is more than 1.5 m,
(iii) the perch height is more than 1.5 m,
(iv) the perch height is more than 1.5 m, given that the perch diameter is greater than 10 cm.

Define the events A and B as follows:

A: The perch diameter is greater than 10 cm.
B: The perch height is more than 1.5 m.

We can read the answers direct from the table.

(i) $P(A) = \frac{57}{207} = 0.28$ (to 2 d.p.)
(ii) $P(A|B) = \frac{22}{86} = 0.26$ (to 2 d.p.)
(iii) $P(B) = \frac{86}{207} = 0.42$ (to 2 d.p.)
(iv) $P(B|A) = \frac{22}{57} = 0.39$ (to 2 d.p.)

Since $P(A) \approx P(A|B)$ and $P(B) \approx P(B|A)$ it appears that perch height and perch diameter are approximately independent.

Example 30

A person is chosen at random from the population. Let A be the event 'the person is female' and let B be the event 'the person is aged at least 80'. Suppose that $P(A) = 0.5$, $P(B) = 0.1$ and $P(A|B) = 0.7$. Let the event C be defined by $C = A \cap B'$.
(i) Describe the event C in real terms.
(ii) Determine $P(A|B')$.

This question is much easier to answer when written in English!
In a certain population, 50% are female, 10% are aged at least 80 and 70% of these aged people are female.
(i) The event C is 'a female aged less than 80'.
(ii) We need to find the probability that someone aged under 80 is female. A simple approach is to form a table. It may also help to give the population a definite size, N, say. The number of females aged 80 or over is therefore $0.7 \times 0.1N = 0.07N$. The remainder of the table is filled by subtraction.

	<80	580	
Males	$0.47N$	$0.03N$	$0.50N$
Females	$0.43N$	$0.07N$	$0.50N$
	$0.90N$	$0.10N$	N

The proportion of females amongst those aged under 80 is therefore $\frac{0.43N}{0.90N} = \frac{43}{90}$. So $P(A|B') = \frac{43}{90}$, which is just less than $\frac{1}{2}$.

Exercises 4g

1 Given that $P(A) = 0.4$, $P(B) = 0.7$, $P(A \cap B) = 0.2$, find (i) $P(A|B)$, (ii) $P(A'|B)$, (iii) $P(A|B')$, (iv) $P(A'|B')$.

2 Given that $P(A) = 0.8$, $P(A|B) = 0.8$, $P(A \cap B) = 0.5$, find (i) $P(B)$, (ii) $P(B|A)$, (iii) $P(A \cup B)$, (iv) $P(A|A \cup B)$, (v) $P(A \cap B|A \cup B)$, (vi) $P(A \cap B|B')$, (vii) $P(A \cap B|A)$.

3 Given that $P(C \cap D) = \frac{1}{4}$, $P(C|D) = \frac{1}{3}$, $P(D|C) = \frac{3}{5}$, find (i) $P(C)$, (ii) $P(D)$, (iii) $P(C|D')$, (iv) $P(C|C \cup D)$.

4 Given that $P(A) = 0.8$, $P(B) = 0.7$, $P(C) = 0.6$, $P(A|B) = 0.8$, $P(C|B) = 0.7$, $P(A \cap C) = 0.48$, determine whether:
(i) A and B are independent,
(ii) A and C are independent,
(iii) B and C are independent.

5 Given that C and D are independent and that $P(C|D) = \frac{2}{3}$, $P(C \cap D) = \frac{1}{3}$, find (i) $P(C)$, (ii) $P(D)$.

6 Given that $P(B) = \frac{4}{5}$, $P(C) = \frac{2}{3}$, $P(A|B) = \frac{1}{2}$, $P(B|A) = \frac{3}{4}$, $P(C|A \cap B) = \frac{1}{3}$, find (i) $P(A \cap B)$, (ii) $P(A \cap B \cap C)$, (iii) $P(A)$, (iv) $P(A \cap B|C)$.

7 Three ordinary unbiased six-sided dice, one red, one green and one blue, are thrown simultaneously. Events R, G, S and T are defined as follows:
R: The score on the red die is 3.
G: The score on the green die is 2.
S: The sum of the scores on the red and the green dice is 4.
T: The total score for the three dice is 5.
Find the following probabilities:
(a) $P(R \cap G)$, (b) $P(S|R)$, (c) $P(R|S)$, (d) $P(R \cup G)$, (e) $P(T)$, (f) $P(S|T)$.

8 On the sunny tropical island of Utopia, one quarter of the large number of adult inhabitants are male and the remainder are female. The island's tourist welcoming committee consists of six individuals drawn at random from the adult inhabitants of the island.
Determine the probability that:
(a) exactly one committee member is male,
(b) all the committee members are female,
(c) at least five committee members are female,
(d) all the committee members are of the same sex,
(e) all the committee members are female, given that it is known that they are all of the same sex.

9 A box contains 5 red balls and 3 white balls. A second box contains 4 red balls and 4 white balls. Two balls are drawn at random from the first box and placed in the second box. One ball is then drawn at random from the 10 balls now in the second box.
Determine the probability that this ball is red.

10 The Green Hand gang used to consist of 12 individuals, of whom 8 were called Smith and 4 were called Jones. One bad year, they fell foul of a rival gang and every month one member of the Green Hand gang was 'eliminated' at random. Determine the probability of each of the following events:
A: Exactly three of the first five eliminated were named Jones.
B: The last two to be eliminated were named Smith.
Determine also $P(A|B)$ and $P(B|A)$.

11 The events A and B are such that $P(A) = \frac{2}{5}$, $P(B) = \frac{1}{6}$ and $P(A \cup B) = \frac{11}{30}$. Show that A and B are neither mutually exclusive nor independent. [WJEC]

12 A box contains ten objects of which 1 is a red ball, 2 are white balls, 3 are red cubes and 4 are white cubes. Three objects are drawn at random from the box, in succession and without replacement. Events B and R are defined as follows:
B: Exactly two of the objects drawn are balls.
R: Exactly one of the objects drawn is red.
Show that $P(B) = \frac{7}{40}$ and calculate $P(R)$, $P(B \cap R)$, $P(B \cup R)$ and $P(B|R)$. [UCLES]

13 A box contains 25 apples, of which 20 are red and 5 are green. Of the red apples, 3 contain maggots and of the green apples, 1 contains maggots. Two apples are chosen at random from the box. Find, in any order,
(i) the probability that both apples contain maggots,
(ii) the probability that both apples are red and at least one contains maggots,
(iii) the probability that at least one apple contains maggots, given that both apples are red,
(iv) the probability that both apples are red given that at least one apple is red. [UCLES]

14 A golfer observes that, when playing a particular hole at his local course, he hits a straight drive on 80 per cent of the occasions when the weather is not windy but only on 30 per cent of the occasions when the weather is windy. Local records suggest that the weather is windy on 55 per cent of all days.

(i) Show that the probability that, on a randomly chosen day, the golfer will hit a straight drive at the hole is 0.525.

(ii) Given that he fails to hit a straight drive at the hole, calculate the probability that the weather is windy. [JMB]

15 The events A and B are such that
$P(A') = \frac{1}{4}$,
$P(A|B) = \frac{1}{3}$,
$P(A \cup B) = \frac{2}{3}$,
where A' denotes the event "A does not occur".
Find (i) $P(A)$, (ii) $P(A \cap B)$, (iii) $P(B)$, (iv) $P(A|B')$,
where B' denotes the event "B does not occur".
Determine whether A and B are independent.
[Answers may be given as fractions in their lowest terms.] [O&C]

16 A game is played with an ordinary six-sided die. A player throws this die, and if the result is 2, 3, 4 or 5, that result is the player's score. If the result is 1 or 6, the player throws the die a second time and the sum of the two numbers resulting from both throws is the player's score. Events A and B are defined as follows:

A: the player's score is 5, 6, 7, 8 or 9;
B: the player has two throws.

Show that $P(A) = \frac{1}{3}$.

Find (i) $P(A \cap B)$, (ii) $P(A \cup B)$, (iii) $P(A|B)$, (iv) $P(B|A')$. [UCLES]

17 A bag contains 4 red counters and 6 green counters. Four counters are drawn at random from the bag, without replacement. Calculate the probability that

(i) all the counters drawn are green,

(ii) at least one counter of each colour is drawn,

(iii) at least two green counters are drawn,

(iv) at least two green counters are drawn, given that at least one of each colour is drawn.

State with a reason whether or not the events 'at least two green counters are drawn' and 'at least one counter of each colour is drawn' are independent. [UCLES]

18 A bag contains 5 white balls and 3 red balls. Two players, A and B, take turns at drawing one ball from the bag at random, and balls drawn are not replaced. The player who first gets two red balls is the winner, and the drawing stops as soon as either player has drawn two red balls. Player A draws first. Find the probability

(i) that player A is the winner on his second draw,

(ii) that player A is the winner, given that the winning player wins on his second draw,

(iii) that neither player has won after two draws, given that A draws a red ball on his first draw. [UCLES]

19 For married couples the probability that the husband has passed his driving test is $\frac{7}{10}$ and the probability that the wife has passed her driving test is $\frac{1}{2}$. The probability that the husband has passed, given that the wife has passed, is $\frac{14}{15}$. Find the probability that, for a randomly chosen married couple, the driving test will have been passed by

(*a*) both of them,

(*b*) only one of them,

(*c*) neither of them.

If two married couples are chosen at random, find the probability that only one of the husbands and only one of the wives will have passed the driving test. [ULSEB]

20 Write down an expression involving probabilities for $P(B|A)$, the probability of event B given that event A occurs.
Alison and Brenda play a tennis match in which the first player to win two sets wins the match. In tennis no set can be drawn. The probability that Alison wins the first set is $\frac{1}{3}$; for sets after the first, the probability that Alison wins the set is $\frac{3}{5}$ if she won the preceding set, but is only $\frac{1}{4}$ if she lost the preceding set.
With the aid of a suitable diagram, or otherwise, determine the probability that

(i) the match lasts for just two sets,

(ii) Alison wins the match given that it lasts for just two sets,

(iii) Alison wins the match,

(iv) Alison wins the match given that it goes to three sets,

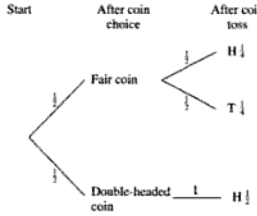
(v) if Alison wins the match, then she does so in two sets. [JMB(P)]

4.20 The total probability theorem

A simple example is provided by the following problem (see Example 24):

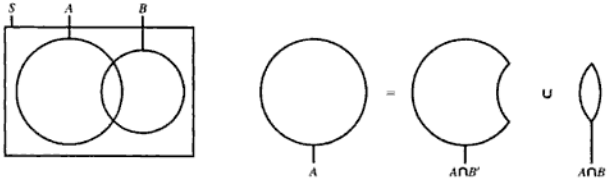
> A statistician has a fair coin and a double-headed coin. She chooses one of the coins at random and tosses it. Determine the probability that she obtains a head.

We can illustrate this situation with a probability tree:



The total probability that she obtains a head is $\frac{1}{4}$, the sum of the two branches of the tree that end with the outcome 'Head'. (This probability could also be deduced by noting that the two coins have four sides between them and that three of the four equally likely sides are heads.)

The total probability theorem is equivalent to the statement that the whole is the sum of its parts. A simple illustration of the general idea is provided by the following.



Translating the diagram into probability statements that use the fact that $A \cap B$ and $A \cap B'$ are mutually exclusive, we have

$$\begin{aligned} \mathrm{P}(A) &= \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B') \\ &= \{\mathrm{P}(B) \times \mathrm{P}(A|B)\} + \{\mathrm{P}(B') \times \mathrm{P}(A|B')\}. \end{aligned}$$

In this case A consists of just two 'slices', $A \cap B$ and $A \cap B'$. The result generalises easily to m 'slices' as follows. Suppose that $B_1, B_2, \ldots, B_m$ are m mutually exclusive and exhaustive events in the sample space S. Let A be some other event in S. Then

$$\mathrm{P}(A) = \sum_{i=1}^{m} \mathrm{P}(A \cap B_i) \tag{4.14}$$

and hence, using Equation (4.12), we get the total probability theorem that, for mutually exclusive and exhaustive events $B_1, B_2, \ldots, B_m$,

$$\mathrm{P}(A) = \sum_{i=1}^{m} \mathrm{P}(B_i) \times \mathrm{P}(A|B_i) \tag{4.15}$$

Example 31

Of those students who do well in Physics, 80% also do well in Mathematics. Of those who do not do well in Physics, only 30% do well in Mathematics. If 40% do well in Physics, what proportion do well in Mathematics?

Define the events A, B_1 and B_2 as follows:

A: Does well in Mathematics.
B_1: Does well in Physics.
B_2: Does not do well in Physics.

The information given tells us that $P(A|B_1) = 0.8$, $P(A|B_2) = 0.3$ and $P(B_1) = 0.4$. From the latter we can deduce that $P(B_2) = 0.6$. The events B_1 and B_2 are mutually exclusive and exhaustive, so, using Equation (4.15):

$$\begin{aligned} P(A) &= \{P(B_1) \times P(A|B_1)\} + \{P(B_2) \times P(A|B_2)\} \\ &= (0.4 \times 0.8) + (0.6 \times 0.3) \\ &= 0.50 \end{aligned}$$

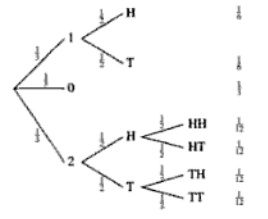
Thus half the students do well in Mathematics.

Example 32

Here is an example involving both balls being drawn from a box and coins being tossed! Suppose that a box contains 3 balls numbered, respectively, 0, 1 and 2. A ball is drawn at random from the box and is found to have the number n, say. We now toss n coins.
What is the probability that we get exactly one head?

We begin by drawing a probability tree and we define events:

A: Exactly one head is obtained.
B_i: The ball chosen is numbered i, where $i = 0, 1$, or 2.



As the diagram shows, $P(B_0) = P(B_1) = P(B_2) = \frac{1}{3}$, while $P(A|B_0) = 0$, $P(A|B_1) = \frac{1}{2}$ and $P(A|B_2) = (\frac{1}{2} \times \frac{1}{2}) + (\frac{1}{2} \times \frac{1}{2}) = \frac{1}{2}$.

The total probability of the event A is given by:

$$\begin{aligned} P(A) &= \{P(B_0) \times P(A|B_0)\} + \{P(B_1) \times P(A|B_1)\} + \{P(B_2) \times P(A|B_2)\} \\ &= (\tfrac{1}{3} \times 0) + (\tfrac{1}{3} \times \tfrac{1}{2}) + (\tfrac{1}{3} \times \tfrac{1}{2}) \\ &= 0 + \tfrac{1}{6} + \tfrac{1}{6} \\ &= \tfrac{1}{3} \end{aligned}$$

The probability that we get exactly one head is $\frac{1}{3}$.

Example 33

A car is made in three versions: 2-door, 4-door and hatchback. The proportions of the three types made are 25%, 40% and 35% respectively. Each version of the car has either a 1400 cc engine or a 1600 cc engine. Of the 2-door version, 70% have 1400 cc engines. The proportions for the 4-door and hatchback versions are 40% and 35% respectively.

In a publicity stunt the car makers choose an owner at random to receive a prize of free car-servicing for the lifetime of the car. Determine the probability that the owner's car has a 1600 cc engine.

Define the events A, B_1, B_2 and B_3 as follows:

A: Owner's car has a 1600 cc engine.
B_1: Owner's car is the 2-door version.
B_2: Owner's car is the 4-door version.
B_3: Owner's car is the hatchback version.

The events B_1, B_2 and B_3 are mutually exclusive and exhaustive, so using Equation (4.15):

$$\begin{aligned} P(A) &= \{P(B_1) \times P(A|B_1)\} + \{P(B_2) \times P(A|B_2)\} + \{P(B_3) \times P(A|B_3)\} \\ &= (0.25 \times 0.3) + (0.4 \times 0.6) + (0.35 \times 0.65) \\ &= 0.5425 \end{aligned}$$

The probability that the owner's car has a 1600 cc engine is approximately 54%.

Exercises 4h

1 A vehicle insurance company classifies drivers as A, B or C according to whether or not they are a good risk, a medium risk or a poor risk with regard to having an accident. The company estimates that A constitutes 30% of drivers that are insured and B constitutes 50%. The probability that a class A driver will have one or more accidents in any 12-month period is 0.01, the corresponding values for B and C being 0.03 and 0.06 respectively.

(a) Find the probability that a motorist, chosen at random, is assessed as a class C risk and will have one or more accidents in a 12-month period.

(b) Find the probability that a motorist, chosen at random, will have one or more accidents in a 12-month period.

(c) The company sells a policy to a customer and within 12 months the customer has an accident. Find the probability that the customer is a class C risk.

(d) If a policy holder goes 10 years without an accident and accidents in each year are independent of those in other years, show that the probabilities that the policy holder belongs to each of the classes can be expressed, to 2 decimal places, in the ratio 2.71 : 3.69 : 1.08. [AEB 90]

2 The events A and B are such that $P(A) = x + 0.2$, $P(B) = 2x + 0.1$, $P(A \cap B) = x$.

(*a*) Given that $P(A \cup B) = 0.7$, find the value of x and state the values of $P(A)$ and $P(B)$.

(*b*) Verify that the events A and B are independent.

The events A and C are mutually exclusive, $P(A \cup B \cup C) = 1$ and $P(B|C) = 0.4$.

(*c*) Find the values of $P(B \cap C)$ and $P(C)$.

(*d*) Giving a reason, state whether or not the events B and C are independent. [ULSEB]

3 For the two events A and B, $P(A|B) = \frac{5}{11}$, $P(A \cup B) = \frac{9}{10}$, $P(B) = x$.

(*a*) Write $P(A \cap B)$ in terms of x and hence show that $P(A) = \frac{9}{10} - \frac{6x}{11}$.

It is given that $P(A \cap B) = 2P(A \cap B')$.

(*b*) Find an equation for x.

(*c*) Deduce that $x = \frac{11}{15}$.

For the two events A and B and a third event C,

$P(A \cup B \cup C) = 1$,

A and C are mutually exclusive,

B and C are independent.

(*d*) Taking $P(B \cap C) = y$, form an equation for y and hence show that $P(C) = \frac{3}{8}$.

(*e*) Find the value of $P(A \cup C)$. [ULSEB]

4 (i) Events A and B are such that $P(A) = \frac{2}{5}$, $P(B) = \frac{1}{4}$ and $P(A \cup B) = \frac{11}{20}$.

Determine whether or not the events A and B are

(*a*) independent,

(*b*) mutually exclusive.

A third event C is such that $P(A \cup C) = \frac{7}{10}$, $P(B \cup C) = \frac{3}{4}$ and $P(A \cap C) = 2P(B \cap C)$.

(*c*) Find $P(C)$ and determine whether or not the events B and C are independent.

(ii) A biased die is constructed so that each of the numbers 3 and 4 is twice as likely to occur as the numbers 1, 2, 5 and 6. Find

(*a*) the probability of throwing a 4,

(*b*) the probability of throwing a 4, given that the throw is greater than 2.

Two such dice are thrown.

(*c*) Find the probability that the sum of the numbers thrown is 7. [ULSEB]

5 (i) The events A and B are such that $P(A|B) = \frac{7}{10}$, $P(B|A) = \frac{7}{15}$ and $P(A \cup B) = \frac{3}{5}$. Find the values of

(*a*) $P(A \cap B)$,

(*b*) $P(A' \cap B)$.

(ii) A hand of four cards is to be drawn without replacement and at random from a pack of fifty two playing cards. Giving your answer in each case to three significant figures, find the probabilities that this hand will contain

(*a*) four cards of the same suit,

(*b*) either two aces and two kings *OR* two aces and two queens. [ULSEB]

6 (a) Use the fact that $P(A) = P(A \cap B) + P(A \cap B')$ to show that $P(A'|B) = 1 - P(A|B)$.

(b) It is given that events A and B have non-zero probabilities, and that $P(A|B) = P(A)$.

(i) Show that $P(B|A) = P(B)$.

(ii) Use the result in (a) to show that $P(A'|B) = P(A')$.

(iii) Given also that $P(B) \neq 1$, show also that $P(A|B') = P(A)$ and $P(A'|B') = P(A')$.

The Reverend Thomas Bayes (1701–61) was a Nonconformist minister in Tunbridge Wells, Kent. He was elected a Fellow of the Royal Society in 1742. The theorem (described in the next Section) that bears his name has led to the development of an approach to statistics that runs parallel to much of the material in later chapters of this book. This approach is referred to as 'Bayesian Statistics' and its advocates are referred to as 'Bayesians'. Ironically, the theorem was contained in an essay that did not appear until after his death and was largely ignored at the time.

4.21 Bayes' theorem

In introducing the idea of conditional probability we effectively asked the question:

> Given that event B has occurred in the past, what is the probability that event A will occur?

We now consider the following 'reverse' question.

> Given that the event A has just occurred, what is the probability that it was preceded by the event B?

As an example, consider the following problem.

> A statistician has a fair coin and a double-headed coin. She chooses one of the coins at random and tosses it. She obtains a head. Determine the probability that the coin that she tossed was double-headed.

We have looked at this situation before. We found that the total probability of a head was made up of a contribution of $\frac{1}{2} \times 1$ from the double-headed coin and $\frac{1}{2} \times \frac{1}{2}$ from the fair coin, giving a total probability of $\frac{3}{4}$. Two-thirds of this total is associated with the selection of the double-headed coin (because $\frac{1}{2} \times 1$ equals $\frac{1}{2}$, which is two-thirds of $\frac{3}{4}$). Expressing this in different words, on two-thirds of the occasions that a head is obtained the double-headed coin has been tossed. The required probability is therefore $\frac{2}{3}$.

If you found the last paragraph difficult to follow, fear not! We now develop a general result, beginning with a restatement of Equation (4.12):

$$\mathrm{P}(A) \times \mathrm{P}(B|A) = \mathrm{P}(B) \times \mathrm{P}(A|B)$$

Dividing through by $\mathrm{P}(A)$ we get:

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(B) \times \mathrm{P}(A|B)}{\mathrm{P}(A)} \tag{4.16}$$

Suppose that, instead of a single event, B, there were m alternative previous events that could have happened, namely, $B_1, B_2, \ldots, B_m$. Assume that, as was the case with the total probability theorem, these events are mutually exclusive and exhaustive. From Equation (4.16):

$$\mathrm{P}(B_j|A) = \frac{\mathrm{P}(B_j) \times \mathrm{P}(A|B_j)}{\mathrm{P}(A)}$$

and, on substituting for $\mathrm{P}(A)$ using Equation (4.15), we get **Bayes' theorem**:

$$\mathrm{P}(B_j|A) = \frac{\mathrm{P}(B_j) \times \mathrm{P}(A|B_j)}{\sum_{i=1}^{m} \{\mathrm{P}(B_i) \times \mathrm{P}(A|B_i)\}} \tag{4.17}$$

You may not believe it, but this is not as bad as it looks – the denominator is, after all, simply $\mathrm{P}(A)$.

Note

- In Equation (4.17) it should be noted that the numerator, $\mathrm{P}(B_j) \times \mathrm{P}(A|B_j)$, is one of the terms in the sum in the denominator, $\sum_{i=1}^{m} \{\mathrm{P}(B_i) \times \mathrm{P}(A|B_i)\}$.

Example 34

A statistician has a fair coin and a double-headed coin. She chooses one of the coins at random and tosses it. She obtains a head. Determine the probability that the coin that she tossed was double-headed.

This is the problem that we answered rather long-windedly at the beginning of this section! We now use a formal approach using Bayes' theorem. We define the events A, B_1 and B_2 as follows:

A: A head is obtained.
B_1: The fair coin is chosen.
B_2: The double-headed coin is chosen.

We want $\mathrm{P}(B_2|A)$ and we know the following probabilities: $\mathrm{P}(B_1) = \frac{1}{2}$, $\mathrm{P}(B_2) = \frac{1}{2}$, $\mathrm{P}(A|B_1) = \frac{1}{2}$, $\mathrm{P}(A|B_2) = 1$. Using Bayes' theorem we have:

$$\mathrm{P}(B_2|A) = \frac{\mathrm{P}(B_2) \times \mathrm{P}(A|B_2)}{\mathrm{P}(B_1) \times \mathrm{P}(A|B_1) + \mathrm{P}(B_2) \times \mathrm{P}(A|B_2)}$$
$$= \frac{\frac{1}{2} \times 1}{(\frac{1}{2} \times \frac{1}{2}) + (\frac{1}{2} \times 1)} = \frac{\frac{1}{2}}{\frac{1}{4} + \frac{1}{2}} = \frac{2}{3}$$

The good thing about Bayes' theorem is that (once the events have been carefully defined!) we do not need to think!

Example 35

According to a firm's internal survey, of those employees living more than 2 miles from work, 90% travel to work by car. Of the remaining employees, only 50% travel to work by car. It is known that 75% of employees live more than 2 miles from work.
Determine:
(i) the overall proportion of employees who travel to work by car,
(ii) the probability that an employee who travels to work by car lives more than 2 miles from work.

Define the events A, B_1 and B_2 as follows:

A: Travels to work by car.
B_1: Lives more than 2 miles from work.
B_2: Lives not more than 2 miles from work.

The events B_1 and B_2 are mutually exclusive and exhaustive, with $\mathrm{P}(B_1) = 0.75$, $\mathrm{P}(B_2) = 0.25$, $\mathrm{P}(A|B_1) = 0.9$ and $\mathrm{P}(A|B_2) = 0.5$.

(i) From the total probability theorem:

$$\mathrm{P}(A) = \{\mathrm{P}(B_1) \times \mathrm{P}(A|B_1)\} + \{\mathrm{P}(B_2) \times \mathrm{P}(A|B_2)\}$$
$$= (0.75 \times 0.9) + (0.25 \times 0.5) = 0.675 + 0.125 = 0.8$$

so 80% of employees travel to work by car.

(ii) From Bayes' theorem:

$$\mathrm{P}(B_1|A) = \frac{\mathrm{P}(B_1) \times \mathrm{P}(A|B_1)}{\mathrm{P}(A)} = \frac{0.75 \times 0.9}{0.8} = 0.843\,75$$

so the probability that an employee, who travels to work by car, lives more than 2 miles from work, is about 0.84 (to 2 d.p.).

An alternative approach involves constructing the following table from the information in the question:

	More than 2 miles	Not more than 2 miles	Total
Travels by car	67.5	12.5	80.0
Does not travel by car	7.5	12.5	20.0
Total	75.0	25.0	100.0

The entries are percentages of the workforce. The first entry, 67.5%, is obtained by calculating the value corresponding to 90% of the 75% who live more than 2 miles from work (using $0.90 \times 0.75 = 0.675$).

(i) The answer is the first row total, 80%.

(ii) The answer is the proportion of the first row that are contained in the top left cell of the table, namely $\frac{67.5}{80} = 0.84$ (to 2 d.p.).

Example 36

A box contains three coins. Two coins are fair, but the third coin is double-headed. A coin is chosen at random and tossed.

(i) Determine the probability that a head is obtained.

(ii) If a head is obtained, determine the probability that it was the double-headed coin that was tossed.

We provide three alternative answers. One uses the formality of Bayes' theorem, one uses a probability tree and the last uses a 'common-sense' approach. The first is the recommended answer!

Define the events A, B_1 and B_2 as follows:

A: A head is obtained.
B_1: A fair coin was tossed.
B_2: The double-headed coin was tossed.

The events B_1 and B_2 are mutually exclusive and exhaustive, with $\mathrm{P}(B_1) = \frac{2}{3}$ and $\mathrm{P}(B_2) = \frac{1}{3}$. Also $\mathrm{P}(A|B_1) = \frac{1}{2}$ and $\mathrm{P}(A|B_2) = 1$.

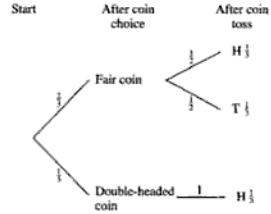
(i) $$\mathrm{P}(A) = \{\mathrm{P}(B_1) \times \mathrm{P}(A|B_1)\} + \{\mathrm{P}(B_2) \times \mathrm{P}(A|B_2)\}$$
$$= \left(\tfrac{2}{3} \times \tfrac{1}{2}\right) + \left(\tfrac{1}{3} \times 1\right) = \tfrac{1}{3} + \tfrac{1}{3} = \tfrac{2}{3}$$

The probability of obtaining a head is $\frac{2}{3}$.

(ii) $$\mathrm{P}(B_2|A) = \frac{\mathrm{P}(B_2) \times \mathrm{P}(A|B_2)}{\mathrm{P}(A)} = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}$$

Given that a head is obtained, the probability that it was the double-headed coin that was tossed is $\frac{1}{2}$.

We can see the various possibilities quite easily using a probability tree.



In this case there are three alternative outcomes, all with probability $\frac{1}{3}$. Two correspond to getting a head, hence P(Head) $= \frac{2}{3}$. Of these two outcomes one corresponds to the case where the double-headed coin was tossed and hence the second of the required probabilities is $\frac{1}{2}$.

An alternative argument is as follows. The three coins have six sides between them. The side actually seen is equally likely to be any of the six. Since four of the sides are heads, the probability of obtaining a head is $\frac{4}{6} = \frac{2}{3}$. Since two of the four heads are on the double-sided coin, the probability that it was this coin that was tossed is $\frac{2}{4} = \frac{1}{2}$. This type of argument is perfectly acceptable *when it is correct*! However, it is easy to go wrong – it is safer to follow the formulae!

Exercises 4i

1 It is given that B_1 and B_2 are mutually exclusive and exhaustive, and $P(A|B_1) = 0.3$, $P(A|B_2) = 0.4$, $P(B_1) = 0.4$.
Find (i) $P(B_1|A)$, (ii) $P(B_2|A)$, (iii) $P(B_1|A')$, (iv) $P(B_2|A')$.

2 It is given that $P(A) = 0.3$, $P(B) = 0.2$, $P(C) = 0.5$, $P(A \cap B) = 0$, $P(B \cap C) = 0$, $P(C \cap A) = 0$. It is also given that $P(D|A) = 0.1$, $P(D|B) = 0.4$, $P(D|C) = 0.6$.
Find (i) $P(A|D)$, (ii) $P(A'|D)$, (iii) $P(A|D')$, (iv) $P(A'|D')$.

3 A bag contains 7 white balls and 3 black balls. A white box contains 5 green balls and 2 red balls. A black box contains 3 green balls and 1 red ball. A ball is taken at random from the bag, and if this ball is white a ball is taken at random from the white box. If it is black a ball is taken at random from the black box. Given that the ball taken from the box is red, determine the probability that the box is coloured white.

4 A factory has three machines making large numbers of components. 10% of the components made by machine I are faulty. The corresponding figures for machines II and III are 5% and 1% respectively. The proportions of the total output produced by machines I, II and III are 50%, 30% and 20% respectively.
(i) A randomly chosen component is found to be faulty.
Find the probability that it was made by machine I.
Find also the probability that it was not made by machine II.
(ii) A randomly chosen component is found not to be faulty.
Find the probability that it was made by machine I.

5 Suppose that on one-third of the days of the year some rain falls on my garden. Suppose also that when it rains there is a probability of 0.7 that my barometer will be indicating rain, but when it does not rain there is a probability of 0.1 that my barometer nevertheless indicates rain.

(continued)

(a) Determine the probability that, on a randomly chosen day of the year, my barometer indicates rain.
(b) Given that my barometer is indicating rain, determine the probability that it is actually raining.

6 In an examination, the probabilities of three candidates, Aloysius, Bertie and Claude, solving a certain problem are $\frac{4}{5}$, $\frac{3}{4}$ and $\frac{2}{3}$, respectively. Calculate the probability that the examiner will receive from these candidates:
(a) one, and only one, correct solution,
(b) not more than one correct solution,
(c) at least one correct solution.
Given that the examiner receives exactly one correct solution, determine the probability that this solution was provided by Bertie.

7 A test for a particular disease has the following characteristics. If someone has the disease the probability of a positive test is 90%, and the probability of a negative result (a 'false negative') is 10%. If someone does not have the disease the probability of a positive test (a 'false positive') is 20%, and the probability of a negative result is 80%. The proportion of the population that has the disease is denoted by p. A person is chosen at random from the population and tested. Given that the result of the test is positive, find, in terms of p, the probability P that the person has the disease.
Verify that when $p = 0.05$ the value of P is a little less than 20%.
Sketch the graph of P against p and comment on the results.

8 Four machines A, B, C and D produce respectively 30%, 30%, 15% and 25% of the total number of items from a factory. The percentages of defective output of these machines are 1%, 1.5%, 3% and 2% respectively. Given that an item is to be selected at random from the total output, find the probability that the item will be defective. An item is selected at random and is found to be defective. Find the probability that the item was produced by machine A.
[ULSEB(P)]

9 In a simple model of the weather in October, each day is classified as either fine or rainy. The probability that a fine day is followed by a fine day is 0.8. The probability that a rainy day is followed by a fine day is 0.4. The probability that 1 October is fine is 0.75.
(i) Find the probability that 2 October is fine and the probability that 3 October is fine.
(ii) Find the conditional probability that 3 October is rainy, given that 1 October is fine.
(iii) Find the conditional probability that 1 October is fine, given that 3 October is rainy. [UCLES]

10 (a) Give an equation involving probabilities which is equivalent to the statement
(i) the events L and M are *mutually exclusive*,
(ii) the events L, M and N are *exhaustive*.
If the events L, M and N are mutually exclusive as well as being exhaustive, write down an equation relating $P(L)$, $P(M)$ and $P(N)$.

(b) A city Passenger Transport Executive (PTE) carries out a survey of the commuting habits of its city centre workers. The PTE discovers that 40% of commuters travel by bus, 25% travel by train and the remainder use private vehicles.
Of those who travel by bus, 60% have a journey of less than 5 miles and 30% have a journey of between 5 and 10 miles.
Of those who travel by train, 30% have a journey of between 5 and 10 miles and 60% have a journey of more than 10 miles.
Of those who use private vehicles, 20% travel less than 5 miles, with the same percentage travelling more than 10 miles.
By organising the above information in a suitable table or diagram, or otherwise, determine the probability that a commuter chosen at random
(i) travels by bus for a journey of less than 5 miles,
(ii) has a journey of more than 10 miles,
(iii) travels by bus or has a journey of more than 10 miles,
(iv) uses a private vehicle given that the commuter travels between 5 and 10 miles. [JMB(P)]

11 State in words the relationship between two events E and F when

(*a*) $P(E \cap F) = P(E).P(F)$,
(*b*) $P(E \cap F) = 0$.

Given that $P(E) = \frac{1}{3}$, $P(F) = \frac{1}{2}$, $P(E' \cap F) = \frac{1}{2}$, find

(*c*) the relationship between E and F,
(*d*) the value of $P(E|F)$,
(*e*) the value of $P(E' \cap F')$.

A boy always either walks to school or goes by bus. If one day he goes to school by bus, then the probability that he goes by bus the next day is $\frac{7}{10}$. If one day he walks to school, then the probability that he goes by bus the next day is $\frac{2}{5}$. Given that he walks to school on a particular Tuesday, draw a tree diagram and hence find the probability that he will go to school by bus on Thursday of that week.

Given that the boy walks to school on both Tuesday and Thursday of that week, find the probability that he will also walk to school on Wednesday.

[You may assume that the boy will not be absent from school on Wednesday or Thursday of that week.] [ULSEB]

Chapter summary

- The **probability** of the event E is denoted by $P(E)$.
- The event 'E does not occur' is the **complementary event** and is denoted by E'.

  $P(E') = 1 - P(E)$
- The event 'At least one of events A or B occurs' is the **union** of events A and B and is denoted by $A \cup B$.
- The event 'Both A and B occur' is the **intersection** of events A and B and is denoted by $A \cap B$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- If events A and B are **mutually exclusive** then:

  $P(A \cap B) = 0$
  $P(A \cup B) = P(A) + P(B)$: the **addition rule**.
- If events A and B are **exhaustive** then:

  $P(A \cup B) = 1$
  $P(A) + P(B) = 1 + P(A \cap B)$
- If events A and B are **mutually exclusive and exhaustive** then:

  $P(A \cap B) = 0$
  $P(A) + P(B) = 1$
- If events A and B are **independent** then:

  $P(A \cap B) = P(A) \times P(B)$: the **multiplication rule**.

- *Orderings*
  - **Factorials** $n! = n \times (n-1) \times \cdots \times 1; \quad 0! = 1$
    The number of ways that n distinct objects can be arranged in order is $n!$
  - If a set of n objects comprises a objects of one type, b of another, etc., then

    the number of distinct orderings is $\dfrac{n!}{a!b!\ldots}$
  - When r objects are chosen from a group of n unlike objects, with ordering being important, the number of distinct **permutations** is:

$$^nP_r = n \times (n-1) \times \cdots \times (n-r+1) = \frac{n!}{(n-r)!}$$

  - When the order of drawing is unimportant, the number of possible collections (**combinations**) of r objects drawn from n is:

$$^nC_r = \binom{n}{r} = \frac{n \times (n-1) \times \cdots \times (n-r+1)}{r \times (r-1) \times \cdots \times 1} = \frac{n!}{r!(n-r)!}$$

$$\binom{n}{r} = \binom{n}{n-r}; \quad \binom{n}{0} = \binom{n}{n} = 1$$

- $P(B|A)$ denotes the probability that the event B occurs (or has occurred) given the information that the event A occurs (or has occurred).
  $P(B|A)$ is known as the **conditional probability** of B **given** A.

- $$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- $$P(A \cap B) = P(A) \times P(B|A) = P(B)P(A|B)$$

- **Statistical independence**: each of the following statements implies all the others.
  - Events A and B are statistically independent.
  - $P(A|B) = P(A)$
  - $P(B|A) = P(B)$
  - $P(A \cap B) = P(A) \times P(B)$ – the **multiplication rule**.

- **Physically independent** events are statistically independent.

- **The total probability theorem**:
  If events $B_1, B_2, \ldots, B_n$ are mutually exclusive and exhaustive, then

$$P(A) = \sum P(A \cap B_i) = \sum \{P(B_i) \times P(A|B_i)\}$$

- **Bayes' theorem**:

$$P(B_j|A) = \frac{P(B_j) \times P(A|B_j)}{\sum \{P(B_i) \times P(A|B_i)\}}$$

Exercises 4j (Miscellaneous)

1 In Ruritania all the cars are made by a single firm and vary only in their colouring. Six different colours are available (including 'communist red'). The same numbers of cars are painted in each of the six colours. Assuming that, when travelling on the roads, the colours of the cars occur in random order, determine the probability that:

(a) the first six cars to pass Rudolf are all of different colours,

(b) the second car is the same colour as the first, but the next 5 are all of different colours to their predecessors,

(c) the first two cars have different colours, the third is the same colour as one or other of the first two, and the next four cars are all of different colours to their predecessors,

(d) at least 8 cars pass Rudolf before all 6 colours are encountered.

(e) none of the first six cars to pass Rudolf were painted in 'communist red'.

2

	Small	Medium	Large
White	40	35	20
Blue	25	30	15
Cream	10	20	5

Table 1

Table 1 shows the distribution by size and colour of shirts in a batch of 200. A shirt is to be selected at random from the batch. Calculate the probability that it will be

(*a*) small,

(*b*) either blue or white.

Two shirts are to be selected at random, without replacement, from the large shirts. Calculate, to 4 decimal places, the probability that

(*c*) both shirts will be white,

(*d*) one shirt will be white and one will be cream. [ULSEB]

3 In the Upper Sixth Statistics class there are two boys and four girls, while in the Lower Sixth Statistics class there are four boys and six girls. Two different pupils are chosen at random from **each** of the two classes. Calculate the probabilities that the four chosen consist of

(i) two boys from the Upper Sixth and two girls from the Lower Sixth,

(ii) two boys and two girls. [WJEC]

4 In a computer game played by a single player, the player has to find, within a fixed time, the path through a maze shown on a computer screen. On the first occasion that a particular player plays the game, the computer shows a simple maze, and the probability that the player succeeds in finding the path in the time allowed is $\frac{3}{4}$. On subsequent occasions, the maze shown depends on the result of the previous game. If the player succeeded on the previous occasion, the next maze is harder, and the probability that the player succeeds is one half of the probability of success on the previous occasion. If the player failed on the previous occasion, a simple maze is shown and the probability of the player succeeding is again $\frac{3}{4}$. The player plays three games.

(i) Show that the probability that the player succeeds in all three games is $\frac{27}{512}$.

(ii) Find the probability that the player succeeds in exactly one of the games.

(iii) Find the probability that the player does not have two consecutive successes. [UCLES(P)]

5 The probability that a particular man will survive the next twenty-five years is 0.6, and independently, the probability that the man's wife will survive the next twenty-five years is 0.7. Calculate the probability that in twenty-five years' time

(i) only the man will be alive,

(ii) at least one will be alive. [WJEC]

6 Show that, for any two events E and F,

$$P(E \cup F) = P(E) + P(F) - P(E \cap F).$$

Express in words the meaning of $P(E|F)$. Given that E and F are independent events, express $P(E \cap F)$ in terms of $P(E)$ and $P(F)$, and show that E' and F are independent. In a college, 60 students are studying one or more of the three subjects Geography, French and English. Of these, 25 are studying Geography, 26 are studying French, 44 are studying English, 10 are studying Geography and French, 15 are studying French and English, and 16 are studying Geography and English.

(*continued*)

Write down the probability that a student chosen at random from those studying English is also studying French.
Determine whether or not the events "studying Geography" and "studying French" are independent.
A student is chosen at random from all 60 students. Find the probability that the chosen student is studying all three subjects. [ULSEB]

7 Students in a class were given two statistics problems to solve, the second of which was harder than the first. Within the class $\frac{4}{5}$ of the students got the first one correct and $\frac{7}{12}$ got the second one correct. Of those students who got the first one correct, $\frac{3}{5}$ got the second one correct. One student was chosen at random from the class.
Let A be the event that the student got the first problem correct and B be the event that the student got the second one correct.
(*a*) Express in words the meaning of $A \cap B$ and of $A \cup B$.
(*b*) Find $P(A \cap B)$ and $P(A \cup B)$.
(*c*) Given that the student got the second problem right, find the probability that the first problem was solved correctly.
(*d*) Given that the student got the second problem wrong, find the probability that the first problem was solved correctly.
(*e*) Given that the student got the first problem wrong, find the probability that the student also got the second problem wrong. [ULSEB]

8 Two porcelain factories A and B produce cheap china cups in equal numbers. If closely examined, a cup from A will be found flawless with probability $\frac{1}{4}$, but one from B with probability $\frac{1}{2}$. Jim picks up two cups from a batch in a shop. The shopkeeper says that all the cups in the batch come from the same factory, but the batch is equally likely to come from factory A or from factory B.
(i) What is the probability that the first cup Jim examines is flawless?
(ii) Given that the first cup is flawless, what is the conditional probability that the batch came from factory A?
(iii) Unfortunately, Jim drops the second cup before he can examine it. Given that the first cup was flawless, find the probability that the second cup was also flawless before the accident. [SMP]

9 A high jumper estimates the probabilities that she will be able to clear the bar at various heights, on the basis of her experience in training. These are given in the table:

Height	Probability of success at each attempt
1.60 m	1
1.65 m	0.6
1.70 m	0.2
1.75 m	0

In a competition she is allowed up to three attempts to clear the bar at each height. If she succeeds, the bar is raised by 5 cm and she is allowed three attempts at the new height; and so on. It is assumed that the result of each attempt is independent of all her previous attempts.
(i) Show that the probability that she will be successful at 1.65 m is 0.936.
(ii) Calculate the probability that, if she is successful at 1.65 m, she will not be successful at 1.70 m.
Hence find the probabilities that, in the competition, the height she jumps will be recorded as
(*a*) 1.60 m, (*b*) 1.65 m, (*c*) 1.70 m. [SMP]

10 In a sales campaign, a petrol company gives each motorist who buys their petrol a card with a picture of a film star on it. There are 10 different pictures, one each of 10 different film stars, and any motorist who collects a complete set of all 10 pictures gets a free gift. On any occasion when a motorist buys petrol, the card received is equally likely to carry any one of the 10 pictures in the set.
(i) Find the probability that the first four cards the motorist receives all carry different pictures.
(ii) Find the probability that the first four cards received result in the motorist having exactly three different pictures.
(iii) Two of the ten film stars in the set are X and Y. Find the probability that the first four cards received result in the motorist having a picture of X or Y (or both).
(iv) At a certain stage the motorist has collected nine of the ten pictures. Find the least value of n such that
P(at most n more cards are needed to complete the set) > 0.99. [UCLES]

11 The staff employed by a college are classified as academic, administrative or support. The following table shows the numbers employed in these categories and their sex.

	Male	Female
Academic	42	28
Administrative	7	13
Support	26	9

A member of staff is selected at random.
A is the event that the person selected is female.
B is the event that the person selected is academic staff.
C is the event that the person selected is administrative staff.
($\bar{A}$ is the event not A, $\bar{B}$ is the event not B, $\bar{C}$ is the event not C.)

(a) Write down the values of
(i) $P(A)$,
(ii) $P(A \cap B)$,
(iii) $P(A \cup \bar{C})$,
(iv) $P(\bar{A}|C)$.

(b) Write down one of the events which is
(i) not independent of A,
(ii) independent of A,
(iii) mutually exclusive of A.
In each case justify your answer.

(c) Given that 90% of academic staff own cars, as do 80% of administrative staff and 30% of support staff,
(i) what is the probability that a staff member selected at random owns a car?
(ii) A staff member is selected at random and found to own a car. What is the probability that this person is a member of the support staff?
[AEB 91]

12 On one of his travels, Gulliver landed on an island inhabited by equal numbers of two groups of people – the Veracians who always told the truth and the Confusians who answered questions truthfully with probability $\frac{2}{3}$, independently for each question. The Veracians and Confusians were indistinguishable with regard to features, dress, etc. One afternoon Gulliver was lost on the island and on meeting a local inhabitant asked the following two questions:
"Is it night or day?"
"Which is the way to the nearest town?"
(i) Find the probability that the answer to the first question was correct.
(ii) Given that the answer to the first question was correct, calculate the conditional probability that the answer to the second question was correct. [JMB]

13 Three children, *A*, *B* and *C*, are not good at keeping secrets. The probability that *A* will tell any secret to any other child is p. Similarly, for *B* and *C* the probabilities of telling secrets are q and r respectively. *A* knows a secret. Firstly *A* meets *B*, then *A* meets *C* and finally *B* meets *C*.
(i) Show that, after these three meetings, the probability that *B* knows the secret and *C* does not is $p(1-p)(1-q)$.
It is known that $p = \frac{1}{2}$, $q = \frac{1}{3}$, $r = \frac{1}{4}$.
(ii) Show that the probability that, after these three meetings, both *B* and *C* know the secret is $\frac{19}{28}$.
(iii) After these three meetings a fourth child *D*, who never tells secrets, first meets *B* and then meets *C*. Find the probabilities that, after these five meetings,
(*a*) *A*, *B* and *C* know the secret and *D* does not,
(*b*) all four children know the secret,
(*c*) just three of the children know the secret.
[Assume independence of events throughout. Answers may be left as fractions in their lowest terms.] [O&C]

5 Probability distributions and expectations

I am giddy: expectation whirls me round. The imaginary relish is so sweet that it enchants my sense

Troilus and Cressida, William Shakespeare.

This chapter is concerned with discrete random variables. Recall that a variable is described as being a **random variable** if its value is the result of a random observation or experiment, and that **discrete** implies that a list of its possible numerical values could be made. Here are some examples:

Discrete random variable	Possible values
The number obtained when rolling a fair six-sided die	1, 2, 3, 4, 5, 6
The number of heads obtained when four fair coins are tossed	0, 1, 2, 3, 4
The amount (in £) won in a lottery having prizes of 50p, £5 and £50	0, 0.5, 5, 50
The net gain (in £) from buying a 25p ticket in the above lottery	−0.25, 0.25, 4.75, 49.75
The number of rainy days in May	0, 1, ..., 31
The number of heads obtained when a single fair coin is tossed once	0, 1
The number of tosses of a fair coin until a head is obtained	1, 2, 3, ... (no limit!)

In each case the possible outcomes can be written down as a list of numerical values. These values do not have to be positive, nor do they have to be integers. Usually, but not always, the list is limited to just a few values.

5.1 Notation

We write:

RANDOM VARIABLES	as e.g.	X, Y, Z
observed values	as e.g.	x, y, z

This leads to a statement such as:

$$\mathrm{P}(X = x) = \tfrac{1}{4}$$

which should be read as:

The probability that the random variable X takes the value x is $\frac{1}{4}$.

We can link this statement to the probability of an event by defining the event A as 'the random variable X takes the particular value x', so $\mathrm{P}(A) = \frac{1}{4}$.

To simplify formulae we will often replace the cumbersome $\mathrm{P}(X = x)$ by the simpler P_x. The alternative $p(x)$ is sometimes used.

5.2 Probability distributions

Suppose we roll a biased die which has sides numbered 1 to 6. Define the random variable X to be 'the number showing on the top of the die'. We know two things:

1 The observed value of X must be 1, 2, 3, 4, 5 or 6.
2 On a given roll the random variable X can only take *one* of those values.

These correspond to statements that the six outcomes are both exhaustive and mutually exclusive, hence:

$$P_1 + P_2 + \cdots + P_6 = 1$$

Generalising, for a discrete random variable X that can take only the distinct values $x_1, x_2, \ldots, x_m$:

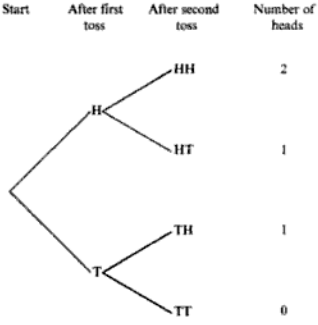
$$\sum_{i=1}^{m} P_{x_i} = 1 \tag{5.1}$$

The sizes of $P_{x_1}, P_{x_2}, \ldots,$ show how the total probability of 1 is *distributed* amongst the possible values of X. The most likely value for X will be the one with the highest probability. This is analogous to a frequency distribution, and, since the quantities are probabilities, the values $P_{x_1}, P_{x_2}, \ldots,$ are said to define a **probability distribution**.

Example 1

Tabulate the probability distribution of the number of heads obtained when a fair coin is tossed twice.

Let X be the random variable 'the number of heads obtained'. The possible values are 0, 1 and 2. The simplest way of finding the required probabilities is to use a probability tree from which we obtain the required table.



Number of heads, x	0	1	2
P_x	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Jean-le-Rond d'Alembert (1717–83) was found abandoned as a new-born infant near the church of Saint Jean-le-Rond in Paris. The gendarme who discovered the baby chose the name of the church for the baby. Despite this inauspicious beginning the boy did well! At the age of 24 he was admitted to the French Academy (the equivalent of the British Royal Society). He is best known for his work on kinetics in connection with what is now known as d'Alembert's principle. However, probabilists recall his name best because his answer to the previous example was wrong! He argued falsely that, since there were three possibilities, each probability must be $\frac{1}{3}$!

Practical

d'Alembert's error was to assume that the three possibilities were equally likely. To verify that they are not, toss two coins a total of twenty times. Draw up a tally chart of the number of heads (0, 1 or 2) obtained. Do you believe d'Alembert? If he had seen the combined results from your class, he would have spotted his error!

The probability function

For many situations it will not be necessary to make a list of all m probabilities, in order to specify the probability distribution, because some simple all-embracing formula (sometimes called the **probability function**) can be found.

Example 2

Obtain a formula for the probability distribution of the random variable X defined as 'the result of rolling a fair six-sided die'.

Each of the six possible values for X has probability $\frac{1}{6}$, so we can write:

$$P_x = \frac{1}{6} \qquad (x = 1, 2, \ldots, 6)$$

Illustrating probability distributions

As always in statistics, it is a good idea to draw pictures whenever possible. Since a discrete random variable can only take discrete values, a bar chart is appropriate, with the y-axis measuring probability.

Example 3

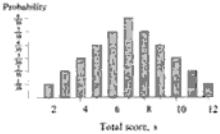
The random variable X is defined as 'the sum of the scores shown by two fair six-sided dice'.

Tabulate the probability distribution of X and draw an appropriate diagram.

We begin by drawing up a table showing the 36 possible outcomes, all of which (since the dice are fair) are equally likely:

		First die					
		1	2	3	4	5	6
Second die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Entries in table are the sums of the numbers shown by the two dice.



By inspection of the table (look along the NE-SW diagonals!) we can see that, of the 36 equally likely possibilities, there is just 1 possibility leading to the outcome $X = 2$, so $P_2 = \frac{1}{36}$. The most likely value for X is 7, which has probability $\frac{6}{36} = \frac{1}{6}$.

The full distribution is tabulated below.

Value of x	2	3	4	5	6	7	8	9	10	11	12
P_x	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Exercises 5a

In Questions **1–9**, find the set of possible values of the random variable X, and draw up a table showing P_x (where $P_x = P(X = x)$) for each value of x. The distributions will be used in Exercises 5c and 5d.

1 A box contains 3 red marbles and 5 green marbles. Two marbles are taken at random without replacement, and X is the number of green marbles obtained.

2 A box contains 3 red marbles and 5 green marbles. Two marbles are taken at random with replacement, and X is the number of green marbles obtained.

3 A fair coin has the number '1' on one face and the number '2' on the other. The coin is thrown with a fair die and X is the sum of the scores.

4 In a raffle, 20 tickets are sold and there are two prizes. One ticket number is drawn at random and the corresponding ticket earns a £10 prize. A second, different, ticket number is drawn at random, and the corresponding ticket earns a £3 prize. The prize earned by a particular one of the original 20 tickets is £X.

5 Two cards are drawn at random (without replacement) from a pack of playing cards, and X is the number of Hearts obtained.

6 A fair die is thrown and X is the reciprocal of the score (i.e. 'one over the score').

7 Two fair dice, one red and the other green, are thrown and X is the score on the red die minus the score on the green die.

8 Two fair dice, one red and the other green, are thrown and X is the positive difference in the scores, (i.e. the modulus of the random variable in the preceding example).

9 Packets of 'Hidden Gold' cornflakes are sold for £1.20 each. One in twenty of the packets contains a £1 coin. A shopper buys two packets and £X is the net cost of the two packets.

10 Which of the following experiments give a discrete random variable? (You are not asked to find any probabilities.)

(i) A book is chosen at random from a shelf with 50 books and its author noted.

(*continued*)

(ii) A book is chosen at random from a shelf with 50 books and the number of pages noted.

(iii) A book is chosen at random from a shelf with 50 books and the fifth letter on the tenth page is noted.

(iv) A pupil is chosen at random from a particular class and the pupil's name is noted.

(v) A pupil is chosen at random from a particular class and the pupil's height is recorded to the nearest inch.

(vi) The number of cars passing a given point on the road between 0800 and 0900 tomorrow.

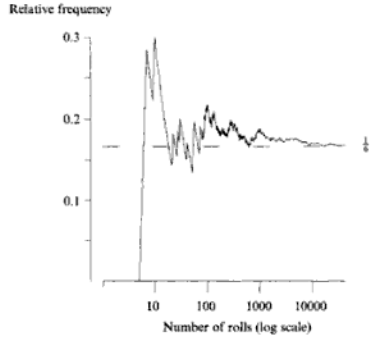
(vii) The colour of the first car to pass a given point after 0900 tomorrow.

(viii) The time, after 0900 tomorrow, recorded to the nearest second, at which the telephone first rings in the local Town Hall.

(ix) A point is chosen at random in the x–y plane and the distance from the origin is recorded to the nearest mm.

Estimating probability distributions

As we noted in Section 4.1 (p. 92), probabilities can be thought of as being the limiting values of relative frequencies. If we concentrate on a single outcome, such as obtaining a six when a die is rolled, and plot relative frequency against number of rolls, then we get a graph such as the following, which was obtained using the random number generator of a computer to simulate the rolling of a die.



Note how the 'wiggles' die away as the number of rolls increases and the relative frequency becomes increasingly close to its limiting value of $\frac{1}{6}$. A summary of the results for all six outcomes is given below:

Number of rolls	Relative frequency of 1	2	3	4	5	6
36	0.222	0.083	0.167	0.167	0.139	0.222
216	0.185	0.139	0.167	0.162	0.185	0.162
1296	0.156	0.171	0.168	0.159	0.176	0.170
7776	0.161	0.164	0.173	0.160	0.168	0.173
46656	0.168	0.164	0.167	0.168	0.165	0.168
Target	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

As the sample size (the number of rolls) increases, so the relative frequencies converge ever more closely on the theoretical probabilities, and the observed distribution of the possible outcomes converges on the theoretical probability distribution.

Computer project

Write your own computer (or calculator) program to simulate the rolling of a die. To convert a random number r having a value between 0 and 1 into the score d on a die, set $d = 1 + INT(6 * r)$, *where INT is a function that truncates a decimal to an integer (e.g.* $INT(5.8) = 5$*). Examine the relative frequencies as you increase the sample size – if they are all converging on* $\frac{1}{6}$ *then the program works!*

Practical

This practical needs a little care, since it involves drawing pins! Take ten(unsquashed) drawing pins and drop them on to a flat surface. Count the number of drawing pins, x, that land with their point in the air. Repeat the experiment twenty times.

Draw a bar chart of the results and determine the relative frequency of the outcome $x = 5$.

Combine your results with four other members of the class and recalculate the relative frequency.

What do you suppose is the numerical value of P_5?

Project

Car number plates provide a useful guide to the age of cars. The relationship between number plate and age is not perfect, of course, since some owners have 'cherished' number plates that they transfer from old cars to new ones. Furthermore, two cars with number plates of the same 'age' can differ in age by as much as 365 days (in a leap year!). Nevertheless, a rough indication of the age distribution of cars can be obtained. A convenient way of avoiding most of the problems of deciding on the age of a car is to define the random variable X to be 'the age of the car in completed years as indicated by the registration'. Thus cars registered in the current year correspond to $X = 0$.

A number of interesting questions now arise! Is it the case that the age distribution of the cars on a dual carriageway (company cars, speeding executives, etc) is the same as that for cars in the supermarket car park (shoppers using older(?) 'second' cars). Does the age distribution vary according to the time of day? This could be the case if the 'first car' leaves early for work and the 'second car' leaves later for the shops. You will be able to think of other possibilities. Several hundred observations will be required before any differences can be detected with confidence.

The cumulative distribution function

This is an alternative function for summarising a probability distribution. The function F is defined by

$$F(x_0) = P(X \leqslant x_0) = \sum_{x \leqslant x_0} P(X = x)$$

Example 4

Obtain the cumulative distribution function for the random variable X defined as 'the result of rolling a fair six-sided die'.

The following formula does the trick:

$$P(X \leqslant x) = \begin{cases} 0 & x < 1 \\ \frac{1}{6}m & m \leqslant x < (m+1);\ m = 1, 2, \ldots, 5 \\ 1 & x \geqslant 6 \end{cases}$$

since, for example:

$$P(X \leqslant 3) = P(X = 1) + P(X = 2) + P(X = 3) = \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{3}{6}$$

5.3 Some special discrete probability distributions

The two most important discrete distributions are the binomial and Poisson distributions, which will be discussed in Chapters 7 and 8. We now look at some others.

The discrete uniform distribution

Here the random variable X is equally likely to take any of k values $x_1, x_2, \ldots, x_k$, so that the distribution is summarised by:

$$P_{x_i} = c \qquad (i = 1, 2, \ldots, k)$$

where c is a constant. Can c take any value that we please? Certainly not! Equation (5.1) specified that the probabilities must sum to 1 and so c is determined by the necessity that:

$$c + c + \cdots + c = 1$$

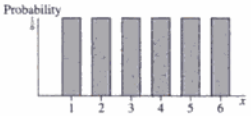
which implies that $c = \dfrac{1}{k}$. The distribution is properly specified by:

$$P_{x_i} = \frac{1}{k} \qquad (i = 1, 2, \ldots, k)$$

Example 5

The most familiar example occurs when X is defined as 'the score obtained when a fair six-sided die is rolled'. In this case $k = 6$. The distribution is tabulated below:

Value of X	1	2	3	4	5	6
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$



James Bernoulli (1654–1705) was a member of an extremely talented Swiss family. The most famous family members were James, his brother John and his nephews Nicholas and Daniel – though there were seven Bernoullis who would deserve a mention in a mathematician's *Who's Who!* James was 21 when he graduated (in Theology) from the University of Basel. He returned to the university as a lecturer (in Physics) when he was 29 and became a Professor of Mathematics at the age of 33. His principal work, *Ars Conjectandi* (The Art of Conjecture), was a treatise on probability.

The Bernoulli distribution

The Bernoulli distribution is very simple! It refers to a random variable X that can take only the values 0 and 1:

$$P_0 = 1 - p \qquad P_1 = p$$

An example of the random variable X is 'the number of heads obtained on a single toss of a bent coin', where the probability of a head is p. The importance of this simple distribution will become apparent in Chapter 7.

5.4 The geometric distribution

We can again use coin-tossing as an illustration. Suppose that we have a bent penny with P(Head) $= p$ and P(Tail) $= 1 - p$, with $0 < p < 1$. This time we embark on a succession of tosses and define the random variable X to be the number of tosses up to and including the first head (a 'Success').

Evidently $P_1 = p$, since this is the probability of an immediate head. For X to be equal to 2 we must obtain a tail on the first toss and a head on the second toss. Thus:

$$\begin{aligned} P_2 &= \text{P(Tail then Head)} \\ &= \text{P(Tail)} \times \text{P(Head)} \qquad \text{physically independent events} \\ &= (1-p)p \end{aligned}$$

Similarly, for X to be equal to x, we must obtain a sequence of $(x - 1)$ tails followed by a head. Each tail occurs with probability $(1 - p)$ so that we get the general result:

$$P_x = (1-p)^{x-1}p \qquad (x = 1, 2, \ldots) \tag{5.2}$$

This general result, which holds for all positive integer values of x, defines a **geometric distribution**.

For a fair penny, $p = \frac{1}{2}$. In a recent five-match test series the English test captain lost the first four tosses, but won the fifth. Using Equation (5.2) we see that the probability of his having to wait this long for a win during his next sequence of tosses is:

$$\left(1 - \tfrac{1}{2}\right)^4 \tfrac{1}{2} = \tfrac{1}{32}$$

Notes

- The distribution is called *geometric* because the successive probabilities, p, $(1-p)p$, $(1-p)^2p$, ... form a **geometric progression** with first term p and common ratio $(1 - p)$.

- Writing q for $(1-p)$, and noting that $0 < q < 1$:

$$\sum_{x=1}^{\infty} P_x = p(1 + q + q^2 + q^3 + \cdots)$$
$$= p\frac{1}{1-q} \quad \text{sum to infinity of a geometric progression}$$
$$= 1 \quad \text{since } q = 1 - p$$

This shows that the total probability being distributed is equal to 1 as required. It also proves that a success will occur eventually – if you have just had 3000 failures, don't worry! Providing $0 < p < 1$, you will get a success eventually (if you don't die of exhaustion first ...).

- The distribution is sometimes written as:

$$P_x = (1-p)^x p \qquad (x = 0, 1, 2, \ldots)$$

Notation

To save having to write: 'The random variable X has a geometric distribution with the probability of a "success" being p for each trial', we write:

$$X \sim \text{Geo}(p)$$

The symbol $\sim$ means 'has distribution' and 'Geo' is used as a shorthand for 'geometric'.

Cumulative probabilities

To calculate $P(X \leqslant x)$ we note that this means that at least one of the first x trials must have been a success. The complement to this event is that all x trials were failures. If the probability of a failure is $(1-p)$ then the probability of x failures is $(1-p)^x$. Writing q for $(1-p)$ we have

$$P(X \leqslant x) = 1 - q^x \tag{5.3}$$

Similarly:

$$P(X < x) = 1 - q^{x-1}$$
$$P(X > x) = q^x$$
$$P(X \geqslant x) = q^{x-1}$$

Note

- We can also prove the result in Equation (5.3) as follows:

$$P(X \leqslant x) = P(X = 1) + P(X = 2) + \cdots + P(X = x)$$
$$= p + pq + \cdots + pq^{x-1}$$
$$= p(1 + q + \cdots + q^{x-1})$$

The bracketed terms are a geometric series with sum $\dfrac{1-q^x}{1-q}$. Since $p = (1-q)$, this establishes the given result.

Example 6

Only 1% of the vehicles leaving a motorway are prepared to give lifts to hitch-hikers. George Nerdowell arrives at a motorway exit and sticks out his thumb. Determine the probability that at least four vehicles fail to stop for him (i.e. that he doesn't get a lift until at least vehicle 5).

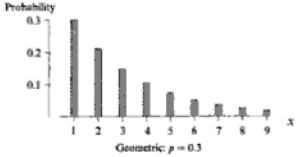
George will keep his thumb stuck out until he obtains a lift. So each vehicle is either a 'Success' (with probability, p, equal to 0.01), or a 'Failure' (with probability, q, equal to 0.99). Let X be the number of vehicles up to and including the vehicle that gives George a lift. The question requires us to calculate $P(X > 4)$:

$$P(X > 4) = q^4 = 0.99^4 = 0.961 \text{ (to 3 d.p.)}$$

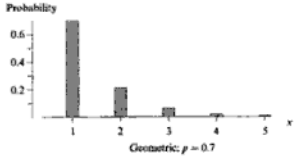
The probability that at least four vehicles fail to stop for him is about 0.96.

A paradox!

Assuming that $0 < p < 1$, all geometric distributions have a similar shape: an infinite sequence of ever smaller probabilities. The rate of decline in the size of the probabilities depends upon the value of p, but the mode (the most probable value) is at $x = 1$ in each case.



The practical consequences of this result are, to say the least, surprising! Suppose, for example, that I decide that I will stand outside my house until I see a red sports car. Clearly I may have to stand there for a long time, since red sports cars are not all that common. Consider therefore the following question: 'What is the most probable number of cars that pass my house up to and including the red sports car?'. The situation is geometric, with the value of p being rather small. Nevertheless, the previous result still holds and the answer to the question is that the most probable number of cars is just 1!



Note

- This result can easily be misinterpreted! The probability of the specific outcome 1 is $P_1 = p$, so $P(X > 1) = 1 - p$. The rarer the event of interest is (i.e. the smaller the value of p is), the more likely is the observed value to be greater than 1. Nevertheless, 1 remains the most probable single value.

Practical

Roll a normal six-sided die repeatedly until a 6 is obtained. Record the number of tosses required. Repeat a further 9 times. Pool your results with those of your neighbours, or with those for the entire class. You should find, as predicted, that the mode is at 1 – though some people may have had to roll as many as 20 times in order to get a 6!

Practical

A pack of cards is required for this exercise. Begin by shuffling the pack so that the cards are in a random order and draw a card at random from the pack. Replace the card, shuffle once again and repeat the procedure. Continue with this process until a 'Success' is obtained. The event to be called a 'Success' could be 'a Spade' ($p = \frac{1}{4}$), 'an Ace' ($p = \frac{1}{13}$) or whatever is of interest. However, it is inadvisable to choose a very rare event such as 'the Ace of Spades' ($p = \frac{1}{52}$), unless there is a great deal of time available! Record the number of cards required (x).

Repeat the entire procedure a number of times and pool your results with those of the rest of the class. Compare the class relative frequencies with the theoretical probabilities. There should be reasonable agreement, particularly for low values of x. In Chapter 13 we shall look at ways of testing how good the agreement is.

Exercises 5b

1 A book has pages numbered 1 to 300. A page is chosen at random and X is the last digit of the page number.
(i) Find the probability distribution of X.
(ii) The first digit of the page number is Y. Does Y have a discrete uniform distribution? Find the distribution of Y.

2 In Ludo it is necessary to throw a six with a single fair die in order to start. The number of throws needed to obtain the first six is N. Find the probability distribution of N. Find (i) $P(N > 6)$, (ii) $P(N \leqslant 5)$.

3 Determine the distribution of X, where X is the random variable denoting the number of sixes obtained on a single throw of a fair die.

4 The random variable X has a distribution which is both uniform and Bernoulli. Describe the distribution.

5 The random variable X is the number of heads obtained when a fair coin is thrown twice. Show that the distribution of X is (i) not uniform, (ii) not Bernoulli, (iii) not geometric.

6 Packets of 'Hidden Gold' cornflakes are sold for £1.20 each. One in twenty of the packets contains a £1 coin.
(i) A shopper goes on buying packets until a packet containing £1 is obtained. Find the probability distribution of the number of packets bought.
(ii) A shopper buys a single packet. Find the probability distribution of the number of coins obtained.

5.5 Expectations

In Chapter 2 we saw that the mean of a frequency distribution is given by:

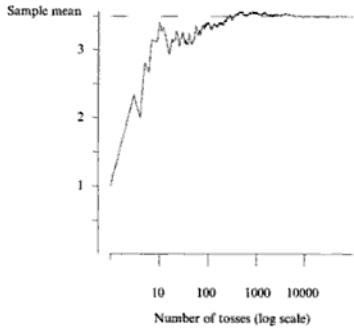
$$\bar{x} = \frac{1}{n}\Sigma f_j x_j$$

where f_j is the frequency with which the value x_j occurred, n is the total number of observations, and the summation is over all the values of x_j. The formula can be rewritten as:

$$\bar{x} = \sum \left\{ x_j \left(\frac{f_j}{n} \right) \right\}$$

which emphasises that, in the summation, each value of x is multiplied by its relative frequency.

As the sample size increases, what happens to $\bar{x}$? We can get an idea by re-examining the die-rolling results (see Section 5.2, p. 147) and plotting the value of $\bar{x}$ against the number of rolls.



We can see that, as in the case of a relative frequency, after some initial oscillations it appears to be settling down to some value. We call this limiting value the **expectation** of X and denote it by $E(X)$. We can think of it as the long-term average value of X. Whilst $\bar{x}$ is approaching $E(X)$, each relative frequency is approaching the corresponding population probability. To determine the value of $E(X)$, therefore, we calculate:

$$E(X) = \Sigma x P_x \tag{5.4}$$

where the summation is over all possible values of X. Because of its derivation as the limiting value of the sample mean we see that:

$E(X)$ is the population mean value of X.

Notes

- The expectation of X does not have to be equal to an integer, nor does it have to be one of the possible values for X. We do not require these features for a sample mean, so there is no reason to require them for a population mean.
- A related concept is **expected frequency**: if the probability of an event is p and we have n observations, then the expected frequency of the event is np.

Example 7

The random variable X can only take the values 2 and 5. Given that the value 5 is twice as likely as the value 2, determine the expectation of X.

Suppose we denote the probability that X equals 2 by p. Then the probability that X equals 5 is $2p$. Since these are the only possible values for X, the sum of their probabilities is 1: $p + 2p = 1$. Since $3p = 1$ it follows that $p = \frac{1}{3}$. The expectation of X is therefore given by:

$$\begin{aligned} E(X) &= (2 \times P_2) + (5 \times P_5) \\ &= \left(2 \times \tfrac{1}{3}\right) + \left(5 \times \tfrac{2}{3}\right) \\ &= \tfrac{2}{3} + \tfrac{10}{3} \\ &= 4 \end{aligned}$$

The random variable X has a long-term average value of 4, though no individual values of X will be equal to that value.

Example 8

Determine the expectation of the random variable X, which has probability distribution given below.

Value of X	0	1	2	3
Probability	0.3	0.4	0.2	0.1

$$\begin{aligned} E(X) &= (0 \times P_0) + (1 \times P_1) + (2 \times P_2) + (3 \times P_3) \\ &= 0 + 0.4 + 0.4 + 0.3 \\ &= 1.1 \end{aligned}$$

The expectation of X is 1.1.

Practical

Roll a die four times, recording your results using a tally chart. Calculate the sample mean. Compare your results with other members of the class. You should find that almost everyone has a sample mean between 2 and 5.

Now roll the die a further thirty-six times and calculate the sample mean for the combined set of forty values. How variable are people's results now? You should find that most people have obtained values in the range 3 to 4. As the sample size increases so the sample mean becomes less likely to deviate far from 3.5.

Calculate a sample mean for the entire class.

Expected value or expected number

Sometimes either '**expected value**' or '**expected number**' is used in place of 'expectation' – these are all synonyms for one another. Whichever term is used, the numerical value that is being sought can be thought of as being *the long-term average value.*

Note

- This is just one of several places where Statistics has 'borrowed' a word from the ordinary English vocabulary but subtly altered its meaning – the 'expected value of X', using the statistical meaning of the phrase, does not have to be a value of X that is actually 'expected', using the everyday interpretation of the word 'expected'!

Example 9

In a multiple-choice paper, each question is followed by four alternative answers. The candidate is asked to ring one of these answers. If the answer ringed is correct, then the candidate gains 3 marks, but if the answer is incorrect the candidate loses 1 mark. Determine the expected value of the mark gained per question by the candidate if:

(i) the candidate chooses an answer at random,

(ii) the candidate knows that one of the incorrect answers is incorrect and chooses at random from the remaining three possibilities.

Comment on the results in each case.

Let X be the number of marks gained.

(i) The probability distribution of X is:

$$P_3 = \tfrac{1}{4} \qquad P_{-1} = \tfrac{3}{4}$$

so that:

$$\mathrm{E}(X) = \{3 \times (\tfrac{1}{4})\} + \{(-1) \times (\tfrac{3}{4})\} = \tfrac{3}{4} - \tfrac{3}{4} = 0$$

The examination marking scheme has been designed so that the expected mark obtained by someone who knows nothing and guesses every question will be zero.

(ii) The revised probability distribution, after the elimination of one of the incorrect answers is:

$$P_3 = \tfrac{1}{3} \qquad P_{-1} = \tfrac{2}{3}$$

so that:

$$\mathrm{E}(X) = \{3 \times (\tfrac{1}{3})\} + \{(-1) \times (\tfrac{2}{3})\} = 1 - \tfrac{2}{3} = \tfrac{1}{3}$$

Since $\mathrm{E}(X)$ is greater than 0, if one or more of the possibilities can be eliminated as being certainly incorrect, then there will be an advantage in guessing the answer. If the candidate were to guess, under these conditions, the answers to lots of questions, then the average gain would be one-third of a mark per question.

Expectation of X^2

We have seen that, essentially, $\mathrm{E}(X)$ is the long-term average value of the random variable X. In the same way $\mathrm{E}(X^2)$ is the long-term average value of X^2. The value of $\mathrm{E}(X^2)$ is calculated using:

$$\mathrm{E}(X^2) = \Sigma x^2 P_x \tag{5.5}$$

where the summation is over all possible values of X.

Note

- We will usually find that:

$$\mathrm{E}(X^2) \neq [\mathrm{E}(X)]^2$$

Example 10

Calculate the expected value of X^2, where X is the value obtained from rolling a fair six-sided die.

$$\begin{aligned}\mathrm{E}(X^2) &= 1^2 \times P_1 + 2^2 \times P_2 + \cdots + 6^2 \times P_6 \\ &= \tfrac{1}{6} + \tfrac{4}{6} + \cdots + \tfrac{36}{6} \\ &= \tfrac{91}{6}\end{aligned}$$

The expected value of X^2 is about 15 and is *not* simply the square of $\mathrm{E}(X)$ (which would have given the answer $3.5^2 \approx 12$).

Exercises 5c

Find $E(X)$ and $E(X^2)$ for Questions **1–9** below which are based on Questions **1–9** of Exercises 5a.

1 A box contains 3 red marbles and 5 green marbles. Two marbles are taken at random without replacement, and X is the number of green marbles obtained.

2 A box contains 3 red marbles and 5 green marbles. Two marbles are taken at random with replacement, and X is the number of green marbles obtained.

3 A fair coin has the number '1' on one face and the number '2' on the other. The coin is thrown with a fair die and X is the sum of the scores.

4 In a raffle, 20 tickets are sold and there are two prizes. One ticket number is drawn at random and the corresponding ticket earns a £10 prize. A second, different ticket number is drawn at random, and the corresponding ticket earns a £3 prize. The prize earned by a particular one of the original 20 tickets is £X.

5 Two cards are drawn at random (without replacement) from a pack of playing cards, and X is the number of Hearts obtained.

6 A fair die is thrown and X is the reciprocal of the score (i.e. 'one over the score').

7 Two fair dice, one red and the other green, are thrown and X is the score on the red die minus the score on the green die.

8 Two fair dice, one red and the other green, are thrown and X is the positive difference in the scores (i.e. the modulus of the random variable in the preceding example).

9 Packets of 'Hidden Gold' cornflakes are sold for £1.20 each. One in twenty of the packets contains a £1 coin. A shopper buys two packets and £X is the net cost of the two packets.

5.6 The variance

We have seen that, as a sample gets larger and larger so its properties will generally come increasingly to resemble those of the corresponding population. In particular we have:

Sample		Population
Relative frequency, $\frac{f_x}{n}$	$\longrightarrow$	Probability, P_{x_j}
Sample mean, $\bar{x}$	$\longrightarrow$	Population mean, $E(X)$
to these we now add:		
Sample variance, σ_n^2	$\longrightarrow$	Population variance, $\text{Var}(X)$

Two equivalent expressions for the population variance are:

$$\text{Var}(X) = E\left[\{X - E(X)\}^2\right] = E(X^2) - \{E(X)\}^2 \qquad (5.6)$$

Notes

- In practice, the word 'population' is often omitted and we simply write 'the variance of X'.
- The variance can never be negative, i.e. $\text{Var}(X) \geqslant 0$.

Example 11

The random variable X has probability distribution given by:

Value of X	2	5
Probability	0.4	0.6

Determine the variance of X.

We first calculate the expectation of X:

$$E(X) = (2 \times 0.4) + (5 \times 0.6) = 3.8$$

We next calculate $E(X^2)$:

$$E(X^2) = (2^2 \times 0.4) + (5^2 \times 0.6) = 16.6$$

Finally, using Equation (5.6), we get:

$$\text{Var}(X) = E(X^2) - \{E(X)\}^2 = 16.6 - 3.8^2 = 2.16$$

Example 12

The random variable X has probability distribution given by:

Value of X	2	3	4
Probability	p	p	$1 - 2p$

Show that X has variance equal to $p(5 - 9p)$.

We first calculate the expectation of X:

$$E(X) = (2 \times p) + (3 \times p) + \{4 \times (1 - 2p)\} = 4 - 3p$$

We next calculate $E(X^2)$:

$$E(X^2) = (2^2 \times p) + (3^2 \times p) + \{4^2 \times (1 - 2p)\} = 16 - 19p$$

Finally, using Equation (5.6), we get:

$$\begin{aligned}\text{Var}(X) &= E(X^2) - \{E(X)\}^2 = (16 - 19p) - (4 - 3p)^2 \\ &= 5p - 9p^2 = p(5 - 9p)\end{aligned}$$

as required.

Example 13

The random variable X has the Bernoulli distribution:

$$P_0 = 1 - p \qquad P_1 = p$$

Find the expectation and variance of X.

Finding $E(X)$ is straightforward:

$$E(X) = \{0 \times (1 - p)\} + \{1 \times p\} = p$$

To determine the variance of X we use Equation (5.6) and first find $E(X^2)$:

$$E(X^2) = \{0^2 \times (1 - p)\} + \{1^2 \times p\} = p$$

Hence:

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}(X^2) - \{\mathrm{E}(X)\}^2 \\ &= p - (p)^2 \\ &= p(1-p)\end{aligned}$$

A Bernoulli distribution with parameter p has mean p and variance $p(1-p)$.

Notes

- The geometric distribution with parameter p has mean $\frac{1}{p}$ and variance $\frac{1-p}{p^2}$. The proof is difficult and is omitted.
- If X is a random variable having a uniform distribution with the k equally likely values $x_1, \ldots, x_k$ then

$$\mathrm{E}(X) = \frac{1}{k}\sum_{i=1}^{k} x_i \qquad \mathrm{Var}(X) = \frac{1}{k}\sum_{i=1}^{k} x_i^2 - \left(\frac{1}{k}\sum_{i=1}^{k} x_i\right)^2$$

If the values for X are $1, \ldots, k$ then these formulae simplify to

$$\mathrm{E}(X) = \tfrac{1}{2}(k+1) \qquad \mathrm{Var}(X) = \tfrac{1}{12}(k^2 - 1)$$

Project

The nature of this project, which involves the use of a telephone directory, will depend on your location. In a country district, you should choose some large town from those covered by the directory; in a city, you should choose some well-defined large sub-area. Open the directory at random and start at the top of the left-hand page (unless it is an advertisement!). Count the number of subscribers until the first that you encounter with a number in your selected town (or sub-area). The reason for choosing a large town is so that your counting does not take too long! Record your value. Repeat the process until you have a total of 50 values.
Summarise your data using a bar chart (if most values are small) or a histogram (if most values are large).
Does your data look as though it could be described by a geometric distribution?

Calculate the sample mean, $\bar{x}$. Recall that the population mean is equal to $\frac{1}{p}$, assuming a geometric distribution. Deduce an estimate of the proportion of subscribers in the directory that reside in your chosen town or sub-area. (This is a rather involved way of estimating this proportion – how else might you have estimated it?)

5.7 The standard deviation

The population standard deviation is simply the square root of the population variance.

Example 14

The discrete random variable X has probability distribution given by:

$$P_x = \begin{cases} kx^3 & (x = 1, 2, 3) \\ 0 & \text{otherwise} \end{cases}$$

Determine, correct to 3 decimal places, the values of:
(i) the constant k,
(ii) the expectation of X,
(iii) the standard deviation of X.

(i) In order to find the value of k we use the fact that the probabilities sum to 1. Thus:

$$k(1^3 + 2^3 + 3^3) = 1$$

The sum of the left-hand side is $36k$ and hence $k = \frac{1}{36}$.

(ii) The expectation is $E(X)$, which is given by:

$$\begin{aligned} E(X) &= (1 \times k) + (2 \times 8k) + (3 \times 27k) = (1 + 16 + 81)k \\ &= 98k = \tfrac{98}{36} \\ &= 2.722 \text{ (to 3 d.p.)} \end{aligned}$$

(iii) To calculate the standard deviation we must first calculate the variance. Since $E(X)$ is not an integer we use Equation (5.6) and begin by calculating $E(X^2)$:

$$\begin{aligned} E(X^2) &= (1^2 \times k) + (2^2 \times 8k) + (3^2 \times 27k) = (1 + 32 + 243)k \\ &= 276k = \tfrac{276}{36} \end{aligned}$$

Hence:

$$\text{Var}(X) = E(X^2) - \{E(X)\}^2 = \tfrac{276}{36} - (\tfrac{98}{36})^2 \approx 0.2562$$

The standard deviation of X is therefore $\sqrt{0.2562} = 0.506$ (to 3 d.p.).

Note

- In order to achieve a desired accuracy of 3 decimal places, intermediate calculations have used fractions wherever practicable and have otherwise worked with extended accuracy so as to reduce round-off errors.

Exercises 5d

Find the variance and standard deviation of X in each of Questions **1**–**9** below, which have been seen earlier in Exercises 5a and 5c.

1 A box contains 3 red marbles and 5 green marbles. Two marbles are taken at random without replacement, and X is the number of green marbles obtained.

2 A box contains 3 red marbles and 5 green marbles. Two marbles are taken at random with replacement, and X is the number of green marbles obtained.

3 A fair coin has the number '1' on one face and the number '2' on the other. The coin is thrown with a fair die and X is the sum of the scores.

4 In a raffle, 20 tickets are sold and there are two prizes. One ticket number is drawn at random and the corresponding ticket earns a £10 prize. A second, different, ticket number is drawn at random, and the corresponding ticket earns a £3 prize. The prize earned by a particular one of the original 20 tickets is £X.

5 Two cards are drawn at random (without replacement) from a pack of playing cards, and X is the number of Hearts obtained.

6 A fair die is thrown and X is the reciprocal of the score (i.e. 'one over the score').

7 Two fair dice, one red and the other green, are thrown and X is the score on the red die minus the score on the green die.

8 Two fair dice, one red and the other green, are thrown and X is the positive difference in the scores (i.e. the modulus of the random variable in the preceding example).

9 Packets of 'Hidden Gold' cornflakes are sold for £1.20 each. One in twenty of the packets contains a £1 coin. A shopper buys two packets and £X is the net cost of the two packets.

10 Show that if X has a discrete uniform distribution on the integers $1, 2, \ldots, n$ then:

$$E(X) = \tfrac{1}{2}(n+1)$$
$$Var(X) = \tfrac{1}{12}(n^2 - 1)$$

11 The random variable X has a geometric distribution with variance 6.
Find $E(X)$.

12 The random variable Y has a Bernoulli distribution with standard deviation $\frac{3}{10}$.
Find the possible values of the expectation of Y.

13 A woman removes the labels from 3 tins of tomato soup and from 4 tins of peaches. She sends the labels off to the manufacturers in order to win herself a huggable teddy bear. Delighted with the prospect of the forthcoming bear, she forgets to mark the tins, which, devoid of their labels, then appear identical. The next week she is entertaining guests and requires a tin of peaches. She chooses tins at random, opening each in turn until a tin of peaches has been located. Let p_r be the probability that it is the rth tin that first contains peaches.
Determine the values of p_1, p_2, p_3 and p_4.
Determine the expectation and variance of the number of tins that are opened.

5.8 Greek notation

All branches of Mathematics display a liking for Greek symbols and Statistics is no exception! Conventionally the symbols μ (pronounced 'mu') and σ (a lower-case 'sigma') are reserved for the population mean and population standard deviation, with σ^2 being used for the variance. Thus, when studying the random variable X we may write:

$$E(X) = \mu \tag{5.7}$$

$$Var(X) = \sigma^2 \tag{5.8}$$

Note

- A useful guide as to whether, for a random variable X, we have calculated μ and σ^2 incorrectly, is provided by noting the following:
  - The population mean, μ, must have a value lying between the smallest and largest possible values for X.
  - If the range of possible values of X is finite then it usually has a magnitude of between 3σ and 6σ.

Example 15

Verify that the calculations in Example 14 seem reasonable.

In the previous example we calculated the expectation of X as being approximately 2.7 which does lie between 1 and 3, the extreme values that are possible for X. There is therefore no obvious indication that we calculated $E(X)$ incorrectly.

The range of the possible values of X is $3 - 1 = 2$. This is about 4 times 0.5, our calculated standard deviation, so it provides no suggestion of an incorrect calculation.

Chapter summary

- A **random variable** is denoted by a CAPITAL letter, e.g. X; an observed value by a lower-case letter, e.g. x.
- The **probability** of the value x, $P(X = x)$, is denoted by P_x.
- Probabilities **sum to 1**. For a discrete random variable that can take values $x_1, x_2, \ldots, x_n$

$$P_{x_1} + \cdots + P_{x_n} = 1$$

- **Expectation** (population mean):

$$E(X) = \Sigma x P_x \qquad E(X^2) = \Sigma x^2 P_x$$

- **Variance**: $\mathrm{Var}(X) = E[(X - \mu)^2] = E(X^2) - \{E(X)\}^2$
- *Special distributions*:
  - **Bernoulli**:

$$P_0 = 1 - p \qquad P_1 = p$$

$$E(X) = p \qquad \mathrm{Var}(X) = p(1 - p)$$

  - **Geometric** Geo(p) ($0 < p < 1$, $q = 1 - p$):

$$P_x = q^{x-1}p \qquad x = 1, 2, \ldots$$

$$P(X \leqslant x) = 1 - q^x \qquad P(X < x) = 1 - q^{x-1}$$

$$P(X > x) = q^x \qquad P(X \geqslant x) = q^{x-1}$$

$$E(X) = \frac{1}{p} \qquad \mathrm{Var}(X) = \frac{q}{p^2}$$

Exercises 5e (Miscellaneous)

1 A typically nutty statistician performs the following experiment. He first tosses a fair tetrahedron whose sides are numbered 0, 1, 2 and 3. When it lands, three sides are visible. Let n be the number on the fourth side. The statistician now tosses n unbiased coins.
Let X be the number of heads obtained.
(a) Obtain the probability distribution of X.
(b) Show that X has expectation $\frac{3}{4}$. Find the variance of X.
(c) Suppose that on a particular occasion the statistician has obtained exactly one head. Determine the probability that, on that occasion, $n = 2$.

2 Peter sets out with two 50p coins and three 10p coins in his pocket. When he comes to pay for goods at a shop, he finds that two of the coins are missing. Find
(i) the probability that he can pay for £1 worth of goods,
(ii) the expected total value of the money lost.
[Assume that each coin is equally likely to be lost.] **[SMP]**

3 An unbiased six-sided die has numbers 1, 1, 1, 2, 2, 3 printed on its faces. It is thrown twice.
(i) By drawing a tree diagram, or otherwise, find the probability that the total score is 4.
(ii) Find the expected value of the total score. **[SMP]**

4 In the first trial of a random experiment the probability of a successful outcome is $\frac{2}{5}$. In the second trial the probability of a successful outcome will be $\frac{1}{5}$ if the outcome of the first trial was successful, and will be $\frac{4}{5}$ if the outcome of the first trial was not successful. Determine the probability distribution and the mean value of the number of successful outcomes that will be obtained in the first two trials of the random experiment. [WJEC]

5 The probability distribution of a discrete random variable X is given by

$$P(X = r) = kr, \quad r = 1, 2, 3, \ldots, n,$$

where k is a constant. Show that

$$k = \frac{2}{n(n+1)}$$

and find, in terms of n, the mean of X. [JMB]

6 An experiment is carried out with three coins. Two of the coins are fair, so that the probability of obtaining a 'head' on any throw is $\frac{1}{2}$, while the third coin is biased so that the probability of obtaining a 'head' on any throw is $\frac{1}{4}$.

The three coins are thrown, and events A and B are defined as follows:

A occurs if all three coins show the same result;
B occurs if the biased coin shows a 'head'.

Find (i) $P(A)$, (ii) $P(A \cup B)$, (iii) $P(A' \cap B)$.

The random variable N denotes the number of 'heads' showing as a result of the experiment being carried out. Tabulate the probability distribution of N, and hence or otherwise calculate $E(N)$. [UCLES]

7 A circular card is divided into 3 sectors scoring 1, 2, 3 and having angles 135°, 90°, 135° respectively. On a second circular card, sectors scoring 1, 2, 3 have angles 180°, 90°, 90°, respectively. Each card has a pointer pivoted at its centre. After being set in motion, the pointers come to rest independently in random positions. Find the probability that

(i) the score on each card is 1,
(ii) the score on at least one of the cards is 3.

The random variable X is the larger of the two scores if they are different, and their common value if they are the same.
Show that $P(X = 2) = \frac{9}{32}$.
Show that $E(X) = \frac{75}{32}$ and find $Var(X)$. [UCLES]

8 (*a*) A regular tetrahedron is a solid with four faces, all identical. What is the probability, if it is tossed in the air, that it will land on a given face?

A dice is made by numbering the four faces 1, 2, 3, 4. The random variable X represents the number on the face on which the tetrahedron lands. The mean of X is 2.5. Calculate the variance of X.

(*b*) A circular disc is marked with a '1' on one side and a '2' on the other. Y is the random variable representing the number showing on the visible face of the disc when it is tossed and lands. Demonstrate that the mean and variance of Y are $1\frac{1}{2}$ and $\frac{1}{4}$ respectively.

(*c*) The disc and the dice are tossed together. The sum of the outcomes is recorded. Z is the random variable representing independent sums of X and Y. Write down the probability distribution for Z, and use this to calculate its mean and variance. [UODLE(P)]

9 A woman is waiting for a taxi to come into view so that she can hail it. Explain why the Geometric distribution is appropriate to model the number of vehicles she sees up to and including the first taxi.

If 5% of the vehicles in the locality are taxis:

(*a*) write down the mean number of vehicles up to and including the first taxi;

(*b*) calculate, giving your answer to three significant figures, the probability that the first taxi is the 6th vehicle to come into view;

(*c*) calculate the probability that the first taxi is among the first 6 vehicles she sees. [UODLE]

6 Expectation algebra

Oft expectation fails, and most oft there where most it promises

All's Well That Ends Well, William Shakespeare

In this chapter we derive some very useful results that are concerned with transformations of a single random variable and with combining information on several random variables. Where appropriate we shall use the simplifying notation of P_x for $P(X = x)$. We begin with an example that foreshadows some of these results.

Example 1

Suppose that the discrete random variable X has probability distribution given by:

$$P_0 = P_1 = 0.4 \qquad P_2 = 0.2$$

The random variable Y is defined by $Y = 2X - 1$.
Determine the mean and variance of X and of Y.
Comment on the results.

The simplest approach is to make a table of the probabilities and the possible values for X and Y:

Probability	0.4	0.4	0.2
Value of X	0	1	2
Value of $Y = 2X - 1$	−1	1	3

$$E(X) = (0 \times 0.4) + (1 \times 0.4) + (2 \times 0.2) = 0.8$$
$$E(Y) = \{(-1) \times 0.4\} + (1 \times 0.4) + (3 \times 0.2) = 0.6$$

In order to obtain the variances, we first calculate $E(X^2)$ and $E(Y^2)$:

$$E(X^2) = (0^2 \times 0.4) + (1^2 \times 0.4) + (2^2 \times 0.2) = 1.2$$
$$E(Y^2) = \{(-1)^2 \times 0.4\} + (1^2 \times 0.4) + (3^2 \times 0.2) = 2.6$$

Hence:

$$\text{Var}(X) = 1.2 - 0.8^2 = 1.2 - 0.64 = 0.56$$
$$\text{Var}(Y) = 2.6 - 0.6^2 = 2.6 - 0.36 = 2.24$$

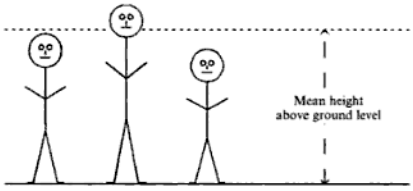
Comparing the values obtained we find:

$$E(Y) = E(2X - 1) = 2E(X) - 1$$
$$\text{Var}(Y) = \text{Var}(2X - 1) = 2^2 \times \text{Var}(X)$$

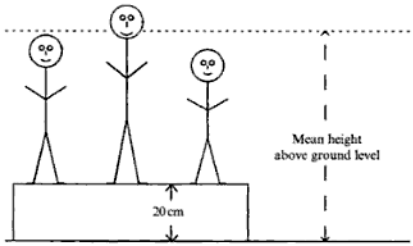
These connections between $E(X)$ and $E(Y)$ and between $\text{Var}(X)$ and $\text{Var}(Y)$ are not coincidental, but are examples of general results that we derive later in the chapter.

6.1 $E(X+a)$ and $Var(X+a)$

Suppose that the random variable X refers to the distance (in cm) between the top of a person's head and ground level. A group of three people is illustrated below.



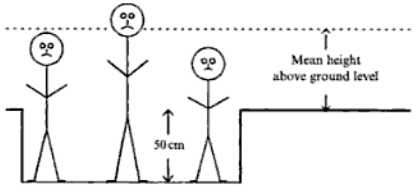
The three people now stand on a platform that is 20 cm high, as shown below.



Since all are standing on the platform, the value of X for each person has increased by 20 cm and therefore the average value of X has increased by 20 cm. However, the new values of X are no more variable than the old ones, since the individual differences from the mean height are the same as they were previously. Generalising these results to a platform of height a cm we have the results:

$$E(X+a) = E(X) + a \tag{6.1}$$
$$Var(X+a) = Var(X) \tag{6.2}$$

Suppose instead that the people now stand in a pit that is 50 cm deep.



Each X value has been reduced by 50 cm, so the average X value is reduced by that amount. However, the variability of the values of X is again unaffected. Generalising to a pit of depth a cm we have:

$$E(X-a) = E(X) - a \tag{6.3}$$
$$Var(X-a) = Var(X) \tag{6.4}$$

We now prove the first of these results algebraically, for a discrete random variable. Let $Y = X + a$, where a is any constant (either positive or negative). We want $E(Y)$ and $Var(Y)$.

Now:

$$E(Y) = E(X + a) = \Sigma(x + a)P_x$$

where the summation is over all possible values of X. Thus:

$$\begin{aligned} E(X + a) &= \Sigma xP_x + \Sigma aP_x \\ &= E(X) + a\Sigma P_x \qquad \text{by definition of } E(X) \\ &= E(X) + a \qquad \text{since } \Sigma P_x = 1 \end{aligned}$$

which proves the first result.
The second result can be established in a similar way.

Example 2

A lottery ticket costs 10p. There are 10 000 tickets for the lottery, which has a top prize of £100 and 9 runner-up prizes of £10 each. Determine the expected gain or loss resulting from the purchase of a ticket.

Let X be the random variable denoting the amount (in £) won by a ticket. Assuming all the tickets are sold, the probability distribution of X is given by:

Value of X	100	10	0
Probability	0.0001	0.0009	0.9990

Hence:

$$E(X) = (100 \times 0.0001) + (10 \times 0.0009) + (0 \times 0.999) = 0.01$$

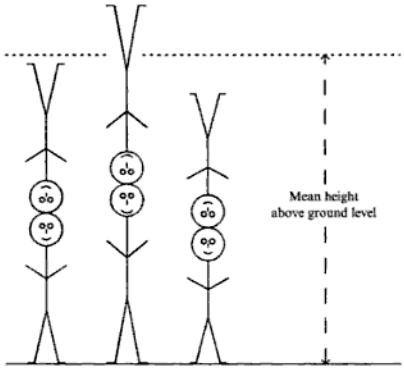
Since the lottery ticket costs £0.10, we want the expectation of $Y = X - 0.10$. Using the general result:

$$E(Y) = E(X - 0.10) = E(X) - 0.10 = -0.09$$

On average, therefore, the purchase of a ticket will result in a loss of 9p. Of course this need not deter us! We can probably afford to gamble on losing 10p, even with these very unfavourable circumstances, for the slight chance of being £99.90 better off.

6.2 E(aX) and Var(aX)

For simplicity, suppose that $a = 2$ and that each of the people illustrated in the original diagram was one of a pair of identical twins, remarkably adept at gymnastics – as illustrated opposite.



Let the random variable Y be the distance from the top of the twin's feet to ground level. Obviously, for each pair of twins, $Y = 2X$, where X was the height of one of the twins. So the average of the Y-values must be double that of the X-values. The general result, true for any random variable and for any constant a (positive or negative) is:

$$\mathrm{E}(aX) = a\mathrm{E}(X) \tag{6.5}$$

We will also need the result:

$$\mathrm{E}(a^2X^2) = a^2\mathrm{E}(X^2) \tag{6.6}$$

These results follow immediately from Equations (5.4) and (5.5) (pp. 154, 156).

It is obvious from the figure that the Y-values are a great deal more variable than the X-values, but to find out exactly how much more variable requires some algebra.

Let $Y = aX$ and denote $\mathrm{E}(X)$ by μ, so that:

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

Then:

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{E}(Y^2) - \{\mathrm{E}(Y)\}^2 \\ &= \mathrm{E}(a^2X^2) - (a\mu)^2 \\ &= a^2\mathrm{E}(X^2) - a^2\mu^2 \\ &= a^2\{\mathrm{E}(X^2) - \mu^2\} \\ &= a^2\mathrm{Var}(X)\end{aligned}$$

The general result is therefore that:

$$\mathrm{Var}(aX) = a^2\mathrm{Var}(X) \tag{6.7}$$

6.3 $E(aX + b)$ and $Var(aX + b)$

Combining the results of the previous two sections we have the result that, if a and b are any two constants, then:

$$E(aX + b) = E(aX) + b \quad \text{by Equation (6.1)} \tag{6.8}$$
$$= aE(X) + b \quad \text{by Equation (6.5)} \tag{6.9}$$

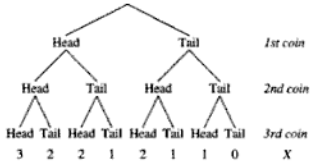
and also:

$$Var(aX + b) = Var(aX) \quad \text{by Equation (6.2)} \tag{6.10}$$
$$= a^2 Var(X) \quad \text{by Equation (6.7)} \tag{6.11}$$

Example 3

At a fairground there is the following game. The player pays 20p in order to toss three coins. The stall-holder pays the player (in pence) 10 times the number of heads that the player obtains.
Determine the mean and variance of the player's net loss.

Let X be the random variable indicating the number of heads obtained. We require the mean and variance of the net loss which (in pence) is given by $Y = 20 - 10X$.



Assuming that the coins are fair, each of the 8 alternatives has probability $\frac{1}{8}$, so that the probability distribution for X is:

Value of X	0	1	2	3
Probability	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

We can calculate $E(X)$ from its definition, or, more simply, we can note that since the distribution is symmetric, $E(X)$ will be equal to the mid-range, which in this case is $\frac{3}{2}$. Also:

$$E(X^2) = (0^2 \times \tfrac{1}{8}) + (1^2 \times \tfrac{3}{8}) + (2^2 \times \tfrac{3}{8}) + (3^2 \times \tfrac{1}{8}) = \tfrac{24}{8} = 3$$

Hence:

$$Var(X) = 3 - (\tfrac{3}{2})^2 = \tfrac{3}{4}$$

Using the general results we therefore get:

$$E(Y) = E(20 - 10X) = 20 - 10E(X) = 20 - 10 \times (\tfrac{3}{2}) = 5$$

which implies an average net loss of 5p a go, and:

$$Var(Y) = Var(20 - 10X) = (-10)^2 Var(X) = 100 \times \tfrac{3}{4} = 75$$

Exercises 6a

1 Given that $E(X) = 4$, $Var(X) = 2$, find
(i) $E(3X + 6)$, (ii) $Var(3X + 6)$, (iii) $E(6 - 3X)$,
(iv) $Var(6 + 3X)$.

2 Given that $E(X) = 3$, $Var(X) = 4$, find the expectation and variance of (i) $X - 2$,
(ii) $3X + 1$, (ii) $2 - 3X$.

3 Given that $E(Y) = \frac{1}{4}$ and $Var(Y) = \frac{1}{4}$, find $E(\frac{1}{3}Y)$ and $Var(\frac{1}{3}Y)$.

4 Given that the expectation of X is -2, and the standard deviation of X is 9, find
(i) $E\{(X + 2)^2\}$,
(ii) $E(X^2)$,
(iii) $E\{(X - 1)(X + 3)\}$.

5 Given that $E(3Y + 2) = 8$ and $Var(4 - 2Y) = 12$, find the expected value and the variance of Y.

6 Given that $E(Z) = 0$, $Var(Z) = 1$ and $Y = 3Z - 4$, find $E(Y)$ and $Var(Y)$.

7 The random variable U has mean 10 and standard deviation 5. The random variable V is defined by $V = \frac{1}{2}(U + 5)$. Find the mean and standard deviation of V.

8 It costs £30 to hire a car for the day, and there is a mileage charge of 10p per mile. The distance travelled in a day has expectation 200 miles and standard deviation 20 miles. Find the expectation and standard deviation of the cost per day.

9 Given that $E(X) = \mu$ and $Var(X) = \sigma^2$, find two pairs of values for the constants a and b such that $E(aX + b) = 0$ and $Var(aX + b) = 1$.

10 Given that $E(X) = \mu$, $Var(X) = \sigma^2$ and a is a constant, show that:

$$E\{(X - a)^2\} = (\mu - a)^2 + \sigma^2$$

Hence show that, as a varies, $E\{(X - a)^2\}$ is least when $a = \mu$ and find the least value.

11 The random variable T has mean 5 and variance 16.
Find two pairs of values for the constants c and d such that $E(cT + d) = 100$ and $Var(cT + d) = 144$.

12 Find $E(2S - 6)$ and $Var(2S - 6)$, where S is the score resulting from a single throw of an unbiased die.

13 The random variable Y takes the values $-1, 0, 1$ with probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$, respectively.
Find $E(10Y + 10)$ and $Var(10Y + 10)$.

6.4 Expectations involving more than one variable

We now quote a result that is rather obvious, but is surprisingly difficult to prove, namely that for two random variables, X and Y:

$$E(X + Y) = E(X) + E(Y) \tag{6.12}$$

Combining this result with those from the previous sections we have the more general results:

$$E(aX + bY + c) = aE(X) + bE(Y) + c \tag{6.13}$$

$$E(R + S + T + U) = E(R) + E(S) + E(T) + E(U) \tag{6.14}$$

Var($X + Y$)

If X and Y are independent, then:

$$Var(X + Y) = Var(X) + Var(Y) \tag{6.15}$$

Combining this result with Equation (6.11) we get the more general result that, if X and Y are independent, then:

$$\text{Var}(aX + bY + c) = a^2\text{Var}(X) + b^2\text{Var}(Y) \tag{6.16}$$

Once again, these results can be extended to cases involving more than two random variables. For example, if R, S, T and U are all mutually independent, then:

$$\text{Var}(R + S + T + U) = \text{Var}(R) + \text{Var}(S) + \text{Var}(T) + \text{Var}(U) \tag{6.17}$$

Note

- A particular case of Equation (6.16) that should be noted is:

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) \tag{6.18}$$

Example 4

Two fair six-sided dice are rolled. One die has its sides numbered 0, 0, 0, 1, 1, 2; the other die has its sides numbered 2, 2, 3, 3, 4, 4. Determine the mean and variance of Z, the total of the numbers shown by the dice.

Let X and Y be the numbers shown by the two dice.
We are interested in $Z = X + Y$. We require $\text{E}(Z)$ and $\text{Var}(Z)$.

We can use the result $\text{E}(Z) = \text{E}(X) + \text{E}(Y)$. Also, since the two dice are independent of one another, $\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y)$.

For X we have the probability distribution:

$$\text{P}(X = 0) = \tfrac{3}{6} \qquad \text{P}(X = 1) = \tfrac{2}{6} \qquad \text{P}(X = 2) = \tfrac{1}{6}$$

Hence:

$$\text{E}(X) = (0 \times \tfrac{3}{6}) + (1 \times \tfrac{2}{6}) + (2 \times \tfrac{1}{6}) = \tfrac{4}{6} = \tfrac{2}{3}$$

Also:

$$\text{E}(X^2) = (0^2 \times \tfrac{3}{6}) + (1^2 \times \tfrac{2}{6}) + (2^2 \times \tfrac{1}{6}) = 1$$

so that:

$$\text{Var}(X) = \text{E}(X^2) - \{\text{E}(X)\}^2 = 1 - (\tfrac{2}{3})^2 = \tfrac{5}{9}$$

For Y we have the probability distribution:

$$\text{P}(Y = 2) = \text{P}(Y = 3) = \text{P}(Y = 4) = \tfrac{1}{3}$$

By symmetry $\text{E}(Y) = 3$. Also:

$$\text{E}(Y^2) = (2^2 \times \tfrac{1}{3}) + (3^2 \times \tfrac{1}{3}) + (4^2 \times \tfrac{1}{3}) = \tfrac{29}{3}$$

so that:

$$\text{Var}(Y) = \text{E}(Y^2) - \{\text{E}(Y)\}^2 = \tfrac{29}{3} - 3^2 = \tfrac{2}{3}$$

Thus:

$$\text{E}(Z) = \text{E}(X) + \text{E}(Y) = \tfrac{2}{3} + 3 = \tfrac{11}{3}$$

and:

$$\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) = \tfrac{5}{9} + \tfrac{2}{3} = \tfrac{11}{9}$$

An alternative (long!) approach

A check on the above results is provided by tackling the distribution of Z head on! Again let X and Y be the numbers shown by the two dice. Since X and Y are independent, the probability of the outcome (say) $X = 0$ and $Y = 2$ is the product of their separate probabilities: $\frac{1}{6} \times \frac{1}{3} = \frac{1}{18} = \frac{1}{6}$. A convenient summary of the 9 possible value combinations is as follows:

Probabilities

First die \ Second die	2	3	4
0	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$
1	$\frac{2}{18}$	$\frac{2}{18}$	$\frac{2}{18}$
2	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$

Totals

First die \ Second die	2	3	4
0	2	3	4
1	3	4	5
2	4	5	6

The distribution for Z is therefore:

Value of Z	2	3	4	5	6
Probability	$\frac{1}{18}$	$\frac{2}{18}$	$\frac{6}{18}$	$\frac{2}{18}$	$\frac{1}{18}$

Hence:

$$\mathrm{E}(Z) = (2 \times \tfrac{1}{18}) + (3 \times \tfrac{2}{18}) + (4 \times \tfrac{6}{18}) + (5 \times \tfrac{2}{18}) + (6 \times \tfrac{1}{18})$$
$$= \tfrac{66}{18} = \tfrac{11}{3}$$

as before.

Also:

$$\mathrm{E}(Z^2) = (2^2 \times \tfrac{1}{18}) + \cdots + (6^2 \times \tfrac{1}{18}) = \tfrac{264}{18} = \tfrac{44}{3}$$

So that:

$$\mathrm{Var}(Z) = \mathrm{E}(Z^2) - \{\mathrm{E}(Z)\}^2 = \tfrac{44}{3} - (\tfrac{11}{3})^2 = \tfrac{11}{9}$$

Exercises 6b

1 The independent random variables X and Y are such that $\mathrm{E}(X) = 5$, $\mathrm{E}(Y) = 7$, $\mathrm{Var}(X) = 3$ and $\mathrm{Var}(Y) = 4$. Determine the mean and variance of the random variables U, V and W defined by:

$$U = 2X, \quad V = X + Y, \quad W = X - 2Y$$

2 It is given that $\mathrm{E}(X) = 3$, $\mathrm{Var}(X) = 16$, $\mathrm{E}(Y) = 4$, $\mathrm{Var}(Y) = 9$ and that X and Y are independent.
Find:
(i) $\mathrm{E}(X + Y)$, (ii) $\mathrm{Var}(X + Y)$, (iii) $\mathrm{E}(4X - 3Y)$, (iv) $\mathrm{Var}(4X - 3Y)$, (v) $\mathrm{E}(\frac{1}{4}X + \frac{1}{3}Y)$, (vi) $\mathrm{Var}(\frac{1}{4}X + \frac{1}{3}Y)$.

3 It is given that X_1 and X_2 are independent, and $\mathrm{E}(X_1) = \mathrm{E}(X_2) = \mu$, $\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = \sigma^2$. Find $\mathrm{E}(\bar{X})$ and $\mathrm{Var}(\bar{X})$, where $\bar{X} = \frac{1}{2}(X_1 + X_2)$.

4 It is given that $\mathrm{E}(X) = -5$, $\mathrm{Var}(X) = 25$, $\mathrm{E}(Y) = 8$, $\mathrm{Var}(Y) = 9$, and that X and Y are independent.
Find:
(i) $\mathrm{E}(X^2)$ and $\mathrm{E}(Y^2)$, (ii) $\mathrm{E}(3X^2 + 4Y^2)$.

5 H is the number of heads obtained when an unbiased coin is thrown and S is the score obtained when an unbiased die is thrown.
The random variable X is defined by $X = 2H - 6S$.
Find the expectation and variance of X.

6 A bank has two branches in Camchester. The number of customers on a Monday at the High Street branch has mean 100 and standard deviation 15. The number of customers on a Monday at the Station Road branch has mean 50 and standard deviation 20.
Find the mean and standard deviation of the total number of customers at both branches on a Monday. State any assumption that you need to make in order to be able to answer the question.

$E(X_1 + X_2)$ and $Var(X_1 + X_2)$

An important application of the previous results concerns the case where the random variables X and Y are replaced by two variables, X_1 and X_2 which are independent, but have identical probability distributions. In other words X_1 and X_2 share the properties that:

$$P(X_1 = x) = P(X_2 = x) \quad \text{(for all values of } x\text{)}$$
$$E(X_1) = E(X_2)$$
$$Var(X_1) = Var(X_2)$$

Denoting the common value of the population mean by μ, and the common variance by σ^2, then, using Equations (6.3) and (6.15), we have:

$$E(X_1 + X_2) = \mu + \mu = 2\mu$$

and:

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2) = \sigma^2 + \sigma^2 = 2\sigma^2$$

The difference between $2X$ and $X_1 + X_2$

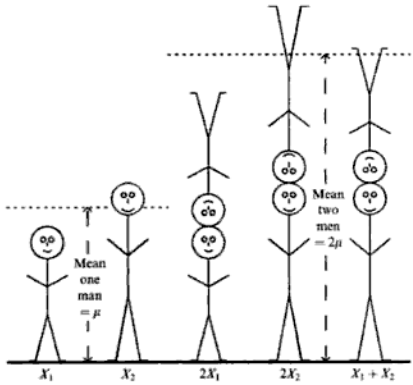
Suppose that each of the random variables X, X_1 and X_2 has mean μ and variance σ^2. Gathering the previous results together we have:

$$E(2X) = 2E(X) = 2\mu \qquad E(X_1) + E(X_2) = 2\mu$$

which is what we would expect. However, the results for the variances are not so accommodating:

$$Var(2X) = 2^2 Var(X) = 4\sigma^2 \qquad \text{but} \qquad Var(X_1 + X_2) = 2\sigma^2$$

Why is there a difference? To see the answer, consider the acrobats once again.



Suppose that the observed value of X_1 is less than μ: then $2X_1$ must certainly be less than 2μ. Likewise, if the observed value of X_2 is greater than μ, then $2X_2$ must be greater than 2μ.

However, on some occasions that X_1 is smaller than μ, X_2 will be larger than μ. Whenever this happens, the total of the values of X_1 and X_2 is likely to be quite close to 2μ. In this case therefore there is an opportunity for central values that does not exist in the previous case – hence the distribution is less variable.

Practical

In order to verify that there really is a difference between $2X$ and $X_1 + X_2$, we can perform two simple experiments using dice.

1. *Roll an ordinary die 25 times. On each roll double the score before recording it on a tally chart.*
   Calculate the values of the sample mean and variance.
2. *Roll a pair of dice 25 times. On each roll record the total of the two dice on a second tally chart.*
   Calculate the values of the sample mean and variance.

Verify that the two sample means are about equal, whereas the first sample variance is about twice the second.
To see why this has occurred draw a barchart of the outcomes of the first experiment and superimpose (using a different colour or different shading) the bar chart for the second experiment.

Example 5

The independent random variables X_1 and X_2 each have the probability distribution: $P_2 = 0.4$, $P_3 = 0.6$
Determine the values of (i) $\text{Var}(X_1)$, (ii) $\text{Var}(2X_1)$, (iii) $\text{Var}(X_1 + X_2)$.

For each X-variable we have:

$$\text{E}(X) = (2 \times 0.4) + (3 \times 0.6) = 2.6$$
$$\text{E}(X^2) = (2^2 \times 0.4) + (3^2 \times 0.6) = 7.0$$

Hence:

$$\text{Var}(X) = \text{E}(X^2) - \{\text{E}(X)\}^2 = 7.0 - 2.6^2 = 0.24$$

We can now answer the various questions:

(i) $\text{Var}(X_1) = 0.24$
(ii) $\text{Var}(2X_1) = 2^2\text{Var}(X_1) = 4 \times 0.24 = 0.96$
(iii) $\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) = 2 \times 0.24 = 0.48$

We can also obtain the final result by considering the distribution of $Y = X_1 + X_2$ directly:

$$\text{P}(Y = 4) = 0.4^2 = 0.16, \qquad \text{P}(Y = 6) = 0.6^2 = 0.36$$

and hence:

$$\text{P}(Y = 5) = 1 - 0.16 - 0.36 = 0.48$$

Hence:

$$\text{E}(Y) = (4 \times 0.16) + (5 \times 0.48) + (6 \times 0.36) = 5.2$$
$$\text{E}(Y^2) = (4^2 \times 0.16) + (5^2 \times 0.48) + (6^2 \times 0.36) = 27.52$$

and so:

$$\text{Var}(X_1 + X_2) = \text{Var}(Y) = \text{E}(Y^2) - \{\text{E}(Y)\}^2 = 27.52 - 5.2^2 = 0.48$$

6.5 The expectation and variance of the sample mean

Suppose we take a total of m samples, each of n independent observations, on the random variable X. Each sample will have a first observation, a second observation, and so on. Denote the jth observation in the ith sample by x_{ij}. The observations are summarised in the following table:

Sample number	1st observation	⋯	jth observation	⋯	nth observation
1	x_{11}	⋯	x_{1j}	⋯	x_{1n}
2	x_{21}	⋯	x_{2j}	⋯	x_{2n}
i	x_{i1}	⋯	x_{ij}	⋯	x_{in}
m	x_{m1}	⋯	x_{mj}	⋯	x_{mn}

Consider the first observations of all the samples that we take: $x_{11}, x_{21}, \ldots, x_{m1}$. These values vary because of random variation. For the ith sample, x_{i1} can be thought of as an observation on the random variable 'the first observation' which we denote by X_1. In the same way we can define a further $(n-1)$ random variables: $X_2, X_3, \ldots, X_n$. Since the observations are independent and are all observations of the same underlying random variable X, the n random variables $X_1, \ldots, X_n$ are independent and identically distributed, with their common distribution being that of X.

Denote the sample mean for the ith sample by $\bar{x}_i$. Because of random variation the values of $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m$ will also vary. There is therefore yet another lurking random variable, namely 'the sample mean', which we will denote by $\bar{X}$. Evidently:

$$\begin{aligned}\bar{X} &= \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \\ &= \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n\end{aligned}$$

Suppose that $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Using the result concerning expectations of sums of random variables, we have:

$$\begin{aligned}\mathrm{E}(\bar{X}) &= \mathrm{E}\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n\right) \\ &= \frac{1}{n}\mathrm{E}(X_1) + \frac{1}{n}\mathrm{E}(X_2) + \cdots + \frac{1}{n}\mathrm{E}(X_n) \\ &= \frac{1}{n}\mu + \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu \\ &= n\left(\frac{1}{n}\mu\right) \\ &= \mu\end{aligned}$$

The expectation of the sample mean is therefore the population mean – a pleasing result.

Since the random variables $X_1, X_2, \ldots, X_n$ are mutually independent:

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n\right) \\
&= \mathrm{Var}\left(\frac{1}{n}X_1\right) + \mathrm{Var}\left(\frac{1}{n}X_2\right) + \cdots + \mathrm{Var}\left(\frac{1}{n}X_n\right) \\
&= \left(\frac{1}{n}\right)^2 \mathrm{Var}(X_1) + \left(\frac{1}{n}\right)^2 \mathrm{Var}(X_2) + \cdots + \left(\frac{1}{n}\right)^2 \mathrm{Var}(X_n) \\
&= \left(\frac{1}{n}\right)^2 \sigma^2 + \left(\frac{1}{n}\right)^2 \sigma^2 + \cdots + \left(\frac{1}{n}\right)^2 \sigma^2 \\
&= n\left(\frac{1}{n}\right)^2 \sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

This is an important result because, for $n > 1$, this tells us that the sample mean is much less variable than are the individual observations. We can also see that the variance decreases as n increases, so that the sample mean is more and more likely to be close to the population mean as the sample size increases.

Example 6

The discrete random variable X has probability distribution $P_x = \dfrac{4-x}{10}$ for $x = 0, 1, 2, 3$.

Determine the variance of the sample mean (i) when the sample size is 2, (ii) when the sample size is 16.

In tabular form the probability distribution of X is as follows:

x	0	1	2	3
P_x	$\frac{4}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	$\frac{1}{10}$

The probabilities sum to 1, so it seems that we have interpreted the formula correctly! To answer the question we must first obtain the variance of a single observation on X. Now:

$$\mathrm{E}(X) = (0 \times 0.4) + (1 \times 0.3) + (2 \times 0.2) + (3 \times 0.1) = 1.0$$
$$\mathrm{E}(X^2) = (0^2 \times 0.4) + (1^2 \times 0.3) + (2^2 \times 0.2) + (3^2 \times 0.1) = 2.0$$

so that:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \{\mathrm{E}(X)\}^2 = 2.0 - (1.0)^2 = 1.0$$

From the general formula for a sample of size n we therefore have the answers (i) $\mathrm{Var}(\bar{X}) = \frac{1}{2}$, and (ii) $\mathrm{Var}(\bar{X}) = \frac{1}{16}$.

Note

- The square root of the variance of the sample mean, $\dfrac{\sigma}{\sqrt{n}}$, is often called the **standard error of the mean**, or simply the **standard error**. The same terms may be used for the corresponding sample value $\dfrac{s}{\sqrt{n}}$.

Exercises 6c

1 A random variable has expectation 12 and standard deviation 3. A sample of 81 observations is taken.
Find the expectation and variance of the sample mean.

2 An unbiased die is thrown 100 times and the score is observed.
Find the expectation and standard error of the mean score.

3 An unbiased die is thrown until the first six is obtained. The number X of throws needed to obtain the first six is observed. The process is repeated 50 times, giving 50 observations of X. Denoting the sample mean by $\bar{X}$, find the mean and standard error of $\bar{X}$.

4 A random variable Y takes the value 1 and 10, with probabilities p and $1-p$ respectively. 200 observations of Y are taken and the sample mean is $\bar{Y}$.
Find expressions for the mean and variance of $\bar{Y}$.

5 The mean weight of a soldier may be taken to be 90 kg, and the standard deviation may be taken to be 10 kg. 250 soldiers are on board an aircraft.
Find the expectation and variance of their average weight. State any assumption necessary.
Hence, or otherwise, find the mean and standard deviation of the total weight of the soldiers.

6 A random variable V has mean 150 and standard deviation 2. A random sample of n observations of V is taken.
Find the smallest value of n such that the standard error of the sample mean is less than 0.1.

7 A random variable R has mean 12 and variance 3. A random sample of n observations of R is taken.
Find the smallest value of n such that the expected value of the sample total exceeds 1000, and find the variance of the sample total for this value of n.

8 A computer program generates, with equal probabilities, one of the three numbers 0, 1 or 2. The variables X, Y and Z result from three independent runs of the program. If m is the mean of X, Y and Z, calculate the mean and variance of m.
If M is the median of X, Y and Z show that $P(M=0)=\frac{7}{27}$. Deduce the values of $P(M=2)$ and of $P(M=1)$. Hence determine the mean and variance of M.
If U is the largest of X, Y and Z calculate the mean and variance of U. [SMP]

6.6 The unbiased estimate of the population variance

In Chapter 2 we introduced the quantity s^2, given, for a sample of n observations $x_1, x_2, \ldots, x_n$, by the formula:

$$s^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$$

The phrase 'unbiased estimate' simply means that the corresponding random variable has expectation equal to σ^2, the population variance:

$$E\left[\frac{1}{n-1}\Sigma(X_i - \bar{X})^2\right] = \sigma^2$$

where, as in the previous section, X_i is the random variable 'the ith observation' and $\bar{X}$ is the random variable 'the sample mean'.

Chapter summary

- **Expectations and variances of functions of X:**

$$\begin{aligned} \mathrm{E}(X+a) &= \mathrm{E}(X)+a \\ \mathrm{E}(aX) &= a\mathrm{E}(X) \\ \mathrm{Var}(X+a) &= \mathrm{Var}(X) \\ \mathrm{Var}(aX) &= a^2\mathrm{Var}(X) \end{aligned}$$

- **Expectations of combinations of random variables:**

$$\begin{aligned} \mathrm{E}(X+Y) &= \mathrm{E}(X)+\mathrm{E}(Y) \\ \mathrm{E}(aX+bY+c) &= a\mathrm{E}(X)+b\mathrm{E}(Y)+c \\ \mathrm{E}(R+S+T+U) &= \mathrm{E}(R)+\mathrm{E}(S)+\mathrm{E}(T)+\mathrm{E}(U) \end{aligned}$$

- **Variances of combinations of *independent* random variables:**

$$\begin{aligned} \mathrm{Var}(aX+bY+c) &= a^2\mathrm{Var}(X)+b^2\mathrm{Var}(Y) \\ \mathrm{Var}(X-Y) &= \mathrm{Var}(X)+\mathrm{Var}(Y) \\ \mathrm{Var}(R+S+T+U) &= \mathrm{Var}(R)+\mathrm{Var}(S)+\mathrm{Var}(T)+\mathrm{Var}(U) \end{aligned}$$

- **Multiples and sums of random variables:**
  Combinations of identically distributed random variables having mean μ and variance σ^2

$$\begin{aligned} \mathrm{E}(2X) &= 2\mu \quad \text{and} \quad \mathrm{E}(X_1)+\mathrm{E}(X_2) = 2\mu \\ \mathrm{Var}(2X) &= 4\sigma^2 \quad \text{but} \quad \mathrm{Var}(X_1+X_2) = 2\sigma^2 \end{aligned}$$

- **Expectation and variance of sample mean:**

$$\mathrm{E}(\bar{X}) = \mu \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Exercises 6d (Miscellaneous)

1 At a certain institution, students living off-campus travel to campus on foot (0, 30%), by bicycle (2, 20%), by car (4, 30%) or by bus (6, 20%). The bracketed figures indicate the numbers of wheels of the mode of transport and the percentage of students involved.

(a) Let X be the number of wheels utilised by a randomly chosen student.
Determine $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

(b) Let S be the total number of wheels utilised by two randomly chosen students who travel independently of one another.
Determine the probability distribution of S.
State the mean and variance of S.

(c) A third student is now randomly chosen. This student travels independently of the two previously chosen students. Let the total number of wheels utilised by the three students be denoted by T.
Show that $\mathrm{P}(T=4) = 0.117$.
Find $\mathrm{P}(S=2|T=4)$.

(d) Let W be the event that at least one of the three students walks.
Find $\mathrm{P}(T=4|W)$ and $\mathrm{P}(W|T=4)$.

2 (i) Six fuses, of which two are defective and four are good, are to be tested one after another in random order until both defective fuses are identified. Find the probability that the number of fuses that will be tested is
(*a*) three,
(*b*) four or fewer.

(*continued*)

(ii) A random variable R takes the integer value r with probability $p(r)$ where

$p(r) = kr^3, \quad r = 1, 2, 3, 4,$
$p(r) = 0, \quad$ otherwise.

Find

(*a*) the value of k, and display the distribution on graph paper,

(*b*) the mean and the variance of the distribution,

(*c*) the mean and variance of $5R - 3$.

[ULSEB]

3 A darts player practises throwing a dart at the bull's-eye on a dart board. Independently for each throw, her probability of hitting the bull's-eye is 0.2. Let X be the number of throws she makes, up to and including her first success.

(a) Find the probability that she is successful for the first time on her third throw.

(b) Write down the distribution of X, and give the name of this distribution.

(c) Find the probability that she will have at least 3 failures before her first success.

(d) Show that the mean value of X is 5. (You may assume the result $\sum_{r=1}^{\infty} rq^{r-1} = \frac{1}{(1-q)^2}$ when $|q| < 1$.)

On another occasion the player throws the dart at the bull's-eye until she has two successes. Let Y be the number of throws that she makes up to and including her second success. Given that $\mathrm{Var}(X) = 20$, determine the mean and the variance of Y, and find the probability that $Y = 4$. [ULSEB]

4 A random variable X takes values $-1, 0, +1$ with probabilities $p, q, 2p$, respectively, and can take no other values.

(i) Express q in terms of p.

(ii) Find, in terms of p, the expected value and standard deviation of X.

(iii) If X_1 and X_2 are independent random variables each having the same distribution as X, find the probability distribution of $Y = X_1 + X_2$ and find $\mathrm{E}(Y)$, giving your answers in terms of p. [UCLES]

5 A coin and a six-faced die are thrown simultaneously. The random variable X is defined as follows:

If the coin shows a head,
then X is the score on the die.

If the coin shows a tail,
then X is twice the score on the die.

Find the expected value, μ, of X and show that $\mathrm{P}(X < \mu) = \frac{7}{12}$.

Show that $\mathrm{Var}(X) = \frac{427}{48}$.

The experiment is repeated and the sum of the two values obtained for X is denoted by Y. Find $\mathrm{P}(Y = 4)$ and $\mathrm{E}(Y)$. [UCLES]

6 The random variable X takes values $-2, 0, 2$ with probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ respectively. Find $\mathrm{Var}(X)$ and $\mathrm{E}(|X|)$.

The random variable Y is defined by $Y = X_1 + X_2$, where X_1 and X_2 are two independent observations of X. Find the probability distribution of Y.

Find $\mathrm{Var}(Y)$ and $\mathrm{E}(Y + 3)$. [UCLES]

7 The probability of there being X unusable matches in a full box of Surelite matches is given by

$\mathrm{P}(X = 0) = 8k, \mathrm{P}(X = 1) = 5k,$
$\mathrm{P}(X = 2) = \mathrm{P}(X = 3) = k, \mathrm{P}(X \geqslant 4) = 0.$

Determine the constant k and the expectation and variance of X.

Two full boxes of Surelite matches are chosen at random and the total number Y of unusable matches is determined. Calculate $\mathrm{P}(Y > 4)$, and state the values of the expectation and variance of Y. [UCLES]

8 Alfred and Bertie play a game, each starting with cash amounting to £100. Two dice are thrown. If the total score is 5 or more then Alfred pays £x, where $0 < x \leqslant 8$, to Bertie. If the total score is 4 or less then Bertie pays £$(x + 8)$ to Alfred. Show that the expectation of Alfred's cash after the first game is £$\frac{1}{3}(304 - 2x)$.

Find the expectation of Alfred's cash after six games.

Find the value of x for the game to be fair, i.e. for the expectation of Alfred's winnings to equal the expectation of Bertie's winnings.

Given that $x = 3$, find the variance of Alfred's cash after the first game. [UCLES]

7 The binomial distribution

To be or not to be: that is the question

Hamlet, William Shakespeare

It is not well known that Hamlet played cricket. He was captain of the local side and his problem was that when it came to calling heads or tails at the start of the match, he was rarely correct – just twice in the last test series of six matches against the Visigoths. The consequences were disastrous: whole families wiped out... He wondered what was the probability of being that unlucky?

Of course, the binomial distribution had not been discovered then! If it had been, then Hamlet would have known that, for 6 independent trials, with the probability of a success being $\frac{1}{2}$ for each trial, the probability of calling correctly just twice was $\binom{6}{2}(\frac{1}{2})^6 = \frac{15}{64}$ (which is about $\frac{1}{4}$, so he wasn't really all that unlucky after all!).

7.1 Derivation

The essential elements of Hamlet's problem, for which we shall develop a general formula, are:

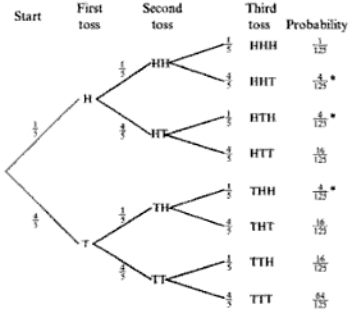
- A fixed number, n, of independent trials.
- Each trial results in either a 'success' or a 'failure'.
- The probability of success, p, is the same for each trial.

We have already encountered problems of this type for which a tree diagram helps.

Example 1

Determine the probability of getting 2 heads in 3 tosses of a bent coin which has $P(\text{Head}) = \frac{1}{5}$.

The tree of possible outcomes is as follows.



There are three possible sequences (indicated by a *) that lead to the outcome 'exactly 2 heads'. Each sequence has probability $\frac{4}{125}$. Hence the total probability of obtaining exactly 2 heads is $\frac{12}{125}$.

Exercises 7a

1 A cube has the letter 'A' on four faces and the letter 'B' on the remaining two faces. It is thrown three times.
Draw an appropriate tree diagram and find the probability that the number of A's obtained is (i) 0, (ii) 1, (iii) 2, (iv) 3.

Find also the probability that the number of B's obtained is (v) 0, (vi) 1, (vii) 2, (viii) 3.

2 Four players each have a pack of cards and, after shuffling each pack, they each turn over the top card of their pack.
By drawing an appropriate tree diagram, find the probability that the number of Hearts obtained is (i) 0, (ii) 2, (iii) 4.

3 A coin is tossed at the start of each cricket match in a series of 4 Test matches. One captain tosses and the other calls 'Heads' or 'Tails', at random. Find the probability that the toss is called correctly (i) exactly once, (ii) exactly twice.

Suppose the caller always calls 'Heads'. Does this alter the probabilities?
Give a reason for your answer.

4 Use a tree diagram to determine the probability of getting exactly two sixes when three fair dice are rolled one after another.

Would it make any difference to the probability if:

(i) all the dice were rolled at once,
(ii) instead of rolling three different dice, the same die was rolled each time?

5 A woman is trying to light a bonfire. She has only four matches left in her matchbox.
Given that ten per cent of matches break when struck, determine the probability that:
(i) all four matches will break when struck,
(ii) at least one match will not break when struck.
(Assume that all four matches are struck in each case.)

6 Every thousandth visitor to an exhibition is given a voucher for £50.
Assuming that 65% of the visitors to the exhibition are male, find the probability that, out of the first five to be given a voucher, exactly three are male.

Using a tree diagram is only feasible when the number of trials, n, is small. Otherwise we need a formula!

Look back at Example 1. There were three possible sequences leading to the desired outcome, and each sequence had the same probability. The answer we calculated was, in effect:

$$\begin{aligned}\text{P(exactly 2 heads)} &= (\text{Number of sequences}) \times \{\text{P(Head)}\}^2 \times \{\text{P(Tail)}\}^1 \\ &= 3 \times \left(\tfrac{4}{5}\right)^2 \times \left(\tfrac{4}{5}\right)^1 \\ &= \tfrac{12}{125}\end{aligned}$$

This approach works every time.

Example 2

Suppose a (rich) gambler has a biased coin for which the probability of a head is 0.55. He tosses the coin 8 times.
What is the probability of his getting 6 heads?

Using the method above, we get:

$$\begin{aligned}\text{P(6 heads in 8 tosses)} &= (\text{Number of sequences}) \times \{\text{P(Head)}\}^6 \times \{\text{P(Tail)}\}^2 \\ &= (\text{Number of sequences}) \times (0.55)^6 \times (0.45)^2\end{aligned}$$

In this case the number of sequences leading to the desired result happens to be 28, so:

$$\text{P(6 heads in 8 tosses)} = 28 \times (0.55)^6 \times (0.45)^2 = 0.157 \text{ (to 3 d.p.)}$$

The essential question is, for the general case of n independent trials, 'How many sequences in a probability tree lead to exactly r successes?' To answer this question, note that the r successes can be the result of any combination of r of the n trials. In Chapter 4 we introduced the notation $\binom{n}{r} = \frac{n \times (n-1) \times \ldots \times (n-r+1)}{r \times (r-1) \times \ldots \times 1}$ to represent the number of ways of choosing r out of n: the number of sequences is therefore $\binom{n}{r}$. (The number '28' used in Example 2 above is $\binom{8}{6}$.)

To illustrate this approach, consider the following problem:

> A marksman fires 10 times at a target.
> Assuming that the outcomes of the shots are independent of one another, and that each shot has probability 0.96 of being a 'bull', determine the probability that the marksman obtains exactly 9 bulls.

In this problem each shot is either a 'success' (a bull) or a 'failure'. Denoting the number of bulls obtained by X, we require $P(X = 9)$, which for convenience we will write as P_9. Since the number of sequences leading to exactly 9 bulls is $\binom{10}{9}$ $(= 10)$, we obtain:

$$P_9 = 10(0.96)^9(0.04)^1 = 0.277 \text{ (to 3 d.p.)}$$

What would have happened if the marksman's probability of obtaining a bull had been 0.92, instead of 0.96? To find out we simply replace 0.96 by 0.92 and 0.04 by 0.08, to get:

$$P_9 = 10(0.92)^9(0.08)^1 = 0.378 \text{ (to 3 d.p.)}$$

As the marksman's probability of a bull changes, so we change the values in the formula. If his probability of getting a bull had been p, then we would have had:

$$P_9 = 10p^9(1-p)^1$$

The generalisation is clear:

> **The probability of obtaining r successes out of n independent trials, when for each trial the probability of a success is p, is:**
>
> $$P_r = \binom{n}{r} p^r (1-p)^{n-r} \qquad (7.1)$$

This result, which provides the definition of the **binomial distribution**, makes no assumptions about the size of r and is therefore valid for all values of r from 0 to n inclusive.

Notes

- Remember $\binom{n}{r} = \binom{n}{n-r}$, $\binom{n}{0} = \binom{n}{n} = 1$ and $p^0 = 1$.
- The quantity $1 - p$ is often written as q.
- Write q for $(1 - p)$ and consider the **binomial expansion** of $(q + p)^n$, which is:

$$(q+p)^n = q^n + \binom{n}{1}q^{n-1}p^1 + \binom{n}{2}q^{n-2}p^2 + \cdots + \binom{n}{n-1}q^1p^{n-1} + p^n$$

  The probabilities $P_0, P_1, \ldots, P_n$ are the successive terms in this expansion. Since $q + p = 1$ this confirms that the sum of the binomial probabilities is 1.
- The most usual error in calculating a binomial probability is to forget that, in order for there to be *exactly* r successes, there must also be $n - r$ failures. The $(1 - p)^{n-r}$ factor must not be omitted from the formula!

- The binomial distribution can be used as a model for sampling *with replacement* from a population of any size.
- Only if a (finite) population is very large, can the binomial distribution be used as a model for sampling *without replacement*.

Example 3

According to a motoring magazine, in Britain, Japanese cars account for 5% of the cars on the road. Whilst held up in a traffic jam I occupy my time by examining the cars racing past on the other side of the road. Assuming that the magazine is correct, determine the probability that, of the first 50 cars that pass me, 4 are Japanese.

Each car is either Japanese (a 'success') or not Japanese. Assuming that the traffic jam is not immediately outside a car manufacturing plant, the 50 cars can be assumed to be a random sample of the cars on the road. The population of cars is sufficiently large for us to use the binomial distribution.

The number of trials, n is 50, since 50 cars are examined. The probability of a 'success', p, is 0.05 and the value of r is 4. Hence the required probability is:

$$\binom{50}{4}(0.05)^4(0.95)^{46} = 0.136 \text{ (to 3 d.p.)}$$

Example 4

Four cards are drawn at random from an ordinary pack of 52 cards. Determine the probability that precisely three are Spades (i) if the four are drawn *without* replacement, (ii) if the four are drawn one-at-a-time *with* replacement.

(i) The pack contains 13 Spades and 39 other cards. The probability that the first card drawn is a Spade is $\frac{13}{52} = \frac{1}{4}$. However, the probability of the second card drawn being a Spade depends upon the outcome of the first card drawn. If the first is a Spade then the probability of the second being a spade is $\frac{12}{51}$, whereas if the first is not a Spade then the probability of the second being a Spade is $\frac{13}{51}$. This situation was discussed in Section 4.17 (p. 95) and the required probability is:

$$\frac{\binom{13}{3} \times \binom{39}{1}}{\binom{52}{4}} = \frac{286 \times 39}{270\,725} = 0.041 \text{ (to 3 d.p.)}$$

(ii) In this case, for each draw the probability of getting a Spade is constant at $\frac{13}{52} = \frac{1}{4}$. The binomial distribution is now appropriate and the probability is:

$$\binom{4}{3}\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^1 = \frac{3}{64} = 0.047 \text{ (to 3 d.p.)}$$

The probability obtained in the 'with replacement' case is appreciably larger than that obtained in the 'without replacement' case.

Exercises 7b

1 The number of successes in n independent trials is X. The probability of a success in each trial is p.
(i) Given that $n = 10$, $p = \frac{1}{4}$, find $P(X = 3)$.
(ii) Given that $n = 8$, $p = \frac{1}{4}$, find $P(X = 6)$.
(iii) Given that $n = 12$, $p = \frac{1}{3}$, find $P(X \leqslant 3)$.
(iv) Given that $n = 11$, $p = \frac{4}{5}$, find $P(X \geqslant 9)$.
(v) Given that $n = 7$, $p = \frac{1}{2}$, find $P(3 \leqslant X \leqslant 5)$.

2 Five per cent of bluebells (confusingly) have white flowers. The remainder have blue flowers. Determine the probability that a random sample of ten bluebell plants includes exactly one with white flowers.

3 In a telephone poll 22% of the respondents believed in astrology and 78% did not. Assuming that the same proportions apply to the whole population, find the probability that in a random sample of 10 people, less than 20% believe in astrology.
Comment on the validity of the extrapolation from the poll to the population.

4 There are 15 students in a class.
Assuming that each student is equally likely to have been born on any day of the week, find the probability that three or fewer were born on a Monday.
Find also the probability that four or more were born on a Tuesday.

5 Two parents each have the gene for cystic fibrosis. For each of their children, the probability of developing cystic fibrosis is $\frac{1}{4}$. If there are four children, find the probability that exactly two develop cystic fibrosis.

6 A pair of dice is thrown 20 times.
Find the probability of getting a double six at least 3 times.

7 When the Romans decimated a population they lined up the men and executed every tenth man. Six brothers stood in random places in the line. Find the probability that:
(i) none were executed,
(ii) four or more escaped execution.

8 A large box contains a mixture of three different types of bolt, in equal numbers. Another box contains the nuts for the bolts. Each nut only fits a bolt of the same type. A nut and a bolt are chosen at random and checked to see if they match (i.e. they are of the same type). The process is repeated 12 times. Find the probability that more than 4 matches are obtained.

9 Driving to work I have to negotiate three sets of traffic lights. I have observed that each of these shows green for 0.45 of the time, red for 0.45 of the time and amber for the remaining time.
Assuming that the colours of the traffic lights are independent of one another and of the time at which I reach them, determine the probability that exactly two of the lights force me (a law-abiding citizen) to stop (by showing either amber or red).

10 The characters in a film are classed as being either 'Good', 'Bad', or 'Ugly'. The proportions in these classes are, respectively, 0.4, 0.4 and 0.2. Seven of the characters have red hair. Assuming that class and hair colour are independent, determine the probability that exactly two of these characters are 'Ugly'.

Project

Cars provide a very convenient set of easily collected data with which to test how well a binomial model works. Suppose we define a success to be 'a number plate in which the last digit is a 3, a 6 or a 9'. Assuming that all 10 digits from 0 to 9 are equally likely, the probability of a success is therefore 0.3. In a sample of five cars the probability of observing, for example, 3 successes, should be:

$$\binom{5}{3}(0.3)^3(0.7)^2 = 0.132 \text{ (to 3 d.p.)}$$

Does it really work out like this? To find out, we noted the last digits of a sequence of 200 cars that passed by.

The first car that passed was an old banger. Its number was GJG 1944, which ends in a '4' – an immediate failure. After 15 cars had passed our records (the last numbers) looked like this:

Groups of five cars	*4, 9, 1, 7, 4*	*3, 5, 5, 6, 0*	*4, 4, 2, 0, 8*
Numbers of successes	*1*	*2*	*0*

When all the data had been collected the numbers of successes that we had obtained were as follows:

1,2,0,0,0 2,0,3,1,0 1,1,2,2,4 0,1,1,1,2
3,2,0,1,2 0,1,1,0,1 2,3,4,2,1 0,1,2,1,1

Our next step was to summarise the values using a tally chart, so that we could subsequently tabulate the results as follows:

Number of successes	*0*	*1*	*2*	*3*	*4*	*5*
Observed frequency	*10*	*15*	*10*	*3*	*2*	*0*
Relative frequency	*0.250*	*0.375*	*0.250*	*0.075*	*0.050*	*0.000*
Theoretical probability	*0.168*	*0.360*	*0.309*	*0.132*	*0.028*	*0.002*

All in all the model has not done too badly! The largest observed proportion corresponds to the case '1' as predicted, and the fact that we never observed a '5' should not be a surprise.

This project can easily be varied. For example, the value of n need not be five. Similarly, the value of p can easily be altered: for example, we could set p to be 0.15 by defining a 'success' to be a number plate ending in some number between '00' and '14', inclusive – ignoring single digit number plates.

Decide on values for n and p and collect some of your own. If class results are pooled together, then (assuming that the class members obtained independent samples) the agreement with the theoretical model should be even better.

Practical

We have seen that coin-tossing provides a simple example of a binomial situation. If a fair coin is tossed six times, and the random variable X denotes the number of heads obtained, then:

$$P_r = \binom{6}{r}(0.5)^r(0.5)^{6-r} \qquad r = 0, 1, \ldots, 6$$

$$= \binom{6}{r}(0.5)^6$$

the probabilities of the various values of r are:

Outcome, r	*0*	*1*	*2*	*3*	*4*	*5*	*6*
Probability (to 3 d.p.)	*0.016*	*0.094*	*0.234*	*0.313*	*0.234*	*0.094*	*0.016*

Toss a coin six times and record the number of heads, r, that you obtain. Repeat a further nine times and compare the relative frequencies for your outcomes with the theoretical probabilities. The resemblance is unlikely to be close since ten observations is a very small sample.

Combine your results with those of the rest of the class to get a larger amount of data. You should find the overall class results closely resemble those predicted by the binomial distribution.

In Chapter 13 we shall see how to use precise methods to test the goodness of fit between a theoretical distribution and an observed set of data.

Practical

Take out one suit from a pack of cards. Shuffle these cards and choose one card at random. Replace the card and repeat a further four times. Record the number of times (out of the five) that you obtain a court card (a Jack, Queen or King). For example, suppose the original card is a 7, and the next four are respectively 9, 3, King and 7. A court card has occurred on just one of the five occasions, so the outcome of this experiment is a '1'.

Repeat the entire process to obtain a total of twenty observations, each with a value between 0 and 5, inclusive. Combine your results with those of your neighbours in class and calculate the relative frequencies of the outcomes in your combined sample of results. Calculate the theoretical probabilities for this situation and compare them with your relative frequencies.

In the above experiment the cards were replaced after their values had been noted. Repeat the experiment without replacing the cards. This is most easily done by choosing five cards from the collection of thirteen and noting the number of court cards.
Compare your results with those obtained previously.

Calculator practice

If your calculator has a random number generator facility and can be set to show a fixed number of decimal places, then the generator can be tested for randomness as follows. Set the display to show (say) 6 decimal places. Generate a random number and count the number of 9s that it contains. The number of 9s is an observation from a binomial distribution with $n = 6$ *and* $p = 0.1$. *Repeat the process a further 99 times, summarise your results using a tally chart and compare the observed proportions of the outcomes 0 to 6 with those predicted by the binomial model. If they appear to be very different then either something (your calculations?) is wrong or you have been very unlucky!*

Computer project

With a computer it is easy to write a program that will both generate the random numbers (as in the previous calculator project) and count the number of occurrences of 9s (or anything else that takes one's fancy). The advantage of the computer is that it can do this a really large number of times. Advanced programmers will arrange for the output to list both the observed proportions and the theoretical probabilities. If a graphical display is available then pictures can also be created.

7.2 Notation

To save having to write: 'The random variable X has a binomial distribution. There are n independent trials. The probability of a "success" is p for each trial', we write:

$$X \sim \mathrm{B}(n, p)$$

Here, the symbol $\sim$ means 'has distribution' and 'B' is used as a shorthand for 'binomial'.

The quantities n and p are called the **'parameters'** of the distribution; they are the quantities whose values are required in order to specify the distribution completely.

7.3 'Successes' and 'failures'

Some good news! It does not matter which of the two possible outcomes we think of as being a 'success' – the calculations will be the same. For example, suppose I play a game of chance with an opponent and suppose my probability of winning is p. Obviously my 'successes' are my opponent's 'failures' and vice versa.

	Me	Opponent
P(success)	p	$q = 1 - p$
Number of successes	r	$n - r$

Example 5

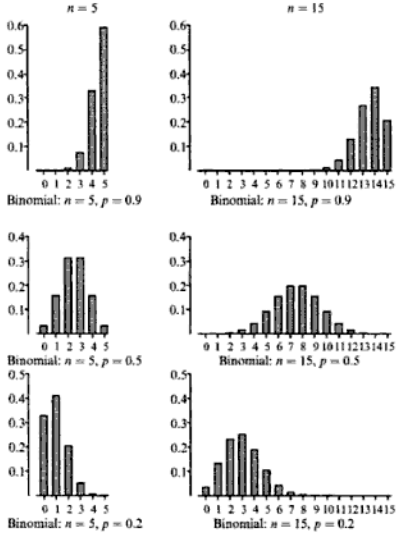
In Example 3, when we required the probability of observing four Japanese cars in a random sample of 50 cars, we defined a 'success' to be 'a Japanese car'. Suppose instead that we define a 'success' to be 'a *non*-Japanese car'. Thus n is 50, the probability of a 'success', p, is 0.95 and the value of r, the required number of 'successes', is now 46. The required probability is:

$$P_{46} = \binom{50}{46} (0.95)^{46}(0.05)^4 = 0.136 \text{ (to 3 d.p.)}$$

which is the same value that we obtained previously.

7.4 The shape of the distribution

The shape of the binomial distribution depends upon the value of p. When $p = \frac{1}{2}$, this means that a 'success' is just as likely as a 'failure'. So, for example, the probability of obtaining two 'successes' (and hence $n - 2$ 'failures') is equal to the probability of obtaining two 'failures' (and hence $n - 2$ 'successes'); when $p = \frac{1}{2}$ the binomial distribution is symmetric. For other values of p the distribution is asymmetric (skewed), with a mode near np (see the examples in the diagram).



Project

An important requirement for town planning is a knowledge of the type of traffic that uses the principal roads. In particular, planners need to know the proportion of vehicles that are cars (as opposed to lorries, vans, etc).

Suppose that you are part of the local transport authority. Go to a busy road junction and record the identities of the vehicles that pass you. For every car that passes record a 1, and for every other vehicle record a 0. Set out your records carefully in groups of ten vehicles, so that it might look like this:

1111110111 1100111011 1000111110....

Continue until you have recorded the identities of 200 vehicles. Now count the numbers of cars in each group of 10 vehicles. In the example above, the three groups contained 9, 7 and 6 cars respectively. Summarise your data, initially using a tally chart, and then using a frequency distribution, which might look like this:

Number of cars in group (r)	*10*	*9*	*8*	*7*	*6*
Number of groups with this number of cars	*8*	*5*	*4*	*2*	*1*
Relative frequencies	*0.40*	*0.25*	*0.20*	*0.10*	*0.05*

Now count the total number of cars that you have observed. In the example above, this is:

$$(10 \times 8) + (9 \times 5) + (8 \times 4) + (7 \times 2) + (6 \times 1) = 177$$

Finally determine your personal estimate of p, the probability of a vehicle being a car. In the example this is $\frac{177}{200} = 0.885$. *You can now calculate the theoretical probabilities of seeing* $10, 9, 8, \ldots$ *cars in a group of 10 vehicles, based on your personal estimate of p. In our example the binomial model predicts that the proportion of groups consisting of all cars would be:*

$$\binom{10}{10}(0.885)^{10}(0.115)^{0} = 0.295 \text{ (to 3 d.p.)}$$

while the proportion with nine cars will be:

$$\binom{10}{9}(0.885)^{9}(0.115)^{1} = 0.383 \text{ (to 3 d.p.)}$$

and so on.

An extension to this project is to compare (using, for example, a paired bar chart) the theoretical probabilities with your observed relative frequencies. If your relative frequencies for the small values of r are unexpectedly large then this suggests that lorries (and buses!) travel in convoys.

Naturally your personal estimate of p will not be very accurate. Therefore, as a final task, the comparisons may be repeated using the pooled data from the entire class. (For this to be sensible, the observations taken by each class member should be independent of one another, e.g. at slightly different times at the same road junction.) Does it appear that the binomial model is appropriate, or do 'non-cars' (and cars) appear to cluster?

7.5 Tables of binomial distributions

Tables of binomial distributions may take any of the following forms:

(i) tables of $P(X = r)$,
(ii) tables of $P(X \geqslant r)$,
(iii) tables of $P(X \leqslant r)$.

Tables vary from book to book (tables of the third form are given in the Appendix, p. 437). The tables recommended or supplied by the various examination boards also vary. You should be careful to become familiar with the tables relevant to your examination: there is no point in wasting time in calculating a value if it is in the tables supplied!

Here is an extract from a table (of the third type) that gives the values of $P(X \leqslant r)$:

n	r	p
		0.10
5	0	0.5905
	1	0.9185
	2	0.9914
	3	0.9995
	4	
	5	

The table gives the cumulative probabilities correct to four decimal places. For any combination of n and p, once a value greater than 0.999 95 has been reached it will be shown in the table as a blank.

In the fragment of table given we see that, if $X \sim B(5, 0.1)$, then $P(X \leqslant 2) = 0.9914$.

In order to use the tables to find probabilities for individual values of r, or for other types of inequality, we need the following relations:

$$\begin{aligned} P(X < r) &= P(X \leqslant r-1) \\ P_r = P(X = r) &= P(X \leqslant r) - P(X \leqslant r-1) \\ P(X > r) &= 1 - P(X \leqslant r) \\ P(X \geqslant r) &= 1 - P(X \leqslant r-1) \end{aligned}$$

Notes

- It is easy to confuse $P(X < r)$ with $P(X \leqslant r)$, so questions should always be read very carefully.
- It is important to be familiar with the tables that are available for your examinations. The tables given in this book are just a convenience!
- In this book the values of p in the table range from 0.05 to 0.5 only. For values of p greater than 0.5 we must interchange the definitions of success and failure. Thus, if $Y \sim B(n, q)$ and $q = 1 - p$, then:

  $$P(X = r) = P(Y = n - r)$$

  so that, for example:

  $$P(X \geqslant r) = P(Y \leqslant n - r)$$

- Inevitably the tables do not provide for *every* combination of n and p. If a value is required that cannot be obtained directly from the tables then, if possible, it should be calculated from the formula rather than by interpolation in the tables (since this may not be very accurate).

Example 6

Given that $X \sim B(5, 0.1)$ find (i) $P(X < 2)$, (ii) $P(X = 2)$, (iii) $P(X > 2)$, (iv) $P(X \geqslant 2)$.

(i) $P(X < 2) = P(X \leqslant 1) = 0.9185$
(ii) $P(X = 2) = P(X \leqslant 2) - P(X \leqslant 1) = 0.9914 - 0.9185 = 0.0729$
(iii) $P(X > 2) = 1 - P(X \leqslant 2) = 1 - 0.9914 = 0.0086$
(iv) $P(X \geqslant 2) = 1 - P(X \leqslant 1) = 1 - 0.9185 = 0.0815$

Example 7

Given that $X \sim B(5, 0.9)$, determine (i) $P(X \geqslant 2)$, (ii) $P(X \leqslant 2)$.

In this case the value of p is 0.9, which is greater than 0.5, so we must work instead with Y, where $Y \sim B(5, 0.1)$.
(i) $P(X \geqslant 2) = P[Y \leqslant (5-2)] = P(Y \leqslant 3) = 0.9995$
(ii) $P(X \leqslant 2) = P(Y \geqslant 3) = 1 - P(Y \leqslant 2) = 1 - 0.9914 = 0.0086$

Exercises 7c

Use the tables that you will use in your examination, or the tables in the Appendix at the back of this book, to answer the following questions.

1 Given that $X \sim B(8, 0.3)$, find (i) $P(X \leqslant 4)$, (ii) $P(X > 6)$.

2 Given that $X \sim B(10, 0.4)$, find (i) $P(X \geqslant 7)$, (ii) $P(X = 6)$, (iii) $P(X < 5)$.

3 Given that $X \sim B(15, 0.7)$, find (i) $P(X \geqslant 9)$, (ii) $P(X \leqslant 11)$.

4 Given that $X \sim B(12, 0.6)$, find $P(5 \leqslant X \leqslant 8)$.

5 When serving at tennis, the probability that Holly Hitter gets the first service in court is 30%. If the first service is a fault (i.e. does not go in court), there is a second service and the probability that the second service goes in court is 90%. Find the probability that out of 20 first services more than 10 go in court.

Show that the probability of a double fault (i.e. neither service goes in court) is 0.07.

Find the probability that out of 20 (combined) services more than 3 are double faults.

6 University student Joe Sleepwell often misses 9 o'clock lectures through oversleeping. The probability that he oversleeps is 0.4.
Find the probability that, in a nine-week term, with two 9 o'clock lectures each week, he misses more than half of them.

7.6 The expectation and variance of a binomial random variable

We want to find the values of $E(X)$ and $Var(X)$, where $X \sim B(n, p)$. It is a little difficult to calculate these quantities using the formula for $P(X = r)$. Instead we note that:

$$X = Y_1 + Y_2 + \cdots + Y_n$$

where Y_i is the number of successes (0 or 1) obtained on the ith trial. Now $Y_1, Y_2, \ldots$ are independent Bernoulli random variables of the type studied in Chapter 5. We found there that a Bernoulli random variable with parameter p has expectation equal to p and variance equal to $p(1 - p)$. Combining this information with that from Chapter 6 on the expectations of sums of random variables, we have:

$$\begin{aligned} E(X) &= E(Y_1 + Y_2 + \cdots + Y_n) \\ &= E(Y_1) + E(Y_2) + \cdots + E(Y_n) \\ &= (p + p + \cdots + p) \\ &= np \end{aligned}$$

Similarly, writing q for $(1 - p)$:

$$\begin{aligned} Var(X) &= Var(Y_1) + Var(Y_2) + \cdots + Var(Y_n) \\ &= pq + pq + \cdots + pq \\ &= npq \end{aligned}$$

Thus:

a random variable having a $B(n,p)$ distribution has expectation np and variance npq.

Later results (Sections 10.11 and 10.8) show that (if n is reasonably large, and p and q are not too small) on about 95% of occasions the observed value of a random variable X having a $B(n, p)$ distribution will lie in the range: mean ± 2 standard deviations, i.e. in the range $np \pm 2\sqrt{npq}$.

Note

- The Bernoulli distribution is really a special case of the binomial distribution in which $n = 1$.

Example 8

A very lazy candidate has done no revision for his multiple-choice statistics exam and guesses the answer to each of the 40 questions. Given that each question offers four alternative answers, only one of which is correct, determine the mean and variance of X, the number of correct answers obtained.

The probability, p, that the candidate guesses the correct answer to a question is $\frac{1}{4}$. The situation is binomial, since, for each question, the candidate is either correct or not correct. Thus $X \sim B(40, \frac{1}{4})$. Hence:

$$E(X) = np = 40 \times 0.25 = 10$$

and:

$$\text{Var}(X) = npq = E(X) \times 0.75 = 7.5$$

Exercises 7d

1 Determine the expectation and variance of a binomial random variable X for which $n = 50$ and $p = 0.2$.

2 Two boys are throwing skimmers. The probability that a skimmer thrown by Alec will hop 5 times (a success!) is 0.2, whereas for Bill the probability is 0.1. Both boys throw 10 skimmers. Determine:
(i) the expectation and variance of the number of successes obtained by Alec,
(ii) the expectation and variance of the number of successes obtained by Bill,
(iii) the expectation and variance of the total number of successes obtained by the two boys.

3 A die is thrown 10 times. Let X be the number of sixes obtained.
Find μ, the expected number of sixes.
Find also $\text{Var}(X)$ and $P(X < \mu)$.

4 A motorist making a regular journey to work finds that she is delayed at a particular level crossing once in five journeys, on average. Using a binomial model, find the expected number of journeys that are delayed at the level crossing in a month when she makes 22 journeys to work, and find also the probability that she is delayed on fewer than 4 occasions.
Comment on the appropriateness of the binomial model.

5 Published articles in medical journals indicate that, on average, 35 out of 100 patients having a lumbar puncture will suffer SSH ('Severe Spinal Headache'). Twelve patients are given a lumbar puncture.
Using a binomial model, find the expected number of patients who will suffer SSH, and find also the standard deviation.
Find the probability that four or more of the twelve patients will suffer SSH.

6 The random variable X has a binomial distribution with mean 12 and variance 3.
Find $P(X \geqslant 14)$.

7 It is given that $Y \sim B(9, p)$ and that the standard deviation of Y is $\frac{9}{10}$.
Find the possible values of p.
For each value of p find $P(Y = 4)$, giving 3 significant figures in your answers.

8 The Post Office claims that 92% of first-class letters are delivered the next day after posting. A company selects 20 letters at random from a large batch of first-class letters in order to determine the number X that were delivered the next day.
Find the expectation μ and the standard deviation σ of X.
Find $P(|X - \mu| < 1)$, and $P(|X - \mu| < 2\sigma)$.

9 The random variable X is such that $X \sim B(n, p)$. It is known that $\frac{\text{Var}(X)}{E(X)} = 0.3$ and that X has mean 10.5.
Determine the values of n and p.

10 The random variable X is such that $X \sim B(n, 0.5)$.
Determine the smallest value of n for which the ratio of the standard deviation of X to the mean of X is less than 1 to 10.

Chapter summary

- **The binomial distribution** applies to situations in which each outcome is either a 'success' or a 'failure'. If n independent trials each have probability p of being a success and X denotes the number of successes, then:
  - $X \sim \mathrm{B}(n, p)$
  - $\mathrm{P}(X = r) = P_r = \binom{n}{r} p^r q^{n-r} \qquad r = 0, 1, \ldots, n$

    where $q = (1 - p)$
  - $\mathrm{E}(X) = np$
  - $\mathrm{Var}(X) = npq$

Exercises 7e (Miscellaneous)

1 The germination of cactus seeds is not easy. From experience Mr Thorn, the expert cactus grower, knows that on average only 40% germinate. An intrepid collector returns from a very dry desert with six seeds of a previously unknown type of cactus.

(i) Determine the probability that only 1 seed germinates.

(ii) Determine the most likely number of germinating seeds.

2 A lorry carrying a large number of boxes of eggs is involved in an accident. Each box contains six eggs. After the accident the contents of a random sample of 100 boxes are examined and the numbers of broken eggs (x) are recorded. The numbers of boxes (n) containing various numbers of broken eggs are given in the table below.

x	0	1	2	3	4	5	6
n	31	37	22	7	2	1	0

From this frequency distribution estimate p, the proportion of broken eggs.

Calculate, correct to 1 d.p., the expected frequencies to be expected from a binomial distribution having this value of p.
(Expected frequencies are obtained by multiplying the theoretical probabilities by the sample size.)

3 A company produces electrical components, some of which are defective. The proportion of defectives is usually low, but if the proportion reached 10% then the company would want to know that this had happened in order to adjust the machine. A random sample of n components is therefore examined.
Given that the proportion of defectives currently being produced is indeed equal to 10%, determine an expression, in terms of n, for the probability that the sample contains no defectives.
Denoting this probability by P_0, determine the smallest value of n for which P_0 is less than 5%.

4 There is room for 53 passengers on flight ZJ142. Tickets are sold at £130. If a ticket is sold, but the would-be passenger is unable to fly, the airline pays back none of the money to the passenger. From past experience it is known that only 95% of ticket purchasers actually fly. If a ticket is sold to a passenger, but no seat is available then there is an average cost to the airline of £200.
By calculating the net revenue to an airline that results from its selling N tickets, for values of N from 53 to 57, determine the value of N that maximises the expected revenue.

5 On average, it is found that the failure rate for germination of geranium seeds, sold in packets of ten, is 0.8 seeds per packet. Find

(*a*) the variance of the number of seeds per packet that fail to germinate,

(*b*) the probability, to 3 decimal places, that a packet chosen at random will contain more than one seed that fails to germinate.

[ULSEB]

6 A boojum is a rare mammal which inhabits tropical seas and spends most of the time under the water. An ecological expedition suspects that there is a boojum in a certain area and attempts to obtain photographic evidence. The technique used is such that if a boojum is present, the probability that it will be visible on any particular photograph is $\frac{1}{4}$. A member of the expedition takes six photographs. If there is a boojum in the area, show that the probability that it will not be visible on any of the photographs is about 0.178.

Five members of the expedition independently search the area, each taking six photographs by the same technique. What is the probability that at least two of them will succeed in photographing the boojum? [SMP]

7 A reader of a magazine enters for a competition in the magazine, in which the competitors have to choose the correct answers to a number of questions. There are five suggested answers for each question, but the reader is completely unskilful and selects an answer at random to each question, so that, for each question, the probability of choosing the right answer is $\frac{1}{5}$. For a competition with 12 questions, find the probability of the reader getting more than 3 correct answers, giving three decimal places in your answer.

[UCLES(P)]

8 A rifle competition is entered by teams of four people. Each person in a team fires one shot at the target. The table below shows the number of points for the number of hits by a team.

Number of hits	0	1	2	3	4
Number of points	0	2	4	8	16

For a particular team, each member of the team independently has probability 0.7 of hitting the target. Find:

(i) the probability of the team hitting the target r times, for each $r = 0, 1, 2, 3, 4$;
(ii) the team's expected points score;
(iii) the team's most likely points score;
(iv) the probability that the team scores more than 6 points. [O&C]

9 In each trial of a random experiment the probability that the event A will occur is 0.6 and the probability that the event B will occur is 0.5. It is known that A and B are independent.

(i) Calculate the probability that at least one of A and B will occur in a single trial.
(ii) Using the tables provided, or otherwise, find the probability that in twenty independent trials the event A will occur exactly twelve times; give your answer correct to three decimal places.

[WJEC]

10 (*a*) A class of 16 pupils consist of 10 girls, 3 of whom are left-handed, and 6 boys, only one of whom is left-handed. Two pupils are to be chosen at random from the class to act as monitors. Calculate the probabilities that the chosen pupils will consist of

(i) one girl and one boy,
(ii) one girl who is left-handed and one boy who is left-handed,
(iii) two left-handed pupils,
(iv) at least one pupil who is left-handed.

(*b*) The probability of a manufactured item being defective is 0.1. A batch consisting of a very large number of the items is inspected as follows. A random sample of five items is chosen. If this sample contains no defective item then the batch is accepted, while if the sample includes 3 or more defectives it is rejected. If the sample includes either 1 or 2 defectives then a second random sample of five items is chosen from the batch. The batch will be accepted if this second sample contains no defective item and will be rejected otherwise. Calculate, correct to three decimal places, the probabilities that

(i) the first sample will result in the batch being accepted,
(ii) the first sample will result in the batch being rejected,
(iii) the second sample will be necessary and will result in the batch being accepted. [WJEC]

8 The Poisson distribution

In the United States there is more space where nobody is than where anybody is. That is what makes America what it is

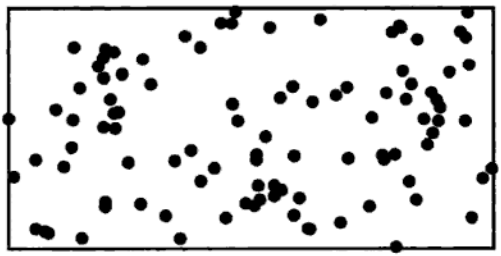
The Geographical History of America, Gertrude Stein

8.1 The Poisson process

In a Poisson distribution (capital P because the distribution is named after Siméon Poisson, of whom more anon) the random variable is a count of events occurring *at random* in regions of time or space. 'At random' here has a very particular and strict definition: the occurrences of the events are required to be distributed through time or space so as to satisfy the following:

- Whether or not an event occurs at a particular point in time or space is independent of what happens elsewhere.
- At all points in time the probability of an event occurring within a small fixed interval of time is the same. This also applies to the occurrence of events in small regions of space.
- There is no chance of two events occurring at precisely the same point in time or space.

Events that obey these requirements are said to be described by a **Poisson process**. Typically, in a spatial Poisson process, there appear to be haphazardly arranged clusters of points as well as wide-open spaces.



A spatial Poisson process

Examples of real-life Poisson processes are the following:

- The points in time at which a given piece of radioactive substance emits a charged particle.
- The points in space occupied by the micro-organisms in a random sample of well-stirred water taken from a pond.

A Poisson distribution describes the probabilities of the associated counts:

- The number of particles emitted in a minute by the radioactive substance.
- The number of micro-organisms in 1 ml of pond water.

There are many other situations in which a Poisson distribution provides a good approximation for small periods of time or for regions of space that are not too large:

- The number of phone calls received on a randomly chosen day.
- The number of cars passing in a randomly chosen five-minute period on a road with no traffic lights or long queues (assuming such a road exists!).
- The number of currants in a randomly chosen currant bun.
- The number of accidents in a factory during a randomly chosen week.
- The number of typing errors on a randomly chosen page of a manuscript.
- The number of daisies in a randomly chosen square metre of playing field.

Two somewhat bloodthirsty classic examples that appear prominently in older Statistics books are:

- The numbers of bomb craters in equal-sized areas of wartime London.
- The numbers of deaths of cavalrymen caused by horse kicks. These data were collected with military precision each year for each of the various Prussian army corps!

8.2 The form of the distribution

The formula for a Poisson distribution involves one of the 'magic numbers' of mathematics, the number $e = 2.718\,28\ldots$

$$P_r = P(X = r) = \frac{\lambda^r e^{-\lambda}}{r!} \qquad r = 0, 1, 2, \ldots \tag{8.1}$$

where λ, pronounced 'lambda', is a positive number. For $r = 0$ this becomes:

$$P_0 = e^{-\lambda}$$

since $0! = 1$ and $\lambda^0 = 1$ for all values of λ.

For a Poisson distribution:

$$E(X) = \mathrm{Var}(X) = \lambda$$

One way of defining the value of e is via the expression:

$$e^c = 1 + \frac{c}{1!} + \frac{c^2}{2!} + \frac{c^3}{3!} + \frac{c^4}{4!} + \cdots$$

so that:

$$e = e^1 = 1 + \frac{1}{1} + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \cdots$$

We can use the definition of e^c to verify that the probabilities of the Poisson distribution do indeed sum to 1:

$$\begin{aligned} P_0 + P_1 + P_2 + \cdots &= \left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \cdots\right) e^{-\lambda} \\ &= e^{\lambda} e^{-\lambda} \\ &= 1 \end{aligned}$$

Notes

- The notation $X \sim \mathrm{Po}(\lambda)$ may be used as a shorthand for 'X has a Poisson distribution with parameter λ'.
- Any symbol could be used for the parameter of the distribution. Some authorities use μ. We have used λ so as to avoid the potential confusion of having a variance equal to μ.
- Almost all scientific calculators have a button marked e^x which enables easy calculation of these probabilities without the need to remember the value of e.

Example 1

Between 6 p.m. and 7 p.m., Directory Enquiries receives calls at the rate of 2 per minute.
Assuming that the calls arrive at random points in time, determine the probability that:
(i) 4 calls arrive in a randomly chosen minute,
(ii) 6 calls arrive in a randomly chosen two-minute period.

Since calls arrive at random points in time, a Poisson process is being described.

(i) Let X be the number of calls that arrive in a randomly chosen minute. Since the mean number of calls per 1-minute period is 2, we put $\lambda = 2$. Hence:

$$P(X = 4) = \frac{2^4 e^{-2}}{4!} = 0.090 \text{ (to 3 d.p.)}$$

(ii) Let Y be the number of calls that arrive in a randomly chosen two-minute period. The mean number of calls per two-minute period is 4, so we put $\lambda = 4$. Hence:

$$P(Y = 6) = \frac{4^6 e^{-4}}{6!} = 0.104 \text{ (to 3 d.p.)}$$

Example 2

In a certain disease a small proportion of the red blood corpuscles display a tell-tale characteristic. A test consists of taking a random sample of 2 ml of a person's blood and counting the number of distinctive corpuscles. A count of five or more is taken to be an indication that the person has the disease. Mrs Wretched has the disease: the mean number of distinctive corpuscles in her blood is 1.6 per ml.
Determine the probability that a randomly chosen sample of 2 ml of her blood will contain five or more of the distinctive corpuscles.
Does the test appear to be a good one?

Let X be the number of distinctive corpuscles in 2 ml of Mrs Wretched's blood. On average, 2 ml of her blood contains 3.2 distinctive corpuscles. Assuming that the corpuscles are haphazardly distributed through her blood, the random variable X has a Poisson distribution with $\lambda = 3.2$. We want $P(X \geqslant 5)$. Since there is no upper bound to the possible values for X, it is 'infinitely' simpler to calculate the probability of the complementary event, $(X \leqslant 4)$:

$$\begin{aligned} P(X \leqslant 4) &= P_0 + \cdots + P_4 \\ &= \left(1 + \frac{(3.2)^1}{1!} + \frac{(3.2)^2}{2!} + \frac{(3.2)^3}{3!} + \frac{(3.2)^4}{4!}\right) e^{-3.2} \\ &= 0.781 \text{ (to 3 d.p.)} \end{aligned}$$

Hence: $P(X \geqslant 5) = 1 - 0.781 = 0.219$

Hence the probability that 2 ml of Mrs Wretched's blood contains five or more of the distinctive corpuscles is 0.219 (to 3 d.p.). There is a chance of nearly 80% that the test will fail to suggest that Mrs Wretched has the disease – the test is not very good.

Exercises 8a

In Questions **1–5**, the random variable X has a Poisson distribution with mean λ.

1 Given that $\lambda = 2$, find (i) $P(X = 0)$, (ii) $P(X = 1)$, (iii) $P(X = 2)$, (iv) $P(X \leqslant 2)$, (v) $P(X \geqslant 2)$.

2 Given that $\lambda = 0.5$, find (i) $P(X < 3)$, (ii) $P(2 \leqslant X \leqslant 4)$, (iii) $P(X \geqslant 3)$.

3 Given that $\lambda = 5$, find (i) $P(X = 5)$, (ii) $P(X < 5)$, (iii) $P(X > 5)$.

4 Given that $\lambda = 1.4$ find $P(X = 1, 3 \text{ or } 5)$.

5 Given that $\lambda = 2.1$ and $P(X = r) = 0.1890$, find the value of r.

6 The number of currants in a randomly chosen currant bun can be modelled as a random variable having a Poisson distribution with mean 5.6.
Find the probability that a randomly chosen currant bun contains (i) fewer than 4 currants, (ii) more than 4 currants.

7 The number of accidents in a randomly chosen week at a factory can be modelled by a Poisson distribution with mean 0.7.
Find the probability that there are more than two accidents in a randomly chosen week.

8 The number of emergency calls received by a Gas Board in a randomly chosen day can be modelled by a Poisson distribution with mean 3.4.
Find the probability that, in a randomly chosen day, the number of emergency calls received is 5 or more.

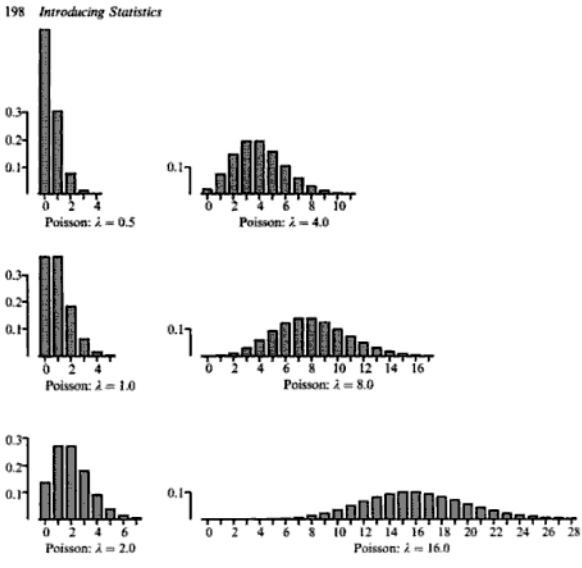
9 Buttercups are randomly distributed across a playing field with the probability of a randomly chosen square metre of the field containing precisely r buttercups being:

$$\frac{2^r e^{-2}}{r!}, \qquad r = 0, 1, 2, \ldots$$

Determine the probability that a randomly chosen region of area 0.5 square metres contains precisely one buttercup.

8.3 The shape of a Poisson distribution

When $\lambda < 1$ the distribution has mode at $x = 0$ and is very skewed. As λ increases so the distribution takes on a more symmetrical appearance. Note that the diagrams are truncated – in each case it is *possible* for the Poisson random variable to take a larger value than those shown. However, although the range of possible values is infinite, the results given later in Section 10.12 (p. 283) suggest that in practice 95% of values will lie between $\lambda - 2\sqrt{\lambda}$ and $\lambda + 2\sqrt{\lambda}$ (i.e. in the range: mean ± 2 standard deviations).



Practical

A chess board and a tube of Smarties® (or something similar) are required for this project. The idea is to toss the Smarties one at a time on to the chess board and, when all are on the board (you may need several attempts!) to count the numbers of Smarties in each of the 64 squares. Providing the board is reasonably large and you didn't cheat, the arrangement of the Smarties should approximate a spatial Poisson process.

Check by drawing a bar chart of the 64 observations and calculating their mean and variance, which should be reasonably similar.

Project

Providing a road is not so busy that queues of traffic form, traffic flow may be modelled by a Poisson process. To investigate this, choose a reasonably busy road and count the numbers of cars (or lorries, or bicycles, or whatever) that pass in a particular direction in a period of one minute. Repeat for a complete half-hour. It will be easier to work in pairs, with one person counting and the other timing and recording. Choose a dry day and warm clothing!

Represent your results using a bar chart.

Calculate the sample mean and variance.

If the stream of cars does form a Poisson process then the mean and variance should be quite similar. If the variance is much larger than the mean then this will suggest that there is appreciable clumping of the cars due to slower cars holding up faster ones or to the presence of a nearby roundabout or traffic light.

Computer project

Write a computer program to generate a sequence of 10000 random numbers, each with a value between 0 and 1. Test each number to see if its value is less than 0.002 (a 'success'). If the mth number in the sequence is a 'success' record the value of m. You should end up with about 20 'successes'. (Why?). Now draw a line 10 inches long on a piece of paper, with one end corresponding to the start of your simulation. Suppose your first 'success' occurred on the 876th random number. Illustrate this by placing a dot 0.876 inches (approximately!) from the start of the line. Suppose now that the second 'success' is obtained on the 83rd subsequent random number. This will be illustrated by a second dot at a point 0.959 inches (since 876 + 83 = 959) from the start of the line. Continue to add subsequent 'successes' in this way.

When all the 'successes' have been marked you will have an illustration of a linear Poisson process somewhat similar to that shown earlier. Computer wizards, whose computers have graphical capabilities, may wish to get the computer to draw this display.

Calculator practice

If you have a graphical calculator with programming facilities then the previous computer project can be carried out (more slowly!) on the calculator. You could arrange for each line of the screen to correspond to a single 'time sequence'. When the program stops, your display of alternative time sequences will be indistinguishable from a realisation of a spatial Poisson process.

8.4 Tables for Poisson distributions

As with tables of the binomial distribution, tables for Poisson distributions may occur in a variety of forms. The tables provided in the Appendix (p. 438) are tables of $P(X \leqslant r)$, for various values of λ. Our tables give probabilities correct to four decimal places. Cumulative probabilities exceeding 0.999 95 are omitted from the table. A (rearranged) extract from the table is given below:

λ	r								
	0	1	2	3	4	5	6	7	8
1.4	0.2466	0.5918	0.8335	0.9463	0.9857	0.9968	0.9994	0.9999	

This table refers to the case $\lambda = 1.4$ and shows $P(X \leqslant r)$ for $r = 0, 1, \ldots, 7$. Thus $P(X \leqslant 4) = 0.9857$. There is no entry corresponding to $r = 8$ since the cumulative probability exceeds 0.999 95 and is therefore 1.0000 to 4 d.p.

In order to use the tables to find probabilities for individual values of r, or for other types of inequality, we need the following relations:

$$P(X < r) = P(X \leqslant r - 1)$$
$$P_r = P(X = r) = P(X \leqslant r) - P(X \leqslant r - 1)$$
$$P(X > r) = 1 - P(X \leqslant r)$$
$$P(X \geqslant r) = 1 - P(X \leqslant r - 1)$$

Notes

- It is easy to confuse $P(X < r)$ with $P(X \leqslant r)$, so questions should always be read very carefully!
- Inevitably the tables do not provide for *every* value of λ. If a probability is required for a value of λ that is not included in the tables then, if possible, it should be calculated from the formula rather than by interpolation in the tables.
- It is important to be familiar with the tables that are available for your examinations. The tables given in this book are just a convenience!

Example 3

Tadpoles are scattered randomly through a pond at the rate of 14 per litre. A random sample of 0.1 litre is examined.
What is the probability that it will contain more than 3 tadpoles?

Assuming a Poisson distribution (since the tadpoles are distributed at random in space) with mean 1.4 per 0.1 litre, we require:

$$1 - (P_0 + P_1 + P_2 + P_3)$$

which is:

$$1 - P(X \leqslant 3) = 1 - 0.9463$$

and so the probability that the sample contains more than 3 tadpoles is 0.054 (to 3 d.p.).

Example 4

The random variable Y has a Poisson distribution with mean 1.4. Determine the probability that Y takes a value greater than 4, but less than 7.

The question is asking for $P_5 + P_6$.
Using the cumulative tables we calculate this as:

$$P(Y \leqslant 6) - P(Y \leqslant 4) = 0.9994 - 0.9857$$

and so the probability that Y takes a value greater than 4, but less than 7, is 0.014 (to 3 d.p.).

Example 5

Use tables of cumulative Poisson probabilities to determine $(3 < X \leqslant 7)$, where X has a Poisson distribution with mean 1.4.

The question requires $P_4 + P_5 + P_6 + P_7$, which we calculate as:

$$P(X \leqslant 7) - P(X \leqslant 3) = 0.9999 - 0.9463.$$

The required probability is 0.054 (to 3 d.p.).

Exercises 8b

In Questions **1–5**, the random variable X has a Poisson distribution with mean λ. Use the tables that you will use in your examination, or the tables in the Appendix at the back of this book, to answer the following questions.

1 Given that $\lambda = 3$ find (i) $P(X \leqslant 5)$, (ii) $P(X < 7)$.

2 Given that $\lambda = 0.9$ find (i) $P(X \geqslant 3)$, (ii) $P(X > 4)$.

3 Given that $\lambda = 1.2$ find (i) $P(2 < X < 5)$, (ii) $P(2 \leqslant X \leqslant 5)$.

4 Find λ given that $P(X \leqslant 5) = 0.9896$.

5 Find λ given that $P(X > 4) = 0.0527$.

6 Weak spots occur at random in the manufacture of a certain cable at an average rate of 1 per 100 metres.
If X represents the number of weak spots in 100 metres of cable, write down the distribution of X.
Lengths of this cable are wound onto drums. Each drum carries 50 metres of cable.
Find the probability that a drum will have 3 or more weak spots.
A contractor buys five such drums.
Find the probability that two have just one weak spot and the other three have none.
[AEB(P)91]

Siméon Denis Poisson (1781–1840) was a French mathematician who is variously described as lively and extremely hard-working. His principal interest lay in aspects of mathematical physics. His major work on probability was entitled *Researches on the probability of criminal and civil verdicts*. In this long book (over 400 pages) only about one page is devoted to the derivation of the distribution that bears his name! Poisson derived the distribution as a limiting form of the binomial (see below). He is quoted as having said that 'Life is good for only two things: to study mathematics and to teach it'. No comment!

8.5 The Poisson approximation to the binomial

If X has a binomial distribution with parameters n and p, and if n is large and p is near 0, then the distribution of X is closely approximated by a Poisson distribution with mean np.

Notes

- The approximation should only be used when it is not feasible to calculate the required probability exactly.
- The usual guidelines for the use of the approximation are that n should be greater than 50 and that p should be less than 0.1. These are not strict rules. All that can be said with confidence is that:
  - the smaller p, the better.
  - the larger n, the better.

Example 6

The discrete random variable X has a binomial distribution with $n = 60$ and $p = 0.02$.
Determine $P(X = 1)$ (i) exactly, (ii) using a Poisson approximation.

(i) The exact binomial probability is given by:

$$P(X = 1) = \binom{60}{1}(0.02)^1(0.98)^{59} = 0.364 \text{ (to 3 d.p.)}$$

(ii) Since n is quite large and p is small, we can expect the Poisson approximation to be quite accurate. Setting $\lambda = np = 60 \times 0.02 = 1.2$, we have:

$$P(X = 1) \approx \frac{1.2}{1!}e^{-1.2} = 0.361 \text{ (to 3 d.p.)}$$

The approximation is indeed an accurate one.

Example 7

Past experience suggests that 0.4% of peaches show signs of mildew on arrival at the market. Occasionally, if storage conditions are faulty, the proportion of mildewed peaches may be much higher than this. Assuming that the conditions of individual fruit are independent of one another, and that the proportion of mildewed peaches is the usual 0.4%, determine the probability that a carton of 250 individually packed peaches contains more than three that show signs of mildew. What conclusions would you draw if a randomly chosen carton was found to contain 5 mildewed peaches?

This is a binomial situation. The parameters are $n = 250$ and $p = \frac{0.4}{100} = 0.004$ (not 0.4 which corresponds to 40%, and would mean that nearly half were mildewed!). The question asks for 'more than three', which means 4, 5, ..., 250. It is much easier to consider the complementary event that there are 0, 1, 2 or 3 mildewed peaches. Although it is feasible to calculate the required probability directly, using the binomial distribution, it is much easier to use the Poisson approximation with $\lambda = np = 250 \times 0.004 = 1$. The individual probabilities are tabulated below:

No. of peaches	0	1	2	3	3 or less
Exact binomial prob.	0.3671	0.3686	0.1843	0.0612	0.9813
Poisson approximation	0.3679	0.3679	0.1839	0.0613	0.9810

There is about a 2% chance that there are more than three peaches showing signs of mildew.

If a randomly sampled carton was found to contain five mildewed peaches then this would strongly suggest that the storage conditions had been faulty.

Practical

This is another practical involving the rolling of dice. Each member of the class should roll a die twice (or two dice once) and report a 'success' if two sixes are obtained. The total number of successes for the class should be recorded and the exercise repeated twenty times so as to give twenty observations from a binomial distribution with parameters n (the class size) and $p = \frac{1}{36}$.

Compare the observed relative frequencies with the exact binomial probabilities.

Calculate the approximating Poisson probabilities and compare them with the exact values.

Practical

Ordinary packs of playing cards are required for this practical. The event of interest is that a single card drawn from a pack is the Ace of Spades. Each class member should have 26 attempts at striking lucky, with the card chosen being returned to the pack, and the pack being shuffled between attempts. Since $n = 26$ and $p = \frac{1}{52}$, the approximating Poisson distribution has $\lambda = \frac{1}{2}$. About 60% of the class should not see the Ace of Spades at all, but about 9% (where do these percentages come from?) should see the Ace more than once.

Exercises 8c

In Questions 1–3 use the Poisson approximation to find the required probability concerning the random variable X which has a binomial distribution with parameters n and p.

1 Given that $n = 40$, $p = 0.1$, find (i) $P(X \leqslant 3)$, (ii) $P(X \geqslant 3)$.

2 Given that $n = 100$, $p = 0.02$, find (i) $P(X \geqslant 2)$, (ii) $P(X < 4)$.

3 Given that $n = 55$, $p = \frac{1}{11}$, find $P(3 \leqslant X \leqslant 6)$.

4 Screws are packed in boxes of 200. For each screw the probability that it is faulty is 0.4%. Using a suitable approximation, find the probability that a box contains at most two faulty screws.

5 For a beginner taking photos, the probability that a photo is 'useless' is 0.1 and the probability that it is 'brilliant' is 0.05. The beginner takes 72 photos.
Use a suitable approximation to find the probability that:
(i) at least 3 photos are brilliant,
(ii) at most 3 photos are useless.

6 The proportion of red sports cars is 1 per 200 cars in the country as a whole. There are 500 cars in a car park.
Assuming these to be a random sample from the population, use a Poisson approximation to determine the probability that there are exactly 5 red sports cars in the car park.

7 A rare type of error in the printing of postage stamps is such that in a random sample of 1000 stamps there will be on average 2 stamps displaying the error. Using a Poisson approximation, calculate the probability of there being exactly one stamp displaying the error in a random sample of 100 stamps.
For a sample of this size, state the mean and variance of the number of stamps displaying the error.

8 Thirty digits are taken at random from a table of random numbers. Find the exact binomial probabilities of obtaining (i) one seven, (ii) two sevens, (iii) three sevens. Find the same probabilities using the Poisson approximation and compare the results.

9 A charity runs a prize draw every week, and each person who buys a ticket has a chance of 1 in 1000 of winning the prize. A contributor buys a ticket each week for 50 weeks.
Using a suitable approximation, find:
(i) the probability that she wins at least one prize
(ii) the smallest number of weeks that she must buy tickets in order that the probability of winning at least one prize exceeds 0.9.

10 A machine produces resistors of which 99% are up to standard. They are packed in boxes each containing 200 resistors. Using a suitable approximation, find the probability that a randomly chosen box contains at least 198 resistors that are up to standard.

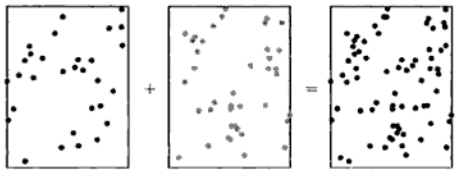
11 A hockey team consists of 11 players. It may be assumed that, on every occasion, the probability of any one of the regular members of the team being unavailable for selection is 0.15, independently of all the other members. Calculate, giving three significant figures in your answers, the probability that, on a particular occasion,
(i) exactly one regular member is unavailable,
(ii) more than two regular members are unavailable.
Taking the probability that more than 3 regular members are unavailable as 0.07, write down, for a season in which 50 matches are played, the expected value of the number of matches for which more than 3 regular members are unavailable.
Use a suitable Poisson distribution to find an approximation for the probability that, in the course of a season, more than 3 regular players will be unavailable at most twice. [UCLES]

8.6 Sums of independent Poisson random variables

If X and Y are independent Poisson random variables with parameters λ and μ, respectively, then the random variable Z, defined by $Z = X + Y$, is a Poisson random variable with parameter $\lambda + \mu$.

A direct proof of this result is tedious, but the result is obvious once we consider the Poisson process background to the distribution.

Mixing together two random patterns produces another random pattern!



A Poisson distribution refers to counts of 'events' scattered at random in time or space. Suppose we have a collection of m red objects and n green objects which are all identical apart from their colour. We scatter these objects at random over a square region of area A. Focusing on the red objects alone we see a spatial Poisson process with rate m per unit area. Likewise, focusing on the green objects, we see a random arrangement with a rate of n per unit area. A colour-blind person would see a combined set of randomly distributed objects at a rate of $(m+n)$ objects per unit area.

Example 8

An observer is standing beside a road. Both cars and lorries pass the observer at random points in time. On average there are 300 cars per hour, while the mean time between lorries is five minutes. Determine the probability that exactly 6 vehicles pass the observer in a one-minute period.

Since the question refers to 'random points in time' a Poisson distribution is appropriate both for cars and for lorries. The mean rate for lorries is 12 an hour, so the combined rate is 312 vehicles per hour, which corresponds to 5.2 vehicles per minute. The required probability is therefore:

$$\frac{(5.2)^6 e^{-5.2}}{6!} = 0.151 \text{ (to 3 d.p.)}$$

Chapter summary

- The Poisson distribution refers to the counts of items that occur at **random points in time or space** (a Poisson process).
- **Distribution**: $P_r = \dfrac{\lambda^r e^{-\lambda}}{r!} \qquad r = 0, 1, 2, \ldots$
- **Expectation and variance**: $\mathrm{E}(X) = \mathrm{Var}(X) = \lambda$
- **Approximation of binomial**: A binomial distribution with parameters n (large) and p (small) may be approximated by a Poisson distribution with parameter np.
- **Additivity**: If X and Y are independent Poisson random variables, with parameters λ and μ respectively, then $X + Y$ has a Poisson distribution with parameter $\lambda + \mu$.

Exercises 8d (Miscellaneous)

1 The numbers of emissions per minute from two radioactive objects A and B are independent Poisson variables with means 0.65 and 0.45 respectively.
Find the probabilities that:
(i) in a period of three minutes there are at least three emissions from A,
(ii) in a period of two minutes there is a total of less than four emissions from A and B together.

2 The number of customers per hour entering a jeweller's shop has a Poisson distribution. For the first hour after opening the mean is 0.7 per hour and for the next three hours the mean is 1.3 per hour.
Find the probability that there are between 4 and 6 (inclusive) customers entering the shop in the first four hours.

3 In a particular form of cancer, deformed blood corpuscles occur at random at the rate of 10 per 1000 corpuscles.
Use an appropriate approximation to determine the probability that a random sample of 200 corpuscles taken from a cancerous area will contain no deformed corpuscles.
How large a sample should be taken in order to be 99% certain of there being at least one deformed corpuscle in the sample?

4 The number of telephone calls arriving per minute at a small telephone exchange has a Poisson distribution with mean 2.25. Find, correct to three decimal places, the probability that
(i) exactly 2 calls arrive in a minute,
(ii) more than 4 calls arrive in a period of 2 minutes. [WJEC]

5 The numbers of emissions per minute from two radioactive substances, A and B, are independent and have Poisson distributions with means 2.8 and 3.25, respectively.
Find, correct to three decimal places, the probabilities that in a period of one minute there will be
(i) exactly three emissions from A,
(ii) one emission from one of the two substances and two emissions from the other substance. [WJEC]

6 Independently for each page of a printed book the number of errors occurring has a Poisson distribution with mean 0.2. Find, correct to three decimal places, the probabilities that
(i) the first page will contain no error,
(ii) four of the first five pages will contain no error,
(iii) the first error will occur on the third page. [WJEC]

7 (i) The discrete random variable X has probability function given by

$$p(x) = \begin{cases} (\frac{1}{2})^x & x = 1,2,3,4,5, \\ C & x = 6, \\ 0 & \text{otherwise,} \end{cases}$$

where C is a constant.
Determine the value of C and hence the mode and arithmetic mean of X.

(ii) A process for making plate glass produces small bubbles (imperfections) scattered at random in the glass, at an average rate of four small bubbles per $10\,m^2$.
Assuming a Poisson model for the number of small bubbles, determine to 3 decimal places, the probability that a piece of glass 2.2m × 3.0m will contain

(*a*) exactly two small bubbles,

(*b*) at least one small bubble,

(*c*) at most two small bubbles.

Show that the probability that five pieces of glass, each 2.5m by 2.0m will all be free of bubbles is e^{-10}.
Find, to 3 decimal places, the probability that five pieces of glass, each 2.5m by 2.0m, will contain a total of at least ten small bubbles.
[ULSEB]

8 Serious delays on a certain railway line occur at random, at an average rate of one per week.
Show that the probability of at least four serious delays occurring during a particular 4-week period is 0.567, correct to 3 decimal places.
Taking a year to consist of thirteen 4-week periods, find the probability that, in a particular year, there are at least ten of these 4-week periods during which at least 4 serious delays occur.
Given that the probability of at least one serious delay occurring in a period of n weeks is greater than 0.995, find the least possible integer value of n. [UCLES]

9 In a certain country it is known that 35% of the adult population have some knowledge of a foreign language. If 10 adults are chosen at random from this population, find the probability that

(i) at least one of those chosen will have some knowledge of a foreign language,

(ii) at most three of those chosen will have some knowledge of a foreign language.

For one particular foreign language, only a very small proportion r% of the adult population have some knowledge of it. It is required to select n adults at random, where n is chosen so that the probability of obtaining at least one adult having some knowledge of the language is to be 0.99, as nearly as possible. Use a suitable Poisson approximation to show that $n \approx \dfrac{460.5}{r}$.
For the case when $r = \frac{1}{2}$ and $n = 921$, find the probability that precisely four adults in the sample will have some knowledge of the language. [UCLES]

10 A biased cubical die is such that the probability of any one face landing uppermost is proportional to the number on that face. Thus, if X denotes the score obtained in one throw of this die,

$$P(X = r) = kr, \qquad r = 1,2,3,4,5,6,$$

where k is a constant.
(i) Find the value of k.
(ii) Show that $E(X) = 4\frac{1}{3}$, and find $Var(X)$.
This die is thrown 80 times, and the scores are noted. Use an appropriate Poisson distribution to estimate the probability of at least four "ones" being scored. [UCLES]

11 The number of telephone calls X made by a daughter D to her mother in each week has a Poisson distribution with mean 2, whilst the number of telephone calls Y made by her brother B in each week has a Poisson distribution with mean 1. Show that

$$(n+1)P(X = n+1) = 2P(X = n)$$

and

$$(n+1)P(Y = n+1) = P(Y = n),$$

$n = 0, 1, 2 \ldots$.

Assuming that X and Y are independent, find the probability, to 2 decimal places, that in a given week,

(a) neither B nor D makes a call,

(b) B and D make an equal number of calls not exceeding 2 calls each,

(c) B makes less than 4 calls, but makes more calls than D. [ULSEB]

12 (a) The independent Poisson random variables X and Y have means of 2.5 and 1.5, respectively. Obtain the mean and variance of the random variables
(i) $X - Y$, (ii) $2X + 5$.
For each of these random variables give a reason why the distribution is not Poisson.

(b) A car salesman receives £60 commission for each *new* car that he sells and £40 for each *used* car that he sells. The weekly number of *new* cars that he sells has a Poisson distribution with mean 3 and, independently, the weekly number of *used* cars that he sells has a Poisson distribution with mean 2.
(i) Determine the probability that he sells more than two *new* cars in a week.
(ii) Determine the probability that he sells no more than one car in a week.
(iii) Determine the probability that his commission in a week is exactly £100.
(iv) Calculate the mean and standard deviation of the salesman's weekly commission. [JMB]

13 Independently for each page the number of typing errors per page in the first draft of a novel has a Poisson distribution with mean 0.4.

(*a*) Calculate, correct to five decimal places, the probabilities that
(i) a randomly chosen page will contain no error,
(ii) a randomly chosen page will contain 2 or more errors,
(iii) the third of three randomly chosen pages will be the first to contain an error.

(*b*) Write down an expression for the probability that each of n randomly chosen pages will contain no error. Hence find the largest n for which there is a probability of at least 0.1 that each of the n pages contains no error.
Independently for each page the number, Y, of typing errors per page in the first draft of a Mathematics textbook also has a Poisson distribution.

(*c*) Given that $P(Y = 2) = 2P(Y = 3)$
(i) find $E(Y)$,
(ii) show that $P(Y = 5) = 4P(Y = 6)$.

(*d*) One page is chosen at random from the first draft of the novel and one page is chosen at random from the first draft of the Mathematics textbook. Calculate, correct to three decimal places, the probability that exactly one of the two chosen pages will contain no error. [WJEC]

14 In a double-sampling scheme an initial sample of 50 items is taken from the large batch under investigation and the number m of defectives is noted. If $m = 0$, the whole batch is accepted without further testing. If $m = 1$ or 2, a second sample, this time of 100 items, is taken from the batch and the number n of defectives noted. The whole batch is now accepted if $m + n \leqslant 4$. In all other cases the batch is rejected. For a batch with 1% defective items, use suitable Poisson approximations to estimate
(i) the probability of the batch being accepted,
(ii) the expected number of items sampled. [SMP]

15 In a certain area, the probability of a randomly selected cow dying from 'mad cow' disease is 0.04.
(i) Calculate the probability that in a random sample of 18 cows exactly 2 will die from the disease.
(ii) Find the probability that in a random sample of 20 cows more than 2 will die from the disease.
(iii) Find the probability that in a random sample of 50 cows between 2 and 5 (inclusive) will die from the disease.
(iv) When a random sample of n cows is taken, the probability that at least one cow will die from the disease exceeds 0.99. Find the smallest value of n.
(v) Use a distributional approximation to find the probability that in a random sample of 150 cows fewer than 7 will die from the disease. [WJEC]

16 Manufactured articles are packed in boxes each containing 200 articles, and, on average, $1\frac{1}{2}\%$ of all articles manufactured are defective. A box which contains 4 or more defective articles is substandard. Using a suitable approximation, show that the probability that a randomly chosen box will be substandard is 0.353, correct to three decimal places.
A lorry-load consists of 16 boxes, randomly chosen. Find the probability that a lorry-load will include at most 2 boxes that are

(*continued*)

substandard, giving three decimal places in your answer.
A warehouse holds 100 lorry-loads. Show that, correct to two decimal places, the probability that exactly one of the lorry-loads in the warehouse will include at most 2 substandard boxes is 0.06. [UCLES]

17 A randomly chosen doctor in general practice sees, on average, one case of a broken nose per year and each case is independent of other similar cases.
(i) Regarding a month as a twelfth part of a year,
(*a*) show that the probability that, between them, three such doctors see no cases of a broken nose in a period of one month is 0.779, correct to three significant figures,
(*b*) find the variance of the number of cases seen by three such doctors in a period of six months.
(ii) Find the probability that, between them, three such doctors see at least three cases in one year.
(iii) Find the probability that, of three such doctors, one sees three cases and the other two see no cases in one year. [UCLES]

9 Continuous random variables

I've measured it from side to side: 'Tis three feet long, and two feet wide

The Thorn, William Wordsworth

Chapters 5 to 8 have focused on discrete random variables: quantities whose values are unpredictable but for which a list of the possible values can be made. Continuous random variables differ in that no such list is feasible, though the range of values can be described. Here are some examples:

Continuous random variable	Possible range of values
The height of a randomly chosen 18-year-old male student	1.3 m to 2.3 m
The true mass of a '1 kg' bag of sugar	990 g to 1010 g
The time interval between successive earthquakes of magnitude >7 on the Richter scale	Any (positive) time

The measurements all refer to **physical** quantities. The number of distinct values is limited only by the inefficiency of our measuring instruments. Since there are an uncountable number of possible values that a continuous random variable might take, the probability of any particular value is zero. Instead, we assign probabilities to (arbitrarily small) ranges of values.

If a continuous random variable is measured rather inaccurately, then we will treat it as though it is a discrete random variable.

Age of randomly chosen male (aged under 100) measured to nearest 10 years.	→	Treat as a discrete random variable with 11 categories.

Conversely, if a discrete random variable has a great many possible outcomes, then we may treat it as though it is a continuous random variable

Mark on exam paper (An integer in the range 0 to 100.)	→	Treat as a continuous variable on the interval [0, 100].

Because of this easy transition between the two types of variable, the ideas of expectation and the formulae interrelating expectations that were derived in Chapters 5 and 6 carry over to continuous variables – more of this anon!

9.1 Histograms and sample size

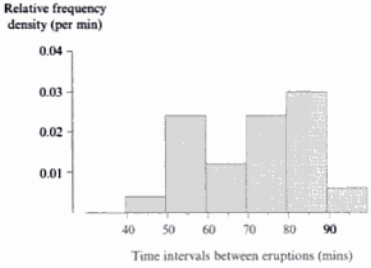
In Chapter 1 the histogram was introduced as being the appropriate method for displaying data on a continuous variable. The crucial part of the instructions for drawing a histogram was that *area should be proportional to frequency*. We now develop that idea by requiring that:

area should be proportional to *relative* frequency.

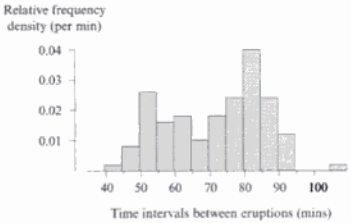
As an illustration we consider some data concerning the geyser known as 'Old Faithful', which is situated in Yellowstone National Park in Wyoming, USA. This geyser is a great tourist attraction because of the regularity of its eruptions of steam. In August 1985 the geyser was watched continuously for a fortnight, with the times between its eruptions being recorded to the nearest minute. The first 50 times are shown below.

80	71	57	80	75	77	60	86	77	56
81	50	89	54	90	73	60	83	65	82
84	54	85	58	79	57	88	68	76	78
74	85	75	65	76	58	91	50	87	48
93	54	86	53	78	52	83	60	87	49

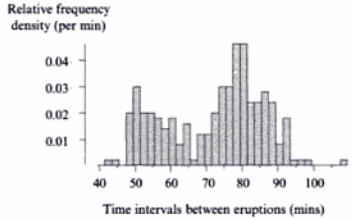
Using classes of width 10 minutes (with class boundaries at 39.5, 49.5, etc.) we can represent these data using a histogram. There are just two observations out of the 50 that fall in the 39.5–49.5 class, so the relative frequency density for that class is $\frac{2}{50} = 0.04$. Since the class width is 10 minutes, the relative frequency density is $\frac{0.04}{10} = 0.004$ per minute. The histogram looks like this:



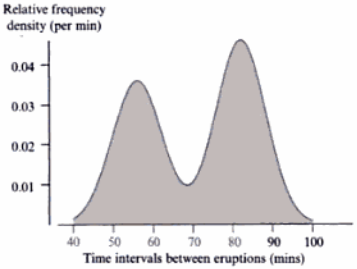
The histogram has a fairly chunky appearance! We now increase the sample size to 100 observations (twice the original size) and illustrate the combined set of data using classes of width 5 minutes (half the original size). With the same vertical scale, the area of the shaded region is the same as before.



Finally, we add in a further 150 observations, raising the total to 250, and now illustrate it using classes that are one-fifth of the original width, so that the area of the shaded region remains the same once again.



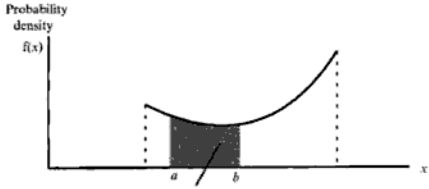
Comparing the sequence of histograms it is easy to see that, as we increase the amount of data, so we increase the precision with which we can see the outline of the histogram. What would happen if we had not 250 observations, but 2500 or 25 000? There would still be the odd bit of random variation, but it seems likely that the dominating effect would be of a histogram with a remarkably smooth outline. With a very large sample (assuming 'Old Faithful' was still working faithfully!) we might obtain a diagram that appeared to be outlined by a smooth curve something like this:



It is clear that 'Old Faithful' behaves in a rather odd fashion! The periods between eruptions are either short (around 50–60 minutes) or long (around 75–90 minutes), with durations of around 66 minutes being rather unusual.

9.2 The probability density function, f

The data from 'Old Faithful' suggested a general result: as we allow the sample size to increase (with correspondingly narrower class intervals), the outline of a histogram will usually converge on a smooth curve. The areas of the individual sections of a histogram represent relative frequencies. We know that as the sample size increases so sample relative frequencies approach the corresponding population probabilities. The area of any section under the curve therefore represents a probability.



Probability that X lies between a and b

When the curve is close to the x-axis the probability associated with a unit range of x is small, whereas when the curve is distant from the axis, the probability is much larger. The height of the curve represents the rate at which probability is accumulated as we move along the x-axis. The curve is the graph of the **probability density function**, written as **pdf** for short and the function is usually denoted by the letter f.

Properties of the pdf

We already know what these are!

1 Since we cannot have negative probabilities, the graph of f cannot dip below the x-axis:

$$\mathrm{f}(x) \geqslant 0 \tag{9.1}$$

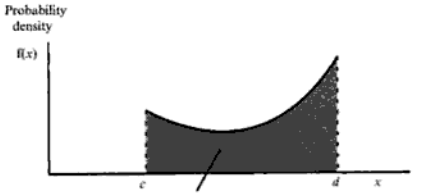
2 The probability of X taking a value in the interval (a, b) is given by the corresponding area. Since the area between any curve and the x-axis is given by the integral of that curve with respect to x, we therefore have:

$$\mathrm{P}(a < X < b) = \int_a^b \mathrm{f}(x)\,\mathrm{d}x \tag{9.2}$$

3 The total of a set of relative frequencies is, by definition, equal to 1. The same is true for probabilities. The total area between the graph of f(x) and the x-axis is therefore 1.

$$\int_c^d \mathrm{f}(x)\,\mathrm{d}x = 1 \tag{9.3}$$

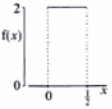
where the limits of the integral are the end-points of the interval for which f is non-zero.



Total area equals 1

Notes

- Suppose k is somewhere between c and d, and let a be just less than k and let b be just greater than k. As a and b get closer to k, the value of the integral in Equation (9.2) approaches zero, so in the limit, $P(X = k) = 0$. This is an entirely general result and implies that:
  – we need not be fussy about whether we write $P(X < k)$ or $P(X \leqslant k)$, since the two have the same value.
- If f has a unique maximum when $x = M$, then M is called the **mode**.
  – Often the mode can be located by examination of a sketch of $f(x)$.
- The function f measures probability *density*, not probability. Although $f(x)$ usually has values less than 1, this need not be the case. For example:



$$f(x) = \begin{cases} 2 & 0 < x < \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

  defines a proper probability density function that integrates to 1.
- Problems involving probability density functions often require the calculation of areas. Instead of using calculus it is often much quicker to use standard geometric results. In particular:

  A triangle of height h and base b has area $\frac{1}{2}hb$

  A trapezium with parallel sides of lengths k and l at a distance d apart has area $\frac{1}{2}d(k + l)$.

Example 1

The continuous random variable X has probability density function given by:

$$f(x) = \begin{cases} \frac{1}{2}x & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Determine $P(X > 1)$.

The statement that $f(x) = 0$ 'otherwise' merely emphasises that attention may safely be restricted to the interval (0, 2).

Glancing at the diagram we can see that the area corresponding to $P(X > 1)$ is greater than half of the total area between $f(x)$ and the x-axis, so that the required probability will be greater than 0.5. If our calculations give a value smaller than 0.5 then we must have made an error (possibly in the diagram!).

Method 1: Calculus

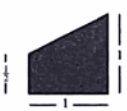
The required probability is:

$$\int_1^2 \frac{x}{2}\,dx = \left[\frac{x^2}{4}\right]_1^2 = \frac{4-1}{4} = \frac{3}{4}$$

So $P(X > 1) = 0.75$, which, as anticipated, is considerably greater than 0.5.

Method 2: Geometry

The required probability is given by the area of a trapezium having parallel sides of lengths $\frac{1}{2}$ and 1 at a distance apart of $2 - 1 = 1$. The area corresponding to $P(X > 1)$ is therefore equal to:



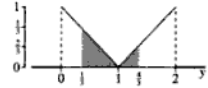
$$\tfrac{1}{2} \times 1 \times (\tfrac{1}{2} + 1) = \tfrac{3}{4}$$

confirming the result found by calculus.

Example 2

The continuous random variable Y has probability density function given by:

$$f(y) = \begin{cases} |y-1| & 0 < y < 2 \\ 0 & \text{otherwise} \end{cases}$$



Determine $P(\frac{1}{3} < Y < \frac{4}{3})$.

This time the diagram suggests that the answer will be less than 0.5.

Method 1: Calculus

We begin by rewriting the formula for the pdf so that the | signs are not needed:

$$f(y) = \begin{cases} 1-y & 0 < y \leqslant 1 \\ y-1 & 1 < y < 2 \\ 0 & \text{otherwise} \end{cases}$$

We now use the result that:

$$P(\tfrac{1}{3} < Y < \tfrac{4}{3}) = P(\tfrac{1}{3} < Y \leqslant 1) + P(1 < Y < \tfrac{4}{3})$$

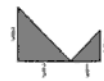
We therefore need:

$$\begin{aligned} \int_{\frac{1}{3}}^{1} (1-y)\,dy + \int_{1}^{\frac{4}{3}} (y-1)\,dy &= \left[y - \frac{y^2}{2}\right]_{\frac{1}{3}}^{1} + \left[\frac{y^2}{2} - y\right]_{1}^{\frac{4}{3}} \\ &= \left\{\left(1 - \frac{1}{2}\right) - \left(\frac{1}{3} - \frac{1}{18}\right)\right\} + \left\{\left(\frac{16}{18} - \frac{4}{3}\right) - \left(\frac{1}{2} - 1\right)\right\} \\ &= \left(\frac{1}{2} - \frac{5}{18}\right) + \left(\frac{-8}{18} - \frac{-1}{2}\right) \\ &= \frac{4}{18} + \frac{1}{18} = \frac{5}{18} \end{aligned}$$

The probability that Y takes a value between $\frac{1}{3}$ and $\frac{4}{3}$ is $\frac{5}{18}$, or 0.278 (to 3 d.p.).

Method 2: Geometry

In this case we have two triangles. At $y = \frac{1}{3}$, $f(y) = \frac{2}{3}$, so the left-hand triangle has both height and base equal to $\frac{2}{3}$ and therefore has area equal to:



$$\tfrac{1}{2} \times \tfrac{2}{3} \times \tfrac{2}{3} = \tfrac{2}{9}$$

At $y = \frac{4}{3}$, $f(y) = \frac{1}{3}$, so this triangle has area equal to:

$$\tfrac{1}{2} \times \tfrac{1}{3} \times \tfrac{1}{3} = \tfrac{1}{18}$$

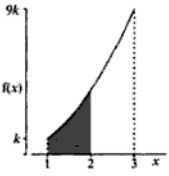
The total area of the two triangles is therefore $\frac{2}{9} + \frac{1}{18} = \frac{5}{18}$, which agrees with the answer obtained by calculus.

Example 3

The continuous random variable X has pdf given by:

$$f(x) = \begin{cases} kx^2 & 1 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

Determine (i) the value of the constant k, (ii) $P(X < 2)$.



We will not attempt a geometric solution in this case since a curve is involved.

(i) To find k we use the fact that f integrates to 1:

$$\int_1^3 kx^2\,dx = \left[\frac{kx^3}{3}\right]_1^3$$
$$= \frac{k}{3}(27 - 1)$$
$$= \frac{26k}{3}$$

Since we know that the integral is equal to 1, it follows that $k = \frac{3}{26}$.

(ii) $$P(X < 2) = \int_1^2 kx^2\,dx$$
$$= \left[\frac{kx^3}{3}\right]_1^2$$
$$= \frac{k}{3}(8 - 1)$$
$$= \frac{7k}{3}$$
$$= \frac{7}{26}$$

The probability that X takes a value less than 2 is $\frac{7}{26}$, or 0.269 (to 3 d.p.).

Calculator practice

If you have a calculator that can perform numerical integration then you should check that you know the appropriate instructions. Practice by checking that the answers to Examples 1–3 are correct.

Exercises 9a

In Questions **1–6**, the continuous random variable X has pdf f and k is a constant.

1 Given that:

$$f(x) = \begin{cases} kx^2 + \frac{1}{6} & 0 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f and find (i) the value of k, (ii) $P(X < 1)$, (iii) $P(X > 2)$.

2 Given that:

$$f(x) = \begin{cases} kx^2 & -2 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f and find (i) the value of k, (ii) $P(X > 1)$, (iii) $P(|X| < 1)$, (iv) $P(-1 < X < 0)$.

3 Given that:

$$f(x) = \begin{cases} \frac{1}{2}x & 1 < x < k \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f and find (i) the value of k, (ii) $P(\frac{3}{2} < X < 2)$, (iii) $P(X > 3)$, (iv) $P(2 < X < 3)$, (v) the mode.

4 Given that:

$$f(x) = \begin{cases} 1 - kx & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Sketch the graph of f and find
(i) the value of the constant k,
(ii) $P(X \leqslant 1)$.

5 Given that:

$$f(x) = \begin{cases} 10k & -0.05 < x < 0.05 \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f and find
(i) the value of k, (ii) $P(X > 0.1)$,
(iii) $P(X < 0.025)$.

6 Given that:

$$f(x) = \begin{cases} kx(6 - x) & 2 < x < 5 \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f and find
(i) the value of k, (ii) the mode, m,
(iii) $P(X < m)$.

7 For each of the following functions, state, giving a reason, whether or not there is a value of the constant k for which the function can be a pdf. Sketch graphs may help. You do not have to find the value of k.

(i) $f(x) = \begin{cases} kx & -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

(ii) $f(x) = \begin{cases} kx^2 & -1 < x < 2 \\ 0 & \text{otherwise} \end{cases}$

(iii) $f(x) = \begin{cases} 1 + kx & 0 < x < 3 \\ 0 & \text{otherwise} \end{cases}$

(iv) $f(x) = \begin{cases} 4 + x^2 & -k < x < k \\ 0 & \text{otherwise} \end{cases}$

8 A garage is supplied with petrol once a week. Its volume of weekly sales, X, in thousands of gallons, is distributed with probability density function $f(x)$ given by:

$$f(x) = \begin{cases} kx(1 - x)^2 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Determine the value of the constant k.

Determine an expression for the probability that the sales are less than c hundred gallons.

Determine the value of this probability for each of $c = 7$, 7.5 and 8.

Hence, or otherwise, determine an appropriate capacity for the garage's tank, if it is to have a probability of only about 0.05 of being exhausted in a given week.

9.3 The cumulative distribution function, F

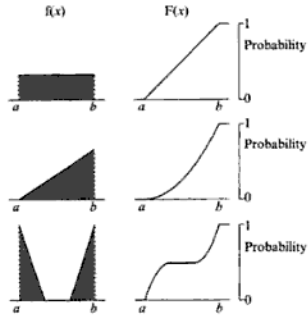
The cumulative distribution function is often referred to as the **distribution function** or, more simply, as the **cdf**. The graph of a cdf may be thought of as the limiting form of the cumulative frequency polygon (Section 1.14, p. 22). The function is defined by:

$$F(x) = P(X \leqslant x) = P(X < x) \tag{9.4}$$

and is related to the function f by:

$$F(b) = \int_{-\infty}^{b} f(x)\,dx \tag{9.5}$$

The lower limit of the integral is given as $-\infty$, but is in effect the smallest attainable value of X. In each of the following three diagrams, $P(X < a) = 0$ and $P(X > b) = 0$. The area under the graph of each density function is equal to 1.



Notes

- Since it is always impossible to have a value of X smaller than $-\infty$ or larger than ∞:
  - $F(-\infty) = 0$
  - $F(\infty) = 1$

  (Strictly $F(-\infty)$ means 'the limiting value of $F(x)$ as x approaches $-\infty$' and $F(\infty)$ is similarly defined.)
- As x increases so $F(x)$ either increases or remains constant, but never decreases. The third diagram shows that this also applies to cases where f is discontinuous.
- F is a continuous function, even if f is discontinuous.
- Useful relations are:

$$P(c < X < d) = F(d) - F(c) \tag{9.6}$$

$$P(X > x) = 1 - F(x) \tag{9.7}$$

Example 4

The continuous random variable X has pdf $f(x)$, given by:

$$f(x) = \begin{cases} 1 & 2 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

Determine the form of $F(x)$.

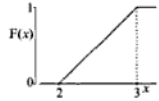
For $b \leqslant 2$, $F(b) = 0$, since there is no chance that X will take a value less than or equal to 2. Similarly, for $b \geqslant 3$, $F(b) = 1$, since it is certain that X takes a value less than 3.

For $2 \leqslant b \leqslant 3$, we use the definition:

$$F(b) = P(X \leqslant b) = \int_2^b 1 \, dx$$

$$= \left[x\right]_2^b$$

$$= (b - 2)$$

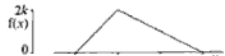


The form of $F(x)$ is therefore:

$$F(x) = \begin{cases} 0 & x \leqslant 2 \\ x - 2 & 2 \leqslant x \leqslant 3 \\ 1 & x \geqslant 3 \end{cases}$$

Example 5

The continuous random variable X has pdf given by:



$$f(x) = \begin{cases} 2k(x-1) & 1 < x < 2 \\ k(4-x) & 2 < x < 4 \\ 0 & \text{otherwise} \end{cases}$$

where k is a constant.
Determine (i) the value of k, (ii) the form of $F(x)$.

Either geometry or calculus could be used to answer both parts of the question. The simplest procedure is to use geometry to find the first answer and calculus to find the second.

(i) To find the value of k we use the fact that the total area of the region between the graph of the probability density function and the x-axis is equal to 1.

The sketch reveals that the region of interest is a triangle. Since $f(x) = 2k$ at $x = 2$, and since the triangle has base equal to $(4 - 1) = 3$, the triangle has area $\frac{1}{2} \times 2k \times 3 = 3k$. The area is known to be equal to 1 and therefore $k = \frac{1}{3}$.

(ii) $F(x) = 0$ for $x \leqslant 1$ and $F(x) = 1$ for $x \geqslant 4$, since X only takes values in the interval 1 to 4. We consider the two intervals $[1, 2]$ and $[2, 4]$ separately since f has a different form in each interval.

For $1 \leqslant b \leqslant 2$:

$$\begin{aligned} F(b) &= \int_1^b 2k(x-1)\,dx \\ &= \left[2k\frac{(x-1)^2}{2}\right]_1^b \\ &= k(b-1)^2 \\ &= \frac{(b-1)^2}{3} \end{aligned}$$

In particular, $P(X < 2) = F(2) = \frac{1}{3}$.

For $2 \leqslant b \leqslant 4$, we write:

$$\begin{aligned} F(b) &= P(X \leqslant b) = P(X \leqslant 2) + P(2 \leqslant X \leqslant b) \\ &= \tfrac{1}{3} + P(2 \leqslant X \leqslant b) \end{aligned}$$

We therefore need $P(2 \leqslant X \leqslant b)$:

$$\begin{aligned} P(2 \leqslant X \leqslant b) &= \int_2^b k(4-x)\,dx \\ &= \left[-k\frac{(4-x)^2}{2}\right]_2^b \\ &= -\frac{k}{2}\{(4-b)^2 - (4-2)^2\} \\ &= \frac{1}{6}\{4 - (4-b)^2\} \end{aligned}$$

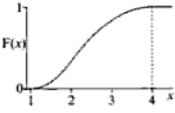
Hence, for $2 \leqslant b \leqslant 4$:

$$F(b) = \frac{1}{3} + \frac{4 - (4-b)^2}{6}$$

$$= \frac{6 - (4-b)^2}{6}$$

As a check, note that F(4) does indeed equal the maximum value, 1, and that F(2) equals $\frac{1}{3}$, the value obtained previously. The complete description of F(x) is therefore:

$$F(x) = \begin{cases} 0 & x \leqslant 1 \\ \dfrac{(x-1)^2}{3} & 1 \leqslant x \leqslant 2 \\ 1 - \dfrac{(4-x)^2}{6} & 2 \leqslant x \leqslant 4 \\ 1 & x \geqslant 4 \end{cases}$$



The median, m

The **median**, m, is the value that bisects the distribution in the sense that X is equally likely to be smaller or larger than m. Hence:

$$\int_{-\infty}^{m} f(x)\,dx = \int_{m}^{\infty} f(x)\,dx = 0.5 \qquad (9.8)$$

Notes

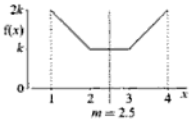
- If the graph of f is symmetric about the line $x = x_0$, then $m = x_0$.
- **Percentiles** and **quartiles** are defined similarly. For example, the 90th percentile is the solution of $F(x) = 0.90$, while the upper quartile is the solution of $F(x) = 0.75$.

Example 6

The continuous random variable X has pdf given by:

$$f(x) = \begin{cases} k(3-x) & 1 < x < 2 \\ k & 2 < x < 3 \\ k(x-2) & 3 < x < 4 \\ 0 & \text{otherwise} \end{cases}$$

where k is a constant.
Determine the median.



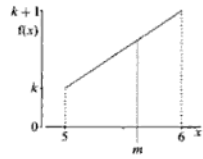
The diagram shows that the pdf is symmetric about the line $x = 2.5$, which implies that the median is 2.5. We do not need to know the value of k.

Example 7

The continuous random variable X has pdf given by:

$$f(x) = \begin{cases} k-5+x & 5 < x < 6 \\ 0 & \text{otherwise} \end{cases}$$

where k is a constant.
Determine the median.



Method 1: Calculus

This time the pdf is not symmetric. We must begin by determining the value of k, which we do by using the fact that the total area is 1. The area is given by:

$$\int_5^6 (k-5+x)\,dx = \left[(k-5)x + \frac{x^2}{2}\right]_5^6$$

$$= \left\{6(k-5) + \frac{6^2}{2}\right\} - \left\{5(k-5) + \frac{5^2}{2}\right\}$$

$$= (k-5) + 18 - \frac{25}{2}$$

$$= k + \frac{1}{2}$$

Since $(k + \frac{1}{2}) = 1$ it follows that $k = \frac{1}{2}$ and so:

$$f(x) = x - \tfrac{9}{2} \qquad \text{for } 5 < x < 6$$

To find the median, m, we need to solve $F(m) = \frac{1}{2}$. Hence m is the solution of:

$$\frac{1}{2} = \int_5^m \left(x - \frac{9}{2}\right) dx$$

$$= \left[\frac{x^2}{2} - \frac{9x}{2}\right]_5^m$$

$$= \left(\frac{m^2}{2} - \frac{9m}{2}\right) - \left(\frac{25}{2} - \frac{45}{2}\right)$$

$$= \frac{m^2 - 9m + 20}{2}$$

Rearranging we get:

$$m^2 - 9m + 19 = 0$$

which has solution:

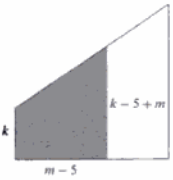
$$m = \frac{9 \pm \sqrt{81 - 76}}{2}$$

We need the root to be between 5 and 6 (since this is the range of possible values for X) and so it is the larger root, $\frac{1}{2}(9 + \sqrt{5})$, that is relevant. To three significant figures, the median is 5.62.

Method 2: Geometry

The region of interest is a trapezium with parallel sides of lengths k and $(k+1)$. The 'distance apart' is $(6-5)=1$, and so the area is:

$$\frac{1}{2}\times 1\times\{k+(k+1)\}=k+\frac{1}{2}$$



Since this area must be 1, we again conclude that $k=\frac{1}{2}$.

To find m we proceed in a similar way, by considering the smaller trapezium having parallel sides corresponding to the cases $x=5$ and $x=m$. These sides have lengths k and $k-5+m$, so the area of this trapezium is:

$$\frac{1}{2}\times(m-5)\times\{k+(k-5+m)\}=\frac{1}{2}(m-5)(2k-5+m)=\frac{1}{2}(m-5)(m-4)$$

We wish to choose m so that this area $=\frac{1}{2}$. Therefore, m is the solution of:

$$(m-5)(m-4)=1$$

which on rearranging gives:

$$m^2-9m+19=0$$

as before.

Example 8

The continuous random variable X has pdf given by:

$$\mathrm{f}(x)=\begin{cases} ax & 1<x\leqslant 3\\ c(4-x) & 3\leqslant x<4\\ 0 & \text{otherwise}\end{cases}$$

where a and c are constants.

Determine the values of a and c, and also the median, m.

Method 1: Calculus

We begin by noting the duplicate definition at $x=3$. This implies that:

$$\mathrm{f}(3)=3a=c$$

We replace c by $3a$ in the subsequent calculations.

As usual we begin by using the fact that the total area must equal 1. This total area is given by:

$$\begin{aligned}\int_1^3 ax\,\mathrm{d}x+\int_3^4 3a(4-x)\,\mathrm{d}x&=\left[a\frac{x^2}{2}\right]_1^3+\left[-3a\frac{(4-x)^2}{2}\right]_3^4\\&=a\left(\frac{9}{2}-\frac{1}{2}\right)+\left\{0-\left(-3a\times\frac{1}{2}\right)\right\}\\&=4a+\frac{3a}{2}\\&=\frac{11a}{2}\end{aligned}$$

and we therefore conclude that $a=\frac{2}{11}$ (and hence that $c=\frac{6}{11}$).

A glance at the diagram shows that the median is less than 3. We therefore solve the equation:

$$\int_1^m ax\,\mathrm{d}x=\frac{1}{2}$$

Now:

$$\int_1^m ax\,dx = \left[a\frac{x^2}{2}\right]_1^m = \frac{am^2}{2} - \frac{a}{2} = \frac{a(m^2-1)}{2} = \frac{m^2-1}{11}$$

Thus m is the solution of:

$$\frac{m^2-1}{11} = \frac{1}{2}$$

Multiplying through by 11 and rearranging we get:

$$m^2 = 1 + \frac{11}{2} = 6.5$$

and, since m is evidently positive:

$$m = \sqrt{6.5} = 2.55 \text{ (to 3 s.f.)}$$

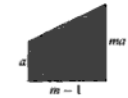
Method 2: Geometry

The total area is made up of a trapezium corresponding to the region between $x = 1$ and $x = 3$ and a triangle for the remainder. The parallel sides of the trapezium have lengths a and $3a$, with the 'distance apart' being $(3 - 1) = 2$. The triangle has height $3a$ and base $(4 - 3) = 1$, so the combined area is:

$$\frac{1}{2} \times (3-1) \times (a+3a) + \frac{1}{2} \times 3a \times (4-3) = 4a + \frac{3}{2}a = \frac{11}{2}a$$

giving $a = \frac{2}{11}$, as before.

Seeing that the median is less than 3, we need only consider the trapezium bounded by $x = 1$ and $x = m$, with sides of length a and ma, respectively. This trapezium therefore has area:



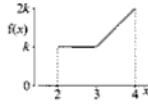
$$\frac{1}{2} \times (m-1) \times (a+ma) = \frac{1}{2}a(m-1)(1+m)$$
$$= \frac{1}{11}(m-1)(m+1) = \frac{1}{11}(m^2-1)$$

and so, as before, m is the solution of:

$$\frac{m^2-1}{11} = \frac{1}{2}$$

Example 9

The continuous random variable X has pdf given by:



$$f(x) = \begin{cases} k & 2 < x < 3 \\ k(x-2) & 3 < x < 4 \\ 0 & \text{otherwise} \end{cases}$$

Determine the value of the median, m.

Method 1: Calculus

We begin by finding the value of k. The total area, corresponding to the total probability of 1, is given by:

$$\int_2^3 k\,\mathrm{d}x + \int_3^4 k(x-2)\,\mathrm{d}x = \Big[kx\Big]_2^3 + \left[\frac{k(x-2)^2}{2}\right]_3^4$$
$$= (3k-2k) + \left(\frac{4k}{2} - \frac{k}{2}\right)$$
$$= k + \frac{3k}{2}$$
$$= \frac{5k}{2}$$

Since $\frac{5}{2}k = 1$, we see that $k = \frac{2}{5}$.

The diagram shows that the median is greater than 3. We can confirm this by noting that $\mathrm{P}(X < 3)$ is given by:

$$\int_2^3 k\,\mathrm{d}x = \Big[kx\Big]_2^3 = (3k-2k) = k = \frac{2}{5}$$

The value of m therefore satisfies the equation:

$$\int_3^m k(x-2)\,\mathrm{d}x = \tfrac{1}{2} - \tfrac{2}{5} = \tfrac{1}{10}$$

The integral is equal to:

$$\left[\frac{k(x-2)^2}{2}\right]_3^m = \frac{k(m-2)^2}{2} - \frac{k}{2} = \frac{k\{(m^2-4m+4)-1\}}{2}$$
$$= \frac{m^2-4m+3}{5}$$

and hence m is the solution of:

$$\frac{m^2-4m+3}{5} = \frac{1}{10}$$

or:

$$2(m^2-4m+3) = 1$$

which simplifies to:

$$2m^2 - 8m + 5 = 0$$

and has solution:

$$m = \frac{8 \pm \sqrt{8^2 - 4\times 2\times 5}}{2\times 2} = 2 \pm \sqrt{1.5}$$

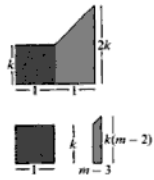
Since $m > 3$, we need the larger solution: the median is 3.22 (to 2 d.p.).

Method 2: Geometry

In this case the region of interest consists of a rectangle of sides k and 1 together with a trapezium with parallel sides of length k and $2k$, and 'distance apart' 1. The total area is therefore:

$$(k\times 1) + \frac{1}{2}\times 1\times(k+2k) = k + \frac{3}{2}k = \frac{5}{2}k$$

from which we (very quickly!) have found that, since the total area must be 1, the value of k must be $\frac{2}{5}$.



Substituting for k, we note that the area of the rectangle is $\frac{2}{5}$, which is less than $\frac{1}{2}$. It follows that the median, m, must exceed 3. We need to add to the rectangle a (thin!) trapezium with parallel sides of length k and $k(m-2)$, and 'distance apart' $(m-3)$.

The thin trapezium has area $\frac{1}{2} \times (m-3) \times \{k + k(m-2)\}$. We need this to have area $(\frac{1}{2} - \frac{2}{5}) = \frac{1}{10}$, and hence m is the solution of:

$$\tfrac{1}{2} \times (m-3) \times \{k(m-1)\} = \tfrac{1}{5} \times (m-3) \times (m-1) = \tfrac{1}{10}$$

Multiplying through by 10, this becomes:

$$2(m-3)(m-1) = 1$$

which finally simplifies to the equation obtained previously:

$$2m^2 - 8m + 5 = 0$$

from which we found that the median was 3.22 (to 2 d.p.).

Exercises 9b

In Questions **1–10**, the pdf and cdf of the continuous random variable X are f and F respectively, and k is a constant.

1 Given that:

$$f(x) = \begin{cases} kx + \frac{1}{4} & 1 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f.
Find: (i) the value of k, (ii) $F(x)$,
(iii) $P(X > 2)$, (iv) the median of X.

2 Given that:

$$f(x) = \begin{cases} -kx & -2 < x < 0 \\ kx & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f.
Find: (i) the value of k, (ii) the median of X,
(iii) $F(x)$, (iv) $P(0 < X < 1)$.

3 Given that:

$$f(x) = \begin{cases} \frac{1}{2} & 1 < x < 2 \\ k & 2 < x < 4 \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f.
Find:
(i) the value of k,
(ii) $F(x)$,
(iii) the tenth percentile of X,
(iv) the eightieth percentile of X.

4 It is given that:

$$f(x) = \begin{cases} k(x+2) & -2 < x < 0 \\ \frac{1}{2}k(3-x) & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Find:
(i) the value of k, (ii) $F(x)$
(iii) $P(-1 < X < 1)$, (iv) $P(1 < X < 3)$.

5 Given that:

$$f(x) = \begin{cases} kx^2 & -2 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f and find:
(i) the value of k, (ii) $F(x)$, (iii) the mode,
(iv) the median of X.

6 Given that:

$$f(x) = \begin{cases} k(x+2)^2 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f and find:
(i) the value of k, (ii) $F(x)$,
(iii) the median of X.

7 Given that:

$$f(x) = \begin{cases} kx^2 - c & 1 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f.
(i) Find, in terms of k, the maximum possible value for the constant c.
(ii) With this value for c, find the corresponding value of k.
(iii) With these values for k and c, find $F(x)$.

8 Given that:

$$f(x) = \begin{cases} 2x + k & 3 < x < 4 \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f and find:
(i) the value of k, (ii) $F(x)$,
(iii) the lower and upper quartiles of X.

9 Given that:

$$f(x) = \begin{cases} k & 0 < x < 1 \\ 4k & 1 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f and find:
(i) the value of k, (ii) $F(x)$,
(iii) the difference between the median and the fifth percentile of X.

10 Given that:

$$f(x) = \begin{cases} 2(1-x) & 0 < x < k \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f and find:
(i) the value of k, (ii) $F(x)$,
(iii) the median of X.

11 At Wetville, the proportion of the sky covered in cloud, S, has pdf f given by:

$$f(s) = \begin{cases} k(3+s) & 0 < s < 1 \\ 0 & \text{otherwise} \end{cases}$$

Sketch the graph of f and find:
(i) the value of k, (ii) $F(s)$, (iii) $P(S > 0.5)$,
(iv) the median of S.

12 The time, T hours, required to erect a type of wooden garden shed has pdf f given by:

$$f(t) = \begin{cases} kt^2 & 5 < t < 8 \\ 0 & \text{otherwise} \end{cases}$$

Sketch the graph of f and find:
(i) the value of k, (ii) $F(t)$,
(iii) the probability that it takes between 6 hours and 7 hours 20 minutes to erect a shed.

13 The continuous random variable Z has pdf f given by:

$$f(z) = \begin{cases} a + bz & 0 < z < 1 \\ 0 & \text{otherwise} \end{cases}$$

It is given that $F(0.5) = 0.6$.
(a) Determine the values of a and b.
(b) Determine the median and the mode of Z.

9.4 Expectation and variance

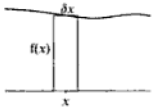
The formula for the sample mean, $\bar{x}$, can be written as:

$$\bar{x} = \sum x_j \left(\frac{f_j}{n}\right)$$

where $\frac{f_j}{n}$ is the relative frequency of the value x_j. As the sample increases in size, so, for a *discrete* random variable, $\frac{f_j}{n}$ converges on the corresponding population probability. However, for a *continuous* random variable, a probability can only be associated with a *range* of values, for example:

$$P\left[\left(x - \frac{\delta x}{2}\right) < X < \left(x + \frac{\delta x}{2}\right)\right] \approx f(x) \times \delta x$$

since the thin rectangle that approximates this probability has width δx and height $f(x)$.



The analogue of $\sum x_j \left(\frac{f_j}{n}\right)$ is therefore $\Sigma x\, f(x)\delta x$, where the latter summation is over a huge number of values of x, each separated by a small amount δx. As we decrease the size of δx so the value of $\Sigma x\, f(x)\delta x$ converges on $\int x\, f(x)\, dx$. Hence, for a continuous variable X, the population mean is the expectation $E(X)$ given by:

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx \qquad (9.9)$$

The limits of the integral are given as $-\infty$ and ∞, but are in effect the largest and smallest attainable values of X. By a similar argument:

$$\mathrm{E}[\mathrm{g}(X)] = \int_{-\infty}^{\infty} \mathrm{g}(x)\mathrm{f}(x)\,\mathrm{d}x \tag{9.10}$$

In particular:

$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2\,\mathrm{f}(x)\,\mathrm{d}x \tag{9.11}$$

and:

$$\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2\,\mathrm{f}(x)\,\mathrm{d}x \tag{9.12}$$

where $\mu = \mathrm{E}(X)$, though it is usually easier to calculate $\mathrm{Var}(X)$ using:

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

All the results of Chapter 6 continue to hold:

$$\begin{aligned}
\mathrm{E}(X+a) &= \mathrm{E}(X) + a \\
\mathrm{E}(aX) &= a\mathrm{E}(X) \\
\\
\mathrm{Var}(X+a) &= \mathrm{Var}(X) \\
\mathrm{Var}(aX) &= a^2\mathrm{Var}(X) \\
\\
\mathrm{E}(X+Y) &= \mathrm{E}(X) + \mathrm{E}(Y) \\
\mathrm{E}(aX+bY+c) &= a\mathrm{E}(X) + b\mathrm{E}(Y) + c \\
\mathrm{E}(R+S+T+U) &= \mathrm{E}(R) + \mathrm{E}(S) + \mathrm{E}(T) + \mathrm{E}(U)
\end{aligned}$$

For independent random variables we also have:

$$\begin{aligned}
\mathrm{Var}(aX+bY+c) &= a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) \\
\mathrm{Var}(R+S+T+U) &= \mathrm{Var}(R) + \mathrm{Var}(S) + \mathrm{Var}(T) + \mathrm{Var}(U)
\end{aligned}$$

Note

- If f is symmetric about the line $x = c$ and X has expectation $\mathrm{E}(X)$, then $\mathrm{E}(X) = c$. The median is also c, of course.

Example 10

The continuous random variable X has pdf given by:

$$\mathrm{f}(x) = \begin{cases} \dfrac{3(1-x^2)}{4} & -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Determine $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.
Denoting these quantities by μ and σ^2, respectively, determine the probability that an observed value of X has a value in the interval $(\mu-\sigma, \mu+\sigma)$.

We see from the sketch that f is symmetric about the line $x = 0$, hence $\mu = \mathrm{E}(X) = 0$.

To calculate $\mathrm{Var}(X)$ we need $\mathrm{E}(X^2)$:

$$\mathrm{E}(X^2) = \int_{-1}^{1} x^2 \times \frac{3(1-x^2)}{4}\,\mathrm{d}x$$

$$= \frac{3}{4}\int_{-1}^{1} (x^2 - x^4)\,\mathrm{d}x$$

$$= \frac{3}{4}\left[\frac{x^3}{3} - \frac{x^5}{5}\right]_{-1}^{1}$$

$$= \frac{3}{4}\left[\left(\frac{1}{3} - \frac{1}{5}\right) - \left\{\frac{(-1)}{3} - \frac{(-1)}{5}\right\}\right]$$

$$= \frac{3}{2}\left(\frac{1}{3} - \frac{1}{5}\right)$$

$$= \frac{3}{2} \times \frac{2}{15}$$

$$= \frac{1}{5}$$

Hence $\sigma^2 = \mathrm{Var}(X) = \mathrm{E}(X^2) - \{\mathrm{E}(X)\}^2 = \frac{1}{5}$.

Note that, because of symmetry, we could instead have calculated:

$$\mathrm{E}(X^2) = 2\int_{0}^{1} \frac{3}{4}x^2(1-x^2)\,\mathrm{d}x$$

$$= \frac{3}{2}\int_{0}^{1} (x^2 - x^4)\,\mathrm{d}x$$

$$= \frac{3}{2}\left[\frac{x^3}{3} - \frac{x^5}{5}\right]_{0}^{1}$$

$$= \frac{3}{2}\left(\frac{1}{3} - \frac{1}{5}\right)$$

$$= \frac{3}{2} \times \frac{2}{15}$$

$$= \frac{1}{5}$$

which is slightly quicker!

The probability of an observed value of X having a value in the interval $(\mu-\sigma, \mu+\sigma)$ is therefore:

$$\begin{aligned}
\mathrm{P}(\mu-\sigma<X<\mu+\sigma) &= \mathrm{P}\left(-\frac{1}{\sqrt{5}}<X<\frac{1}{\sqrt{5}}\right)\\
&= \frac{3}{4}\int_{-\frac{1}{\sqrt{5}}}^{\frac{1}{\sqrt{5}}}(1-x^2)\,\mathrm{d}x\\
&= \frac{3}{2}\int_{0}^{\frac{1}{\sqrt{5}}}(1-x^2)\,\mathrm{d}x \qquad \text{by symmetry}\\
&= \frac{3}{2}\left[x-\frac{x^3}{3}\right]_0^{\frac{1}{\sqrt{5}}}\\
&= \frac{3}{2}\left(\frac{1}{\sqrt{5}}-\frac{1}{3}\times\frac{1}{5\sqrt{5}}\right)\\
&= \frac{3}{2}\times\frac{15-1}{15\sqrt{5}}\\
&= \frac{3\times 14}{2\times 15\sqrt{5}}\\
&= \frac{7}{5}\sqrt{\frac{1}{5}}\\
&= 0.626 \text{ (to 3 d.p.)}
\end{aligned}$$

The required probability is 0.626, which is illustrated in the sketch.

Example 11

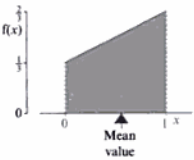
The continuous random variable X has pdf given by:

$$\mathrm{f}(x)=\begin{cases}\dfrac{2(x+1)}{3} & 0<x<1\\ 0 & \text{otherwise}\end{cases}$$

Determine $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.
Determine also the probability that two independent observed values of X both have values below the mean.

Since $\mathrm{f}(x)$ is not symmetric we need to carry out the integration:

$$\begin{aligned}
\mathrm{E}(X) &= \int_0^1 x\times\frac{2(x+1)}{3}\,\mathrm{d}x\\
&= \frac{2}{3}\int_0^1 (x^2+x)\,\mathrm{d}x\\
&= \frac{2}{3}\left[\frac{x^3}{3}+\frac{x^2}{2}\right]_0^1\\
&= \frac{2}{3}\left(\frac{1}{3}+\frac{1}{2}\right)\\
&= \frac{2}{3}\times\frac{5}{6}\\
&= \frac{5}{9}
\end{aligned}$$



A glance at the sketch suggests that the shaded area *would* balance on a fulcrum placed at $x=\frac{5}{9}$.

In Section 2.17 (p. 60) we noted that, as a guide, the range of a random variable usually takes a value between 3σ and 6σ, where σ is the standard deviation. Since the range here is equal to 1 it follows that we expect that σ will lie between $\frac{1}{3}$ and $\frac{1}{6}$, and hence that $\text{Var}(X)$ $(= \sigma^2)$ will lie between $\frac{1}{9}$ and $\frac{1}{36}$ (i.e. between 0.111 and 0.028).

In order to calculate the variance we need $\text{E}(X^2)$:

$$\begin{aligned}\text{E}(X^2) &= \int_0^1 x^2 \times \frac{2(x+1)}{3}\,\text{d}x \\ &= \frac{2}{3}\int_0^1 (x^3 + x^2)\,\text{d}x \\ &= \frac{2}{3}\left[\frac{x^4}{4} + \frac{x^3}{3}\right]_0^1 \\ &= \frac{2}{3}\left(\frac{1}{4} + \frac{1}{3}\right) \\ &= \frac{2}{3} \times \frac{7}{12} = \frac{7}{18}\end{aligned}$$

Hence:

$$\text{Var}(X) = \text{E}(X^2) - \{E(X)\}^2 = \frac{7}{18} - \left(\frac{5}{9}\right)^2 = \frac{63}{162} - \frac{50}{162} = \frac{13}{162}$$

Since $\frac{13}{162} = 0.080$ (to 3 d.p.) which lies comfortably in the anticipated range of (0.028, 0.111), we have no indication of having made an error.

The probability that an observed value of X is less than the mean is given by:

$$\begin{aligned}\text{P}\left(X < \frac{5}{9}\right) &= \int_0^{\frac{5}{9}} \frac{2(x+1)}{3}\,\text{d}x \\ &= \frac{2}{3}\left[\frac{x^2}{2} + x\right]_0^{\frac{5}{9}} \\ &= \frac{2}{3}\left(\frac{25}{162} + \frac{5}{9}\right) \\ &= \frac{2}{3} \times \frac{115}{162} = \frac{115}{243}\end{aligned}$$

The probability that two independent observed values of X are both smaller than the mean is therefore (using the multiplication rule) $(\frac{115}{243})^2 = 0.224$ (to 3 d.p.).

Example 12

The continuous random variable X has pdf given by:

$$\text{f}(x) = \begin{cases} x & 0 < x < 1 \\ 2 - x & 1 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Determine $\text{E}(X)$ and $\text{Var}(X)$.

Denoting these quantities by μ and σ^2, respectively, determine the probability that an observed value of X lies in the interval $(\mu - \sigma, \mu + \sigma)$.

We see from the sketch that $f(x)$ is symmetric about the line $x = 1$, **so** $\mu = E(X) = 1$.

We now need $E(X^2)$ and must take care, since the form of $f(x)$ depends upon the value of x:

$$\begin{aligned}
E(X^2) &= \int_0^1 x^2 x\,dx + \int_1^2 x^2(2-x)\,dx \\
&= \int_0^1 x^3\,dx + \int_1^2 (2x^2 - x^3)\,dx \\
&= \left[\frac{x^4}{4}\right]_0^1 + \left[\left(2\times\frac{x^3}{3}\right) - \frac{x^4}{4}\right]_1^2 \\
&= \frac{1}{4} + \left\{\left(2\times\frac{2^3}{3}\right) - \frac{2^4}{4}\right\} - \left\{\left(2\times\frac{1}{3}\right) - \frac{1}{4}\right\} \\
&= \frac{1}{4} + \frac{16}{3} - \frac{16}{4} - \frac{2}{3} + \frac{1}{4} \\
&= \frac{14}{3} - \frac{14}{4} \\
&= \frac{7}{6}
\end{aligned}$$

Hence:

$$\sigma^2 = \mathrm{Var}(X) = \frac{7}{6} - 1^2 = \frac{1}{6}$$

Calculation of the probability that an observed value of X lies in the interval $(\mu - \sigma, \mu + \sigma)$ is made easier by exploiting the symmetry of $f(x)$ and calculating the equivalent quantity $2P(\mu - \sigma < X < \mu)$:

$$\begin{aligned}
P(\mu - \sigma < X < \mu) &= \int_{\mu-\sigma}^{\mu} f(x)\,dx \\
&= \int_{1-\frac{1}{\sqrt{6}}}^{1} x\,dx \\
&= \left[\frac{x^2}{2}\right]_{1-\frac{1}{\sqrt{6}}}^{1} \\
&= \frac{1}{2}\left[1 - \left(1 - \frac{1}{\sqrt{6}}\right)^2\right] \\
&= \frac{1}{2}\left[1 - \left(1 - 2\frac{1}{\sqrt{6}} + \frac{1}{6}\right)\right] \\
&= \frac{1}{2}\left(2\frac{1}{\sqrt{6}} - \frac{1}{6}\right) \\
&= 0.3249 \text{ (to 4 d.p.)}
\end{aligned}$$

The probability that an observed value of X falls in the interval $(\mu - \sigma, \mu + \sigma)$ is therefore $2 \times 0.3249 = 0.650$ (to 3 d.p.).

Exercises 9c

In Questions **1–8** the continuous random variable X has pdf f.

1 It is given that:

$$f(x) = \begin{cases} \frac{1}{2}x & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Find (i) $E(X)$, (ii) $E(X^2)$, (iii) $Var(X)$.
Find also (iv) $P[X < E(X)]$.

2 It is given that:

$$f(x) = \begin{cases} \frac{1}{4}(3x^2 - 6x + 4) & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Find (i) $E(X)$ and (ii) $Var(X)$.

3 It is given that:

$$f(x) = \begin{cases} \frac{1}{4} & 0 < x < 1 \\ \frac{1}{2} & 1 < x < 2 \\ \frac{1}{4} & 2 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

Find (i) $E(X)$ and (ii) $Var(X)$.

4 It is given that:

$$f(x) = \begin{cases} \frac{2}{3}x & 0 < x < 1 \\ \frac{1}{3} & 2 < x < 4 \\ 0 & \text{otherwise} \end{cases}$$

Find (i) $E(X)$ and (ii) $Var(X)$.
Find also (iii) the median of X,
(iv) $P[X < E(X)]$, (v) $E(X^3)$.

Two independent observations of X are taken.
Find (vi) the probability that one of the observations exceeds the mean and the other is less than the median.

5 Given that:

$$f(x) = \begin{cases} 4 & 0 < x < \frac{1}{4} \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f and find (i) $E(X)$,
(ii) $E(2X + 4)$, (iii) $Var(X)$, (iv) $Var(2X + 4)$.

6 Given that:

$$f(x) = \begin{cases} 2x & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f.
The random variable Y is defined by
$Y = 4X + 2$.
Find (i) the expectation of Y,
(ii) the variance of Y.

7 Given that:

$$f(x) = \begin{cases} kx^3 & 2 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

sketch the graph of f and find
(i) the value of k, (ii) $E(X)$, (iii) $Var(X)$.
The independent continuous random variables X_1 and X_2 have the same distribution as X.
Find (iv) $E(X_1 - X_2)$, (v) $Var(X_1 - X_2)$.

8 Given that:

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Determine (i) the mean of X,
(ii) the variance of X.
The random variable Y is defined as the sum of 12 independent random variables that each has the same distribution as X.
Determine (iii) the mean of Y,
(iv) the variance of Y.

9 The amount of chemical, W g, produced by a reaction, has pdf given by:

$$f(w) = \begin{cases} k\{4 - (4 - w)^2\} & 3 < w < 6 \\ 0 & \text{otherwise} \end{cases}$$

Sketch the graph of f and find
(i) the value of k, (ii) $E(W)$, (iii) $P[W > E(W)]$.

10 The random variable X has probability density function given by:

$$f(x) = \begin{cases} x & 0 \leqslant x \leqslant 1 \\ k - x & 1 \leqslant x \leqslant 2 \\ 0 & \text{otherwise} \end{cases}$$

Find k.
Find also the mean, μ, and show that the variance, σ^2, is equal to $\frac{1}{6}$.
Determine the probability that a future observation lies in the interval $(\mu - \sigma, \mu)$.

11 The amount of cloud cover is measured on a scale from 0 to 1. In this country a reasonable model for the amount of cloud cover, X, at midday during the spring is provided by the probability density function f, defined below.

$$f(x) = \begin{cases} 8(x - 0.5)^2 + c & 0 \leqslant x \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the value of the constant c.
Sketch the graph of $f(x)$ for $0 \leqslant x \leqslant 1$.
State the mean cloud cover.
Determine, correct to 3 decimal places, the value x_0 which is such that $P(X > x_0) = 0.95$.

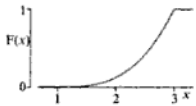
9.5 Obtaining f from F

Since F can be obtained by integrating f, f can be obtained by differentiating F. The value of f(b) is therefore the slope of F at the point where $x = b$.

Example 13

The random variable X has cdf given by:

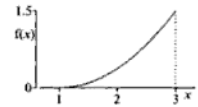
$$F(x) = \begin{cases} 0 & x \leqslant 1 \\ \dfrac{(x-1)^3}{8} & 1 \leqslant x \leqslant 3 \\ 1 & x \geqslant 3 \end{cases}$$



Find the form of the pdf of X.

Evidently f(x) is equal to 0 for $x < 1$ and for $x > 3$, since F(x) is unchanging in these regions. Writing F′(x) for $\dfrac{dF(x)}{dx}$, for the interval $1 < x < 3$ we calculate:

$$f(x) = F'(x)$$
$$= \frac{3}{8}(x-1)^2$$



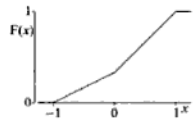
Hence:

$$f(x) = \begin{cases} \dfrac{3}{8}(x-1)^2 & 1 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

Example 14

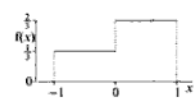
The continuous random variable X has cdf given by:

$$F(x) = \begin{cases} 0 & x \leqslant -1 \\ \alpha x + \alpha & -1 \leqslant x \leqslant 0 \\ 2\alpha x + \alpha & 0 \leqslant x \leqslant 1 \\ 3\alpha & x \geqslant 1 \end{cases}$$



where α is a constant.
Determine (i) the value of α, (ii) the pdf of X.

We know that F(∞) = 1. Hence $3\alpha = 1$, implying that $\alpha = \frac{1}{3}$. Also it is clear that f(x) = 0 for $x < -1$ and for $x > 1$. Differentiating F(x) for each of the remaining intervals, and substituting for α, we get:



$$f(x) = \begin{cases} \frac{1}{3} & -1 < x < 0 \\ \frac{2}{3} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Exercises 9d

In Questions **1–7** the pdf and cdf of the continuous random variable X are f and F respectively.

1 It is given that:

$$F(x) = \begin{cases} 0 & x \leqslant 1 \\ \dfrac{2(x-1)}{x} & 1 \leqslant x \leqslant 2 \\ 1 & x \geqslant 2 \end{cases}$$

Find (i) $f(x)$, (ii) $E(X)$.

2 It is given that:

$$F(x) = \begin{cases} 0 & x \leqslant 1 \\ a + bx^2 & 1 \leqslant x \leqslant 3 \\ 1 & x \geqslant 3 \end{cases}$$

Find the values of the constants (i) a and (ii) b.
Find (iii) $E(X)$ and (iv) $\text{Var}(X)$.

3 It is given that:

$$F(x) = \begin{cases} 0 & x \leqslant a \\ k(x-a)^2 & a \leqslant x \leqslant 2a \\ 1 & x \geqslant 2a \end{cases}$$

Find k in terms of a.
It is also given that $P(X > 2) = \frac{3}{4}$.
(i) Find the value of a.
(ii) Find $f(x)$.
(iii) Find the median of X.

4 It is given that:

$$F(x) = \begin{cases} 0 & x \leqslant 4 \\ \frac{1}{4}(x-4) & 4 \leqslant x \leqslant 8 \\ 1 & x \geqslant 8 \end{cases}$$

(i) Find $f(x)$ and sketch its graph.
Find also:
(ii) the median of X,
(iii) the lower and upper quartiles of X,
(iv) $E(X)$.

5 It is given that:

$$F(x) = \begin{cases} 0 & x \leqslant -4 \\ \frac{1}{8}(x+4) & -4 \leqslant x \leqslant 0 \\ \frac{1}{2} & 0 \leqslant x \leqslant 4 \\ \frac{1}{8}x & 4 \leqslant x \leqslant 8 \\ 1 & x \geqslant 8 \end{cases}$$

Find (i) $f(x)$ and sketch its graph.
Find also:
(ii) $E(X)$,
(iii) $P[|X - E(X)| < 2]$,
(iv) $\text{Var}(X)$.

6 It is given that:

$$F(x) = \begin{cases} 0 & x \leqslant 0 \\ \frac{1}{4}x & 0 \leqslant x \leqslant 2 \\ a + bx & 2 \leqslant x \leqslant 3 \\ 1 & x \geqslant 3 \end{cases}$$

Find:
(i) the constants a and b, (ii) $f(x)$.
Sketch the graph of f.
(iii) Find the lower and upper quartiles of X.

7 It is given that:

$$F(x) = \begin{cases} 0 & x \leqslant 0 \\ \frac{1}{2}x^2 & 0 \leqslant x \leqslant 1 \\ a + bx^3 & 1 \leqslant x \leqslant 2 \\ 1 & x \geqslant 2 \end{cases}$$

Find:
(i) the constants a and b, (ii) $f(x)$.
Sketch the graph of f.
Find:
(iii) the mode, (iv) the median,
(v) the mean of X.

9.6 Distribution of a function of a random variable

If we know the pdf for the random variable X, then we can often deduce the pdf for a simple function of X. Denote the function of X by Y and let f_X and f_Y be the pdfs of the two variables, with the corresponding cdfs being F_X and F_Y. The determination of f_Y from f_X proceeds in three stages:

$$f_X \rightarrow F_X \rightarrow F_Y \rightarrow f_Y$$

Providing that Y is an increasing (or decreasing) function of X, this can be carried out fairly easily. A description of the general method follows two simple examples.

Example 15

The continuous random variable X has pdf $f_X(x)$, given by

$$f_X(x) = \begin{cases} 1 & 2 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

The random variable Y is given by $Y = 2X + 3$. Determine the pdf and cdf of Y.

The only possible values for X lie in the interval $2 < X < 3$. When $X = 2$, $Y = 2 \times 2 + 3 = 7$. Similarly, when $X = 3$, $Y = 9$. Since Y is an increasing function of X, we have found that $7 < Y < 9$ and hence, for values outside this interval $f_Y(y) = 0$.

The next step is to find the cdf of X. We found this earlier, in Example 4:

$$F_X(x) = P(X \leqslant x) = \begin{cases} 0 & x \leqslant 2 \\ x - 2 & 2 \leqslant x \leqslant 3 \\ 1 & x \geqslant 3 \end{cases}$$

We now deduce the cdf of Y, $F_Y(y)$. For values of y in $7 \leqslant y \leqslant 9$ we argue as follows:

$$\begin{aligned} F_Y(y) = P(Y \leqslant y) &= P[(2X + 3) \leqslant y] = P[2X \leqslant (y - 3)] \\ &= P\left[X \leqslant \tfrac{1}{2}(y - 3)\right] = F_X\left[\tfrac{1}{2}(y - 3)\right] \\ &= \tfrac{1}{2}(y - 3) - 2 = \tfrac{1}{2}(y - 7) \end{aligned}$$

For values of y in the interval $7 < y < 9$ the pdf $f_Y(y)$ is the derivative of $F_Y(y)$ which is just $\frac{1}{2}$. The complete pdf of Y is therefore:

$$f_Y(y) = \begin{cases} \frac{1}{2} & 7 < y < 9 \\ 0 & \text{otherwise} \end{cases}$$

and the complete cdf of Y is:

$$F_Y(y) = P(Y \leqslant y) = \begin{cases} 0 & y \leqslant 7 \\ \frac{1}{2}(y - 7) & 7 \leqslant y \leqslant 9 \\ 1 & y \geqslant 9 \end{cases}$$

Example 16

Suppose that, with X as in the previous example, we look at W, defined by $W = X^2$. We want the pdf of W, f_W.

This time the interval $2 < X < 3$ corresponds to $4 < W < 9$. We begin by determining $F(x)$ which (as before) is

$$F_X(x) = P(X \leqslant x) = \begin{cases} 0 & x \leqslant 2 \\ x - 2 & 2 \leqslant x \leqslant 3 \\ 1 & x \geqslant 3 \end{cases}$$

For values of w in $4 \leqslant w \leqslant 9$:

$$\begin{aligned} F_W(w) = P(W \leqslant w) &= P(X^2 \leqslant w) \\ &= P(X \leqslant \sqrt{w}) = F_X(\sqrt{w}) \\ &= \sqrt{w} - 2 \end{aligned}$$

and differentiating with respect to w, $F_W(w) = \dfrac{1}{2\sqrt{w}}$. The full pdf is therefore

$$f_W(w) = \begin{cases} \dfrac{1}{2\sqrt{w}} & 4 < w < 9 \\ 0 & \text{otherwise} \end{cases}$$

The general method is as follows. Suppose X has pdf f_X and cdf F_X, and suppose that y, equal to $g(x)$, is an increasing function of x. From calculus, this is equivalent to stating that $g'(x)$ (the derivative of $g(x)$ with respect to x) is always positive. This means that g has an inverse function, which we shall denote by h, which is also an increasing function. This implies that

$$Y = g(X) \quad \text{is equivalent to} \quad X = h(Y)$$

and that the condition

$$Y \leqslant y \quad \text{is equivalent to} \quad h(Y) \leqslant h(y)$$

Since $X = h(Y)$

$$Y \leqslant y \quad \text{is equivalent to} \quad X \leqslant h(y)$$

Proceeding as before

$$F_Y(y) = P(Y \leqslant y) = P[X \leqslant h(y)] = F_X[h(y)]$$

Differentiating with respect to y using the chain rule, we obtain the general formula

$$f_Y(y) = f_X[h(y)] \times h'(y)$$

since $F'_Y(y) = f_Y(y)$ and $F'_X(x) = f_X(x)$.

In Example 15 above, $f_X(x) = 1$ for $2 < x < 3$, and $h(y) = \frac{1}{2}(y - 3)$. Thus $h'(y) = \frac{1}{2}$, giving $f_Y(y) = \frac{1}{2}$ for $7 < y < 9$, as before.

In Example 16 above, $h(w) = \sqrt{w}$ and $h'(w) = \dfrac{1}{2\sqrt{w}}$, so that $f_W(w) = \dfrac{1}{2\sqrt{w}}$, for $4 < w < 9$, as before.

If g is a decreasing function and h is the inverse of g, then $h'(y)$ is negative and

$$F_Y(y) = P(Y \leqslant y) = P[X \geqslant h(y)] = 1 - F_X[h(y)]$$

and the final formula becomes

$$f_Y(y) = -f_X[h(y)] \times h'(y)$$

Combining the two results, we see that, if Y is either an increasing function of X, or a decreasing function of X, then

$$f_Y(y) = f_X[h(y)] \times |h'(y)|$$

where $Y = g(X)$, $X = h(Y) = g^{-1}(Y)$.

Exercises 9e

1 The continuous random variable X has pdf $f_X(x)$, given by

$$f_X(x) = \begin{cases} \frac{1}{3} & 2 < x < 5 \\ 0 & \text{otherwise} \end{cases}$$

The random variable Y is given by $Y = 3X - 1$. Determine the pdf of Y.

2 The continuous random variable X has pdf $f_X(x)$, given by

$$f_X(x) = \begin{cases} \frac{1}{3} & 2 < x < 5 \\ 0 & \text{otherwise} \end{cases}$$

The random variable Y is given by $Y = X^2 + 3$. Determine the pdf of Y.

3 The continuous random variable X has pdf $f_X(x)$, given by

$$f_X(x) = \begin{cases} \frac{2}{3}(x+1) & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

The random variable Y is given by $Y = (1 - X)^2$. Determine the pdf of Y.

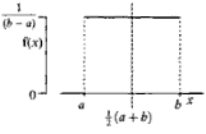
4 The continuous random variable X has pdf $f_X(x)$, given by

$$f_X(x) = \begin{cases} \frac{1}{21}x^2 & 1 < x < 4 \\ 0 & \text{otherwise} \end{cases}$$

The random variable Y is given by $Y = \sqrt{X} - 1$. Determine the pdf of Y.

9.7 The uniform (rectangular) distribution

We encountered a random variable having a uniform distribution in Example 4 of this chapter. Its characteristic is that, for the entire range of possible values of X (from a to b, say), f is constant:



$$f(x) = \begin{cases} \dfrac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases} \qquad (9.13)$$

Between a and b the probability density is **uniform** and the resulting shape is **rectangular**. The rectangle has width $(b-a)$ and height $\dfrac{1}{b-a}$, so that its area (equal to *height* × *width*) is equal to 1, as required.

Since the probability density is symmetrical about the line $x = \frac{1}{2}(a+b)$, the mean, $E(X)$, and the median, m, are both equal to $\frac{1}{2}(a+b)$. The cumulative distribution function, F, is given by:

$$\begin{aligned} F(c) = P(X \leqslant c) &= \int_a^c \frac{1}{b-a}\,dx \\ &= \left[\frac{x}{b-a}\right]_a^c \\ &= \frac{c-a}{b-a} \end{aligned}$$

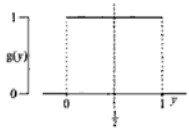
Formally, therefore, we have:

$$F(x) = \begin{cases} 0 & x \leqslant a \\ \dfrac{x-a}{b-a} & a \leqslant x \leqslant b \\ 1 & x \geqslant b \end{cases} \qquad (9.14)$$

To find the variance of X, we use the transformation:

$$Y = \frac{X-a}{b-a}$$

which amounts to a translation followed by an enlargement. It follows that the distribution of Y is also uniform, but with a revised range. Since $X = a$ gives $Y = 0$ and $X = b$ gives $Y = 1$, the range for Y is from 0 to 1. The probability density function of Y, g say, is given by:



$$g(y) = \begin{cases} 1 & 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

We see that $E(Y) = \frac{1}{2}$, while $E(Y^2)$ is given by:

$$\begin{aligned} E(Y^2) &= \int_0^1 y^2\,dy \\ &= \left[\frac{y^3}{3}\right]_0^1 \\ &= \frac{1}{3} \end{aligned}$$

Practical

Any type of round-off error is likely to have a uniform distribution. An 'instantaneous' example is provided by a glance at a watch! On a given signal, all members of the class should record the number of seconds shown by their watches. The numbers recorded are likely to be observations from a continuous uniform distribution with range 0 to 60, though, since they are recorded as integers, this distribution is being approximated by its discrete counterpart.
Summarise the class data using a stem-and-leaf diagram.
Does the uniform distribution seem appropriate?

You can easily generate more observations from your own watch by recording the numbers of seconds that it shows at odd times during the day (i.e. every now and then, when you remember – this could be a way of livening up the lessons in which you are not studying Statistics!).

Exercises 9f

1 The continuous random variable X has a uniform distribution on the interval $0 < x < 2$.
Find (i) the pdf of X, (ii) the cdf of X.

The random variable Y is defined by $Y = 2X$.
Find (iii) $P(Y < y)$ and hence (iv) $g(y)$, where g is the pdf of Y.

Verify that $E(Y) = 2E(X)$.

2 The continuous random variable U is uniformly distributed on the interval $a < u < b$.
Given that $E(U) = 4$ and $Var(U) = 3$, find
(i) a and b, (ii) $P(U > 5)$.

3 The continuous random variable T is uniformly distributed on the interval $0 < t < 100$.
(i) Find $P(20 < T < 60)$.
(ii) Denoting the mean and standard deviation of T by μ and σ respectively, find $P(|T - \mu| < \sigma)$.

4 The continuous random variable S is uniformly distributed on the interval $c < s < d$.
Given that $P(S < 3) = \frac{1}{4}$ and $P(S < 7) = \frac{3}{4}$, find c and d.

5 The continuous random variable Y is uniformly distributed on the interval $0 < y < 2$ and $X = 3Y + 4$.
Show that X is uniformly distributed on an interval $a < x < b$, giving the values of a and b.

6 (a) A pointed arrow is thrown on to a table and the continuous random variable A is the angle (acute or obtuse and measured in degrees) between the direction of the arrow and due north, measured so that $0 < A < 180$.
Find (i) $E(A)$, (ii) $Var(A)$.

(b) A pointed arrow is thrown on to a table and the continuous random variable B is the bearing (in degrees) of the direction of the arrow, measured so that $0 < B < 360$.
Find (i) $E(B)$, (ii) $Var(B)$.

7 Many calculators and computers generate random numbers which are approximately uniformly distributed on the interval $0 < u < 1$.
Let U be such a random variable.
(a) It is desired to find constants h and k such that $X = hU + k$ is uniformly distributed on the interval $a < x < b$.
Find h and k in terms of a and b.
(b) It is desired to find constants r and s such that $rU + s$ is uniformly distributed with mean μ and standard deviation σ.
Find r and s in terms of μ and σ.

(Strictly these random numbers have a discrete uniform distribution, since they only contain a fixed number of decimal places, say 9. As the difference between neighbouring numbers is 10^{-9}, the distribution may be taken to be continuous.)

8 Mrs Parent occasionally allows her daughter to borrow her car. When Mrs Parent leaves the car at home after driving it, the amount of petrol in the tank is uniformly distributed between 10 litres and 50 litres. When her daughter leaves the car at home after having borrowed it, the amount of petrol in the tank is uniformly distributed between 0 litres and 20 litres. (Mrs Parent is none too pleased!)

(continued)

Exercises 9g (Miscellaneous)

1 The probability density function f of a continuous random variable X is given by

$$f(x) = \begin{cases} kx(2-x) & 0 \leqslant x \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

Show that $k = \frac{3}{2}$ and calculate

(i) the mean and variance of X,

(ii) $P(X < \frac{1}{2})$,

(iii) the probability that all of three independent observed values of X will be less than $\frac{1}{2}$,

(iv) $P(X > \frac{1}{4} | X < \frac{1}{2})$. [WJEC]

2 The continuous random variable X has probability density function f given by

$$f(x) = \begin{cases} k(4-x^2) & \text{for } 0 \leqslant x \leqslant 2, \\ 0 & \text{otherwise,} \end{cases}$$

where k is a constant. Show that $k = \frac{3}{16}$ and find the values of $E(X)$ and $Var(X)$.

Find the cumulative distribution function of X, and verify by calculation that the median value of X is between 0.69 and 0.70.

Find also $P(0.69 < X < 0.70)$, giving your answer correct to one significant figure. [UCLES]

3 The continuous random variable X has probability density function given by

$$f(x) = \begin{cases} \dfrac{k}{x} & \text{for } 1 \leqslant x \leqslant 9, \\ 0 & \text{otherwise} \end{cases}$$

where k is a constant. Giving your answers correct to three significant figures where appropriate,

(i) find the value of k, and find also the median value of X,

(ii) find the mean and variance of X,

(iii) find the cumulative distribution function, F, of X, and sketch the graph of $y = F(x)$. [UCLES]

4 The continuous random variable X has cumulative distribution function given by

$$F(x) = \begin{cases} 0, & \text{for } x < 0, \\ 2x - x^2, & \text{for } 0 \leqslant x \leqslant 1, \\ 1, & \text{for } x > 1. \end{cases}$$

(i) Find $P(X > \frac{1}{2})$.

(ii) Find the value of q such that $P(X < q) = \frac{1}{4}$.

(iii) Find the probability density function of X, and sketch its graph.

(iv) Find $E(X)$. [UCLES(P)]

5 The continuous random variable U has a uniform distribution on $0 < u < 1$. The random variable X is defined as follows:

$X = 2U$ when $U \leqslant \frac{3}{4}$,

$X = 4U$ when $U > \frac{3}{4}$.

(i) Give a reason why X cannot take values between $\frac{3}{2}$ and 3, and write down the values of $P(0 < X \leqslant \frac{3}{2})$ and $P(3 < X < 4)$.

(ii) Sketch the complete graph of the probability density function of X.

(iii) Find the lower quartile q of X, i.e. the value of q such that $P(X < q) = \frac{1}{4}$.

(iv) Three independent observations are taken of X. Find the probability that they all exceed q.

(v) Show that $E(X) = \frac{21}{16}$ and find $E(X^2)$. [UCLES]

6 The total amount of fuel used by a road haulage firm in a month is a random variable X (thousands of gallons) which has the following probability density function.

$$f(x) = \begin{cases} cx & 0 < x < 1, \\ c(3-x) & 1 \leqslant x < 3, \\ 0 & \text{otherwise,} \end{cases}$$

(a) Find the value of c.

(b) Find the probability that the firm uses less than 900 gallons in a month.

(c) Find the probability that the firm uses between 900 and 1600 gallons a month.

(d) Given that the firm used over 900 gallons in a particular month, find the probability that over 2000 gallons were used during the month.

(e) The supplier of the fuel charges the firm £1.20 per gallon for the first 900 gallons supplied per month, £1.10 per gallon for the next 700 gallons and £1.00 per gallon for the remainder. Find the probability that the monthly cost exceeds £2250. [AEB 90]

7 The amount of vegetables eaten by a family in a week is a random variable W kg. The probability density function is given by

$$f(w) = \begin{cases} \dfrac{20}{5^5} w^3(5-w) & 0 \leqslant w \leqslant 5, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the cumulative distribution function of W.

(continued)

(*b*) Find, to 3 decimal places, the probability that the family eats between 2kg and 4kg of vegetables in one week.

(*c*) Given that the mean of the distribution is $3\frac{1}{3}$, find, to 3 decimal places, the variance of W.

(*d*) Find the mode of the distribution.

(*e*) Verify that the amount, m, of vegetables which is such that the family is equally likely to eat more or less than m in any week is about 3.431 kg.

(*f*) Use the information above to comment on the skewness of the distribution.

[ULSEB]

8 The overall mark obtained by a ski-jumper is based on the distance, X metres, jumped in excess of 80 metres and marks for style, Y.

Assuming that X is a continuous random variable with probability density function

$$f(x) = \begin{cases} ax & 0 \leqslant x \leqslant 10, \\ 0 & \text{elsewhere,} \end{cases}$$

and that Y is a discrete random variable with probability function

$$g(y) = \begin{cases} by & y = 1, 2, 3, 4, 5, \\ 0 & \text{elsewhere,} \end{cases}$$

find a and b.

Evaluate

(a) the expected distance jumped in excess of 80 metres,

(b) the expected style marks obtained by the ski-jumper,

(c) the expected overall mark for the ski-jumper if the total mark, T, is given by $T = X^2 + \frac{1}{2}Y$.

Assuming X and Y are independent, determine the probability that the ski-jumper exceeds 85 metres and obtains more than 3 style marks for a particular jump. [AEB 91]

10 The normal distribution

Thank God we're normal, normal, normal
Thank God we're normal
Yes, this is our finest shower!

The Entertainer, John Osborne

The normal distribution is the most important of all distributions because it describes the situation in which very large values are rather rare, very small values are rather rare, but middling values are rather common. Since this is a good description of lots of things the 'normal' distribution is indeed normal!

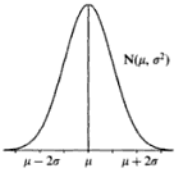
Here are some examples:

- Heights and weights (bean poles and Humpty Dumpties are not too common!)
- Times taken by students to run 100 m.
- The precise volumes of lager in 'pints' of lager at the local pub.

The distribution can also be applied as an approximation in the case of some discrete variables:

- Marks obtained by students on an A-level paper.
- The IQ scores of the population.

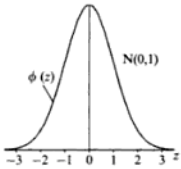
We will see later that the distribution can also be used as an approximation to the binomial and Poisson distributions.
Formally, a normal distribution is a unimodal symmetric continuous distribution having two parameters, μ (the mean) and σ^2 (the variance). Because of the symmetry, the mean is equal to both the mode and the median. As a shorthand we refer to a $N(\mu, \sigma^2)$ distribution – note that it is $N(\mu, \sigma^2)$ and *not* $N(\mu, \sigma)$.



10.1 The standard normal distribution

Since for any distribution, changes in μ and σ can be regarded as changes of location and scale, all normal distributions can be related to a single reference distribution, the so-called **standard normal** distribution, which has mean 0 and variance 1. Traditionally the random variable with this distribution is denoted by Z. Hence:

$$Z \sim N(0, 1)$$

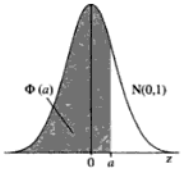


The pdf for Z is usually designated by ϕ (a lower-case Greek letter, pronounced 'phi'):

$$\phi(z) \propto e^{-\frac{1}{2}z^2} \qquad -\infty < z < \infty$$

The corresponding distribution function is denoted by Φ (the capital letter version of ϕ and also pronounced 'phi'):

$$\Phi(a) = P(Z \leqslant a) = P(Z < a) = \int_{-\infty}^{a} \phi(z)\ dz$$

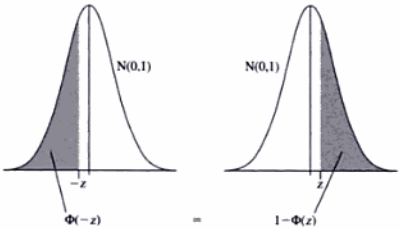


10.2 Tables of $\Phi(z)$

Although $\phi(z)$ looks simple, it cannot be integrated explicitly and so tables are required for $\Phi(z)$. The tables available vary in style quite considerably. You should make sure that you are familiar with the tables supplied for your particular examination.

Most tables give the values of either $\Phi(z)$ or $1 - \Phi(z)$, and generally do so only for non-negative values of z. The tables may refer to $\Phi(z)$ as $P(z)$ or to $1 - \Phi(z)$ as $Q(z)$, as shown in the diagrams.

The tables given at the back of this book are tables of $\Phi(z)$ for $z \geqslant 0$. Values for negative values of z can be obtained by using the symmetry of the distribution:



$$\Phi(-z) = 1 - \Phi(z) \qquad (10.1)$$

Here is an extract from the first column of the table of values of $\Phi(z)$ given in the Appendix (p. 439):

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
0.0	0.5000	1.0	0.8413	2.0	0.9772
0.2	0.5793	1.2	0.8849	2.2	0.9861
0.4	0.6554	1.4	0.9192	2.4	0.9918
0.6	0.7257	1.6	0.9452	2.6	0.9953
0.8	0.7881	1.8	0.9641	2.8	0.9974

The first entry in the table, $\Phi(0) = 0.5000$ is one we already knew! All normal distributions are symmetric about their mean, and the standard normal distribution has mean 0.

Example 1

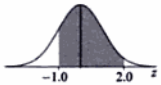
$$\begin{aligned} P(1.0 < Z < 2.0) &= \Phi(2.0) - \Phi(1.0) \\ &= 0.9772 - 0.8413 \\ &= 0.1359 \end{aligned}$$

Using $Q(z)$, we would have:

$$P(1.0 < Z < 2.0) = Q(1.0) - Q(2.0) = 0.1587 - 0.0228 = 0.1359$$

Example 2

$$\begin{aligned} P(-1.0 < Z < 2.0) &= \Phi(2.0) - \Phi(-1.0) \\ &= 0.9772 - \{1 - \Phi(1.0)\} \\ &= 0.9772 - 1 + 0.8413 \\ &= 0.8185 \end{aligned}$$



Using $Q(z)$, we would have:

$$\begin{aligned} P(-1.0 < Z < 2.0) &= \{1 - Q(1.0)\} - Q(2.0) \\ &= (1 - 0.1587) - 0.0228 = 0.8185 \end{aligned}$$

Every problem involving normal distributions can be solved using either of Φ or Q. This is the last example for which we will provide both solutions, since there is no essential difference between one solution and the other.

This problem also illustrates the limitations of 4-figure tables. The quantities 0.9772 and 0.8413, which are correct to four decimal places, would be 0.97725 and 0.84134 if expressed to five decimal places. With the extra accuracy the final result would be 0.81859 which rounds to 0.8186 and not to our stated answer of 0.8185. However, there is no need to worry too much about this! A reasonable tolerance is always built into marking schemes, and in real life one would be quoting only one or two significant figures – not four. We can't tell the difference between 82% and 80%!

Example 3

$$\begin{aligned} P(|Z| > 1.2) &= P(Z > 1.2) + P(Z < -1.2) \\ &= \Phi(-1.2) + \{1 - \Phi(1.2)\} \\ &= \{1 - \Phi(1.2)\} + \{1 - \Phi(1.2)\} \\ &= 2\{1 - \Phi(1.2)\} \\ &= 2(1 - 0.8849) = 2 \times 0.1151 = 0.2302 \end{aligned}$$

Example 4

The random variable Z has a standard normal distribution. Determine the value of a, where $P(Z < a) = 0.9953$.

From the table we see that $\Phi(2.6) = 0.9953$: hence $a = 2.6$.

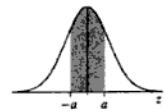
Example 5

The random variable Z has a standard normal distribution.
Determine the value of a, where $P(Z > a) = 0.2743$.

If $P(Z > a) = 0.2743$, then $P(Z < a) = 1 - 0.2743 = 0.7257$. Scanning through the tables we see that $P(Z < 0.6) = 0.7257$. Hence $a = 0.6$.

Example 6

The random variable Z has a standard normal distribution.
Determine the value of a, where $P(|Z| < a) = 0.5762$.



If $P(|Z| < a) = 0.5762$, then $P(0 < Z < a) = \frac{1}{2}(0.5762) = 0.2881$ and hence $P(Z < a) = 0.5 + 0.2881 = 0.7881$. The tables show that $\Phi(0.8) = 0.7881$. Hence $a = 0.8$.

Example 7

The random variable Z has a standard normal distribution.
Determine the value of a, where $P(a < Z < 1.6) = 0.7865$.

From the tables we know that $P(Z < 1.6) = 0.9452$.
Since:

$$P(a < Z < 1.6) = P(Z < 1.6) - P(Z < a)$$

it follows that:

$$\begin{aligned} P(Z < a) &= P(Z < 1.6) - P(a < Z < 1.6) \\ &= 0.9452 - 0.7865 \\ &= 0.1587 \end{aligned}$$

This is less than 0.5 and therefore corresponds to a negative value of z. We need to use Equation (10.1). Instead of finding the value of z corresponding to 0.1587, we find the value that corresponds to $1 - 0.1587 = 0.8413$. From the table we find that this is the value 1.
Since $\Phi(-z) = 1 - \Phi(z)$, we have therefore deduced that $\Phi(-1) = 1 - 0.8413 = 0.1587$.
The required value of a is -1.

Exercises 10a

In Questions **1–5**, the random variable Z has a standard normal distribution, with mean zero and variance 1.
Use the table given in Section 10.2 (p. 243) to answer the following questions.

1 Find:
(i) $P(Z < 1.2)$, (ii) $P(Z > 1.8)$,
(iii) $P(Z < -1.4)$, (iv) $P(Z < -0.8)$.

2 Find:
(i) $P(2.2 < Z < 2.8)$, (ii) $P(-1.2 < Z < 0.4)$,
(iii) $P(-1.8 < Z < -0.2)$.

3 Find:
(i) $P(|Z| < 0.8)$, (ii) $P(|Z| > 1.6)$,
(iii) $P(0.6 < |Z| < 2.2)$.

4 Find a such that:
(i) $P(Z < a) = 0.9192$, (ii) $P(Z < a) = 0.3446$,
(iii) $P(Z > a) = 0.8849$, (iv) $P(Z > a) = 0.0047$,
(v) $P(1 < Z < a) = 0.1039$,
(vi) $P(a < Z < -0.8) = 0.1760$.

5 Find a such that:
(i) $P(|Z| < a) = 0.4514$,
(ii) $P(|Z| > a) = 0.1096$.

10.3 Probabilities for other normal distributions

Suppose $X \sim N(\mu, \sigma^2)$ and let Y be defined by $Y = aX + b$, where a and b are constants. It can be shown that the distribution of Y is also normal, in fact

$$Y \sim N(a\mu + b,\ a^2\sigma^2)$$

If we take $a = \frac{1}{\sigma}$ and $b = -\frac{\mu}{\sigma}$, and put $Z = Y$, so that

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

then $Z \sim N(0, 1)$. This is equivalent to the change of location and scale referred to in Section 10.1. Thus:

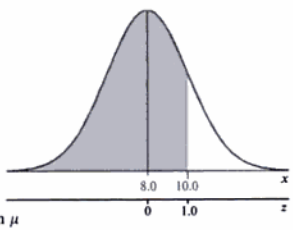
$$\begin{aligned} P(X < x) = P(X - \mu < x - \mu) &= P\left(\frac{X - \mu}{\sigma} < \frac{x - \mu}{\sigma}\right) \\ &= P\left(Z < \frac{x - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{x - \mu}{\sigma}\right) \end{aligned}$$

The relevant value for Z is simply the value of $\frac{X - \mu}{\sigma}$ when X is replaced by the value of interest.

As an example, if $X \sim N(8, 4)$ and we want $P(X < 10)$, then this is given by:

$$\Phi\left(\frac{10 - 8}{2}\right) = \Phi(1) = 0.8413$$

The link between the normal distribution for X and the standard normal distribution for Z is conveniently summarised by showing two scales.



The value of zero for z always occurs under the mean μ for x. A value of k for z would occur under the value $\mu + k\sigma$ for x, where σ^2 is the variance of the distribution of X.

Example 8

The random variable $X \sim N(3.4, 0.09)$.
Determine $P(X < 3.1)$.

Now $X < 3.1$ corresponds to $Z < \frac{3.1 - 3.4}{\sqrt{0.09}} = -1$, so that:

$$\begin{aligned} P(X < 3.1) &= P(Z < -1) \\ &= 1 - P(Z < 1) \\ &= 1 - 0.8413 \\ &= 0.1587 \end{aligned}$$

Example 9

The random variable $X \sim N(21, 16)$.
Determine $P(17.8 < X < 29.0)$.

Now $X < 29.0$ corresponds to $Z < \dfrac{29.0 - 21}{\sqrt{16}} = 2$

and $X > 17.8$ corresponds to $Z > \dfrac{17.8 - 21}{\sqrt{16}} = -0.8$, so that:

$$\begin{aligned} P(17.8 < X < 29.0) &= P(-0.8 < Z < 2) \\ &= \Phi(2) - \Phi(-0.8) \\ &= \Phi(2) - \{1 - \Phi(0.8)\} \\ &= 0.9772 - 1 + 0.7881 \\ &= 0.7653 \end{aligned}$$

Example 10

The random variable $X \sim N(50, 36)$.
Denoting $E(X)$ by μ, determine $P(|X - \mu| > 4.8)$.

We are given that X has mean 50. Hence $\mu = 50$.

$$\begin{aligned} P(|X - \mu| > 4.8) &= P\left(\frac{|X - \mu|}{\sqrt{36}} > \frac{4.8}{\sqrt{36}}\right) \\ &= P(|Z| > 0.8) \\ &= 1 - P(|Z| < 0.8) \\ &= 1 - \{P(Z < 0.8) - P(Z < -0.8)\} \\ &= 1 - \{0.7881 - (1 - 0.7881)\} \\ &= 0.4238 \end{aligned}$$

Project

This practical needs a sensitive pair of scales or balance. ***The aim is to*** *study the variability in the masses of some packaged goods.* ***A good choice*** *would be packets of crisps – how variable are their masses?* ***Does it seem*** *as though a normal distribution is appropriate?*

This project can be extended to compare different brands ***and flavours.*** *Is there any evidence that some brands are more variable than* ***others? Is*** *there a difference between plain crisps and flavoured crisps? In each* ***case*** *represent your data using a histogram. For comparing flavours* ***or types,*** *box–whisker diagrams will be useful.*
Some ways of formally testing for differences will be discussed ***later in*** *Chapters 12 and 13.*

Exercises 10b

Use the table given in Section 10.2 (p. 243) to answer the following questions.

1 Given that $X \sim N(12, 9)$, find:
(i) $P(X > 15)$, (ii) $P(X < 16.8)$,
(iii) $P(X < 8.4)$, (iv) $P(X > 9.6)$.

2 Given that $X \sim N(50, 100)$, find:
(i) $P(36 < X < 62)$, (ii) $P(40 < X < 50)$,
(iii) $P(56 < X < 70)$, (iv) $P(38 < X < 42)$.

3 Given that $X \sim N(1.6, 4)$, find:
(i) $P(X > 0)$, (ii) $P(X < -1.6)$,
(iii) $P(|X| < 2)$, (iv) $P(0 < X < 2)$.

4 Given that $X \sim N(-4, 25)$, find:
(i) $P(X > 0)$, (ii) $P(-5 < X < -2)$,
(iii) $P(-2 < X < 1)$, (iv) $P(|X| > 1)$.

5 IQ scores are normally distributed with mean 100 and standard deviation 15.
Determine the proportion of people with an IQ:
(a) below 118,
(b) above 112,
(c) below 94,
(d) above 73,
(e) between 100 and 112,
(f) between 73 and 118,
(g) between 73 and 94.

10.4 Finer detail in the tables of $\Phi(z)$

The table of $\Phi(z)$ in Section 10.2 was very abbreviated. We now give a (very short) section from the full table given in the Appendix (p. 439):

z	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
											ADD								
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359	4	8	12	16	20	24	28	32	36
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753	4	8	12	16	20	24	28	32	36
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141	4	8	12	15	19	23	27	31	35
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517	4	7	11	15	19	22	26	30	34

As an example of the use of this table, suppose we require $P(Z < 0.100)$. We look first at the row with 0.1 in the first column. Since the value in the second decimal place of 0.100 is a '0', we look in the column headed '0' and find the value '.5398' implying a probability of 0.5398.

If we require instead $P(Z < 0.140)$ then we look at the value in the row labelled 0.1, and the column headed '4', which gives 0.5557. Some further examples:

$$P(Z < 0.070) = 0.5279$$
$$P(Z < 0.260) = 0.6026$$

All the previous examples referred to cases where the z had a '0' in the third decimal place. For other cases we need to modify the value found using the values given in the right-hand section of the table.

For example, if we require $P(Z < 0.175)$ then we must first find $P(Z < 0.170)$. Looking in the main part of the table we find '.5675'. For the adjustment due to the third decimal place (which is 5) we look at the right-hand column headed '5'. The value in this column for the row labelled 0.1 is 20. This value '20' is a shorthand for 0.0020. The instruction at the top of this section is ADD. Therefore the required probability is $0.5675 + 0.0020 = 0.5695$. Here are some further examples:

$$P(Z < 0.246) = 0.5948 + 0.0023 = 0.5971$$
$$P(Z < 0.302) = 0.6179 + 0.0007 = 0.6186$$

The tables can also be used 'in reverse'. For example, to find the value of z which is such that $P(Z < z) = 0.6443$, we look through the table for the probability 0.6443 and observe that this corresponds to $z = 0.37$. A negative value for z would be signalled by a probability less than 0.5000. For example, if $P(Z < z) = 0.3783$, we look instead for the probability $1 - 0.3783 = 0.6217$, which corresponds to $z = 0.31$. The required z-value is therefore -0.31.

An example of a case where the given probability does not appear in the table is given by part (v) of Example 11.

Notes

- The tables report probabilities as, for example, '.5948'. This is done to save space in the table: the probability should be reported as '0.5948'.
- It is easy to get confused over the decimal places. One guide is that after adding the adjustment from the right-hand side of the table, the sum should not be larger than the next item in the body of the table. For example, if we had calculated $P(Z < 0.302)$ as $0.6179 + 0.0070 = 0.6249$, then we would know there was an error because this value is greater than $P(Z < 0.31) = 0.6217$.
- When the tables are of $Q(z)$, the adjustment factors given in the right-hand section will have to be *subtracted* from the value given in the body of the table.
- Some tables do not contain the fine detail provided by the right-hand side of our table. For these tables the user has to interpolate 'by hand'.
- Some tables use a type of shorthand to deal with probabilities that are very close to either 0 or 1. Thus:

  $.0^43 \rightarrow 0.000\,03$

  $.9^37 \rightarrow 0.9997$

Example 11

The random variable Z has a normal distribution with mean 0 and variance 1.
Determine (i) $P(Z < 0.27)$, (ii) $P(Z < 0.345)$, (iii) $P(Z > 0.004)$,
(iv) the value of a for which $P(Z < a) = 0.6217$,
(v) the value of b for which $P(Z < b) = 0.6000$.

(i) From the table, $\Phi(0.27) = 0.6064$.
(ii) From the main part of the table, $\Phi(0.34) = 0.6331$. From the fifth column of the additional part of the table we need to add '19' (i.e. 0.0019) and so the required probability is $0.6331 + 0.0019 = 0.6350$.
(iii) From the table $\Phi(0.004) = 0.5000$ with an addition of '16', so $P(Z < 0.004) = 0.5016$. Hence $P(Z > 0.004) = 1 - 0.5016 = 0.4984$.
(iv) Scanning through the main part of the table we find that $\Phi(0.31) = 0.6217$. Thus $a = 0.31$.
(v) Scanning through the table we see that $\Phi(0.25) = 0.5987$, while $\Phi(0.26) = 0.6026$. The required value for a must lie between 0.25 and 0.26. If we use the supplementary part of the table we find that $\Phi(0.253) = 0.5999$ and $\Phi(0.254) = 0.6002$ so the value of a must be about 0.2533.

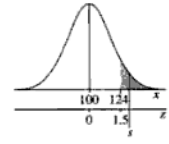
Example 12

The random variable X has a normal distribution with mean 12 and variance 5.
Determine $P(X < 13)$.

$$P(X < 13) = P\left(\frac{X-12}{\sqrt{5}} < \frac{13-12}{\sqrt{5}}\right) = P\left(Z < \frac{1}{\sqrt{5}}\right) = \Phi(0.4472)$$
$$= 0.673 \text{ (to 3 d.p.)}$$

Example 13

The entrance qualification for membership of a society is that the aspiring member should score highly on a particular IQ test. The scores obtained on the test have a normal distribution, with mean 100.0 and standard deviation 16. All those obtaining 124.0 or more automatically join the society.
Determine the median score of the members of the society.



The median score obtained by those taking the test is 100.0, since the distribution of scores is normal. However, it is the median of those who are admitted that is required. We need first to find the proportion of those taking the test that obtain scores of 124.0 or more and then divide this proportion in two. As usual, a diagram helps.

The proportion of those taking the test that are admitted to the society is $1 - \Phi\left(\frac{124 - 100}{16}\right) = 1 - \Phi(1.5) = 0.0668$. We therefore require the score that is exceeded by $\frac{1}{2}(0.0668) = 0.0334$ of the population.

Using tables of $\Phi(z)$, we need to find the value of z corresponding to $\Phi(z) = 1 - 0.0334 = 0.9666$. The required value is 1.833 and the required median score, s, is therefore the solution of:

$$\frac{s - 100}{16} = 1.833$$

The median score of the members of the society is therefore:
$s = 100 + 16 \times 1.833 \approx 129.3$.

Exercises 10c

Use the tables in the Appendix, or the tables that you will use in your examination, to answer the following questions.

1 Given that $Z \sim N(0, 1)$, find:
(i) $P(Z < 0.932)$, (ii) $P(Z > 1.235)$,
(iii) $P(Z < -1.414)$, (iv) $P(Z > -0.519)$.

2 Given that $X \sim N(0, 1)$, find:
(i) $P(X > 3.213)$, (ii) $P(X < 3.615)$,
(iii) $P(X < -2.841)$, (iv) $P(X > -2.818)$.

3 Given that $Y \sim N(3.7, 2.4)$, find:
(i) $P(Y > 4)$, (ii) $P(Y < 4.5)$,
(iii) $P(3.1 < Y < 4.2)$, (iv) $P(2.8 < Y < 3.5)$.

4 Given that $X \sim N(23, 12)$, find:
(i) $P(X < 25)$, (ii) $P(20 < X < 25)$,
(iii) $P(X > 27)$, (iv) $P(23 < X < 30)$.

5 Given that $X \sim N(3, 4)$, find:
(i) $P(3X < 7)$, (ii) $P(\frac{1}{2}X < 2)$,
(iii) $P(2X + 1 > 10)$, (iv) $P(3 - X < 2)$.

6 The mass of a small loaf of bread is normally distributed with mean 500 g and standard deviation 20 g.
Find the probability that a randomly chosen loaf has a mass:
(i) not exceeding 475 g,
(ii) not less than 495 g,
(iii) at most 510 g,
(iv) at least 515 g.

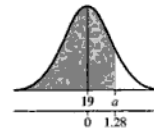
7 Chicken eggs have mean mass 60 g with standard deviation 15 g, and the distribution of their masses may be taken to be normal. Eggs of less than 45 g are classed as 'small'. The remainder are classed as either 'standard' or 'large'. It is desired that these two classifications should occur with approximately equal frequency.
Suggest the mass at which the division between standard and large should be made.

Example 15

The random variable $X \sim N(19, 49)$.
Determine the value of a which is such that $P(X < a) = 0.90$.

$$P(X < a) = P\left(\frac{X-\mu}{\sigma} < \frac{a-\mu}{\sigma}\right)$$
$$= \Phi\left(\frac{a-\mu}{\sigma}\right)$$

We want to find the value of a such that $\Phi\left(\frac{a-19}{7}\right) = 0.90$, which means that there is 10% in the upper tail. But we know, from the table, that the upper 10% point of a standard normal distribution is 1.282. The inescapable conclusion is that:



$$\frac{a-19}{7} = 1.282$$

which implies that $a = 19 + (7 \times 1.282) = 27.974$. Hence, to 1 d.p., $a = 28.0$.

Example 16

A machine is supposed to cut up logs into pieces that are each 2 metres long. However, the machine is an old one, and while the pieces that it produces do have an *average* length of 2 metres, 10% of the pieces that it produces are less than 1.95 metres long.
Assuming that the lengths produced are normally distributed, determine the proportion of the pieces that are longer than 2.10 metres.

Let X be the random variable corresponding to the length of a log.
We have the information:

$$X \sim N(2.00, \sigma^2)$$
$$P(X < 1.95) = 0.10$$

We need to find $P(X > 2.10)$ and we must evidently first determine the value of σ. To do this we relate the known probability to the standard normal distribution.

Now:

$$P(X < 1.95) = \Phi\left(\frac{1.95-2.00}{\sigma}\right)$$

and also:

$$0.10 = \Phi(-1.282)$$

using the information in the table of percentage points. Hence it must be the case that:

$$\Phi(-1.282) = \Phi\left(\frac{1.95-2.00}{\sigma}\right)$$

which implies that:

$$-1.282 = \frac{1.95-2.00}{\sigma}$$

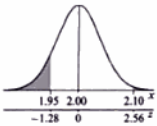
Hence:

$$\sigma = \frac{-0.05}{-1.282}$$

$$= 0.0390$$

We now know that $X \sim N(2.00, 0.0390^2)$. In order to find $P(X > 2.10)$ we first find $P(X < 2.10)$ which equals:

$$\Phi\left(\frac{2.10 - 2.00}{0.039}\right) = \Phi(2.564)$$



Using the tables in the Appendix, the corresponding probability is found to be 0.9948. The required probability is therefore $1 - 0.9948 = 0.0052$, in other words just over 0.5% of the pieces of wood have lengths greater than 2.10 metres.

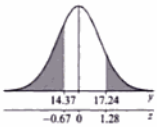
Example 17

The random variable Y has a normal distribution with mean μ and variance σ^2.
Given that 10% of the values of Y exceed 17.24 and that 25% of the values of Y are less than 14.37, find the values of μ and σ.

We know from the tables of percentage points that the upper 10% point of a standard normal distribution is 1.282 and the lower 25% point is −0.674. The values of μ and σ are therefore the solutions of:

$$\frac{17.24 - \mu}{\sigma} = 1.282$$

$$\frac{14.37 - \mu}{\sigma} = -0.674$$



Multiplying through by σ, and subtracting, we get:

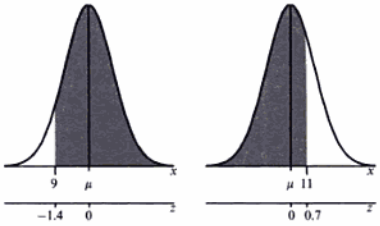
$$\begin{array}{rcl} 17.24 - \mu & = & 1.282\sigma \\ 14.37 - \mu & = & -0.674\sigma \\ \hline 2.87 & = & 1.956\sigma \end{array}$$

Thus $\sigma = 1.47$ and hence $\mu = 17.24 - 1.282\sigma = 15.4$.

Example 18

The random variable X has a normal distribution. It is known that $P(X > 9) = 0.9192$ and that $P(X < 11) = 0.7580$.
Determine $P(X > 10)$.

We are not told the values of μ or σ, so must begin by finding these, using the two pieces of information that we do have. An initial sketch shows that μ must lie between 9 and 11, and, since 0.9192 is larger than 0.7580, the mean must be slightly above 10.



Now $1 - 0.9192 = 0.0808$, so we need to search the tables of $\Phi(z)$ to find the z-values corresponding to $\Phi(z) = 0.0808$ and $\Phi(z) = 0.7580$. These values are -1.400 and 0.700, respectively. We now solve the simultaneous equations:

$$\frac{9-\mu}{\sigma} = -1.400$$

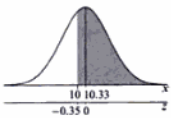
$$\frac{11-\mu}{\sigma} = 0.700$$

Multiplying through by σ we get:

$$9 - \mu = -1.400\sigma$$
$$11 - \mu = 0.700\sigma$$

Subtracting one equation from the other we get $2 = 2.100\sigma$, so that $\sigma = \frac{20}{21}$ (approximately).
Substituting in either equation we get $\mu = \frac{31}{3}$.
We want $P(X > 10)$, so first find the corresponding value of z:

$$z = \frac{10 - \frac{31}{3}}{\frac{20}{21}} = -\frac{7}{20}$$



Hence:

$$P(X > 10) = 1 - \Phi(-\tfrac{7}{20}) = 1 - \{1 - \Phi(0.35)\} = 0.6368$$

The probability that X exceeds 10 is 0.637 (to 3 d.p.).

Exercises 10d

Use the tables of percentage points in Section 10.5 (p. 252), or the tables that you will use in your examination, to answer the following questions.

1 Given that $Z \sim N(0, 1)$, find a such that:
(i) $P(Z < a) = 0.97$, (ii) $P(Z > a) = 0.05$,
(iii) $P(Z < a) = 0.001$, (iv) $P(Z > a) = 0.99$,
(v) $P(Z > a) = 0.0001$, (vi) $P(Z < a) = 0.999$.
Give your answers to 3 decimal places.

2 Given that $X \sim N(20, 25)$, find a such that:
(i) $P(X < a) = 0.97$, (ii) $P(X > a) = 0.05$,
(iii) $P(X < a) = 0.001$, (iv) $P(X > a) = 0.99$,
(v) $P(X > a) = 0.0001$, (vi) $P(X < a) = 0.999$.
Give your answers to 3 decimal places.

3 Given that $X \sim N(\mu, 2.5)$ and that $P(X > 3.5) = 0.970$, find μ.

4 Given that $X \sim N(\mu, 0.5)$ and that $P(X < -1.2) = 0.050$, find μ.

5 Given that $X \sim N(32.4, \sigma^2)$ and that $P(X > 45.2) = 0.300$, find σ.

6 Given that $X \sim N(-7.21, \sigma^2)$ and that $P(X < 0) = 0.900$, find σ.

7 Given that $X \sim N(\mu, \sigma^2)$, $P(X > 0) = 0.800$ and $P(X < 5) = 0.700$, find μ and σ.

10.8 General properties

Gauss proposed that if one makes a number of independent observations on the value of some quantity then:

1. A positive error of given magnitude should be as probable as a negative error of the same magnitude.
2. Large errors should be less likely than small errors.
3. The mean of the observations should be the most likely value of the quantity being measured.

The consequence of (1) is that the distribution is *symmetric* and therefore has mean equal to median. The consequence of (2) is that both are equal to the mode. The three propositions taken together led Gauss to deduce that a random error, X, was likely to have pdf:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad -\infty < x < \infty$$

which is, in fact, the pdf for a $N(\mu, \sigma^2)$ random variable.
Here π and e have their usual values (3.14... and 2.78...) and the two parameters μ and σ^2 are the mean and the variance of the distribution.

Notes

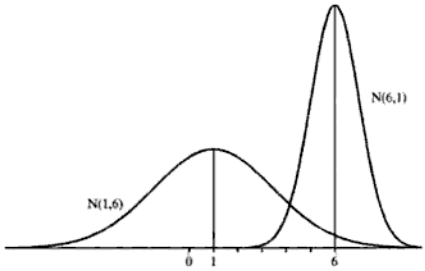
- Although the nominal range for X is infinite, most values (about 99.7%!) of X fall in the interval:

  $$\mu - 3\sigma, \mu + 3\sigma$$

  while 95% fall in the interval:

  $$\mu - 2\sigma, \mu + 2\sigma$$

  which is sometimes referred to as the **2 sigma rule**.
- For a normal distribution, changes in the values of μ and σ can be regarded simply as changes of location and scale. The fundamental shape is unaltered, though its appearance will vary.



Calculator practice

If your calculator is able to evaluate integrals numerically then you will be able to satisfy yourself that the area under the normal curve is indeed equal to 1. Choose any values that you please for μ and σ. Try a variety of upper and lower limits for the integration and observe the way in which the area slowly approaches 1 as the limits become further apart.

10.9 Linear combinations of independent normal random variables

Normal random variables behave very nicely!

If X and Y are two independent normally distributed random variables, and if a and b are constants, then $aX + bY$ also has a normal distribution.

The mean and variance of the resulting normal distribution follows from the results of Section 9.4 (p. 226):

$$\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$$
$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)$$

Example 19

The random variables X and Y are independent with $X \sim \mathrm{N}(2, 3)$ and $Y \sim \mathrm{N}(6, 4)$. The random variable W is defined by $W = X + Y$. Determine (i) $\mathrm{P}(W > 8)$, (ii) $\mathrm{P}(W > 10)$.

We begin by determining the mean and variance of W:

$$\mathrm{E}(W) = \mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y) = 2 + 6 = 8$$
$$\mathrm{Var}(W) = \mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 3 + 4 = 7$$

so $W \sim \mathrm{N}(8, 7)$.

(i) Since W has a normal distribution with mean 8 we can write down immediately that $\mathrm{P}(W > 8) = \frac{1}{2}$.

(ii) To find $\mathrm{P}(W > 10)$ we need the z-value corresponding to $w = 10$:

$$z = \frac{10 - 8}{\sqrt{7}} = 0.756$$

Hence:

$$\mathrm{P}(W > 10) = 1 - \Phi(0.756) = 1 - 0.7752 = 0.225 \text{ (to 3 d.p.).}$$

Example 20

The random variables X and Y are independent with $X \sim \mathrm{N}(3, 1)$ and $Y \sim \mathrm{N}(7, 5)$. The random variable W is defined by $W = Y - 2X$. Determine the probability that W takes a positive value.

We begin by determining the mean and variance of W:

$$\mathrm{E}(W) = \mathrm{E}(Y - 2X) = \mathrm{E}(Y) - 2\mathrm{E}(X) = 7 - (2 \times 3) = 1$$
$$\mathrm{Var}(W) = \mathrm{Var}(Y - 2X) = \mathrm{Var}(Y) + (-2)^2\,\mathrm{Var}(X)$$
$$= 5 + (4 \times 1) = 9$$

so $W \sim \mathrm{N}(1, 9)$.

To find $\mathrm{P}(W > 0)$ we need the z-value corresponding to $w = 0$:

$$z = \frac{0 - 1}{\sqrt{9}} = -\tfrac{1}{3}$$

Hence:

$$\mathrm{P}(W > 0) = 1 - \Phi(-\tfrac{1}{3}) = \Phi(\tfrac{1}{3}) = 0.6304$$

The probability that W takes a positive value is 0.630 (to 3 d.p.).

Example 21

The random variables X and Y are independent with $X \sim N(16,4)$ and $Y \sim N(8, 9)$.
Find (i) $P(X - 2Y > 0)$, (ii) $P(X + 2Y < 30)$.

(i) Denote $X - 2Y$ by V. Then V has a normal distribution with mean:

$$E(X) - 2E(Y) = 16 - (2 \times 8) = 0$$

and variance:

$$\text{Var}(X) + \{(-2)^2 \times \text{Var}(Y)\} = 4 + (4 \times 9) = 40$$

We require $P(V > 0)$. Since $E(V) = 0$, the required probability is $\frac{1}{2}$.

(ii) Denote $X + 2Y$ by W. Then W has a normal distribution with mean:

$$E(X) + 2E(Y) = 16 + (2 \times 8) = 32$$

and variance:

$$\text{Var}(X) + \{2^2 \times \text{Var}(Y)\} = 4 + (4 \times 9) = 40$$

We require $P(W < 30)$. The corresponding z-value is $\frac{30 - 32}{\sqrt{40}} = -0.316$. The required probability is therefore $\Phi(-0.316) = 0.376$ (to 3 d.p.).

Example 22

The random variable $X \sim N(16, 4)$. The independent random variables X_1 and X_2 have the same distribution as X. The random variable $\bar{X}$ is defined by $\bar{X} = \frac{1}{2}(X_1 + X_2)$.
Determine $P(\bar{X} > 18)$.

The distribution of $\bar{X}$ is normal with mean $\frac{1}{2}(16 + 16) = 16$ and variance $(\frac{1}{4})(4 + 4) = 2$. The z-value of interest is therefore $\frac{18 - 16}{\sqrt{2}} = 1.414$ and hence the required probability is

$1 - \Phi(1.414) = 1 - 0.9213 = 0.0787 = 0.079$ (to 3 d.p.).

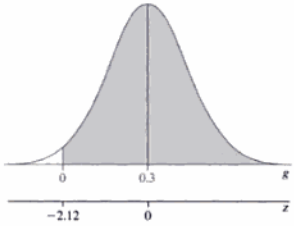
Example 23

The diameter in mm, X, of the circular mouth of a bottle has a normal distribution with mean 20 and standard deviation 0.1. The diameter in mm, Y, of the circular cross-section of a glass stopper, has a normal distribution with mean 19.7 and standard deviation 0.1.
Determine the probability that a randomly chosen stopper will fit in the mouth of a randomly chosen bottle.

We require $P(X > Y)$, which looks like a rather difficult quantity to calculate. However, $P(X > Y) = P(X - Y > 0)$ and $X - Y$ is a linear combination of independent normal random variables and therefore has a

normal distribution. Let $G = X - Y$, where G mm is the gap between the diameters of the stopper and the mouth. Now
$\mathrm{E}(G) = \mathrm{E}(X) - \mathrm{E}(Y) = 20 - 19.7 = 0.3$
and $\mathrm{Var}(G) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 0.1^2 + 0.1^2 = 0.02$, so:

$$G \sim \mathrm{N}(0.3,\ 0.02)$$



The z-value of interest is $\dfrac{0 - 0.3}{\sqrt{0.02}} = -2.121$, so:

$$\mathrm{P}(G > 0) = 1 - \mathrm{P}(G < 0) = 1 - \Phi(-2.121) = \Phi(2.121) = 0.9830$$

The probability that a randomly chosen stopper will fit inside the mouth of a randomly chosen bottle is 0.983 (to 3 d.p.).

Extension to more than two variables

This follows immediately. Suppose that W, X and Y are independent normal random variables, and that a, b and c are constants. Consider the random variable U defined by:

$$U = aW + bX + cY$$

and let $V = aW + bX$. From the previous result we know that V also has a normal distribution. Thus we can write:

$$U = V + cY$$

and, since the right-hand side is once again a linear combination of independent normal random variables, it follows that U has a normal distribution. The results of Section 9.4 (p. 226) give the mean and variance of U:

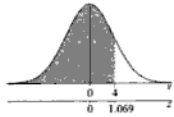
$$\mathrm{E}(U) = a\mathrm{E}(W) + b\mathrm{E}(X) + c\mathrm{E}(Y)$$
$$\mathrm{Var}(U) = a^2\mathrm{Var}(W) + b^2\mathrm{Var}(X) + c^2\mathrm{Var}(Y)$$

This argument can be extended indefinitely.

A linear combination of any number of independent normal random variables has a normal distribution.

Example 24

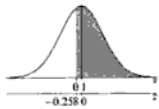
The independent normal random variables W, X and Y each have expectation 0 and variance 1. The random variable V is defined by $V = W + 2X + 3Y$.
Determine the probability that V is less than 4.



The random variable V has expectation 0 and variance $1 + (2^2 \times 1) + (3^2 \times 1) = 14$. The z-value of interest is $\dfrac{4-0}{\sqrt{14}} = 1.069$.
The required probability is $\Phi(1.069) = 0.857$ (to 3 d.p.).

Example 25

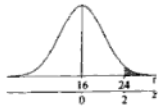
The independent random variables W, X and Y are such that $W \sim N(3, 5)$, $X \sim N(5, 5)$ and $Y \sim N(7, 5)$.
Determine the probability that the sum of W and X exceeds Y.



Let $V = W + X - Y$. Then V has a normal distribution with mean equal to $3 + 5 - 7 = 1$ and variance equal to $5 + 5 + \{(-1)^2 \times 5\} = 15$. We want $P(V > 0)$. The z-value of interest is therefore $\dfrac{0-1}{\sqrt{15}} = -0.258$. The required probability is $1 - \Phi(-0.258) = \Phi(0.258) = 0.602$ (to 3 d.p.).

Example 26

The normal random variable X has expectation and variance equal to 1.
Determine the probability that 16 independent values of X have a total in excess of 24.



Denote the random variables corresponding to the 16 values by $X_1, X_2, \ldots, X_{16}$. Let T be given by $T = X_1 + \cdots + X_{16}$. Then T has a normal distribution with mean and variance given by:

$$E(T) = E(X_1) + \cdots + E(X_{16}) = 16$$

$$\text{Var}(T) = \text{Var}(X_1) + \cdots + \text{Var}(X_{16}) = 16$$

Hence $T \sim N(16, 16)$. The z-value of interest is $\dfrac{24-16}{\sqrt{16}} = 2$. The required probability is $1 - \Phi(2) = 0.0228 = 0.023$ (to 3 d.p.).

Example 27

The total mass (in g) of a packet of biscuits is made up of the mass (in g) of the packaging, Y, and the masses (in g) of the fifteen biscuits $X_1, \ldots, X_{15}$. The mass of a biscuit has a normal distribution, with mean 30 and standard deviation 1, while the mass of the packaging has a normal distribution with mean 5 and standard deviation 0.2.
Determine the probability that the total mass of a packet of biscuits lies between 450 and 460 g.

Let W denote the total mass. So:

$$W = Y + X_1 + \cdots + X_{15}$$

and hence:

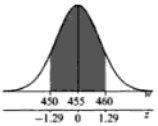
$$\mathrm{E}(W) = 5 + (15 \times 30) = 455$$

Assuming that the masses of the sixteen items are independent of one another, we also have:

$$\mathrm{Var}(W) = 0.2^2 + (15 \times 1^2) = 15.04$$

so that $W \sim \mathrm{N}(455, 15.04)$.

The z-values corresponding to the values of interest, 450 and 460, are $\frac{460 - 455}{\sqrt{15.04}}$ and $\frac{450 - 455}{\sqrt{15.04}}$, respectively. These expressions simplify to 1.289 and -1.289, respectively.



Hence:

$$\begin{aligned}\mathrm{P}(450 < W < 460) &= \mathrm{P}(W < 460) - \mathrm{P}(W < 450)\\ &= \Phi(1.289) - \Phi(-1.289)\\ &= 2\Phi(1.289) - 1\\ &= (2 \times 0.9013) - 1\\ &= 0.8026\end{aligned}$$

The probability that the total mass of a packet of biscuits lies between 450 and 460 g is 0.803 (to 3 d.p.).

Distribution of the mean of normal random variables

One particular linear combination of normal random variables is of special interest. Let $X_1, \ldots, X_n$ be n independent normal random variables, each with a distribution having mean μ and variance σ^2. The random variable representing the sample mean is given by

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$

From the previous results we see that $\bar{X}$ has a normal distribution with mean and variance given by

$$\mathrm{E}(\bar{X}) = \frac{1}{n}(\mu + \cdots + \mu) = \frac{1}{n} n\mu = \mu$$

$$\mathrm{Var}(\bar{X}) = \left(\frac{1}{n}\right)^2 (\sigma^2 + \cdots + \sigma^2) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Example 28

The random variable X has a normal distribution with mean μ and variance σ^2. The value u is such that $\mathrm{P}(X > u) = 0.25$. The random variable $\bar{X}$ denotes the mean of 16 independent observations of X. Determine the probability that $\bar{X}$ exceeds u.

This looks like an impossible question unless we are told the values of μ and σ^2. We shall see! We know that

$$0.25 = \mathrm{P}(X > u) = \mathrm{P}\left(\frac{X-\mu}{\sigma} > \frac{u-\mu}{\sigma}\right)$$

In other words

$$\Phi\left(\frac{u-\mu}{\sigma}\right) = 0.75$$

However, from the tables we find that

$$\Phi(0.674) = 0.75$$

which implies that

$$\frac{u-\mu}{\sigma} = 0.674$$

We are asked to find $\mathrm{P}(\bar{X} > u)$. Since we know that $\bar{X}$ has a normal distribution with mean μ and variance $\frac{1}{16}\sigma^2$, we can write

$$\mathrm{P}(\bar{X} > u) = \mathrm{P}\left(\frac{\bar{X}-\mu}{\frac{1}{4}\sigma} > \frac{u-\mu}{\frac{1}{4}\sigma}\right) = 1 - \Phi\left(\frac{u-\mu}{\frac{1}{4}\sigma}\right)$$

But

$$\frac{u-\mu}{\frac{1}{4}\sigma} = 4\left(\frac{u-\mu}{\sigma}\right)$$
$$= 4 \times 0.674 = 2.696$$

The required probability is therefore

$$1 - \Phi(2.696) = 0.0035$$

Both μ and σ have disappeared! The result is therefore a general one for samples of size 16 from any normal distribution. It shows the extent to which the mean of a sample is less variable than an individual observation: a value exceeded by 25% of observations is exceeded by the mean of a sample of size 16 on only about 1 occasion in 3000.

Exercises 10e

1 Given that $X \sim \mathrm{N}(4, 9)$, $Y \sim \mathrm{N}(7, 4)$, and that X and Y are independent, find:
(i) $\mathrm{P}(X + Y < 15)$, (ii) $\mathrm{P}(Y - X > 1)$,
(iii) $\mathrm{P}(X > Y)$, (iv) $\mathrm{P}(X > 10 - Y)$.

2 Given that $X \sim \mathrm{N}(-1, 5)$, $Y \sim \mathrm{N}(2, 7)$ and that X and Y are independent, find:
(i) $\mathrm{P}(3X + 2Y > 1)$, (ii) $\mathrm{P}(Y < 2X + 6)$,
(iii) $\mathrm{P}(X + 2Y + 1 < 0)$, (iv) $\mathrm{P}(Y > X)$.

3 Given that $X \sim \mathrm{N}(-1, 4)$, $Y \sim \mathrm{N}(1, 4)$, and that X and Y are independent, find:
(i) $\mathrm{P}(2X > 3Y)$, (ii) $\mathrm{P}(X + Y > 0)$,
(iii) $\mathrm{P}(50X + 100Y > 300)$,
(iv) $\mathrm{P}(100Y - 50X > 200)$.

4 It is given that $X \sim \mathrm{N}(7, 5)$ and that $X_1, X_2, \ldots, X_{10}$ are ten independent observations of X. Let $S = X_1 + X_2 + \cdots + X_{10}$. Find:
(i) $\mathrm{P}(10X_1 > 80)$, (ii) $\mathrm{P}(S > 80)$,
(iii) $\mathrm{P}(10X_{10} < 50)$, (iv) $\mathrm{P}(S < 50)$.

5 The mass of a biscuit is a normal variable with mean 50 g and standard deviation 4 g. A packet contains 10 randomly selected biscuits. The mass of the packing material is a normal random variable with mean 40 g and standard deviation 3 g and is independent of the masses of the biscuits. Find the probability that the total mass of a packet is less than 566 g.

6 A toy company manufactures plastic nuts whose internal diameters are normally distributed with mean 50 mm and standard deviation 4 mm. The company also manufactures plastic bolts whose external diameters are normally distributed with mean 48 mm and standard deviation 3 mm.
By considering the difference between their diameters, determine the probability that a randomly chosen bolt can be inserted into a randomly chosen nut.

7 The amount of jam in a standard jar has a normal distribution with mean 340 g and standard deviation 10 g. The mass of the jar has a normal distribution with mean 150 g and standard deviation 8 g.
Find the probability that:
(i) a randomly chosen jar of jam has total mass exceeding 500 g,
(ii) a randomly chosen pack of 20 jars of jam has a total mass exceeding 10 000 g.

8 Three women and four men enter a lift. Assume that women have masses that are normally distributed with mean 60 kg and standard deviation 10 kg, and that men have masses that are normally distributed with mean 80 kg and standard deviation 15 kg.
Find the probability that the total mass of the seven people in the lift exceeds 550 kg.

9 The masses of tins of baked beans are normally distributed with mean 416 g and standard deviation 1.5 g. A random sample of 10 tins are selected. Determine the probability that their mean mass exceeds 415 g.

10 The masses of biscuits are normally distributed. Ginger nuts have a mean mass of 10.1 g with a standard deviation of 0.1 g, digestives have a mean mass of 10 g with a standard deviation of 0.08 g. Random samples of 8 ginger nuts and 12 digestives are obtained. Determine the probability that the mean mass of the ginger nuts is greater than that of the digestives.

11 The random variable X has a normal distribution with mean μ and variance σ^2. The value l is such that $P(X < l) = 0.4$. The random variable $\bar{X}$ denotes the mean of 25 observations of X. Determine the probability that $\bar{X}$ exceeds l.

12 The continuous random variable X is normally distributed with mean 212.6 and standard deviation 2. Calculate, correct to three decimal places, the probabilities that
(i) a randomly observed value of X will lie between 212 and 213,
(ii) the mean of four randomly observed observations of X will exceed 213. [WJEC]

13 [In this question give three places of decimals in each answer.] The mass of tea in "Supacuppa" teabags has a normal distribution with mean 4.1 g and standard deviation 0.12 g. The mass of tea in "Bumpacuppa" teabags has a normal distribution with mean 5.2 g and standard deviation 0.15 g.
(i) Find the probability that a randomly chosen Supacuppa teabag contains more than 4.0 g of tea.
(ii) Find the probability that, of two randomly chosen Supacuppa teabags, one contains more than 4.0 g of tea and one contains less than 4.0 g of tea.
(iii) Find the probability that five randomly chosen Supacuppa teabags contain a total of more than 20.8 g of tea.
(iv) Find the probability that the total mass of tea in five randomly chosen Supacuppa teabags is more than the total mass of tea in four randomly chosen Bumpacuppa teabags. [UCLES]

14 Monto sherry is sold in bottles of two sizes – standard and large. For each size, the content, in litres, of a randomly chosen bottle is normally distributed with mean and standard deviation as given in the table.

	Mean	Standard deviation
Standard bottle	0.760	0.008
Large bottle	1.010	0.009

(i) Show that the probability that a randomly chosen standard bottle contains less than 0.750 litres is 0.1056, correct to 4 places of decimals.
(ii) Find the probability that a box of 10 randomly chosen standard bottles contains at least three bottles whose content are each less than 0.750 litres. Give three significant figures in your answer.
(iii) Find the probability that there is more sherry in four randomly chosen standard bottles than in three randomly chosen large bottles. [UCLES]

21 The length, in cm, of a rectangular tile is a normal variable with mean 19.8 and standard deviation 0.1. The breadth, in cm, is an independent normal variable with mean 9.8 and standard deviation 0.1.

(i) Find the probability that the sum of the lengths of five randomly chosen tiles exceeds 99.4 cm.

(ii) Find the probability that the breadth of a randomly chosen tile is less than one half of the length.

(iii) S denotes the sum of the lengths of 50 randomly chosen tiles and T denotes the sum of the breadths of 90 randomly chosen tiles. Find the mean and variance of $S - T$. [UCLES]

22 A group of students are weighing lead weights said to have a nominal mass of 10 grams. They discover that the weights produced by manufacturer A have a mean mass of 9.82 grams and standard deviation 0.1 gram.

(*a*) Using the Normal distribution and these values as estimates of the population parameters, calculate the probability that a randomly chosen weight from manufacturer A has a mass of 10 grams or more.

(*b*) Similar weights from manufacturer B are known to have a mass which is Normally distributed with mean 10.05 grams and standard deviation 0.05 gram. Calculate the probability that a randomly chosen weight from manufacturer A has a mass which is greater than the mass of a randomly chosen weight from manufacturer B. You are expected to make clear the parameters of the distribution you use in answering this part of the question. [UODLE(P)]

23 Mass-produced laminated-wood beams are constructed using five layers of wood. A study of the ends of individual layers reveals that the thickness of each of the two outside layers is normally distributed with mean 52 mm and standard deviation 3 mm. The thickness of the ends of each of the three middle layers is also normally distributed but with mean 31 mm and standard deviation 2 mm. In the assembly of the beams both the two outside layers and the three middle layers are selected at random.

Find the mean thickness of the beam ends.

Show that, correct to two decimal places, the standard deviation of the thickness of the beam ends is 5.48 mm.

Determine the probability that the thickness of an end of a beam exceeds 200 mm.

Each end of a beam is fitted into a slot in a steel plate. The widths of the slots are normally distributed with mean 206 mm and standard deviation 2 mm, independently of the thicknesses of the beam ends.

Determine the probability that an end of a beam will fit into its slot.

Hence, assuming that the thicknesses of the two ends of a beam are independent, determine the probability that both ends of a beam will fit into their slots.

The mean width of the slots in the steel plates can be altered without affecting the standard deviation. Determine the mean slot width that will ensure that 95% of the beams fit into *both* their slots. [JMB]

Pierre Simon Laplace (1749–1827) was eulogized by Poisson as being 'the Newton of France'. He had been elected to membership of the French Royal Academy of Sciences by the age of 24, and during his long life held a large number of influential positions including being a professor at the Ecole Militaire during the time that Napoleon was a student. His greatest interest was in celestial mechanics, which involves, amongst other things, the accurate determination of the positions of heavenly bodies. The paper in which Laplace derived the Central Limit Theorem was read to the Academy in 1810, and was a direct consequence of the work by Gauss in the previous year. In France the normal distribution is often referred to as the **Laplacean distribution**.

10.10 The Central Limit Theorem

An informal statement of this extremely important theorem is as follows:

> Suppose $X_1, X_2, \ldots, X_n$ are n independent random variables, each having the *same* distribution.
>
> Then, as n increases, the distributions of $X_1 + \cdots + X_n$ and of $\dfrac{X_1 + \cdots + X_n}{n}$ come increasingly to resemble normal distributions.

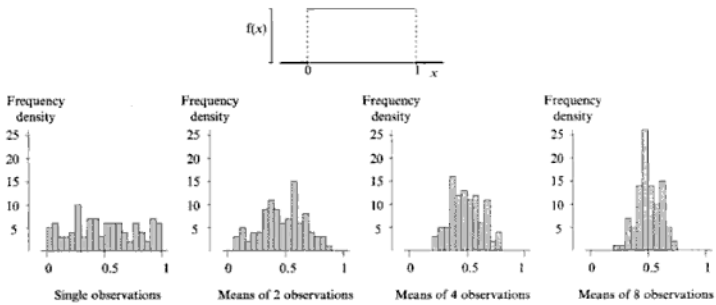
The importance of the **Central Limit Theorem** lies in the fact that:

- The common distribution of the X-variables is not stated – it can be almost *any* distribution.
- In most cases the resemblance to a normal distribution holds for remarkably small values of n.
- Totals and means are quantities of interest!

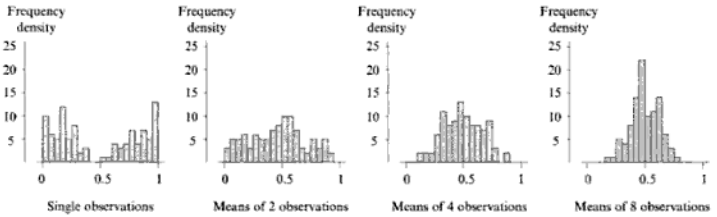
As an example, consider the following data which constitute part of a random sample of observations on a random variable having a continuous uniform distribution in the interval (0, 1).

Original observations	0.020	0.706	0.536	0.580
Means of pairs	0.363		0.558	
Means of fours		0.4605		
Original observations	0.290	0.302	0.776	0.014
Means of pairs	0.296		0.395	
Means of fours		0.3455		

The successive diagrams below show (i) the original distribution, (ii) a histogram of the first 50 randomly chosen single observations from that distribution, (iii) a histogram of the first 50 means of pairs of observations from the distribution, (iv) the same for groups of four observations, and finally, (v) the same for groups of eight observations. As the group size increases so the means become increasingly clustered in a symmetrical fashion about 0.5 (the mean of the original population).

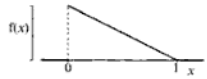


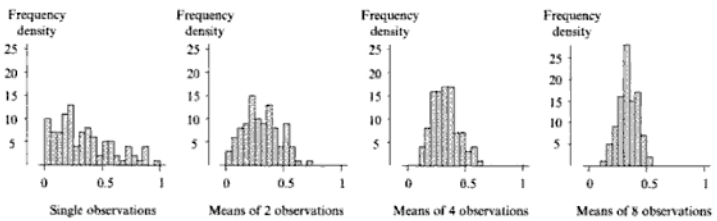
As a second example, consider successive averages of observations from a V-shaped distribution:



Here the original distribution has a trough in the middle, whereas the normal distribution has a peak, but once we start looking at averages of even as few as $n = 2$ observations, a peak starts to appear.

As a final example we look at a triangular distribution, which is quite heavily skewed. In this case the histograms of the means of two observations and four observations still appear skewed, but this skewness has almost vanished by the time we are working with means of eight observations.





Note

- It is clear that the practical consequences of the Central Limit Theorem were understood well before the time of Laplace. A 16th-century German treatise on surveying instructs the surveyor to establish the length of a rood (a standard unit of length) in the following manner:

  'Stand at the door of a church on a Sunday and bid 16 men to stop, tall ones and small ones, as they happen to pass out when the service is finished; then make them put their left feet one behind the other, and the length thus obtained shall be a right and lawful rood to measure and survey the land with, and the 16th part of it shall be a right and lawful foot.'

The distribution of the sample mean, $\bar{X}$

Denote the ith observation in a sample by x_i. Different samples would give different values for x_i (see Section 6.5, p. 174). Thus x_i is an observation on a random variable that we will denote by X_i. Suppose that X_1 has expectation μ and variance σ^2. The same will be true for each of $X_1, X_2, \ldots, X_n$, which are therefore identically distributed random variables, each with expectation μ and variance σ^2. We define $\bar{X}$ by:

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

and we showed in Section 6.5 that $\bar{X}$ has expectation μ and variance $\frac{\sigma^2}{n}$. By the Central Limit Theorem, $\bar{X}$ has an approximate normal distribution for large n, and so:

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

Standardising, we get the equivalent result that:

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathrm{N}(0, 1)$$

Notes

- If the distribution of the individual X variables is normal, then, since $\bar{X}$ is then a linear combination of normally distributed random variables, the result is true even for small values of n.
- The equivalent result for a sum is that:
  $\Sigma X_i \sim \mathrm{N}(n\mu, n\sigma^2)$

Practical

This is a slightly tiring practical! Roll a die and record the value obtained. Roll the die again ... and again ... – about 200 rolls should suffice! Record each outcome and total groups of two and four as indicated below:

Singles	*2*	*4*	*4*	*2*	*2*	*1*	*6*	*2*
Sums of 2		*6*		*6*		*3*		*8*
Sums of 4			*12*				*11*	
Singles	*1*	*4*	*4*	*5*	*1*	*5*	*4*	*4*
Sums of 2		*5*		*9*		*6*		*8*
Sums of 4			*14*				*14*	

Draw up a frequency distribution for the original values and also for the two sets of sums. Illustrate the three distributions using bar charts. In each case, state the smallest and largest values that could possibly have occurred. Compare these extreme values with the ranges of values actually observed and comment on the results.

Computer project

The random numbers generated by a computer (or a calculator) may be thought of as independent observations from a uniform distribution on the interval (0, 1). Write a program to examine the distribution of the means of k observations. A convenient way of keeping count of the values generated is as follows.

1 *Let XBAR be an array of size 100, and set all members of the array to 0. (The choice of 100 is arbitrary, but the array needs to be quite large in order to cope with the case where k is large.)*

Now, for each sample, repeat 2 to 5:

2 *Calculate the sample mean, which will have a value between 0 and 1.*

3 *Multiply the mean by 100 (the array size) to give a value between 0 and 100.*

4 *Calculate the integer part of this quantity. Call this M.*

5 *Add 1 to the Mth item in the array XBAR.*

This was the method used to produce the first of the three earlier examples. The result will be a set of counts for each of 100 classes, of width 0.01. As k increases it will be found that increasing numbers of these classes will have zero frequencies.

Computer project

To simulate from other than uniform distributions requires a little more work! Consider, for example, the third example (the triangular density function) for which the pdf was given by:

$$f(x) = \begin{cases} 2(1-x) & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

This corresponds to the cdf:

$$F(x) = \begin{cases} 0 & x \leqslant 0 \\ x(2-x) & 0 \leqslant x \leqslant 1 \\ 1 & x \geqslant 1 \end{cases}$$

Suppose that u is a number in the range (0,1). If we solve the equation:

$$u = x(2-x)$$

which can be written as:

$$x^2 - 2x + u = 0$$

choosing the root in the interval (0,1), we get:

$$x = 1 - \sqrt{(1-u)}$$

We now know that, with this choice of value for x,

$$\mathrm{P}(X \leqslant x) = \mathrm{P}(U \leqslant u)$$

where U is a random variable having a uniform distribution with range (0,1).

Successive uniform random numbers $u_1, u_2, \ldots$ can therefore be used to generate successive 'random observations' from the triangular distribution, if the latter are chosen to have the values $1 - \sqrt{(1-u_1)}$, $1 - \sqrt{(1-u_2)}$, and so on.

Adjust the program that you devised for the previous project and verify that this approach works!

Example 29

The continuous random variable X has mean 5 and variance 25. A random sample of 100 observations are taken on X.
Determine the probability that the sample mean exceeds 5.4.

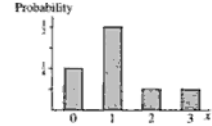
The random variable $\bar{X}$, corresponding to the sample mean, has expectation 5 and variance $\frac{25}{100} = \frac{1}{4}$. By the Central Limit Theorem it has an approximate normal distribution.

The z-value of interest is therefore $\dfrac{5.4-5}{\sqrt{\frac{1}{4}}} = 0.8$. Since $\Phi(0.8) = 0.788$ the required probability is 0.212.

Example 30

The discrete random variable X has probability distribution given by $P(X = 0) = \frac{1}{4}$, $P(X = 1) = \frac{1}{2}$, $P(X = 2) = P(X = 3) = \frac{1}{8}$.
Determine an approximation to the probability that a random sample of 500 observations on X will have a total less than 520, giving the answer to the nearest percentage point.

x	0	1	2	3
P_x	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{8}$



We begin by determining the expectation and variance of X:

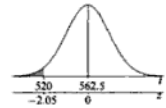
$$E(X) = \left(0 \times \frac{1}{4}\right) + \left(1 \times \frac{1}{2}\right) + \left(2 \times \frac{1}{8}\right) + \left(3 \times \frac{1}{8}\right) = \frac{9}{8}$$

and:

$$E(X^2) = \left(0^2 \times \frac{1}{4}\right) + \left(1^2 \times \frac{1}{2}\right) + \left(2^2 \times \frac{1}{8}\right) + \left(3^2 \times \frac{1}{8}\right) = \frac{17}{8}$$

so that $\mathrm{Var}(X) = \frac{17}{8} - \frac{81}{64} = \frac{55}{64}$. Denoting the random variable corresponding to the total of the 500 observations by T, we have $E(T) = (500 \times \frac{9}{8}) = 562.5$ and $\mathrm{Var}(T) = (500 \times \frac{55}{64}) = \frac{6875}{16}$.

By the Central Limit Theorem the distribution of T will be approximately normal. The z-value of interest is $\dfrac{520 - 562.5}{\sqrt{\frac{6875}{16}}} = -2.050$.



Hence the approximate probability is $1 - \Phi(2.050) = 0.0202$, which is approximately 2%.
(An improved approximation would use a continuity correction – see Section 10.11, p. 277.)

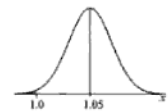
Example 31

Bags of rice are marked as containing 1 kg of rice. In reality, the mean mass of rice per bags is 1.05 kg. The mass of rice varies from bag to bag, and has a standard deviation of 20 g.
Making a suitable assumption, estimate the proportion of bags that contain less than 1 kg of rice.

No mention has been made of a probability distribution for the mass of the contents. However, the total mass of a bag is the aggregate of the masses of thousands of individual grains of rice and we can reasonably assume that the mass of a bag, X kg, has a normal distribution.

The question mentions both kg and g as units of weight: we will work in kg, noting that 20 g = 0.02 kg. Thus $X \sim N(1.05, 0.02^2)$. The z-value corresponding to $x = 1$ is $\dfrac{1.00 - 1.05}{0.02}$, which equals -2.5. Hence:

$$P(X < 1.00) = \Phi(-2.5) = 0.006\,21$$



This value was obtained from the table given in the Appendix. The probability is reassuringly low – only about 1 in 160 bags actually have masses less than the stated value.

Example 32

A builder orders 200 planks of walnut and 50 planks of mahogany. The mean and standard deviation of the masses (in kg) of walnut planks are 15 and 1, respectively. The corresponding figures for the mahogany planks are 20 and 1.1, respectively.
Assuming that the planks delivered to the builder are random samples from the populations of planks, determine the probability that the wood delivered has a total mass of wood that is:
(i) less than 4000 kg, (ii) between 3980 kg and 4000 kg.

From the Central Limit Theorem, the total mass of the walnut planks has an approximate normal distribution. The same is true for the mahogany planks. Also, since linear combinations of independent normal random variables have a normal distribution, the combined mass, X, has an approximate normal distribution.

(i) If we denote the masses of the walnut planks by $W_1, \ldots, W_{200}$ and the masses of the mahogany planks by $M_1, \ldots, M_{50}$, then we can write:

$$X = W_1 + \cdots + W_{200} + M_1 + \cdots + M_{50}$$

Thus:

$$\begin{aligned} \mathrm{E}(X) &= \mathrm{E}(W_1) + \cdots + \mathrm{E}(W_{200}) + \mathrm{E}(M_1) + \cdots + \mathrm{E}(M_{50}) \\ &= (200 \times 15) + (50 \times 20) \\ &= 3000 + 1000 \\ &= 4000 \end{aligned}$$

The probability that the wood delivered has a total mass of less than 4000 kg is therefore exactly $\frac{1}{2}$.

(ii) In a similar way, we see that:

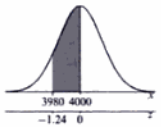
$$\mathrm{Var}(X) = (200 \times 1^2) + (50 \times 1.1^2) = 200 + 60.5 = 260.5$$

The z-value corresponding to $x = 3980$ is therefore $\dfrac{3980 - 4000}{\sqrt{260.5}}$

which equals -1.239 (to 3 d.p.). Thus:

$$\begin{aligned} \mathrm{P}(3980 < X < 4000) &= \Phi(0) - \Phi(-1.239) \\ &= 0.5 - (1 - 0.8924) = 0.3924 \end{aligned}$$

The probability that the total mass of wood delivered is between 3980 kg and 4000 kg is 0.392 (to 3 d.p.).



Exercises 10f

1 The random variable X has mean 15 and variance 25. The random variable $\bar{X}$ is the mean of a random sample of 70 observations on X. State the approximate distribution of $\bar{X}$ and hence find, approximately, $\mathrm{P}(15 < \bar{X} < 16)$.

2 The random variable Y has mean 50 and standard deviation 20. The random variable $\bar{Y}$ is the mean of a random sample of n observations on Y. Find, approximately, $\mathrm{P}(45 < \bar{Y} < 55)$ in the cases (i) $n = 50$, (ii) $n = 100$.

3 The random variable W has mean 20 and variance 72. The random variable S is defined to be the sum of 80 independent observations on W.
Find, approximately, (i) $\mathrm{P}(S > 1700)$, (ii) $\mathrm{P}(1400 < S < 1700)$.

12 Explain briefly how you acquired empirical evidence for the Central Limit theorem.
The weights of the trout at a trout farm are normally distributed with mean 1 kg and standard deviation 0.25 kg.
(*a*) Find, to 4 decimal places, the probability that a trout chosen at random will weigh more than 1.25 kg.
(*b*) Two trout are chosen independently and at random. If their total weight is denoted by X kg, find $E(X)$ and $\mathrm{Var}(X)$.
(*c*) If $\bar{Y}$ kg represents the mean weight of a sample of 10 trout chosen at random at the farm, state the distribution of $\bar{Y}$. Evaluate the mean and the variance of this distribution.
Find, to 3 decimal places, the probability that the mean weight of a sample of 10 trout chosen at random will be less than 0.9 kg. [ULSEB(P)]

13 The contents of bags of oats are normally distributed with mean 3.05 kg, standard deviation 0.08 kg.
(a) What proportion of bags contain less than 3.11 kg?
(b) What proportion of bags contain between 3.00 and 3.15 kg?
(c) What weight is exceeded by the contents of 99.9% of the bags?
(d) If 6 bags are selected at random what is the probability that the mean weight of the contents will be between 3.00 and 3.15 kg?
The weight of the bags, when empty, is normally distributed with mean 0.12 kg, standard deviation 0.02 kg. Full bags are packed into boxes each of which holds 6 bags.
(e) What is the distribution of the weight in a box, i.e. 6 bags together with their contents? Assume that the weight of all bags and contents are independent of each other.
(f) Within what limits will the weight in a box lie with probability 0.9? [AEB 93]

14 The lengths of the petals of a particular variety of flower are approximately normally distributed with mean 32 mm and standard deviation 5 mm.
(i) Explain briefly why the assumption of a normal distribution for the lengths of the petals may be reasonable even though such a length cannot possibly be negative.
(ii) Calculate the proportions of all petals that have lengths
(*a*) greater than 34 mm,
(*b*) between 29 mm and 38 mm.
(iii) If the lengths of the petals were to be measured correct to the nearest millimetre, calculate the proportion of the petals whose measured lengths would be 35 mm or less.
(iv) Find the length, correct to the nearest mm, which is exceeded by 60% of all petals.
(v) Calculate the probability that the mean of the lengths of a random sample of ten petals will be greater than 33 mm.
(vi) Determine how many petals should be sampled to ensure that there is a probability of at least 0.95 that the sample mean length will be within 1 mm of 32 mm, the mean length of all petals. [WJEC]

15 (a) The percentage of a metal extracted from a batch of raw material is normally distributed with mean 53.5% and standard deviation 2.5%.
Determine the probability that for any batch selected at random the percentage of metal extracted from that batch
(i) exceeds 58%,
(ii) lies between 50% and 60%.
Two batches are selected at random. Find the probability that both batches will have between 50% and 60% of metal extracted from them.
Find the minimum number of batches which need to be randomly sampled to ensure that the probability of the sample mean of the batches being within 1.5% of 53.5% is at least 0.9.
(b) In a blending process two liquids, A and B, are poured into the same randomly chosen empty container. The volumes of the containers are normally distributed with mean 9.2 cm^3 and standard deviation 0.55 cm^3.
The volume of liquid A poured in is normally distributed with mean 4 cm^3 and standard deviation 0.32 cm^3. Liquid B is added independently and its volume is normally distributed with mean 5 cm^3 and standard deviation 0.45 cm^3.
Determine the probability that the contents of the container do not overflow. [AEB 92]

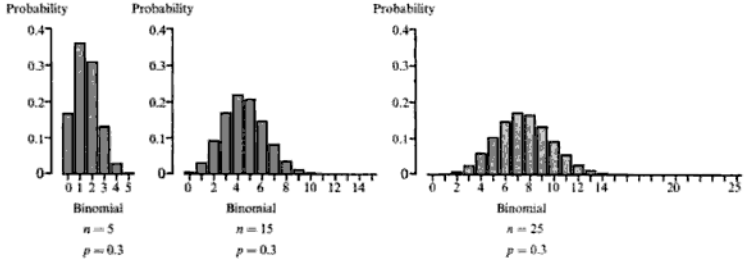
Abraham de Moivre (1667–1754) was a French Protestant who emigrated to London in 1688. By the age of 30 he had been elected a Fellow of the Royal Society as a consequence of his work in various branches of mathematics. His first book, the *Doctrine of Chances*, deals with various aspects of probability, and in its second edition (1738) he carried out the essential calculations that lead to the approximation of a binomial distribution with $p = \frac{1}{2}$ by a normal distribution with mean $\frac{1}{2}n$. It is said that, during his final illness, he noted that, each day, he needed a quarter of an hour's more sleep than on the preceding day. He was thus able to predict accurately the day of his death – when he needed 24 hours of sleep.

10.11 The normal approximation to a binomial distribution

Suppose $X \sim \mathrm{B}(n, p)$. In Section 7.6 (p. 190) we noticed that we could write:

$$X = Y_1 + Y_2 + \cdots + Y_n$$

where the Y-variables had independent Bernoulli distributions, each with parameter p. By the Central Limit Theorem, the sum of independent identically distributed random variables has an approximate normal distribution. For large n, therefore, the binomial distribution must resemble a normal distribution. This is illustrated in the diagram, which shows three binomial distributions having the same value of p (0.3) but differing values of n.

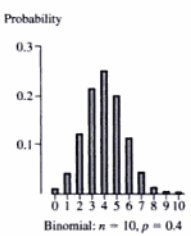


The limiting normal distribution must have the same mean and variance as its binomial counterpart and hence, if we denote the normal counterpart of X by W:

$$X \sim \mathrm{B}(n, p) \rightarrow W \sim \mathrm{N}(np, npq)$$

where $q = 1 - p$.

The normal distribution is continuous, with probabilities associated with all small intervals between $-\infty$ and ∞. The binomial is discrete, with 'chunks' of probability, like slices of a slab of butter, associated with each integer between 0 and n, inclusive. If the bars of binomial probability really were made of butter what would happen if we trod on them? They would spread out sideways – an equal amount on each side. This is precisely how we deal with the move from X to W – we imagine that the probability originally associated with the single point value x becomes identified with the interval $(x - \frac{1}{2}, x + \frac{1}{2})$.



The normal approximation is:

$$P(X = x) \approx P(x - \tfrac{1}{2} < W < x + \tfrac{1}{2}) \tag{10.2}$$

where $W \sim N(np, npq)$. The adjustment by $\frac{1}{2}$ in each direction is referred to as using the **continuity correction**.

To calculate the approximate probability, we must transform to the standard normal distribution by writing:

$$Z = \frac{W - np}{\sqrt{npq}}$$

Hence:

$$P(X = x) \approx \Phi\left(\frac{(x + \frac{1}{2}) - np}{\sqrt{npq}}\right) - \Phi\left(\frac{(x - \frac{1}{2}) - np}{\sqrt{npq}}\right) \tag{10.3}$$

Example 33

Using the normal approximation, determine the probability that exactly 30 heads are obtained when a fair coin is tossed 64 times.

Here $n = 64$, $p = q = 0.5$, so:

$$\begin{aligned} P(X = 30) &\approx \Phi\left(\frac{30.5 - 32}{4}\right) - \Phi\left(\frac{29.5 - 32}{4}\right) \\ &= \Phi(-0.375) - \Phi(-0.625) \\ &= 0.3538 - 0.2660 \\ &= 0.0878 \end{aligned}$$

(This value agrees, to four decimal places, with the exact value calculated using the binomial distribution.)

Inequalities

To see how the approximation works for inequalities we will look at $P(X \leqslant x)$:

$$\begin{aligned} P(X \leqslant x) &= P(X = 0) + \cdots + P(X = x) \\ &\approx P(-\tfrac{1}{2} < W < \tfrac{1}{2}) + \cdots + P(x - \tfrac{1}{2} < W < x + \tfrac{1}{2}) \\ &= P(-\tfrac{1}{2} < W < x + \tfrac{1}{2}) \\ &= \Phi\left(\frac{(x + \frac{1}{2}) - np}{\sqrt{npq}}\right) - \Phi\left(\frac{-\frac{1}{2} - np}{\sqrt{npq}}\right) \end{aligned}$$

For sufficiently large n the second term will be close to zero and we then have:

$$\mathrm{P}(X \leqslant x) \approx \Phi\left(\frac{(x+\frac{1}{2}) - np}{\sqrt{npq}}\right) \tag{10.4}$$

Similarly

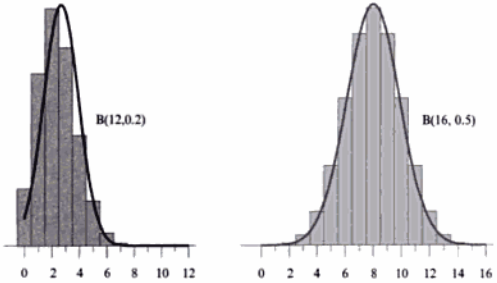
$$\mathrm{P}(X \geqslant x) \approx 1 - \Phi\left(\frac{(x-\frac{1}{2}) - np}{\sqrt{npq}}\right) \tag{10.5}$$

These formulae need adjustment if the inequalities are strict:

$$\mathrm{P}(X < x) \approx \Phi\left(\frac{(x-\frac{1}{2}) - np}{\sqrt{npq}}\right) \tag{10.6}$$

$$\mathrm{P}(X > x) \approx 1 - \Phi\left(\frac{(x+\frac{1}{2}) - np}{\sqrt{npq}}\right) \tag{10.7}$$

It is impossible to give precise rules for when the normal approximation should be used. All one can say is that the approximation is better for p near to $\frac{1}{2}$ and that it improves as n increases.



As the diagram shows, the approximation of the skewed B(12, 0.2) distribution is rather poor, particularly in the lower tail (where the normal approximation anticipates an appreciable proportion of negative outcomes!). By contrast, the approximation of the symmetric B(16, 0.5) distribution is quite impressive.

On the grounds that accurate working is better than inaccurate working (!) binomial probabilities should be calculated exactly if this is feasible in the time available.

Notes

- The approximation is most accurate when $p = \frac{1}{2}$ because this is the case when the binomial distribution is symmetric.
- A continuity correction is needed whenever a discrete distribution is being approximated by a continuous distribution.
- If $X \sim \mathrm{B}(n, p)$ and a normal approximation is valid then about 95% of observations on X will lie in the interval $(np - 2\sqrt{npq},\ np + 2\sqrt{npq})$.

13 (a) The random variable X follows a binomial distribution with parameters n and p.
(i) Prove that $E[X] = np$.
(ii) Show that

$$P(X = r) = \frac{(n - r + 1)p}{(1 - p)r} \times P(X = r - 1).$$

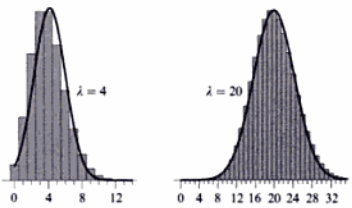
Hence, given that X follows a binomial distribution with $E[X] = 5$ and $P(X = 4) = 1.75P(X = 3)$, find n and p.

(b) A manufacturer of wine glasses sells them in presentation boxes of twelve. Records show that three in a hundred glasses are defective. Find the probability that a randomly chosen box of glasses contains
(i) no defective glasses,
(ii) at least two defective glasses.
Find the probability that a consignment of 10 000 such glasses contains at most 250 defective glasses. [AEB 89]

14 Describe, briefly, the conditions under which the binomial distribution $B(n, p)$ may be approximated by
(*a*) a normal distribution,
(*b*) a Poisson distribution,
giving the parameters of each of the approximate distributions.
Among the blood cells of a certain animal species, the proportion of cells which are of type A is 0.37 and the proportion of cells that are of type B is 0.004. Find, to 3 decimal places, the probability that in a random sample of 8 blood cells at least 2 will be of type A. Find, to 3 decimal places, an approximate value for the probability that
(*c*) in a random sample of 200 blood cells the combined number of type A and type B cells is 81 or more,
(*d*) there will be 4 or more cells of type B in a random sample of 300 blood cells. [ULSEB]

10.12 The normal approximation to a Poisson distribution

We saw in Section 8.3 that the shape of a Poisson distribution depends upon the value of its parameter λ. Although the distribution is always skewed, as λ increases this skewness becomes less visible and the distribution increasingly resembles the normal in appearance.



A Poisson random variable X with parameter λ has expectation and variance both equal to λ. The approximating normal random variable, Y, therefore has a $N(\lambda, \lambda)$ distribution. As in the case of the approximation to a binomial distribution, a **continuity correction** is required:

$$P(X = x) \approx P(x - \tfrac{1}{2} < Y < x + \tfrac{1}{2})$$

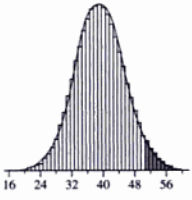
After standardising to a N(0,1) distribution we get:

$$P(X = x) \approx \Phi\left(\frac{(x + \frac{1}{2}) - \lambda}{\sqrt{\lambda}}\right) - \Phi\left(\frac{(x - \frac{1}{2}) - \lambda}{\sqrt{\lambda}}\right)$$

Example 38

The numbers of accidents per day on a given stretch of road are found to have a Poisson distribution with mean 1.4.
Determine the probability that more than 50 accidents occur during a 4-week period.

The number of accidents, X, occurring during a 4-week period has a Poisson distribution with mean $(1.4 \times 28) = 39.2$. We require $P(X > 50)$. Without a computer or programmable calculator it is not feasible to calculate this probability exactly. However, since λ is quite large the normal approximation should be sufficiently accurate.



The relevant value of z is $\dfrac{50.5 - 39.2}{\sqrt{39.2}} = 1.805$. Hence:

$$\begin{aligned} P(X > 50) &= 1 - P(X \leqslant 50) \\ &\approx 1 - \Phi(1.805) \\ &= 1 - 0.9645 \\ &= 0.0355 \end{aligned}$$

On about 3.6% of 4-week periods more than 50 accidents will occur.

Exercises 10h

1 It is given that X has a Poisson distribution with mean 50.
Using a suitable approximation, find $P(X < 40)$.

2 It is given that Y has a Poisson distribution with mean 30.
Using a suitable approximation, find $P(Y > 20)$.

3 In deal planks the number of knots per metre has a Poisson distribution with mean 3.2.
Use a suitable approximation to find the probability that two 5 m lengths contain a total of at least 40 knots.

4 The number of faulty light bulbs returned to a shop in a week has a Poisson distribution with mean 0.7. Using a suitable approximation, find the probability that in a period of 50 weeks not more than 45 faulty bulbs are returned.

5 Large boulders deposited by glaciers are known as erratics. On average a square kilometre of glacial valley contains 25 erratics.
Using a normal approximation to a Poisson distribution, find the probability that a randomly chosen square kilometre contains between 15 and 35 (inclusive) erratics.

6 When pond water is examined under the microscope there are nasty bugs to be seen. The average concentration of these bugs is three per millilitre.
Determine the probability that a random sample of 5 millilitres of pond water contains exactly 14 bugs:
(a) using the Poisson distribution,
(b) using the normal approximation to the Poisson distribution.
Determine also the probability that the number of bugs to be found in a random sample of 200 millilitres of pond water is between 560 and 620 *inclusive*.

7 The number of cars arriving at a petrol station in a period of t minutes may be assumed to have a Poisson distribution with mean $\frac{7}{20}t$.
(i) Find, to three decimal places, the probability that fewer than 6 cars will arrive in a 10 minute period.
(ii) Find, to three decimal places, the probability that exactly 3 cars will arrive in a 5 minute period.
(iii) Find, to two decimal places, an approximate value for the probability that more than 24 cars will arrive in an hour.
[JMB]

8 Analysis of the scores in football matches in a local league suggests that the total number of goals scored in a randomly chosen match may be modelled by the Poisson distribution with parameter 2.7. The numbers of goals scored in different matches are independent of one another.

(i) Find the probability that a match will end with no goals scored.

(ii) Find the probability that 4 or more goals will be scored in a match.

One Saturday afternoon, 11 matches are played in the league.

(iii) State the expected number of matches in which no goals are scored.

(iv) Find the probability that there are goals scored in all 11 matches.

(v) State the distribution for the total number of goals scored in the 11 matches. Using a suitable approximating distribution, or otherwise, find the probability that more than 30 goals are scored in total. **[MEI]**

9 (i) X, Y and Z are random variables having Poisson distributions with means a, b and $a+b$ respectively. X and Y are independent. Show that

$$P(X+Y=2)=P(Z=2).$$

(ii) The number of printing errors on any page of a certain book has a Poisson distribution with mean 0.4.

(1) Find the probability that the total number of errors in the first 10 pages is exactly 3.

(2) Find the probability that the total number of errors in the first 10 pages is more than 3.

(3) N is the smallest integer such that the probability of there being more than 2 errors in the first N pages is greater than 0.88. Verify that $N=13$.

(4) The book has 250 pages. If there are more than 110 errors then the publishers will have the book corrected and reprinted. Use a suitable approximation to find the probability of this happening. **[O&C]**

10 Independently for each year, the number of road accidents per year at a certain blackspot may be regarded as having a Poisson distribution with 2.3 as mean. Calculate, correct to three decimal places, the probabilities that

(i) more than two accidents will occur in a given year,

(ii) the first accident after 31st December 1989 will occur during 1992,

(iii) exactly two of the six years 1990 to 1995 inclusive will be free from accidents.

Given that the sum of independent Poisson variables is also a Poisson variable, use a suitable approximation to calculate, to three decimal places, the probability that there will be at least 30 accidents in a given ten-year period. **[JMB]**

11 The number of flaws in a length of cloth, λ m long, produced on a certain machine, has a Poisson distribution with mean 0.04λ.

(i) Find, to three decimal places, the probability that a 10 m length of cloth has fewer than 2 flaws.

(ii) Find, to three decimal places, the probability that a 100 m length of cloth has more than 4 flaws.

(iii) Find, to two decimal places, an approximate value for the probability that a 1000 m length of cloth has at least 46 flaws.

(iv) Given that the cost of rectifying X flaws in a 1000 m length of cloth is X^2 pence, find the expected value of this cost. **[JMB]**

12 Data files on computers have sizes measured in megabytes. When files are sent from one computer to another down a communications link, the number of errors has a Poisson distribution. On average, there is one error for every 10 megabytes of data.

(i) Find the probability that a 3 megabyte file is transmitted

(A) without error,

(B) with 2 or more errors.

(ii) Show that a file which has a 95% chance of being transmitted without error is a little over half a megabyte in size.

A commercial organisation transmits 1000 megabytes of data per day.

(iii) State how many errors per day they will incur on average.

Using a suitable approximating distribution, show that the number of errors on any randomly chosen day is virtually certain to be between 70 and 130. **[MEI]**

Chapter summary

- **Notation:** $N(\mu, \sigma^2)$denotes a random variable having a normal distribution with mean μ and variance σ^2.
  - The standard normal random variable, Z, has a $N(0, 1)$ distribution.
  - $P(Z < z)$ is denoted by $\Phi(z)$.
  - $\Phi(-z) = 1 - \Phi(z)$.
  - If $X \sim N(\mu, \sigma^2)$ then $P(X < x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$.
- The mean of a set of n independent identically distributed random variables has a distribution that increasingly resembles the normal distribution as n increases (the **Central Limit Theorem**). The same is true for the sum.
- A **linear combination of independent normal random variables** has a normal distribution.
- If $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$, then the sample mean $\bar{X}$, defined by $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$, is $N\left(\mu, \frac{\sigma^2}{n}\right)$
- **Approximation to binomial distribution:**
  If $X \sim B(n, p)$ with $np > 5$ and $n(1-p) > 5$ then:
  $$P(X = x) \approx P(x - \tfrac{1}{2} < W < x + \tfrac{1}{2})$$
  where $W \sim N(np, npq)$ and $q = 1 - p$.
- **Approximation to Poisson distribution:**
  If X has a Poisson distribution with parameter λ, and if $\lambda > 25$ then:
  $$P(X = x) \approx P(x - \tfrac{1}{2} < W < x + \tfrac{1}{2})$$
  where $W \sim N(\lambda, \lambda)$.

Exercises 10i (Miscellaneous)

1 'Disks R Us' purchases a consignment of 50 000 grade A computer disks and 25 000 grade B disks. The probability that a grade A disk contains at least one bad sector is 0.0001, whereas, for a grade B disk this probability is 0.0005. Determine the approximate probability that the consignment contains between 15 and 20 (inclusive) disks having bad sectors.
Grade A disks come in two colours: black and red. Assuming that the colours of the disks in the consignment were independent of one another, and given that the proportion of red disks in the population is 40%, determine the probability that fewer than 20 200 of the grade A disks were red.

2 In a large shipment of peaches, 10% are bad. In most cases the peaches have gone bad because of bruising. However, 10% of the bad peaches can be attributed to the presence of an insect called a 'peach-borer'.
(a) Determine the probability that a random sample of ten peaches contains precisely two bad peaches.
(b) Using a Poisson approximation, determine the approximate probability that a random sample of 100 peaches contains more than eight bad peaches.
(c) Using a Poisson approximation, determine the approximate probability that a random

(continued)

sample of 100 peaches contains no more than one that has gone bad because of a peach-borer.

(d) Using a normal approximation, determine the approximate probability that a random sample of 1000 peaches contains more than eight peaches that have gone bad because of peach-borers.

3 A newspaperman sells papers at random points in time: the number of papers sold in an hour is an observation from a Poisson distribution with mean 50.

(a) Determine the probability that he sells more than 180 in a 3-hour period.

(b) If I have just purchased a paper from him, what is the probability that it will be at least 2 minutes until he sells another?

(c) Given that it is already 5 minutes since his last sale, what is the probability that it will be at least 2 more minutes before his next sale?

4 The random variables X_1 and X_2 are both normally distributed such that $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$. Given that $\mu_1 < \mu_2$ and $\sigma_1^2 < \sigma_2^2$, sketch both distributions on the same diagram.

State the '2σ rule' for a normal random variable. Explain how you used, or could have used, a normal distribution in a project.

The weights of vegetable marrows supplied to retailers by a wholesaler have a normal distribution with mean 1.5 kg and standard deviation 0.6 kg. The wholesaler supplies 3 sizes of marrow:

Size 1, under 0.9 kg,
Size 2, from 0.9 kg to 2.4 kg,
Size 3, over 2.4 kg.

Find, to 3 decimal places, the proportions of marrows in the three sizes.

Find, in kg to one decimal place, the weight exceeded on average by 5 marrows in every 200 supplied.

The prices of the marrows are 16p for Size 1, 40p for Size 2 and 60p for Size 3. Calculate the expected total cost of 100 marrows chosen at random from those supplied. **[ULSEB]**

5 Give **two** reasons why the normal distribution is so important in statistics.

In a study of the dimensions of fibre glass particles produced by manufacturing process *A*, the particle diameters, in micrometers (μm), were established as being normally distributed with a mean of 1.52 μm and a standard deviation of 0.44 μm.

A filter with porosity 0.80 will remove all particles with diameters greater than 0.80 μm. Calculate the proportion of particles removed by such a filter.

Determine the maximum porosity of a filter that will remove at least 99% of the particles.

An analysis of fibre glass particles from process *B* revealed that 28.1% had diameters greater than 2.60 μm and that 10.2% had diameters less than 1.30 μm. Assuming these diameters to be normally distributed, determine their mean and standard deviation.

In fact the two manufacturing processes, *A* and *B*, are in the same building. The number of fibre glass particles produced by *B* is three times the number produced by *A*. Determine the proportion of particles in the building not removed by a filter with a porosity of 0.80. **[JMB]**

6 Describe the main features of a normal distribution. Give an example of an experiment that you would expect to produce a random variable that is normally distributed.

A small firm has three machines producing ball bearings. The diameters of bearings produced by each machine are normally distributed. The firm rejects as undersize all bearings with diameter less than 9.490 mm, and rejects as oversize all bearings with diameter greater than 9.520 mm.

Bearings produced on Machine I have diameters with mean 9.506 mm and standard deviation 0.006 mm. Calculate

(i) the percentage of bearings produced on this machine that are rejected as undersize,

(ii) the percentage of bearings produced on this machine that are considered acceptable.

Machine II produces bearings with a mean diameter of 9.504 mm, of which 2.28% are rejected as oversize. Calculate the standard deviation of the bearings produced on Machine II.

Of the bearings produced on Machine III, 0.5% are rejected as undersize and 4.35% are rejected as oversize. Calculate the mean and the standard deviation of the bearing diameters from Machine III. **[JMB]**

7 The lifetimes of Surecell batteries are normally distributed with mean 200 hours and standard deviation 25 hours.

(*continued*)

Calculate, to 3 decimal places, the probability that a battery chosen at random will
(*a*) last longer than 230 hours,
(*b*) have a lifetime between 190 and 210 hours.
The manufacturers of Surecell batteries wish to offer a guaranteed life T hours on their batteries. They conduct an experiment on batteries they use in the factory. It is found that 9.85% of batteries have a lifetime less than T hours.
(*c*) Calculate, to 2 decimal places, the value of T.
Batteries are removed from the production line and placed in groups of 6.
(*d*) Calculate, to 3 decimal places, the probability that, in a group of 6 batteries, exactly 2 batteries will have a lifetime less than T hours.
Batteries are packed in boxes of 6.
(*e*) Use your answer to (*d*) to calculate, to 3 decimal places, the probability that, in a batch of ten boxes, exactly three boxes will contain 2 batteries with lifetimes less than T. [ULSEB]

8 The weights of pieces of home made fudge are normally distributed with mean 34 g and standard deviation 5 g.
(a) What is the probability that a piece selected at random weighs more than 40 g?
(b) For some purposes it is necessary to grade the pieces as small, medium or large. It is decided to grade all pieces weighing over 40 g as large and to grade the heavier half of the remainder as medium. The rest will be graded as small. What is the upper limit of the small grade?
(c) A bag contains 15 pieces of fudge chosen at random. What is the distribution of the total weight of fudge in a bag?
What is the probability that the total weight is between 490 g and 540 g?
(d) What is the probability that the total weight of fudge in a bag containing 15 pieces exceeds that in another bag containing 16 pieces? [AEB 91]

9 Part of an assembly requires the fitting of a cylinder through a circular hole in a metal plate. It is known that the diameters of the cylinders, D_c, are distributed with mean 24.96 mm and standard deviation 0.04 mm and the diameters of the holes, D_h, are distributed with mean 25.00 mm and standard deviation 0.03 mm.
(a) Find the mean and standard deviation of the difference, $D_h - D_c$, between the diameters of randomly chosen components.
(b) Assuming that both distributions are normal and the components are chosen at random, find the percentage of cases for which the cylinder will not fit the hole.
(c) A plate is chosen at random and then a cylinder is chosen randomly. This is done 4 *more* times. Find the probability that in 3 of the cases the cylinder will not fit its plate.
(d) The percentage of cases for which the cylinder will not fit its plate is to be fixed at 5%. If the standard deviation remains unchanged, determine the increase in mean diameter of hole needed to meet this requirement.
(e) The maximum tolerance – the amount by which D_h exceeds D_c – is set at 0.16 mm. Using the increased mean diameter of the hole, find the percentage of assemblies that do not satisfy this tolerance. [AEB 90]

10 In a survey into working practices, the distance walked each day by a postal worker delivering mail in a residential district was recorded. For each weekday (Monday to Friday) the distribution had mean 12 km and standard deviation 0.9 km. For Saturdays the distribution had mean 10 km and standard deviation 0.5 km. No mail was delivered on Sundays. The distances walked on different days may be assumed to be independent of each other and normally distributed.
(i) Find the probability that, on a randomly chosen Saturday, the postal worker walked between 8.5 km and 11 km.
(ii) Find the probability that, in a randomly chosen week, the postal worker walked further on the Saturday than on the Friday.
(iii) Find the probability that, in a randomly chosen week, the mean daily distance walked by the postal worker for the six-day period was less than 11 km. [UCLES]

11 A machine makes metal rods. A rod is oversize if its diameter exceeds 1.05 cm. It is found from experience that 1% of the rods produced by the machine are oversize. The diameters of the rods are normally distributed with mean 1.00 cm and standard deviation σ cm. Find the value of σ, giving 3 decimal places in your answer.

(*continued*)

Two hundred rods are chosen at random. Using a suitable approximation, find the probability that four or more of the rods are oversize, giving 3 decimal places in your answer.
Another machine makes metal rings. The internal diameters of the rings are normally distributed with mean $(1.00 + 2\sigma)$ and standard deviation 2σ cm, where σ has the value found in the first paragraph. Find the probability that a randomly chosen ring can be threaded on a randomly chosen rod, giving 3 decimal places in your answer. [UCLES]

12 The mass of flour in bags produced by a particular supplier is normally distributed with mean μ grammes and standard deviation 7.5 grammes, where the actual value of μ may be set accurately by the supplier. Any bag containing less than 500 grammes of flour is said to be underweight. The trading standards inspector takes a sample of n bags of flour at random from the bags packed by the supplier. The supplier will be prosecuted if the mean mass of flour in the n bags is less than 500 grammes.

(i) Given that $\mu = 505$ and $n = 10$, find the probability that the supplier will be prosecuted.

(ii) The supplier wishes to ensure that, when $n = 10$, the probability of being prosecuted is not greater than 0.001. Calculate, to one decimal place, the least value at which μ should be set.

(iii) The inspector wishes to ensure that, if the supplier's mean setting produces 80% of bags underweight, the chance that he will escape prosecution is less than one in a thousand. Determine the least value of n that the inspector can use in taking his sample. [JMB]

13 A food packaging company produces tins of baked beans. The empty cans have weights which are normally distributed with mean 40 grams and standard deviation 3.5 grams. The weights of the contents are independent of the weights of the cans and are normally distributed with mean 450 grams and standard deviation 12 grams. Find, to two decimal places, the probability that the total weight of a randomly chosen can of beans is greater than 500 grams.
Show that, in approximately 91% of the cans of beans, the weight of the contents is more than ten times the weight of the empty can. It is decided to change the procedure for packing beans into cans. The weights of the empty cans have the same distribution as before. The cans are filled with beans, so that the total weight is independent of the weight of the can and is a normal variable with mean 490 grams and standard deviation 12.5 grams. Calculate the mean and (to two decimal places) the standard deviation of the weights of the contents of the cans, and explain briefly the significance to a consumer of the change on the packing system. [JMB]

14 An octahedral die has eight faces numbered from 1 to 8. The random variable X is the score obtained when the die is thrown. The bias of the die is such that

$$P(X = r) = c \text{ for } r = 1, 2, 3, 4, 5$$
$$P(X = r) = d \text{ for } r = 6, 7, 8$$
$$P(X < 6) = P(X \geqslant 6).$$

(i) Find the values of c and d, show that $E(X) = 5$ and find the variance of X.

(ii) The die is thrown twice. Calculate the probability that the sum of the two scores is 10.

(iii) The random variable Y is the sum of the scores when this die is thrown 48 times. Find the mean and variance of Y. Assuming that Y has a normal distribution with this mean and variance, find the probability that Y lies between 220 and 260 inclusive. [O&C]

15 A class of 35 third-form pupils conduct a physics experiment in which each measures the time for one complete swing of a pendulum. The experiment is repeated until each pupil has six measurements. The mean time for a complete swing was 1.015 seconds and the standard deviation of the times was 0.045 second. Using the Normal distribution and these values as estimates of the population parameters calculate:

(*a*) the probability that a recorded time is less than 1.1 seconds;

(*b*) the number of recordings of a time less than 1.0 seconds;

(*c*) the number of recorded times that are more than two standard deviations away from the mean time. [UODLE(P)]

(v) It is now required that the probability of a plate being less than 31 cm thick must be 0.95. If the mean thickness of the plates cannot be changed, what value of the standard deviation is required? [O&C]

26 A component has a length which is normally distributed with a mean of 15 cm and a standard deviation of 0.05 cm. An acceptable component is one whose length is between 14.92 cm and 15.08 cm inclusive. The cost of production is 50p per component. An acceptable component can be sold for £1. Undersized components can be sold for scrap at 10p each, and oversized components can be corrected at an additional cost of 20p each and then sold as acceptable. Find the expected profit per 1000 components.

Of these components with acceptable length, the company estimates that 6 in every 1000 are defective in some other way.

(*a*) If X represents the number of defective items in a sample, state the distribution associated with X.

A customer is considering buying some of these components, but will only place an order if there is less than 5% risk that a sample of 150 components contains more than 3 defectives.

(*b*) Use the Poisson approximation to decide whether or not this customer is likely to place an order.

(*c*) State why a Poisson approximation is appropriate in this situation. [ULSEB]

Note

- The distinction between 'large' and 'small' sample sizes is arbitrary, but, typically, 'large' is taken in this context to mean 30 or more observations.

Normal distribution with known variance

A sample of n observations is taken from a $N(\mu, \sigma^2)$ distribution. We denote the random variable corresponding to the sample mean by $\bar{X}$. Since $\bar{X}$ is a linear combination of independent normal random variables, it too has a normal distribution. From Section 6.5 (p. 174) we know that $\bar{X}$ has expectation μ and variance $\frac{\sigma^2}{n}$. Hence:

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

Supposing, for the moment, that μ was known, we could work with the random variable Z, given by:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where the quantity $\frac{\sigma}{\sqrt{n}}$ is often called the **standard error** of the mean.

Since the distribution of Z is known to be N(0,1) we find, by looking at the table of percentage points for a standard normal distribution, that:

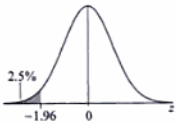
$$P(Z > 1.96) = 0.025$$

from which it follows that:

$$P(Z < -1.96) = 0.025$$

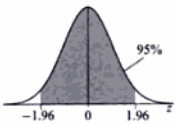
and hence that:

$$P(|Z| < 1.96) = 0.95$$



Substituting for Z, this implies that:

$$P\left(\frac{|\bar{X} - \mu|}{\frac{\sigma}{\sqrt{n}}} < 1.96\right) = 0.95$$



Multiplying the inequality through by $\frac{\sigma}{\sqrt{n}}$, this statement becomes:

$$P\left(|\bar{X} - \mu| < 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

In words, this states that the probability that the distance between μ and $\bar{X}$ is less than $1.96\frac{\sigma}{\sqrt{n}}$ is 0.95. We can conveniently rewrite this as:

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

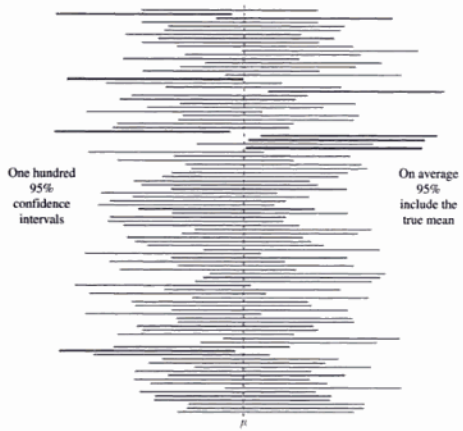
Note that, despite its present appearance, this is still a probability statement concerning the random variable $\bar{X}$. It is *not* a probability statement about μ which is a constant (albeit an unknown constant).

Suppose we now collect our n observations on X and compute the sample mean, $\bar{x}$. The interval:

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) \tag{11.1}$$

is called a **95% symmetric confidence interval** for μ. Often the adjective symmetric is omitted and we just write **95% confidence interval**. The two limiting values that define the interval are known as the **95% confidence limits**.

As the diagram shows, different samples will lead to different values of $\bar{x}$ and hence to different 95% confidence intervals: on average, 95% will include the true population value.



If we wish to be more confident that our interval includes the true value of μ, all we need do is to replace 1.96 by a larger value. This will make the intervals wider! If we wish to have a smaller interval, then we must either take a larger sample or be less confident that the interval includes μ.

The most common percentage points used in the construction of symmetric confidence intervals based on the normal distribution are given in the table below.

Degree of confidence	90%	95%	98%	99%
Percentage point	1.645	1.960	2.326	2.576

Example 1

A machine cuts metal tubing into pieces. It is known that the lengths of the pieces have a normal distribution with standard deviation 4 mm. After the machine has undergone a routine overhaul, a random sample of 25 pieces is found to have a mean length of 146 cm. Assuming the overhaul has not affected the variance of the tube lengths, determine a 99% symmetric confidence interval for the population mean length.

Working in centimetres, the confidence interval is:

$$\left(146 - 2.576 \times \frac{0.4}{\sqrt{25}},\ 146 + 2.576 \times \frac{0.4}{\sqrt{25}}\right)$$

which simplifies to (145.79, 146.21). This particular interval either does or does not include the true population mean length – we cannot say which is true! What we *can* say is that 99% of the intervals constructed in this way will include the true population mean length.

Unknown population distribution, known population variance, large sample

By the Central Limit Theorem (Section 10.10, p. 270), the distribution of $\bar{X}$ will be approximately normal:

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

This case is therefore equivalent to the last case and nothing further need be added.

Unknown population distribution, unknown population variance, large sample

Once again, from the Central Limit Theorem, we can assume that the distribution of $\bar{X}$ is approximately normal. In place of the unknown population variance, σ^2, we use s^2, the unbiased estimate of the population variance (as an approximation). If the sample size is reasonably large (say 30 or more) then the approximation should not be bad. The 95% confidence interval for μ becomes:

$$\left(\bar{x} - 1.96\frac{s}{\sqrt{n}},\ \bar{x} + 1.96\frac{s}{\sqrt{n}}\right) \tag{11.2}$$

Note

- A more accurate procedure using the t-distribution is discussed in Section 11.6 (p. 308).

Example 2

A random sample of 64 sweets is selected. The sweets are found to have a mean mass of 0.932 g, and the value of s is 0.100 g.
Determine an approximate 99% confidence interval for the population mean mass.

The confidence interval will be approximate, since the population variance is unknown and we will use $s(= 0.100)$ in place of σ. The percentage point for a 99% symmetric confidence interval is 2.576, and so the interval becomes:

$$\left(0.932 - 2.576 \times \frac{0.100}{\sqrt{64}},\ 0.932 + 2.576 \times \frac{0.100}{\sqrt{64}}\right)$$

which simplifies to:

(0.900, 0.964)

giving the 99% confidence limits correct to three decimal places.

Example 3

A random sample of 100 men is measured and they are found to have heights (x cm) summarised by $\Sigma x = 17\,280$ and $\Sigma x^2 = 2\,995\,400$. Determine an unbiased estimate of the population variance. Determine also an approximate 98% symmetric confidence interval for the population mean. Give your answers correct to one decimal place.

The unbiased estimate of the population variance is s^2, given by

$$s^2 = \frac{1}{99}\left\{2\,995\,400 - \frac{17\,280^2}{100}\right\} = 95.11$$

which is 95.1 to one decimal place.

The corresponding approximate 98% symmetric confidence interval is:

$$\left(172.8 - 2.326\sqrt{\frac{95.11}{100}},\ 172.8 + 2.326\sqrt{\frac{95.11}{100}}\right)$$

which simplifies to:

(170.5, 175.1)

Example 4

Stingy Stephen takes a random sample of 20 observations from a population with unknown mean μ and unknown variance σ^2. His sample has a mean of 16.2 and an unbiased estimate of the population variance equal to 27.34. Independently, Gorgeous Gertie takes a random sample of 16 observations from the same population. Her sample has a mean of 18.0 and an unbiased estimate of the population variance equal to 35.40. Combining their results to give a single sample, obtain an approximate 95% confidence interval for the population mean, giving the confidence limits correct to two decimal places.

In order to obtain the combined mean and combined variance we need to find the overall sum of the 36 observations and also the overall sum of squares.

Obtaining the overall sum is easy:

$$\Sigma x = (20 \times 16.2) + (16 \times 18.0) = 324.0 + 288.0 = 612.0$$

The combined mean is therefore $\frac{1}{36} \times 612.0 = 17.0$.

Since:

$$s^2 = \frac{1}{n-1}\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}$$

simple algebraic manipulation gives:

$$\Sigma x^2 = (n-1)s^2 + \frac{(\Sigma x)^2}{n}$$

The combined sum of squares for our data is therefore:

$$\Sigma x^2 = \left\{(19 \times 27.34) + \frac{324^2}{20}\right\} + \left\{(15 \times 35.40) + \frac{288^2}{16}\right\}$$
$$= 519.46 + 5248.8 + 531.00 + 5184.0$$
$$= 11\,483.26$$

The unbiased estimate of the population variance for the combined sample of 36 observations is:

$$\frac{1}{35}\left(11\,483.26 - \frac{612^2}{36}\right) = \frac{1079.26}{35} = 30.836$$

An approximate 95% confidence interval for the population mean is therefore given by:

$$\left(17.0 - 1.96\sqrt{\frac{30.836}{36}},\ 17.0 + 1.96\sqrt{\frac{30.836}{36}}\right)$$

which simplifies to:

(15.19, 18.81)

Poisson distribution, large mean

When the mean of a Poisson distribution is large, the normal distribution provides a reasonable approximation (see Section 10.12, p. 283). Since the variance of a Poisson distribution is equal to its mean, there is no need to estimate its value from the data. We simply use the value of the sample mean as its estimate. In this case, therefore, the 95% confidence interval for the population mean is given (approximately) by:

$$\left(\bar{x} - 1.96\sqrt{\frac{\bar{x}}{n}},\ \bar{x} + 1.96\sqrt{\frac{\bar{x}}{n}}\right) \qquad (11.3)$$

Example 5

An environmentalist takes a random sample of water from a river. She discovers that her 100 ml sample contains 64 organisms of a particular (undesirable!) type. Give a 99% confidence interval for the mean number of these organisms in a litre of this river water.

We must first obtain a confidence interval for a water sample of the size obtained. We can then scale this to the required size. The 99% confidence interval for 100 ml is:

$$(64 - 2.576\sqrt{64},\ 64 + 2.576\sqrt{64})$$

since in this case $n = 1$. This interval simplifies to (43.4, 84.6).
The required confidence interval for a litre of the river water is therefore (434, 846).

Exercises 11a

1 The random variable X has a normal distribution with mean μ and variance 9. A random sample of 10 observations of X has mean 8.2.
Find:
(i) a 95% symmetric confidence interval for μ,
(ii) a 99% symmetric confidence interval for μ.

2 The random variable Y has a normal distribution with mean μ and unknown variance. A random sample of 200 observations of Y gives $\sum y_i = 541.2$, $\sum y_i^2 = 1831.42$.
Find:
(i) a 90% symmetric confidence interval for μ,
(ii) a 98% symmetric confidence interval for μ.

3 The random variable W has a distribution with mean μ and unknown variance. A random sample of 150 observations of W gives $\sum w_i = 1601$, $\sum w_i^2 = 18048$.
Giving your answers to two decimal places, find:
(i) a 90% symmetric confidence interval for μ,
(ii) a 95% symmetric confidence interval for μ.

4 The number of telephone calls arriving at a school was monitored on 10 randomly chosen days. The total number of calls was 1053. Assuming a Poisson distribution, find a 95% symmetric confidence interval for the mean number of calls per day.

5 A field of area 7000 m^2 is sown with grass seed. Fifteen non-overlapping squares, each of side 0.1 m are chosen at random and the number of seeds falling on each square is counted. The results are summarised by $\sum x = 2874$.
Assuming a Poisson distribution, find a 90% symmetric confidence interval for:
(i) the mean number of seeds per square metre,
(ii) the number of seeds on the whole field.

6 The weights of 4-month-old pigs are known to be normally distributed with standard deviation 4 kg. A new diet is suggested and a sample of 25 pigs given this new diet have an average weight of 30.42 kg.
Determine a 99% confidence interval for the mean weight of 4-month-old pigs that are fed this diet.

7 The result X of a stress test is known to be a normally distributed random variable with mean μ and standard deviation 1.3. It is required to have a 95% symmetric confidence interval for μ with total width less than 2. Find the least number of tests that should be carried out to achieve this. [ULSEB(P)]

8 The frequency table below summarises the lengths of time in minutes that it took to service an aeroplane between flights on 24 occasions chosen at random.

Time (centre of interval)	55	60	65	70	75
Frequency	2	5	8	6	3

(i) Find the mean and standard deviation of these data.
(ii) Assuming that this sample comes from a normally distributed population with the same standard deviation as you have found in (i), find symmetric 98% confidence limits for the population mean. [O&C]

9 Packets of soap powder are filled by a machine. The weights of powder (to the nearest gram) in 32 packets chosen at random are summarised below.

Weight	999	1000	1001	1002	1003	1004
Packets	1	7	12	8	3	1

Find
(i) the amount by which the mean exceeds 1000 g
(ii) the standard deviation
(iii) the standard error of the mean.
Assuming that this sample comes from a normally distributed population, find, correct to the nearest 0.1 g, 99.8% symmetrical confidence limits for the population mean. [O&C]

10 A plant produces steel sheets whose weights are known to be normally distributed with a standard deviation of 2.4 kg. A random sample of 36 sheets had a mean weight of 31.4 kg. Find 99% confidence limits for the population mean. [ULSEB]

11 A random sample of 80 electrical elements produced by a manufacturer have resistances $x_1, x_2, \ldots, x_{80}$ ohms, where $\sum x_i = 790$, and $\sum x_i^2 = 7821$.

(continued)

(i) Calculate unbiased estimates of the mean and the variance of the resistances of the elements produced by the manufacturer.
(ii) Use a normal distribution to calculate approximate 98% confidence limits for the mean resistance of the elements produced by the manufacturer. [WJEC]

12 Every week a boy buys a packet of his favourite sweets. Each packet carries the statement: "Average contents 150 sweets". Suspecting that this is not the case, the boy decides to count the number of sweets, x, in each of the 52 packets bought during a given year, and finds that $\sum x = 7540$ and $\sum x^2 = 1\,104\,775$.
Calculate
(i) an unbiased estimate of the mean number, μ, of sweets in a packet,
(ii) an unbiased estimate of the variance of the number of sweets in a packet,
(iii) an approximate symmetrical 95% confidence interval for μ. [JMB(P)]

13 A machine is regulated to dispense liquid into cartons in such a way that the amount of liquid dispensed on each occasion is normally distributed with a standard deviation of 20 ml. Find 99% confidence limits for the mean amount of liquid dispensed if a random sample of 40 cartons had an average content of 266 ml. [ULEAC]

14 Describe the work you did to obtain empirical evidence to demonstrate the Central Limit Theorem. State the parameters of your distributions.

$\bar{X}$ is the mean of a large random sample of size n_1 from a population with mean μ_1 and variance σ_1^2.

$\bar{Y}$ is the mean of a large random sample of size n_2 from a population with mean μ_2 and variance σ_2^2.

State the form of the sampling distribution of $(\bar{Y} - \bar{X})$, giving its mean and variance.

Buildrite and Constructall are two building firms. The amount, X thousand pounds, paid to Buildrite by each of 100 randomly chosen customers is summarised as follows:

$$\sum x = 160, \quad \sum x^2 = 265$$

Find approximate 95% symmetrical confidence limits for the amount paid per customer to Buildrite.

The amount paid to Constructall by each customer was Y thousand pounds. Based on a random sample of 200 customers, unbiased estimates of the mean and variance of Y were 1.8 and 0.3216 respectively. Find, to the nearest pound, approximate 90% confidence limits for the value by which the mean amount paid per customer to Constructall exceeds that paid to Buildrite. [ULSEB]

11.4 Confidence interval for a population proportion

Suppose that a random sample of n observations is taken from a population in which the proportion of successes is p and the proportion of failures is q $(= 1 - p)$. Suppose the number of successes in the sample is denoted by r (an observation on the random variable R). The observed proportion of successes is $\frac{r}{n}$, which is denoted by $\hat{p}$, so that $\hat{p} = \frac{r}{n}$ with the corresponding random variable, $\hat{P}$, being given by $\hat{P} = \frac{R}{n}$.

The random variable R has a binomial distribution with parameters n and p and therefore $\mathrm{E}(R) = np$ and $\mathrm{Var}(R) = npq$. Hence:

$$\mathrm{E}(\hat{P}) = \mathrm{E}\left(\frac{R}{n}\right) = \frac{1}{n}\mathrm{E}(R) = \frac{1}{n}np = p$$

which shows that $\hat{P}$ is an unbiased estimator of p. Its variance is given by:

$$\mathrm{Var}(\hat{P}) = \mathrm{Var}\left(\frac{R}{n}\right) = \left(\frac{1}{n}\right)^2 \mathrm{Var}(R) = \left(\frac{1}{n}\right)^2 npq = \frac{pq}{n}$$

Example 6

An importer has ordered a large consignment of tomatoes. When it arrives he examines a randomly chosen sample of 50 boxes and finds that 12 contain at least one bad tomato. Assuming that these boxes may be regarded as being a random sample from the boxes in the consignment, obtain an approximate 99% confidence interval for the proportion of boxes containing at least one bad tomato, giving your confidence limits correct to three decimal places.

We have $\hat{p} = 0.24$, $\hat{q} = 0.76$. The percentage point is 2.576. The confidence interval is therefore:

$$\left(0.24 - 2.576\sqrt{\frac{0.24 \times 0.76}{50}},\ 0.24 + 2.576\sqrt{\frac{0.24 \times 0.76}{50}}\right)$$

which simplifies to:

$$(0.084,\ 0.396)$$

or from about 8% to 40%.

Example 7

It is known that p, the proportion of voters supporting the Conservative party, is (at the time of writing!) about 40%. A market research organisation intends to interview a random sample of n voters, and wishes to ensure that the probability is about 0.90 that its sample estimate of the proportion of Conservative voters lies within two percentage points of the population percentage.
What size sample (to the nearest hundred) should the organisation take? Assume that all voters interviewed do reveal which party they support!

The requirement implies that the 90% confidence interval for p should take the form:

$$(\hat{p} - 0.02,\ \hat{p} + 0.02)$$

We therefore choose n so that:

$$1.645\sqrt{\frac{\hat{p}\hat{q}}{n}} \approx 0.02$$

Rearranging, and taking $\hat{p}$ to be 0.4, we get:

$$n \approx \frac{1.645^2 \times 0.4 \times 0.6}{0.02^2} = 1623.6$$

so that a sample size of about 1600 people should be satisfactory.

As ν increases so the corresponding t_ν-distribution increasingly resembles the limiting standard normal distribution (which corresponds to $\nu = \infty$). When ν is 30 or more, the differences between the t_ν-distribution and the normal distribution are very slight – which explains why the normal distribution could continue to be used for cases where n was large.

Notes

- The result 'T has a t_{n-1}-distribution' requires that $X_1, \ldots, X_n$ have independent and identical normal distributions.
- The phrase 'degrees of freedom' is used because of a link between the t-distribution and the chi-squared distribution (which is introduced in Chapter 13).

Tables of the t-distribution

Since the use of the t-distribution is largely confined to situations involving pre-specified tail probabilities, the tables concentrate on giving the percentage points (which depend on ν, the number of degrees of freedom) for a limited number of cases. There is some variation in the way in which the tables are set out and you should make sure that you are familiar with the tables available to you.

Here is a brief extract from the table given in the Appendix (p. 441) which gives values of t such that $P(T < t) = p\%$, where T has a t_ν-distribution:

	$p(\%)$								
ν	75	90	95	97.5	99	99.5	99.75	99.9	99.95
1	1.000	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
∞	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Note

- A problem with tables of the t-distribution is finding the correct column. You may find it helpful to begin by locating the corresponding normal percentage point, which will be given in the final row of the table.

Example 8

The random variable T has a t-distribution with 3 degrees of freedom. Determine the values of t for which:

(i) $P(T < t) = 0.999$,

(ii) $P(T < t) = 0.25$,

(iii) $P(|T| > t) = 0.05$,

(iv) $P(|T| < t) = 0.98$.

The unbiased estimate of the population variance, s^2, is given by:

$$s^2 = \frac{1}{15}\left\{15.13 - \frac{13.3^2}{16}\right\} = 0.271\,625$$

so that $s = 0.521\,18$ and $\bar{x} = \frac{13.3}{16} = 0.831\,25$. The 99% symmetric confidence interval is therefore:

$$\left(0.831\,25 - 2.947 \times \frac{0.521\,18}{\sqrt{16}},\ 0.831\,25 + 2.947 \times \frac{0.521\,18}{\sqrt{16}}\right)$$

which simplifies to:

$$(0.447,\ 1.215)$$

A 99% symmetric confidence interval for the population mean is from 0.447 g to 1.215 g.

Note that the intermediate working has been carried out using far more than just the three decimal places required for the answer. Premature rounding of intermediate calculations is liable to adversely affect final accuracy.

Example 11

Ten students independently performed an experiment to estimate the value of π. Their results were:

3.12, 3.16, 2.94, 3.33, 3.00, 3.11, 3.50, 2.81, 3.02, 3.10

(i) Calculate the sample mean and the value of s^2.

(ii) Stating any necessary assumption that you make, calculate a 95% symmetric confidence interval for π based on these data, giving the confidence limits correct to two decimal places.

(iii) Estimate the minimum number of results that would be needed if it is required that the width of the resulting 95% symmetric confidence interval should be at most 0.02.

(i) The data are summarised by $\Sigma x = 31.09$ and $\Sigma x^2 = 97.0011$, giving $\bar{x} = 3.109$ and $s^2 = 0.038\,032$.

(ii) We have to assume that the underlying distribution is normal with mean π. The percentage point of the t_9-distribution is 2.262, leading to the symmetric confidence interval:

$$\left(3.109 - 2.262\sqrt{\frac{0.038\,032}{10}},\ 3.109 + 2.262\sqrt{\frac{0.038\,032}{10}}\right)$$

which simplifies to give a 95% symmetric confidence interval for π as (2.97, 3.25).

(iii) The width of a symmetric confidence interval is:

$$2c\frac{s}{\sqrt{n}}$$

where c is the percentage point. With 10 observations the width was 0.28, which is much greater than the desired 0.02. Far more observations will be required to achieve the desired accuracy. Since n will be large, the value of c will be that for the limiting normal distribution, in other words 1.96.

We do not know what value will be obtained for s^2, so our best guess is the value provided by the present sample, namely 0.038 032. To find the required value for n we must solve the equation:

$$2 \times 1.96 \times \sqrt{\frac{0.038\,032}{n}} = 0.02$$

The solution is:

$$n = \frac{2^2 \times 1.96^2 \times 0.038\,032}{0.02^2} = 1461.04$$

and hence the required number of observations (rounding up to the next integer) is estimated as being 1462.

Project

How many words are there in this book? Choose a number of pages at random and count (or estimate) the number of words on each chosen page. Assume that these numbers may be regarded as arising from a (discretised) normal distribution.
Calculate their variance and hence obtain a 95% symmetric confidence interval for the number of words in the book.

Exercises 11d

1 The random variable X has a normal distribution with mean μ. A random sample of 10 observations of X is taken and gives $\sum x_i = 83.3$, $\sum x_i^2 = 721.41$.
Find:
(i) a 95% confidence interval for μ,
(ii) a 99% confidence interval for μ.

2 The quantity of milk in a bottle may be assumed to have a normal distribution.
A random sample of 16 bottles was taken and the quantity of milk was measured, with the following results, in ml.

1005, 1003, 998, 1001, 1002, 999, 1000, 1001, 1007, 1003, 1010, 1001, 1003, 1002, 1005, 995

Find a 99% confidence interval for the mean quantity of milk in a bottle, giving your answers to 2 decimal places.

3 A random sample of 12 hollyhock plants, grown from the seeds in a particular packet, was taken, and the height of each plant was measured, in m. The results are summarised by $\sum x_i = 28.43$, $\sum x_i^2 = 88.4704$.
Making a suitable assumption about the distribution of heights, which should be stated, find a 90% confidence interval for the mean height of hollyhock plants grown from that packet.

4 A lorry is transporting a large number of red apples. As it passes over a bump in the road 10 apples fall off its back and are collected by a passing boy. The masses (in g) of the fallen apples are summarised by $\sum(x - 100) = 23.7$, $\sum(x - 100)^2 = 1374.86$.
Treating the fallen apples as being a random sample, determine a symmetric 99% confidence interval for the mean mass of a red apple, stating any assumptions that you have made. [UCLES(P)]

5 The total costs (in £) of the telephone calls from an office during six randomly chosen weeks of the year are given below.

113.20, 87.60, 109.40,
131.20, 201.10, 142.90

Regarding these values as being independent observations from a normal distribution, obtain a symmetric 99% confidence interval for the mean weekly cost of telephone calls made from the office. [UCLES(P)]

6 The speed at which a baseball is thrown is measured (in km h^{-1}) at the instant that it leaves the pitcher's hand. The results for 10 randomly chosen throws on a cool day are summarised by $\sum(x_i - 128) = 7.9$, $\sum(x_i - 128)^2 = 338.4$, where x_i is the speed of throw i.
Assuming that these results are observations from a normal distribution, obtain unbiased estimates of the mean and variance of this distribution, and obtain a symmetric 99% confidence interval for the mean. [UCLES(P)]

7 A customer obtained a trial supply of wire from a manufacturer and measured the breaking strength, y N, of each of a random sample of 12 lengths of wire, obtaining the results shown below.

80.2	83.5	76.2	79.2	88.7	90.2
93.4	75.1	87.2	83.4	82.6	81.2

($\sum y = 1000.9$, $\sum y^2 = 83\,826.27$)

Use the sample data to obtain a symmetric 99% confidence interval for the mean breaking strength of lengths of wire from the manufacturer. State any distributional assumptions you have made in obtaining your confidence interval.

Explain carefully the meaning of 99% *confidence* as applied to an interval in this context. [JMB(P)]

8 In a classroom experiment to estimate the mean height, μ cm, of seventeen-year-old boys, the heights, x cm, of 10 such pupils were obtained. The data were summarised by $\sum x = 1727$, $\sum x^2 = 298\,834$.

(i) Find the mean and variance of the data, and use them to find the symmetrical 95% confidence interval for μ. State clearly but briefly the two important assumptions which you need to make.

A large experiment is planned using the heights of 150 seventeen-year-old boys.

(ii) What effect will the use of a larger sample have on the width of the confidence interval for μ? Identify two distinct mathematical reasons for this effect.

(iii) To what extent are the assumptions made in (i) still necessary with the larger sample size? [MEI]

3 The random variable X is distributed normally with mean μ and variance σ^2.
Write down the distribution of the sample mean $\bar{X}$ of a random sample of size n.

Records from a dental practice showed that during 1991 the number of minutes per visit spent in the dentist's chair can be taken to be normally distributed with mean 14.5 minutes and standard deviation 2.9 minutes.

(*a*) Calculate an interval within which 90% of the times spent in the dentist's chair will lie.

In 1992 it was assumed that the standard deviation remained unchanged, and the distribution can be assumed to be normal. A random sample of 16 consultations gave the following times in minutes.

13.2	18.7	14.9	12.1
11.6	17.2	10.6	9.4
14.6	12.9	11.2	13.5
12.9	11.8	14.1	12.5

(*b*) For 1992, calculate a 95% confidence interval for the mean length of visit to the dentist. [ULEAC(P)]

4 A piece of apparatus used by a chemist to determine the weight of impurity in a chemical is known to give readings that are approximately normally distributed with a standard deviation of 3.2 mg per 100 g of chemical.

(*a*) In order to estimate the amount of impurity in a certain batch of the chemical, the chemist takes 12 samples, each of 100 g, from the batch and measures the amount of impurity in each sample. The results, obtained in mg/100 g are as follows:

7.6	3.4	13.7	8.6	5.3	6.4
11.6	8.9	7.8	4.2	7.1	8.4

(i) Find 95% central confidence limits for the mean weight of impurity present in a 100 g unit from the batch.

(ii) The chemist calculated a 95% confidence interval for the mean weight of impurity of 100 g units from the batch. The interval was of the form $-\infty \leqslant \text{mean} \leqslant \alpha$. Find the value of α. Suggest why the chemist might prefer to use the value α rather than the limits in (i).

(iii) Calculate an interval within which approximately 90% of the measured weights of impurity of 100 g units from the batch will lie.

(*b*) Estimate how many samples of 100 g the scientist should take in order to be 95% confident that an estimate of the mean weight of impurity per 100 g is within 1.5 mg of the true value.

(*c*) Over a period of months the chemist found that of 150 samples of the chemicals, 18 yielded a level of impurity which was unacceptable.
Calculate an approximate 95% confidence interval for the proportion of samples having an unacceptable level of impurity. [AEB 91]

5 Sugar is produced and bagged by a large company. In a random sample of 80 bags, 18 were found to be bags on which the printing was not clear. Calculate an approximate 95% confidence interval for the proportion of bags with unclear printing.

Explain what you understand by a 95% confidence interval in this context.

The sugar produced is classified as granulated or castor and the masses of the bags of both types are known to be normally distributed. The mean of the masses of bags of granulated sugar is 1022.51 g and the standard deviation for both types of sugar is 8.21 g.

Calculate an interval within which 90% of the masses of bags of granulated sugar will lie.

A sample of 10 bags of castor sugar had masses, measured to the nearest gram, as follows.

1062	1008	1027	1031	1011
1007	1072	1036	1029	1041

Find a 99% confidence interval for the mean mass of bags of castor sugar.

To produce a bag of castor sugar of mass x g costs, in pence,

$$(32 + 0.023x)$$

and it is sold for 65p.

If the company produces 10,000 bags of castor sugar per day, derive a 99% confidence interval for its daily profit from castor sugar. [AEB 92]

6 (i) Explain briefly, referring to your projects where possible, what you understand by a 90% confidence interval.

A normal population has variance 25. Find the size of the smallest sample which could be taken from the population so that the symmetrical 90% confidence interval for the mean has width less than 3 units.

(ii) Rainfall records in a certain town show that it rains on average 2 days in every 5. Taking Monday as the first day of the week, find, to 3 significant figures, the probability that, in a given week,

(*a*) the first 3 days will be without rain and on the remaining days there will be rain,

(*b*) rain will fall on exactly 4 days in the week,

(*c*) Friday will be the first day on which it rains.

Find, to 3 decimal places, the probability that there will be rain in that town on exactly 160 days in a given year of 365 days. [ULSEB]

7 People attending a particular theatre during a week of performances were asked to complete a questionnaire. One of the questions asked the person to indicate his/her age-group. A random sample of 400 of the completed questionnaires produced the following grouped frequency distribution of the ages of the respondents.

Age	Under 25	25–39	40–49
Number of people	28	75	80

Age	50–59	60 or over	Total
Number of people	150	67	400

(*a*) Estimate the proportion of the people who attended the theatre who were under 30 years old.

(*b*) Estimate the median and the semi-interquartile range of the ages of the people who attended the theatre, giving each answer to the nearest month.

(*c*) State why it is not possible to obtain a reliable estimate of the mean age from the information given in the table.

(*d*) Give a reason why the median would be preferred to the mean as a representative value of the average age of the people who attended the theatre.

(*e*) Calculate approximate 95% confidence limits for the proportion of the people who attended the theatre who were 50 years old or more. [WJEC]

8 A random sample of ten quartz watches of a particular make were tested for accuracy over a period of four weeks.
The times, in seconds, gained by the ten watches were: -3, $+7$, $+2$, $+6$, $+8$, $+2$, -3, $+6$, $+11$, $+8$.

(i) Calculate unbiased estimates of the mean, μ, and the variance, σ^2, of the times gained over a period of four weeks.

(ii) Stating any assumptions you make, find a 95% confidence interval for the mean time gained by such watches over a period of four weeks.

(iii) Another random sample of ten of the watches was taken and the times, in seconds, they gained over a period of four weeks gave unbiased estimates of μ and σ^2 equal to 3.8 and 23.4, respectively.
Use the combined set of twenty observations to determine a 95% confidence interval for the mean time gained by such watches over a period of four weeks. [WJEC(P)]

12 Hypothesis tests

It is a good morning exercise for a research scientist to discard a pet hypothesis every day before breakfast. It keeps him young

On Aggression, Konrad Lorenz

Having nothing better to do, you decide to weigh some tins of extraordinarily cheap RIPOFF baked beans. To your amazement, the twelve tins have an average mass that is 10 g less than the mass stated on the tins. Should you report the manufacturers to the authorities? Still thinking about this, you visit the local casino to play Roulette. There are 37 numbers on the wheel and you keep betting on number 1. Should you be surprised that on 25 successive occasions you lose? Perhaps the wheel is biased? You go home in despair to eat baked beans (and learn about hypothesis tests – also known as **significance tests**).

12.1 The null and alternative hypotheses

The first stage of any hypothesis test is to write down the two hypotheses. Usually the **null hypothesis** specifies *a particular value* for some population parameter whereas the **alternative hypothesis** specifies *a range of values*. Here are some examples:

Parameter	Null hypothesis	Alternative hypothesis
Mean, μ	$\mu = 435$	$\mu \neq 435$
Proportion, p	$p = \frac{1}{37}$	$p \neq \frac{1}{37}$
Mean, μ	$\mu = 435$	$\mu < 435$
Proportion, p	$p = \frac{1}{37}$	$p < \frac{1}{37}$

To save writing out 'null hypothesis' and 'alternative hypothesis' lots of times, we denote the hypotheses by H_0 and H_1, respectively. Thus the first pair of hypotheses in the table above would become:

H_0: $\mu = 435$ $\quad$ H_1: $\mu \neq 435$

The first part of this chapter concentrates on cases where the sample size, n, is large. The situations considered are the following:

Unknown parameter	Sample statistic	Random variable	Condition	Distribution
μ	$\bar{x}$	$\bar{X}$	Normal distribution, σ^2 known	$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ (exact)
μ	$\bar{x}$	$\bar{X}$	Any distribution, σ^2 known, n large	$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$
μ	$\bar{x}$	$\bar{X}$	Any distribution, σ^2 unknown, n large	$\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim N(0,1)$
λ	x	X	Poisson distribution, λ large	$\frac{X-\lambda}{\sqrt{\lambda}} \sim N(0,1)$
p	$\hat{p}$	$\hat{P}$	n large	$\frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$

In each case the null hypothesis specifies a value for the unknown parameter. Using this value we can determine the probabilities of events of interest (such as the sample mean being greater than 450, or the sample proportion being less than 0.01). This enables us to develop rules for deciding whether or not to accept the null hypothesis.

Example 1

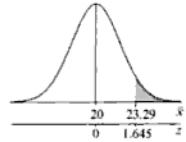
Ten independent observations are to be taken from a $N(\mu, 40)$ distribution. The hypotheses are H_0: $\mu = 20$, H_1: $\mu > 20$. The following procedure has been proposed:
'Reject H_0 (and accept H_1) if $\bar{X} > 23.29$; accept H_0 otherwise'.
Assuming H_0, determine the probabilities of accepting and rejecting H_0 when using this procedure.

Assuming that $\mu = 20$ the distribution of $\bar{X}$ is $N(20, \frac{40}{10})$ so that:

$$\frac{\bar{X} - 20}{2} \sim N(0, 1)$$

Thus:

$$P(\bar{X} > 23.29) = P\left(Z > \frac{23.29 - 20}{2}\right) = P(Z > 1.645)$$



where $Z \sim N(0, 1)$. This tail probability is 5%. Hence, assuming H_0, the probabilities of accepting and rejecting H_0, when using the procedure, are 95% and 5%, respectively.

Note

- In English law the prisoner in the dock is considered to be innocent until 'proven' guilty. In the same way, the null hypothesis is accepted until the evidence suggests that, compared to the alternative hypothesis, it is implausible.

12.2 Critical regions and significance levels

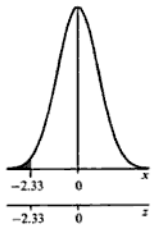
The set of values that leads to the rejection of H_0 in favour of H_1 is called the **rejection region** or the **critical region**. The set of values that leads to the acceptance of H_0 is referred to as – wait for it! – the **acceptance region**.

When the population parameter has the value specified by H_0, the probability that H_0 is nevertheless rejected in favour of H_1 is called the **significance level**. Changing the significance level changes the size of the critical region. In Example 1, the significance level was 5% and the critical region was values of $\bar{x}$ greater than 23.29. In this context 23.29 would be described as the **critical value**.

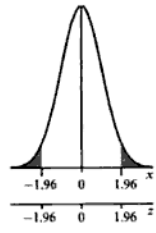
Hypothesis tests in which H_1 involves either a '>' sign (as in Example 1) or a '<' sign are called **one-tailed tests**. The critical regions in these cases involve values in the corresponding tail of the distribution specified by H_0.

Hypothesis tests in which H_1 involves a '$\neq$' sign are called **two-tailed tests**. In these cases the 'critical region' actually consists of two regions – one in each tail of the distribution specified by H_0.

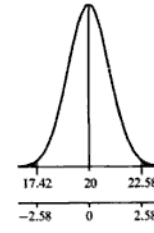
Three examples of critical regions (with probabilities shown shaded) are illustrated below for the case of a single observation on a random variable X having a normal distribution with variance 1. As usual, it is convenient to work with Z, where $Z = \frac{X - \mu}{\sigma}$, and so both x and z scales are shown.



$H_0: \mu = 0$
$H_1: \mu < 0$
1% level
Critical region: $z < -2.33$



$H_0: \mu = 0$
$H_1: \mu \neq 0$
5% level
Critical region: $|z| > 1.96$



$H_0: \mu = 20$
$H_1: \mu \neq 20$
1% level
Critical region: $|z| > 2.58$

In the examples above, the critical values of z are given, in addition to those of x. In the context of hypothesis tests, z is often referred to as the **test statistic**. If the value of z falls in the critical (rejection) region then the result is said to be '**significant**'. If the significance level were $\alpha\%$, then the result would be described as being '**significant at the** $\alpha\%$ **level**'.

Notes

- The most commonly chosen significance levels are 5%, 1% and 0.1%. Note, however, that professional statisticians regard 'significance at the 5% level' as being no more than an indicator that further sampling should take place.
- A result that is significant at the $\alpha\%$ level is also significant at the $\beta\%$ level, for all $\beta > \alpha$.
- Smaller significance levels result in smaller rejection regions.

12.3 The general test procedure

Following the determination of the underlying probability distribution, the full test procedure is as follows:

1 Write down H_0 and H_1.
2 Determine the appropriate test statistic and the distribution of the corresponding random variable (using the parameter value specified by H_0).
3 Determine the significance level.
4 Determine the acceptance and rejection regions.

Now collect the data

5 Calculate the value of the test statistic.
6 Determine the outcome of the test.

It is important to decide upon the critical (rejection) region *before* looking at the actual data so as not to be accidentally biased. We might otherwise have carefully selected our region so as to get a 'significant' result! This would be cheating!

12.4 Test for mean, known variance, normal distribution or large sample

Evidence concerning the value of the population mean is provided by the sample mean. If the population variance is known to be σ^2, and the null hypothesis

specifies a mean μ, then, by the Central Limit Theorem (see Section 10.10, p. 270), for a large sample of size n, the distribution of $\bar{X}$ is approximately

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

If the individual observations are themselves normally distributed, then this result is exact and n need not be large.

Example 2

A random sample of 36 observations is to be taken from a distribution with variance 100. In the past the distribution has had a mean of 83.0, but it is believed that recently the mean may have changed.

(i) Using a 5% significance level, determine an appropriate test of the null hypothesis, H_0, that the mean is 83.0.
When the sample is actually taken it is found to have a mean of 86.2. Does this provide significant evidence against H_0?

(ii) Suppose it is known that, if the population mean has changed, then it can only have increased.
How would this knowledge affect the conclusions?

(i) We will go through the test procedure one stage at a time.

1 *Write down H_0 and H_1*
There is no suggestion in the initial question that any change can only be in one direction. The test is therefore two-tailed:

H_0: $\mu = 83$

H_1: $\mu \neq 83$

2 *Determine the appropriate test statistic and the distribution of the corresponding random variable (using the parameter value specified by H_0).*
The sample size is sufficiently large for us to assume that the distribution of $\bar{X}$ is approximately normal. Since $\sigma^2 = 100$ and $n = 36$, the appropriate test statistic is:

$$z = \frac{\bar{x} - 83.0}{\sqrt{\frac{100}{36}}}$$

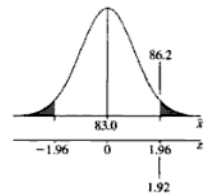
Assuming H_0, z is an observation from a standard normal distribution.

3 *Determine the significance level.*
The question specifies 5%.

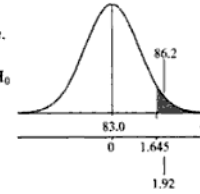
4 *Determine the acceptance and rejection regions.*
The test is two-tailed. Since $P(Z > 1.96) = 0.025$, and $P(Z < -1.96) = 0.025$, an appropriate procedure is to accept H_0 if z lies in the interval $(-1.96, 1.96)$ and otherwise to reject H_0 in favour of H_1.

5 *Calculate the value of the test statistic.*
Since $\bar{x} = 86.2$, $z = 1.92$.

6 *Determine the outcome of the test.*
Since z lies in the interval $(-1.96, 1.96)$, we accept H_0. In other words there is no significant evidence, at the 5% level, that the mean has changed from its previous value of 83.0. Note that this does *not* imply that the mean is unchanged; simply that the mean of our particular sample did not happen to fall in the rejection region.



(ii) If it is known that the population mean cannot have decreased then we will only be persuaded to reject H_0 if $\bar{x}$ is unusually large. The test is now one-tailed with H_1: $\mu > 83.0$. Since $P(Z > 1.645) = 0.05$, an appropriate procedure is now to reject H_0 in favour of H_1 if z is greater than 1.645.



Since 1.92 *is* greater than 1.645, we reject the null hypothesis and accept the alternative hypothesis. In other words, we now have significant evidence, at the 5% level, that the population mean has increased from its previous value.

Exercises 12a

1 Jars of honey are filled by a machine. It has been found that the quantity of honey in a jar has mean 460.3 g, with standard deviation 3.2 g. It is believed that the machine controls have been altered in such a way that, although the standard deviation is unaltered, the mean quantity may have changed. A random sample of 60 jars is taken and the mean quantity of honey per jar is found to be 461.2 g. State suitable null and alternative hypotheses, and carry out a test using a 5% level of significance.

2 Observations of the time taken to test an electrical circuit board show that it has mean 5.82 minutes with standard deviation 0.63 minutes. As a result of the introduction of an incentive scheme, it is believed that the inspectors may be carrying out the test more quickly. It is found that, for a random sample of 150 tests, the mean time taken is 5.68 minutes.
State suitable null and alternative hypotheses.
Assuming that the population variance remains unchanged, carry out a test at the 5% significance level.

3 A lightbulb manufacturer has established that the life of a bulb has mean 95.2 days with standard deviation 10.4 days. Following a change in the manufacturing process which is intended to increase the life of a bulb, a random sample of 96 bulbs has mean life 96.6 days.
State suitable hypotheses.
Assuming that the population standard deviation is unchanged, test whether there is significant evidence, at the 1% level, of an increase in life.

4 The length of string in the balls of string made by a particular manufacturer has mean μ m and variance 27.4 m^2. The manufacturer claims that $\mu = 300$. A random sample of 100 balls of string is taken and the sample mean is found to be 299.2 m. Test whether this provides significant evidence, at the 3% level, that the manufacturer's claim overstates the value of μ. [UCLES(P)]

5 Climbing rope produced by a manufacturer is known to be such that one-metre lengths have breaking strengths that are normally distributed with mean 170.2 kg and standard deviation 10.5 kg.
A new component material is added to the ropes being produced. The manufacturer believes that this will increase the mean breaking strength without changing the standard deviation. A random sample of 50 one-metre lengths of the new rope is found to have a mean breaking strength of 172.4 kg. Perform a significance test at the 5% level to decide whether this result provides sufficient evidence to confirm that the mean breaking strength is increased. State clearly the null and alternative hypotheses which you are using. [ULSEB(P)]

6 The distance driven by a long distance lorry driver in a week is a normally distributed variable having mean 1130 km and standard deviation 106 km.
New driving regulations are introduced and, in the first 20 weeks after their introduction, he drives a total of 21900 km. Assuming that the standard deviation of the weekly distances he drives is unchanged, test, at the 10% level of significance, whether his mean weekly driving distance has been reduced. State clearly your null and alternative hypotheses. [ULSEB(P)]

7 In a large population of chickens, the distribution of the mass of a chicken has mean μ kg and standard deviation σ kg. A random sample of 100 chickens is taken from the population. The mean mass for the sample is $\bar{X}$ kg. State the approximate distribution of $\bar{X}$, giving its mean and standard deviation.
The sample values are summarised by $\Sigma x = 189.1$, where x kg is the mass of a chicken. Given that, in fact, $\sigma = 0.71$, test, at the 1% level of significance, the null hypothesis $\mu = 1.75$ against the alternative hypothesis $\mu > 1.75$, stating whether you are using a one-tail or a two-tail test and stating your conclusion clearly.
[UCLES(P)]

8 A normal distribution has unknown mean μ and known variance σ^2. A random sample of n observations from the distribution has sample mean $\bar{x}$. The null hypothesis $\mu = \mu_0$ is being tested. Find, in terms of μ_0, σ and n, the set of values of $\bar{x}$ for which $\mu = \mu_0$ is rejected in favour of $\mu \neq \mu_0$ at the 1% level of significance. Find also, in terms of $\bar{x}$, σ and n, the set of values of μ_0 for which the hypothesis $\mu = \mu_0$ is rejected in favour of $\mu < \mu_0$ at the 5% level of significance. [UCLES(P)]

9 A fruit grower uses a machine to sort apples into various grades. Grade C apples have weights uniformly distributed in the interval 100 to 110 grams. Find the variance of the weight of a grade C apple.
Ten randomly chosen grade C apples are packed in a bag. Using the central limit theorem, find an approximate value for the probability that the weight of the ten apples in the bag exceeds 1030 grams.
The grower suspects that the machine is not working correctly and that the mean weight, μ grams, of a grade C apple may be less than 105 grams. Devise a test, at the 10% level of significance, based on the weight of the apples in five randomly chosen bags, each containing ten apples, of the null hypothesis $\mu = 105$, with alternative hypothesis $\mu < 105$. [UCLES]

10 Every day Wombles collect litter from Wimbledon Common. They take it home, weigh it (in Womblegrams) and record the daily total. The recorded daily totals for a randomly chosen week during the last year were

173, 149, 181, 151, 178, 185, 194.

Assuming that these figures are independent observations from the population distribution of daily totals, obtain an unbiased estimate of the population mean and show that the unbiased estimate of the population variance is 289.
A Scottish relation, MacWomble, claims that they will find more litter if they have porridge for breakfast. During the first week that they have porridge they collect a daily average of 180.0 Womblegrams of litter. Assuming a normal distribution, with variance 289, test whether this week's daily average is significantly greater, at the 5% level, than that of the week whose daily results are given in the first paragraph. [UCLES]

11 'Kruncho' biscuits have weights that are normally distributed with mean 10 g and standard deviation 1.5 g. If the biscuits are sold in packets of 16, what distribution do the weights of randomly chosen packets follow?
Following maintenance adjustments to the moulding equipment (that are not thought to affect the standard deviation of the biscuit weights) an inspector finds that the average weight of a random sample of 25 packets is 156.9 g. Examine whether there is significant evidence that the adjustments have affected the mean weight of a biscuit.
If the inspector were to weigh a sample of 100 packets, determine over what range of average weights he should conclude that the adjustments have had a significant effect. [SMP]

12 The variables $X_1, X_2, \ldots, X_{12}$ are independent with common probability density

$$f(x) = \begin{cases} 1 & \text{for } 0 \leqslant x \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

Give the mean and variance of X_1 and deduce the mean and variance of the variable $Y = X_1 + X_2 + \cdots + X_{12}$. What is the approximate distribution of Y?
As a check on the random number generator of a microcomputer the following sample of ten values of Y was obtained:

4.85, 5.11, 8.06, 4.20, 6.04,
4.82, 6.28, 5.68, 5.49, 5.58.

Use a test based on the normal distribution to determine whether the mean of these values differs significantly from the expected value.
[SMP]

In real life the choice is not usually so clear cut! Suppose, for example,that we have a situation such as the following:

> The mean breaking strength of a type of climbing rope is 200 kg. Scientists make an adjustment to the method of construction which, they claim, will result in an increase in the breaking strength. A random sample of 12 pieces of the new rope are tested ...

This appears very straightforward. We would use the hypotheses:

H_0: $\mu = 200$ kg

H_1: $\mu > 200$ kg

Suppose now that the 12 pieces of new rope have the following breaking strengths:

187, 196, 193, 187, 194, 193, 197, 194, 191, 195, 194, 199

We evidently do not reject H_0 in favour of H_1 – but would we really want to accept H_0? The new rope appears to have a mean breaking strength of about 193 or 194, and not 200. Some statisticians argue that, because of this type of situation, one-tailed tests should never be used. However, in the context of exam questions they certainly *can* be used.

12.6 Test for mean, large sample, variance unknown

The unbiased estimate of the population variance is given by:

$$s^2 = \frac{1}{n-1}\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}$$

If the sample size is large then this should be a reasonably accurate estimate of σ^2. For such large samples the Central Limit Theorem will also apply and hence, assuming the population mean is μ as specified by H_0, the distribution of $\bar{X}$ is approximately:

$$N\left(\mu, \frac{s^2}{n}\right)$$

The approximation improves as n increases, but should not be used for cases where $n < 30$.

Example 3

In an experiment on people's perception, a class of 100 students were given a piece of paper which was blank except for a line 120 mm long. The students were asked to judge by eye the centre point of the line, and to mark it. The students then measured the distance, x, between the left-hand end of the line and their mark. Working with $y = x - 60$, the results are summarised by $\Sigma y = -143.5$, $\Sigma y^2 = 1204.00$.
Determine whether there is significant evidence, at the 1% level, of any overall bias in the students' perception of the centre of the lines.

It is simplest to work with $Y = X - 60$.

1 *Write down H_0 and H_1.*

The test is two-tailed since there is no suggestion in the question that any bias will necessarily be to the left. With the mean of Y denoted by μ, the hypotheses are therefore:

H_0: $\mu = 0$

H_1: $\mu \neq 0$

2 *Determine the appropriate test statistic and the distribution of the corresponding random variable (using the parameter value specified by H_0).*
Since σ^2 is unknown, but n is large (100), we use:

$$s^2 = \frac{1204.00 - \frac{(-143.5)^2}{100}}{99} = 10.0816$$

The test statistic is therefore:

$$z = \frac{\bar{y} - 0}{\sqrt{\frac{10.0816}{100}}}$$

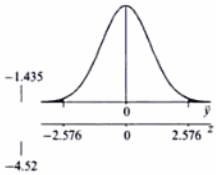
which, assuming H_0, will be an observation from an approximate standard normal distribution.

3 *Determine the significance level.*
The question specifies 1%.

4 *Determine the acceptance and rejection regions.*
The test is two-tailed. Since $P(Z > 2.576) = 0.005$, and $P(Z < -2.576) = 0.005$, an appropriate procedure is to accept H_0 if z lies in the interval $(-2.576, 2.576)$ and otherwise to reject H_0 in favour of H_1.

5 *Calculate the value of the test statistic.*
Since $\bar{y} = -1.435$, $z = -4.52$.

6 *Determine the outcome of the test.*
Since $z < -2.576$, we reject H_0 and accept H_1. There is significant evidence, at the 1% level, that the students' results are biased. Indeed, the result would also have been deemed significant at the 0.001% level!



Practical

Why not try out the experiment in Example 3 in class? In order to have a sufficiently large sample size it may be necessary for everyone to divide two lines. A good idea is to make the line 5 inches long and then to measure to the marked 'centre' point in millimetres. Changing the unit of measurement helps to avoid 'accidental' cheating in which the recorded answer is miraculously correct!

If time allows, there is much scope for experimentation. For example, does the length of the line affect accuracy?
Does the angle of inclination of the line make a difference?

Exercises 12b

1 The mean IQ score is adjusted to be 100 for each age group of the population. A random sample of 3-year-old children is given vitamin supplements for five years. At the end of the period the 180 children have mean IQ score 102.4. The value of s^2 is 219.4.
Test whether there is significant evidence at the 1% level to support the theory that vitamin supplements increase IQ scores.

2 An inspector wishes to determine whether eggs sold as Size 1 have mean weight 70.0 g. She weighs a sample of 200 eggs and her results are summarised by $\Sigma x = 13\,824$, $\Sigma x^2 = 957\,320$, where x is the weight of an egg in grams.
Test whether there is significant evidence, at the 1% level, that the mean weight is not 70.0 g.

3 Rumour has it that the average length of a leading article in the 'Daily Intellectual' is 960 words. As part of a project, a student counts the number of words in each of 55 randomly chosen leading articles from the paper. His results give $\Sigma x = 51\,452$, $\Sigma x^2 = 49\,146\,729$. Test, at the 10% significance level, the truth of the rumour.

4 A teacher notes the time that she takes to drive to school. She finds that, over a long period, the mean time is 24.5 minutes. After a new bypass is opened, she notes the time on 72 randomly chosen journeys to school. Her results are summarised by $\Sigma(x - 20) = 215$, $\Sigma(x - 20)^2 = 3234$, where x minutes is the time for a journey.
Using a 5% significance level, test whether the journey now takes less time.

5 A supermarket manager investigated the lengths of time that customers spent shopping in the store. The time, x minutes, spent by each of a random sample of 150 customers was measured and it was found that $\Sigma x = 2871$, $\Sigma x^2 = 60\,029$. Test, at the 5% level of significance, the hypothesis that the mean time spent shopping by customers is 20 minutes, against the alternative that it is less than this. [UCLES(P)]

6 An electronic device is advertised as being able to retain information stored in it "for 70 to 90 hours" after power has been switched off. In experiments carried out to test this claim, the retention time in hours, X, was measured on 250 occasions, and the data obtained is summarised by $\Sigma(x - 76) = 683$ and $\Sigma(x - 76)^2 = 26\,132$. The population mean and variance of X are denoted by μ and σ^2 respectively.
(i) Show that, correct to one decimal place, an unbiased estimate of σ^2 is 97.5.
(ii) Test the hypothesis that $\mu = 80$ against the alternative hypothesis that $\mu < 80$, using a 5% significance level. [UCLES(P)]

12.7 Test for large Poisson mean

If X has a Poisson distribution with a large mean, λ, then the distribution of X is well approximated by:

$$N(\lambda, \lambda)$$

providing a continuity correction is used (see Section 10.12, p. 283).

There is no need to consider the sample size. If there are n observations from a Poisson distribution with hypothesised mean λ, then their sum may be considered as a single observation from a Poisson distribution with mean $n\lambda$. This is because the sum of independent Poisson random variables is a Poisson random variable (see Section 8.6, p. 204).

Example 4

In a particular river a certain micro-organism occurs at an average rate of 10 per millilitre. A random sample of 0.5 litres of water is taken from a nearby stream and is found to contain 3478 micro-organisms. Does this provide significant evidence, at the 5% level, of a difference in the incidence of the micro-organisms between the stream and the river?

1 *Write down H_0 and H_1.*
If the incidence in the stream were the same as that in the river, then 0.5 litres (i.e. 500 millilitres) of stream water would contain an average of $10 \times 500 = 5000$ micro-organisms. The question refers to a 'difference' and there is no implication that a low count was

anticipated when sampling the stream; we can take the alternative hypothesis to be a two-sided one.

$H_0: \lambda = 5000$

$H_1: \lambda \neq 5000$

2 *Determine the appropriate test statistic and the distribution of the corresponding random variable (using the parameter value specified by H_0).*
We assume that the micro-organisms are randomly distributed in the stream water, so that a Poisson distribution is appropriate. The single count, x, is therefore an observation from a Poisson distribution with mean 5000. The test statistic is therefore:

$$z = \frac{x - 5000}{\sqrt{5000}}$$

When the population mean is indeed 5000, z will be an observation from an approximate standard normal distribution.
The approximation is improved by introducing a continuity correction which would reduce the magnitude of the numerator in the expression for z by 0.5.

3 *Determine the significance level.*
The question prescribes a significance level of 5%.

4 *Determine the acceptance and rejection regions.*
The test is two-tailed. Since $P(Z > 1.645) = 0.025$, and $P(Z < -1.645) = 0.025$, an appropriate procedure is to accept H_0 if z lies in the interval $(-1.645, 1.645)$ and otherwise to reject H_0 in favour of H_1.

5 *Calculate the value of the test statistic.*
Using $x = 3478$ and introducing the continuity correction we calculate z:

$$z = \frac{(3478 + 0.5) - 5000}{\sqrt{5000}} = -21.52$$

6 *Determine the outcome of the test.*
Since -21.5 is considerably(!) less than -1.645, we need have no hesitation in rejecting H_0 at the 5% significance level. There can be no real doubt that there is a difference in the incidence of micro-organisms between the stream and the river.

Exercises 12c

1 Rolls of plastic sheeting from a given manufacturer have been established to have minor faults at an average rate of 0.32 per metre. A 100-metre roll is obtained from a second manufacturer and is found to have 27 minor faults.
Is there significant evidence, at the 10% level, that the second manufacturer's plastic sheeting has fewer faults per metre than the first?

2 A traffic survey shows that, between 9 a.m. and 10 a.m., cars pass a particular census point at an average rate of 4.5 per minute. After the opening of a supermarket in the vicinity, the total number of cars passing the census point, between 9 a.m. and 10 a.m. on 5 days, is found to be 1258.
Test, at the 1% level of significance, whether there is evidence of a change in the rate at which cars pass the census point.

3 A rail company claims that 4.3% of its trains are late. A Passenger Association believes this to be an under-estimate, and carries out a check on a random sample of 500 trains, finding that 30 trains are late.
Test the Association's belief, using a 5% significance level.

4 At a small telephone exchange, the number of calls arriving in a period of t minutes has a Poisson distribution with mean λt, where λ is an unknown constant. Use a 5 per cent significance level to test the null hypothesis H_0: $\lambda = 1$ against the alternative hypothesis H_1: $\lambda > 1$, when 74 calls arrive in 1 hour.
[JMB(P)]

12.8 Test for proportion, large sample

With a sample of size n that contains r 'successes', evidence concerning the population success probability, p, is provided by the sample proportion $\hat{p}$, defined by $\hat{p} = \frac{r}{n}$, with the corresponding random variable being denoted by $\hat{P}$. When n is large, the normal approximation to the binomial distribution is valid. Writing $q = 1 - p$, the distribution of $\hat{P}$ is approximately:

$$N\left(p, \frac{pq}{n}\right)$$

(see Section 11.4, p. 301). The approximation is improved by the use of a continuity correction.

Example 5

A golf professional sells wooden tees. The type that he usually sells are very brittle, and 25% break on the first occasion that they are used. The golfers are not very pleased about this, so the golf professional buys a batch of 'Longlast' tees (which are supposed to last longer!). The professional chooses a random sample of 100 of these tees and tries them out. Only 18 break on the first occasion that they are used.
Does this provide significant evidence, at the 1% level, that the proportion of 'Longlast' tees that break on the first occasion they are used is less than 25%?

In this case a 'success' is a breakage!

1 *Write down H_0 and H_1.*

H_0: $p = 0.25$
H_1: $p < 0.25$

2 *Determine the appropriate test statistic and the distribution of the corresponding random variable (using the parameter value specified by H_0).*

6 In a public opinion poll, 1000 randomly chosen electors were asked whether they would vote for the "Purple Party" at the next election and 357 replied "Yes". The leader of the "Purple Party" believes that the true proportion is 0.4. Test, at the 8% level, whether he is overestimating his support. [UCLES(P)]

7 The owner of a large apple orchard states that 10% of the apples on the trees in his orchard have been attacked by birds. A random sample of 2500 apples is picked and 274 apples are found to have been attacked by birds. Test, at the 8% significance level, whether there is significant evidence that the owner has understated the proportion of the apples on the trees in his orchard that have been attacked. State your hypotheses clearly. [UCLES(P)]

8 A drug company tested a new pain-relieving drug on a random sample of 100 headache suffers. Of these, 75% said that their headache was relieved by the drug. With the currently marketed drug, 65% of users say that their headache is relieved by it. Test, at the 4% level, whether the new drug will have a greater proportion of satisfied users. [UCLES(P)]

9 A questionnaire was sent to a large number of people, asking for their opinions about a proposal to alter an examination syllabus. Of the 180 replies received, 134 were in favour of the proposal. Assuming that the people replying were a random sample from the population, test, at the 5% level, the hypothesis that the population proportion in favour of the proposal is 0.7 against the alternative that it is more than 0.7. [UCLES(P)]

10 A schoolmaster wishes to estimate how many of the 36 pupils in his class smoke at least one cigarette every day. Since they may not answer truthfully if he asks this question directly, each pupil is asked to carry out the following procedure:

Toss a coin, concealing the outcome from all but yourself. If you obtain a head then answer "yes". If you obtain a tail then answer "yes" only if you smoke at least one cigarette every day, otherwise answer "no".

Using this procedure, the pupils may be assumed to answer truthfully, and the number who answer "yes" is 24.

(i) Given that the probability of a head is $\frac{1}{2}$, estimate the proportion of pupils in the class who smoke at least one cigarette every day.

(ii) Using a normal distribution, test, at the 4% significance level, the null hypothesis that there is no pupil in the class who smokes at least one cigarette every day, stating clearly your alternative hypothesis. [UCLES]

12.9 Test for mean, small sample, variance unknown

When the sample size is small, we have to use the t-distribution rather than the normal distribution (see Sections 11.5 and 11.6, pp. 305–8). The test statistic, t, is defined by:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where μ is the population mean specified by the null hypothesis. The distribution of the corresponding random variable, T, is t_{n-1}.

Notes

- The distribution of T is only exactly a t-distribution if the population is normal.
- Strictly, the t-distribution should be used in preference to the normal whenever s is used in place of σ, and not only when n is small.

Example 6

Bottles of wine are supposed to contain 75 cl of wine. An inspector takes a random sample of six bottles of wine and determines the volumes of their contents, correct to the nearest half millilitre. Her results are:

747.0, 751.5, 752.0, 747.5, 748.0, 748.0

Determine whether these results provide significant evidence, at the 5% level, that the population mean is less than 75 cl.

It is simplest to work in millilitres. The target quantity, 75 cl, is $\frac{75}{100}$ of a litre, which is the same as 750 millilitres.

1 *Write down H_0 and H_1.*
The test is one-tailed. The hypotheses are:

H_0: $\mu = 750$

H_1: $\mu < 750$

2 *Determine the appropriate test statistic and the distribution of the corresponding random variable (using the parameter value specified by H_0).*
Since σ^2 is unknown, we must calculate s^2. The numbers are simpler if we work with y, given by $y = x - 750$. This transformation does not alter the variability of the observations, which become:

$-3.0, 1.5, 2.0, -2.5, -2.0, -2.0$

These are summarised by $\Sigma y = -6.0$ and $\Sigma y^2 = 29.50$, so that:

$$s^2 = \frac{1}{5}\left\{29.50 - \frac{(-6.0)^2}{6}\right\} = 4.70$$

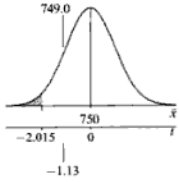
The test statistic is therefore:

$$t = \frac{\bar{x} - 750}{\sqrt{\frac{4.70}{6}}}$$

which, assuming H_0, is an observation from a t_5-distribution.

3 *Determine the significance level.*
The question specifies 5%.

4 *Determine the acceptance and rejection regions.*
The test is one-tailed. The upper 5% point of a t_5-distribution is 2.015, and hence, by symmetry, the lower 5% point is -2.015. An appropriate procedure is therefore to accept H_0 if t is greater than -2.015 and otherwise to reject H_0 in favour of H_1.



5 *Calculate the value of the test statistic.*
Since $\bar{x} = 749.0$, $t = -1.13$.

6 *Determine the outcome of the test.*
Since $t > -2.015$, we accept H_0: there is no significant evidence, at the 5% level, that the population mean is less than 75 cl.

Exercises 12e

1 A manufacturer claims that the mean lifetime of the light bulbs he produces is at least 1200 hours. A random sample of 10 bulbs is taken and the lifetimes are observed. The results are summarised by $\Sigma(x - 1000) = 1890.0$ and $\Sigma(x - 1000)^2 = 362\,050.2$, where x is measured in hours. Assuming the lifetimes of the bulbs to be normally distributed, and using a 5% significance level, test whether there are grounds to dispute this claim. [UCLES(P)]

2 A national company owns a chain of laboratories at which routine chemical tests are carried out. In order to ensure the accuracy of the analyses a sample with a known (but undisclosed) sulphur content of 10.57 grams/litre is sent to each laboratory for testing by its senior analyst. The results from 10 such laboratories are shown below.

Lab	1	2	3	4
Result (x)	10.37	10.49	10.51	10.39

Lab	5	6	7	8
Result (x)	10.56	10.56	10.70	10.46

Lab	9	10
Result (x)	10.44	10.59

These results are summarised by
$\Sigma(x - 10.5) = 0.07$,
$\Sigma(x - 10.5)^2 = 0.0897$.
Assuming that the errors of analysis are normally distributed, test, at the 5% significance level, whether there is any indication of an overall bias in these results. [UCLES(P)]

3 A sample of eight containers is selected at random from a large batch. The containers have powder contents with masses x g,
1998.5, 2000.4, 1999.9, 2005.8,
2011.5, 2007.6, 2001.3, 2002.4,
which are summarised by
$\Sigma(x - 2000) = 27.4$ and
$\Sigma(x - 2000)^2 = 233.52$.
Assuming a normal distribution for the masses of the contents, show that there is significant evidence, at the 5% level, that the mean mass of the contents of the containers in this batch is greater than 2000 g. [UCLES(P)]

4 After a nuclear accident, government scientists measured radiation levels at 20 randomly chosen sites in a small area. The measuring instrument used is calibrated so as to measure the ratio of present radiation to the previous known average radiation in that small area. The measurements are summarised by $\Sigma x_i = 22.8$, $\Sigma x_i^2 = 27.55$. Making suitable assumptions, test, at the 5% level, the hypothesis that there has been no increase in the radiation level. [UCLES(P)]

5 A random sample of size n is taken from a normal distribution with mean μ and variance σ^2. The statistic z is defined by $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$,
where $\bar{x}$ is the sample mean. A statistician who wants to use the statistic z to test a hypothesis about μ does not know the population variance σ^2 and so replaces σ in the statistic z by an estimate of σ.

(i) State what estimate of σ the statistician should use.

(ii) Name the distribution of z and the distribution of the statistic which results from z when σ is replaced by the estimate of σ.

(iii) Sketch these two distributions on the same diagram. [UCLES(P)]

6 A marmalade manufacturer produces thousands of jars of marmalade each week. The mass of marmalade in a jar is an observation from a normal distribution having mean 455 g and standard deviation 0.8 g. Determine the probability that a randomly chosen jar contains less than 454 g.
Following a slight adjustment to the filling machine, a random sample of 10 jars is found to contain the following masses (in g) of marmalade:
454.8, 453.8, 455.0, 454.4, 455.4,
454.4, 454.4, 455.0, 455.0, 453.6.

(i) Assuming that the variance of the distribution is unaltered by the adjustment, test, at the 5% significance level, the hypothesis that there has been no change in the mean of the distribution.

(ii) Assuming that the variance of the distribution may have altered, obtain an unbiased estimate of the new variance and, using this estimate, test, at the 5% significance level, the hypothesis that there has been no change in the mean of the distribution. [UCLES(P)]

Another disadvantage arises in connection with the small-sample cases considered later in Sections 12.13 and 12.14 (pp. 341–5). Since Professors of Statistics disagree over how to report such tail probabilities, we will stick with our original procedure!

12.11 Hypothesis tests and confidence intervals

There is a simple rule that usually works in the case of a two-sided alternative hypothesis:

> If a c% symmetric confidence interval excludes the population value of interest, then the null hypothesis that the population parameter takes this value will be rejected at the $100(1-c)$% level.

For example, if the symmetric 95% confidence interval for a population mean, μ, is (83.0, 85.1), then the null hypothesis that $\mu = 85.2$ will be rejected at the 5% level since the interval excludes 85.2. Indeed *any* hypothesised value for μ that is greater than 85.1, or is less than 83.0 will be rejected at the 5% level. Conversely, the hypothesis that μ takes *any* specific value in the range (83.0, 85.1) will be accepted at the 5% level.

Example 7

A machine cuts wood to form stakes, which are supposed to be 2 metres long. A random sample of 40 stakes is taken, the stakes are accurately measured, and their lengths (x cm) are summarised (using a coding method with reference value 200 cm) by:

$$\Sigma(x-200) = 41.56, \qquad \Sigma(x-200)^2 = 107.4673$$

Determine a 95% confidence interval for the mean stake length.
Test, at the 5% significance level, the null hypothesis that the population mean is 2 metres against the alternative that this is not the case.

The unbiased estimate of the population variance is given by:

$$s^2 = \frac{1}{39}\left(107.4673 - \frac{(41.56)^2}{40}\right) = 1.6484$$

The sample size is sufficiently large that the distribution of the sample mean can be taken to be normal (by the Central Limit Theorem). There will be little loss of accuracy in treating $\dfrac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ as having a N(0,1) distribution. The 95% confidence interval is therefore:

$$200 + \frac{41.56}{40} \pm 1.96\sqrt{\frac{1.6484}{40}} = (200.64, 201.44)$$

Since the interval excludes 200.0, the hypothesis that the mean is 2 metres is rejected, at the 5% significance level, in favour of the alternative that this is not the case.

Notes

- The rule does not work perfectly in the case of a binomial proportion because the variance used in the calculation of the confidence interval $\left(\frac{\hat{p}\hat{q}}{n}\right)$ will usually be slightly different from that used in the context of a hypothesis test $\left(\frac{pq}{n}\right)$.

Exercises 12f

1 Jars of honey are filled by a machine. It has been found that the quantity of honey in a jar has mean 460.3 g, with standard deviation 3.2 g. It is believed that the machine controls have been altered in such a way that, although the standard deviation is unaltered, the mean quantity may have changed. A random sample of 60 jars is taken and the mean quantity of honey per jar is found to be 461.2 g. State suitable null and alternative hypotheses, and carry out a test using a 5% level of significance:
(a) using the *p*-value method,
(b) by finding an appropriate confidence interval.

2 An inspector wishes to determine whether eggs sold as Size 1 have mean weight 70.0 g. She weighs a sample of 200 eggs and her results are summarised by $\Sigma x = 13\,824$, $\Sigma x^2 = 957\,320$, where x is the weight of an egg in grams. Test whether there is significant evidence, at the 1% level, that the mean weight is not 70.0 g:
(a) using the *p*-value method,
(b) by finding an appropriate confidence interval.

3 Rumour has it that the average length of a leading article in the 'Daily Intellectual' is 960 words. As part of a project, a student counts the number of words in each of 55 randomly chosen leading articles from the paper. His results give $\Sigma x = 51\,452$, $\Sigma x^2 = 49\,146\,729$. Test, at the 10% significance level, the truth of the rumour:
(a) using the *p*-value method,
(b) by finding an appropriate confidence interval.

4 Explain what you understand by the term 'Central Limit Theorem', illustrating your answer with reference to any experiment you may have conducted.
In 1988 a meteorologist recorded the length of time (hours) the sun shone at her work station for each of the 31 days during December. She then calculated the mean daily figure for that month. Her data can be summarised as
$\sum x_i = 44.48$, $\sum x_i^2 = 83.5008$,
where $i = 1$ to 31 and x_i represents the daily sunshine in hours.
(*a*) Write down a point estimate for the mean daily hours of sunshine. Calculate an unbiased estimate for the variance of the daily sunshine. Hence find the 'standard error' of the mean.
(*b*) Calculate a 95% confidence interval for the expected hours of sunshine for a day in December. In December 1989, the sun shone for a total of 62.62 hours. Is this sufficient evidence to suggest that there was a change in the average daily sunshine? Justify your response. [UODLE]

5 A politician, speaking to a journalist, claims that school leavers in his constituency have, on average, 6 GCSEs. The journalist checks the claim by interviewing a random sample of 100 school leavers. The data he obtains are summarised below; x denotes the number of GCSEs per person.

$$n = 100 \quad \sum x = 431 \quad \sum x^2 = 2578$$

(i) Obtain the mean and standard deviation of the data.
(ii) Construct a 95% confidence interval for the mean number of GCSEs per person.
(iii) Explain, without further calculation, whether or not the politician's claim is consistent with the journalist's findings.
(iv) By considering again the mean and standard deviation of the data as calculated in (i), explain why the number of GCSEs per person seems unlikely to be Normally distributed. Show in a sketch a possible shape for the distribution.
(v) Explain whether lack of Normality does or does not invalidate the confidence interval found in (ii). [MEI]

6 A credit card company is interested in estimating the proportion of card holders who, at some time, have carried a non-zero balance at the end of a month and so have incurred interest charges. A random sample of 400 credit card holders reveals that 168 have, at some time, incurred interest charges. Calculate an approximate 99% confidence interval for the proportion of all credit card holders who have, at some time, incurred interest charges.
Hence comment on the claim that this proportion is 0.5. [AEB(P) 90]

7 Packets of baking powder have a nominal weight of 200 g. The distribution of weights is normal and the standard deviation is 7 g. Average quantity legislation states that, if the nominal weight is 200 g

(i) the average weight must be at least 200 g,
(ii) not more than 2.5% of packages may weigh less than 191 g,
(iii) not more than 1 in 1000 packages may weigh less than 182 g.

A random sample of 30 packages had the following weights

218,	207,	214,	189,	211,	206,
203,	217,	183,	186,	219,	213,
207,	214,	203,	204,	195,	197,
213,	212,	188,	221,	217,	184,
186,	216,	198,	211,	216,	200

(a) Calculate a 95% confidence interval for the mean weight.
(b) Find the proportion of packets in the sample weighing less than 191 g and use your result to calculate an approximate 95% confidence interval for the proportion of all packets weighing less than 191 g.
(c) Assuming that the mean is at the lower limit of the interval calculated in (a), what proportion of packets would weigh less than 182 g?
(d) Discuss the suitability of the packets from the point of view of the average quantity system. A simple adjustment will change the mean weight of future packages. Changing the standard deviation is possible, but very expensive. Without carrying out any further calculations, discuss any adjustments you might recommend. [AEB 90]

8 A food processor produces large quantities of jars of jam. In each batch, the gross weight of a jar is known to be normally distributed with standard deviation 7.5 g. (The gross weight is the weight of the jar plus the weight of the jam.) The gross weights, in grams, of a random sample from a particular batch were:

514,	485,	501,	486,	502,
496,	509,	491,	497,	501,
506,	486,	498,	490,	484,
494,	501,	506,	490,	487,
507,	496,	505,	498,	499.

(a) Estimate the proportion of this batch with gross weight over 500 g. Calculate an approximate 95% confidence interval for this proportion.
(b) Calculate a 90% confidence interval for the mean gross weight of this batch.

The weight of an empty jar is known to be normally distributed with mean 40 g and standard deviation 4.5 g. It is independent of the weight of the jam.

(c) (i) What is the standard deviation of the weight of the jam in a batch of jars?
(ii) Assuming that the mean gross weight is at the upper limit of the confidence interval calculated in (b), calculate limits within which 99% of the weights of the contents would lie.
(d) The jars are claimed to contain 454 g of jam. Comment on this claim as it relates to this batch of jars. [AEB 94]

12.12 Type I and Type II errors

The statistician's life is not a happy one! When conducting hypothesis tests there are two types of error that may occur, which are summarised in the table below.

		Our decision	
		We accept H_0	We reject H_0
Reality	H_0 correct	Correct!	TYPE I ERROR
	H_0 incorrect	TYPE II ERROR	Correct!

As the table shows, a **Type I error** is made if a correct null hypothesis is rejected. The probability of this error is under our control since:

$$\text{P(Type I error)} = \text{significance level} \tag{12.1}$$

Calculation of the probability of a **Type II error** is not so straightforward, since the probability depends on the extent to which H_0 is false. If H_0 is only slightly incorrect then we may not notice that it is wrong and the probability of a Type II error will be large. On the other hand, if H_0 is nothing like correct then the probability of a Type II error will be low.

In a more positive frame of mind, rather than asking about the probability of making an error, we can ask how good a test is at detecting a false null hypothesis. This is known as the **power** of a test. Formally:

$$\text{power} = 1 - \text{P(Type II error)} \tag{12.2}$$

The general procedure

This closely follows that for the construction of hypothesis tests:

1 Write down the two hypotheses, for example, H_0: $\mu = \mu_0$ and H_1: $\mu > \mu_0$.
2 Determine the appropriate test statistic and the distribution of the corresponding random variable (using the parameter value specified by H_0).
3 Determine the significance level. This is P(Type I error).
4 Determine the acceptance and rejection regions.

Consider now the case when the value of the parameter is not that specified by H_0

5 Determine the distribution of the random variable corresponding to the test statistic given $\mu = \mu_1$, say.
6 Calculate the probability of an outcome falling in the acceptance region (given $\mu = \mu_1$). This is P(Type II error) for the case where $\mu = \mu_1$.

Notes

- As usual, it is sensible to avoid premature rounding during intermediate calculations.
- When calculating the probability of a Type II error for a test concerning a proportion, remember that a change in the value of p will change the value of the quantity pq that occurs in the variance of the test statistic.

Example 8

A machine is supposed to fill bags with 38 kg of sand. It is known that the quantities in the bags vary and have a standard deviation of 0.5 kg. When a new employee starts using the machine it is standard practice to determine the masses of a random sample of 20 bags taken from the first batch produced by the employee in order to verify that the mean of the machine has been set correctly.

Determine an appropriate test procedure, given that it is desired that the probability of a Type I error should be 4%.

Suppose that an employee has set the machine so that it fills bags with an average of μ kg.

Determine the probability of a Type II error in the cases $\mu = 38.1$ and $\mu = 38.4$.

---

1 *Write down H_0 and H_1.*

The test is two-tailed with the hypotheses being:

H_0: $\mu = 38.0$

H_1: $\mu \neq 38.0$

2 *Determine the appropriate test statistic and the distribution of the corresponding random variable (using the parameter value specified by H_0).*
The appropriate test is one that uses the sample mean, $\bar{x}$. Assuming H_0, by the Central Limit Theorem the distribution of $\bar{X}$ is approximately:

$$N\left(38.0, \frac{(0.5)^2}{20}\right)$$

so that the appropriate test statistic is:

$$z = \frac{\bar{x} - 38.0}{\sqrt{\frac{0.250}{20}}}$$

3 *Determine the significance level. This is P(Type I error).*
This is given as 4% (which makes a change from the more usual 5%).

4 *Determine the acceptance and rejection regions.*
Tables show that the upper 2% point of a standard normal random variable is 2.054. The acceptance region (in terms of z) is therefore $(-2.054, 2.054)$. In order to calculate the probability of a Type II error we will need this as an interval for $\bar{x}$:

$$\left(38.0 - 2.054\sqrt{\frac{0.250}{20}},\ 38.0 + 2.054\sqrt{\frac{0.250}{20}}\right)$$

which simplifies to (37.7704, 38.2296). This is the acceptance region for $\bar{x}$. Values of $\bar{x}$ outside this interval are in the rejection region.

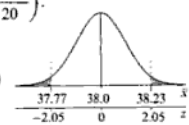
Consider now the case $\mu = 38.1$

5 *Determine the distribution of the random variable corresponding to the test statistic.*
The distribution of $\bar{X}$ is:

$$N\left(38.1, \frac{0.250}{20}\right)$$

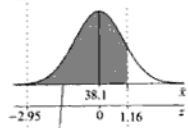
6 *Calculate the probability of an outcome falling in the acceptance region. This is P(Type II error).*
We require $P(37.7704 < \bar{X} < 38.2296)$, given that $\bar{X} \sim N\left(38.1, \frac{0.250}{20}\right)$.
Now:

$$P(\bar{X} < 38.2296) = P\left(Z < \frac{38.2296 - 38.1}{\sqrt{\frac{0.250}{20}}}\right) = P(Z < 1.159)$$



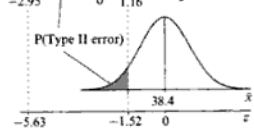
where $Z \sim N(0, 1)$. Similarly:

$$P(\bar{X} < 37.7704) = P\left(Z < \frac{37.7704 - 38.1}{\sqrt{\frac{0.250}{20}}}\right) = P(Z < -2.948)$$



Thus:

$$\begin{aligned} P(37.7704 < \bar{X} < 38.2296) &= \Phi(1.159) - \Phi(-2.948) \\ &= 0.8767 - (1 - 0.9984) \\ &= 0.8751 \end{aligned}$$



When the mean is only slightly too large (38.1, rather than 38.0) the probability of a Type II error is high, being 0.875 (to three decimal places).

Consider now the case $\mu = 38.4$. We need to calculate:

$$\frac{38.2296 - 38.4}{\sqrt{\frac{0.250}{20}}} = -1.524$$

and:

$$\frac{37.7704 - 38.4}{\sqrt{\frac{0.250}{20}}} = -5.631$$

so that now:

$$\begin{aligned} P(37.7704 < \bar{X} < 38.2296) &= \Phi(-1.524) - \Phi(-5.631) \\ &= (1 - 0.9362) - 0 \\ &= 0.0638 \end{aligned}$$

When a more substantial shift in the mean is considered, the probability of a Type II error is considerably reduced. When $\mu = 38.4$ the value is 0.064 (to three decimal places).

Example 9

A coin, believed to be fair, is tossed 100 times. The hypothesis that the coin is fair will be accepted if the number of heads obtained lies between 40 and 60, inclusive.
Determine the probability of a Type I error.
Determine also the probability of a Type II error for the case where the probability of a head is 0.6.
State the power of the test in this case.

In this question the acceptance region is given, but it is still useful to follow through the general procedure. We will let X be the number of heads obtained and let p be the probability of a head.

1 *Write down H_0 and H_1.*
The hypotheses are:
H_0: $p = 0.5$
H_1: $p \neq 0.5$.

2 *Determine the appropriate test statistic and the distribution of the corresponding random variable (using the parameter value specified by H_0).*
The test statistic is the number of heads obtained.

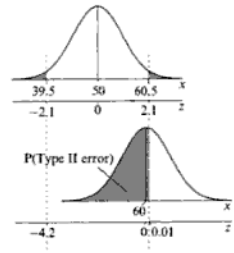
3, 4 *Determine the acceptance and rejection regions and the significance level*
We are given that the acceptance region is $40 \leqslant X \leqslant 60$. We require P(Type I error) which is therefore equal to $1 - P(40 \leqslant X \leqslant 60)$. Assuming H_0, $X \sim B(100, 0.5)$. Since the number of trials is large we can use the normal approximation, together with continuity

corrections, to determine the required probability. The distribution of X is approximated by:

$$N(50, 100 \times 0.5 \times 0.5) = N(50, 25)$$

Hence, using continuity corrections:

$$\begin{aligned}\text{P(Type I error)} &\approx 1 - \left\{\Phi\left(\frac{60.5 - 50}{5}\right) - \Phi\left(\frac{39.5 - 50}{5}\right)\right\} \\ &= 1 - \{\Phi(2.1) - \Phi(-2.1)\} \\ &= 1 - \{0.9821 - (1 - 0.9821)\} \\ &= 0.0358\end{aligned}$$



The probability of a Type I error (the significance level) is about 3.6%.

Consider now the case $p = 0.6$.

5 *Determine the distribution of the random variable corresponding to the test statistic.*

The normal approximation becomes

$$N(60, 100 \times 0.6 \times 0.4) = N(60, 24)$$

Note that the variance has slightly altered as a consequence of the change in the value of p.

6 *Calculate the probability of an outcome falling in the* ***acceptance region****. This is P(Type II error).*

The probability of a type II error is given by:

$$\begin{aligned}P(40 \leqslant X \leqslant 60) &\approx \Phi\left(\frac{60.5 - 60}{\sqrt{24}}\right) - \Phi\left(\frac{39.5 - 60}{\sqrt{24}}\right) \\ &= \Phi(0.102) - \Phi(-4.185) \\ &= 0.5406 - 0 \\ &= 0.5406\end{aligned}$$

The probability of a Type II error is about 54.1%.

The power of the test is 1 − P(Type II error) = **1 − 0.5406 = 0.460** (to 3 decimal places).

Example 10

The random variable X has a binomial distribution with unknown p.
Devise a test of H_0: $p = 0.4$ against H_1: $p < 0.4$.

The test is to be based on 240 observations, and should have a significance level of about 1%.
Determine the approximate probability of a Type II error for your test in the case where $p = 0.3$.

1 *Write down H_0 and H_1.*

H_0: $p = 0.4$

H_1: $p < 0.4$

2 *Determine the appropriate test statistic and the distribution of the corresponding random variable (using the parameter value specified by H_0).*
Assuming H_0, the distribution of X is approximated by:

$$N(240 \times 0.4,\ 240 \times 0.4 \times 0.6) = N(96, 57.6)$$

3 *Determine the significance level. This is P(Type I error).*
The significance level is to be about 1%. It may not be *exactly* 1% because the binomial distribution is a discrete distribution with probability that comes in 'chunks'.

4 *Determine the acceptance and rejection regions.*
For a standard normal random variable Z, $P(Z < -2.326) = 0.01$. The appropriate critical value for X is therefore:

$$96 - (2.326\sqrt{57.6}) = 78.35$$

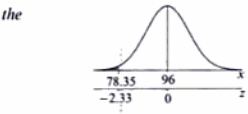
However, the only possible values for X are integers. The resulting procedure is:

Accept H_0 unless the observed value of X is 78 or less, in which case H_0 should be rejected in favour of H_1.

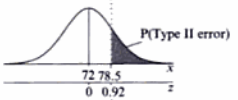
Suppose now that $p = 0.3$

5 *Determine the distribution of the random variable corresponding to the test statistic.*
Using the normal approximation the distribution of X is approximately:

$$N(240 \times 0.3,\ 240 \times 0.3 \times 0.7) = N(72, 50.4)$$



6 *Calculate the probability of an outcome falling in the acceptance region. This is P(Type II error).*
The probability of a Type II error is the probability of observing a value greater than 78. Using a continuity correction, we proceed as follows:



$$P(X > 78) \approx 1 - \Phi\left(\frac{78.5 - 72}{\sqrt{50.4}}\right)$$

$$= 1 - \Phi(0.916)$$

$$= 1 - 0.8201$$

$$= 0.1799$$

The probability of a Type II error, when the value of p is 0.3, is 0.180 (to three decimal places).

Exercises 12g

1 It is given that $X \sim N(\mu, 16)$. It is desired to test the null hypothesis $\mu = 12$ against the alternative hypothesis $\mu > 12$, with the probability of a Type I error being 1%. A random sample of 15 observations of X is taken and the sample mean $\bar{X}$ is taken to be the test statistic.
(i) Find the acceptance and rejection regions.
(ii) For the case $\mu = 15$, find the probability of a Type II error and the power of the test.

2 It is given that $Y \sim N(\mu, 25)$. It is desired to test the null hypothesis $\mu = 20$ against the alternative hypothesis $\mu < 20$, with the probability of a Type I error being 5%. A random sample of 100 observations of Y is taken and the sample mean $\bar{Y}$ is taken to be the test statistic.
(i) Find the acceptance and rejection regions.
(ii) For the case $\mu = 19$, find the probability of a Type II error and the power of the test.

3 The temperature of an item taken from a freezer cabinet is X°C. X may be taken to be a normal variable with mean μ and standard deviation 1.8. A random sample of 11 items is taken from the cabinet, and the mean $\bar{X}$ of their temperatures is to be used as test statistic. It is desired to test the null hypothesis $\mu = -5.5$ against the alternative hypothesis $\mu \neq -5.5$, with the probability of a Type I error equal to 0.10.
(i) Find the acceptance region.
(ii) For the case $\mu = -7.0$, find the probability of a Type II error and the power of the test.

4 The random variable X is normally distributed with mean μ and standard deviation 11. The null hypothesis $\mu = 52$ is to be tested against the alternative hypothesis $\mu > 52$ using a 5 per cent significance level. The mean $\bar{X}$ of a random sample of 150 observations of X is to be used as the test statistic.
(i) Find the range of values of the test statistic which lie in the critical region.
(ii) When $\mu = 54$ calculate, to two decimal places, the probability of a type-2 error and the power of the test. [JMB]

5 *[Use a t-distribution in this question.]*
The haemoglobin levels (in g/100l) of a random sample of ten elderly male cancer patients are as follows:

13.6, 11.9, 13.4, 12.4, 12.4,
13.2, 12.7, 15.7, 14.8, 12.0

Extensive evidence suggests that for healthy elderly males, the mean haemoglobin level is 13.0. The null hypothesis, H_0, is therefore that the mean haemoglobin level of the population of elderly male cancer patients is 13.0.

(a) State what is meant by a *Type I error*. Suppose that the alternative hypothesis, H_1, is that the mean haemoglobin level is not 13.0. Use a 5% significance level to test the null hypothesis.

(b) State what is meant by a *Type II error*. Show that, if the true mean haemoglobin level is 15.0, then the power of the test used in (a) is approximately 0.99.

(c) Explain how the test used in (a) would be modified if the alternative hypothesis had been that the mean haemoglobin level is less than 13.0.
What would the outcome of the modified test have been?

(d) Explain carefully what is meant by the phrase *confidence interval*.
Obtain a symmetric 99% confidence interval for the mean haemoglobin level of patients of the type sampled.

6 A bag contains a very large number of marbles, identical except for their colour. Of these, an unknown proportion p are red. It is required to test the null hypothesis

$$H_0: p = 0.3$$

against the alternative hypothesis

$$H_1: p < 0.3$$

In order to perform the test, a random sample of 100 marbles is taken and the number X of red marbles noted. The distribution of X is to be approximated by a normal distribution.
(i) If the significance level is 10%, determine whether the null hypothesis should be accepted in the case when $X = 25$.
(ii) If the significance level of the test is to be as close as possible to 5%, find the critical region in the form $0 \leqslant X \leqslant a$, where a is an integer.
(iii) Calculate the power of the test in the case when the critical region is $0 \leqslant X \leqslant 24$ and $p = 0.2$. [JMB]

12.13 Hypothesis tests for a proportion based on a small sample

The difficulties associated with this type of hypothesis test are best illustrated with an example. Consider the following problem.

> The standard treatment for a particular disease is successful on only 40% of occasions. A new treatment is introduced that is supposed to be better. Initially the treatment is given to just ten patients: the treatment is successful eight times.
> Does this provide significant evidence, at the 5% level, that the new treatment is significantly better than the standard treatment?

This clearly requires a hypothesis test, so let us follow through our standard procedure.

1 *Write down H_0 and H_1*
We will ignore the fact that the results of the new treatment might have appeared worse than the standard treatment, since there appears to be a strong expectation that the new treatment is at least as good as the old. So the test is one-tailed and the hypotheses are:

$H_0: p = 0.4$
$H_1: p > 0.4$

2 *Determine the appropriate test statistic and the distribution of the corresponding random variable (using the parameter value specified by H_0).*
For a test of a hypothesis about p it is natural to use $\hat{p}$. However, we cannot use a normal approximation since n is small, so it is simpler to work directly with the number of 'successes', x. Assuming H_0, the distribution of the corresponding random variable, X, is:

$B(10, 0.4)$

3 *Determine the significance level.*
The question specifies 5%.

4 *Determine the acceptance and rejection regions.*
The test is one-tailed: only if we see a large value of x are we going to reject H_0 in favour of H_1. We therefore need to look at the probabilities of large values of x occurring under the conditions specified by H_0. From tables of the B(10, 0.4) distribution (Appendix, p. 437) we have the following:

r	$P(X = r)$	$P(X \geqslant r)$
10	0.0001	0.0001
9	0.0016	0.0017
8	0.0106	0.0123
7	0.0425	0.0548
6	0.1115	0.1662

There is a problem here – probability comes in chunks! Suppose that we decide on the following strategy.

> Reject H_0 in favour of H_1 if $X \geqslant 7$, accept H_0 otherwise.

The significance level was defined as the probability of an observation falling into the rejection region by chance under the conditions specified by H_0. The rejection region is the set of values 7, 8, 9 and 10 and the significance level is therefore 5.48% and not 5%. A significance level of exactly 5% is not obtainable. In this case we refer to '5%' as the **nominal significance level**, with '5.48%' being the **actual significance level** corresponding to the rejection region selected.

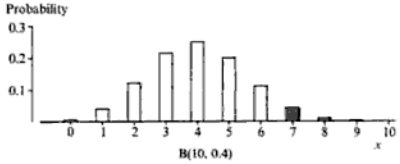
A conservative approach would be to define the critical region so that the actual significance level was not greater than the nominal significance level. In the present example this would imply using the following strategy.

Reject H_0 in favour of H_1 if $X \geqslant 8$, accept H_0 otherwise.

This strategy has an actual significance level of 1.23%.

This approach is adopted in some textbooks (and, implicitly, by some syllabuses), but we prefer an 'average' approach. This defines the rejection region so as to obtain an actual significance level as close to the nominal level as possible, without the restriction that it may not exceed it. Applying this strategy in lots of cases should lead to an average significance level close to the nominal value. In the present case this approach means working with an actual significance level of 5.48%.

Whichever approach is adopted, it is good practice to report the actual significance level, rather than the nominal significance level, when reporting the final decision.



5 *Calculate the value of the test statistic.*
The observed value of X was 8.

6 *Determine the outcome of the test.*
Since 8 lies in the rejection region $\{7, 8, 9, 10\}$ we can state that there is significant evidence, at an actual level of 5.48%, that the new treatment is successful in more than 40% of cases.

Notes

- The distinction between the nominal and actual significance levels is glossed over in some texts (and in some syllabuses!). We recommend that, where possible, the actual significance level is reported.
- Sometimes the possible significance levels are quite remote from the nominal level. For a nominal 5% test, the achievable significance levels may be 3% and 7%. If the outcome of the test is unaffected by the level chosen, then this does not matter. If the outcome is significant at the 7% level, but not at the 3% level, then the two results should be reported – the results might reasonably be described as inconclusive, if really the 5% level is required.
- The problem caused by discreteness also affects two-tailed tests. For a two-tailed 5% hypothesis test for a continuous distribution, it is simple to assign 2.5% to each tail. However, this is unlikely to be true for a discrete distribution, and we recommend applying the 'average' approach to each tail separately.

Example 11

According to a genetic theory, $\frac{1}{4}$ of a certain group of plants should have red flowers. A random sample of 12 plants is examined. Six have red flowers. Does this provide significant evidence, at a nominal 5% level, to reject the hypothesis?

---

1 *Write down H_0 and H_1.*
The test is two-tailed. The hypotheses are:

$H_0: p = 0.25$

$H_1: p \neq 0.25$

2 *Determine the appropriate test statistic and the distribution of the corresponding random variable (using the parameter value specified by H_0).*
We use X, the number of plants with red flowers as our test statistic. Assuming H_0, the distribution of X is:

$B(12, 0.25)$

3 *Determine the significance level.*
The question specifies 5%.

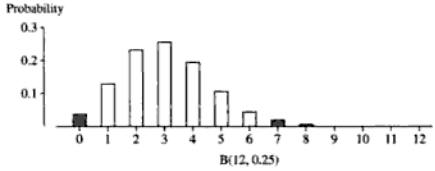
4 *Determine the acceptance and rejection regions.*
The test is two-tailed, so we need to consider the cumulative probabilities associated with each tail of the B(12, 0.25) distribution:

Upper tail			Lower tail		
r	$P(X = r)$	$P(X \geqslant r)$	r	$P(X = r)$	$P(X \leqslant r)$
12	0.0000	0.0000	0	0.0317	0.0317
11	0.0000	0.0000	1	0.1267	0.1584
10	0.0000	0.0000			
9	0.0004	0.0004			
8	0.0024	0.0028			
7	0.0115	0.0143			
6	0.0401	0.0544			

In the upper tail the nearest that we can get to 2.5% is 1.43%. In the lower tail the nearest that we can get is 3.17%. We therefore propose the decision rule:

> Reject H_0 in favour of H_1 if the observed value of X is either 0 or at least 7. Otherwise accept H_0.

The actual significance level is $1.43\% + 3.17\% = 4.60\%$.



5 *Calculate the value of the test statistic.*
The observed value of X was 6.

6 *Determine the outcome of the test.*
Since 6 lies in the acceptance region $\{1, 2, 3, 4, 5, 6\}$ we accept the hypothesis H_0: $p = 0.25$, using an actual significance level of 4.60%.

12.14 Hypothesis tests for a Poisson mean based on a small sample

The problems here are essentially the same as those of the previous section. However, since Poisson distributions have infinite range, it is always sensible to work upwards from the outcome zero. The following example illustrates the procedure.

Example 12

A company uses a large number of floppy disks. At random intervals in time, disks develop faults: on average 0.4% of black disks fail per month. The company also has blue disks. During a randomly chosen nine-month period a random sample of 100 blue disks develop a total of 7 faults. Is there significant evidence, at the 5% level, that the failure rate of the blue disks is not 0.4% per month?

1 *Write down H_0 and H_1.*
The test is two-tailed. For convenience we will work with X, the total number of faults in 100 disks during a nine-month period. This has a Poisson distribution since faults occur at random intervals in time. The hypotheses are therefore:
H_0: Rate $= 0.4\%$
H_1: Rate $\neq 0.4\%$

2 *Determine the appropriate test statistic and the distribution of the corresponding random variable (using the parameter value specified by H_0).*
Assuming H_0, the distribution of X is Poisson with mean $0.004 \times 100 \times 9 = 3.6$

3 *Determine the significance level.*
The question specifies 5%.

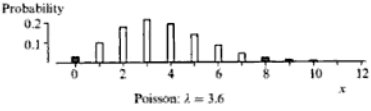
4 *Determine the acceptance and rejection regions.*
The test is two-tailed, so we need to consider the cumulative probabilities associated with each tail of the distribution, using either tables or direct calculation:

r	$P(X = r)$	$P(X \leqslant r)$	$P(X \geqslant r)$
0	0.0273	0.0273	1.0000
1	0.0984	0.1257	0.9727
2	0.1771		0.8743
3	0.2125		0.6973
4	0.1912		0.4848
5	0.1377		0.2936
6	0.0826		0.1559
7	0.0425		0.0733
8	0.0191		0.0308
9	0.0076		0.0117

In the lower tail the nearest that we can get to 2.5% is 2.73%. In the upper tail the nearest that we can get is 3.08%. We therefore propose the decision rule:

> Reject H_0 in favour of H_1 if the observed value of X is either 0 or at least 8. Otherwise accept H_0.

The actual significance level is $2.73\% + 3.08\% = 5.81\%$.



5 *Calculate the value of the test statistic.*
The observed value of X was 7.

6 *Determine the outcome of the test.*
Since 7 lies in the acceptance region $\{1, 2, 3, 4, 5, 6, 7\}$ we accept the null hypothesis (that the rate is 0.4%) using a significance level of 5.81%. The findings do not provide significant evidence that the blue disks have a different failure rate to the black disks.

Exercises 12h

1 A die is suspected of being biased towards the score of 6. It is thrown 10 times and the number n of sixes is observed.
State suitable null and alternative hypotheses and determine the acceptance and rejection regions for a test at a nominal significance level of 10%.

State the actual significance level of the test.

What are the conclusions in the cases
(i) $n = 3$, (ii) $n = 5$?

2 The proportion of £1 coins that bear the motto *Nemo me impune lacessit* is denoted by p. A random sample of 9 coins is taken and the number n that bear the motto is observed. It is desired to test the null hypothesis $p = 0.40$ against the alternative hypothesis $p \neq 0.40$ at a nominal significance level of 10%.
Determine the appropriate acceptance region and the corresponding actual significance level.

3 Two methods are proposed to test whether a coin is biased.

(a) In the first method, the coin will be tossed 10 times and it will be considered biased if at least 8 heads or at least 8 tails are obtained.

(i) Show that the probability of making a type-1 error is approximately 0.11.

(ii) Determine, to two decimal places, the probability of making a type-2 error when the probability of obtaining a head on each toss is actually 0.6.

(b) In the second method, the coin will be tossed 100 times and it will be considered biased if at least 60 heads or at least 60 tails are obtained. Calculate, to two decimal places, an approximate value of the significance level of the test. **[JMB]**

4 Explain the terms *critical region*, *significance level* and *power* in the context of hypothesis testing.

National publicity was given to a university microbiologist's claim that 30% of pre-cooked chicken portions sold in supermarkets are contaminated with listeria. A large supermarket chain arranged to test, in its own laboratory, a random sample of 20 chicken portions from its supplier. Although it believed that the microbiologist was overstating the problem, the supermarket chain decided that it would contest the microbiologist's claim only if fewer than 3 of the chicken portions tested proved to be contaminated with listeria.

(continued)

Considering this as a hypothesis testing problem, state suitable null and alternative hypotheses.

State the critical region and determine the significance level of the test.

Determine the power of the test if 15% of the supermarket chain's chicken portions are contaminated with listeria.

Subsequently the chain commissioned an independent laboratory to carry out tests on a random sample of 120 chicken portions, of which 22 proved to be contaminated.

Using a 1% significance level, conduct a test to decide whether the chain has sufficient evidence to conclude that the listeria contamination affects less than 30% of its chicken portions. [JMB]

5 A certain production process is said to be out of control when the proportion p of its output which is defective exceeds 5%. A test is required to decide between the hypotheses: $H_0 : p = 0.05$ and $H_1 : p > 0.05$. The test suggested is to take a random sample of 20 items and reject H_0 if more than 2 items are defective. Calculate

(i) the significance level of this test,

(ii) the power of the test when 10% of the output is defective.

Without carrying out any further calculations, state briefly why the answers to (i) and (ii) should cause some concern.

Suggest a modification which could be made to the test in order that both the significance level and the power might be improved. [JMB]

6 Dr Zed believes he possesses psychic powers. He claims that, when shown the back of a normal playing card, he has a better than 25% chance of predicting the suit of the card. A statistician is invited to investigate this claim. She asks Dr Zed to make independent predictions of the suits of 20 cards. The number of times Dr Zed predicts correctly is denoted by X and p is the probability that any given prediction is correct. To decide between the hypotheses:

$$H_0 : p = 0.25 \quad \text{and} \quad H_1 : p > 0.25$$

the statistician decides to accept H_0 if $X \leqslant r$, where r is the least integer for which the probability of a type I error is less than 5%. Using appropriate tables, or otherwise, find the value of r.

Dr Zed claims that the true value of p is 0.6. Show that, if this claim is true, the probability that a type II error will be made using the above test is approximately 6%. [JMB]

7 Describe the roles of the null and alternative hypotheses in a test of significance. Explain how to decide whether the use of a one-tail or a two-tail test is appropriate.

Over a long period it has been found that the ratio of females to males attending classical ballet performances is 13 females to 7 males.

(a) On the afternoon of a football cup match, a random sample of 20 people attending a classical ballet performance is found to contain 4 males. Carry out a significance test to determine whether or not the proportion of males attending is lower than usual. State clearly your null and alternative hypotheses, and use a 10% significance level.

(b) At a contemporary ballet performance, a random sample of 100 people attending is found to contain 44 males. Set up null and alternative hypotheses and test whether the mean number of males attending contemporary ballet performances is different from that associated with classical ballet performances. Use a normal approximation and a 5% level of significance. [ULSEB]

8 Explain what is meant by the *null hypothesis* and the *alternative hypothesis* in significance testing.

Explain the difference between a *one-sided* and a *two-sided* test, briefly describing how you would decide which one to use.

The annual number of accidents at a certain road junction could be modelled by a Poisson distribution with mean 8.5. New road markings are introduced at the junction which may reduce the accident rate, and which are not expected to increase it. In order to test their effectiveness it is planned to use the number of accidents at the junction in the following year. You may assume that, after the change in road markings, a Poisson model is still appropriate.

(continued)

12.16 Comparison of two means – known population variances

The random variable X has unknown mean μ_x and known variance σ_x^2. The independent random variable Y has unknown mean μ_y and known variance σ_y^2. The null hypothesis is:

$$H_0: \mu_x = \mu_y, \text{ or equivalently, } \mu_x - \mu_y = 0$$

The alternative hypothesis may be two-sided:

$$H_1: \mu_x \neq \mu_y$$

or one-sided, e.g.

$$H_1: \mu_x > \mu_y$$

Since the hypotheses concern the population means, the test statistic will involve the sample means $\bar{x}$ and $\bar{y}$. Suppose that the samples have sizes n_x and n_y. Then, if X and Y have normal distributions, the same will be true for $\bar{X}$ and $\bar{Y}$:

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x^2}{n_x}\right)$$

$$\bar{Y} \sim N\left(\mu_y, \frac{\sigma_y^2}{n_y}\right)$$

These results also hold, approximately, for large samples from other distributions, because of the Central Limit Theorem. In both cases $\bar{X} - \bar{Y}$ has a normal distribution with mean $\mu_x - \mu_y$ and variance $\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$. Hence:

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim N(0, 1)$$

Assuming H_0, so that $\mu_x - \mu_y = 0$, we can calculate:

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

and then proceed as usual.

Confidence interval for the common mean

According to H_0, $\mu_x = \mu_y$. We denote their common value by μ. All the n_x observations on X (i.e. $x_1, x_2, \ldots, x_{n_x}$) and all the n_y observations on Y (i.e. $y_1, y_2, \ldots, y_{n_y}$) therefore come from populations having the same mean. A natural **pooled estimate of the population mean**, $\hat{\mu}$, is therefore given by:

$$\hat{\mu} = \frac{\sum_{i=1}^{n_x} x_i + \sum_{j=1}^{n_y} y_j}{n_x + n_y} = \frac{n_x\bar{x} + n_y\bar{y}}{n_x + n_y} \qquad (12.3)$$

The distribution of the corresponding random variable is:

$$N\left(\mu, \frac{n_x\sigma_x^2 + n_y\sigma_y^2}{(n_x + n_y)^2}\right)$$

(see the note below).

Following the usual arguments, the corresponding 95% confidence interval for μ is:

$$\left(\hat{\mu} - 1.96\frac{\sqrt{n_x\sigma_x^2 + n_y\sigma_y^2}}{(n_x + n_y)}, \quad \hat{\mu} + 1.96\frac{\sqrt{n_x\sigma_x^2 + n_y\sigma_y^2}}{(n_x + n_y)}\right)$$

If $\sigma_x^2 = \sigma_y^2$ $(= \sigma^2$, say) then, writing $n_x + n_y$ as n, the confidence interval simplifies to become:

$$\left(\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}}, \quad \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Note

- Since $\bar{X}$ has variance $\frac{\sigma_x^2}{n_x}$, the quantity $\frac{n_x\bar{X}}{n_x + n_y}$ has variance:

$$\left\{\frac{n_x}{(n_x + n_y)}\right\}^2 \times \frac{\sigma_x^2}{n_x} = \frac{n_x\sigma_x^2}{(n_x + n_y)^2}$$

Combining this expression with the corresponding expression for the variance of $\frac{n_y\bar{Y}}{(n_x + n_y)}$, results in the expression for the variance given above.

Example 13

Suppose that random samples from two independent normal populations give the following results:

Sample 1: $n_x = 100, \quad \bar{x} = 46.0$
Sample 2: $n_y = 120, \quad \bar{y} = 47.0$

Suppose that the specified significance level is 5%, that the population variances are known to be $\sigma_x^2 = 16.0$ and $\sigma_y^2 = 24.0$, and that the hypotheses being compared are:

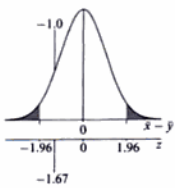
H_0: $\mu_x = \mu_y$
H_1: $\mu_x \neq \mu_y$

Show that H_0 may be accepted, and obtain a symmetric 99% confidence interval for the common population mean.

Assuming H_0, the test statistic z, given by:

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

is an observation from a N(0, 1) distribution. Since the alternative hypothesis is two-sided, and the significance level is 5%, we will accept H_0 only if $-1.96 < z < 1.96$.



The observed value of z is:

$$\frac{46.0 - 47.0}{\sqrt{\frac{16.0}{100} + \frac{24.0}{120}}} = \frac{-1.0}{\sqrt{0.16 + 0.20}} = -1.67$$

We therefore accept, at the 5% significance level, the hypothesis that there is no difference between the means of the two populations.

The pooled estimate of the common population mean is:

$$\frac{n_x\bar{x} + n_y\bar{y}}{n_x + n_y} = \frac{(100 \times 46.0) + (120 \times 47.0)}{220} = 46.545$$

The variance of the corresponding random variable is:

$$\frac{n_x\sigma_x^2 + n_y\sigma_y^2}{(n_x + n_y)^2} = \frac{(100 \times 16.0) + (120 \times 24.0)}{220^2} = 0.092\,56$$

and thus the 99% confidence interval for μ is given by:

$$(46.545 - 2.576 \times \sqrt{0.092\,56},\quad 46.545 + 2.576 \times \sqrt{0.092\,56})$$

which simplifies to (45.76, 47.33).

Example 14

The standard deviation of the scores obtained on a particular test of mathematical ability is known to be 15. A school experiments with a new method of teaching which is supposed to increase general quantitative awareness. A group of 99 students are randomly assigned to one of two classes. The 50 students in the first class are given the new method of teaching, whereas the 49 students in the second class are taught in the standard way.

At the end of the year, the two classes are given the same test of mathematical ability. The mean for the first class is 116.0, whereas that for the second class is 113.1.

Does this provide significant evidence, at the 5% level, that the new method leads to a higher mean performance? State carefully any assumptions made.

We begin at the end! The assumptions that we need to make are the following:

1. The students allocated to the two classes originally had the same average ability (otherwise any apparent differences may be because there were better students in one class than the other).
2. The same degree of effort was put into the two types of teaching. (In practice this is usually *not* the case. A new method usually requires extra effort by all concerned and it may be this effort rather than the method itself that affects the results.)
3. The two classes are assumed to be effectively random samples from the population of national students. If this is not the case, then the results of the experiment cannot be generalised to the wider population.

Assumptions like these may seem pedantic and are often not spelt out. However, if they are *not* all true then the school's experiment is useless as a guide to future results.

Given that the assumptions *are* valid, we can proceed as usual. We denote the mean of the (hypothetical) population of students taught by the new method by μ_x, with the mean for students taught in the normal way being denoted by μ_y. The hypotheses are therefore:

$$H_0: \mu_x = \mu_y$$

$$H_1: \mu_x > \mu_y$$

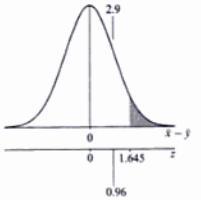
The sample sizes are sufficiently large that we can assume that the sample means are observations from normal distributions. Assuming H_0, the test statistic z, given by:

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{\sigma_x^2}{n_x} + \dfrac{\sigma_y^2}{n_y}}}$$

is therefore an observation from a N(0, 1) distribution. Since the alternative hypothesis is one-sided, and the significance level is 5%, we will accept H_0 only if $z > 1.645$.

The observed value of z is:

$$\frac{116.0 - 113.1}{\sqrt{\frac{15^2}{49} + \frac{15^2}{50}}} = \frac{2.9}{\sqrt{9.092}} = 0.962$$



This value is considerably smaller than 1.645 and we can therefore confidently accept H_0: there is no significant evidence that the mean score using the new teaching method is greater than that using the traditional method. Subject to the assumptions made previously, the school need not switch to the new method.

Exercises 12i

1 The mean of a random sample of 10 observations from a population distributed as $N(\mu_1, 25)$ is 97.3. The mean of a random sample of 15 observations taken from a population distributed as $N(\mu_2, 36)$ is 101.2. Test, at the 5% level, (i) whether $\mu_1 < \mu_2$, (ii) whether $\mu_1 \neq \mu_2$.

2 A random sample of 85 observations is taken from a population with standard deviation 10.2, and the sample mean is 31.2. A random sample of 72 observations is taken from a second population with standard deviation 15.8, and the sample mean is 35.5. Test, at the 1% level, whether the second population has a greater mean than the first.

3 Two wine producers A and B have identical machines that fill bottles of wine. For A the quantity of wine put into a bottle is $(k_A + X)$ cl, where k_A is a constant and X is a normal random variable with mean 0 and standard deviation 0.180. For B the quantity of wine put into a bottle is $(k_B + Y)$ cl, where k_B is a constant and Y has the same distribution as X.

A retailer buys 8 bottles from A and measures the contents in cl. He finds the sample mean is 75.22 cl. He also buys 10 bottles from B and finds that the mean content is 74.91 cl.

Is there significant evidence, at the 5% level, that, on average, bottles from A contain more than bottles from B?

4 A machine assesses the life of a ball-point pen, by measuring the length of a continuous line drawn using the pen. A random sample of 80 pens of brand A have a total writing length of 96.84 km. A random sample of 75 pens of brand B have a total writing length of 93.75 km. Assuming that the standard deviation of the writing length of a single pen is 0.15 km for both brands, test at the 5% level, whether the writing lengths of the two brands differ significantly.

Francis Ysidro Edgeworth 1845–1926 was an Irishman who obtained degrees in Classics from Trinity College, Dublin and Oxford University. On leaving Oxford he studied commercial law, becoming a barrister in 1877.

At the same time as his law studies, Edgeworth educated himself in mathematics and in 1880 he became lecturer in logic at King's College, London. Subsequently he turned his attention to Probability and Statistics, and in 1885 he delivered a paper entitled 'Methods of Statistics' to the Royal Statistical Society.

Edgeworth's early statistical work was concerned with formulating two-sample tests of means (though not with the structure shown in this chapter). His major contributions to Statistics were in the areas of correlation and regression, which will be encountered in Chapter 14.

Equivalent expressions that make use of the quantities calculated by statistical calculators are:

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2} \tag{12.6}$$

and:

$$s_p^2 = \frac{n_x\sigma_{n_x}^2 + n_y\sigma_{n_y}^2}{n_x + n_y - 2} \tag{12.7}$$

where $\sigma_{n_x}^2$ and $\sigma_{n_y}^2$ are the two sample variances.

What happens next depends on the sizes of the samples.

Large sample sizes

If the sample sizes are large then, because of the Central Limit Theorem, the distributions of $\bar{X}$ and $\bar{Y}$ will be approximately normal, so that, assuming H_0:

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\sigma^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \sim N(0, 1)$$

If the sample sizes are very large then s_p^2, the pooled estimate of the common variance, will be an excellent approximation to the unknown σ^2. A natural test statistic is therefore z, given by:

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{s_p^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \tag{12.8}$$

Assuming H_0, z may be considered to be an observation from a N(0, 1) distribution and a test procedure can be constructed in the usual way.

Note

- If the sample sizes are very large, but the assumption that $\sigma_x^2 = \sigma_y^2$ cannot be made, then, assuming H_0, it is reasonable to base a test procedure on the test statistic:

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

as in Section 12.16 (p. 348).

Since the sample sizes are large and this is only an approximation, either s_x^2 and s_y^2 or $\sigma_{n_x}^2$ and $\sigma_{n_y}^2$ can be used in this formula.

Example 15

The marks obtained in a statistics paper by a random sample of 200 male students have $\bar{x} = 54.6$ and $s^2 = 101.3$. On the same paper, an independent random sample of 150 female students had a mean mark of 57.1, with $s^2 = 92.4$. Assuming a common population variance, obtain the pooled estimate of this variance, and test, at the 1% significance level, whether there is significant evidence of a difference in the two population means.

Denoting the mark of a randomly chosen male by X, and the mark of a randomly chosen female by Y, the two hypotheses are:

H_0: $\mu_x = \mu_y$

H_1: $\mu_x \neq \mu_y$

where s_p^2, the pooled estimate of the common variance, is given by:

$$s_p^2 = \frac{n_x\sigma_{n_x}^2 + n_y\sigma_{n_y}^2}{n_x + n_y - 2}$$

$$= \frac{(80 \times 24.21) + 100 \times 43.23)}{178}$$

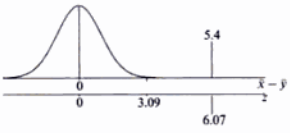
$$= 35.167$$

Hence:

$$z = \frac{74.2 - 68.8}{\sqrt{35.167 \times \left(\frac{1}{80} + \frac{1}{100}\right)}}$$

$$= \frac{5.4}{0.8895}$$

$$= 6.07$$



In this case the sample sizes are so large that a normal approximation can be used. The one-tailed 0.1% point of the standard normal distribution is 3.090. The test procedure is therefore to accept H_0 at the 0.1% significance level, if the value of z is less than 3.090.

Since the observed value, 6.07, is much greater than 3.09, we can confidently reject the null hypothesis in favour of the alternative hypothesis that the apples on Rufus Russett's stall have a greater population mean mass than that of the population of apples sold by Granny Smith.

Project

Are the cars in the local station car park newer than those in the supermarket car park? This could be the case if the 'bread-winner' uses the newest car to drive to the station (and thence to work), while the bread-winner's partner takes an older car to do the shopping.

Assume that all cars with this year's registration letter are 0 years old, that all cars with last year's letter are 1 year old, and so forth. Choose random samples of 50 cars in each situation, record the ages of the cars and perform an appropriate two-sample test to determine whether there is significant evidence to reject the hypothesis that the populations have a common mean.

Exercises 12j

1 A supermarket suspects that the average weight of Grade A melons from supplier X is less than that for Grade A melons from supplier Y. Two random samples are taken and weighed. For 82 melons from X, the results, in kg, are summarised by $\sum x = 58.65$, $\sum x^2 = 51.6460$. For 78 melons from Y, the results are summarised by $\sum y = 61.23$, $\sum y^2 = 55.3425$. Is there evidence at the 5% level to support the supermarket's suspicion:

(i) assuming the population variances are equal,

(ii) without assuming this?

2 In a traffic census drivers are asked the distance, in miles, of their current journey. The figures for a random sample of 120 drivers, between 8 and 9 am, are summarised by $\sum x = 1873$, $\sum x^2 = 56\,285$. The figures for a random sample of 94 drivers, between 1 and 2 pm, are summarised by $\sum y = 1711$, $\sum y^2 = 89\,894$.

Without assuming a common variance, test, at the 10% level, whether the mean distance reported by the 8 to 9 am drivers is less than the mean distance reported by the 1 to 2 pm drivers.

3 Acorns are sown in seed compost A and, after three years, the resulting 105 oak trees have mean height 0.641 m, with the corresponding value of s^2 being 0.0453 m^2. Acorns are also sown in seed compost B and grown in similar circumstances. After three years the 97 trees have mean height 0.578 m, with the corresponding value of s^2 being 0.0712 m^2. Test whether there is significant evidence, at the 5% level, that taller trees are produced in seed compost A:
(i) without assuming that the population variances are equal,
(ii) assuming that the population variances are equal.

4 A consumers' association tests car tyres by running them on a machine until their tread depth reaches a prescribed minimum. 150 tyres of brand A were tested and the equivalent distance, measured in thousands of kilometres, was measured with summary results $\sum(x-30)=974$, $\sum(x-30)^2=10\,051$. The corresponding results for 120 tyres of brand B were $\sum(y-30)=587$, $\sum(y-30)^2=10\,473$. Assuming a common variance, test whether there is significant evidence, at the 5% level, of a difference in the two mean distances.

Small sample sizes

If the sample sizes are not large, then the normal distribution is no longer a reasonable approximation to the distribution of the test statistic. In order to progress we must not only assume a common variance, but also we must:

assume that X and Y have normal distributions.

With this assumption it can be shown that the test statistic t, given by:

$$t=\frac{\bar{x}-\bar{y}}{\sqrt{s_p^2\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}} \qquad (12.9)$$

has a t-distribution, with n_x+n_y-2 degrees of freedom. This is the statistic labelled z in the case of large sample sizes. With the exception of the resulting change in the percentage points (which are found from the t-tables), the test procedure is unchanged.

Example 17

I have two alternative routes to work. The times taken on the 8 randomly chosen occasions that I use route 1 are summarised by $\sum x=182$ and $\sum x^2=4202$, while the times taken on the 12 randomly chosen occasions that I take route 2 are summarised by $\sum y=238$ and $\sum y^2=5108$, with time being measured in minutes.
Assuming that the times taken on either route have normal distributions, with a common variance, determine whether there is significant evidence, at the 5% level, of a difference in the mean times taken on the two routes.

The pooled estimate of the common variance s_p^2 is given by:

$$s_p^2=\frac{\left\{\sum x^2-\frac{1}{n_x}(\sum x)^2\right\}+\left\{\sum y^2-\frac{1}{n_y}(\sum y)^2\right\}}{n_x+n_y-2}$$

$$=\frac{\left(4202-\frac{1}{8}\times 182^2\right)+\left(5108-\frac{1}{12}\times 238^2\right)}{18}$$

$$=24.95$$

The two hypotheses are:

H_0: $\mu_x = \mu_y$

H_1: $\mu_x \neq \mu_y$

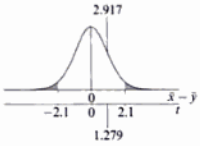
The test statistic is:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s_p^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$$

$$= \frac{\frac{182}{8} - \frac{238}{12}}{\sqrt{24.95 \times \left(\frac{1}{8} + \frac{1}{12}\right)}}$$

$$= \frac{2.917}{\sqrt{5.20}}$$

$$= 1.279$$



Assuming H_0, t is an observation from a t-distribution with 18 degrees of freedom. The two-tailed 5% point of a t_{18}-distribution is 2.101, so that H_0 will be accepted if t falls in the interval $(-2.101, 2.101)$.

The actual value of t is 1.279. There is no reason to reject the null hypothesis that the mean times taken using the two routes are equal.

Exercises 12k

1 The quantities of beer in a random sample of 7 'pints', bought at 'The Sensible Statistician', are measured in litres, and the results are summarised by $\sum x = 4.15, \sum x^2 = 2.4638$. The results for a random sample of 5 'pints' from 'The Mad Mathematician' are summarised by $\sum y = 2.79, \sum y^2 = 1.5585$.
Assuming the population variances are equal, find a pooled estimate of the common variance. Test, at the 5% level, whether there is more beer in a 'pint' from the first pub than the second.

2 A random sample of 10 yellow grapefruit is weighed and the average weight is found to be 201.4 g. The value of an unbiased estimate for the population variance is 234.1 g^2. The corresponding figures for a random sample of 8 pink grapefruit are 221.8 g and 281.9 g^2. Determine, using a 1% level of significance, whether there is a difference in the mean weights of the two kinds of grapefruit.

3 Explain the different circumstances in which a one-tailed or a two-tailed test of significance of a sample mean should be used.
The suppliers of a particular brand of jam claim that their pots of standard size have a mean mass greater than 346 g. A random sample of 8 pots from a particular delivery yielded the following masses, in grams.

342 354 348 344 349 350 347 345

Stating any necessary assumptions, perform a test, at the 10% significance level, to determine whether these data support the manufacturer's claim.
From the next delivery of jam, a random sample of 10 pots yielded the following masses, in grams.

340 341 350 348 342 350 346
344 347 342

Perform a test to determine whether there is evidence, at the 10% significance level, that the mean mass has changed from that of the previous delivery. State any further assumptions required. [UCLES]

4 State the conditions under which a two-sample t-test may validly be applied.
The drug sodium aurothiomalate is sometimes used as a treatment for rheumatoid arthritis. Twenty patients were treated with the drug and, of these, 12 suffered an adverse reaction while 8 did not. The ages, in years, of the two groups were as follows.

Adverse reaction	53 29 53 67 54 57
	51 68 38 44 63 53
No adverse reaction	44 51 64 33
	39 37 41 72

(*continued*)

13 Goodness of fit

To observations which ourselves we make,
We grow more partial for th' observer's sake

Moral Essays, Alexander Pope

Previous chapters have assumed that a particular type of distribution is appropriate for the data given and have focused on estimating and testing hypotheses about the parameter(s) of this distribution. In this chapter the focus switches to the distribution itself, and we ask the question 'Does the data support the assumption that a particular type of distribution is appropriate?'.

Suppose, for example, that we roll an apparently normal six-sided die 60 times and obtain the following observed frequencies:

Outcome	1	2	3	4	5	6
Observed frequency	4	7	16	8	8	17

In this sample (of possible results from rolling the die) there seems to be a rather large number of 3s and 6s. Is this die fair, or is it biased? With a fair die the probability of each outcome is $\frac{1}{6}$. With 60 tosses the expected frequencies would each be $60 \times \frac{1}{6} = 10$:

Outcome	1	2	3	4	5	6
Expected frequency	10	10	10	10	10	10

The question of interest is whether the observed frequencies (O) and the expected frequencies (E) are reasonably close or unreasonably different. We add the differences ($O - E$) to the table:

Observed frequency, O	4	7	16	8	8	17
Expected frequency, E	10	10	10	10	10	10
Difference, $O - E$	−6	−3	6	−2	−2	7

The larger the magnitude of the differences (i.e. ignoring the sign), the more the observed data differs from that expected according to our model (that the die was fair).

Suppose we now roll a second die 660 times, and obtain the following results:

Observed frequency, O	104	107	116	108	108	117
Expected frequency, E	110	110	110	110	110	110
Difference, $O - E$	−6	−3	6	−2	−2	7

This time the observed and expected frequencies seem remarkably close, yet the $O - E$ values are the same as before: it is not simply the size of $O - E$ that matters, but also its size relative to the expected frequency, $\frac{O-E}{E}$.

Combining the ideas that both 'difference' and 'relative size' matter might suggest using the product $(O - E) \times \frac{O-E}{E}$, so that the goodness of fit for outcome i is measured using $\frac{(O_i - E_i)^2}{E_i}$. The smaller this quantity is, the better the fit. An aggregate measure of **goodness of fit** of the model is therefore provided by X^2, defined by:

$$X^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \tag{13.1}$$

- If Z has a N(0,1) distribution, then Z^2 has a χ^2_1 distribution.
- If U and V are independent random variables having χ^2_u and χ^2_v distributions, respectively, then their sum $U + V$ has a χ^2_{u+v} distribution.
- The χ^2_2 distribution is an exponential distribution with mean 2.

Tables of the chi-squared distribution

The usual layout consists of a few selected percentage points giving the columns of the table, with rows referring to different values for v. Here is an extract from the table given in the Appendix (p. 442) at the end of this book:

v	p(%)					
	90	95	97.5	99.0	99.5	99.9
1	2.706	3.841	5.024	6.635	7.879	10.83
2	4.605	5.991	7.378	9.210	10.60	13.82
3	6.251	7.815	9.348	11.34	12.84	16.27
4	7.779	9.488	11.14	13.28	14.86	18.47
5	9.236	11.07	12.83	15.09	16.75	20.52

If X has a χ^2_v distribution, then a tabulated value x is such that $P(X < x) = p\%$. Thus $P(\chi^2_1 < 2.706) = 0.900$, $P(\chi^2_5 > 20.52) = 0.001$ and the upper 1% percentage point of a χ^2_3 distribution is 11.34.

Exercises 13a

1 Find: (i) $P(\chi^2_4 > 11.14)$, (ii) $P(\chi^2_5 < 11.07)$, (iii) $P(\chi^2_3 > 12.84)$, (iv) $P(\chi^2_1 < 6.635)$, (v) $P(\chi^2_1 > 1.96^2)$.

2 Find:
(i) $P(7.779 < \chi^2_4 < 13.28)$,
(ii) $P(11.07 < \chi^2_5 < 16.75)$,
(iii) $P(7.378 < \chi^2_2 < 9.210)$.

3 Find c such that:
(i) $P(\chi^2_4 > c) = 0.005$, (ii) $P(\chi^2_5 > c) = 0.025$,
(iii) $P(\chi^2_1 > c) = 0.100$, (iv) $P(\chi^2_1 < c) = 0.995$,
(v) $P(\chi^2_3 < c) = 0.975$.

4 Verify that the upper percentage points of χ^2_1, given in the table above, are (except for rounding errors) the squares of the corresponding two-tail percentage points of N(0, 1).

5 By finding the cumulative distribution function of an exponential distribution with mean 2, verify the entries in the above table for the case $v = 2$.

13.2 Goodness of fit to prescribed probabilities

The goodness-of-fit statistic proposed earlier was X^2, where:

$$X^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

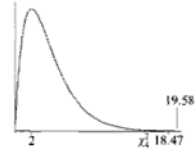
Here O_i and E_i are, respectively, observed and expected frequencies and m is the number of categories being compared. H_0 specifies the probabilities of the various categories, and the expected frequencies are the product of the sample size and these probabilities. The alternative hypothesis is that H_0 is incorrect. Assuming H_0, X^2 is *approximately* an observation from a chi-squared distribution with $m - 1$ degrees of freedom (χ^2_{m-1}).

The null hypothesis is that each student is equally likely to reply. The alternative hypothesis is that the sample of replies is in some way biased.

The university contains a total of 5100 students. The probability that a randomly chosen student belongs to the Arts faculty is therefore $\frac{1100}{5100}$. The expected number of Arts students under the hypothesis that the sample of replies is unbiased would therefore be $\frac{13}{51} \times 300 = 76.47$. The remaining expected frequencies can be calculated in the same way. The results are summarised below:

Faculty	O_i	Probability	E_i	$O_i - E_i$	$\frac{(O_i - E_i)^2}{E_i}$
Arts	101	$\frac{13}{51}$	76.471	24.529	7.868
Engineering	30	$\frac{8}{51}$	47.059	−17.059	6.184
Humanities	69	$\frac{11}{51}$	64.706	4.294	0.285
Law	17	$\frac{5}{51}$	29.412	−12.412	5.238
Science	83	$\frac{14}{51}$	82.353	0.647	0.005
Total	300	1.000	300.000	0	19.580

Since there are 5 faculties the relevant χ^2 distribution has 4 degrees of freedom. The upper 0.1% point of a χ^2_4 distribution is 18.47, which is less than the observed 19.58. There is therefore significant evidence, at the 0.1% level, that the sample is not a fair cross-section of the university. Studying the table it is apparent that the sample contains a higher than expected proportion of Arts students, while the proportions of Engineering and Law students are unduly low.



Example 4

Let S denote the sum of 12 uniform random numbers in (0, 1). It is known that, if the random number generator is working correctly, the distribution of S will be approximately normal, with mean 6 and variance 1.

The distribution of 500 values of S is summarised in the table below:

$s < 4$	$4 \leqslant s < 5$	$5 \leqslant s < 6$	$6 \leqslant s < 7$	$7 \leqslant s < 8$	$8 \leqslant s$
10	75	163	174	66	12

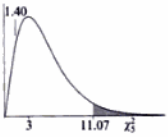
Is there evidence, at the 5% significance level, that the random number generator is working incorrectly?

The null hypothesis is that $S \sim \mathrm{N}(6,1)$, with the alternative being that this is not the case.

$$\text{If } S \sim \mathrm{N}(6,1) \text{ then } \mathrm{P}(S < 4) = \mathrm{P}\left(Z < \frac{4-6}{\sqrt{1}}\right) = \Phi(-2) = 0.0228.$$

For $\mathrm{P}(4 \leqslant S < 5)$ we need to calculate $\mathrm{P}(S < 5) - \mathrm{P}(S < 4)$. This is $\Phi(-1) - \Phi(-2) = 0.1587 - 0.0228 = 0.1359$. We obtain probabilities for the remaining categories in the same way. The calculations are summarised as follows:

Interval	O_i	Probability	E_i	$O_i - E_i$	$\frac{(O_i - E_i)^2}{E_i}$
$s < 4$	10	0.0228	11.40	−1.40	0.172
$4 \leqslant s < 5$	75	0.1359	67.95	7.05	0.731
$5 \leqslant s < 6$	163	0.3413	170.65	−7.65	0.343
$6 \leqslant s < 7$	174	0.3413	170.65	3.35	0.066
$7 \leqslant s < 8$	66	0.1359	67.95	−1.95	0.056
$s \geqslant 8$	12	0.0228	11.40	0.60	0.032
Total	500	1.0000	500.00	0	1.400



On this occasion there are 6 categories and hence 5 degrees of freedom. The value of X^2 (1.400) is much less than the upper 5% point of the χ^2_5 distribution (11.07) and hence there is no significant evidence (at the 5% level) that the random number generator is working incorrectly.

Computer project

The repetitive nature of these calculations suggests the use of a spreadsheet in which successive columns hold the observed frequencies, the expected frequencies, their differences and the contributions to X^2. Use a spreadsheet to reproduce the working of the previous example.

Practical

Roll a die 30 times, recording the results. Now perform a goodness-of-fit test to determine whether there is significant evidence, at the 10% level, of any bias.

Assuming that your die is fair, what proportion of the time would a sample of 30 rolls give rise to a result that was 'significant at the 10% level'?

What proportion of individuals in your class obtained 'significant' results?

Project

In Statistics we frequently use phrases such as 'randomly chosen' or 'at random'. The object of this project is to determine whether people can really choose things 'at random'. If they can't, then tables of random numbers are needed!

Write the letters A B C D E *in a horizontal line on a sheet of paper. Then ask people to 'choose a letter at random'. Record their choice. After you have recorded the choices of at least 25 people, test the hypothesis that all letters are chosen with equal probability. Combine your results with those of the other members of your class. Most research suggests that people are biased towards the ends of lists, and particularly towards the left of a horizontal list. You might like to repeat the experiment with a vertical list.*

(i) Carry out a suitable test, at the 5% significance level, to determine whether the sample supports the manufacturer's claim that the bottles contain amounts which are normally distributed with mean 254.0 ml and standard deviation 2.4 ml. State your conclusions clearly.

(ii) The sample mean is 253.68 ml. Assuming that the population mean and standard deviation are as claimed in (i), calculate the probability that a random sample of size 40 will yield a sample mean of at least 253.68 ml. [UCLES]

13.3 Small expected frequencies

The distribution of X^2 is discrete – the χ^2 distribution is continuous and is simply a convenient approximation which becomes less accurate as the expected frequencies become smaller. The rule often stated for deciding whether the approximation may be used is:

'All expected frequencies must be equal to at least 5'.

If the original categories chosen lead to expected frequencies less than 5, then it will be necessary to combine categories together. This combination may be done on any sensible grounds, but should be done *without reference to the observed frequencies* so as to avoid biasing the results. With numerical data it is natural to combine adjacent categories: for example, we might replace the three categories '7', '8' and '9' by the single category '7–9'.

Note

- The rule given ('at least 5') errs on the safe side. Many researchers happily permit a small proportion of expected frequencies to be less than 5.

Example 5

A test of a random number generator is provided by studying the lengths of 'runs' of digits. The probability of a run of length k (i.e. that a randomly chosen digit is followed by exactly $k-1$ similar digits) is $0.9 \times 0.1^{k-1}$. This is a geometric distribution (see Section 5.4, p. 118).

A sequence of supposedly random numbers are generated, and the following results are obtained:

Length of run	1	2	3	4	5	6 or more
Frequency	8083	825	75	9	1	0

Use a 10% significance level to decide whether these results suggest that there is anything wrong with the random number generator.

The null hypothesis specifies that a run of length k has probability $0.9 \times 0.1^{k-1}$, with the alternative hypothesis stating that the null hypothesis is incorrect.

Run length	O_i	Probability	E_i	$O_i - E_i$	$\frac{(O_i - E_i)^2}{E_i}$
1	8083	0.900	8093.700	−10.700	0.014
2	825	0.090	809.370	15.630	0.302
3	75	0.009	80.937	−5.937	0.435
4	9		8.094		
5 4+	1 10	0.001	0.809 8.993	1.007	0.113
6+	0		0.090		
Total	8993	1.000	8993.000	0	0.864

The expected frequencies are shown in the table. The frequency for run lengths of 6 or more is obtained by subtraction. We then find that the last two expected frequencies are very small and we combine these with the previous category to form a category '4+'.

After combining these categories m becomes 4 and we use the χ^2_3 distribution. The upper 10% point of this distribution is 6.251 which greatly exceeds the observed value (0.864). There is no significant evidence for rejecting the null hypothesis that the random number generator is working correctly.

Exercises 13c

1 A random sample of 50 observations on the discrete random variable X is summarised below:

x	0	1	2	3
Frequency	12	16	20	2

Test, at the 1% significance level, the hypothesis that $X \sim B(3, 0.45)$.

2 A random sample of 100 observations on the discrete random variable X is summarised below:

x	0	1	2	3	4
Frequency	18	36	36	8	2

Test, at the 1% significance level, the hypothesis that $X \sim B(4, 0.3)$.

3 A random sample of 80 observations on the discrete random variable X is summarised below:

x	0	1	2	3	$\geqslant 4$
Frequency	24	30	17	5	4

Test, at the 5% significance level, the hypothesis that X has a Poisson distribution with mean equal to 1.2.

4 A random sample of 150 observations on the continuous variable X is summarised below:

x	<5	5–	10–	15–
Frequency	2	6	24	51

x	20–	25–	30–	$\geqslant 35$
Frequency	35	28	3	1

Test, at the 1% significance level, the hypothesis that $X \sim N(20, 36)$.

5 The null hypothesis is that the random variable X has a binomial distribution with $n = 6$ and $p = \frac{1}{3}$. A random sample of 60 observations gave the following results:

x	0	1	2	3	4	5	6
Frequency	8	16	18	15	3	0	0

Test the null hypothesis using a 5% significance level.

6 Timothy believes that his lucky coin is fair. To test his belief he tosses the coin and counts the numbers of tails between successive heads. If the coin is fair then $P(r \text{ tails}) = (\frac{1}{2})^{r+1}$ for $r = 0, 1, 2, \ldots$

Timothy's results are as follows:

r	0	1	2	3	4	$\geqslant 5$
Frequency	25	18	9	6	2	0

Determine whether there is significant evidence, at the 5% level, that the coin is biased.

7 A certain plant population has flowers of various colours whose proportions are given in the table below. A gardener has 80 of the plants, as specified in the table.

Colour	red	white	pink	orange	yellow
Population	0.4	0.3	0.2	0.05	0.05
Plants	24	28	24	3	1

Test, at the 5% level, the hypothesis that the gardener's plants may be regarded as a random sample from the population.

8 A botanist conjectures that the specimens of a particular plant are growing at random locations in a bog, with mean rate 4 plants per square metre. To test the conjecture she divides a randomly chosen region of the bog into non-overlapping regions of area 0.5 m^2 and counts the numbers of plants in each region. The results were as follows:

No. of plants	0	1	2	3	4	$\geqslant 5$
No. of regions	4	14	9	7	6	0

Test the conjecture using a 5% significance level.

9 A random number program is claimed to produce random integers in the range

00000 to 99999

inclusive. One way of testing this supposed randomness is to let X be the total number of 3s *and* 7s in a five-digit number that has been produced. For example, for 02037 the value of X is 2, and for 30703 it is 3. If the program is working properly, what would you predict the probability distribution of X to be?

A test run produces the following results:

Value of X	0	1	2	3	4	5
Frequency	1001	1302	624	175	23	0

Test the consistency of this sample with the predicted distribution of X, stating the basis of your calculations. **[SMP]**

10 A random variable X has a normal distribution with mean 35 and variance 100. The first table below shows the probability that the value of a single reading, x, lies in some particular interval. Copy and complete this table.

x	less than 10	10–	20–	30–
Probability	0.0062			0.3830

x	40–	50–	60 and above
Probability			

The second table shows the frequency distribution of the times, in seconds, required by 200 ten-year-old children to tie both their shoe laces.

Time	less than 10	10–	20–	30–
Frequency	8	11	40	59

Time	40–	50–	60 and above
Frequency	66	10	6

Perform a χ^2 goodness-of-fit test to show that there is evidence to suggest that the times taken by ten-year-old children to tie both their shoe laces do not follow a normal distribution with mean 35 seconds and standard deviation 10 seconds. **[JMB(P)]**

11 The heights (x) of one hundred police officers recruited to a police force in a particular year are summarised in the table below.

Height (cm)	Frequency
$x < 175$	2
$175 \leqslant x < 177$	15
$177 \leqslant x < 179$	29
$179 \leqslant x < 181$	25
$181 \leqslant x < 183$	12
$183 \leqslant x < 185$	10
$185 \leqslant x$	7

The population of police officers has a mean height of 180 cm with a standard deviation of 3 cm. Test the hypothesis that the distribution of heights is normal.

12 Pseudo-random numbers are generated, usually by computer, using mathematical algorithms. The numbers generated are supposed to mimic the properties of genuine (unpredictable) random numbers, but are called pseudo-random because the computer will usually generate the same sequence each time the computer is switched on.

Various tests for randomness are applied to a set of supposed pseudo-random numbers, for example:

(i) A set of digits should contain each of $0, 1, \ldots, 9$ with approximately equal frequency.

(ii) The number, K, of non-zero digits occurring between successive occurrences of a multiple of 3 should be an observation from a geometric distribution:

$$P(K = k) = \frac{2^k}{3^{k+1}}, \text{ for } k = 0, 1, \ldots$$

Using the X^2 goodness-of-fit test, perform these two tests of randomness on the following set of 80 digits (working along the rows).

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

13.4 Goodness of fit to prescribed distribution type

We now turn to cases where the null hypothesis states that the data 'has a particular named distribution', but does *not* specify all the parameters of the distribution. A typical example would be the hypothesis:

> H_0: the masses of a certain brand of digestive biscuit have a normal distribution.

The hypothesis does not specify *which* normal distribution, so we choose the most plausible one – this is the one having the same mean and variance as the observed data. Because of this deliberate matching we are imposing constraints on the expected frequencies. The value of X^2 will be a little smaller because of the better fit and this alters the value of ν, the degrees of freedom of the approximating χ^2 distribution. The general rule is:

$$\nu = m - 1 - k \tag{13.2}$$

where m is the number of different outcomes (after amalgamations to eliminate small expected frequencies) and k is the number of parameters estimated from the data. In the cases just considered in Sections 13.2 and 13.3, k was equal to 0.

10 Explain how to calculate the number of degrees of freedom associated with a χ^2-test for goodness of fit.

A study of the distribution of lichens found on stone walls in Derbyshire was carried out. As part of the study, 100 randomly chosen sections of wall, all of the same width and all of the same height, were examined. The number, X, of lichens in each section was recorded and the results are summarised in the table.

Number of lichens (x)	0	1	2	3	4	5	6	7
Number of sections	8	21	32	15	12	6	4	2

(i) Calculate the mean and variance of this sample, and state why the results might suggest that X has a Poisson distribution.

(ii) Perform a test, at the 5% significance level, to determine whether the data could be a sample from a Poisson distribution. [UCLES]

11 A shop that repairs television sets keeps a record of the number of sets brought in for repair each day. The numbers brought in during a random sample of 40 days were as follows.

4	0	0	0	2	1	1	0	0	0
0	1	1	0	3	0	0	0	1	0
4	0	0	0	0	0	2	0	1	0
0	0	0	1	1	1	0	2	0	0

Test, at the 5% significance level, the hypothesis that these numbers are observations from a Poisson distribution. [UCLES(P)]

12 Describe briefly how the number of degrees of freedom is calculated in a χ^2 goodness-of-fit test.

The following set of grouped data from 100 observations has mean 1.03. The data are thought to come from a normal distribution with variance 1 but unknown mean. Using an appropriate χ^2-distribution, test this hypothesis at the 1% significance level.

Lower value of grouping interval	$-\infty$	−2.0	−1.5	−1.0	−0.5
Number of observations	0	1	0	6	10

Lower value of grouping interval	0.0	0.5	1.0	1.5	2.0
Number of observations	12	15	23	16	13

Lower value of grouping interval	2.5	3.0	3.5
Number of observations	3	1	0

[UCLES]

13 A weaving mill sells lengths of cloth with a nominal length of 70 m. The customer measured 100 lengths and obtained the following frequency distribution:

Length (m)	Frequency
61–67	1
67–69	16
69–71	26
71–73	19
73–75	20
75–81	18

(a) Use a χ^2 test at the 5% significance level to show that the normal model is not an adequate model for the data.

(b) The contract provides for the mill to pay compensation to the customer for any lengths less than 67 m supplied. Comment on the distribution of the lengths of cloth in the light of this further information. [AEB 89]

13.5 Contingency tables

Often data are collected on several variables at a time. For example, a questionnaire will usually contain more than one question! A table that gives the frequencies for two or more variables simultaneously is called a **contingency table**. Here is an example which shows information on voting:

Introducing STATISTICS

Introducing Statistics has been revised to match the requirements of all the new A Level specifications. It covers in one volume all the statistics required by students taking single-subject Advanced Level Mathematics. It also provides the basis for a first course in statistics in Higher Education.

A clear text is supported by worked examples, exercises, examination questions, and suggestions for practical work. The book also includes biographies, chapter summaries, and statistical tables.

This is one of three texts for single-subject students, the others being:

Introducing Pure Mathematics (2nd edition) by Robert Smedley and Garry Wiseman
ISBN 0 19 914803 1

Introducing Mechanics by Brian Jefferson and Tony Beadsworth
ISBN 0 19 914710 8

For Further Mathematics, Oxford offers:

Further Pure Mathematics by Brian Gaulter and Mark Gaulter
ISBN 0 19 914735 3

Further Mechanics by Brian Jefferson and Tony Beadsworth
ISBN 0 19 914738 8

Understanding Statistics by Graham Upton and Ian Cook
ISBN 0 19 9143491 9

OXFORD
UNIVERSITY PRESS

Orders and enquiries to Customer Services:
tel. 01536 741068 fax 01536 454519

www.oup.com

ISBN 0-19-914801-5
9 780199 148011