

CAMBRIDGE

Mathematics Higher Level
Topic 7 – Option:
Statistics
and Probability

for the IB Diploma

**Paul Fannon, Vesna Kadelburg,
Ben Woolley and Stephen Ward**

Mathematics Higher Level
Topic 7 – Option:
Statistics
and Probability
for the IB Diploma

**Paul Fannon, Vesna Kadelburg,
Ben Woolley and Stephen Ward**



**CAMBRIDGE
UNIVERSITY PRESS**

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

www.cambridge.org
Information on this title: www.cambridge.org/9781107682269

© Cambridge University Press 2013

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 2013

Printed in Poland by Opolgraf

A catalogue record for this publication is available from the British Library

ISBN 978-1-107-68226-9 Paperback

Cover image: Thinkstock

Cambridge University Press has no responsibility for the persistence or
accuracy of URLs for external or third-party internet websites referred to in
this publication, and does not guarantee that any content on such websites is,
or will remain, accurate or appropriate. Information regarding prices, travel
timetables and other factual information given in this work is correct at
the time of first printing but Cambridge University Press does not guarantee
the accuracy of such information thereafter.

NOTICE TO TEACHERS

Worksheets and copies of them remain in the copyright of Cambridge University Press
and such copies may not be distributed or used in any way outside the
purchasing institution.

Contents

How to use this book	v
Acknowledgements	viii
Introduction	1
1 Combining random variables	2
1A Adding and multiplying all the data by a constant	2
1B Adding independent random variables	5
1C Expectation and variance of the sample mean and sample sum	9
1D Linear combinations of normal variables	12
1E The distribution of sums and averages of samples	16
2 More about statistical distributions	22
2A Geometric distribution	22
2B Negative binomial distribution	24
2C Probability generating functions	27
2D Using probability generating functions to find the distribution of the sum of discrete random variables	32
3 Cumulative distribution functions	38
3A Finding the cumulative probability function	38
3B Distributions of functions of a continuous random variable	43
4 Unbiased estimators and confidence intervals	48
4A Unbiased estimates of the mean and variance	48
4B Theory of unbiased estimators	51
4C Confidence interval for the population mean	55
4D The t-distribution	60
4E Confidence interval for a mean with unknown variance	63
5 Hypothesis testing	71
5A The principle of hypothesis testing	71
5B Hypothesis testing for a mean with known variance	78
5C Hypothesis testing for a mean with unknown variance	81
5D Paired samples	85
5E Errors in hypothesis testing	88

6 Bivariate distributions	97
6A Introduction to discrete bivariate distributions	97
6B Covariance and correlation	100
6C Linear regression	107
7 Summary and mixed examination practice	115
Answers	123
Appendix: Calculator skills sheets	131
A Finding probabilities in the t-distribution	
CASIO	132
TEXAS	133
B Finding t-scores given probabilities	
CASIO	134
TEXAS	135
C Confidence interval for the mean with unknown variance (from data)	
CASIO	136
TEXAS	137
D Confidence interval for the mean with unknown variance (from stats)	
CASIO	138
TEXAS	140
E Hypothesis test for the mean with unknown variance (from data)	
CASIO	141
TEXAS	143
F Hypothesis test for the mean with unknown variance (from stats)	
CASIO	144
TEXAS	146
G Confidence interval for the mean with known variance (from data)	
CASIO	148
TEXAS	149
H Confidence interval for the mean with known variance (from stats)	
CASIO	150
TEXAS	152
I Hypothesis test for the mean with known variance (from stats)	
CASIO	154
TEXAS	155
J Finding the correlation coefficient and the equation of the regression line	
CASIO	157
TEXAS	158
Glossary	159
Index	163

How to use this book

Structure of the book

This book covers all the material for Topic 7 (Statistics and Probability Option) of the Higher Level Mathematics syllabus for the International Baccalaureate course. It assumes familiarity with the core Higher Level material (Syllabus Topics 1 to 6), in particular Topic 5 (Core Statistics and Probability) and Topic 6 (Core Calculus). We have tried to include in the main text only the material that will be examinable. There are many interesting applications and ideas that go beyond the syllabus and we have tried to highlight some of these in the ‘From another perspective’ and ‘Research explorer’ boxes.

The book is split into seven chapters. Chapter 1 deals with combinations of random variables and requires familiarity with Binomial, Poisson and Normal distributions; we recommend that it is covered first. Chapters 2 and 3 extend your knowledge of random variables and probability distributions, and use differentiation and integration; they can be studied in either order. Chapters 4 and 5 develop the main theme of this Option: using samples to make inferences about a population. They require understanding of the material from chapter 1. Chapter 6, on bivariate distributions, is largely independent of the others, although it requires understanding of the concept of a hypothesis test. Chapter 7 contains a summary of all the topics and further examination practice, with many of the questions mixing several topics – a favourite trick in IB examinations.

Each chapter starts with a list of learning objectives to give you an idea about what the chapter contains. There is also an introductory problem, at the start of the topic, that illustrates what you will be able to do after you have completed the topic. You should not expect to be able to solve the problem, but you may want to think about possible strategies and what sort of new facts and methods would help you. The solution to the introductory problem is provided at the end of the topic, at the start of chapter 7.

Key point boxes

The most important ideas and formulae are emphasised in the ‘KEY POINT’ boxes. When the formulae are given in the Formula booklet, there will be an icon: ; if this icon is not present, then the formulae are **not** in the Formula booklet and you may need to learn them or at least know how to derive them.

Worked examples

Each worked example is split into two columns. On the right is what you should write down. Sometimes the example might include more detail than you strictly need, but it is designed to give you an idea of what is required to score full method marks in examinations. However, mathematics is about much more than examinations and remembering methods. So, on the left of the worked examples are notes that describe the thought processes and suggest which route you should use to tackle the question. We hope that these will help you with any exercise questions that differ from the worked examples. It is very deliberate that some of the questions require you to do more than repeat the methods in the worked examples. Mathematics is about thinking!

Signposts

There are several boxes that appear throughout the book.

Theory of knowledge issues



Every lesson is a Theory of knowledge lesson, but sometimes the links may not be obvious. Mathematics is frequently used as an example of certainty and truth, but this is often not the case. In these boxes we will try to highlight some of the weaknesses and ambiguities in mathematics as well as showing how mathematics links to other areas of knowledge.

From another perspective



The International Baccalaureate® encourages looking at things in different ways. As well as highlighting some international differences between mathematicians these boxes also look at other perspectives on the mathematics we are covering: historical, pragmatic and cultural.

Research explorer



As part of your course, you will be asked to write a report on a mathematical topic of your choice. It is sometimes difficult to know which topics are suitable as a basis for such reports, and so we have tried to show where a topic can act as a jumping-off point for further work. This can also give you ideas for an Extended essay. There is a lot of great mathematics out there!

Exam hint

EXAM HINT

Although we would encourage you to think of mathematics as more than just learning in order to pass an examination, there are some common errors it is useful for you to be aware of. If there is a common pitfall we will try to highlight it in these boxes. We also point out where graphical calculators can be used effectively to simplify a question or speed up your work, often referring to the relevant calculator skills sheet in the back of the book.

Fast forward / rewind



Mathematics is all about making links. You might be interested to see how something you have just learned will be used elsewhere in the course, or you may need to go back and remind yourself of a previous topic. These boxes indicate connections with other sections of the book to help you find your way around.

How to use the questions

Calculator icon



You will be allowed to use a graphical calculator in the final examination paper for this Option. Some questions can be done in a particularly clever way by using one of the graphical calculator functions, or cannot be realistically done without. These questions are marked with a calculator symbol.

The colour-coding

The questions are colour-coded to distinguish between the levels.

Black questions are drill questions. They help you practise the methods described in the book, but they are usually not structured like the questions in the examination. This does not mean they are easy, some of them are quite tough.

Each differently numbered drill question tests a different skill. Lettered subparts of a question are of increasing difficulty. Within each lettered part there may be multiple roman-numeral parts ((i), (ii), ...), all of which are of a similar difficulty. Unless you want to do lots of practice we would recommend that you only do one roman-numeral part and then check your answer. If you have made a mistake then you may want to think about what went wrong before you try any more. Otherwise move on to the next lettered part.

- █ Green questions are examination-style questions which should be accessible to students on the path to getting a grade 3 or 4.
- █ Blue questions are harder examination-style questions. If you are aiming for a grade 5 or 6 you should be able to make significant progress through most of these.
- █ Red questions are at the very top end of difficulty in the examinations. If you can do these then you are likely to be on course for a grade 7.
- █ Gold questions are a type that are *not* set in the examination, but are designed to provoke thinking and discussion in order to help you to a better understanding of a particular concept.

At the end of each chapter you will see longer questions typical of the second section of International Baccalaureate® examinations. These follow the same colour-coding scheme.

Of course, these are just **guidelines**. If you are aiming for a grade 6, do not be surprised if you find a green question you cannot do. People are never equally good at all areas of the syllabus. Equally, if you can do all the red questions that does not guarantee you will get a grade 7; after all, in the examination you have to deal with time pressure and examination stress!

These questions are graded relative to our experience of the final examination, so when you first start the course you will find all the questions relatively hard, but by the end of the course they should seem more straightforward. Do not get intimidated!

We hope you find the Statistics and Probability Option an interesting and enriching course. You might also find it quite challenging, but do not get intimidated, frequently topics only make sense after lots of revision and practice. Persevere and you will succeed.

The author team.

Acknowledgements

The authors and publishers are grateful for the permissions granted to reproduce materials in either the original or adapted form. While every effort has been made, it has not always been possible to identify the sources of all the materials used, or to trace all copyright holders.

If any omissions are brought to our notice, we will be happy to include the appropriate acknowledgements on reprinting.

IB exam questions © International Baccalaureate Organization. We gratefully acknowledge permission to reproduce International Baccalaureate Organization intellectual property.

Cover image: Thinkstock

Diagrams are created by Ben Woolley.

TI-83 fonts are reproduced on the calculator skills sheets with permission of Texas Instruments Incorporated.

Casio fonts are reproduced on the calculator skills sheets with permission of Casio Electronics Company Ltd (to access downloadable Casio resources go to www.casio.co.uk/education and <http://edu.casio.com/dll>).

Introduction

In this Option you will learn:

- how to combine information from more than one random variable
- how to predict the distribution of the mean of a sample
- about more distributions used to model common situations
- about the probability generating function: an algebraic tool for combining probability distributions
- about the cumulative distribution: the probability of a variable being less than a particular value
- how to estimate information about the population from a sample
- about hypothesis testing: how to decide if new information is significant
- how to make predictions based upon data.

Introductory problem

A school claims that their average International Baccalaureate (IB) score is 34 points. In a sample of four students the scores are 31, 31, 30 and 35 points. Does this suggest that the school was exaggerating?

As part of the core syllabus, you should have used statistics to find information about a population using a sample, and you should have used probability to predict the average and standard deviation of a given distribution. In this topic we extend both of these ideas to answer a very important question: does any new information gathered show a significant change, or could it just have happened by chance?

The statistics option is examined in a separate, one-hour paper. There will be approximately five extended-response questions based mainly upon the material covered in the statistics option, although any aspect of the core may also be included.

In this chapter you will learn:

- how multiplying all of your data by a constant or adding a constant changes the mean and the variance
- how adding or multiplying together two independent random variables changes the mean and the variance
- how we can apply these ideas to making predictions about the average or the sum of a sample
- about the distribution of linear combinations of normal variables
- about the distribution of the sum or average of lots of observations from any distribution.

1 Combining random variables

If you know the average height of a brick, then it is fairly easy to guess the average height of two bricks, or the average height of half of a brick. What is less obvious is the variation of these heights.

Even if we can predict the mean and the variance of this random variable this is not enough to find the probability of it taking a particular value. To do this, we also need to know the *distribution* of the random variable. There are some special cases where it is possible to find the distribution of the random variable, but in most cases we meet the enormous significance of the normal distribution; if the sample is large enough, the sample average will (nearly) always follow a normal distribution.

1A Adding and multiplying all the data by a constant

The average height of the students in a class is 1.75 m and their standard deviation is 0.1 m. If they all then stood on their 0.5 m tall chairs then the new average height would be 2.25 m, but the range, and any other measure of variability, would not change, and so the standard deviation would still be 0.1 m. If we add a constant to all the variables in a distribution, we add the same constant on to the expectation, but the variance does not change:

$$E(X + c) = E(X) + c$$

$$\text{Var}(X + c) = \text{Var}(X)$$

If, instead, each student were given a magical growing potion that doubled their heights, the new average height would be 3.5 m, and in this case the range (and any similar measure of variability) would also double, so the new standard deviation would be 0.2 m. This means that their variance would change from 0.01 m^2 to 0.04 m^2 .

If we multiply a random variable by a constant, we multiply the expectation by the constant and multiply the variance by the constant squared:

$$E(aX) = aE(X)$$

$$\text{Var}(aX) = a^2 \text{Var}(X)$$

These ideas can be combined together:

KEY POINT 1.1

$$E(aX + c) = aE(X) + c$$

$$\text{Var}(aX + c) = a^2 \text{Var}(X)$$

EXAM HINT

It is important to know that this only works for the structure $aX + c$ which is called a linear function. So, for example, $E(X^2)$ cannot be simplified to $[E(X)]^2$ and $E(|X|)$ is not equivalent to $|E(X)|$.

Worked example 1.1

A piece of pipe with average length 80 cm and standard deviation 2 cm is cut from a 100 cm length of water pipe. The leftover piece is used as a short pipe. Find the mean and standard deviation of the length of the short pipe.

Define your variables

L = crv 'length of long pipe'
 S = crv 'length of short pipe'

Write an equation to connect the variables

$$S = 100 - L$$

Apply expectation algebra

$$\begin{aligned} E(S) &= E(100 - L) \\ &= 100 - E(L) \\ &= 100 - 80 = 20 \end{aligned}$$

So the mean of S is 20 cm

$$\begin{aligned} \text{Var}(S) &= \text{Var}(100 - L) \\ &= (-1)^2 \text{Var}(L) \\ &= \text{Var}(L) = 4 \text{ cm}^2 \end{aligned}$$

So the standard deviation of S is also 2 cm.

EXAM HINT

Even if the coefficients are negative, you will always get a positive variance (since square numbers are always positive). If you find you have a negative variance, something has gone wrong!

The result regarding $E(aX + b)$ stated in Key point 1.1 represents a more general result about the expectation of a function of a random variable:

KEY POINT 1.2

For a discrete random variable:

$$E(g(X)) = \sum_i g(x_i)p_i$$

For a continuous random variable with probability density function $f(x)$:

$$E(g(X)) = \int g(x)f(x)dx$$

Worked example 1.2

The continuous random variable X has probability density e^x for $0 < x < \ln 2$. The random variable Y is related to X by the function $Y = e^{-2X}$. Find $E(Y)$.

Use the formula for the expectation of a function of a variable

Use the laws of exponents

$$E(e^{-2x}) = \int_0^{\ln 2} e^{-2x} \times e^x dx$$

$$= \int_0^{\ln 2} e^{-x} dx$$

$$= [-e^{-x}]_0^{\ln 2}$$

$$= -e^{-\ln 2} - (-e^0)$$

$$= -\frac{1}{2} + 1$$

$$= \frac{1}{2}$$

Exercise 1A

1. If $E(X) = 4$ find:

- | | |
|-------------------------------------|-----------------------------------|
| (a) (i) $E(3X)$ | (ii) $E(6X)$ |
| (b) (i) $E\left(\frac{X}{2}\right)$ | (ii) $E\left(\frac{3X}{4}\right)$ |
| (c) (i) $E(-X)$ | (ii) $E(-4X)$ |
| (d) (i) $E(X+5)$ | (ii) $E(X-3)$ |
| (e) (i) $E(5-2X)$ | (ii) $E(3X+1)$ |

2. If $\text{Var}(X) = 6$ find:

- | | |
|--|--|
| (a) (i) $\text{Var}(3X)$ | (ii) $\text{Var}(6X)$ |
| (b) (i) $\text{Var}\left(\frac{X}{2}\right)$ | (ii) $\text{Var}\left(\frac{3X}{4}\right)$ |
| (c) (i) $\text{Var}(-X)$ | (ii) $\text{Var}(-4X)$ |
| (d) (i) $\text{Var}(X+5)$ | (ii) $\text{Var}(X-3)$ |
| (e) (i) $\text{Var}(5-2X)$ | (ii) $\text{Var}(3X+1)$ |

3. The probability density function of the continuous random variable Z is kz for $1 < z \leq 3$.

- Find the value of k .
- Find $E(Z)$.
- Find $E(6Z + 5)$.
- Find the exact value of $E\left(\frac{1}{1+Z^2}\right)$.

[10 marks]

1B Adding independent random variables

A tennis racquet is made by adding together two components; the handle and the head. If both components have their own distribution of length and they are combined together randomly then we have formed a new random variable: the length of the racquet. It is not surprising that the average length of the whole racquet is the sum of the average lengths of the parts, but with a little thought we can reason that the standard deviation will be less than the sum of the standard deviation of the parts. To get either extremely long or extremely short tennis racquets we must have extremes in the same direction for both the handle and the head. This is not very likely. It is more likely that either both are close to average or an extreme value is paired with an average value or an extreme value in one direction is balanced by another.

KEY POINT 1.3

Linear Combinations

$$E(a_1X_1 \pm a_2X_2) = a_1E(X_1) \pm a_2E(X_2)$$

$$\text{Var}(a_1X_1 \pm a_2X_2) = a_1^2\text{Var}(X_1) + a_2^2\text{Var}(X_2)$$

The result for variance is only true if X and Y are independent.

There is a similar result for the product of two independent random variables:

KEY POINT 1.4

If X and Y are independent random variables then:

$$E(XY) = E(X)E(Y)$$



We could write the whole of statistics only using standard deviation, without referring to variance at all, where $\sigma(aX + bY + c) = \sqrt{a^2\sigma^2(X) + b^2\sigma^2(Y)}$. However, as you can see, the concept of standard deviation squared occurs very naturally. Is this a sufficient justification for the concept of variance?

It is not immediately obvious that if $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ then the standard deviation of $(X + Y)$ will always be less than the standard deviation of X plus the standard deviation of Y . This is an example of one of many interesting inequalities in statistics. Another is that $E(X^2) \geq [E(X)]^2$ which ensures that variance is always positive. If you are interested in proving these types of inequalities you might like to look at the Cauchy-Schwarz inequality.



EXAM HINT

Notice in particular that, if X and Y are independent:

$$\begin{aligned}\text{Var}(X - Y) &= (1^2) \times \text{Var}(X) + (-1)^2 \times \text{Var}(Y) \\ &= \text{Var}(X) + \text{Var}(Y)\end{aligned}$$

The result extends to more than two variables.

Worked example 1.3

The mean thickness of the base of a burger bun is 1.4 cm with variance 0.02 cm².

The mean thickness of a burger is 3.0 cm with variance 0.14 cm².

The mean thickness of the top of the burger bun is 2.2 cm with variance 0.2 cm².

Find the mean and standard deviation of the total height of the whole burger and bun, assuming that the thickness of each part is independent.

Define your variables

$X = \text{crv 'Thickness of base'}$
 $Y = \text{crv 'Thickness of burger'}$
 $Z = \text{crv 'Thickness of top'}$
 $T = \text{crv 'Total thickness'}$

Write an equation to connect the variables

$$T = X + Y + Z$$

Apply expectation algebra

$$\begin{aligned}E(T) &= E(X + Y + Z) \\ &= E(X) + E(Y) + E(Z) = 6.6 \text{ cm}\end{aligned}$$

So the mean of T is 6.6 cm

$$\begin{aligned}\text{Var}(T) &= \text{Var}(X + Y + Z) \\ &= \text{Var}(X) + \text{Var}(Y) + \text{Var}(Z) \\ &= 0.36 \text{ cm}^2\end{aligned}$$

So standard deviation of T is 0.6 cm

X and Y have to be independent (see Key point 1.3) but this does not mean that they have to be drawn from different populations. They could be two different observations of the same population, for example the heights of two different people added together. This is a different variable from the height of

one person doubled. We will use a subscript to emphasise when there are repeated observations from the same population:

$X_1 + X_2$ means adding together two different observations of X

$2X$ means observing X once and doubling the result.

The expectation of both of these combinations is the same, $2E(X)$, but the variance is different.

From Key point 1.3:

$$\begin{aligned}\text{Var}(X_1 + X_2) &= \text{Var}(X_1) + \text{Var}(X_2) \\ &= 2\text{Var}(X)\end{aligned}$$

From Key point 1.1:

$$\text{Var}(2X) = 4\text{Var}(X)$$

So the variability of a single observation doubled is greater than the variability of two independent observations added together. This is consistent with the earlier argument about the possibility of independent observations ‘cancelling out’ extreme values.

Worked example 1.4

In an office, the mean mass of the men is 84 kg and standard deviation is 11 kg. The mean mass of women in the office is 64 kg and the standard deviation is 6 kg. The women think that if four of them are picked at random their total mass will be less than three times the mass of a randomly selected man. Find the mean and standard deviation of the difference between the sums of four women’s masses and three times the mass of a man, assuming that all these people are chosen independently.

Define your variables

$X = \text{crv ‘Mass of a man’}$

$Y = \text{crv ‘Mass of a woman’}$

$D = \text{crv ‘Difference between the mass of 4 women and 3 lots of 1 man’}$

Write an equation to connect the variables

$$D = Y_1 + Y_2 + Y_3 + Y_4 - 3X$$

Apply expectation algebra

$$\begin{aligned}E(D) &= E(Y_1) + E(Y_2) + E(Y_3) + E(Y_4) - 3E(X) \\ &= 4 \text{ kg}\end{aligned}$$

$$\begin{aligned}\text{Var}(D) &= \text{Var}(Y_1) + \text{Var}(Y_2) + \text{Var}(Y_3) + \text{Var}(Y_4) + (-3)^2 \times \text{Var}(X) \\ &= 1233 \text{ kg}^2\end{aligned}$$

So the standard deviation of D is 35.1 kg

Finding the mean and variance of D is not very useful unless you also know the distribution

► of D . In Sections 1D and 1E you will see that this can be done in certain circumstances. ►

We can then go on to calculate probabilities of different values of D .

Exercise 1B

1. Let X and Y be two independent variables with $E(X) = -1$, $\text{Var}(X) = 2$, $E(Y) = 4$ and $\text{Var}(Y) = 4$. Find the expectation and variance of:

- (a) (i) $X - Y$ (ii) $X + Y$
 (b) (i) $3X + 2Y$ (ii) $2X - 4Y$
 (c) (i) $\frac{X - 3Y + 1}{5}$ (ii) $\frac{X + 2Y - 2}{3}$

Denote by X_i, Y_i independent observations of X and Y .

- (d) (i) $X_1 + X_2 + X_3$ (ii) $Y_1 + Y_2$
 (e) (i) $X_1 - X_2 - 2Y$ (ii) $3X - (Y_1 + Y_2 - Y_3)$

- 2.** If X is the random variable ‘mass of a gerbil’ explain the difference between $2X$ and $X_1 + X_2$.

4. The average mass of a man in an office is 85 kg with standard deviation 12 kg. The average mass of a woman in the office is 68 kg with standard deviation 8 kg. The empty lift has a mass of 500 kg. What is the expectation and standard deviation of the total mass of the lift when 3 women and 4 men are inside?

[6 marks]

5. A weighted die has mean outcome 4 with standard deviation 1. Brian rolls the die once and doubles the outcome. Camilla rolls the die twice and adds her results together. What is the expected mean and standard deviation of the difference between their scores? [7 marks]

[7 marks]

6. Exam scores at a large school have mean 62 and standard deviation 28. Two students are selected at random. Find the expected mean and standard deviation of the difference between their exam scores. [6 marks]

[6 marks]

7. Adrian cycles to school with a mean time of 20 minutes and a standard deviation of 5 minutes. Pamela walks to school with a mean time of 30 minutes and a standard deviation of 2 minutes. They each calculate the total time it takes them to get to school over a five-day week. What is the expected mean and standard deviation of the difference in the total weekly journey times, assuming journey times are independent? [7 marks]

[7 marks]

8. In this question the discrete random variable X has the following probability distribution:

x	1	2	3	4
$P(X = x)$	0.1	0.5	0.2	k

- (a) Find the value of k .
- (b) Find the expectation and variance of X .
- (c) The random variable Y is given by $Y = 6 - X$.
Find the expectation and the variance of Y .
- (d) Find $E(XY)$ and explain why the formula $E(XY) = E(X)E(Y)$ is not applicable to these two variables.
- (e) The discrete random variable Z has the following distribution, independent of X :

z	1	2
$P(Z = z)$	p	$1-p$

If $E(XZ) = \frac{35}{8}$ find the value of p .

[14 marks]

1C Expectation and variance of the sample mean and sample sum

When calculating the mean of a sample of size n of the variable X we have to add up n independent observations of X then divide by n . We give this **sample mean** the symbol \bar{X} and it is itself a random variable (as it might change each time it is observed).

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

$$= \frac{1}{n} X_1 + \frac{1}{n} X_2 + \dots + \frac{1}{n} X_n$$

This is a linear combination of independent observations of X , so we can apply the rules of the previous section to get the following very important results:

KEY POINT 1.5

$$E(\bar{X}) = E(X)$$

$$\text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n}$$

The first of these results seems very obvious; the average of a sample is, on average, the average of the original variable, but you will see in chapter 4 that this is not the case for all sample statistics.

The result actually goes further than that; it contains what economists call 'The law of diminishing returns'. The standard deviation of the mean is proportional to $\frac{1}{\sqrt{n}}$, so going from a sample of 1 to a sample of 20 has a much bigger impact than going from a sample of 101 to a sample of 120.



The second result demonstrates why means are so important; their standard deviation (which can be thought of as a measure of the error caused by randomness) is smaller than the standard deviation of a single observation. This proves mathematically what you probably already knew instinctively, that finding an average of several results produces a more reliable outcome than just looking at one result.

Worked example 1.5

Prove that if \bar{X} is the average of n independent observations of X then $\text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n}$.

Write \bar{X} in terms of X_i

Apply expectation algebra

Since X_1, X_2, \dots are all observations of X

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

$$= \frac{1}{n}(X_1 + X_2 + \dots + X_n)$$

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right)$$

$$= \frac{1}{n^2} \text{Var}(X_1 + X_2 + \dots + X_n)$$

$$= \frac{1}{n^2} (\text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n))$$

$$= \frac{1}{n^2} \left(\underbrace{\text{Var}(X) + \text{Var}(X) + \dots + \text{Var}(X)}_{n \text{ times}} \right)$$

$$= \frac{1}{n^2} (n \text{Var}(X))$$

$$= \frac{\text{Var}(X)}{n}$$

We can apply similar ideas to the sample sum.

KEY POINT 1.6

For the sample sum:

$$\text{E}\left(\sum_{i=1}^n X_i\right) = n\text{E}(X) \text{ and } \text{Var}\left(\sum_{i=1}^n X_i\right) = n\text{Var}(X)$$

Exercise 1C

1. A sample is obtained from n independent observations of a random variable X . Find the expected value and the variance of the sample mean in the following situations:
 - (a) (i) $E(X) = 5$, $\text{Var}(X) = 1.2$, $n = 7$
(ii) $E(X) = 6$, $\text{Var}(X) = 2.5$, $n = 12$
 - (b) (i) $E(X) = -4.7$, $\text{Var}(X) = 0.8$, $n = 20$
(ii) $E(X) = -15.1$, $\text{Var}(X) = 0.7$, $n = 15$
 - (c) (i) $X \sim N(12, 3^2)$, $n = 10$
(ii) $X \sim N(8, 0.6^2)$, $n = 14$
 - (d) (i) $X \sim N(21, 6.25)$, $n = 7$
(ii) $X \sim N(14, 0.64)$, $n = 15$
 - (e) (i) $X \sim B(6, 0.5)$, $n = 10$
(ii) $X \sim B(12, 0.3)$, $n = 8$
 - (f) (i) $X \sim Po(6.5)$, $n = 20$
(ii) $X \sim Po(8.2)$, $n = 15$
2. Find the expected value and the variance of the total of the samples from the previous question.
3. Eggs are packed in boxes of 12. The mass of the box is 50 g. The mass of one egg has mean 12.4 g and standard deviation 1.2 g. Find the mean and the standard deviation of the mass of a box of eggs. [4 marks]
4. A machine produces chocolate bars so that the mean mass of a bar is 102 g and the standard deviation is 8.6 g. As a part of the quality control process, a sample of 20 chocolate bars is taken and the mean mass is calculated. Find the expectation and variance of the sample mean of these 20 chocolate bars. [5 marks]
5. Prove that $\text{Var}\left(\sum_{i=1}^n X_i\right) = n\text{Var}(X)$. [4 marks]
6. The standard deviation of the mean mass of a sample of 2 aubergines is 20 g smaller than the standard deviation in the mass of a single aubergine. Find the standard deviation of the mass of an aubergine. [5 marks]
7. A random variable X takes values 0 and 1 with probability $\frac{1}{4}$ and $\frac{3}{4}$, respectively.
 - (a) Calculate $E(X)$ and $\text{Var}(X)$.

A sample of three observations of X is taken.

- (b) List all possible samples of size 3 and calculate the mean of each.
(c) Hence complete the probability distribution table for the sample mean, \bar{X} .

\bar{x}	0	$\frac{1}{3}$	$\frac{2}{3}$	1
$P(\bar{X} = \bar{x})$	$\frac{1}{64}$			

- (d) Show that $E(\bar{X}) = E(X)$ and $\text{Var}(\bar{X}) = \frac{\text{Var}(X)}{3}$. [14 marks]

8. A laptop manufacturer believes that the battery life of the computers follows a normal distribution with mean 4.8 hours and variance 1.7 hours². They wish to take a sample to estimate the mean battery life. If the standard deviation of the sample mean is to be less than 0.3 hours, what is the minimum sample size needed? [5 marks]
9. When the sample size is increased by 80, the standard deviation of the sample mean decreases to a third of its original size. Find the original sample size. [4 marks]

1D Linear combinations of normal variables

Although the proof is beyond the scope of this course, it turns out that any *linear combination of normal variables* will also follow a normal distribution. We can use the methods of Section C to find out the parameters of this distribution.

KEY POINT 1.7

If X and Y are random variables following a normal distribution and $Z = aX + bY + c$ then Z also follows a normal distribution.

Worked example 1.6

If $X \sim N(12, 15)$, $Y \sim N(1, 18)$ and $Z = X + 2Y + 3$ find $P(Z > 20)$.

Use expectation algebra

$$E(Z) = E(X) + 2 \times E(Y) + 3 = 17$$
$$\text{Var}(Z) = \text{Var}(X) + 2^2 \times \text{Var}(Y) = 87$$

State distribution of Z

$$Z \sim N(17, 87)$$
$$P(Z > 20) = 0.626 \text{ (from GDC)}$$

Worked example 1.7

If $X \sim N(15, 12^2)$ and four independent observations of X are made find $P(\bar{X} < 14)$.

Express \bar{X} in terms of observations of X

Use expectation algebra

State distribution of \bar{X}

$$\bar{X} = \frac{X_1 + X_2 + X_3 + X_4}{4}$$

$$E(\bar{X}) = \frac{E(X)}{4} = 15$$

$$\text{Var}(\bar{X}) = \frac{\text{Var}(X)}{4} = 36$$

$$\bar{X} \sim N(15, 36)$$

$$P(\bar{X} < 14) = 0.434 \text{ (from GDC)}$$

The Poisson distribution is scaleable. If the number of butterflies on a flower in 10 minutes follows a Poisson distribution with mean (expectation) m , then the number of butterflies on a flower in 20 minutes follows a Poisson distribution with mean $2m$ and so on. We can interpret this as meaning that the sum of two Poisson variables is also Poisson. However, this only applies to sums of Poisson distributions, not differences or multiples or linear combinations.

Exercise 1D

1. If $X \sim N(12, 16)$ and $Y \sim N(8, 25)$, find:
 - (a) (i) $P(X - Y > -2)$ (ii) $P(X + Y < 24)$
 - (b) (i) $P(3X + 2Y > 50)$ (ii) $P(2X - 3Y > -2)$
 - (c) (i) $P(X > 2Y)$ (ii) $P(2X < 3Y)$
 - (d) (i) $P(X > 2Y - 2)$ (ii) $P(3X + 1 < 5Y)$
 - (e) (i) $P(X_1 + X_2 > 2X_3 + 1)$ (ii) $P(X_1 + Y_1 + Y_2 < X_2 + 12)$
 - (f) (i) $P(\bar{X} > 13)$ where \bar{X} is the average of 12 observations of X
(ii) $P(\bar{Y} < 6)$ where \bar{Y} is the average of 9 observations of Y
2. An airline has found that the mass of their passengers follows a normal distribution with mean 82.2 kg and variance 10.7 kg^2 . The mass of their hand luggage follows a normal distribution with mean 9.1 kg and variance 5.6 kg^2 .
 - (a) State the distribution of the total mass of a passenger and their hand luggage and find any necessary parameters.
 - (b) What is the probability that the total mass of a passenger and their luggage exceeds 100 kg? [5 marks]

EXAM HINT

Make sure you do not confuse the standard deviation and the variance!

- 3.** Evidence suggests that the times Aaron takes to run 100 m are normally distributed with mean 13.1 s and standard deviation 0.4 s. The times Bashir takes to run 100 m are normally distributed with mean 12.8 s and standard deviation 0.6 s.
- Find the mean and standard deviation of the difference (Aaron – Bashir) between Aaron's and Bashir's times.
 - Find the probability that Aaron finishes a 100 m race before Bashir.
 - What is the probability that Bashir beats Aaron by more than 1 second? [7 marks]
- 4.** A machine produces metal rods so that their length follows a normal distribution with mean 65 cm and variance 0.03 cm^2 . The rods are checked in batches of six, and a batch is rejected if the mean length is less than 64.8 cm or more than 65.3 cm.
- Find the mean and the variance of the mean of a random sample of six rods.
 - Hence find the probability that a batch is rejected. [5 marks]
- 5.** The lengths of pipes produced by a machine is normally distributed with mean 40 cm and standard deviation 3 cm.
- What is the probability that a randomly chosen pipe has a length of 42 cm or more?
 - What is the probability that the average length of a randomly chosen set of 10 pipes of this type is 42 cm or more?
- [6 marks]
- 6.** The masses, $X \text{ kg}$, of male birds of a certain species are normally distributed with mean 4.6 kg and standard deviation 0.25 kg. The masses, $Y \text{ kg}$, of female birds of this species are normally distributed with mean 2.5 kg and standard deviation 0.2 kg.
- Find the mean and variance of $2Y - X$.
 - Find the probability that the mass of a randomly chosen male bird is more than twice the mass of a randomly chosen female bird.
 - Find the probability that the total mass of three male birds and 4 female birds (chosen independently) exceeds 25 kg.
- [11 marks]
- 7.** A shop sells apples and pears. The masses, in grams, of the apples may be assumed to have a $N(180, 12^2)$ distribution and the masses of the pears, in grams, may be assumed to have a $N(100, 10^2)$ distribution.
- Find the probability that the mass of a randomly chosen apple is more than double the mass of a randomly chosen pear.
 - A shopper buys 2 apples and a pear. Find the probability that the total mass is greater than 500 g. [10 marks]

- 8.** The length of a cornsnake is normally distributed with mean 1.2 m. The probability that a randomly selected sample of 5 cornsnakes having an average of above 1.4 m is 5%. Find the standard deviation of the length of a cornsnake. [6 marks]
- 9.** (a) In a test, boys have scores which follow the distribution $N(50, 25)$. Girls' scores follow $N(60, 16)$. What is the probability that a randomly chosen boy and a randomly chosen girl differ in score by less than 5?
 (b) What is the probability that a randomly chosen boy scores less than three quarters of the mark of a randomly chosen girl? [10 marks]
- 10.** The daily rainfall in Algebraville follows a normal distribution with mean μ mm and standard deviation σ mm. The rainfall each day is independent of the rainfall on other days.
 On a randomly chosen day, there is a probability of 0.1 that the rainfall is greater than 8 mm.
 In a randomly chosen 7-day week, there is a probability of 0.05 that the *mean* daily rainfall is less than 7 mm.
 Find the value of μ and of σ . [7 marks]
- 11.** Anu uses public transport to go to school each morning. The time she waits each morning for the transport is normally distributed with a mean of 12 minutes and a standard deviation of 4 minutes.
- On a specific morning, what is the probability that Anu waits more than 20 minutes?
 - During a particular week (Monday to Friday), what is the probability that
 - her total morning waiting time does not exceed 70 minutes?
 - she waits less than 10 minutes on exactly 2 mornings of the week?
 - her average morning waiting time is more than 10 minutes?
 - Given that the total morning waiting time for the first four days is 50 minutes, find the probability that the average for the week is over 12 minutes.
 - Given that Anu's average morning waiting time in a week is over 14 minutes, find the probability that it is less than 15 minutes.

[20 marks]

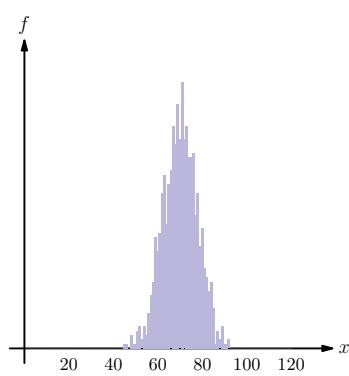
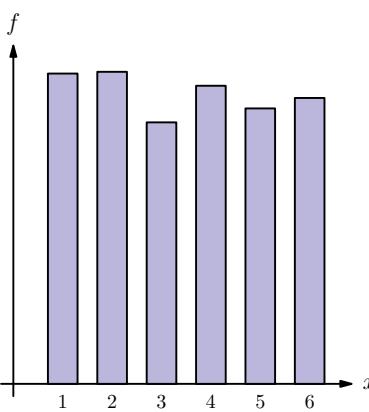
1E The distribution of sums and averages of samples

In this section we shall look at how to find the distribution of the sample mean or the sample total, even if we do not know the original distribution.

The graph alongside shows 1000 observations of the roll of a fair die.

It seems to follow a uniform distribution quite well, as we would expect.

However, if we look at the sum of 2 dice 1000 times the distribution looks quite different.



There are many other distributions which have a similar shape, such as the Cauchy distribution. To show that these sums form a normal distribution we need to use moment generating functions, which are well beyond this course.



The sum of 20 dice is starting to form a more familiar shape. The sum seems to form a normal distribution. This is more than a coincidence. If we sum enough independent observations of any random variable, the result will follow a normal distribution. This result is called the **Central Limit Theorem** or CLT. We generally take 30 to be a sufficiently large sample size to apply the CLT.

As we saw in Section 1D, if a variable is normally distributed then a multiple of that variable will also be normally distributed.

Since $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ it follows that the mean of a sufficiently large sample is also normally distributed. Using Key point 1.5 where $E(\bar{X}) = E(X)$ and $\text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n}$, we can predict which normal distribution is being followed:

KEY POINT 1.8

Central Limit Theorem

For any distribution if $E(X) = \mu$, $\text{Var}(X) = \sigma^2$ and $n \geq 30$, then the approximate distributions of the sum and the mean are given by:

$$\begin{aligned}\sum_{i=1}^n X_i &\sim N(n\mu, n\sigma^2) \\ \bar{X} &\sim N\left(\mu, \frac{\sigma^2}{n}\right)\end{aligned}$$

Worked example 1.8

Esme eats an average of 1900 kcal each day with a standard deviation of 400 kcal. What is the probability that in a 31-day month she eats more than 2000 kcal per day on average?

Check conditions for CLT are met

Since we are finding an average over 31 days we can use the CLT.

State distribution of the mean

$$\bar{X} \sim N\left(1900, \frac{400^2}{31}\right)$$

Calculate the probability

$$P(\bar{X} > 2000) = 0.0820 \text{ (3SF from GDC)}$$

Exercise 1E

1. The random variable X has mean 80 and standard deviation 20.

State where possible the approximate distribution of:

- (a) (i) \bar{X} if the sample has size 12.
(ii) \bar{X} if the sample has size 3.
(b) (i) \bar{X} if the average is taken from 100 observations.
(ii) \bar{X} if the average is taken from 400 observations.

(c) (i) $\sum_{i=1}^{i=50} X_i$ (ii) $\sum_{i=1}^{i=150} X_i$

2. The random variable Y has mean 200 and standard deviation 25.

A sample of size n is found. Find, where possible, the probability that:

- (a) (i) $P(\bar{Y} < 198)$ if $n = 100$ (ii) $P(\bar{Y} < 198)$ if $n = 200$
(b) (i) $P(\bar{Y} < 190)$ if $n = 2$ (ii) $P(\bar{Y} < 190)$ if $n = 3$
(c) (i) $P(|\bar{Y} - 195| > 10)$ if $n = 100$
(ii) $P(|\bar{Y} - 201| > 3)$ if $n = 400$

(d) (i) $P\left(\sum_{i=1}^{i=50} Y_i > 10500\right)$ (ii) $P\left(\sum_{i=1}^{i=150} Y_i \leq 29500\right)$

3. Random variable X has mean 12 and standard deviation 3.5.

A sample of 40 independent observations of X is taken. Use the Central Limit Theorem to calculate the probability that the mean of the sample is between 13 and 14. [5 marks]

4. The weight of a pomegranate, in grams, has mean 145 and variance 96. A crate is filled with 70 pomegranates. What is the probability that the total weight of the pomegranates in the crate is less than 10 kg? [5 marks]

5. Given that $X \sim Po(6)$, find the probability that the mean of 35 independent observations of X is greater than 7. [6 marks]

6. The average mass of a sheet of A4 paper is 5 g and the standard deviation of the masses is 0.08 g.

- Find the mean and standard deviation of the mass of a ream of 500 sheets of A4 paper.
- Find the probability that the mass of a ream of 500 sheets is within 5 g of the expected mass.
- Explain how you have used the Central Limit Theorem in your answer. [7 marks]

7. The times Markus takes to answer a multiple choice question are normally distributed with mean 1.5 minutes and standard deviation 0.6 minutes. He has one hour to complete a test consisting of 35 questions.

- Assuming the questions are independent, find the probability that Markus does not complete the test in time.
- Explain why you did not need to use the Central Limit Theorem in your answer to part (a). [6 marks]

8. A random variable has mean 15 and standard deviation 4. A large number of independent observations of the random variable is taken. Find the minimum sample size so that the probability that the sample mean is more than 16 is less than 0.05. [8 marks]

Summary

- When adding and multiplying all the data by a constant:
 - the expectation of variables generally behaves as you would expect:

$$\begin{aligned} E(aX + c) &= aE(X) + c \\ E(a_1 X_1 \pm a_2 X_2) &= a_1 E(X_1) \pm a_2 E(X_2) \end{aligned}$$

- the variance is more subtle:

$$\begin{aligned} \text{Var}(aX + c) &= a^2 \text{Var}(X) \\ \text{Var}(a_1 X_1 \pm a_2 X_2) &= a_1^2 \text{Var}(X_1) + a_2^2 \text{Var}(X_2) \end{aligned}$$

when X_1 and X_2 are independent.

- A more general result about the expectation of a function of a discrete random variable is: $E(g(X)) = \sum_i g(x_i)p_i$. For a continuous random variable with probability density function $f(x)$: $E(g(X)) = \int g(x)f(x)dx$.
- For the sum of independent random variables: $E(a_1 X_1 \pm a_2 X_2) = a_1 E(X_1) \pm a_2 E(X_2)$
 $\text{Var}(a_1 X_1 \pm a_2 X_2) = a_1^2 \text{Var}(X_1) + a_2^2 \text{Var}(X_2)$, note that $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$.
- For the product of two independent variables: $E(XY) = E(X)E(Y)$.
- For a sample of n observations of a random variable X , the **sample mean** \bar{X} is a random variable with mean $E(\bar{X}) = E(X)$ and variance $\text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n}$.

- For the sample sum $E\left(\sum_{i=1}^n X_i\right) = nE(X)$ and $\text{Var}\left(\sum_{i=1}^n X_i\right) = n\text{Var}(X)$.
- When we combine different variables we do not normally know the resulting distribution. However there are two important exceptions:
 - A *linear combination of normal variables* also follows a normal distribution. If X and Y are random variables following a normal distribution and $Z = aX + bY + c$ then Z also follows a normal distribution.
 - The sum or mean of a large sample of observations of a variable follows a normal distribution, irrespective of the original distribution – this is called the **Central Limit Theorem**. For *any* distribution if $E(X) = \mu$, $\text{Var}(X) = \sigma^2$ and $n \geq 30$ then the approximate distributions are given by:

$$\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2)$$

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

Mixed examination practice 1

This chapter does not usually have its own examination questions, so the examples below are parts of longer examination questions.

1. X is a random variable with mean μ and variance σ^2 . Y is a random variable with mean m and variance s^2 . Find in terms of μ , σ , m and s :

 - (a) $E(X - 2Y)$
 - (b) $\text{Var}(X - 2Y)$
 - (c) $\text{Var}(4X)$
 - (d) $\text{Var}(X_1 + X_2 + X_3 + X_4)$ where X_i is the i th observation of X .

[4 marks]
2. The heights of trees in a forest have mean 16 m and variance 60 m². A sample of 35 trees is measured.

 - (a) Find the mean and variance of the average height of the trees in the sample.
 - (b) Use the Central Limit Theorem to find the probability that the average height of the trees in the sample is less than 12 m.

[5 marks]
3. The number of cars arriving at a car park in a five minute interval follows a Poisson distribution with mean 7, and the number of motorbikes follows Poisson distribution with mean 2. Find the probability that exactly 10 vehicles arrive at the car park in a particular five minute interval.

[4 marks]
4. The number of announcements posted by a head teacher in a day follows a normal distribution with mean 4 and standard deviation 2. Find the mean and standard deviation of the total number of announcements she posts in a five-day week.

[3 marks]
5. The masses of men in a factory are known to be normally distributed with mean 80 kg and standard deviation 6 kg. There is an elevator with a maximum recommended load of 600 kg. With 7 men in the elevator, calculate the probability that their combined weight exceeds the maximum recommended load.

[5 marks]
6. Davina makes bracelets using purple and yellow beads. Each bracelet consists of seven randomly selected purple beads and four randomly selected yellow beads. The diameters of the beads are normally distributed with standard deviation 0.4 cm. The average diameter of a purple bead is 1.5 cm and the average diameter of a yellow bead is 2.1 cm. Find the probability that the length of the bracelet is less than 18 cm.

[7 marks]

- 7.** The masses of the parents at a primary school are normally distributed with mean 78 kg and variance 30 kg^2 , and the masses of the children are normally distributed with mean 33 kg and variance 62 kg^2 . Let the random variable P represent the combined mass of two randomly chosen parents and the random variable C the combined mass of four randomly chosen children.
- Find the mean and variance of $C - P$.
 - Find the probability that four children have a mass of more than two parents. [6 marks]
- 8.** X is a random variable with mean μ and variance σ^2 . Prove that the expectation of the mean of three observations of X is μ but the standard deviation of this mean is $\frac{\sigma}{\sqrt{3}}$. [7 marks]
- 9.** An animal scientist is investigating the lengths of a particular type of fish. It is known that the lengths have standard deviation 4.6 cm. She wishes to take a sample to estimate the mean length. She requires that the standard deviation of the sample mean is smaller than 1, and that the standard deviation of the total length of the sample is less than 22. What is the smallest sample size she could take? [6 marks]
- 10.** The marks in a Mathematics test are known to follow a normal distribution with mean 63 and variance 64. The marks in an English test follow a normal distribution with mean 61 and variance 71.
- Find the probability that a randomly chosen mark in English is higher than a randomly chosen Mathematics mark.
 - Find the probability that the mean of 12 English marks is higher than the mean of 12 Mathematics marks. [9 marks]
- 11.** The masses of loaves of bread have mean 802 g and standard deviation σ . The probability that a box containing 40 loaves of bread has mass under 32 kg is 0.146. Find the value of σ . [7 marks]

In this chapter you will learn about:

- the probability distribution describing the number of trials until a success occurs: the geometric distribution
- the probability distribution describing the number of trials until a fixed number of successes occur: the negative binomial distribution
- an algebraic function which can help us to combine probability distributions: the probability generating function.

EXAM HINT

You can find geometric probabilities and cumulative probabilities on your calculator.

2 More about statistical distributions

When we meet a random situation that we wish to model we could return to the ideas of random variables covered in the core syllabus and write out a list of all possible outcomes and their probabilities. However, as with the Binomial and Poisson distributions, it is often easier to simply recognise a situation and apply a known distribution to it. In this chapter we shall meet two new distributions which can be used to model more situations, and then meet a technique which can be used to combine distributions together.

2A Geometric distribution

If there is a series of independent trials with only two possible outcomes and an unchanging probability of success, then the geometric distribution models the number of trials x until the first success. It only depends upon p , the probability of a success. If X follows a geometric distribution we use the notation $X \sim \text{Geo}(p)$.

To calculate the probability of X taking any particular value, x , we use the fact that there must be $x - 1$ consecutive failures (each with probability $q = 1 - p$) followed by a single success. This gives the following probability mass function.

KEY POINT 2.1

If $X \sim \text{Geo}(p)$ then $P(X = x) = pq^{x-1}$ for $x = 1, 2, 3, \dots$



It is useful to apply similar ideas to get a result for $P(X > x)$. For this situation to occur we must have started with x consecutive failures, therefore $P(X > x) = q^x$.

You are not required to know the derivation of the expectation and variance of the geometric distribution, you only need to use the results, which are:

KEY POINT 2.2

If $X \sim \text{Geo}(p)$ then:

$$\text{E}(X) = \frac{1}{p} \text{ and } \text{Var}(X) = \frac{q}{p^2}$$

Worked example 2.1

(a) A normal six-sided die is rolled. What is the probability that the first '3' occurs

- (i) on the fifth throw? (ii) after the fifth throw?

(b) What is the expected number of throws it will take until a 3 occurs?

Define variables

Identify the distribution

Apply the formula for $P(X = x)$

Apply the formula for $P(X > x)$

Apply the formula for $E(X)$

(a) (i) $X = \text{'Number of throws until the first 3'}$

$$X \sim \text{Geo}\left(\frac{1}{6}\right)$$

$$\begin{aligned} P(X = 5) &= \frac{1}{6} \times \left(\frac{5}{6}\right)^4 \\ &= 0.0804 \text{ (3SF)} \end{aligned}$$

$$\begin{aligned} \text{(ii)} \quad P(X > 5) &= \left(\frac{5}{6}\right)^5 \\ &= 0.402 \text{ (3SF)} \end{aligned}$$

$$\text{(b)} \quad E(X) = \frac{1}{\left(\frac{1}{6}\right)} = 6$$

Exercise 2A

1. Find the following probabilities:

$$\text{(a) (i)} \quad P(X = 5) \text{ if } X \sim \text{Geo}\left(\frac{1}{3}\right)$$

$$\text{(ii)} \quad P(X = 7) \text{ if } X \sim \text{Geo}\left(\frac{1}{10}\right)$$

$$\text{(b) (i)} \quad P(X \leq 5) \text{ if } X \sim \text{Geo}\left(\frac{1}{4}\right)$$

$$\text{(ii)} \quad P(X < 4) \text{ if } X \sim \text{Geo}\left(\frac{2}{3}\right)$$

$$\text{(c) (i)} \quad P(X > 10) \text{ if } X \sim \text{Geo}\left(\frac{1}{6}\right)$$

$$\text{(ii)} \quad P(X \geq 20) \text{ if } X \sim \text{Geo}(0.06)$$

(d) (i) The first boy born in a hospital on a given day is the 4th baby born (assuming no multiple births).

(ii) A prize contained in 1 in 5 crisp packets is first won with the 8th crisp packet.

2. Find the expected mean and standard deviation of:
- (i) $\text{Geo}\left(\frac{1}{3}\right)$ (ii) $\text{Geo}(0.15)$
 - (i) The number of attempts to hit a target with an arrow (there is a 1 in 12 chance of hitting the target on any given attempt).
(ii) The number of times a die must be rolled up to and including the first time a multiple of 3 is rolled.
3. The probability of passing a driving test on any given attempt is 0.4 and the attempts are independent of each other.
- Find the probability that you pass the driving test on your third attempt.
 - Find the expected average number of attempts needed to pass the driving test. [5 marks]
4. There are 12 green and 8 yellow balls in a bag. One ball is drawn from the bag and replaced. This is repeated until a yellow ball is drawn.
- Find the expected mean and variance of the number of balls drawn.
 - Find the probability that the number of balls drawn is at most one standard deviation from the mean. [7 marks]
5. If $X \sim \text{Geo}(p)$, prove that $\sum_{i=1}^{\infty} P(X = i) = 1$. [4 marks]
6. If $X \sim \text{Geo}(p)$, find the mode of X . [3 marks]
7. If $T \sim \text{Geo}(p)$ and $P(T = 4) = 0.0189$, find the value of p . [3 marks]
8. $Y \sim \text{Geo}(p)$ and the variance of Y is 3 times the mean of Y . Find the value of p . [3 marks]
9. (a) If $X \sim \text{Geo}\left(\frac{3}{4}\right)$, find the smallest value of x such that $P(X = x) < 10^{-6}$.
(b) Find the smallest value of x such that $P(X > x) < 10^{-6}$. [5 marks]
10. Prove that the standard deviation of a variable following a geometric distribution is always less than its mean. [5 marks]

In some countries the name 'geometric distribution' refers to a distribution that models the number of failures before the first success. This is not the convention used in the IB, but it does demonstrate that mathematics is not an absolutely universal language.



2B Negative binomial distribution

The negative binomial distribution is an extension of the geometric distribution. It models the number of trials before the r th success. If X follows a negative binomial distribution, we

write this as $X \sim NB(r, p)$, where p is the probability of success for each trial.

For X to take a particular value, x , there must be $r - 1$ successes in the first $x - 1$ trials followed by a success on the x th trial. But the probability of $r - 1$ successes in $x - 1$ trials can be found using the binomial distribution, and then we multiply this result by p . This gives the following probability mass function.

KEY POINT 2.3

If $X \sim NB(r, p)$, then

$$P(X = x) = \binom{x-1}{r-1} p^r q^{x-r}$$

for $x = r, r+1, \dots$

Expectation algebra can be used to link the mean and variance of the negative binomial distribution to the mean and variance of the geometric distribution. See Exercise 2B, question 8.

The results are:

KEY POINT 2.4

If $X \sim NB(r, p)$ then:

$$E(X) = \frac{r}{p} \text{ and } \text{Var}(X) = \frac{rq}{p^2}$$



Why is the negative binomial distribution given this name? The answer has to do with the extension of the binomial expansion into negative powers.

Worked example 2.2

A voucher is placed in $\frac{2}{11}$ of all cereal boxes of a particular brand. Three of these vouchers can be exchanged for a toy.

- Find the probability that exactly 8 boxes of this cereal need to be opened to get enough vouchers for a toy.
- Find the expected number of boxes which need to be opened to get enough vouchers for a toy.

Define variables

Identify the distribution

Apply the formula for $P(X = x)$

Apply the formula for $E(X)$

(a) $X = \text{'Number of boxes opened until 3 vouchers found'}$

$$X \sim NB\left(3, \frac{2}{11}\right)$$

$$\begin{aligned} P(X = 8) &= \binom{8-1}{3-1} \left(\frac{2}{11}\right)^3 \left(\frac{9}{11}\right)^5 \\ &= 0.0463 \text{ (3SF)} \end{aligned}$$

$$(b) E(X) = \frac{3}{\left(\frac{2}{11}\right)} = 16.5$$

Exercise 2B

- 1.** Find the probabilities:
 - (a) (i) $P(X = 3)$ if $X \sim NB(2, 0.8)$
(ii) $P(X = 7)$ if $X \sim NB(3, 0.3)$
 - (b) (i) $P(X = 3)$ if $X \sim NB\left(5, \frac{9}{10}\right)$
(ii) $P(X = 4)$ if $X \sim NB\left(7, \frac{2}{3}\right)$
 - (c) (i) $P(X \leq 4)$ if $X \sim NB\left(3, \frac{4}{5}\right)$
(ii) $P(X > 5)$ if $X \sim NB\left(3, \frac{1}{2}\right)$
 - (d) (i) Seeing your 3rd six on the 10th roll of a die.
(ii) Getting your 5th head on the 9th flip of a coin.
 - (e) (i) Taking fewer than 6 attempts to roll your second one on a die.
(ii) Taking more than 5 attempts to pick your second heart from a standard suit of cards (with cards being replaced).
- 2.** Find the expected mean and standard deviation of:
 - (a) (i) $NB(2, 0.8)$ (ii) $X \sim NB(3, 0.3)$
 - (b) (i) $NB\left(n, \frac{1}{n}\right)$ (ii) $NB\left(2n+1, \frac{1}{2n}\right)$
 - (c) (i) The number of rolls required to roll 3 sixes on a standard die.
(ii) The number of tosses required to get 5 tails using a fair coin.
- 3.** A magazine publisher promotes his magazine by putting a concert ticket at random in one out of every five magazines. If you need 4 tickets to take friends to the concert, what is the probability that you will find your last ticket when you buy the 20th magazine? [4 marks]
- 4.** In a party game, players need to either sing or draw. 30 pieces of paper are placed in a hat, with an equal number of 'sing' and 'draw' instructions. Players take turns to take an instruction at random and then return it to the hat. Find the probability that the fifth person to sing is the tenth player. [4 marks]
- 5.** Given that $X \sim NB(4, 0.4)$:
 - (a) State the mean and variance of X .
 - (b) Find the mode of X . [5 marks]

6. A discrete random variable X follows the distribution $\text{NB}(r, p)$. If the sum of the mean and the variance of X is 10, find and simplify an expression for r in terms of p . [4 marks]

7. In a casino game Ruben rolls a die and whenever a one or a six is rolled he receives a token. The game ends when Ruben has received y tokens; he then receives $\$x$, where x is the number of rolls he has made.

- (a) The probability of the game ending on the sixth roll is $\frac{40}{729}$.

Find the value of y .

- (b) The casino wishes to make an average profit of $\$3$ per game.
How much should it charge to play the game?
(c) What is the standard deviation in the casino's profit per game? [7 marks]

8. Let X_1, X_2, \dots, X_{12} be independent random variables each having a geometric distribution with probability of success p .

$$\text{Let } Y = \sum_{i=1}^r X_i$$

- (a) Explain why the random variable Y has a negative binomial distribution.
(b) Hence prove that the variance of the negative binomial

$$\text{distribution } \text{NB}(r, p) \text{ is } \frac{rq}{p^2}. \quad [6 \text{ marks}]$$

2C Probability generating functions

We have found formulae for the mean and variance resulting from adding independent variables. However it is also useful to calculate *probabilities* of the sum of independent variables. For discrete variables we can use a technique called a **Probability Generating Function**, which links probability with algebra and calculus.

Suppose that X and Y are discrete random variables, which can only take positive integer values. If the variable Z is their sum then we can work out the probability that $Z = 2$. This could happen in three different ways: ($A = 0$ and $B = 2$) or ($A = 1$ and $B = 1$) or ($A = 2$ and $B = 0$).

If X and Y are independent we can then write that:

$$\begin{aligned} P(Z=2) &= P(A=0)P(B=2) + \\ &\quad P(A=1)P(B=1) + P(A=2)P(B=0) \end{aligned}$$

This may remind you of a situation where you multiply two long polynomials together, for example:

$$(a_0 + a_1t + a_2t^2)(b_0 + b_1t + b_2t^2)$$

We can write the coefficient of t^2 in the result as $a_0b_2 + a_1b_1 + a_2b_0$.

This suggests that there may be some benefit in writing discrete probability distributions as polynomials with the coefficient of t^x being the probability of the random variable taking the value x .

KEY POINT 2.5

The **probability generating function** of the discrete random variable X is given by:

$$G(t) = \sum_x P(X=x)t^x$$



In this expression t has no real-world meaning. It is a *dummy variable* used to keep track of the value X is taking.

An alternative definition, which we shall only use for some proofs involving generating functions, is:

KEY POINT 2.6

The probability generating function of the discrete random variable X is also given by:

$$G(t) = E(t^X)$$



This is a direct result of the expectation of a function of a variable (Key point 1.2).

Worked example 2.3

Write down the probability generating function for the distribution below.

x	2	3	5	6	7
$P(X=x)$	0.1	0.4	0.3	0.15	0.05

$$G(t) = 0.1t^2 + 0.4t^3 + 0.3t^5 + 0.15t^6 + 0.05t^7$$

Although these generating functions have several interesting features, as polynomials they are a very complicated way of writing a probability distribution. They become much more powerful when we can rewrite the polynomial in a simpler way.

Worked example 2.4

Find and simplify an expression for the probability generating function of the random variable X where $X \sim B\left(4, \frac{1}{3}\right)$.

This can be
recognised
as a binomial
expansion

$$\begin{aligned} P(X=x) &= \binom{4}{x} \left(\frac{2}{3}\right)^{4-x} \left(\frac{1}{3}\right)^x \\ G(t) &= \left(\frac{2}{3}\right)^4 + 4\left(\frac{2}{3}\right)^3 \left(\frac{1}{3}\right)t + 6\left(\frac{2}{3}\right)^2 \left(\frac{1}{3}\right)^2 t^2 + 4\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^3 t^3 + \left(\frac{1}{3}\right)^4 t^4 \\ &= \left(\frac{2}{3}\right)^4 + 4\left(\frac{2}{3}\right)^3 \left(\frac{1}{3}t\right) + 6\left(\frac{2}{3}\right)^2 \left(\frac{1}{3}t\right)^2 + 4\left(\frac{2}{3}\right)\left(\frac{1}{3}t\right)^3 + \left(\frac{1}{3}t\right)^4 \\ &= \left(\frac{2}{3} + \frac{1}{3}t\right)^4 \end{aligned}$$

Most of the important properties of the probability generating function come from its polynomial form, but in most applications we will try to use it in some other form.

The first property comes from considering $G(1)$:

$$G(1) = P(X=0) + P(X=1) \times 1 + P(X=2) \times 1^2 \dots$$

But this is just the sum of the probabilities of any value of X occurring, which is one.

KEY POINT 2.7

For any probability generating function:

$$G(1) = 1$$

The second property comes from considering the derivative of $G(t)$ with respect to t :

$$\begin{aligned} G'(t) &= P(X=0) \times 0 + P(X=1) \times 1 + P(X=2) \times 2t + \\ &\quad P(X=3) \times 3t^2 \dots \end{aligned}$$

Using the same method as above, by setting $t=1$ we get a known expression:

$$\begin{aligned} G'(1) &= P(X=0) \times 0 + P(X=1) \times 1 + P(X=2) \times 2 + \\ &\quad P(X=3) \times 3 \dots \end{aligned}$$

KEY POINT 2.8

$$G'(1) = E(X)$$

If we differentiate the definition of a probability generating function twice and then set $t = 1$ we get:

$$\begin{aligned} G''(t) &= \sum x(x-1)t^{x-2}P(X=x) \\ G''(1) &= \sum x(x-1)P(X=x) \\ &= \sum (x^2 - x)P(X=x) \\ &= \sum x^2 P(X=x) - \sum x P(X=x) \\ &= E(X^2) - E(X) \end{aligned}$$

Therefore, since $\text{Var}(X) = E(X^2) - [E(X)]^2$:

KEY POINT 2.9

$$\text{Var}(X) = G''(1) + G'(1) - (G'(1))^2$$

Worked example 2.5

If $G(t) = e^{3t-3}$ find $E(X)$ and $\text{Var}(X)$.

Find $G'(t)$ and $G''(t)$

$$G'(t) = 3e^{3t-3}$$

$$G''(t) = 9e^{3t-3}$$

Use formula for expectation

$$E(X) = G'(1) = 3$$

Use formula for variance

$$\begin{aligned} \text{Var}(X) &= G''(1) + G'(1) - (G'(1))^2 \\ &= 9 + 3 - 3^2 \\ &= 3 \end{aligned}$$

As well as finding the expectation and the variance we can use probability generating functions to find probabilities. We want to isolate just one coefficient in the polynomial and we can do this by differentiating until the coefficient we want is a constant term, and then setting $t = 0$:

For example:

$$\begin{aligned} G(t) &= P(X=0) + P(X=1)t + P(X=2)t^2 \dots \\ \text{therefore } G(0) &= P(X=0). \end{aligned}$$

$$\begin{aligned} G'(t) &= P(X=1) + P(X=2) \times 2t + P(X=3) \times 3t^2 \dots \\ \text{therefore } G'(0) &= P(X=1). \end{aligned}$$

$$\begin{aligned} G''(t) &= P(X=2) \times 2 + P(X=3) \times 6t + P(X=4) \times 12t^2 \dots \\ \text{therefore } G''(0) &= 2P(X=2). \end{aligned}$$

In general we find the following probability mass function.

KEY POINT 2.10

$$P(X=n) = \frac{1}{n!} G^{(n)}(0)$$

Exercise 2C

1. Find the probability generating function for each of the following distributions:

(a)	X	1	2	3	4	5	6
	$P(X=x)$	0.5	0.2	0.1	0	0.05	0.05

(b)	X	1	2	3	4
	$P(X=x)$	0.3	0.3	0.3	0.1

2. For each of the following probability generating functions find $P(X=1)$:

- (a) (i) $G(t) = 0.6 + 0.4t$ (ii) $G(t) = 0.3 + 0.4t + 0.3t^2$
- (b) (i) $G(t) = 0.1t^2 + 0.9$ (ii) $G(t) = t$
- (c) (i) $G(t) = \frac{(1+t)^2}{4}$ (ii) $G(t) = \frac{(1+t)^3}{8}$
- (d) (i) $G(t) = e^{5t-5}$ (ii) $G(t) = \frac{0.4t}{1-0.6t}$

3. A discrete random variable can take any value in \mathbb{N} . It has a probability generating function of $G(t) = e^{a(t-1)}$. Find the mean and the variance in terms of a . [5 marks]

4. A discrete random variable Y can take the values 2, 3, 4, ... and

has a probability generating function $G(t) = \frac{\frac{1}{9}t^2}{\left(1 - \frac{2t}{9}\right)^2}$.

- (a) Find the probability that $Y=2$.
- (b) Find the expectation and variance of Y . [4 marks]

5. Prove that if $X \sim B(n, p)$ then the probability generating function of X is $(q+pt)^n$ where $q=1-p$. [4 marks]

6. A discrete random variable X has a probability generating function $G(t) = Ae^{(t^2)}$.

- (a) Find the value of A .
- (b) Find $E(X)$.
- (c) Find $\text{Var}(X)$. [7 marks]

7. A random variable X has a probability generating function $G(t)$. Show that the probability that X takes an even value is $\frac{1}{2}(1+G(-1))$. [4 marks]

8. A random variable X has probability generating function $G(t) = \frac{k-1}{k-t}$.

(a) Prove by induction that $\frac{d^n}{dt^n} G(t) = \frac{n!(k-1)}{(k-t)^{n+1}}$.

- (b) Hence or otherwise find the probability distribution of X in terms of k .

- (c) Find the expectation and variance of X in terms of k .

[14 marks]

9. A discrete random variable X has probability generating function $G_X(t)$. If $Y = aX + b$ show that the probability generating function of Y is given by $G_Y(t) = t^b G_X(t^a)$.

Hence prove that $E(Y) = aE(X) + b$ and that $\text{Var}(Y) = a^2\text{Var}(X)$.

[13 marks]

2D Using probability generating functions to find the distribution of the sum of discrete random variables

Each of the discrete distributions you already know has a probability generating function:

KEY POINT 2.11

Distribution	Probability generating function
$B(n, p)$	$(q + pt)^n$
$\text{Geo}(p)$	$\frac{pt}{1 - qt}$
$\text{Po}(\lambda)$	$e^{\lambda(t-1)}$

We now return to the original purpose of probability generating functions: finding the probability distribution of the sum of independent random variables.

When we have two distinct generating functions for the random variables X and Y we shall label them as $G_X(t)$ and $G_Y(t)$. We can find the probability generating function of $Z = X + Y$ by using the definition of probability generating functions given in Key point 2.6:

$$G_Z(t) = E(t^Z) = E(t^{X+Y}) = E(t^X \times t^Y) = E(t^X) \times E(t^Y) = G_X(t) \times G_Y(t)$$

In the penultimate step we used Key point 1.4 which requires that X and Y are independent. We can therefore state that:

KEY POINT 2.12

If $Z = X + Y$ where X and Y are independent:

$$G_Z(t) = G_X(t) \times G_Y(t)$$

Worked example 2.6

Find the probability generating function of the negative binomial function.

The negative binomial distribution is the sum of r geometric distributions:

$$\text{If } X \sim NB(r, p) \text{ and } Y_i \sim Geo(p) \text{ then } X = \sum_{i=1}^{i=r} Y_i$$

Therefore the probability generating function is the product of the generating functions of r geometric distributions:

$$G(t) = \left(\frac{pt}{1-qt} \right)^r$$

Exercise 2D

1. A football team gets three points for a win, one point for a draw and no points for a loss.

St Atistics football team win 40% of their matches, draw 30% and lose the rest. X is the number of points they receive from one game.

- (a) Find the probability generating function for X .
- (b) St Atistics play ten matches in their season. The results of their matches are independent. Find the probability generating function of Y , their total number of points.
- (c) Find the expected number of points at the end of the season. [7 marks]

2. Prove using generating functions that if X and Y are independent random variables then $E(X+Y) = E(X) + E(Y)$. [5 marks]

3. Prove that the sum of two Poisson variables also follows a Poisson distribution. [4 marks]

4. If $X \sim B(n, p)$ and $Y \sim B(m, p)$ prove that $X + Y$ also follows a binomial distribution and state its parameters. [5 marks]

5. A textbook contains short questions, worth one mark each, and long questions worth four marks each. 30% of questions are short questions. Let M be the number of marks for answering one question correctly.

- (a) Find the probability generating function for M .
Caroline selects eight questions at random, and answers them all correctly. Let T be her total number of marks.
(b) Write down the probability generating function for T .
(c) Show that she cannot score exactly 15 marks. [12 marks]

Summary

- In this chapter we have met two new distributions:
 - The number of trials until the first success (geometric).
 - The number of trials until a specified number of successes (negative binomial).
- For both of these distributions we found the probability mass function and a formula for the expectation and variance, all of which are in the Formula booklet:
 - For the geometric distribution, if $X \sim \text{Geo}(p)$ then:
 $P(X = x) = pq^{x-1}$ for $x = 1, 2, 3, \dots$
 $E(X) = \frac{1}{p}$ and $\text{Var}(X) = \frac{q}{p^2}$
 - For the negative binomial distribution, if $X \sim \text{NB}(r, p)$ then:
 $P(X = x) = \binom{x-1}{r-1} p^r q^{x-r}$ for $x = r, r+1, \dots$
(q is the probability of a ‘failure’, so $q = 1 - p$)
 $E(X) = \frac{r}{p}$ and $\text{Var}(X) = \frac{rq}{p^2}$
- We met a new technique for writing probability distributions called a **probability generating function**. The probability generating function of the discrete random variable X is given by
 $G(t) = \sum_x P(X = x)t^x$ and $G(t) = E(t^X)$.
For any probability generating function $G(1) = 1$.
- We saw how probability generating functions can be used to find the expectation and variance of a random variable: $G'(1) = E(X)$ and $\text{Var}(X) = G''(1) + G'(1) - (G'(1))^2$.
- We saw that each of the discrete distributions we already know has a probability generating function:

Distribution	Probability generating function
$B(n, p)$	$(q + pt)^n$
$\text{Geo}(p)$	$\frac{pt}{1-qt}$
$\text{Po}(\lambda)$	$e^{\lambda(t-1)}$

- We saw how probability generating functions can be used to find the probability mass function of a sum of independent random variables:
 - $P(X = n) = \frac{1}{n!} G^{(n)}(0)$
 - If $Z = X + Y$ where X and Y are independent, $G_Z(t) = G_X(t) \times G_Y(t)$.

Mixed examination practice 2

 You might want to remind yourself of the binomial, Poisson and normal distributions  before reading on.

- 1.** A bag contains a large number of coloured pens, $\frac{1}{3}$ of which are red. Find the probability that:
 - (a) I have to select exactly 3 pens before I get a red one.
 - (b) I have to select at least 3 pens before I get a red one.

[6 marks]
- 2.** The probability generating function of the discrete random variable X is given by $G(t) = k(1 + 2t + t^2)$.
 - (a) Find the value of k .
 - (b) Find the mode of X .

[4 marks]
- 3.** Sweets are sold in packets of 20. The probability that a sweet is a fizzy cola bottle is 0.2.
 - (a) Find the probability that a pack of sweets contains exactly 5 fizzy cola bottles.
 - (b) Find the probability that I have to buy 10 packets of sweets before I get 4 with exactly 5 fizzy cola bottles.

[8 marks]
- 4.** The masses of apples are normally distributed with mean 136 g and standard deviation 27 g.
 - (a) Find the probability that an apple has a mass of more than 150 g.
 - (b) Find the probability that in a pack of six apples at least two have a mass of more than 150 g.
 - (c) What is the expected number of apples I need to buy before I get two which have a mass of more than 150 g?

[7 marks]
- 5.** Given that $X \sim \text{Geo}(p)$ and that $P(X \leq 10) = 0.175$, find the value of p .

[4 marks]
- 6.** (a) Show that if the random variable X has a probability generating function $G(t)$ then the probability of X taking an odd value is $\frac{1}{2}(1 - G(-1))$.
(b) X is a random variable such that $X \sim B(10, 0.2)$. Write down the probability generating function of X .
(c) Y is a random variable such that $Y \sim B(12, 0.25)$. If $Z = X + Y$, write down the probability generating function of Z .
(d) Find the probability that Z is even.

[12 marks]

- 7.** A fair six-sided die is rolled repeatedly.
- Find the probability that 5 sixes are obtained from 20 rolls.
 - Find the probability that the 5th six is obtained on the 20th roll.
 - Given that the 2nd six is obtained on the 6th roll, find the probability that 5 sixes are obtained from 20 rolls.
 - Given that 5 sixes are obtained from 20 rolls, find the probability that the 2nd six was rolled on the 6th roll.
- [12 marks]
- 8.** Ian has joined a new social networking site. In order to join a particular group he needs to get nine invitations. The probability that he receives an invitation on any given day is 0.8, independently of other days (he never gets more than one invitation in a day).
- What is the expected number of days Ian has to wait before he can join the group?
 - Find the probability that Ian will first be able to join the group on the 14th day.
 - Given that after 10 days he has had 8 invitations, find the probability that he will first be able to join the group on the 14th day.
 - Ian joins the group as soon as he receives his 9th invitation. Given that Ian has joined the group on the 14th day, find the probability that he received his first invitation on the first day.
- [12 marks]
- 9.** The number of letters Naomi receives in a week follows a Poisson distribution with mean 5.
- Find the probability that in a particular week she receives more than an average number of letters.
 - What is the expected number of weeks she has to wait before she receives more than an average number of letters in a week?
 - Naomi wants to know how long she needs to wait until she has received more than an average number of letters 5 times.
 - Find the probability that she has to wait exactly 12 weeks.
 - What is the most likely number of weeks she has to wait?
- [9 marks]

- 10.** Random variable X has distribution $\text{NB}(r, p)$.

- Show that $\frac{P(X = x+1)}{P(X = x)} = \frac{x(1-p)}{x-r+1}$.
- (i) Show that $P(X = x+1) > P(X = x)$ when $x < \frac{r-1}{p}$ and $P(X = x+1) < P(X = x)$ when $x > \frac{r-1}{p}$.

(ii) Deduce that X is bimodal only if $\frac{r-1}{p}$ is an integer.

(c) $X \sim NB(9, p)$ has modes 12 and 13. Find the value of p . [11 marks]

11. A discrete random variable X has probability generating function $G(t) = kte^t$.

(a) Find the value of k .

(b) Prove by induction that $G^{(n)}(t) = ke^n(n+t)$.

(c) Hence find $P(X=7)$. [12 marks]

In this chapter you will learn:

- how to convert between the probability mass function, $P(X = x)$, and the cumulative distribution function, $P(X \leq x)$
- how to convert between the probability density function and the cumulative distribution function
- how to use the cumulative distribution function to find the median and quartiles
- how to use the cumulative distribution function to find the distribution of the function of a random variable.

3 Cumulative distribution functions

Cumulative distribution functions give the probability of a random variable being less than or equal to a particular value. They allow us quickly to find a range of values of a discrete variable. In the past these functions were the only way of tabulating probabilities for continuous random variables, but today we can use our graphical display calculators (GDC) to do this for us. However, the cumulative distribution functions are still a very important tool for working with continuous variables because they connect directly to probabilities, unlike the probability density function.

3A Finding the cumulative probability function

For a discrete variable with probability mass function $P(X = x)$, the cumulative probability function is found by adding up all of the probabilities of values less than or equal to the given value. Frequently the cumulative distribution function will only be defined over a finite domain. At the bottom end of the domain and below it must take the value 0 and at the top end of the domain and above it must take the value 1.

KEY POINT 3.1

$$\text{For a discrete distribution } P(X \leq x) = \sum_{i=-\infty}^{i=x} p_i$$

There is a similar result for a continuous variable. If the probability density function is $f(x)$ we usually write the cumulative distribution function as $F(x)$.

KEY POINT 3.2

$$\text{For a continuous distribution } P(X \leq x) = F(x) = \int_{-\infty}^x f(t) dt$$

Since integration can be ‘undone’ by differentiation, we can find the probability density function from $F(x)$:

KEY POINT 3.3

$$f(x) = \frac{d}{dx} F(x)$$

Worked example 3.1

Find the cumulative distribution function (cdf) of a continuous random variable X , which has a probability density function $f(x) = e^x$ for $0 < x < \ln 2$.

State $F(x)$ when x is below and above the range in which $f(x)$ is defined

Use integration to find the cdf. The lower limit is zero since the pdf is zero below this point

If $x \leq 0$: $F(x) = 0$
If $x \geq \ln 2$: $F(x) = 1$

$$\begin{aligned} \text{If } 0 < x < \ln 2: \\ F(x) &= \int_0^x e^t dt \\ &= [e^t]_0^x \\ &= e^x - 1 \end{aligned}$$

Once we have the cumulative distribution function we can use it to find the median, quartiles and any other percentiles, since the p th percentile is defined as the value x such that $P(X \leq x) = p\%$.

We can write this as $F(x) = \frac{p}{100}$.

Worked example 3.2

The continuous random variable X has a cumulative distribution function:

$$F(x) = \begin{cases} 0 & x \leq 0 \\ x^2 & 0 < x < 1 \\ 1 & x \geq 1 \end{cases}$$

- (a) Find the probability density function of X .
- (b) Find the lower quartile of X .

pdf is derivative of cdf

$$(a) f(x) = \frac{d}{dx} F(x) = 2x \quad \text{if } 0 < x < 1$$

and zero otherwise

Lower quartile is 25th percentile

$$(b) \text{ At the lower quartile:}$$

$$\begin{aligned} F(x) &= 0.25 \\ x^2 &= 0.25 \\ x &= \pm 0.5 \end{aligned}$$

Decide which solution to choose

$f(x)$ is non-zero only if $0 < x < 1$ therefore $x = 0.5$

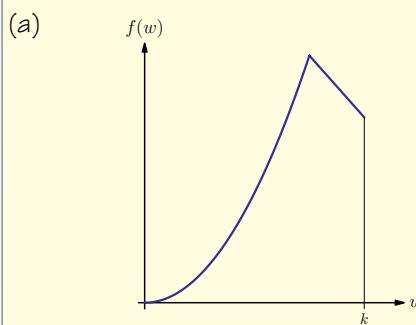
All of these techniques may be applied to a function which is defined piecewise.

Worked example 3.3

The continuous random variable W is defined by the probability density function $f(w)$

$$f(w) = \begin{cases} \frac{w^2}{27} & 0 \leq w \leq 3 \\ \frac{7}{12} - \frac{w}{12} & 3 \leq w \leq k \end{cases}$$

- (a) Sketch the probability density function.
- (b) Find the value of k .
- (c) Find $E(W)$.
- (d) Find the median of w .
- (e) Find the mode of w .



(b) The area under the curved section is

$$\int_0^3 \frac{w^2}{27} dw = \left[\frac{w^3}{81} \right]_0^3 = \frac{1}{3}$$

The remaining area is $\frac{2}{3}$ so

$$\begin{aligned} & \int_3^k \frac{7}{12} - \frac{w}{12} dw = \frac{2}{3} \\ \Rightarrow & \left[\frac{7w}{12} - \frac{w^2}{24} \right]_3^k = \frac{2}{3} \\ \Rightarrow & \left(\frac{7k}{12} - \frac{k^2}{24} \right) - \left(\frac{21}{12} - \frac{9}{24} \right) = \frac{2}{3} \\ \Rightarrow & \frac{7k}{12} - \frac{k^2}{24} = \frac{49}{24} \\ \Rightarrow & 0 = k^2 - 14k + 49 \\ & = (k-7)^2 \\ \text{So } k & = 7. \end{aligned}$$

Use the formula for $E(W)$ split up over the different domains

$$(c) E(W) = \int_0^3 w \times \frac{w^2}{27} dw + \int_3^7 w \times \left(\frac{7}{12} - \frac{w}{12} \right) dw$$

continued...

Use GDC to evaluate the definite integrals

$$= \frac{3}{4} + \frac{26}{9}$$
$$= \frac{131}{36} \approx 3.64$$

The median is the point where $P(W < m) = \frac{1}{2}$. The area under the curved section of the pdf is $\frac{1}{3}$ so the median must lie in the second section

(d) We need

$$\int_0^m f(w) dw = \frac{1}{2}$$
$$\frac{1}{3} + \int_3^m \frac{7-w}{12} dw = \frac{1}{2}$$
$$\left[\frac{7}{12}w - \frac{w^2}{24} \right]_3^m = \frac{1}{6}$$
$$\Rightarrow \left(\frac{7m}{12} - \frac{m^2}{24} \right) - \left(\frac{21}{12} - \frac{9}{24} \right) = \frac{1}{6}$$
$$0 = m^2 - 14m + 37$$

Using the quadratic equation

$$m = \frac{14 \pm \sqrt{196 - 148}}{2}$$
$$= 7 \pm 2\sqrt{3}$$

But the median lies between 3 and 7 so the median is $7 - 2\sqrt{3} \approx 3.54$

The mode corresponds to the highest point on the graph

(e) From the sketch, the mode is when $W = 3$.

Exercise 3A

1. Find the cumulative distribution function for each of the following distributions:

(a) (i) $P(X = x) = \frac{1}{5}$ for $x = 1, 2, 3, 4, 5$

(ii) $P(X = x) = \frac{1}{10}$ for $x = 1, 2, 3, \dots, 10$

(b) (i) $P(X = x) = \frac{1}{4}$ for $x = 3, 4, 5, 6$

(ii) $P(X = x) = \frac{1}{10}$ for $x = 0, 0.1, 0.2, \dots, 0.9$

2. Find the cumulative distribution function for each of the following probability density functions, and hence find the median of the distribution:

(a) (i) $f(x) = \begin{cases} 2-2x & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

(a) $\lim_{n \rightarrow \infty} a_n = \alpha$

$$= \frac{b_n x^n + b_{n-1} x^{n-1} + \dots + b_1 x + b_0}{a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0}$$

$$= \frac{\cancel{a_n} x^n + b_{n-1} x^{n-1} + \dots + b_1 x + b_0}{\cancel{a_n} x^n + \cancel{a_{n-1}} x^{n-1} + \dots + \cancel{a_1} x + a_0}$$

$$= \frac{x^n + \frac{b_{n-1}}{a_n} x^{n-1} + \dots + \frac{b_1}{a_n} x + \frac{b_0}{a_n}}{x^n + 1}$$

$$= \frac{1 + \frac{b_{n-1}}{a_n} x^{-1} + \dots + \frac{b_1}{a_n} x^{-n+1} + \frac{b_0}{a_n} x^{-n}}{1 + \frac{1}{x}}$$

$$\therefore \lim_{n \rightarrow \infty} a_n = \alpha$$

P(A|B) = P(A ∩ B) / P(B)

3. For each of the following continuous cumulative probability functions, find the probability density function and the median:

(a) (i) $F(x) = \begin{cases} 0 & x < 1 \\ x-1 & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$

(ii) $F(x) = \begin{cases} 0 & x < 0 \\ 3x & 0 \leq x < \frac{1}{3} \\ 1 & x \geq \frac{1}{3} \end{cases}$

(b) (i) $F(x) = \begin{cases} 0 & x < 1 \\ x^2 - x & 1 \leq x < \frac{1+\sqrt{5}}{2} \\ 1 & x \geq \frac{1+\sqrt{5}}{2} \end{cases}$

(ii) $F(x) = \begin{cases} 0 & x < 0 \\ \sin x & 0 \leq x < \frac{\pi}{2} \\ 1 & x \geq \frac{\pi}{2} \end{cases}$

4. A discrete random variable has the cumulative distribution function $P(X \leq x) = \frac{x(x+1)(2x+1)}{1224}$ for $x = 1, 2, 3, \dots, n$.

(a) Find $P(X = 3)$.
 (b) Find n . [5 marks]

5. Find the exact value of the 80th percentile of the continuous random variable Y which has pdf $f(y) = \frac{1}{y}$ for $1 < y < e$. [4 marks]

6. (a) If $P(X = x) = \frac{x}{10}$ for $x = 1, 2, 3, 4$ find $P(X \leq x)$.
 (b) Find the median of X . [5 marks]

42 Topic 7 – Option: Statistics and Probability

7. (a) If $P(Y = y) = \frac{y}{22}$ for $y = 4, 5, 6, 7$ find $P(Y \leq y)$.
 (b) Find the mode of Y . [4 marks]

8. A continuous variable X has cumulative distribution function:

$$F(x) = \begin{cases} 0 & x < 0 \\ e^{2x} - 1 & 0 \leq x < k \\ 1 & x \geq k \end{cases}$$

- (a) Find the value of k .
 (b) Find the probability density function for x .
 (c) Find the median of the distribution. [6 marks]

9. $P(X \leq x) = \frac{x^3}{1000}$ for $x = 1, 2, 3, \dots, n$.

- (a) Find the value of n .
 (b) Find the probability mass function of X . [4 marks]

10. The continuous random variable X has the probability density function:

$$f(x) = \begin{cases} ax^3 & 0 \leq x < 1 \\ \frac{a}{x} & 1 \leq x < 2 \\ 0 & \text{otherwise} \end{cases}$$

- (a) Find the value of the parameter a .
 (b) Find the expectation of X .
 (c) Find the cumulative distribution function of X .
 (d) Find the median of X .
 (e) Find the lower quartile of X .
 (f) What is the probability that X lies between the median and the lower quartile? [25 marks]

3B Distributions of functions of a continuous random variable

Using this discrete distribution

x	-1	0	1
$P(X = x)$	$\frac{3}{11}$	$\frac{1}{11}$	$\frac{7}{11}$

you can find the distribution of a random variable Y which is related to X by the formula $Y = X^2 + 3$ by simply listing all the

possible values of Y and their probabilities (remembering that $y = 4$ when $x = 1$ or -1):

y	3	4
$P(Y=y)$	$\frac{1}{11}$	$\frac{10}{11}$

There are many situations where we would like to do the same thing with a continuous random variable but this is much more difficult as we cannot access probabilities directly using the probability density function. We must use the cumulative function instead and then differentiate it to find the probability density function.

Worked example 3.4

X is the crv 'length of the side of a square' and X has pdf $f(x) = \frac{1}{2}$ for $1 < x < 3$. Find the probability density function of Y , the area of the square.

We need to relate $F(x)$ to $G(y)$

The cdf of X is

$$F(x) = \frac{x}{2} - \frac{1}{2}, \quad 1 < x < 3$$

The cdf of Y is $G(y)$

$$\begin{aligned} G(y) &= P(Y < y) \\ &= P(X^2 < y) \end{aligned}$$

Use the fact that $Y = X^2$

$$= P(-\sqrt{y} < X < \sqrt{y})$$

Solve the inequality

$$= P(X < \sqrt{y}) - P(X < -\sqrt{y})$$

Write in terms of cumulative probabilities

$$= F(\sqrt{y}) - F(-\sqrt{y})$$

Write in terms of the cdf of X

$$= \frac{\sqrt{y}}{2} - \frac{1}{2} - 0$$

Remember that $F(x) = 0$ when $x < 1$

This is only true if $1 < \sqrt{y} < 3$
i.e. $1 < y < 9$

Consider the domain of $F(x)$

$$g(y) = \frac{d}{dy} \left(\frac{\sqrt{y}}{2} - \frac{1}{2} \right) = \frac{1}{4\sqrt{y}}, \quad 1 < y < 9$$

pdf is the derivative of cdf

EXAM HINT

This manipulation is challenging. Thankfully, it has only rarely appeared on examination questions.

The general method for finding the probability density function is:

KEY POINT 3.4

If X has cdf $F(x)$ for $a < x < b$ and $Y = g(X)$ (where $g(X)$ is a 1-to-1 function) then the probability density function of Y , $h(y)$, is given by:

- Relating $H(y)$ to $F(g^{-1}(y))$ by rearranging the inequality in $P(Y \leq X) = P(g(X) \leq y)$.
- Differentiating $H(y)$ with respect to y .
- Writing the domain of $h(y)$ by solving the inequality $a < g^{-1}(y) < b$.

Exercise 3B

- 1.** X is a continuous random variable with pdf

$$f(x) = \frac{4}{x^5} \text{ for } x > 1.$$

If $Y = \frac{1}{X^2}$, show that Y has pdf

$$g(y) = 2y, \quad 0 < y < 1.$$

[7 marks]

- 2.** The volume (V) of a spherical soap bubble follows a continuous uniform distribution: $f(v) = \frac{1}{10}$ for $v \in (0, 10)$.

(a) Find the cumulative distribution function of V .

(b) Hence find the probability density function of R , the radius of the bubble.

[6 marks]

- 3.** X is a continuous random variable with pdf

$$f(x) = \frac{3}{26}x^2, \quad 1 < x < 3.$$

(a) Find the cumulative distribution function of X .

(b) If $Y = \frac{1}{X}$, find the probability that $Y > \frac{3}{4}$.

(c) Find the probability density function of Y .

[13 marks]

- 4.** X is a continuous random variable with pdf

$$f(x) = 1 \quad 0 < x < 1$$

Three independent observations of X are made. Find the probability density function of Y where $Y = \max(X_1, X_2, X_3)$.

[4 marks]

Summary

- The cumulative distribution function gives the probability of the random variable taking a value less than or equal to x .
- For a discrete distribution with probability mass function $P(X = x)$:
$$P(X \leq x) = \sum_{i=-\infty}^{i=x} p_i$$
- For a continuous distribution with pdf $f(x)$:
$$P(X \leq x) = F(x) = \int_{-\infty}^x f(t) dt \text{ and } f(x) = \frac{d}{dx} F(x)$$
- The main uses of cumulative distribution functions are finding percentiles of a distribution and converting from a distribution of one continuous variable to a distribution of a function of that variable.
- If X has cdf $F(x)$ for $a < x < b$ and $Y = g(X)$ (where $g(X)$ is a 1-to-1 function) then the probability density function of Y , $h(y)$, is given by:
 - Relating $H(y)$ to $F(g^{-1}(y))$ by rearranging the inequality in $P(Y \leq X) = P(g(X) \leq y)$.
 - Differentiating $H(y)$ with respect to y .
 - Writing the domain of $h(y)$ by solving the inequality $a < g^{-1}(y) < b$.

Mixed examination practice 3

1. The continuous random variable Y has probability density $g(y) = ky(1-y)$, $0 < y < 1$.
Find the cumulative distribution function of Y . [6 marks]
2. The continuous random variable X has the probability density function $f(x) = x + k - 5$, $5 \leq x \leq 6$.

 - (a) Find the cumulative distribution function of X .
 - (b) Find the exact value of the median of X . [9 marks]
3. If the continuous random variable X has pdf $f(x) = \frac{3}{4}(1-x^2)$, $-1 < x < 1$ find the interquartile range of X . [7 marks]
4. The continuous random variable X has cdf $F(x) = \frac{1}{8}(8x - x^2 - 7)$, $1 < x < 3$.
Find the probability that in four observations of X more than two observations take a value of less than two. [5 marks]
5. The continuous random variable X has cdf $F(x) = cx^3$, $a < x < b$. The median of $F(x)$ is $\sqrt[3]{4}$. Find the values of a , b and c . [6 marks]
6. The number of beta particles emitted from a radioactive substance is modelled by a Poisson distribution with a mean of 3 emissions per second. X is the discrete random variable ‘Number of emissions in n seconds’.

 - (a) Write down the probability distribution of X .
 - (b) Find an expression for the probability that there are no emissions in a period of n seconds.
 - (c) Y is the continuous random variable ‘Time until first emission’. Using your answer to (b) find the probability density function of Y .
 - (d) Find $P(0.5 < Y < 1)$. [10 marks]

In this chapter you will learn:

- about finding a single value to estimate a population parameter
- about estimating an interval in which a population parameter lies, called a confidence interval
- how to find the confidence interval for the mean when the true variance is known
- how to find the confidence interval for the mean when the true variance is unknown.

Are there other areas of knowledge in which we have to balance usefulness against truth?



We shall look more at the theory of unbiased estimators in Section 4B.



4 Unbiased estimators and confidence intervals

In the statistics sections of the core syllabus, you should have looked exclusively at finding statistics of samples. However, we are often interested in using the sample to infer the parameters for the entire population. Unfortunately, the sample statistic does not always give us the best estimate of the population parameter. Even if we find the best single number to estimate the population parameter it is unlikely to be exactly correct. There are some situations where it is more useful to have a range of values in which we are reasonably certain the population parameter lies. This is called a confidence interval.

4A Unbiased estimates of the mean and variance

Generally the true mean of the whole population is given the symbol μ and the true standard deviation is given the symbol σ . We can only use our sample mean \bar{x} to estimate the population mean μ . Although we do not know how inaccurate this might be, we do know that it is equally likely to be an underestimate or an overestimate. The expected value of the sample mean is the population mean. We say that the sample mean is an **unbiased estimator** of the population mean.

Unfortunately, things are more complicated for the variance. The variance of a sample s_n^2 is a **biased estimator** of σ^2 . This means that the sample variance tends to get the population variance wrong in one particular direction. To illustrate how this happens, we can look at a slightly simpler measure of spread: the range. A sample can never have a larger range than the whole population, but it has a smaller range whenever it does not include both the largest and smallest value in the population. The range of a sample can therefore be expected to underestimate the range of the population. A similar idea applies to variance: s_n^2 tends to underestimate σ^2 .

Fortunately (using some quite complex maths) there is a value we can calculate from the sample which gives an unbiased estimate for the variance. It is given the symbol s_{n-1}^2 .

KEY POINT 4.1

$$s_{n-1}^2 = \frac{n}{n-1} s_n^2 \text{ is an unbiased estimator of } \sigma^2.$$

Unfortunately this does not mean that s_{n-1} is an unbiased estimate of σ , but it is often a very good approximation.

EXAM HINT

 Make sure you always know whether you are being asked to find s_n or s_{n-1} , and how to select the correct option on your calculator.

See Mixed examination practice

► question 4 at the end ► of this chapter for a demonstration of this problem.

Worked example 4.1

The IQ values of ten 12-year-old boys are summarised below:

$$\sum x = 1062, \sum x^2 = 114664.$$

Find the mean and standard deviation of this sample. Assuming this is a representative sample of the whole population of 12-year-old boys, estimate the mean and standard deviation of the whole population.

Use the formulae for \bar{x} and s_n

$$s_{n-1} = \sqrt{\frac{n}{n-1}} s_n$$

$$n = 10 \\ \bar{x} = \frac{1062}{10} = 106.2$$

$$s_n = \sqrt{\frac{114664}{10} - 106.2^2} = 13.7 \text{ (3SF)}$$

$$s_{n-1} = \sqrt{\frac{10}{9}} \times 13.7 = 14.5 \text{ (3SF)}$$

For the whole population we can estimate the mean as 106.2 and the standard deviation as 14.5

Exercise 4A

1. A random sample drawn from a large population contains the following data:

19.3, 16.2, 14.1, 17.3, 18.2.

Calculate an unbiased estimate of:

- The population mean.
- The population variance.

[4 marks]

2. A machine fills tins with beans. A sample of six tins is taken at random.

The tins contain the following amounts (in grams) of beans:

126, 130, 137, 128, 135, 133.

Find:

- The sample standard deviation.
- An unbiased estimate of the population variance from which this sample is taken.

[4 marks]

3. Vitamin F tablets are produced by a machine. The amounts of vitamin F in 30 tablets chosen at random are shown in the table.

Mass (mg)	49.6	49.7	49.8	49.9	50.0	50.1	50.2	50.3
Frequency	1	3	4	6	8	4	3	1

Find unbiased estimates of:

- The mean of the population from which this sample is taken.
- The variance of the population from which this sample is taken.

[5 marks]

4. A sample of 75 lightbulbs was tested to see how long they last. The results were:

Time (hours)	Number of lightbulbs (frequency)
$0 \leq t < 100$	2
$100 \leq t < 200$	4
$200 \leq t < 300$	8
$300 \leq t < 400$	9
$400 \leq t < 500$	12
$500 \leq t < 600$	16
$600 \leq t < 700$	9
$700 \leq t < 800$	8
$800 \leq t < 900$	6
$900 \leq t < 1000$	1

Estimate:

- The sample standard deviation.
- An unbiased estimate of the variance of the population from which this sample is taken. [5 marks]

5. A pupil cycles to school. She records the time taken on each of 10 randomly chosen days. She finds that $\sum x_i = 180$ and $\sum x_i^2 = 68580$ where x_i denotes the time, in minutes, taken on the i th day.

Calculate an unbiased estimate of:

- The mean time taken to cycle to school.
- The variance of the time taken to cycle to school. [6 marks]

6. The standard deviation of a sample is $\frac{4\sqrt{3}}{7}$ of the square root of the unbiased estimate of the population variance. How many objects are in the sample? [4 marks]

4B Theory of unbiased estimators

We can find estimators of quantities other than the mean and the variance. To do this we need a general definition of an unbiased estimator.

KEY POINT 4.2

If a population has a parameter a then the sample statistic \hat{A} is an unbiased estimator of a if $E(\hat{A}) = a$.

We can interpret this to mean that if samples are taken many times and the sample statistic is calculated each time, the average of these values tends towards the true population statistic.

Worked example 4.2

Prove that the sample mean is an unbiased estimate of the population mean.

Define the sample mean as a random variable

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Where X_i each represents the i th independent observation of \bar{X} .

Apply expectation algebra

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right)$$

$$= \frac{1}{n} E(X_1 + X_2 + \dots + X_n)$$

$$= \frac{1}{n} [E(X_1) + E(X_2) + \dots + E(X_n)]$$

continued ...

Use the fact that $E(X) = \mu$, the population mean

$$\begin{aligned} &= \frac{1}{n} (\underbrace{\mu + \mu + \dots + \mu}_{n \text{ times}}) \\ &= \frac{1}{n} n\mu \\ &= \mu \end{aligned}$$

If we have an idea what the estimator might be, we can test it by finding the expectation of that expression. It is often a good idea to first try finding the expectation of the variable and then see if there is an obvious link.

Worked example 4.3

X is a continuous random variable with probability density function $f(x) = \frac{1}{k}$, $1 < x < k+1$.

Find an unbiased estimator for k .

Start by trying $E(X)$

$$E(X) = \int_1^{k+1} \frac{x}{k} dx = \left[\frac{x^2}{2k} \right]_1^{k+1} = \frac{(k+1)^2}{2k} - \frac{1^2}{2k} = \frac{k}{2} + 1$$

This is close to what we need. We can use expectation algebra to find the required expression

$$E(2X) = k + 2$$

$$\therefore E(2X - 2) = k$$

So $2X - 2$ is an unbiased estimator of k

You may be asked to demonstrate that the sample statistic forms a biased estimate for a particular distribution.

Worked example 4.4

A distribution is equally likely to take the values 1 or 3.

- Show that the variance of this distribution is 1.
- List the four equally likely outcomes when a sample of size two is taken from this population.
- Find the expected value of S_2^2 (sample variance for samples of size two) and comment on your result.

$$(a) E(X) = 1 \times \frac{1}{2} + 3 \times \frac{1}{2} = 2$$

$$E(X^2) = 1^2 \times \frac{1}{2} + 3^2 \times \frac{1}{2} = 5$$

$$\text{Var}(X) = 5 - 2^2 = 1$$

(b) Outcomes could be 1,1 or 1,3 or 3,1 or 3,3

continued ...

For each sample of size two, we need to find its variance and its probability, and then find the expected value of the variances

(c)

Sample	Probability	\bar{x}	\bar{x}^2	s_n^2
1,1	$\frac{1}{4}$	1	1	0
1,3	$\frac{1}{4}$	2	5	1
3,1	$\frac{1}{4}$	2	5	1
3,3	$\frac{1}{4}$	9	9	0

$$\begin{aligned} E(s_n^2) &= 0 \times \frac{1}{4} + 1 \times \frac{1}{4} + 1 \times \frac{1}{4} + 0 \times \frac{1}{4} \\ &= \frac{1}{2} \end{aligned}$$

This is not the same as the population variance, so s_n^2 is a biased estimator of σ^2

There may be more than one unbiased estimator of a population parameter. One important way to distinguish between them is **efficiency**. This is measured by the variance of the unbiased estimator. The smaller the variance, the more efficient the estimator is.

Worked example 4.5

- Show that for all values of c the statistic $cX_1 + (1-c)X_2$ forms an unbiased estimate of the population mean of X .
- Find the value of c that maximises the efficiency of this estimator.

An estimator is unbiased if its expected value equals the population mean of X

The most efficient estimator has the smallest variance

$$\begin{aligned} (a) E(cX_1 + (1-c)X_2) &= cE(X_1) + (1-c)E(X_2) \\ &= c\mu + (1-c)\mu \\ &= c\mu + \mu - c\mu \\ &= \mu \end{aligned}$$

Therefore $cX_1 + (1-c)X_2$ forms an unbiased estimator of μ for all values of c .

$$\begin{aligned} (b) \text{Var}(cX_1 + (1-c)X_2) &= c^2 \text{Var}(X_1) + (1-c)^2 \text{Var}(X_2) \\ &= c^2 \sigma^2 + (1-2c+c^2) \sigma^2 \\ &= 2\sigma^2 c^2 - 2\sigma^2 c + \sigma^2 \end{aligned}$$

This is minimised when $\frac{d}{dc}(2\sigma^2 c^2 - 2\sigma^2 c + \sigma^2) = 0$

$$\Rightarrow 4\sigma^2 c - 2\sigma^2 = 0$$

$$\Rightarrow c = \frac{1}{2} \text{ if } \sigma^2 \neq 0$$

So the most efficient estimator is when $c = \frac{1}{2}$

Exercise 4B

1. A bag contains 5 blue marbles and 3 red marbles. Two marbles are selected at random without replacement.
 - (a) Find the sampling distribution of P , the proportion of the sample which is blue.
 - (b) Show that P is an unbiased estimator of the population proportion of blue marbles. [7 marks]

2. The continuous random variable X has probability distribution $f(x) = \frac{3x^2}{k^3}$ $0 < x < k$.
 - (a) Find $E(X)$.
 - (b) Hence find an unbiased estimator for k .
 - (c) A single observation of X is 7. Use your estimator to suggest a value for k . [5 marks]

3. The random variable X can take values 1, 2 or 3.
 - (a) List all possible samples of size two.
 - (b) Show that the maximum of the sample forms a biased estimate of the maximum of the population.
 - (c) An unbiased estimator for the population maximum can be written in the form $k \times \max$, where \max is the maximum of a sample of size two. Write down the value of k . [9 marks]

4. X_1, X_2 and X_3 are three independent observations of the random variable X which has mean μ and variance σ^2 .
 - (a) Show that both $A = \frac{X_1 + 2X_2 + X_3}{4}$ and $B = \frac{X_1 + 2X_2 + 3X_3}{6}$ are unbiased estimators of μ .
 - (b) Show that A is a more efficient estimator than B . [7 marks]

5. Two independent random samples of observations containing n_1 and n_2 values respectively are made of a random variable, X , which has mean μ and variance σ^2 . The means of the samples are denoted by \bar{X}_1 and \bar{X}_2 .
 - (a) Show that $c\bar{X}_1 + (1-c)\bar{X}_2$ is an unbiased estimator of μ .
 - (b) Show that the most efficient estimator of this form is $\frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$. [9 marks]

6. A biased coin has a probability p that it gives a tail when it is tossed. The random variable T is the number of tosses up to and including the second tail.
 - (a) State the distribution of T .
 - (b) Show that $P(T = t) = (t-1)(1-p)^{t-2} p^2$ for $t \geq 2$.
 - (c) Hence show that $\frac{1}{T-1}$ is an unbiased estimator of p . [8 marks]

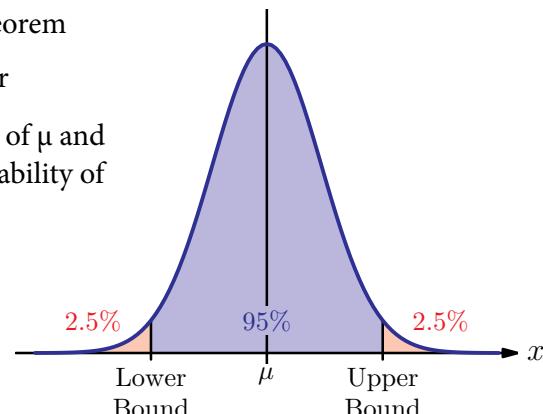
7. Two independent observations X_1 and X_2 are made of a continuous random variable with probability density function
- $$f(x) = \frac{1}{k} \quad 0 \leq x \leq k.$$
- (a) Show that $X_1 + X_2$ forms an unbiased estimator of k .
- (b) Find the cumulative distribution of X .
- (c) Hence find the probability that both X_1 and X_2 are less than m where $0 \leq m \leq k$.
- (d) Find the probability distribution of M , the larger of X_1 and X_2 .
- (e) Show that $\frac{3}{2}M$ is an unbiased estimator of k .
- (f) Find with justification which is the more efficient estimator of k : $X_1 + X_2$ or $\frac{3}{2}M$. [21 marks]

4C Confidence interval for the population mean

A **point estimate** is a single value calculated from the sample and used to estimate a population parameter. However, this can be misleading as it does not give any idea of how certain we are in the value. We want to find an interval which has a specified probability of including the *true* population value of the parameter we are interested in. This interval is called a **confidence interval** and the specified probability is called the **confidence level**. All of the confidence intervals in the IB are symmetrical, meaning that the point estimate is at the centre of the interval. For example, given the data 1, 1, 3, 5, 5, 6 we can find the sample mean as 3.5. However, it is very unlikely that the mean of the population this sample was drawn from is exactly 3.5. We shall see in Section E a method that allows us to say with 95% confidence that the population mean is somewhere between 1.22 and 5.78.

We are first going to look at creating confidence intervals for the population mean μ when the population variance σ^2 is known. This is not a very realistic situation, but it is useful to develop the theory.

Suppose we are estimating μ using a sample statistic \bar{X} . As long as the random variable is normally distributed or the sample size is large enough for the central limit theorem to apply we know that $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. We can use our knowledge of the normal distribution to find, in terms of μ and σ , a region symmetrical about μ which has a 95% probability of containing \bar{x} .



Using the method from the core syllabus we can find the Z-score of the upper bound. Using the symmetry of the situation we find that 2.5% of the distribution is above the upper bound, so the Z-score is $\Phi^{-1}(0.975) = 1.96$ (3SF). We can say that:

$$P(-1.96 < Z < 1.96) = 0.95$$

Converting to a statement about \bar{x} , μ and σ :

$$P\left(-1.96 < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

Rearranging to focus on μ :

$$P\left(\bar{x} - \frac{1.96\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{1.96\sigma}{\sqrt{n}}\right) = 0.95$$

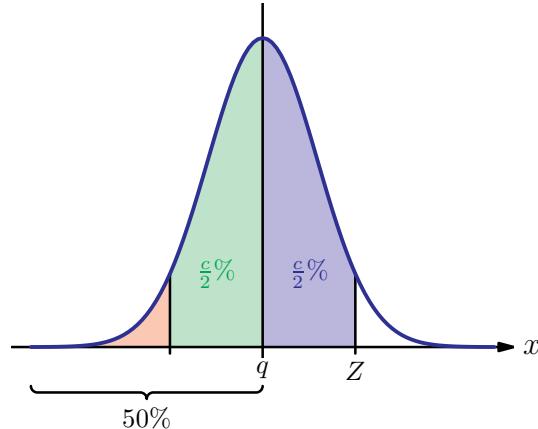
This looks like it is a statement about the probability of μ , but in our derivation we treated μ as a constant so it is meaningless to talk about a probability of μ . This statement is still concerned with the probability distribution of \bar{X} .

So our 95% confidence interval for μ based upon an observation of the sample mean is:

$$\left[\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}\right]$$

We can say that 95% of such confidence intervals contain μ , rather than the probability of μ being in the confidence interval is 95%.

We can generalise this method to other confidence levels. To find a $c\%$ confidence interval we can find the critical Z-value geometrically by thinking about the graph.



From this diagram we can see that the critical Z-value is the one where there is a probability of $0.5 + \frac{c}{100}$ being below it.

KEY POINT 4.3

When the variance is known a $c\%$ confidence interval for μ is:

$$\bar{x} - z \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z \frac{\sigma}{\sqrt{n}} \text{ where } z = \Phi^{-1}\left(0.5 + \frac{c}{100}\right)$$



EXAM HINT

EXAM HINT

Your calculator can find confidence intervals using either sampled data or summary statistics. See Calculator skills sheets C, D, G, and H.



EXAM HINT

The Formula booklet does not tell you how to find z .

Worked example 4.6

The mass of fish in a pond is known to have standard deviation 150 g. The average mass of 96 fish from the pond is found to be 806 g.

- Find a 90% confidence interval for the average mass of all the fish in the pond.
- State, with a reason, whether or not you used the central limit theorem in your previous answer.

Find the Z-score associated with a 90% confidence interval

(a) With 90% confidence we need

$$z = \Phi^{-1}(0.95) = 1.64$$

So confidence interval is $806 \pm 1.64 \times \frac{150}{\sqrt{96}}$ which is [780.9, 831.1]

(b) We did need to use the central limit theorem as we were not told that the mass of fish is normally distributed.

You do not need to know the centre of the interval to find the width of the confidence interval.

KEY POINT 30.4

The width of a confidence interval is $2z \frac{\sigma}{\sqrt{n}}$.

Worked example 4.7

The results in a test are known to be normally distributed with a standard deviation of 20. How many people need to be tested to find a 80% confidence interval with a width of less than 5?

Find the Z-score associated with a 80% confidence interval

With 80% confidence we need $z = \Phi^{-1}(0.9) = 1.28$

Set up an inequality

$$2 \times 1.28 \times \frac{20}{\sqrt{n}} < 5$$

$$\Rightarrow \frac{2 \times 1.28 \times 20}{5} < \sqrt{n}$$

$$\Rightarrow 104.9 < n$$

So at least 105 people need to be tested.

Exercise 4C

- Find z for the following confidence levels:
 - 80%
 - 99%
 - Which of the following statements are true for 90% confidence intervals of the mean?
 - There is a probability of 90% that the true mean is within the interval.
 - If you were to repeat the sampling process 100 times, 90 of the intervals would contain the true mean.
 - Once the interval has been created there is a 90% chance that the next sample mean will be within the interval.
 - On average 90% of intervals created in this way contain the true mean.
 - 90% of sample means will fall within this interval.
 - For a given sample, which will be larger: an 80% confidence interval for the mean or a 90% confidence interval for the mean?
 - Give an example of a statistic for which the confidence interval would not be symmetric about the sample statistic.
 - Find the required confidence interval for the population mean for the following summarised data. You may assume that the data are taken from a normal distribution with known variance.
 - (i) $\bar{x} = 20, \sigma^2 = 14, n = 8, 95\%$ confidence interval
(ii) $\bar{x} = 42.1, \sigma^2 = 18.4, n = 20, 80\%$ confidence interval
 - (i) $\bar{x} = 350, \sigma = 105, n = 15, 90\%$ confidence interval
(ii) $\bar{x} = -1.8, \sigma = 14, n = 6, 99\%$ confidence interval
 - Fill in the missing values in the table. You may assume that the data are taken from a normal distribution with known variance.

	\bar{x}	σ	n	Confidence level	Lower bound of interval	Upper bound of interval
(a) (i)	58.6	8.2	4	90		
	0.178	0.01	12	80		
(b) (i)		4	4		39.44	44.56
		1.2	900		30.30	30.50
(c) (i)		18		95	115.59	124.41
		25		88	1097.3	1102.7
(d) (i)			100	75	-0.601	8.601
			400	90	15.967	16.033
(e) (i)	8	12	14		0.539	
	0.4	0.01		80		0.403

7. The blood oxygen levels of an individual (measured in percent) are known to be normally distributed with a standard deviation of 3%. Based upon six readings Niamh finds that her blood oxygen levels are on average 88.2%. Find a 95% confidence interval for Niamh's true blood oxygen level. [5 marks]
8. The birth weight of male babies in a hospital is known to be normally distributed with variance 2 kg^2 . Find a 90% confidence interval for the average birth weight, if a random sample of ten male babies has an average weight of 3.8 kg. [6 marks]
9. When a scientist measures the concentration of a solution, the measurement obtained may be assumed to be a normally distributed random variable with standard deviation 0.2.
- He makes 18 independent measurements of the concentration of a particular solution and correctly calculates the confidence interval for the true value as $[43.908, 44.092]$. Determine the confidence level of this interval.
 - He is now given a different solution and is asked to determine a 90% confidence interval for its concentration. The confidence interval is required to have a width less than 0.05. Find the minimum number of measurements required.
- [8 marks]
10. A supermarket wishes to estimate the average amount single men spend on their shopping each week. It is known that the amount spent has a normal distribution with standard deviation €22.40. What is the smallest sample required so that the margin of error (the difference between the centre of the interval and the boundary) for an 80% confidence interval is less than €10?
- [5 marks]
11. The masses of bananas are investigated. The masses of a random sample of 100 of these bananas was measured and the average was found to be 168 g. From experience, it is known that the mass of a banana has variance 200 g^2 .
- Find a 95% confidence interval for μ .
 - State, with a reason, whether or not your answer requires the assumption that the masses are normally distributed.
- [6 marks]
12. A physicist wishes to find a confidence interval for the mean voltage of some batteries. He therefore randomly selects n batteries and measures their voltages. Based on his results, he obtains the 90% confidence interval $[8.884V, 8.916V]$. The voltages of batteries are known to be normally distributed with a standard deviation of 0.1V.
- Find the value of n .
 - Assuming that the same confidence interval had been obtained from measuring 49 batteries, what would be its level of confidence?
- [8 marks]

4D The *t*-distribution

In the previous section, we based calculations on the assumption that the population variance was known, even though its mean was not. In reality we commonly need to estimate the population variance from the sample. In our calculations, we then need to use a new distribution instead of the normal distribution. It is called the *t*-distribution.

If the random variable X follows a normal distribution so that $X \sim N(\mu, \sigma^2)$, or if the CLT applies, the Z-score for the mean follows a standardised normal distribution:

$$Z = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

The parameters μ and σ may be unknown, but they are constant; they are the same every time a sample of X is taken.

When the true population standard deviation is unknown, we replace it with our best estimate: s_{n-1} . We then get the *T*-score:

$$T = \frac{\bar{X}_n - \mu}{s_{n-1} / \sqrt{n}}$$

The *T*-score is not normally distributed. The proof of this is beyond the scope of the course, but we can use intuition to suggest how it might be related to the normal distribution:

- The most probable value of T will be zero. As $|T|$ increases, the probability decreases; so it is roughly the same shape as the normal distribution.
- If n is very large, our estimate of the population standard deviation should be very good, so T will be very close to a normal distribution.
- If n is very small, our estimate of the population standard deviation may not be very accurate. The probability of getting a Z-score above 3 or below -3 is very small indeed. However, if s_{n-1} is smaller than σ it is possible that T is artificially increased relative to Z . This means that the probability of getting an extreme value of T ($|T| > 3$) is significant.

From this we can conclude that T follows a different distribution depending upon the value of n . This distribution is called the ***t*-distribution** and it depends only upon the value of n .

KEY POINT 4.5

$$T = \frac{\bar{X}_n - \mu}{s_{n-1} / \sqrt{n}} \sim t_{n-1}$$

The actual formula for the probability density of t_v is

$$f(x) = \frac{(v-1)(v-3)}{2\sqrt{v}(v-2)(v-4)} \cdot \frac{5 \times 3}{4 \times 2} \left(1 + \frac{x^2}{v}\right)^{-\frac{1}{2}(v+1)} \text{ if } v \text{ is even and}$$

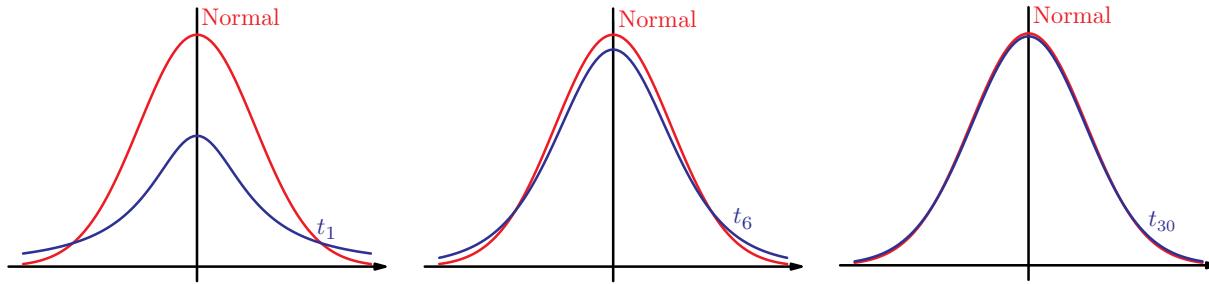
$$f(x) = \frac{(v-1)(v-3)}{\pi\sqrt{v}(v-2)(v-4)} \cdot \frac{4 \times 2}{5 \times 3} \left(1 + \frac{x^2}{v}\right)^{-\frac{1}{2}(v+1)} \text{ if } v \text{ is odd.}$$

This relates to something called the gamma function and is not on the syllabus!

The suffix is $n-1$ because that describes the number of **degrees of freedom** once s_{n-1} has been estimated. It is also given the symbol v . It is nearly always one less than the total number of data items: $v = n - 1$.

The only exception to this is when testing for correlation in Section 6B.

The shapes of these distributions are shown.



If n is large ($n > 30$) we have noted that the t -distribution is approximately the same as the normal distribution but when n is small the t -distribution is distinct from the normal distribution. We still need to have \bar{X}_n following a normal distribution, but with small n we can no longer apply the CLT. Therefore the t -distribution applies to a small sample mean only if the original distribution of X is normal.

There are two types of calculation you need to be able to do with the t -distribution:

- Find the probability that T lies in a certain range.
- Given the cumulative probability $P(T \leq t)$, find the boundary value t .

If $P(T \leq t) = p\%$ then t is called the **p th percentage point** of the distribution.

EXAM HINT

We can use a graphical calculator to find probabilities associated with the t -distribution.

See Calculator skills sheets A and B.

Worked example 4.8

Find the probability that $-1 < T < 3$ if $n = 5$.

$$\nu = n - 1 = 4$$

$$P(-1 < T < 3) = 0.793 \text{ (3SF from GDC)}$$

Worked example 4.9

If $n = 8$, find the value of t such that $P(T < t) = 0.95$.

$$\nu = n - 1 = 7$$

95th percentile point of t_7 is 1.90 (3SF from GDC)

Exercise 4D

1. In each situation below, $T \sim t_v$. (Remember that $v = n - 1$.)

Find the following probabilities:

- | | |
|--------------------------------------|----------------------------------|
| (a) (i) $P(2 < T < 3)$ if $n = 5$ | (ii) $P(-1 < T < 1)$ if $n = 8$ |
| (b) (i) $P(T \geq 5.1)$ if $\nu = 4$ | (ii) $P(T \geq -1.8)$ if $v = 6$ |
| (c) (i) $P(T < -2.4)$ if $n = 12$ | (ii) $P(T < 0.2)$ if $n = 16$ |
| (d) (i) $P(T < 1.9)$ if $n = 20$ | (ii) $P(T > 2.6)$ if $n = 17$ |



2. How does $P(2 < T < 3)$ change as n increases?

3. In each situation below, $T \sim t_v$. Find the values of t :

- | | |
|--|--|
| (a) (i) $P(T < t) = 0.8$ if $n = 13$ | |
| (ii) $P(T < t) = 0.15$ if $n = 9$ | |
| (b) (i) $P(T > t) = 0.75$ if $n = 10$ | |
| (ii) $P(T > t) = 0.3$ if $n = 20$ | |
| (c) (i) $P(T < t) = 0.6$ if $n = 14$ | |
| (ii) $P(T < t) = 0.4$ if $n = 11$ | |

4. If $T \sim t_7$ solve the equation:

$$P(T > -t) + P(T > 0) + P(T > t) + P(T > 2t) = 1.75 \quad [6 \text{ marks}]$$

4E Confidence interval for a mean with unknown variance

When finding an estimate for the population mean we do not know the true population standard deviation; we estimate it from the sample. This means that the statistic $\frac{\bar{x} - \mu}{s_{n-1}/\sqrt{n}}$ does not follow the normal distribution, but rather the t -distribution (as long as X follows a normal distribution). Following a similar analysis to the one in Section 4C we get:

KEY POINT 4.6

When the variance is not known, the $c\%$ confidence interval for the population mean is given by:

$$\bar{x} - t \frac{s_{n-1}}{\sqrt{n}} < \mu < \bar{x} + t \frac{s_{n-1}}{\sqrt{n}}$$

where t is chosen so that $P(T_\nu < t) = 0.5 + \frac{1}{2}\frac{c}{100}$.

EXAM HINT

The Formula booklet does not tell you how to find t .

Worked example 4.10

The sample $\{4, 4, 7, 9, 11\}$ is drawn from a normal distribution. Find the 90% confidence interval for the mean of the population.

Find sample mean and unbiased estimate of σ

From GDC: $\bar{x} = 7$, $s_{n-1} \approx 3.08$

Find the number of degrees of freedom

$\nu = n - 1 = 4$

Find the t -score associated with a 90% confidence interval when $\nu = 4$

95th percentage point of t_4 is 2.132 (from GDC)

Apply formula

$$7 - 2.132 \times \frac{3.08}{\sqrt{5}} < \mu < 7 + 2.132 \times \frac{3.08}{\sqrt{5}}$$

$$\therefore 4.06 < \mu < 9.94 \text{ (3SF)}$$

We are often interested in the difference between two situations, such as 'Are people more awake in the morning or afternoon?' or 'Were the results better in the French or the Spanish examinations?' If we study two different groups to look at this, we risk any observed difference being due to differences between the groups rather than differences caused by the factor being studied. One way to avoid this is to use data which are

naturally paired; the same person in the morning and afternoon, or the same person in the French and Spanish examinations. If we do this we can then simply look at the difference between the paired data and treat this as a single variable.

Worked example 4.11

Six people were asked to estimate the length of a line and the angle at a point. The percentage error in the two measurements was recorded, and it was assumed that the results followed a normal distribution. Find an 80% confidence interval for the average difference between the accuracy of estimating angles and lengths.

Person	A	B	C	D	E	F
Error in length	17	12	9	14	8	6
Error in angle	12	12	15	19	12	8

Define variables

Let $d = \text{error in angle} - \text{error in length}$

Person	A	B	C	D	E	F
	-5	0	6	5	4	2

Find sample mean and unbiased estimate of σ

$\bar{d} = 2, s_{n-1} = 4.04$ (from GDC)

Find the number of degrees of freedom

$v = n - 1 = 5$

Find the t -score associated with a 80% confidence interval when $v = 5$

90th percentage point of t_5 is 1.476

Apply formula

$$2 - 1.476 \times \frac{4.04}{\sqrt{6}} < \mu < 2 + 1.476 \times \frac{4.04}{\sqrt{6}}$$

$$\therefore -0.434 < \mu < 4.43$$

EXAM HINT

Your calculator can find confidence intervals associated with both normal and t -distributions. In your answer, you need to make it clear which distribution and which data you are using. In the above example, you would need to show the table, the values of d and s_{n-1} , state that you are using t -distribution with $v = 5$ and then write down the confidence interval.

Exercise 4E

- Find the required confidence interval for the population mean for the following data, some of which have been summarised. You may assume that the data are taken from a normal distribution.
 - (i) $\bar{x} = 14.1$, $s_{n-1} = 3.4$, $n = 15$, 85% confidence interval
 (ii) $\bar{x} = 191$, $s_{n-1} = 12.4$, $n = 100$, 80% confidence interval
 - (i) $\bar{x} = 18$, $s_n = 2.7$, $n = 10$, 95% confidence interval
 (ii) $\bar{x} = 0.04$, $s_n = 0.01$, $n = 4$, 75% confidence interval
 - (i) $\sum_{i=1}^{15} x_i = 32$, $\sum_{i=1}^{15} x_i^2 = 1200$, 75% confidence interval
 (ii) $\sum_{i=1}^{20} x_i = 18$, $\sum_{i=1}^{20} x_i^2 = 650$, 90% confidence interval
 - (i) $x = \{1, 1, 3, 5, 12, 20\}$, 95% confidence interval
 (ii) $x = \{150, 210, 130, 96, 209\}$, 90% confidence interval

- Find the required confidence intervals for the average difference (after – before) for the data below, given that the data are normally distributed.

(a) 95% confidence interval

Subject	A	B	C	D	E
Before	16	20	20	16	12
After	18	24	18	16	16

(b) 99% confidence interval

Subject	A	B	C	D	E	F
Before	4.2	6.5	9.2	8.1	6.6	7.1
After	5.3	5.5	8.3	9.0	6.1	7.0

- The times taken for a group of children to complete a race are recorded:

t (minutes)	Number of children
$8 \leq t < 12$	9
$12 \leq t < 14$	18
$14 \leq t < 16$	16
$16 \leq t < 20$	20

Assuming that these children are drawn from a random sample of all children, calculate:

- An unbiased estimate of the mean time taken by a child in the race.
- An unbiased estimate of the variance of the time taken.
- A 90% confidence interval for the mean time taken.

[7 marks]

4. Four pupils took a Spanish test before and after a trip to Mexico. Their scores are shown in the table.

	Amir	Barbara	Chris	Delroy
Before trip	12	9	16	18
After trip	15	12	17	18

Find a 90% confidence interval for the average increase in scores after the trip. [4 marks]

5. A garden contains many rose bushes. A random sample of eight bushes is taken and the heights in centimetres were measured and the data were summarised as:

$$\sum x = 943, \sum x^2 = 113005$$

- (a) State an assumption that is necessary to find a confidence interval for the mean height of rose bushes.
- (b) Find the sample mean.
- (c) Find an unbiased estimate for the population standard deviation.
- (d) Find an 80% confidence interval for the mean height of rose bushes in the garden. [9 marks]

6. The mass of four steaks (in grams) before and after being cooked for one minute is measured.

Steak	A	B	C	D
Before cooking	148	167	160	142
After cooking	124	135	134	x

A confidence interval for the mean mass loss was found to include values from 21.5 g to 31.0 g.

- (a) Find the value of x .
- (b) Find the confidence level of this interval. [10 marks]

7. A sample of 3 randomly selected students are found to have a variance of 1.44 hours² in the amount of time they watch television each weekday. Based upon this sample the confidence interval for the mean time a student spends watching television is calculated as [3.66, 7.54].

- (a) Find the mean time spent watching television.
- (b) Find the confidence level of the interval. [8 marks]

8. The random variable X is normally distributed with mean μ . A random sample of 16 observations is taken on X , and it is found that:

$$\sum_{i=1}^{16} (x_i - \bar{x})^2 = 984.15$$

A confidence interval [40.88, 46.72] is calculated for this sample. Find the confidence level for this interval. [8 marks]

9. The lifetime of a printer cartridge, measured in pages, is believed to be approximately normally distributed. The lifetimes of 5 randomly chosen print cartridges were measured and the results were:

120, 480, 370, 650, x

A confidence interval for the mean was found to be [218, 510].

- Find the value of x .
- What is the confidence level of this interval? [8 marks]

10. The temperature of a block of wood 3 minutes after being lifted out of liquid nitrogen is measured and then the experiment is repeated. The results are -1.2°C and 4.8°C .

- Assuming that the temperatures are normally distributed find a 95% confidence interval for the mean temperature of a block of wood 3 minutes after being lifted out of liquid nitrogen.
- A different block of wood is subjected to the same experiment and the results are 0°C and $x^{\circ}\text{C}$ where $x > 0$. Prove that the two confidence intervals overlap for all values of x . [12 marks]

11. In a random sample of three pupils, x_i is the mark of the i th pupil in a test on volcanoes and y_i is the mark of the i th pupil in a test on glaciers. All three pupils sit both tests.

- Show that $\bar{y} - \bar{x}$ is always the same as $\bar{y} - \bar{x}$.
- Give an example to show that the variance of $y - x$ is not necessarily the same as the difference between the variance of y and the variance of x .
- It is believed that the difference between the results in these two tests follows a normal distribution with variance 16 marks. If the mean mark of the volcano test was 23 and the mean mark for the glacier test was 30, find a 95% confidence interval for the improvement in marks from the volcano test to the glacier test. [10 marks]

Summary

- An **unbiased estimator** of a population parameter has an expectation equal to the population parameter: if a is a parameter of a population then the sample statistic \hat{A} is an unbiased estimator of a if $E(\hat{A}) = a$. This means that if samples are taken many times and the sample statistic calculated each time, the average of these values tends towards the true population statistic.
- The sample mean (\bar{X}) is an unbiased estimator of the population mean μ .
- The sample variance (s_{n-1}^2) is a **biased estimator** of the population variance (σ^2), but the value $s_{n-1}^2 = \frac{n}{n-1} s_n^2$ is an unbiased estimate.

- S_{n-1} is not an unbiased estimator of the standard deviation, but it is often a very good approximation.
- There may be more than one unbiased estimator of a population parameter. The **efficiency** of a parameter is measured by the variance of the unbiased estimator; the smaller the variance, the more efficient the estimator is.
- If X follows a normal distribution with mean μ and unknown variance, and if a random sample of n independent observations of X is taken, then it is useful to calculate the

$$T\text{-score: } T = \frac{\bar{X}_n - \mu}{s_{n-1} / \sqrt{n}}$$

This follows a t_{n-1} distribution.

- Rather than estimating a population parameter using a single number (a **point estimate**), we can provide an interval (called the **confidence interval**) that has a specified probability (called the **confidence level**) of including the true population value of the statistic we are interested in:

- The width of a confidence interval is $2z \frac{\sigma}{\sqrt{n}}$.
- If the true population variance is known and the sample mean follows a normal distribution then the $c\%$ confidence interval takes the form $\bar{x} - z \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z \frac{\sigma}{\sqrt{n}}$, where $z = \Phi^{-1}\left(0.5 + \frac{c}{200}\right)$
- If the true population variance is unknown and the population follows a normal distribution then the $c\%$ confidence interval takes the form $\bar{x} - t \frac{s_{n-1}}{\sqrt{n}} < \mu < \bar{x} + t \frac{s_{n-1}}{\sqrt{n}}$, where t is chosen so that $P(T_v < t) = 0.5 + \frac{c}{200}$.

Mixed examination practice 4

1. The mass of a sample of 10 eggs is recorded and the results in grams are:
 62, 57, 84, 92, 77, 68, 59, 80, 81, 72
 Assuming that these masses form a random sample from a normal population, calculate:
 (a) Unbiased estimates of the mean and variance of this population.
 (b) A 90% confidence interval for the mean. [6 marks]

2. From experience we know that the variance in the increase between marks in a beginning of year test and an end of year test is 64. A random sample of four students was selected and the results in the two tests were recorded.

	Alma	Brenda	Ciaron	Dominique
Beginning of year	98	62	88	82
End of year	124	92	120	116

Find a 95% confidence interval for the mean increase in marks from the beginning of year to the end of year. [5 marks]

3. The time (t) taken for a mechanic to replace a set of brake pads on a car is recorded. In a week she changes 14 tyres and $\sum t = 308$ minutes and $\sum t^2 = 7672$ minutes². Assuming that the times are normally distributed, calculate a 98% confidence interval for the mean time taken for the mechanic to replace a set of brake pads. [7 marks]

4. A distribution is equally likely to take the values 1 or 4. Show that s_{n-1} forms a biased estimator of σ . [8 marks]

5. The random variable X is normally distributed with mean μ and standard deviation 2.5. A random sample of 25 observations of X gave the result $\sum x = 315$.
- (a) Find a 90% confidence interval for μ .
 (b) It is believed that $P(X \leq 14) = 0.55$. Determine whether or not this is consistent with your confidence interval for μ . [12 marks]

(© IB Organization 2006)

6. The proportion of fish in a lake which are below a certain size can be estimated by catching a random sample of the fish. The random variable X_1 is the number of fish in a sample of size n_1 which are below the specified size.
- (a) Show that $P_1 = \frac{X_1}{n_1}$ is an unbiased estimator of p .
 (b) Find the variance of P_1 .

A further sample of size n_2 is taken and the random variable X_2 is the number of undersized fish in this sample. Define $P_2 = \frac{X_2}{n_2}$.

- (c) Show that $P_T = \frac{1}{2}(P_1 + P_2)$ is also an unbiased estimator of p .
 (d) For what values of $\frac{n_1}{n_2}$ is P_T a more efficient estimator than either
 of P_1 or P_2 ? [15 marks]

7. A discrete random variable, X , takes values 0, 1, 2 with probabilities

$1 - 2\alpha, \alpha, \alpha$ respectively, where α is an unknown constant $0 \leq \alpha \leq \frac{1}{2}$. A random sample of n observations is made of X . Two estimators are proposed for α . The first is $\frac{1}{3}\bar{X}$, and the second is $\frac{1}{2}Y$ where Y is the proportion of observations in the sample which are not equal to 0.

- (a) Show that $\frac{1}{3}\bar{X}$ and $\frac{1}{2}Y$ are both unbiased estimators of α .
 (b) Show that $\frac{1}{2}Y$ is the more efficient estimator. [13 marks]

5 Hypothesis testing

If you toss a coin 100 times and get 50 heads, you cannot say that the coin was biased; equally if 52 or 56 heads were observed, you would still not be suspicious. However, if there were 90 heads you would probably conclude that the coin was biased. So how many heads would be enough to decide that the coin is really biased? This type of question occurs frequently in real situations: a result may not be exactly what you would expect, but with random variation, results rarely are. You have to decide if the evidence is significant enough to change from the default position; this is called a **hypothesis test**.

5A The principle of hypothesis testing

The basic principle of hypothesis testing is ‘innocent until proven guilty beyond reasonable doubt’. We start from a fall-back position which we will accept if there is no significant evidence against it, this is called the **null hypothesis**, H_0 . We will compare this against our suspicion of how things might be, this is called the **alternative hypothesis**, H_1 .

Worked example 5.1

The labels on cans of soup claim that a can contains 350 ml of soup. A consumer believes that on average, they contain less than 350 ml. State the null and alternative hypotheses.



There are two philosophies for using data to make decisions.

Hypothesis testing is one approach but there is increasing support for another method called Bayesian statistics.

Define variables	$\mu = \text{mean amount of soup in a tin}$
Decide which is the conservative position	$H_0 : \mu = 350$
Decide in which direction suspicion lies	$H_1 : \mu < 350$

Generally the null hypothesis is written as an equality while the alternative hypothesis is written as an inequality. If the alternative hypothesis is only looking for a change in one direction ($>$ or $<$) it is called a **one-tailed test**. If the alternative hypothesis is looking for a change in either direction (\neq) it is called a **two-tailed test**.

In this chapter you will learn:

- how to find out if a mean has changed significantly when the variance is known (a Z-test)
- how to find out if a mean has changed significantly when the variance has been estimated (a t-test)
- about the types of error associated with these decisions.

We must now come up with a way of deciding whether or not the information gathered is significant. In the example of tossing 100 coins it is possible that a fair coin comes down heads 90 times by chance. Based upon this outcome we cannot say with certainty that the coin is biased. However, we can say that this outcome is extremely unlikely while the coin is fair. Before performing the hypothesis test you must decide exactly how unlikely an outcome must be to reject H_0 ; this is called the **significance level**.

We can now outline the general procedure for hypothesis testing.

KEY POINT 5.1

1. Write down H_0 and H_1 .
2. Decide on the significance level.
3. Decide what statistic you are going to calculate, called the **test statistic**.
4. Find the distribution of this statistic *assuming that H_0 is true*.
5. Calculate the test statistic from the sample.
6. Decide whether the test statistic is sufficiently unlikely.
7. Determine the outcome of the test and interpret it in the context of the question.

The hardest stage in this process is usually stage 6. This can be done in one of two ways, the **p-value** or the **critical region**, both of which have their advantages:

The *p-value* method involves finding the probability of the observed test statistic, or more extreme, occurring when H_0 is true. So for example, if you were testing against $\mu > 100$ and you observed a mean of 110 you would find the probability of the mean being greater than or equal to 110 rather than just 110. If you were testing against $\mu \neq 100$ and you observed a mean of 110 you would find the probability of the mean being equal to or above 110 *or* equal to or below 90 (as this is the same distance away from the mean in the opposite direction). If this *p-value* is less than the significance level we reject H_0 .

The *critical region* method finds all the values the test statistic could take so that H_0 is rejected: all the values which have a *p-value* less than the significance level. The values which result in H_0 being rejected form the critical or **rejection regions** and they have a total probability equal to the significance level. All other values lie in the **acceptance region**. The boundary between the two regions is called the **critical value**.

EXAM HINT

This is the method used by your GDC. See Calculator skills sheets E, F and I.

 The main importance of this method is that it can be used to calculate the probabilities of different types of error. See Section 5E.

Once we have collected data we can look at what region it falls in and decide what conclusion to make. There are two standard conclusions:

1. reject H_0
2. do not reject H_0 .

In this first example we use the p -value method.

In section 5E we see that we cannot really accept H_0 as we do not know how likely

we are to be drawing a false conclusion. We are controlling the probability of falsely rejecting H_0 and our conclusions should reflect this.

Worked example 5.2

It is believed that the normal level of testosterone in blood is normally distributed with mean 24 nmol/l and standard deviation 6 nmol/l. Following a race a sprinter gives a sample with 34 nmol/l. Is this sufficiently different (at 5% significance) to suggest that the sprinter's sample is being drawn from a population with a different level of blood testosterone?

Define variables

$X = \text{crv 'level of blood testosterone in a sprinter'}$

$$X \sim N(\mu, \sigma^2)$$

Decide which is the conservative position

$$H_0 : \mu = 24$$

Decide in which direction suspicion lies

$$H_1 : \mu \neq 24$$

State distribution of X under H_0

$$\text{Under } H_0, X \sim N(24, 6^2)$$

Find the p -value remembering that it includes everything further away from the mean than 34 in the direction of H_1

$$\begin{aligned} p\text{-value} &= P(X \geq 34) + P(X \leq 14) \\ &= 0.0478 + 0.0478 \\ &= 0.0956 \end{aligned}$$

Draw conclusion

This p -value is greater than 0.05, so we do not reject H_0 . There is not sufficient evidence to suggest that the level is different.

EXAM HINT

Notice that the hypotheses concern the underlying population parameter μ . You do not need to define conventional terms like μ (being the population mean) since it is within the IB's list of accepted notation.

We can also apply the p -value method to one-tailed tests.

Worked example 5.3

According to a geography textbook the average volume of raindrops globally is normally distributed with variance 0.01 ml^2 and mean 0.4 ml . Misha believes that the volume of raindrops in Brazil is significantly larger than the global average. He measures the volume of a raindrop and finds that it is 0.6 ml . Test at the 5% significance level whether or not his suspicion is correct.

Define variables

$$X = \text{crv 'volume of a raindrop in ml'} \\ X \sim N(\mu, 0.01)$$

Decide which is the conservative position

$$H_0 : \mu = 0.4$$

Decide in which direction suspicion lies

$$H_1 : \mu > 0.4$$

Therefore a one-tailed test.

State distribution of X under H_0

$$\text{Under } H_0, X \sim N(0.4, 0.01)$$

Find the p -value remembering that it includes everything further away from the mean than 0.6 in the direction of H_1

$$p\text{-value} = P(X \geq 0.6) \\ = 0.0228$$

Draw conclusion

This p -value is less than 0.05 , so we reject H_0 . There is evidence that Misha's suspicion is correct.

You may prefer to use the critical region method, and some questions may require you to use it.

Worked example 5.4

A machine produces screws which have a mean length of 6 cm and a standard deviation of 0.2 cm . The controls are knocked and it is believed that the mean length may have changed while the standard deviation stays the same. A single screw is measured. Find the critical region at the 5% significance level.

Define variables

$$X = \text{crv 'length of a screw'} \\ X \sim N(\mu, 0.2^2)$$

Decide which is the conservative position

$$H_0 : \mu = 6$$

Decide in which direction suspicion lies

$$H_1 : \mu \neq 6$$

Therefore a two-tailed test.

continued...

State distribution of X under H_0

Find the x -values for the critical region: For a two-tail test with 5% significance level, the probability in each tail is 2.5%

Under H_0 , $X \sim N(6, 0.2^2)$

$P(X < a) = 0.025 \Rightarrow a = 5.61$
 $P(X > b) = 0.025 \Rightarrow b = 6.39$

State the critical region

The critical region is $X < 5.61$ or $X > 6.39$ (3SF)

You may have noticed that the method in Worked example 5.4 was very similar to the method used in confidence intervals. It is indeed the case that the boundaries for a $c\%$ confidence interval correspond to the critical values for a 2-tailed hypothesis test at $(100 - c)\%$ significance. Unfortunately, we cannot apply the methods from confidence intervals to one-tailed tests. However we can still use the critical region method.

Worked example 5.5

The reaction time in catching a falling rod is believed to be normally distributed with mean 0.9 seconds and standard deviation 0.2 seconds. Xinyi believes that her reaction times are faster than this.

- Find the critical region at the 5% significance level to test Xinyi's claim.
- In a test Xinyi catches the rod after 0.6 seconds. State the conclusion to your hypothesis test.

Define variables

(a) $X = \text{crv 'reaction time'}$
 $X \sim N(\mu, 0.2^2)$

Decide which is the conservative position

$H_0 : \mu = 0.9$

Decide in which direction suspicion lies

$H_1 : \mu < 0.9$

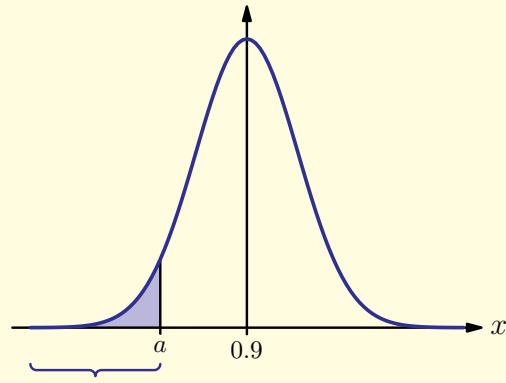
Therefore a one-tailed test.

State distribution of X under H_0

Under H_0 , $X \sim N(0.9, 0.2^2)$

continued ...

Find the x -value associated with 5% significance and one-tailed test



$$P(x < a) = 0.05$$

$$\Rightarrow a = 0.571 \text{ (3SF, from GDC)}$$

The critical region is $x < 0.571$

(b) Observed value falls into acceptance region therefore accept H_0 , there is no significant evidence for Xinyi's claim.

If the observed value is in the critical region, there is evidence to reject this

EXAM HINT

If you are not sure which end to label as the rejection region in a one-tailed test, think about what values would encourage you to accept the alternative hypothesis

Exercise 5A

1. Write null and alternative hypotheses for each of the following situations:
 - (a) (i) The average IQ in a school (μ) over a long period of time has been 102. It is thought that changing the menu in the cafeteria might have an effect upon the average IQ.
(ii) It is claimed that the average size of photos created by a camera (μ) is 1.2 Mb. A computer scientist believes that this figure is inaccurate.
 - (b) (i) A consumer believes that steaks sold in portions of 250 g are on average underweight.
(ii) A careers adviser believes that the average extra amount earned by people with a degree is more than the \$150 000 figure he has been told at a seminar.
 - (c) (i) The mean breaking tension of a brake cable (μ_T) does not normally exceed 3000 N. A new brand claims that it regularly does exceed this value.
(ii) The average time taken to match a fingerprint (μ) is normally more than 28 minutes. A new computer program claims to be able to do better.

- (d) (i) The fraction (p) of toffees in a box of chocolates is advertised as being $\frac{1}{3}$, but Jason thinks that it is more than this.
- (ii) The standard deviation (σ) of measurements of the temperature of meat is thought to have decreased from its previous value of 0.5°C .
2. If it is observed that $x = 10$, find the p -value for each of the following hypotheses, and hence decide the outcome of the hypothesis test at the 5% significance level.
- (a) (i) $X \sim N\left(\mu, \frac{1}{4}\right)$; $H_0 : \mu = 10.8$; $H_1 : \mu \neq 10.8$
(ii) $X \sim N\left(\mu, 5\right)$; $H_0 : \mu = 15$; $H_1 : \mu \neq 15$
- (b) (i) $X \sim N\left(\mu, 7\right)$; $H_0 : \mu = 4$; $H_1 : \mu > 4$
(ii) $X \sim N\left(\mu, 400\right)$; $H_0 : \mu = 40$; $H_1 : \mu < 40$
3. Find the acceptance region for each of the following hypothesis tests when a single value is observed.
- (a) (i) $X \sim N\left(\mu, 5^2\right)$; $H_0 : \mu = 2$; $H_1 : \mu \neq 2$; 5% significance
(ii) $X \sim N\left(\mu, 12^2\right)$; $H_0 : \mu = 16$; $H_1 : \mu \neq 16$; 5% significance
- (b) (i) $X \sim N\left(\mu, 5^2\right)$; $H_0 : \mu = 2$; $H_1 : \mu \neq 2$; 1% significance
(ii) $X \sim N\left(\mu, 12^2\right)$; $H_0 : \mu = 16$; $H_1 : \mu \neq 16$; 10% significance
- (c) (i) $X \sim N\left(\mu, 5^2\right)$; $H_0 : \mu = 2$; $H_1 : \mu > 2$; 5% significance
(ii) $X \sim N\left(\mu, 12^2\right)$; $H_0 : \mu = 16$; $H_1 : \mu > 16$; 5% significance
- (d) (i) $X \sim N\left(\mu, 5^2\right)$; $H_0 : \mu = 2$; $H_1 : \mu < 2$; 5% significance
(ii) $X \sim N\left(\mu, 12^2\right)$; $H_0 : \mu = 16$; $H_1 : \mu < 16$; 5% significance
- (e) (i) $X \sim N\left(\mu, 16\right)$; $H_0 : \mu = -5$; $H_1 : \mu > -5$; 1% significance
(ii) $X \sim N\left(\mu, 100\right)$; $H_0 : \mu = 18$; $H_1 : \mu < 18$; 10% significance
4. The null hypothesis $\mu = 30$ is tested and a value $X = 35$ is observed. Will it have a greater p -value if the alternative hypothesis is $\mu \neq 30$ or $\mu > 30$?
5. A commuter conducts a study and he claims that the average time taken for a train to complete a journey (t) is above 32 minutes. The correct null hypothesis for this is $\mu_t = 32$. What further information would you need before you could write down the *alternative* hypothesis?
-

5B Hypothesis testing for a mean with known variance

One very common and useful parameter whose sample distribution is often known is the mean. If the random variable is normally distributed, or if the mean is taken from a sample large enough to use the CLT then the sample mean follows a normal distribution. More specifically, if $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$ and either of the above conditions is satisfied then

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right). \text{ As long as } \sigma \text{ is known we could either use } \bar{x} \text{ or}$$

the Z -score as the test statistic. This is called a **Z-test**.

If we are given an observed value of \bar{X} we can use it as the test statistic and find the p -value.

EXAM HINT

 If you have the information for the null hypothesis and the observed value of the test statistic then you can use your GDC to perform a Z-test. It will return a p-value which you then need to compare to the significance level. You should state the test statistic and its distribution, as this shows that you have used the correct test. See Calculator skills sheet I.

Worked example 5.6

Standard light bulbs have an average lifetime of 800 hours and a standard deviation of 100 hours. A manufacturer of low energy light bulbs claims that their bulbs' lifetimes have the same standard deviations but that they last longer. A sample of 50 low energy light bulbs have an average lifetime of 829.4 hours. Test the manufacturer's claim at the 5% significance level.

Define variables

$X = \text{crv 'Lifetime of a bulb'}$

$$X \sim N(\mu, 100^2)$$

State hypotheses

$$H_0 : \mu = 800$$

$$H_1 : \mu > 800$$

State the test statistic and its distribution

$$\bar{X} \sim N\left(800, \frac{100^2}{50}\right)$$

Use the calculator to find the p-value

$$p\text{-value} = P(\bar{X} \geq 829.4)$$

$$= 0.0188 \text{ (3SF, from GDC)}$$

Compare to significance level and conclude

$$0.0188 < 0.05$$

Therefore reject H_0 ; there is evidence to support the manufacturer's claim.

We can also find the critical region for a Z-test by using the inverse normal distribution.

Worked example 5.7

The temperature of a water bath is normally distributed with a mean of 60°C and a standard deviation of 1°C . After being serviced it is assumed that the standard deviation is unchanged. The temperature is measured on 5 independent occasions and a test is performed at the 5% significance level to see if the temperature has changed from 60°C . What range of mean temperatures would result in accepting that the temperature has changed?

Define variables

$X = \text{crv 'temperature of water bath'}$

$$X \sim N(\mu, 1)$$

State hypotheses

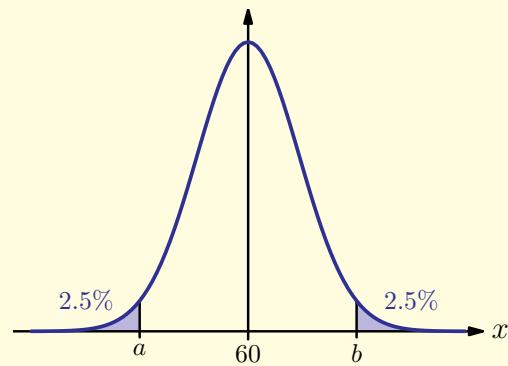
$$H_0 : \mu = 60$$

$$H_1 : \mu \neq 60$$

State test statistic and its distribution

$$\bar{X} \sim N\left(60, \frac{1}{5}\right)$$

Use inverse normal distribution to find the critical values of \bar{X} for the two-tailed region



$$P(\bar{X} < a) = 0.025 \Rightarrow a = 59.1$$

$$P(\bar{X} > b) = 0.025 \Leftrightarrow P(\bar{X} < b = 0.975) \\ \Rightarrow b = 60.9$$

Write down the rejection region

$$\therefore \bar{X} < 59.1 \text{ or } \bar{X} > 60.9$$

Exercise 5B

- In each of the following situations it is believed that $X \sim N(\mu, 100)$. Find the acceptance region in each of the following cases:
 - (i) $H_0 : \mu = 60$; $H_1 : \mu \neq 60$; 5% significance; $n = 16$
 (ii) $H_0 : \mu = 120$; $H_1 : \mu \neq 120$; 10% significance; $n = 30$
 - (i) $H_0 : \mu = 80$; $H_1 : \mu > 80$; 1% significance; $n = 18$
 (ii) $H_0 : \mu = 750$; $H_1 : \mu > 750$; 2% significance; $n = 45$

- (c) (i) $H_0 : \mu = 80.4$; $H_1 : \mu < 80.4$; 10% significance; $n = 120$
(ii) $H_0 : \mu = 93$; $H_1 : \mu < 93$; 5% significance; $n = 400$
2. In each of the following situations it is believed that $X \sim N(\mu, 400)$. Find the p -value of the observed sample mean. Hence decide the result of the test if it is conducted at the 5% significance level:
- (a) (i) $H_0 : \mu = 85$; $H_1 : \mu \neq 85$; $n = 16$; $\bar{x} = 95$
(ii) $H_0 : \mu = 144$; $H_1 : \mu \neq 144$; $n = 40$; $\bar{x} = 150$
 - (b) (i) $H_0 : \mu = 85$; $H_1 : \mu > 85$; $n = 16$; $\bar{x} = 95$
(ii) $H_0 : \mu = 144$; $H_1 : \mu > 144$; $n = 40$; $\bar{x} = 150$
 - (c) (i) $H_0 : \mu = 265$; $H_1 : \mu < 265$; $n = 14$; $\bar{x} = 256.8$
(ii) $H_0 : \mu = 377$; $H_1 : \mu < 377$; $n = 100$; $\bar{x} = 374.9$
 - (d) (i) $H_0 : \mu = 95$; $H_1 : \mu < 95$; $n = 12$; $\bar{x} = 96.4$
(ii) $H_0 : \mu = 184$; $H_1 : \mu > 184$; $n = 50$; $\bar{x} = 183.2$
3. The average height of 18-year-olds in England is 168.8 cm and the standard deviation is 12 cm. Caroline believes that the students in her class are taller than average. To test her belief she measures the heights of 16 students in her class.
- (a) State the hypotheses for Caroline's test.
We can assume that the heights follow a normal distribution and that the standard deviation of heights in Caroline's class is the same as the standard deviation for the whole population.
The students in Caroline's class have an average height of 171.4 cm.
- (b) Test Caroline's belief at the 5% level of significance.
- [6 marks]
4. All students in a large school were given a typing test and it was found that the times taken to type one page of text are normally distributed with mean 10.3 minutes and standard deviation 3.7 minutes. The students were given a month-long typing course and then a random sample of 20 students were asked to take the typing test again. The mean time was 9.2 minutes and we can assume that the standard deviation is unchanged. Test at 10% significance level whether there is evidence that the time the students took to type a page of text had decreased.
- [6 marks]
5. The mean score in Mathematics Higher Level is 4.73 with a standard deviation of 1.21. In a particular school the mean of 50 students is 4.81.
- (a) Assuming that the standard deviation is the same as the whole population, test at the 5% significance level whether the school is producing better results than the international average.
- (b) Does part (a) need the central limit theorem? Justify your answer.
- [8 marks]

6. A farmer knows from experience that the average height of apple trees is 2.7 m with standard deviation 0.7 m. He buys a new orchard and wants to test whether the average height of apple trees is different. He assumes that the standard deviation of heights is still 0.7 m.
- State the hypotheses he should use for his test.
The farmer measures the heights of 45 trees and finds their average.
 - Find the critical region for the test at 10% level of significance.
 - If the average height of the 45 trees is 2.3 m state the conclusion of the hypothesis test. [9 marks]
7. A doctor has a large number of patients starting a new diet in order to lose weight. Before the diet, the weight of the patients was normally distributed with mean 82.4 kg and standard deviation 7.9 kg. The doctor assumes that the diet does not change the standard deviation of the weights. After the patients have been on the diet for a while, the doctor takes a sample of 40 patients and finds their average weight.
- The doctor believes that the average weight of the patients has decreased following the diet. He wishes to test his belief at the 5% level of significance. Find the critical region for this test.
 - Did you use the central limit theorem in your answer to part (a)? Justify your answer.
 - The average weight of the 40 patients after the diet was 78.4 kg. State the conclusion of the test. [11 marks]
8. The school canteen sells coffee in cups claiming to contain 250 ml. It is known that the amount of coffee in a cup is normally distributed with standard deviation 6 ml. Adam believes that on average the cups contain less coffee than claimed. He wishes to test his belief at 5% significance level.
- Adam measures the amount of coffee in 10 randomly chosen cups and finds the average to be 248 ml. Can he conclude that the average amount of coffee in a cup is less than 250 ml?
 - Adam decides to collect a larger sample. He finds the average to be 248 ml again, but this time this is sufficient evidence to conclude at the 1% significance level that the average amount of coffee in a cup is less than 250 ml. What is the minimum sample size he must have used? [12 marks]

5C Hypothesis testing for a mean with unknown variance

In the more realistic situation where we do not know the true population variance, we must use the *T*-score as our test statistic, knowing that it follows a t_{n-1} distribution. This is called a ***t-test***.

See 4D to remind yourself about the *t-distribution*.

EXAM HINT

Your calculator can perform a t -test. You should state the test statistic, its distribution and the p -value from your calculator. If the mean and standard deviation of the sample were not given in the question you should state those too; the calculator will find them in the process of performing the t -test. See Calculator skills sheets E and F.

Worked example 5.8

The label of a pre-packaged steak claims that it has a mass of 250 g. A random sample of 6 steaks is taken and their masses are: 240 g, 256 g, 244 g, 239 g, 245 g, 251 g

Test at the 10% significance level whether the label's claim is accurate, stating any assumptions you need to make.

Define variables

$X = \text{'mass of a steak in g'}$
We assume that $X \sim N(\mu, \sigma^2)$

State hypotheses

$$H_0 : \mu = 250$$
$$H_1 : \mu \neq 250$$

State test statistic and its distribution: use t -test since variance is unknown

$$T = \frac{\bar{X} - 250}{\frac{s_{n-1}}{\sqrt{6}}} \sim t_5$$

Find sample statistics

From GDC:

$$\bar{x} = 245.8$$
$$s_{n-1} = 6.55$$

Find the p -value (use calculator)

$$p\text{-value} = 0.177 \text{ (3SF, from GDC)}$$

Compare to significance level and conclude

$$0.177 > 0.1$$

Therefore do not reject H_0 - there is insufficient evidence to show that the label's claim is inaccurate.

Exercise 5C

- In each of the following situations it is believed that X is normally distributed.
Find the p -value of the observed sample mean. Hence decide the result of the test if it is conducted at the 5% significance level:
 - (i) $H_0 : \mu = 85$; $H_1 : \mu \neq 85$; $n = 30$; $\bar{x} = 92$; $s_{n-1} = 12$
 - (ii) $H_0 : \mu = 122$; $H_1 : \mu \neq 122$; $n = 16$; $\bar{x} = 117$; $s_{n-1} = 8.6$

- (b) (i) $H_0 : \mu = 62$; $H_1 : \mu > 62$; $n = 6$; $\bar{x} = 65$; $s_n = 32.1$
(ii) $H_0 : \mu = 83.4$; $H_1 : \mu < 83.4$; $n = 8$; $\bar{x} = 72$; $s_n = 30.7$
- (c) (i) $H_0 : \mu = 14.7$; $H_1 : \mu \neq 14.7$
Data: 14.7, 14.4, 14.1, 14.2, 15.0, 14.6
(ii) $H_0 : \mu = 79.4$; $H_1 : \mu < 79.4$
Data: 86.4, 79.5, 80.1, 69.9, 75.5

2. John believes that the average time taken for his computer to start is 90 seconds. To test his belief, he records the times (in seconds) taken for the computer to start:
84, 98, 79, 75, 91, 81, 86, 94, 72, 78
(a) State suitable hypotheses.
(b) Test John's belief at the 10% significance level.
(c) Justify your choice of test, including any assumptions required. [8 marks]
3. Michel regularly buys 150 g packets of tea. He has noticed recently that he gets more cups of tea than usual out of one packet, and suspects that the packets contain more than 150 g on average. He weighs eight packets and finds that their mean mass is 153 g and the standard deviation of their masses is 4.2 g.
(a) Find the unbiased estimate of the standard deviation of the masses based on Michel's sample.
(b) Assuming that the masses are normally distributed, test Michel's suspicion at 5% level of significance. [7 marks]
4. The age when 20 babies in a nursery first start to crawl is recorded. The sample has mean 7.1 months and standard deviation 1.2 months. A parenting book claims that the average age for babies crawling is 8 months. Test at the 5% level whether babies in the nursery crawl significantly earlier than average, assuming that the distribution of crawling ages is normal. [7 marks]
5. Ayesha thinks that cleaning the kettle will decrease the amount of time it takes to boil (t). She knows that the average boiling time before cleaning is 48 seconds. After cleaning she boils the kettle 6 times and summarises the results as:

$$\sum t = 283, \sum t^2 = 13369$$

(a) State suitable hypotheses.
(b) Test Ayesha's ideas at the 10% significance level. [8 marks]
6. A national survey of athletics clubs found that the mean time for a 17-year-old athlete to run 100 m is 12.7 s. A coach believes that athletes in his club are faster than average. To test his belief he collects the times for 60 athletes from his club and summarises the results in the following table:

Time (s)	Frequency
11.3–11.7	2
11.7–11.9	3
11.9–12.1	5
12.1–12.3	9
12.3–12.5	12
12.5–12.7	12
12.7–12.9	9
12.9–13.1	5
13.1–13.5	3

- (a) Estimate the mean time for the athletes in the club.
 (b) Find an unbiased estimate of the population standard deviation based on this sample.
 (c) Test the coach's belief at the 5% level of significance.
 (d) Explain how you have used the central limit theorem in your answer to part (c). [10 marks]

7. The lengths of bananas are found to follow a normal distribution with mean 26 cm. Roland has recently changed banana supplier and wants to test whether their mean length is different. He takes a random sample of 12 bananas and obtains the following summary statistics:

$$\sum x = 270, \sum x^2 = 6740$$

- (a) State suitable hypotheses for Roland's test.
 (b) Test at 10% significance level whether the data support the hypothesis that the mean length of Roland's bananas is different from 23 cm.
 (c) Roland's assistant Sabiya suggests that they should test whether the mean length of bananas from the new supplier is *less* than 23 cm.
 (i) State suitable hypotheses for Sabiya's test.
 (ii) Find the outcome of Sabiya's test. [11 marks]

8. Tins of soup claim to contain 300 ml of soup. Aki wants to test if this is an accurate claim. She samples n tins of soup and finds that they have a mean of 303 ml and an unbiased estimate of the population standard deviation of 2 ml.
- (a) State appropriate null and alternative hypotheses.
 (b) For what values of n will Aki reject the null hypothesis at the 5% significance level? [6 marks]
-

5D Paired samples

We often need to ask if a particular factor has a measurable influence. As we saw in Section 4E there are issues associated with studying two different groups so we look for paired samples. If the data come in pairs we can then create another random variable, the difference between the pair, and apply the methods of Sections 5B and 5C to test if the average difference is zero.

We must decide whether to apply a *t*-test or a *Z*-test by checking if the population standard deviation is estimated from the data or given.

Worked example 5.9

The masses of rabbits after a long period of eating only grass is compared to the masses of the same rabbits after a period of eating only ‘Ra-Bites’ pet food. It may be assumed that their masses are normally distributed. The makers of ‘Ra-Bites’ claim that rabbits will get heavier if they eat their food instead of grass. Test this claim at the 10% significance level.

Rabbit	A	B	C	D
Mass on grass diet / kg	2.8	2.6	3.1	3.4
Mass on ‘Ra-Bites’ / kg	3.0	2.8	3.1	3.2

Define variables	$d = \text{mass on ‘Ra-Bites’} - \text{mass on grass}$ $d \sim N(\mu, \sigma^2)$
State hypotheses	$H_0 : \mu_d = 0$ $H_1 : \mu_d > 0$
State test statistic and its distribution: use <i>t</i> -test as the variance is unknown	$T = \frac{\bar{d} - 0}{\frac{s_{n-1}}{\sqrt{4}}} \sim t_3$
Find the differences	Rabbit A B C D d 0.2 0.2 0 -0.2
Find sample statistics	From GDC: $\bar{d} = 0.05$ $s_{n-1} = 0.191$
Find the <i>p</i> -value using calculator	<i>p</i> -value = 0.319 (3SF, from GDC)
Compare to significance level and conclude	$0.319 > 0.1$ Therefore do not reject H_0 - there is insufficient evidence for the maker’s claim.

Exercise 5D

EXAM HINT

You can use your calculator to create a list of differences between the two measurements and then use it to perform the hypothesis test.

1. Test the stated hypotheses at 5% significance. The difference, d , is defined as 'after – before'. You may assume that the data are normally distributed.

(a) $H_0 : \mu_d = 0; H_1 : \mu_d > 0$

Subject	A	B	C	D	E
Before	16	20	20	16	12
After	18	24	18	16	16

(b): $H_0 : \mu_d = 0; H_1 : \mu_d \neq 0$

Subject	A	B	C	D	E	F
Before	4.2	6.5	9.2	8.1	6.6	7.1
After	5.3	5.5	8.3	9.0	6.1	7.0

2. A tennis coach wants to determine whether a new racquet improves the speed of his pupils' serves (faster serves are considered better). He tests a group of 9 children to discover their service speed with their current racquet and with the new racquet. The results are shown in the table below.

Child	A	B	C	D	E	F	G	H	I
Speed with current racquet	58	68	49	71	80	57	46	57	66
Speed with new racquet	72	81	52	59	75	72	48	62	70

- (a) State appropriate null and alternative hypotheses.
 (b) Test at the 5% significance level whether or not the new racquets increase the service speed, justifying your choice of test.

[8 marks]

3. Reading speed of 12-year-olds is measured at different times of the day. It is known that the differences between the reading speed in the morning and in the evening follow a normal distribution with standard deviation of 80 words per minute. Eight 12-year-old pupils from a particular school are tested and their reading speeds in the morning and in the evening are recorded.

Pupil	A	B	C	D	E	F	G	H
Morning	572	421	348	612	364	817	228	350
Evening	421	482	302	687	403	817	220	341

Test at 5% significance level whether there is evidence that for the pupils in this school, the reading in the morning and in the evening are different. State your hypotheses clearly. [8 marks]

4. It is believed that the second harvest of apples from trees is smaller than the first. Ten trees are sampled and the number of apples in the first and second harvests are recorded.

Tree	A	B	C	D	E	F	G	H	I	J
Apples in first harvest	80	72	45	73	68	53	64	48	81	70
Apples in second harvest	75	74	40	67	60	55	58	36	89	60

Stating the null and alternative hypotheses, carry out an appropriate test at the 5% significance level to decide if the claim is justified. [8 marks]

5. Doctor Tosco claims to have found a diet that will reduce a person's weight, on average, by 5 kg in a month. Doctor Crocci claims that the average weight loss is less than this. Ten people use this diet for a month. Their weights before and after are shown below:

Person	A	B	C	D	E	F	G	H	I	J
Weight before (kg)	82.6	78.8	83.1	69.9	74.2	79.5	80.3	76.2	77.8	84.1
Weight after (kg)	75.8	74.1	79.2	65.6	72.2	73.6	76.7	72.9	75.0	79.9

- (a) State suitable hypotheses to test the doctors' claims.
 (b) Use an appropriate test to analyse these data. State your conclusion at:
 (i) the 1% significance level
 (ii) the 10% significance level.
 (c) What assumption do you have to make about the data? [9 marks]

(© IB Organization 2006)

6. It is known that marks on a Mathematics test follow a normal distribution with standard deviation 12 and marks on an English test follow a normal distribution with standard deviation 9.5.

Random variables M and E are defined as follows:

M = mark on Mathematics test

E = mark on English test

Define $D = E - M$.

- (a) State the distribution of D and find its standard deviation.

Marika believes that students at her college get higher marks on average in the English test. The marks of seven students from the college are shown in the table:

Student	P	Q	R	S	T	U	V
Maths	72	61	45	98	82	53	58
English	72	55	55	97	95	72	61

- (b) State suitable null and alternative hypotheses to test whether Marika's belief is justified.
 (c) State your conclusion at the 5% level of significance.

[9 marks]

5E Errors in hypothesis testing

The acceptable conclusions to a hypothesis test are:

1. Sufficient evidence to reject H_0 at the significance level.
2. Insufficient evidence to reject H_0 at the significance level.

It is always possible that these conclusions are wrong.

If the first conclusion is wrong, that is we have rejected H_0 while it was true, it is called a **type I error**.

If the second conclusion is wrong, that is we have failed to reject H_0 when we should have done, it is called a **type II error**.

We cannot eliminate these errors, but we can find the probability that they occur. For a type I error to occur the test statistic must fall within the rejection region while H_0 was true. But we set up the rejection (critical) region to fix this probability as the significance level.

KEY POINT 5.2

When a test statistic follows a continuous distribution the probability of a type I error is equal to the significance level.

If the test statistic follows a discrete distribution, we may not be able to find a critical region so that its probability is exactly equal to the required significance level. In that case the probability of a type I error will be less than or equal to the stated significance level.

KEY POINT 5.3

$$P(\text{type I error}) = P(\text{rejecting } H_0 \mid H_0 \text{ is true})$$

Worked example 5.10

$X \sim \text{Geo}(p)$ and the following hypotheses are tested:

$$H_0 : p = \frac{1}{3}$$

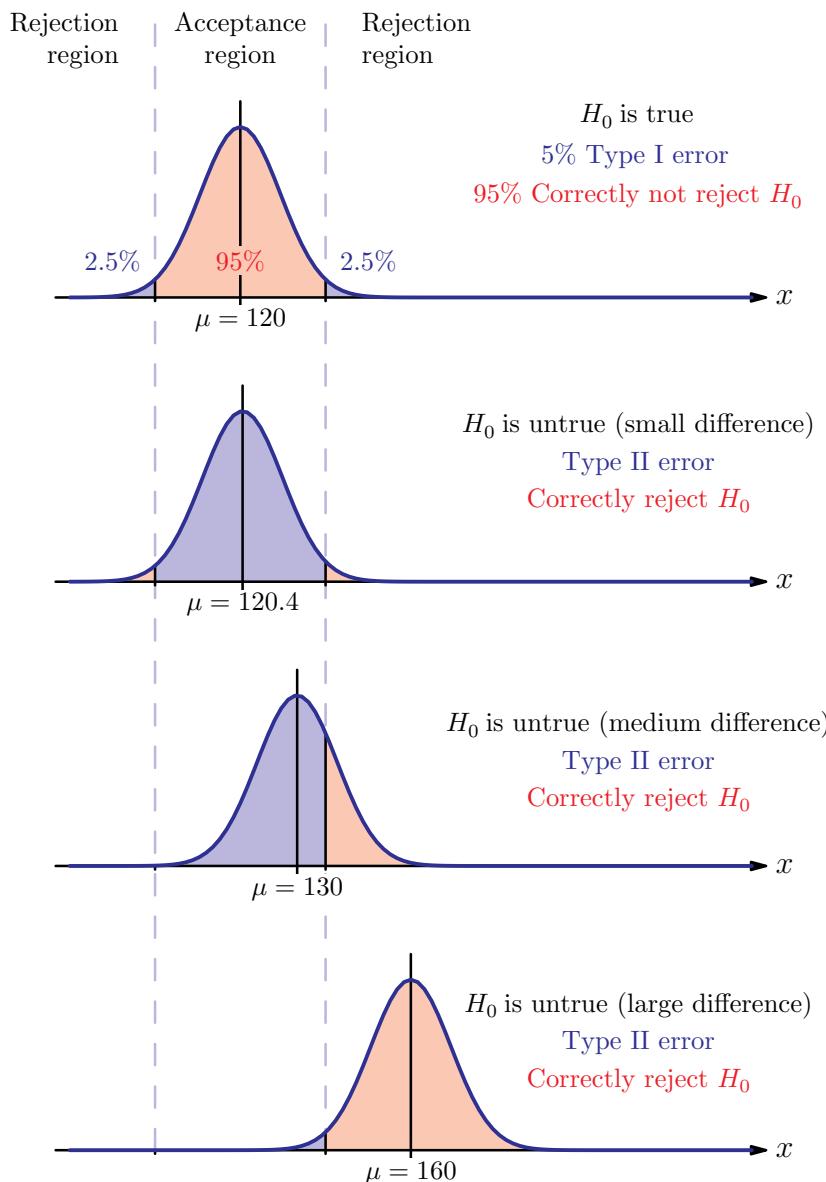
$$H_1 : p < \frac{1}{3}$$

The null hypothesis is rejected if $X \geq 7$. Find the probability of a type I error in this test.

Use definition of type I error

$$\begin{aligned} P\left(\text{rejecting } H_0 \mid p = \frac{1}{3}\right) \\ = P\left(X \geq 7 \text{ where } X \sim \text{Geo}\left(\frac{1}{3}\right)\right) \\ = \left(\frac{2}{3}\right)^6 \\ = 8.78\% \end{aligned}$$

If the true mean is *anything other* than that suggested by the null hypothesis and we have *not* rejected the null hypothesis then we have made a type II error. The probability of a type II error depends upon the true value the population mean takes. Suppose we are testing the null hypothesis $\mu = 120$ with a standard deviation of 10. If the true mean were 160 we would expect to be able to detect this very easily. If the true mean were 120.4 we might have greater difficulty distinguishing this from 120. If we knew the true mean then we could find the probability of an observation of this distribution falling in the acceptance region for H_0 . In the diagrams below, the red regions are where the conclusion is correct, and the blue regions represent errors.



KEY POINT 5.4

$$P(\text{type II error}) = P(\text{not rejecting } H_0 \mid \text{a specific alternative to } H_0)$$

Worked example 5.11

Internet speeds to a particular house are normally distributed with a standard deviation of 0.4 Mbps. The internet provider claims that the average speed of an internet connection has increased above its long term value of 9 Mbps. A sample is taken on 6 occasions and a hypothesis test is conducted at the 1% significance level. Find the probability of a type II error if the true average speed is 9.6 Mbps.

Define variables

$X = \text{crv 'Speed of internet connection'}$

$$X \sim N(\mu, 0.4^2)$$

State hypotheses

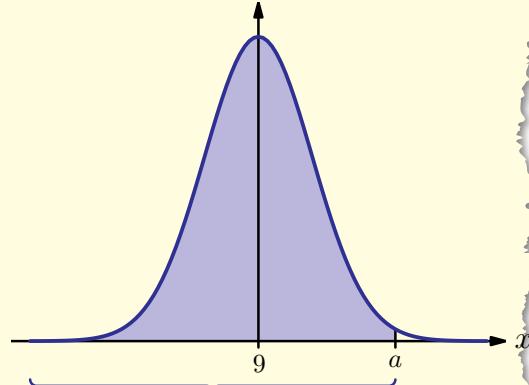
$$H_0 : \mu = 9$$

$$H_1 : \mu > 9$$

State test statistic and its distribution
(assuming H_0 is true)

$$\bar{X} \sim N\left(9, \frac{0.4^2}{6}\right)$$

Decide range of \bar{X} which falls into one-tailed acceptance region



$$P(x < a) = 0.99$$

$$P(\bar{X} < a) = 0.99 \Rightarrow a = 9.38$$

So accept H_0 if $\bar{X} < 9.38$

State the acceptance region

Use the definition of type II error

$$P(\text{type II error}) = P(\bar{X} < 9.38 | \mu = 9.6)$$

$$= P(\bar{X} < 9.38) \text{ where } \bar{X} \sim N\left(9.6, \frac{0.4^2}{6}\right)$$

$$= 8.94\% \text{ (3SF from GDC)}$$

Exercise 5E

1. Find the probability of a type II error for each of the following situations:

(a) (i) $H_0 : X \sim N(\mu, 10^2)$ with $\mu = 84$; $H_1 : \mu \neq 84$;
5% significance; $n = 4$.

In reality, $\mu = 81$

(ii) $H_0 : X \sim N(\mu, 0.4^2)$ with $\mu = 12.6$; $H_1 : \mu \neq 12.6$;
5% significance; $n = 10$.

In reality, $\mu = 12$

(b) (i) $H_0 : X \sim N(\mu, 10^2)$ with $\mu = 84$; $H_1 : \mu \neq 84$;
5% significance; $n = 4$.

In reality, $\mu = 71$

(ii) $H_0 : X \sim N(\mu, 0.4^2)$ with $\mu = 12.6$; $H_1 : \mu \neq 12.6$;
5% significance; $n = 10$.

In reality, $\mu = 12.5$

(c) (i) $H_0 : X \sim N(\mu, 10^2)$ with $\mu = 84$; $H_1 : \mu \neq 84$;
10% significance; $n = 4$.

In reality, $\mu = 81$

(ii) $H_0 : X \sim N(\mu, 0.4^2)$ with $\mu = 12.6$; $H_1 : \mu \neq 12.6$;
10% significance; $n = 10$.

In reality, $\mu = 12$

2. What are the advantages and disadvantages of increasing the significance level of a hypothesis test?

3. John has an eight-sided die and wants to check whether it is biased by looking at the probability, p , of rolling a '4'. He sets up the following hypotheses:

$$H_0 : p = \frac{1}{8}, H_1 : p \neq \frac{1}{8}$$

To test them he decides to roll the die until the first '4' occurs and reject the null hypothesis if the number of rolls is greater than 20 or fewer than 2. Find the probability of making a type I error in John's test. [5 marks]

4. The number of people arriving at a health club follows a Poisson distribution with mean 26 per hour. After a new swimming pool is opened the management want to test whether the number of people visiting the club has increased.

- (a) State suitable null and alternative hypotheses.

They decide to record the number of people arriving at the club during a randomly chosen hour, and to reject the null hypothesis if this number is larger than 52.

- (b) Find the probability of making a type I error in this test.

[6 marks]

5. A long-term study suggests that traffic accidents at a particular junction occur randomly at a constant rate of 3 per week. After new traffic lights are installed it is believed that the number of accidents has decreased. The number of accidents over a 4-week period is recorded.

- Let λ denote the average number of accidents in a 4-week period. State suitable hypotheses involving λ .
- It is decided to reject the null hypothesis if the number of accidents recorded is less than or equal to 7. Find the probability of making a type I error.
- The average number of accidents has in fact decreased to 2.3 per week. Find the probability of making a type II error in the above test. [8 marks]

6. The masses of eggs are known to be normally distributed with standard deviation 7 g. Dhalia wants to test whether eggs produced by her hens have a mass of more than 53 g on average.

- State suitable null and alternative hypotheses to test Dhalia's belief.
- Dhalia weighs 8 eggs and finds that their average mass is 56 g.
- Test at 5% significance level whether Dhalia's eggs weigh more than 53 g on average. State your conclusion clearly.
- Write down the probability of making a type I error in this test.
- What is the smallest average mass of the 8 eggs that would lead Dhalia to reject the null hypothesis?
- Given that the average mass of Dhalia's eggs is actually 51.8 g, find the probability of making a type II error.

[13 marks]

7. A coin is thought to be biased. It is tossed 12 times. The number of tails is denoted by the variable R , and the probability of a tail is given by p .

- State the exact distribution of R and explain why the sampling distribution of R cannot be well approximated by a normal distribution.
- State null and alternative hypotheses in terms of p .
It is required that the probability of a type I error is less than 10%.
- Find the conditions on R for the null hypothesis to be rejected.
- In reality $p = 0.55$. Find the probability of a type II error.
- 8 tails are actually observed. State the conclusion of the test at the 10% significance level.

[14 marks]

- 8.** A population is known to have a normal distribution with a variance of 8 and an unknown mean μ . It is proposed to test the hypotheses $H_0 : \mu = 18$; $H_1 : \mu \neq 18$ using the mean of a sample of size 6.
- Find the appropriate critical regions corresponding to a significance level of:
 - 5%
 - 10%
 - Given that the true population mean is 16.9, calculate the probability of making a type II error when the level of significance is:
 - 5%
 - 10%
- [10 marks]
- 9.** An urn contains 15 marbles, b of which are blue and $(15 - b)$ are red. Peter knows that the value of b is either 5 or 9 but he does not know which. He therefore sets up the hypotheses:
- $$H_0 : b = 5; H_1 : b = 9$$
- To choose which hypothesis to accept, he selects 3 marbles at random without replacement. Let X denote the number of blue marbles selected. He decides to accept H_1 if $X \geq 2$ and to accept H_0 otherwise.
- State the name given to the region $X \geq 2$.
 - Find the probability of making
 - a Type I error;
 - a Type II error.
- [12 marks]
- (© IB Organization 2007)

Summary

- A **hypothesis test** is a way of deciding if observed data are significant.
- A **null hypothesis** (H_0) is the default statement that is accepted if there is no significant evidence against it; it is written as an equality.
- The **alternative hypothesis** (H_1) is a statement that opposes the null hypothesis:
 - it is called a **one-tailed test** when we are looking for a change in one direction ($>$ or $<$).
 - it is called a **two-tailed test** if we are looking for a change in either direction (\neq).
- The **significance level** states how unlikely an outcome must be in order to reject H_0 . If an observed event is less likely than the significance level, it is considered significant.
- The general procedure for hypothesis testing is listed in Key point 5.1.
- To determine if a statistic is sufficiently unlikely, you can use the **p-value** or the **critical region**:
 - the *p*-value is the probability of an observed outcome or more extreme (in the direction of the alternative hypothesis) occurring while the null hypothesis is true. If it is less than the significance level we reject H_0 .

- the critical region (or **rejected region**) is the set of all values of the **test statistic** which would result in the null hypothesis being rejected, i.e. all values with a p -value less than the significance level. They have a total probability equal to the significance level.
- The **Z-test** uses either the \bar{x} or the Z -score as the test statistic and is used for hypothesis testing when we have a mean with a known variance, provided that $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. If we are given an observed value of \bar{X} we can use it as the test statistic and find the p -value; if we use the inverse normal distribution we can also find the critical region for the Z -test.
- The **t-test** uses the T -score as the test statistic knowing that it follows a t_{n-1} distribution and is used for hypothesis testing when the variance for the mean is unknown.
- When we have paired samples, we can create a random variable to be the difference between the pair and then test if the average difference is zero. To determine if you should use a t -test or a Z -test, check if the population standard deviation is unknown or known.
- Hypothesis testing is susceptible to errors:
 - A **type I error** results from rejecting the null hypothesis when it is true:
 $P(\text{type I error}) = P(\text{rejecting } H_0 \mid H_0 \text{ is true})$. When a test statistic follows a continuous distribution the probability of a type I error is equal to the significance level.
 - A **type II error** results from not rejecting the null hypothesis when it is false:
 $P(\text{type II error}) = P(\text{not rejecting } H_0 \mid \text{a specific alternative to } H_0)$.

Mixed examination practice 5

1. A machine cuts carrots into sticks whose length should follow a normal distribution with mean 6.5 cm and standard deviation 0.6 cm. After the machine was cleaned the factory manager wants to check whether the settings are correct. He assumes that the standard deviation of the lengths is unchanged and wants to perform a hypothesis test at the 5% level of significance to determine whether the mean length has changed.

- (a) State suitable null and alternative hypotheses.

The manager measures the lengths of 50 randomly chosen carrot sticks and finds that the sample mean is 6.3 cm.

- (b) What conclusion should the manager draw? Justify your answer.

[6 marks]

2. IQ test scores are assumed to follow a normal distribution with unknown variance. Fifteen students from a particular school are tested and obtained the following scores.

113	102	138	96	106
102	121	113	135	117
125	98	132	115	127

Test at the 5% significance level whether the mean IQ score in this school is higher than 110. State clearly your hypotheses, the type of test used, and the conclusion of your test.

[8 marks]

3. Ten children in a class are given two puzzles to complete. The time taken by each child to solve the puzzles was recorded as follows.

Child	A	B	C	D	E	F	G	H	I	J
Puzzle 1 (min)	10.2	12.9	9.6	14.8	14.3	11.4	10.7	8.3	10.3	10.9
Puzzle 2 (min)	9.7	13.2	8.9	13.6	16.3	12.4	12.5	7.9	10.8	10.6

- (a) For each child, calculate the time taken to solve Puzzle 2 minus the time taken to solve Puzzle 1.
- (b) The teacher believes that both puzzles take, on average, the same time. He believes that the times follow a normal distribution.
- (i) State hypotheses to test this belief.
- (ii) Carry out an appropriate t -test at the 5% significance level and state your conclusion in the context of the problem.

[10 marks]

4. FizzyWhiz is a new drink sold in cans claiming to contain 330 ml. Barbara suspects that the cans contain less drink on average. She tests 30 cans and finds that the mean amount of drink is 326 ml and that the standard deviation of her sample is 6 ml. It can be assumed that the amount of drink in a can is normally distributed.
- Calculate an unbiased estimate of the population standard deviation.
 - Test Barbara's suspicion at the 5% significance level. [7 marks]

5. A machine packs bags of potatoes so that the masses of the bags are normally distributed with mean μ kg and standard deviation 0.15 kg. Initially the machine is adjusted so that $\mu = 2.5$. In order to check that the value of μ has not been changed during a cleaning process, a random sample of 20 bags is selected and their mean mass, \bar{w} , is calculated. The hypotheses used are

$$H_0 : \mu = 2.5, H_1 : \mu \neq 2.5$$

and the critical region is defined to be $\bar{w} < 2.4 \cup \bar{w} > 2.6$.

- Find the significance level of this hypothesis test.
- Given that the actual value of μ is 2.62, find the probability of type II error.

[7 marks]

6. The masses (in kg) of people before and after a new diet are measured.

Person	A	B	C	D	E	F	G	H	I	J
Before	67.3	72.9	61.4	69.6	65.0	84.9	78.7	72.6	70.4	74.2
After	64.8	72.6	59.9	68.4	66.0	83.4	77.8	71.9	69.9	75.0

Apply a hypothesis test at the 5% significance level to check whether the mass has decreased. [9 marks]

7. The random variable X follows a normal distribution with mean μ and standard deviation 4.5. In order to test the null hypothesis $H_0 : \mu = 15$ against the alternative hypothesis $H_1 : \mu \neq 15$, nine observations of X are taken. The mean of this sample is a . Find an expression for the p -value in terms of $\Phi(f(a))$ where $f(a)$ is a function of a .

[5 marks]

6 Bivariate distributions

There are many situations where you measure two quantities for each object of interest. Maybe you want to know the age and weight of a group of teachers, or the marks in Paper 1 and Paper 3 of a group of Higher Level Maths students. We describe such situations as bivariate because there are two variables being measured. Once we have this bivariate data there are usually two important questions to ask:

1. Is there any association between the two variables?
2. If I know the value of one of the variables, can I predict the value of the other variable?

In this chapter you will learn:

- how to describe probability distributions of two linked variables
- how to measure the strength of a relationship between two variables
- how to fit statistical data to a linear model.

6A Introduction to discrete bivariate distributions

A probability distribution is a list of all possible outcomes along with their probabilities. For a bivariate distribution this is easiest to present in a table. An example might be the number of bedrooms in a house on a certain estate and the number of bathrooms in the house. For every different combination we can write down the probability (see table alongside).

Notice that the sum of all of the probabilities in the table is 1.

		Bathrooms	
		1	2
Bedrooms	1	0.1	0
	2	0.1	0.05
	3	0.4	0.1
	4	0.15	0.1

Worked example 6.1

For the data above:

- Find the probability of a randomly selected house having three or more bedrooms.
- Find the probability of having two or more bathrooms given that you have three or more bedrooms.

Combine all the situations with three or more bedrooms

Write the conditional probability required precisely and apply the formula

$$(a) P(3 \text{ or more bedrooms}) = 0.4 + 0.1 + 0.1 + 0.15 = 0.75$$

$$(b) P(2 \text{ bathrooms and } \geq 3 \text{ bedrooms} | \geq 3 \text{ bedrooms}) = \frac{P(2 \text{ bathrooms and } \geq 3 \text{ bedrooms})}{P(\geq 3 \text{ bedrooms})}$$

continued...

Find the probability of the numerator.

$$P(2 \text{ bathrooms and } \geq 3 \text{ bedrooms}) = 0.1 + 0.1 = 0.2$$

$$P(2 \text{ bathrooms and } \geq 3 \text{ bedrooms} | \geq 3 \text{ bedrooms}) = \frac{0.2}{0.75}$$

$$= \frac{4}{15}$$

Expectation and variance of each variable can be found using the formulae given in the Formula booklet. The expectation of XY is found using a similar formula.

Worked example 6.2

For the data above find:

- The expected number of bedrooms.
- The expected value of (number of bedrooms) \times (number of bathrooms).

Write down the distributions of bedrooms

Bedrooms	1	2	3	4
P	0.1	0.15	0.5	0.25

$$E(\text{Bedrooms}) = 1 \times 0.1 + 2 \times 0.15 + 3 \times 0.5 + 4 \times 0.25 = 2.9$$

Calculate bedrooms \times bathrooms for each cell in the table

$$(b) E(\text{Bedrooms} \times \text{Bathrooms}) = \\ 1 \times 1 \times 0.1 + 2 \times 1 \times 0.15 + 3 \times 1 \times 0.4 + 4 \times 1 \times 0.15 \\ + 1 \times 2 \times 0 + 2 \times 2 \times 0.05 + 3 \times 2 \times 0.1 + 4 \times 2 \times 0.1 \\ = 3.7$$

Exercise 6A

1. Find the value of k , $E(X)$, $E(Y)$ and $E(XY)$ in the following distributions:

(a) (i)

		X	
		-1	2
Y	4	0.1	k
	7	0.1	0.2

(ii)

		X	
		0	2
Y	1	0.3	k
	2	0.1	0.2
	3	0.05	0.15

(b) (i)

		X	
		1	2
Y	1	k	$2k$
	2	$3k$	$4k$

(ii)

		X	
		1	2
Y	1	0.1	k
	2	k	0.3

2. For the following distribution find $P(A = 1 | B = 2)$.

		A	
		1	2
B	1	0.15	0.25
	2	0.35	k

[4 marks]

3. A couple decide to keep having children until they have one child of each sex or they have three children, whichever occurs first. G is the random variable 'number of girls' and B is the random variable 'number of boys'.

Assuming they have no multiple births (twins, etc.) and that the probability for each sex is 0.5 for each birth, write down the joint probability distribution of B and G . Hence find $E(G)$.

[7 marks]

4. For the following distribution find $E(A | B = 2)$.

		A	
		1	2
B	1	0.3	0.1
	2	0.1	0.5

[5 marks]

5. A drawer contains three red socks, four blue socks and five green socks. Two socks are drawn at random without replacement. R is the discrete random variable 'number of red socks drawn' and G is the discrete random variable 'number of green socks drawn'.

- (a) Write down the joint probability distribution of R and G .
- (b) Find $P(G = 1 | R = 0)$.
- (c) Find $E(RG)$.
- (d) Show that $E(RG) \neq E(R)E(G)$.

[12 marks]

6. The discrete random variable X can take values 0 or 2. The discrete random variable Y can take the values 1 or 3. Their distribution is defined by:

$$P(X = x \cap Y = y) = k(x + y)$$

- (a) Find $\text{Var}(X)$.
- (b) Find $E(XY)$.

[8 marks]

7. The random variables X and Y can only take the values 0 and 1. $E(X) = 0.4$, $E(Y) = 0.3$ and $E(XY) = 0.1$. Find $P(X = 0 \cap Y = 0)$.

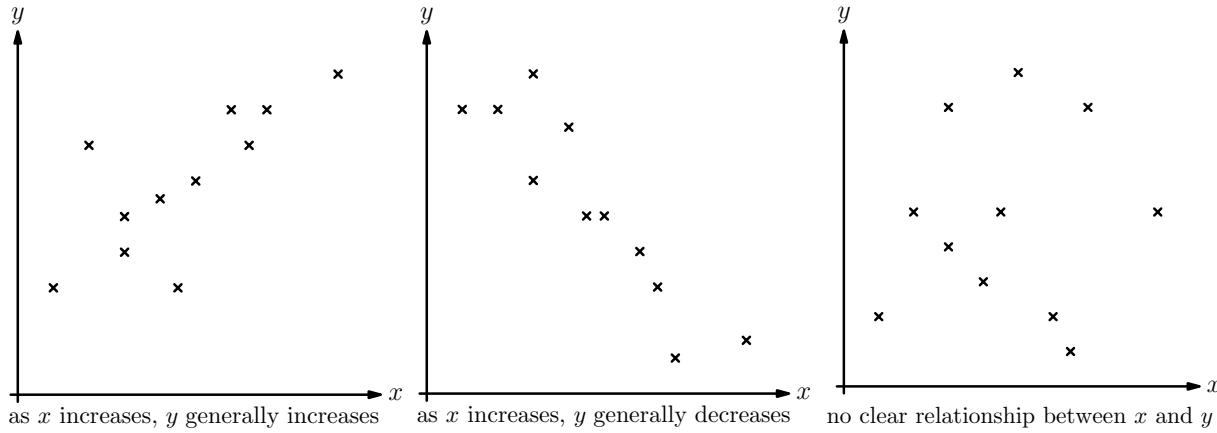
[8 marks]

6B Covariance and correlation

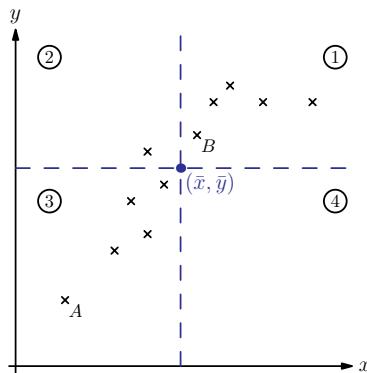
When we have two random variables, their relationship might be independent: when we know one variable it gives us no information about the other one. For example, the IQ and the house number of a randomly chosen person. Alternatively, they may be in a fixed relationship: when we know one variable we know exactly what the other one will be. For example, the length of a side of a cube and the volume of the cube.

In statistics, the relationship is usually somewhere in between, so that if we know one value we can make a better guess at the value of the other variable, but not be absolutely certain: for example, mark in Paper 1 of Maths Higher Level and mark in Paper 2. We use correlation to describe how well the two variables are related.

In this course we focus on linear correlation, which is the extent to which two variables are related by a relationship of the form $Y = mX + c$. If the gradient of the linear relationship is positive we describe the correlation as positive, and if the gradient is negative we describe the correlation as negative. If we have a sample from a bivariate distribution then the relationship is often best illustrated using a scatter diagram.



Rather than simply describing the relationship in words we can find a numerical value to represent the linear correlation. One measure of this is called the **covariance**. The idea behind this comes from splitting a scatter diagram in quadrants around the *mean point*.



If there is a positive linear relationship then we would expect most of the data points to lie in quadrant 1 and quadrant 3. Points lying in these regions should increase our measure of correlation, and points lying in quadrants 2 and 4 should decrease the measure. However, we do not want all points to be treated equally. Point A seems to provide stronger evidence of a positive linear relationship than point B so we would like it to count more. A measure which satisfies this is $\sum (x - \bar{x})(y - \bar{y})$.

In quadrants 1 and 3 both brackets have the same sign, so they make a positive contribution. In quadrants 2 and 4 the brackets have opposite signs so they make a negative contribution.

The average of this measure over n data points is called the

covariance: $\frac{1}{n} \sum (x - \bar{x})(y - \bar{y})$. For a bivariate distribution we can change this into the language of expectation.

KEY POINT 6.1

Covariance formula

$$\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

EXAM HINT

Neither of these definitions is given in the Formula booklet.

An equivalent, but more often used formula for the covariance is given by

KEY POINT 6.2

Alternative covariance formula

$$\text{Cov}(X, Y) = E(XY) - \mu_x \mu_y$$

Worked example 6.3

Prove that $\text{Cov}(X, Y) = E(XY) - \mu_x \mu_y$.

Expand brackets to put into a form where expectation algebra can be used

Use expectation algebra

Since μ_x and μ_y are constant

$E(X) = \mu_x$ and $E(Y) = \mu_y$

$$E[(X - \mu_x)(Y - \mu_y)] = E[XY - \mu_x Y - \mu_y X + \mu_x \mu_y]$$

$$= E(XY) - E(\mu_x Y) - E(\mu_y X) + E(\mu_x \mu_y)$$

$$= E(XY) - \mu_x E(Y) - \mu_y E(X) + \mu_x \mu_y$$

$$= E(XY) - \mu_x \mu_y - \mu_x \mu_y + \mu_x \mu_y$$

$$= E(XY) - \mu_x \mu_y$$

Although this gives a measure of the correlation it is hard to interpret because it depends upon how spread out X and Y are. We can get around this by dividing the covariance by the standard deviations of X and Y . This produces the **correlation coefficient**, ρ (rho).

EXAM HINT

This formula for ρ is not in the Formula booklet.

KEY POINT 6.3

$$\text{Correlation coefficient } \rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

There are many different measures of correlation. This one is often referred to as Pearson's product moment correlation coefficient (PPMCC), developed by Karl Pearson and Sir Francis Galton in the early 20th century. One application of this was an observation that the social standing of the British upper classes was due to a perceived superior genetic makeup. So a mathematical idea led to the field of Eugenics, which supported the sterilisation of those believed to be bad for society.

**Worked example 6.4**

Find the correlation coefficient for the data summarised in this table:

		X	
		4	5
Y	3	0.1	0.3
	7	0.4	0.2

Find the probability distributions of X , then find the expectation and variance

x	4	5
w(X=x)	0.5	0.5

$$E(X) = 4 \times 0.5 + 5 \times 0.5 = 4.5$$

$$E(X^2) = 4^2 \times 0.5 + 5^2 \times 0.5 = 20.5$$

$$\text{Var}(X) = 20.5 - 4.5^2 = 0.25$$

Find the probability distributions of Y , then find the expectation and variance

y	3	7
P(Y=y)	0.4	0.6

$$E(Y) = 3 \times 0.4 + 7 \times 0.6 = 5.4$$

$$E(Y^2) = 3^2 \times 0.4 + 7^2 \times 0.6 = 33$$

$$\text{Var}(Y) = 33 - 5.4^2 = 3.84$$

To find ρ we also need $E(XY)$

$$E(XY) = 3 \times 4 \times 0.1 + 3 \times 5 \times 0.3 + 7 \times 4 \times 0.4 + 7 \times 5 \times 0.2 \\ = 23.9$$

Find ρ

$$\rho = \frac{23.9 - 4.5 \times 5.4}{\sqrt{0.25 \times 3.84}} \\ = -0.408$$

ρ varies between -1 and 1. We can show that when the two variables are independent $\rho = 0$ and when they are linearly related $\rho = \pm 1$.

Worked example 6.5

- Prove that if X and Y are independent random variables then $\rho = 0$.
- Prove that if $Y = mX + c$ then $\rho = \pm 1$.

Think about what facts you can use when the variables are independent

Use expectation algebra on Y

Use expectation algebra on XY

Substitute into the formula for ρ and simplify

Recognise the numerator as $\text{Var}(X)$ and remember that

$$\sqrt{m^2} = |m|$$

(a) If the variables are independent

$$\begin{aligned} E(XY) &= E(X)E(Y) = \mu_x\mu_y \\ \text{Cov}(X, Y) &= E(XY) - \mu_x\mu_y = \mu_x\mu_y - \mu_x\mu_y = 0 \\ \therefore \rho &= 0 \end{aligned}$$

(b) If $Y = mX + c$ then $E(Y) = mE(X) + c$

$$\begin{aligned} \therefore \mu_y &= m\mu_x + c \\ \text{Var}(Y) &= m^2\text{Var}(X) \end{aligned}$$

$$\begin{aligned} E(XY) &= E(mX^2 + cX) \\ &= mE(X^2) + cE(X) \\ &= mE(X^2) + c\mu_x \end{aligned}$$

$$\begin{aligned} \rho &= \frac{mE(X^2) + c\mu_x - \mu_x(m\mu_x + c)}{\sqrt{\text{Var}(X) \times m^2\text{Var}(X)}} \\ &= \frac{mE(X^2) + c\mu_x - m\mu_x^2 - c\mu_x}{\sqrt{m^2[\text{Var}(X)]^2}} \\ &= \frac{m(E(X^2) - \mu_x^2)}{\sqrt{m^2\text{Var}(X)}} \\ &= \frac{m\text{Var}(X)}{|m|\text{Var}(X)} \\ &= \frac{m}{|m|} \\ &= 1 \text{ if } m \text{ is positive or } -1 \text{ if } m \text{ is negative} \end{aligned}$$

In practice we can only estimate ρ for the population by calculating the correlation coefficient of a sample drawn from the bivariate distribution. An unbiased estimate is given by r .

KEY POINT 6.4

An estimate for ρ , the sample product moment correlation coefficient:

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

Worked example 6.6

Find r for the following data:

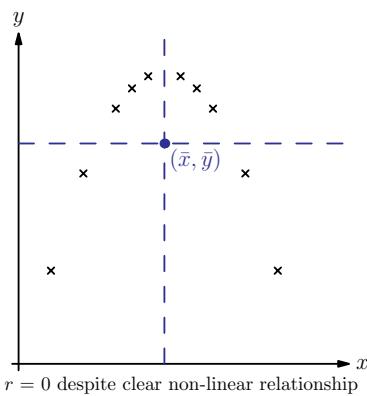
x	y
1	4
3	2
7	9
7	10

Find required sample statistics

$$\begin{aligned}\bar{x} &= 4.5 \\ \bar{y} &= 6.25 \\ \sum x^2 &= 108 \\ \sum y^2 &= 201 \\ \sum xy &= 143 \\ n &= 4 \\ r &= \frac{143 - 4 \times 4.5 \times 6.25}{\sqrt{(108 - 4 \times 4.5^2)(201 - 4 \times 6.25^2)}} \\ &= 0.877 (3SF)\end{aligned}$$

We can use this value of r as an approximate measure of the correlation between the two variables.

Value of r	Interpretation
$r \approx 1$	Strong positive linear correlation
$r \approx 0$	No linear correlation
$r \approx -1$	Strong negative linear correlation



Just because $r = 0$ does not mean that there is no relationship between the two variables, just that there is no *linear* relationship. The graph alongside shows data which have a correlation coefficient of zero, but there is clearly a relationship (the points shown actually lie on a quadratic curve).

The crucial question is how large the correlation coefficient must be before we can say that there is significant correlation between the two variables. To answer this we need to know a probability distribution relating to r . Although the proof is beyond the scope of the course you must know the appropriate sample statistic.

KEY POINT 6.5

Test statistic for $H_0 : \rho = 0$

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

Under the null hypothesis that $\rho = 0$ and assuming that both variables follow a normal distribution then this statistic follows a t -distribution with $n - 2$ degrees of freedom.

KEY POINT 6.6

$$\text{If } \rho = 0, r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$$

This allows us to perform a hypothesis test to see if the observed correlation coefficient provides evidence of correlation.

Worked example 6.7

Test at the 10% significance level whether the data from Worked example 6.6 shows significant evidence of correlation.

State the null and alternative hypotheses

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

Find the value of the test statistic

$$T = 0.877 \sqrt{\frac{4-2}{1-0.877^2}} = 2.587$$

Find the p-value

$$\text{Under } H_0, T \sim t_2$$

$$\therefore p\text{-value} = P(T \geq 2.587) + P(T \leq -2.587) \\ = 0.123 \text{ from GDC}$$

p-value > 10% therefore do not reject H_0 , there is not significant evidence of correlation.

Although $r = 0.877$ seems high, with so few data items it is not significant.

Exercise 6B

1. Find $\text{Cov}(X, Y)$ and ρ for the following distributions:

		X	
		1	2
Y	1	0.75	0.1
	2	0.1	0.05

		X		
		1	2	3
Y	1	0.1	0.05	0.15
	2	0.1	0.05	0.05
	3	0.25	0.25	0

2. Find the sample correlation coefficient for the following data:

- (a) (i) $(-2, 3), (0, 0), (2, 1), (3, 5), (4, 2)$
- (ii) $(3, 15), (17, 9), (22, 10), (33, 7)$
- (b) (i) $\Sigma x = 128, \Sigma x^2 = 2166, \Sigma y = 48, \Sigma y^2 = 400, \Sigma xy = 664, n = 8$
- (ii) $\Sigma x = 122, \Sigma x^2 = 2096, \Sigma y = 140, \Sigma y^2 = 2578, \Sigma xy = 2225, n = 8$

3. Test the null hypothesis $\rho = 0$ based upon the observed correlation coefficients:

- (a) (i) $H_1: \rho \neq 0$, 5% significance, $n = 25$, $r = -0.46$
- (ii) $H_1: \rho \neq 0$, 10% significance, $n = 12$, $r = 0.8$
- (b) (i) $H_1: \rho > 0$, 1% significance, $n = 80$, $r = 0.41$
- (ii) $H_1: \rho < 0$, 5% significance, $n = 50$, $r = -0.52$

4. X and Y follow the following distribution:

		X	
		0	2
Y	1	0.1	k
	3	0.5	0.2

- (a) Find k .
- (b) Find $\text{Cov}(X, Y)$.
- (c) Show that $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$.
[14 marks]

5. In a class of 20 children, the IQ (q) and mass in kg (w) are measured. The findings are summarised as:

$$\sum q = 2197, \sum q^2 = 243929, \sum w = 928, \sum w^2 = 43650, \\ \sum qw = 101762$$

Test at the 10% significance level to see if there is evidence of a correlation between IQ and mass in children. [10 marks]

6. The percentage of people with HIV (p) and the literacy rate (l) of 132 countries were studied in 2004. The summary data are given by:

$$\sum p = 463, \sum p^2 = 8067, \sum l = 8067, \sum l^2 = 890640, \sum lp = 33675$$

Perform a test at the 5% significance level to see if this data shows evidence of positive correlation between the percentage of people with HIV and the literacy rate. [10 marks]

7. (a) Show that if X and Y are independent then $\text{Cov}(X, Y) = 0$.
- (b) (i) Evaluate $\text{Cov}(X, Y)$ if the random variables X and Y have the following distribution:

		X		
		1	2	3
Y	1	a	b	a
	2	b	0	b
	3	a	b	a

- (ii) State, with reasons, whether X and Y are independent.
[12 marks]

Are statistics or emotive stories more persuasive when you decide which charities to support?



8. A sample of bivariate data of size n has correlation coefficient $r = 0.75$. What is the smallest value of n for which this shows evidence of positive correlation at the 5% significance level? [5 marks]
9. Prove that $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$. [6 marks]

6C Linear regression

Once we have established that there is a linear relationship, we can then determine the form of that relationship. To do this we use a method called least squares regression.

Supposing that our x -values are absolutely accurate, we can show how far the points are from the line.

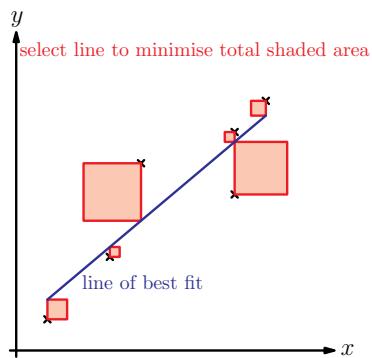
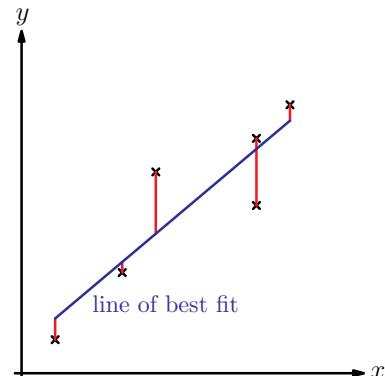
We could try to minimise this total distance by changing the gradient and intercept of the line, but it is actually easier to minimise the area of the *squares* of these distances:

To do so requires some quite advanced calculus. The result is

KEY POINT 6.7

The y -on- x line of best fit is

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



This is called a y -on- x **line of best fit** or a y -on- x **regression line**. The ' y -on- x ' indicates that in the derivation it was assumed that x is known with perfect accuracy and all of the difference between the line and the point is in the y -coordinate. This is usually approximately the case when x is the controlled variable (the variable whose values are predetermined).

If y is the controlled variable then we use an x -on- y line of best fit. This has a slightly different formula.

KEY POINT 6.8

The x -on- y line of best fit is

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

The choice of the appropriate line of best fit depends upon its use.

		Use	
		Estimating y from x	Estimating x from y
Controlled Variable	x	y -on- x	y -on- x
	y	x -on- y	x -on- y
	Neither	y -on- x	x -on- y

There is another expression for the gradient which is easier to remember.

KEY POINT 6.9

Gradient of the y -on- x line of best fit is $\frac{\text{Covariance}}{\text{Variance of } x}$.

Gradient of the x -on- y line of best fit is $\frac{\text{Covariance}}{\text{Variance of } y}$.

Both the y -on- x line of best fit and the x -on- y line of best fit pass through the mean point, the point with coordinates (\bar{x}, \bar{y}) .

Worked example 6.8

Find the y -on- x equation of the regression line of the following data:

$$\Sigma x = 38, \Sigma x^2 = 310, \Sigma y = 54, \Sigma y^2 = 626, \Sigma xy = 436, n = 6$$

Find the mean of x and y

$$\bar{x} = \frac{\sum x}{n} = \frac{38}{6} = \frac{19}{3}$$

$$\bar{y} = \frac{\sum y}{n} = \frac{54}{6} = 9$$

Use the formula

$$y - 9 = \frac{436 - 6 \times \frac{19}{3} \times 9}{310 - 6 \times \left(\frac{19}{3}\right)^2} \left(x - \frac{19}{3}\right)$$

$$= 1.356 \left(x - \frac{19}{3}\right)$$

$$y = 1.356x + 0.413$$

Once you have the line of best fit you can use it to estimate unknown values. However it is important that you do not assume that the pattern continues beyond the data that you are working with. Estimating values beyond the data (extrapolating) requires an assumption that any pattern continues, and should be avoided. Remember also that estimating unknown values from data only makes sense if you have already found evidence of linear correlation between the two variables.

Worked example 6.9

In the data from Worked example 6.8, x is the controlled variable. The values of x vary from 1 to 11. The values of y vary from 4 to 16.

- Assuming that there is significant linear correlation, estimate the value of y when $x = 6$.
- Explain why the regression line should not be used to estimate the value of y when $x = 15$.

(a) When $x = 6$:

$$y = 1.356x + 0.413 \\ = 8.549$$

(b) This would be an extrapolation from the domain of x that has been measured.

Exercise 6C

- Find the required line of best fit for the following data:
 - (i) $(2, -5), (0, 3), (8, 12), (5, 19), (4, 10), (10, 24)$; y -on- x line.
(ii) $(22, 53), (25, 40), (32, 33), (29, 36), (32, 30), (37, 22)$; y -on- x line.
 - (i) $(2, -5), (0, 3), (8, 12), (5, 19), (4, 10), (10, 24)$; x -on- y line.
(ii) $(22, 53), (25, 40), (32, 33), (29, 36), (32, 30), (37, 22)$; x -on- y line.
 - (i) $\Sigma x = 1391, \Sigma x^2 = 81457, \Sigma y = 2174, \Sigma y^2 = 207940$
 $\Sigma xy = 128058, n = 24$; x -on- y line.
(ii) $\Sigma x = 24, \Sigma x^2 = 832, \Sigma y = 30, \Sigma y^2 = 792, \Sigma xy = -462$
 $n = 32$; x -on- y line.
 - (i) $\Sigma d = 112, \Sigma d^2 = 669, \Sigma t = -35, \Sigma t^2 = 153,$
 $\Sigma dt = -222, n = 20$; d -on- t line
(ii) $\Sigma a = 143, \Sigma a^2 = 862, \Sigma b = 186, \Sigma b^2 = 144,$
 $\Sigma ab = 183, n = 25$; b -on- a line.

EXAM HINT

Calculator skills sheet J explains how to use your GDC to find r and the regression line from raw data.

- Use the following data to estimate the value of A when $h = 5$, given that this is within the range of the h values observed and that neither variable is controlled.

$$\Sigma h = 142, \Sigma h^2 = 851, \Sigma A = 247, \Sigma A^2 = 3783, \Sigma Ah = 1800,$$

$$n = 25$$

[5 marks]

3. For a data sample both the x -on- y and y -on- x regression lines are found. Their equations are:

$$x\text{-on-}y: y = 1.2x + 4$$

$$y\text{-on-}x: y = x + 4.6$$

Find the value of \bar{x} . [4 marks]

4. The IQ (q) and results in a maths test (t) for a class of students are summarised below:

$$\Sigma q = 2197, \Sigma q^2 = 243929, \Sigma t = 1124, \Sigma t^2 = 65628,$$

$$\Sigma qt = 125203, n = 20$$

- (a) Find the correlation coefficient and show that there is evidence of significant correlation at the 5% level.
 (b) The test results vary from 36 to 75. Using an appropriate regression line, estimate the IQ of a child who scores 60 in the maths test. [14 marks]

d (m)	v (km/h)
10	12.3
20	17.6
30	21.4
40	23.4
50	25.7
60	26.3

5. The data alongside show the average speeds of cars (v) passing points at 10 m intervals after a junction (d):

- (a) Show at the 10% significance level that there is significant correlation between the average speed and the distance after the junction.
 (b) State, with a reason, which is the controlled variable.
 (c) Using appropriate regression lines find the value of
 (i) d when $v = 20$
 (ii) v when $d = 45$
 (d) Explain why you cannot use your regression line to accurately estimate v when $d = 80$. [13 marks]

6. The data alongside show the height of a cake (h) when baked at different temperatures (T):

- (a) Test at the 5% significance level whether this shows significant evidence of linear correlation between the height of a cake and its baking temperature.
 (b) Find the T -on- h regression line for these data.
 (c) State three reasons why it would be inappropriate to use the regression line found in part (b) to estimate the temperature required to get a cake of height 20 cm. [12 marks]

7. Which of the following statements are true for the bivariate distribution connecting X and Y ?

- (a) If $\rho = 0$ there is no relationship between the two variables.
 (b) If $Y = kX$ then $\rho = 1$.
 (c) If $\rho < 0$ then the gradient of the line of best fit is negative.
 (d) As ρ increases then so does the gradient of the line of best fit.
 (e) If $\rho \approx \pm 1$ then there is a small difference between the y -on- x line of best fit and the x -on- y line of best fit.

8. Data from an experiment are given in the table:

(a) Find the correlation coefficient between:

- (i) y and x
- (ii) y and x^2

(b) Use least squares regression to find a model for the data of the form $y = kx^2 + c$. [6 marks]

x	y
-5	25
7	52
-6	35
-8	62
-4	13
-9	89
0	-3
-6	38

Summary

- Bivariate data is where two variables are being measured for each object of interest, and the data is said to be paired, e.g. age and height.
- The probability distribution (all possible outcomes and their probabilities) for a bivariate distribution is most easily presented in a table.
- The expectation and variance of each variable can be found using:

$$E(X) = \sum_x xP(X=x)$$

$$\text{Var}(X) = E(X^2) - [E(X)]^2 \quad \text{where } E(X^2) = \sum x^2 P(X=x)$$

- The **covariance** is a numerical value used to represent a linear correlation, and can be calculated using:
 - $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$
 - $\text{Cov}(X, Y) = E(XY) - \mu_x \mu_y$
 - $\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$
- The population **correlation coefficient**, ρ , can be found using $\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$
- The sample product moment correlation coefficient, r , is an unbiased estimate of ρ :

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

r can be used as an approximate measure of the correlation between the two variables:
 $r \approx 1$ – strong positive; $r \approx 0$ – no linear correlation; $r \approx -1$ – strong negative.

- We can test for significant correlation using the fact that $r\sqrt{\frac{n-2}{1-r^2}}$ follows a t_{n-2} distribution.

- Least squares regression can be used to determine the form of the linear relationship:

- The y -on- x **line of best fit** is $y - \bar{y} = \left(\frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \right) (x - \bar{x})$.

- The gradient of the y -on- x line of best fit is $\frac{\text{Covariance}}{\text{Variance of } x}$.

- The x -on- y line of best fit is $x - \bar{x} = \left(\frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n y_i^2 - n \bar{y}^2} \right) (y - \bar{y})$.

- The gradient of the x -on- y line of best fit is $\frac{\text{Covariance}}{\text{Variance of } y}$.

Mixed examination practice 6

1. The joint probability distribution of X and Y is shown in the table:

		X	
		1	2
Y	1	0.4	k
	3	k	0.3

Find:

- (a) k
- (b) $\text{Var}(X)$
- (c) $\text{Cov}(X, Y)$

[7 marks]

2. An advertising company wish to test the effectiveness of their advertising. They collect the data on the amount of money spent on advertising (x thousand dollars) and the number of website hits over a week (y thousand) for ten of their internet-based clients. The summary statistics are:

$$\sum x = 68, \sum x^2 = 594, \sum y = 33.7, \sum y^2 = 203, \sum xy = 314$$

- (a) Calculate the value of the correlation coefficient between the amount spent on advertising and the number of website hits.
- (b) Find the equation of the regression line of y -on- x .
- (c) Another internet-based company spend \$4850 on advertising. Estimate how many website hits they should expect to get.
- (d) Does your regression line give a good prediction for the number of website hits for a company that does not spend any money on advertising? Explain your answer.

[13 marks]

3. The owner of a shop selling hats and gloves thinks that his sales are higher on colder days. Over a period of time he records the temperature and the value of the goods sold on a random sample of 8 days:

Temperature (°C)	13	5	10	-2	10	7	-5	5
Sales (£)	345	450	370	812	683	380	662	412

- (a) Calculate the correlation coefficient between the two sets of data.
- (b) Test at the 5% significance level whether the shop owner's belief is supported by the data.
- (c) Suggest one other factor that might cause the sales to vary from day to day.

[10 marks]

4. What is the smallest value of r which will provide significant evidence of positive correlation at the 5% significance level when there are 12 data items?

[5 marks]

5. The regression line of y -on- x is given by $y = 2.8x + 3$ and the regression line of x -on- y is $y = 3.1x + 2.4$. Find the ratios:

(a) $\frac{\bar{x}}{\bar{y}}$
 (b) $\frac{\sigma_x}{\sigma_y}$

[9 marks]

6. A shopkeeper keeps a record of the amount of ice cream sold on a summer's day along with the temperature at noon. He repeats this for several days, and gets the following results.

Temperature ($^{\circ}\text{C}$)	Ice creams sold
26	41
29	51
30	72
24	23
23	29
19	12

- (a) Find the sample correlation coefficient and show that there is significant correlation at the 10% significance level.
 (b) By finding the equation of the appropriate regression line estimate the number of ice creams which would be sold if the temperature were 25°C .
 (c) Explain why it would not be appropriate to use the regression line from part (b) to estimate the number of ice creams sold when the temperature is 0°C .

[12 marks]

7. The variance of the random variable X is 17 and the variance of the random variable Y is 5. Find the maximum possible covariance of X and Y . [5 marks]

7 Summary and mixed examination practice

Introductory problem revisited

A school claims that their average IB score is 34 points. In a sample of four students the scores are 31, 31, 30 and 35 points. Does this suggest that the school was exaggerating?

We can now conduct a hypothesis test to answer this question. First we need to state the significance level. In the absence of any other information, 5% is the default level.

The null hypothesis is $\mu = 34$ and as we are looking for evidence of the school exaggerating the alternative hypothesis is $\mu < 34$. Since we do not know the true variance of the underlying distribution we must use a t -test. This requires the assumption that the underlying distribution is normal, and this seems feasible.

The sample statistics are $\bar{x} = 31.75$ and $s_{n-1} = 2.22$.

Under H_0 , this results in a T -value of -2.03 .

Since there are $4 - 1 = 3$ degrees of freedom the probability of such a value or lower occurring is 0.0677 and this is above our significance level so we cannot reject the null hypothesis. There is no significant evidence that the school was exaggerating.

Summary

This option extends the study of probability distributions from the core study, adding the geometric and negative binomial distributions to model different types of situation. We have also looked at tools which could be used to combine distributions together: the probability generating function and the cumulative distribution function.

The main purpose of this option was finding how to use statistics to decide if new information was significant. Most of these questions relied in some way on the normal distribution so we learnt that the central limit theorem can be used to justify the use of normal distributions in many different situations. To find the parameters of these normal distributions we needed to use expectation algebra.

When finding information from a sample it was important that we estimated population parameters in the best possible way. This led to the idea of an unbiased estimator. We moved from estimating the population mean from a sample as a single number to expressing it as a possible range of values with a defined level of confidence. However, when this was applied to a situation where the true variance was not known we needed to use the t -distribution instead of the normal distribution.

A method similar to constructing confidence intervals allowed us to set up hypotheses tests as a way of making decisions about statistical data. We quantify the probability of a type I error in such tests, but in general we do not control the probability of a type II error and need to be aware of this possibility.

Finally we can now use correlation coefficients to test if there is significant evidence of a relationship between two random variables. Once this has been established a line of best fit can be found and used to make predictions about the relationship.

Mixed examination practice 7

1. Two independent random variables have normal distribution, $X \sim N(5, 2^2)$ and $Y \sim N(3, 5^2)$:
 - (a) Find the mean and variance of $X + 2Y$.
 - (b) State the distribution of $X + 2Y$, including any necessary parameters.
 - (c) Find $P(X + 2Y \geq 15)$. [7 marks]

2. The random variable X has a probability generating function:

$$G(t) = 0.3 + kt + 0.1t^2$$
 - (a) Find the value of k .
 - (b) Find the expectation of X .
 - (c) The random variable $Y = X_1 + X_2$, where X_1 and X_2 are two independent observations of X . Find the probability generating function of Y . [5 marks]

3. It is known that the heights of a certain type of rose bush follow a normal distribution with mean 86 cm and standard deviation 11 cm. Larkin thinks that the roses in her garden have the same standard deviation of heights, but are taller on average. She measures the heights of 12 rose bushes in her garden and finds that their average height is 92 cm.
 - (a) State suitable hypotheses to test Larkin's belief.
 - (b) Showing your method clearly, test at the 5% level of significance whether there is evidence that Larkin's roses are taller than average. [7 marks]

4. The masses of bags of sugar are normally distributed with mean 150 g and standard deviation 12 g.
 - (a) Find the probability that a randomly chosen bag of sugar has a mass of more than 160 g.
 - (b) Find the probability that in a box of 20 bags there are at least two that have a mass of more than 160 g.
 - (c) Dario picks out bags of sugar from a large crate at random. What is the probability that he has to pick up exactly 4 bags before he finds one that has a mass of more than 160 g? [8 marks]

5. The random variables X and Y have the following joint probability distribution:

		X	
		1	4
Y	1	0.1	k
	5	k	0.2

- (a) Find the value of k .
- (b) Find the covariance of X and Y . [7 marks]

6. Sara is investigating how long people can hold their breath under water. She has read that the times should be normally distributed with mean 1.6 minutes. She conducts an experiment with 10 people and records the following summary statistics:

$$\sum t = 15.5, \sum t^2 = 26.5$$

- (a) Find the unbiased estimates of the mean and the standard deviation from Sara's data.
 (b) Perform a suitable test at the 10% level of significance to test whether there is evidence that the average time is less than 1.6 minutes. [9 marks]

7. When Angelo runs a 100 m race he knows that his times are normally distributed with mean 11.3 s and standard deviation 0.35 s. Angelo's coach times his run on eight independent occasions. What is the probability that the average of those times is less than 11 s? [5 marks]

8. (a) If $X \sim B(n, p)$, show that the probability generating function is $(q + pt)^n$ where $q = 1 - p$.
 (b) If X_1 and X_2 are independent observations of X , prove that $X_1 + X_2$ also follows a binomial distribution. [7 marks]

9. Five apple seeds were sown at the same time in different concentrations of fertiliser. After six months, the plants were weighed and the results are given in the table:

Fertiliser concentration (g per litre of soil)	Mass (g)
0	307
5	361
10	402
15	460
20	488

- (a) Calculate the sample correlation coefficient.
 (b) Show that there is evidence of positive correlation at the 5% significance level.
 (c) State and justify whether the fertiliser concentration or mass is the controlled variable.
 (d) By finding a suitable regression line, estimate the smallest concentration of fertiliser required to produce a plant with a mass of 420 g.
 (e) Explain why your line should not be used to estimate the mass of a plant when it is sown in 50 g of fertiliser per litre of soil. [13 marks]

10. (a) The random variable Y is such that $E(3Y + 1) = 10$ and $\text{Var}(5 - 2Y) = 1$.

Calculate:

- (i) $E(Y)$
 (ii) $\text{Var}(Y)$
 (iii) $E(Y - 2)$

- (b) Independent random variables P and Q are such that

$$P \sim N(5, 1) \text{ and } Q \sim N(8, 2).$$

The random variable S is defined by $S = 3Q - 4P$.

Calculate $P(S > 5)$.

[10 marks]

- 11.** The random variable X is normally distributed with unknown mean μ and unknown variance σ^2 . A random sample of 12 observations of X was taken and the 95% confidence interval for μ was correctly calculated as $[6.52, 8.16]$.

- (a) Calculate an unbiased estimate for:

- (i) μ
- (ii) σ^2

- (b) The value of μ is thought to be 8.1, so the following hypotheses are defined:

$$H_0 : \mu = 8.1; H_1 : \mu < 8.1$$

- (i) Find the p -value of the observed sample mean.

- (ii) State your conclusion if the significance level is 10%. [8 marks]

- 12.** Leyton has a biased coin with probability p of showing tails. Given that the probability that he has to toss the coin exactly 10 times before he gets 5 tails is 0.05, find the possible values of p . [5 marks]

- 13.** The random variable X has normal distribution with variance 74. X is measured on 35 independent occasions and the sample mean is found to be 136.

- (a) Find a 95% confidence interval for the mean of X .

- (b) Did you need to use the central limit theorem to answer part (a)? Explain your answer. [5 marks]

- 14.** A teacher claims that she has a new method of teaching spelling which will improve students' performance. A group of six students were given a spelling test before and after being taught by this teacher. Their results are shown in the table below:

Student	A	B	C	D	E	F
Score before	62	38	81	67	82	55
Score after	65	48	79	62	63	67

Assuming that the test scores are normally distributed, carry out an appropriate test to find out whether there is evidence, at the 10% level of significance, that the students' scores have improved. You must make your method and conclusion clear. [7 marks]

- 15.** Chen studies 5 adult dogs, all of the same breed.

Their masses in kg are 45, 42, 51, 48, 44.

- (a) Find an unbiased estimate of the population mean of the masses of adult dogs of this breed.

- (b) Find an unbiased estimate of the population variance of the masses of adult dogs of this breed.

- (c) Find a 90% confidence interval for the population mean of the masses of adult dogs of this breed. [6 marks]

- 16.** The scores on a particular intelligence test are known to have mean 100 and variance 900.
- What is the probability that the average score of a random sample of 50 people is above 110?
 - Explain how you used the central limit theorem in your answer to part (a). [5 marks]
- 17.** A continuous random variable X has probability density function
- $$f(x) = \begin{cases} cx, & 0 \leq x \leq 2 \\ \frac{2c}{9}(x-5)^2, & 2 \leq x \leq 5 \\ 0, & \text{otherwise} \end{cases}$$
- Show that $c = \frac{1}{4}$.
 - Find the cumulative distribution function of X .
 - Write down the value of $P(X \leq 2)$.
 - Find the upper quartile of X . [13 marks]
- 18.** The time taken for a chemical reaction to finish (t) is recorded at five different temperatures (T).
- $$\Sigma t = 149, \Sigma t^2 = 5769, \Sigma T = 200, \Sigma T^2 = 9000, \Sigma tT = 4820$$
- Find the sample correlation coefficient.
 - Test at the 10% significance level whether this coefficient shows significant evidence of correlation.
 - Find the t -on- T line of best fit. [10 marks]
- 19.** The probability distribution of the random variables X and Y is shown in the table:
- | | | X | |
|-----|---|------|------|
| | | 0 | 1 |
| Y | 0 | 0.1 | 0.2 |
| | 1 | 0.3 | 0.2 |
| | 2 | 0.05 | 0.15 |
- Find $E(Y|X=0)$.
 - Find the correlation coefficient ρ for this distribution. [10 marks]
- 20.** A sample of size n is drawn from a normally distributed population with standard deviation 4.6. A 90% confidence interval for the mean was correctly calculated to be [12.7, 13.3]. Find:
- The unbiased estimate of the population mean.
 - The value of n . [5 marks]

- 21.** The random variable X can take the values 1 or 2 with $P(X = x) = \frac{x}{3}$.

- (a) M is the median of a sample of three independent observations of x . Show that M forms a biased estimate of the population mean.
 (b) The statistic kM forms an unbiased estimator of the population mean.
 Find the value k . [8 marks]

- 22.** The random variable W is known to be normally distributed with standard deviation 6. The value of W is measured on eight independent occasions and the mean of the eight observations is 16.3.

- (a) Show that a 95% confidence interval for the mean of W is [12.1, 20.5].
 W is measured on a further ten occasions and the following hypotheses are set up:

$$H_0 : \mu = 16.3, H_1 : \mu \neq 16.3$$

A hypothesis test is carried out and the null hypothesis is rejected if the value of the sample mean falls outside the above confidence interval.

- (b) Find the significance level of this test.
 (c) Find the probability of making a type II error if the true mean is:
 (i) at the bottom end of the confidence interval
 (ii) at the top end of the confidence interval. [9 marks]

- 23.** (a) Tamara repeatedly rolls a fair, six-sided die until she gets a six.

- (i) Find the probability that she has to roll more than 3 times.
 (ii) Name the distribution which models the numbers of rolls required before she gets a six.

- (b) Miguel rolls a fair, six-sided die until he gets 5 sixes.
 (i) Find the probability that Raul has to roll the die exactly 15 times.
 (ii) What is the expected number of rolls Raul needs?
 (iii) What is the most likely number of rolls Raul needs?
 (c) State the relationship between the distributions in parts (a) and (b).
 (d) Explain why $NB(40, 0.4)$ can be approximated by a normal distribution, and find the required mean and variance. [14 marks]

- 24.** A continuous random variable T has probability density function:

$$f(t) = \frac{1}{2} \text{ for } 1 < t < 3$$

- (a) Find the cumulative distribution function of T .
 (b) Two independent observations of X are made. Show that the probability that both are less than 2.5 is $\frac{9}{16}$.
 (c) S is the larger of the two independent observations of T . By considering the cumulative distribution function of S show that S has probability density function $g(s) = \frac{s-1}{2}$, $1 < s < 3$.
 (d) The statistic kS forms an unbiased estimator of the maximum value of T .
 Find the value of k . [15 marks]

- 25.** In this question you may use the fact that if the independent random variables X and Y have Poisson distributions with means λ and μ respectively, and $Z = X + Y$ then Z has a Poisson distribution with mean $(\lambda + \mu)$.

- (a) Given that U_1, U_2, \dots, U_n are independent Poisson random variables each having mean m , use mathematical induction to show that $\sum_{r=1}^n U_r$ has a Poisson distribution with mean nm .
- (b) Random variable W has Poisson distribution with mean 30. By writing W as a sum of independent Poisson random variables with mean 1, explain why the distribution of W is approximately normal. [8 marks]

Answers

Chapter 1

Exercise 1A

1. (a) (i) 12 (ii) 24
 (b) (i) 2 (ii) 3
 (c) (i) -4 (ii) -16
 (d) (i) 9 (ii) 1
 (e) (i) -3 (ii) 13
2. (a) (i) 54 (ii) 216
 (b) (i) 1.5 (ii) 3.375
 (c) (i) 6 (ii) 96
 (d) (i) 6 (ii) 6
 (e) (i) 24 (ii) 54
3. (a) $\frac{1}{4}$ (b) $\frac{13}{6}$
 (c) 18 (d) $\frac{1}{8} \ln 5$

Exercise 1B

1. (a) (i) -5, 6 (ii) 3, 6
 (b) (i) 5, 34 (ii) -18, 72
 (c) (i) -2.4, 1.52 (ii) 5/3, 2
 (d) (i) -3, 6 (ii) 8, 8
 (e) (i) -8, 20 (ii) -7, 30
2. Twice the mass of one gerbil and the total weight of two gerbils
3. (a) 12
 (b) 18
 (c) 12
 (d) 24
4. 1044 kg, 27.7 kg
5. 0, 2.45
6. 0, 39.6
7. 50 minutes, 12.0 minutes

8. (a) 0.2
 (b) $E(X) = 2.5$, $\text{Var}(X) = 0.85$
 (c) $E(Y) = 3.5$, $\text{Var}(Y) = 0.85$
 (d) 7.9. The variables are not independent.
 (e) $p = 0.25$

Exercise 1C

1. (a) (i) 5, 0.171 (ii) 6, 0.208
 (b) (i) -4.7, 0.04 (ii) -15.1, 0.0467
 (c) (i) 12, 0.9 (ii) 8, 0.0257

- (d) (i) 21, 0.893 (ii) 14, 0.0427
 (e) (i) 3, 0.15 (ii) 3.6, 0.315
 (f) (i) 6.5, 0.325 (ii) 8.2, 0.547

2. (a) (i) 35, 8.4 (ii) 72, 30
 (b) (i) -94, 16 (ii) -226.5, 10.5
 (c) (i) 120, 90 (ii) 112, 5.04
 (d) (i) 147, 43.75 (ii) 210, 9.6
 (e) (i) 30, 15 (ii) 28.8, 20.16
 (f) (i) 130, 130 (ii) 123, 123

3. mean = 198.8 g, $\sigma = 4.16$ g
4. E = 102 g, Var = 3.70 g
6. 68.3 g
7. (a) $E(X) = \frac{3}{4}$, $\text{Var}(X) = 0.1875$

(b) 000; mean = 0
 001; mean = $\frac{1}{3}$
 010; mean = $\frac{1}{3}$
 100; mean = $\frac{1}{3}$
 011; mean = $\frac{2}{3}$
 101; mean = $\frac{2}{3}$
 110; mean = $\frac{2}{3}$
 111; mean = 1
 (c) $\frac{9}{64}, \frac{27}{64}, \frac{27}{64}$

8. 19

9. 10

Exercise 1D

1. (a) (i) 0.826 (ii) 0.734
 (b) (i) 0.551 (ii) 0.547
 (c) (i) 0.355 (ii) 0.5
 (d) (i) 0.426 (ii) 0.543
 (e) (i) 0.459 (ii) 0.329
 (f) (i) 0.193 (ii) 0.115
2. (a) N(91.3, 16.3)
 (b) 0.0156
3. (a) 0.3 s, 0.721 s
 (b) 0.339
 (c) 0.166
4. (a) 65 cm, 0.005 cm²
 (b) 0.00235

5. (a) 0.252
(b) 0.0175
6. (a) 0.4 kg, 0.223 kg²
(b) 0.198
(c) 0.0209
7. (a) 0.196
(b) 0.0211
8. 0.272 m
9. (a) 0.208
(b) 0.196
10. $\mu = 7.33 \text{ mm}$, $\sigma = 0.525 \text{ mm}$
11. (a) 0.0228
(b) (i) 0.868
(ii) 0.315
(iii) 0.868
(c) 0.691

ANSWER HINT (c)

Only consider the last day.

(d) 0.645

Exercise 1E

1. (a) (i) Cannot say
(ii) Cannot say
(b) (i) N(80, 4)
(ii) N(80, 1)
(c) (i) N(4000, 20000)
(ii) N(12000, 60000)
2. (a) (i) 0.212
(ii) 0.129
(b) (i) Cannot say
(ii) Cannot say
(c) (i) 0.0228
(ii) 0.0555
(d) (i) 0.00234
(ii) 0.0512
3. 0.0352
4. 0.0336
5. 0.00786
6. (a) mean = 2500 g, SD = 1.79 g
(b) 0.995
(c) We could use normal distribution in part (b).
7. (a) 0.0173
(b) The sum of normal variables is normal.
8. 44

Mixed examination practice 1

1. (a) $\mu - 2m$
(b) $\sigma^2 + 4s^2$
(c) $16\sigma^2$
(d) $4\sigma^2$

2. (a) 16 m, 1.71 m²
(b) 0.00113
3. 0.119
4. mean = 20, SD = 4.47
5. 0.00587
6. 0.249
7. (a) -24 kg, 308 kg²
(b) 0.0857
9. 22
10. (a) 0.432
(b) 0.276
11. 12.0 g

Chapter 2

Exercise 2A

1. (a) (i) 0.0658
(ii) 0.0531
(b) (i) 0.763
(ii) 0.963
(c) (i) 0.162
(ii) 0.309
(d) (i) 0.0625
(ii) 0.0419
2. (a) (i) mean = 3, SD = 2.45
(ii) mean = 6.67, SD = 6.15
(b) (i) mean = 12, SD = 11.5
(ii) mean = 3, SD = 2.45

3. (a) 0.144
(b) 2.5
4. (a) mean = 2.5, SD = 3.75
(b) 0.870

6. 1
7. 0.7 or 0.02
8. 0.25
9. (a) 11
(b) 10

Exercise 2B

1. (a) (i) 0.256
(ii) 0.0972
(b) (i) 0
(ii) 0
(c) (i) 0.819
(ii) 0.5
(d) (i) 0.0465
(ii) 0.137
(e) (i) 0.196
(ii) 0.633
2. (a) (i) mean = 2.5, SD = 0.791
(ii) mean = 10, SD = 4.83
(b) (i) mean = n^2 , SD = $n\sqrt{n-1}$
(ii) mean = $2n(2n+1)$, SD = $\sqrt{2n(2n+1)(2n-1)}$
(c) (i) mean = 18, SD = 9.49
(ii) mean = 10, SD = 3.16
3. 0.0436

4. 0.123
 5. (a) mean = 10, variance = 15
 (b) 8
 6. $r = 10p^2$
 7. (a) $y = 4$

ANSWER HINT (a)

This cannot be solved by manipulating

- (b) \$15
 (c) \$4.90

Exercise 2C

1. (a) $0.5t + 0.2t^2 + 0.1t^3 + 0.05t^5 + 0.05t^6$
 (b) $0.3t + 0.3t^2 + 0.3t^3 + 0.1t^4$
 2. (a) (i) 0.4 (ii) 0.4
 (b) (i) 0 (ii) 1
 (c) (i) 0.5 (ii) 0.375
 (d) (i) 0.0337 (ii) 0.4
 3. $E(X) = a, \text{Var}(X) = a$
 4. (a) $\frac{1}{9}$ (b) $E(X) = 0.472, \text{Var}(X) = 1.13$
 6. (a) $\frac{1}{e}$
 (b) 2
 (c) 4
 8. (b) $P(X = n) = \frac{(k-1)}{k^{n+1}}$
 (c) $E(X) = \frac{1}{k-1}, \text{Var}(X) = \frac{k}{(k-1)^2}$

Exercise 2D

1. (a) $0.3 + 0.3t + 0.4t^3$
 (b) $(0.3 + 0.3t + 0.4t^3)^{10}$
 (c) 15
 4. $B(n+m, p)$
 5. (a) $0.3t + 0.7t^4$
 (b) $(0.3t + 0.7t^4)^8$

Mixed examination practice 2

1. (a) 0.148 (b) 0.444
 2. (a) $\frac{1}{6}$ (b) 2
 3. (a) 0.175 (b) 0.0247

4. (a) 0.302
 (b) 0.584
 (c) 6.62
 5. 0.0191
 6. (b) $(0.8 + 0.2t)^{10}$
 (c) $(0.8 + 0.2t)^{10}(0.75 + 0.25t)^{12}$
 (d) 0.500
 7. (a) 0.129 (b) 0.0324
 (c) 0.227 (d) 0.117
 8. (a) 11.25 (b) 0.0553
 (c) 0.0064 (d) 0.615
 9. (a) 0.384 (b) 2.60
 (c) (i) 0.0927 (ii) 11
 10. (c) $\frac{2}{3}$

11. (a) $\frac{1}{e}$
 (c) 5.11×10^{-4}

Chapter 3

Exercise 3A

1. (a) (i) $\frac{x}{5}$ for $x = 1, 2, 3, 4, 5$
 (ii) $\frac{x}{10}$ for $x = 1, 2, 3, \dots, 10$
 (b) (i) $\frac{x-2}{4}$ for $x = 3, 4, 5, 6$
 (ii) $x + \frac{1}{10}$ for $x = 0, 0.1, 0.2, \dots, 0.9$
 2. (a) (i) $2x - x^2$ $0 < x < 1, \frac{2-\sqrt{2}}{2}$
 (ii) $\frac{x^2-4}{32}$ $2 < x < 6, \sqrt{20}$
 (b) (i) $1 - \cos x$ $0 < x < \frac{\pi}{2}, \frac{\pi}{3}$
 (ii) $\frac{\ln x}{\ln 10}$ $1 < x < 10, \sqrt{10}$
 3. (a) (i) 1 $1 \leq x < 2, \frac{3}{2}$ (ii) 3 $0 \leq x < \frac{1}{3}, \frac{1}{6}$
 (b) (i) $2x - 1$ $1 \leq x < \frac{1+\sqrt{5}}{2}, \frac{1+\sqrt{3}}{2}$
 (ii) $\cos x$ $0 \leq x < \frac{\pi}{2}, \frac{\pi}{6}$

4. (a) $\frac{3}{68}$ (b) 8
 5. $e^{0.8}$
 6. (a) $\frac{x(x+1)}{20}$ for $x = 1, 2, 3, 4$
 (b) 3

7. (a) $\frac{x(x+1)-12}{44}$ for $x=4,5,6,7$

(b) 7

8. (a) $\frac{1}{2}\ln 2$

(b) $2e^{2x} \quad 0 \leq x < \frac{1}{2}\ln 2$

(c) $x = \frac{1}{2}\ln\frac{3}{2}$

9. (a) 10

(b) $\frac{x^3 - (x-1)^3}{1000}$ for $x=1, 2, \dots, 10$

10. (a) $\frac{4}{1+4\ln 2}$

(b) $\frac{24}{5+20\ln 2}$

(c) $f(x) = \begin{cases} 0 & x < 0 \\ \frac{x^4}{1+4\ln 2} & 0 \leq x < 1 \\ \frac{1+4\ln x}{1+4\ln 2} & 1 \leq x < 2 \\ 1 & x > 2 \end{cases}$

(d) $\sqrt{2}e^{-\frac{1}{8}}$

(e) $\sqrt[4]{\frac{1}{4} + \ln 2}$

(f) 0.25

Exercise 3B

2. (a) $\frac{x}{10}, 0 < x < 10$

(b) $\frac{4\pi r^2}{10}, 0 < r < \sqrt[3]{\frac{15}{2\pi}}$

3. (a) $\frac{x^3 - 1}{26}, 1 < x < 3$

(b) $\frac{37}{702}$

(c) $\frac{3}{26y^4}, \frac{1}{3} < y < 1$

4. $3y^2 \quad 0 < y < 1$

Mixed examination practice 3

1. $6\left(\frac{y^2}{2} - \frac{y^3}{3}\right) \quad 0 \leq y \leq 1$

2. (a) $\frac{x^2 - 9x + 20}{2} \quad 5 \leq x \leq 6 \quad$ (b) $\frac{9 + \sqrt{5}}{2}$

3. 0.695

4. 0.519

5. $a=0, b=2, c=\frac{1}{8}$

6. (a) $P(3n)$ (b) e^{-3n}
(c) $f(t)=3e^{-3t} \quad t \geq 0$ (d) 0.173

Chapter 4

Exercise 4A

1. (a) 17.02 (b) 3.97

2. (a) 3.86 (b) 17.9

3. (a) 50.0 (b) 0.0288

4. (a) 210 (b) 44 700

5. (a) 18 (b) 7260

6. 49

Exercise 4B

1. (a) $0 : \frac{3}{28}$
 $\frac{1}{2} : \frac{15}{28}$
 $1 : \frac{5}{14}$

2. (a) $\frac{3k}{4}$ (b) $\frac{4X}{3}$ (c) $\frac{28}{3}$

3. (a) 1,1; 1,2; 1,3;
2,1; 2,2; 2,3;
3,1; 3,2; 3,3

(c) $\frac{27}{22}$

6. (a) $T \sim NB(2, p)$

7. (b) $F(x) = \frac{x}{k} \quad 0 \leq x \leq k$

(c) $\frac{m^2}{k^2}$

(d) $g(m) = \frac{2m}{k^2}$

(f) $\frac{3}{2}M : \text{Var}\left(\frac{3}{2}M\right) = \frac{k^2}{8} < \frac{k^2}{6}$
 $= \text{Var}(X_1 + X_2)$

Exercise 4C

1. (a) 1.28 (b) 2.56

2. (a) False (b) False (c) False
(d) True (e) False

3. 90% interval

4. e.g. Range

5. (a) (i) [17.4, 22.6]
(ii) [40.9, 43.3]
(b) (i) [305, 395]
(ii) [-16.5, 12.9]

	\bar{x}	σ	n	Confidence level	Lower bound of interval	Upper bound of interval
(a) (i)	58.6	8.2	4	90	51.9	65.3
(ii)	0.178	0.01	12	80	0.174	0.182
(b) (i)	42	4	4	80	39.44	44.56
(ii)	30.4	1.2	900	99	30.30	30.50
(c) (i)	120	18	64	95	115.59	124.41
(ii)	1100	25	200	88	1097.3	1102.7
(d) (i)	4	40	100	75	-0.601	8.601
(ii)	16	0.4	400	90	15.967	16.033
(e) (i)	8	12	14	98	0.539	15.46
(ii)	0.4	0.01	16	80	0.397	0.403

7. $[85.8, 90.6]$

8. $[3.06, 4.54]$

9. (a) 94.9% (b) 174

10. 9

11. (a) $[165, 171]$

(b) No; large sample means we can use CLT.

12. (a) 106

(b) 73.7%

Exercise 4D

1. (a) (i) 0.0381 (ii) 0.649
 (b) (i) 0.00349 (ii) 0.939
 (c) (i) 0.0176 (ii) 0.578
 (d) (i) 0.927 (ii) 0.0193
3. (a) (i) 0.873 (ii) -1.11
 (b) (i) -0.703 (ii) 0.533
 (c) (i) 0.870 (ii) 0.542

4. $t = 0.711$

Exercise 4E

1. (a) (i) $[12.8, 15.4]$
 (ii) $[189, 193]$
 (b) (i) $[16.0, 20.0]$
 (ii) $[0.0318, 0.0482]$
 (c) (i) $[-0.653, 4.92]$
 (ii) $[-1.33, 3.13]$
 (d) (i) $[-0.937, 14.9]$
 (ii) $[111, 207]$
2. (a) $[-1.64, 4.84]$
 (b) $[-1.56, 1.40]$
3. (a) 14.7 (b) 7.58
 (c) $[14.1, 15.2]$

4. $[-0.0150, 3.52]$

5. (a) Heights are normally distributed.
 (b) 118
 (c) 16.25
 (d) $[110, 126]$

6. (a) 119
 (b) 90%

7. (a) 5.6 hours
 (b) 85%

8. 83.8%

9. (a) 200
 (b) 80%

10. (a) $[-36.3, 39.9]$

11. (c) $[-11.1, 25.1]$

Mixed examination practice 4

1. (a) 73.2 g, 134 g²
 (b) $[66.5, 79.9]$
2. $[22.7, 38.3]$
3. $[16.1, 27.9]$
5. (a) $[11.8, 13.4]$
 (b) This suggests $\mu = 13.7$, which is not consistent.
6. (b) $p \frac{(1-p)}{n_1}$
 (d) $\frac{1}{3} < \frac{n_1}{n_2} < 3$

Chapter 5

Exercise 5A

1. (a) (i) $H_0: \mu = 102, H_1: \mu \neq 102$
 (ii) $H_0: \mu = 1.2, H_1: \mu \neq 1.2$
- (b) (i) $H_0: \mu = 250, H_1: \mu < 250$
 (ii) $H_0: \mu = 150000, H_1: \mu > 150000$
- (c) (i) $H_0: \mu_T = 3000, H_1: \mu_T > 3000$
 (ii) $H_0: \mu_t = 28, H_1: \mu_t < 28$
- (d) (i) $H_0: p = \frac{1}{3}, H_1: p > \frac{1}{3}$
 (ii) $H_0: \sigma = 0.5, H_1: \sigma < 0.5$
2. (a) (i) 0.110, do not reject H_0
 (ii) 0.0253, reject H_0
- (b) (i) 0.0117, reject H_0
 (ii) 0.0668, do not reject H_0
3. (a) (i) $[-7.80, 11.8]$
 (ii) $[-7.52, 39.5]$
- (b) (i) $[-10.9, 14.9]$
 (ii) $[-3.74, 35.7]$
- (c) (i) $x > 10.2$
 (ii) $x > 35.7$
- (d) (i) $x < -6.22$
 (ii) $x < -3.74$
- (e) (i) $x > 4.31$
 (ii) $x < 5.18$
4. $\mu \neq 30$

Exercise 5B

1. (a) (i) $[55.1, 64.9]$
 (ii) $[117, 123]$
- (b) (i) $\bar{x} < 85.5$
 (ii) $\bar{x} < 753$
- (c) (i) $\bar{x} > 79.2$
 (ii) $\bar{x} > 92.2$
2. (a) (i) 0.0455, reject H_0
 (ii) 0.0578, do not reject H_0
- (b) (i) 0.0228 reject H_0
 (ii) 0.0289, reject H_0
- (c) (i) 0.0625, do not reject H_0
 (ii) 0.147, do not reject H_0
- (d) (i) 0.596, do not reject H_0
 (ii) 0.611, do not reject H_0
3. (a) $H_0: \mu = 168.8, H_1: \mu > 168.8$
 (b) $p = 0.193$. Do not reject H_0 ; no evidence for her belief
4. $p = 0.0918$. Reject H_0 ; sufficient evidence that the time has decreased
5. (a) $p = 0.320$. Do not reject H_0 ; no evidence that results are better
 (b) Yes, distribution unknown
6. (a) $H_0: \mu = 2.7, H_1: \mu \neq 2.7$
 (b) $x < 2.53, x > 2.87$
 (c) Reject H_0 ; evidence that height is different

7. (a) $x < 80.3$
 (b) No; distribution is normal
 (c) Evidence that weight has decreased

8. (a) No ($p = 0.146$)
 (b) 49

Exercise 5C

1. (a) (i) 0.00336, reject H_0
 (ii) 0.0345, reject H_0
- (b) (i) 0.421, do not reject H_0
 (ii) 0.179, do not reject H_0
- (c) (i) 0.203, do not reject H_0
 (ii) 0.351, do not reject H_0
2. (a) $H_0: \mu = 90, H_1: \mu \neq 90$
 (b) $p = 0.0456$, reject H_0 ; evidence that John's belief is wrong
- (c) Assumed times are normally distributed
3. (a) 4.49
 (b) $p = 0.0503$, do not reject H_0 ; no evidence for his suspicion
4. $p = 0.002$, evidence that they crawl earlier
5. (a) $H_0: \mu = 48, H_1: \mu < 48$
 (b) $p = 0.812$, no evidence the time has decreased
6. (a) 12.5
 (b) 0.406
 (c) $p = 2.63 \times 10^{-5}$, evidence that the belief is correct
- (d) Sample mean follows normal distribution
7. (a) $H_0: \mu = 26, H_1: \mu \neq 26$
 (b) $p = 0.147$, no evidence they are different
- (c) (i) $H_0: \mu = 26, H_1: \mu < 26$
 (ii) $p = 0.736$, evidence that they are smaller
8. (a) $H_0: \mu = 300; H_1: \mu \neq 300$
 (b) $n \geq 5$

Exercise 5D

1. (a) $p = 0.121$, do not reject H_0
 (b) $p = 0.830$, do not reject H_0
2. (a) $H_0: \mu_d = 0, H_1: \mu_d > 0$
 (b) Do not reject H_0 ($p > 0.0927$)
3. $H_0: \mu_d = 0, H_1: \mu_d \neq 0$. No evidence they are different ($p = 0.863$)
4. $H_0: \mu_d = 0, H_1: \mu_d < 0$. $p = 0.0352$, reject H_0
5. (a) $H_0: \mu_d = 5, H_1: \mu_d < 5$
 (b) $p = 0.0447$ (i) Accept H_0 (ii) Do not accept H_0
 (c) e.g. Normal distribution, independent
6. (a) Normal, $\sigma = 15.3$
 (b) $H_0: d = 0, H_1: d > 0$
 (c) $p = 0.174$, do not reject H_0

Exercise 5E

1. (a) (i) 0.908
 (ii) 0.00269
- (b) (i) 0.261
 (ii) 0.876
- (c) (i) 0.840
 (ii) 0.000971

3. 0.194

4. (a) $H_0: \lambda = 26, H_1: \lambda > 26$
(b) 2.24×10^{-6} 5. (a) $H_0: \lambda = 12, H_1: \lambda < 12$
(b) 0.0895
(c) 0.6996. (a) $H_0: \mu = 53, H_1: \mu > 53$
(b) Do not reject H_0 ($p = 0.113$) – no evidence that Dhalia's eggs are heavier.
(c) 0.05
(d) 57.1
(e) 0.9837. (a) $R \sim B(12, p)$; n is not large enough
(b) $H_0: p = \frac{1}{2}, H_1: p \neq \frac{1}{2}$
(c) $R \leq 2$ or $R \geq 10$
(d) 0.950
(e) Do not reject H_0 – no evidence that the coin is biased8. (a) (i) $x < 15.7, x > 20.3$ (ii) $x < 16.1, x > 19.9$
(b) (i) 0.849 (ii) 0.7519. (a) Critical region
(b) (i) $\frac{22}{91}$ (ii) $\frac{1}{7}$

Mixed examination practice 5

1. (a) $H_0: \mu = 6.5, H_1: \mu \neq 6.5$
(b) Reject H_0 ($p = 0.0184$)2. $H_0: \mu = 110, H_1: \mu > 110$, t -test $p = 0.0540$, no evidence that the IQ is higher.3. (a) $-0.5, 0.3, -0.7, -1.2, 2, 1, 1.8, -0.4, 0.5, -0.3$
(b) (i) $H_0: \mu_d = 0, H_1: \mu_d \neq 0$
(ii) Do not reject H_0 ($p = 0.480$); Teacher is correct.4. (a) 6.10
(b) Reject H_0 ($p = 6 \times 10^{-4}$). There is evidence for Barbara's suspicion.

5. (a) 0.287% (b) 0.275

6. t -test, $p = 0.0288$, evidence that mass has decreased7. $2\Phi\left(-\frac{15-a}{1.5}\right)$

Chapter 6

Exercise 6A

1. (a) (i) 0.6, $E(X) = 1.4, E(Y) = 4.9, E(XY) = 6.5$
(ii) 0.2, $E(X) = 1.1, E(Y) = 1.7, E(XY) = 2.1$
(b) (i) 0.1, $E(X) = 1.6, E(Y) = 1.7, E(XY) = 2.7$
(ii) 0.3, $E(X) = 1.6, E(Y) = 1.6, E(XY) = 2.1$ 2. $\frac{7}{12}$

		B			
		0	1	2	3
		0	0	0	1/8
G		1	0	1/2	1/8
2		0	1/8	0	0
3		1/8	0	0	0

$$E(G) = 1.25$$

4. $\frac{11}{6}$

5. (a)

		R		
		0	1	2
		$\frac{12}{132}$	$\frac{24}{132}$	$\frac{6}{132}$
G		$\frac{40}{132}$	$\frac{30}{132}$	0
1		$\frac{20}{132}$	0	0

(b) $\frac{5}{9}$ (c) $\frac{5}{22}$ (d) $E(R) = \frac{1}{2}, E(G) = \frac{5}{6}$ 6. (a) $\frac{8}{9}$

(b) 3

7. 0.4

Exercise 6B

1. (a) 0.0275, $\rho = 0.216$ (b) $-0.25, \rho = -0.374$ 2. (a) (i) 0.194 (ii) -0.942
(b) (i) -0.905 (ii) -0.518 3. (a) (i) Reject H_0 ($p = 0.0207$)
(ii) Reject H_0 ($p = 0.00178$)
(b) (i) Reject H_0 ($p = 7.94 \times 10^{-5}$)
(ii) Reject H_0 ($p = 5.43 \times 10^{-5}$)

4. (a) 0.2

(b) -0.32 5. $r = -0.145, p = 0.543$, no correlation6. $r = -0.106, p = 0.113$, no correlation7. (b) (i) 0
(ii) Not independent
(unless $b = 0, a = \frac{1}{4}$)

8. 6

Exercise 6C

1. (a) (i) $y = 2.27x - 0.489$ (ii) $y = -1.88x + 91.0$
(b) (i) $y = 0.283x + 1.86$ (ii) $y = -0.501x + 47.4$
(c) (i) $x - 58.0 = 0.187(y - 90.6)$
(ii) $x - 0.75 = -0.634(y - 0.938)$

- (d) (i) $d - 5.6 = -0.283(t + 1.75)$
(ii) $b - 7.44 = -59.0(a - 5.72)$
2. 3.80
3. 3
4. (a) $r = 0.686, p = 0.417 \times 10^{-4}$
(b) 113
5. (a) $r = 0.962, p = 0.00218$
(b) Distance
(c) (i) 31.2 (ii) 23.9
(d) Extrapolation
6. (a) $r = -0.700, p = 0.122$, no correlation
(b) $T = -20.8h + 689$
(c) No linear correlation; extrapolation; h is not the controlled variable.
7. (a) False (b) False (c) True
(d) False (e) False (no difference)
8. (a) (i) -0.328 (ii) 0.996
(b) $y = 1.11x^2 - 3.55$

Mixed examination practice 6

1. (a) 0.15 (b) 0.2475 (c) 0.195
2. (a) 0.782 (b) $y = 0.645x - 1.01$
(c) 2113 (d) No (negative answer)
3. (a) -0.660
(b) $p = 0.0373$, claim is supported
(c) e.g. day of the week
4. 0.497
5. (a) 0.233 (b) 1.05
6. (a) 0.945 (b) 37
(c) Extrapolation
7. 9.22

Chapter 7

Mixed examination practice 7

1. (a) 11, 104 (b) $N(11, 104)$ (c) 0.347
2. (a) 0.6 (b) 0.8
(c) $(0.3 + 0.6t + 0.1t^2)^2$
3. (a) $H_0: \mu = 86, H_1: \mu > 86$
(b) There is evidence that they are taller ($p = 0.0294$)
4. (a) 0.202 (b) 0.934 (c) 0.103
5. (a) 0.35 (b) -1.38
6. (a) 1.55, 0.524
(b) No sufficient evidence ($p = 0.385$)

7. 0.00767
9. (a) 0.995 (c) Fertiliser concentration
(d) 11.8 (e) Extrapolation
10. (a) (i) 3 (ii) $\frac{1}{4}$ (iii) 1
(b) 0.432
11. (a) (i) 7.34 (ii) 1.67
(b) (i) 0.0330 (ii) Reject H_0
12. 0.297, 0.703
13. (a) [133, 139]
(b) No, distribution of X is normal.
14. t -test, $p = 0.514$, no evidence for improvement
15. (a) 46 (b) 12.5 (c) [42.6, 49.4]
16. (a) 0.00921
(b) We used normal distribution for the sample mean.

$$17. (b) F(x) = \begin{cases} 0, x \leq 0 \\ \frac{1}{8}x^2, 0 \leq x \leq 2 \\ 1 + \frac{1}{54}(x-5)^3, 2 \leq x \leq 5 \\ 1, x > 5 \end{cases}$$

- (c) $\frac{1}{2}$ (d) 2.62
18. (a) -0.989
(b) Yes ($p = 6.96 \times 10^{-4}$)
(c) $t - 29.8 = -1.14(T - 40)$
19. (a) $\frac{8}{9}$ (b) 0.0144
20. (a) 13 (b) 636
21. (b) $\frac{45}{47}$
22. (b) 2.79%
(c) (i) 0.5 (ii) 0.5
23. (a) (i) 0.579 (ii) geometric
(b) (i) 3.61×10^{-4} (ii) 30
(iii) 24 or 25
(c) The distribution in (b) is the sum of five independent observations of the distribution in (a).
(d) It is a sum of a large sample of observations of a geometric distribution.
 $\mu = 100, \sigma^2 = 150$.

$$24. (a) F(t) = \begin{cases} 0, t \leq 1 \\ \frac{1}{2}(t-1), 1 < t < 3 \\ 1, t \geq 3 \end{cases} \quad (d) \frac{9}{7}$$

Appendix: Calculator skills sheets

A Finding probabilities in the <i>t</i> -distribution	
CASIO	132
TEXAS	133
B Finding <i>t</i> -scores given probabilities	
CASIO	134
TEXAS	135
C Confidence interval for the mean with unknown variance (from data)	
CASIO	136
TEXAS	137
D Confidence interval for the mean with unknown variance (from stats)	
CASIO	138
TEXAS	140
E Hypothesis test for the mean with unknown variance (from data)	
CASIO	141
TEXAS	143
F Hypothesis test for the mean with unknown variance (from stats)	
CASIO	144
TEXAS	146
G Confidence interval for the mean with known variance (from data)	
CASIO	148
TEXAS	149
H Confidence interval for the mean with known variance (from stats)	
CASIO	150
TEXAS	152
I Hypothesis test for the mean with known variance (from stats)	
CASIO	154
TEXAS	155
J Finding the correlation coefficient and the equation of the regression line	
CASIO	157
TEXAS	158

A Finding probabilities in the t -distribution

You will need...

- the number of degrees of freedom
- the interval of interest in terms of T

In our example...

- 6
- $[-0.8, 2.4]$

How to do it...

Notes	You should press	You will see
To get to the correct menu	MENU 2 (STAT) F5 (DIST) F2 (t) F2 (tcd) F2 (Var)	Student-t C.D Data : Variable Lower : 0 Upper : 0 df : 0 Save Res: None Execute List Var
Enter limits and degrees of freedom	▼ → 0 . 8 EXE 2 . 4 EXE 6 EXE	Student-t C.D Lower : -0.8 Upper : 2.4 df : 6 Save Res: None Execute None LIST
Calculate	EXE	Student-t C.D P = 0.74624737 t:Low = -0.8 t:UP = 2.4

What to write down...

If $X \sim t_6$, $P(-0.8 \leq X \leq 2.4) = 0.746$ (3SF from GDC)

* These instructions were written based on the CASIO model fx9860G SD and might not be true for other models. If in doubt, refer to your calculator's manual.

A Finding probabilities in the t -distribution

You will need...

- The number of degrees of freedom
- The interval of interest in terms of T

In our example...

- 6
- $[-0.8, 2.4]$

How to do it...

Notes	You should press	You will see
To get to the correct function	2nd VARS 6 (tcdf()	
Enter (lower limit, upper limit, degrees of freedom)	(- 0 , 8 , 2 , 4 , 6 ENTER ENTER ENTER	tcdf(-0.8,2.4,6) .7462473793

What to write down...

If $X \sim t_6$, $P(-0.8 < X < 2.4) = 0.746$ (3SF from GDC)

* These instructions were written based on the TEXAS model TI-84 Plus Silver Edition and might not be true for other models. If in doubt, refer to your calculator's manual.

B Finding t -scores given probabilities

You will need...

- the number of degrees of freedom, ν
- the probability, $P(T > t)$

In our example...

- 6
- 0.28

EXAM HINT

Notice that we use $P(T > t)$ rather than the more common cumulative probability $P(T \leq t)$

How you do it...

Notes	You should press	You will see
To get to the correct menu	MENU 2 (STAT) F5 (DIST) F2 (t) F3 (Inv) F2 (Var)	Inverse Student-t Data : Variable Area : 0 df : 0 Save Res: None Execute List Var
Enter $P(T > t)$ in Area	▼ 0 . 2 8 EXE	
Enter degrees of freedom and find t -score	6 EXE EXE	Inverse Student-t xInv = 0.61684437

What to write down...

If $X \sim t_6$ and $P(X > x) = 0.28$ then $x = 0.617$ (3SF from GDC)

* These instructions were written based on the CASIO model fx9860G SD and might not be true for other models. If in doubt, refer to your calculator's manual.

B Finding t -scores given probabilities

You will need...

- The number of degrees of freedom, ν
- The cumulative probability, $P(T < t)$

In our example...

- 6
- 0.72

How you do it...

Notes	You should press	You will see
To get to the correct function	2nd VARS 4 (invT())	
Enter $P(T < t)$, in 'area' then enter the degrees of freedom	0 . 7 2 , 6) [ENTER]	invT(0.72,6) 0.616844374

What to write down...

If $X \sim t_6$ and $P(X < x) = 0.72$ then $x = 0.617$ (3SF from GDC)

* These instructions were written based on the TEXAS model TI-84 Plus Silver Edition and might not be true for other models. If in doubt, refer to your calculator's manual.

C Confidence interval for the mean with unknown variance (from data)

You will need...

- the sample stored in a list
- the confidence level

In our example...

- {1,3,5,2} stored in List 2
- 90%

How you do it...

Notes	You should press	You will see
To get to the correct menu (make sure you change to 'List' from 'variable' if you need to)	MENU 2 (STAT) F4 (INTR) F2 (t) F1 (1-S)	1-Sample tInterval Data :List C-Level :0.95 List :List1 Freq :1 Save Res :None Execute List Var
Remember that the confidence level must be input as a decimal	▼ 0 . 9 EXE	
Select which list your data are stored in	F1 (LIST) 2 EXE	
You can then say if the frequencies are stored in another list, or if the frequency of each item is 1	▼ F1 (1)	1-Sample tInterval Data :List C-Level :0.9 List :List2 Freq :1 Save Res :None Execute 1 EXE
Find the interval	▼ ▼ EXE	1-Sample tInterval Left =0.74043339 Right =4.7595666 \bar{x} =2.75 s_x =1.70782513 n =4

What to write down...

$$\bar{x} = 2.75, s_{n-1} = 1.71$$

Using t_3 distribution $0.740 < \mu < 4.76$ (3SF from GDC)

* These instructions were written based on the CASIO model fx9860G SD and might not be true for other models. If in doubt, refer to your calculator's manual.

C Confidence interval for the mean with unknown variance (from data)

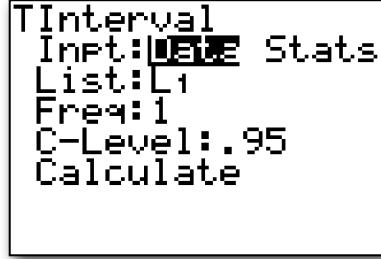
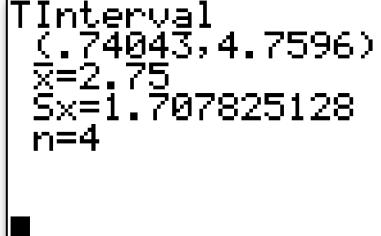
You will need...

- The sample stored in a list
- The confidence level

In our example...

- {1,3,5,2} stored in List 2
- 90%

How you do it...

Notes	You should press	You will see
To get to the correct menu	$\boxed{\text{STAT}}$ \blacktriangleright \blacktriangleright (TESTS) $\boxed{8}$ (Tinterval)	
The default setting is to input data. Move down to enter where the data are stored. If the frequencies are stored in another list you can change that too. By default the frequency of each item is 1	\blacktriangledown $\boxed{2\text{nd}}$ $\boxed{2}$ (L_2) $\boxed{\text{ENTER}}$	
Put in the confidence interval as a decimal	\blacktriangledown \blacktriangledown $\boxed{0}$ $\boxed{.}$ $\boxed{9}$ (C-Level) $\boxed{\text{ENTER}}$ $\boxed{\text{ENTER}}$	

What to write down...

$$\bar{x} = 2.75, s_{n-1} = 1.71$$

Using t_3 distribution $0.740 < \mu < 4.76$ (3SF from GDC)

* These instructions were written based on the TEXAS model TI-84 Plus Silver Edition and might not be true for other models. If in doubt, refer to your calculator's manual.

D Confidence interval for the mean with unknown variance (from stats)

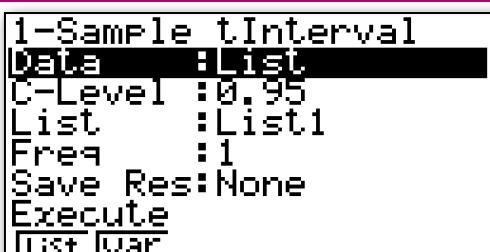
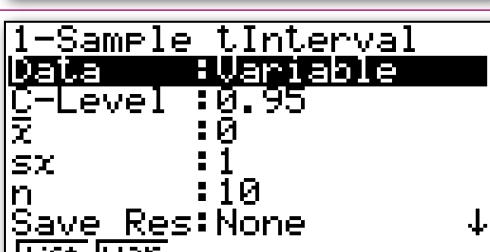
You will need...

- the sample mean (\bar{x})
- unbiased estimate of population standard deviation (s_{n-1})
- the confidence level
- the number of data items (n)

In our example...

- 2.75
- 1.707825128
(stored exactly in A)
- 90%
- 4

How you do it...

Notes	You should press	You will see
To get to the correct menu	DEL 2 (STAT) F4 (INTR) F2 (t) F1 (1-S)	
To change entry to statistics (variable)	F2 (VAR)	
Remember that the confidence level must be input as a decimal	▼ 0 . 9 EXE	
Enter \bar{x}	2 . 7 5 EXE	
Enter standard deviation	ALPHA X,θ,T (A) EXE	
Enter n	4 EXE	

instructions continue on next page →

* These instructions were written based on the CASIO model fx9860G SD and might not be true for other models. If in doubt, refer to your calculator's manual.

continued ...

You can then tell the calculator to find the interval

▼ EXE

1-Sample t Interval	
Left	=0.74043339
Right	=4.7595666
\bar{x}	=2.75
sx	=1.70782513
n	=4

What to write down...

Using the t distribution with $\nu = 3$: $0.740 < \mu < 4.76$ (3SF from GDC)

* These instructions were written based on the CASIO model fx9860G SD and might not be true for other models. If in doubt, refer to your calculator's manual.

D Confidence interval for the mean with unknown variance (from stats)

You will need...

- The sample mean (\bar{x})
- Unbiased estimate of population standard deviation (s_{n-1})
- The confidence level

In our example...

- 2.75
- 1.707825128
(stored exactly in A)
- 90%

How you do it...

Notes	You should press	You will see
To get to the correct menu	[STAT] (Stat) [▼] [▼] (Test) [8] (TInterval)	TInterval Inpt: DATA Stats List:L1 Freq:1 C-Level:.95 Calculate
Move across to change input method to summary statistics	[▼] (Stats) [ENTER]	TInterval Inpt:Data Stats X:0 Sx:1 n:10 C-Level:.95 Calculate
Move down to enter the sample mean	[▼] (\bar{x}) [2] [▼] [7] [5] [ENTER]	
Enter the standard deviation (S_x)	[ALPHA] [MATH] (A) [ENTER]	
Enter the number of data items (n)	[4] [ENTER]	
Put in the confidence interval as a decimal (C-Level)	[0] [.] [9] [ENTER] [ENTER]	TInterval (.74043, 4.7596) $\bar{x}=2.75$ $S_x=1.707825128$ $n=4$

What to write down...

Using the t distribution with $v = 3$: $0.740 < \mu < 4.76$ (3SF from GDC)

* These instructions were written based on the TEXAS model TI-84 Plus Silver Edition and might not be true for other models. If in doubt, refer to your calculator's manual.

E Hypothesis test for the mean with unknown variance (from data)

You will need...

- the sample stored in a list
- the mean under the null hypothesis (μ_0)
- the alternative hypothesis

In our example...

- {1,3,5,2} stored in List 2
- 3.9
- $\mu < \mu_0$

How you do it...

Notes	You should press	You will see
To get to the correct menu	[DEL] [2] (STAT) [F3] (TEST) [F2] (t) [F1] (1-S)	
Set the direction of the alternative hypothesis	▼ [F2] (<)	
Enter the value of the mean under the null hypothesis	▼ [3] [.] [9] [EXE]	
Select which list your data are stored in	[F1] (LIST) [2] [EXE]	
You can then say if the frequencies are stored in another list, or if the frequency of each item is 1	▼ [F1] (1)	
You can then tell the calculator to conduct the test	▼ ▼ [EXE]	

* These instructions were written based on the CASIO model fx9860G SD and might not be true for other models. If in doubt, refer to your calculator's manual.

What to write down...

Under H_0 , $T = \frac{\bar{x} - 3.9}{\sqrt{s_{n-1}/n}} \sim t_{n-1}$

$$\bar{x} = 2.75, s_{n-1} = 1.71, \nu = 3, T = -1.35$$

p-value = 0.135

* These instructions were written based on the CASIO model fx9860G SD and might not be true for other models. If in doubt, refer to your calculator's manual.

E Hypothesis test for the mean with unknown variance (from data)

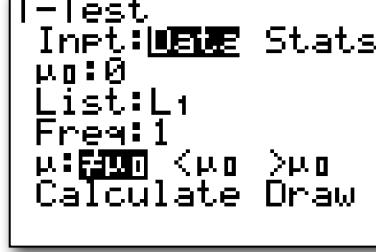
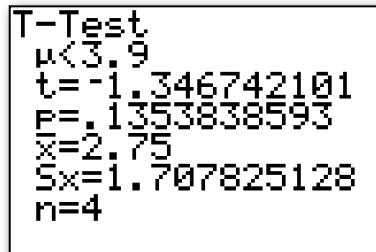
You will need...

- The sample stored in a list
- The mean under the null hypothesis (μ_0)
- The alternative hypothesis

In our example...

- {1,3,5,2} stored in List 2
- 3.9
- $\mu < \mu_0$

How you do it...

Notes	You should press	You will see
To get to the correct menu	[STAT] (Stat) \blacktriangleright (Test) [2] (T-Test)	
The default setting is to input data. Move down to enter the mean under the null hypothesis	\blacktriangledown [3] [.] [9] [ENTER]	
You may need to change the list being used. If the frequencies are stored in another list you can change that too. By default the frequency of each item is 1	[2nd] [2] (L₂) [ENTER]	
Select which alternative hypothesis you wish to test	\blacktriangledown \blacktriangleright (< μ_0) [ENTER] \blacktriangledown (Calculate) [ENTER]	

What to write down...

Under H_0 , $T = \frac{\bar{x} - 3.9}{\sqrt{s_{n-1}/n}} \sim t_{n-1}$

$\bar{x} = 2.75$, $s_{n-1} = 1.71$, $\nu = 3$, $T = -1.35$

p-value = 0.135

* These instructions were written based on the TEXAS model TI-84 Plus Silver Edition and might not be true for other models. If in doubt, refer to your calculator's manual.

F Hypothesis test for the mean with unknown variance (from stats)

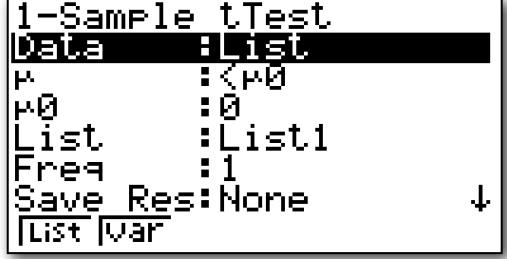
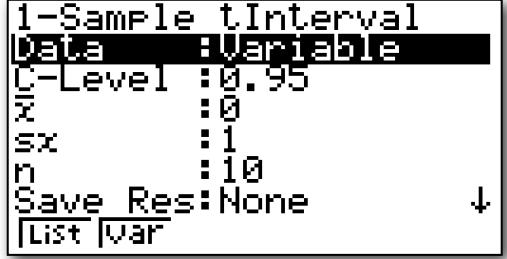
You will need...

- the sample mean (\bar{x})
- unbiased estimate of population standard deviation (s_{n-1})
- the number of data items (n)
- the mean according to the null hypothesis (μ_0)
- the alternative hypothesis

In our example...

- 2.75
- 1.707825128
(stored exactly in A)
- 4
- 3.9
- $\mu < \mu_0$

How you do it...

Notes	You should press	You will see
To get to the correct menu	DEL 2 (STAT) F3 (TEST) F2 (t) F1 (1-S)	
To change entry to statistics	F2 (VARS)	
Set the direction of the alternative hypothesis	▼ F2 (<)	
Enter the value of the mean under the null hypothesis	▼ 3 . 9 EXE	
Enter \bar{x}	2 . 7 5 EXE	
Enter standard deviation	ALPHA X,θ,T (A) EXE	

instructions continue on next page →

* These instructions were written based on the CASIO model fx9860G SD and might not be true for other models. If in doubt, refer to your calculator's manual.

continued ...

Enter n	4 EXE	
You can then tell the calculator to conduct the test	▼ ▼ EXE	1-Sample tTest $\mu < 3.9$ $t = -1.3467421$ $p = 0.13538386$ $\bar{x} = 2.75$ $s_x = 1.70782513$ $n = 4$

What to write down...

Under H_0 , $T = \frac{\bar{x} - 3.9}{\sqrt{s_{n-1}/n}} \sim t_{n-1}$

$$T = -1.35, \nu = 3$$

$$p\text{-value} = 0.135$$

* These instructions were written based on the CASIO model fx9860G SD and might not be true for other models. If in doubt, refer to your calculator's manual.

F Hypothesis test for the mean with unknown variance (from stats)

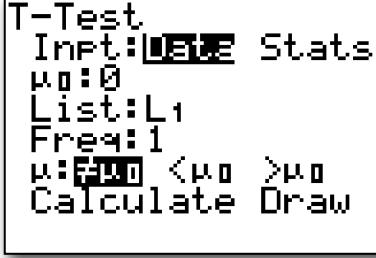
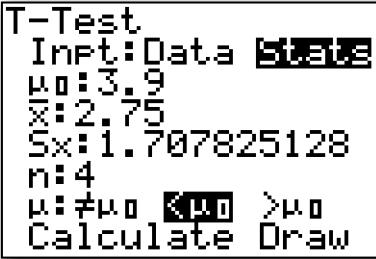
You will need...

- The sample mean
- The unbiased estimate of the sample standard deviation (s_{n-1})
- The number of data items (n)
- The mean under the null hypothesis (μ_0)
- The alternative hypothesis

In our example...

- 2.75
- 1.707825128
(stored exactly in A)
- 4
- 3.9
- $\mu < \mu_0$

How you do it...

Notes	You should press	You will see
To get to the correct menu	[STAT] (Stat) [▼] [▼] (Test) [2] (T-Test)	
Change the setting to input summary statistics	[▼] (Stat) [ENTER]	
Move down to enter the mean under the null hypothesis	[▼] [3] [.] [9] [ENTER]	
Enter the sample mean	[2] [.] [7] [5] [ENTER]	
Enter the standard deviation	[ALPHA] [MATH] (A) [ENTER]	
Enter the number of data items (n)	[4] [ENTER]	
Select which alternative hypothesis you wish to test	[▼] [▼] (< μ_0) [ENTER] [▼] (Calculate) [ENTER]	

instructions continue on next page →

* These instructions were written based on the TEXAS model TI-84 Plus Silver Edition and might not be true for other models. If in doubt, refer to your calculator's manual.

continued ...

T-Test
$\mu < 3.9$
t = -1.346742101
P = .1353838593
$\bar{x} = 2.75$
Sx = 1.707825128
n = 4

What to write down...

Under H_0 , $T = \frac{\bar{x} - 3.9}{\sqrt{s_{n-1}/n}} \sim t_{n-1}$

$$T = -1.35, \nu = 3$$

$$p\text{-value} = 0.135$$

* These instructions were written based on the TEXAS model TI-84 Plus Silver Edition and might not be true for other models. If in doubt, refer to your calculator's manual.

G Confidence interval for the mean with known variance (from data)

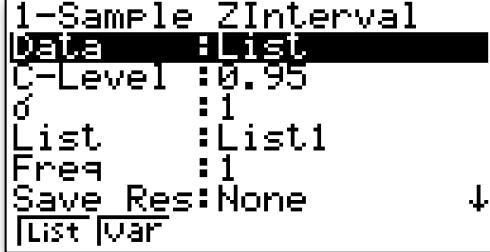
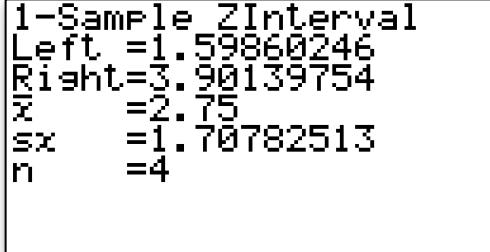
You will need...

- the sample stored in a list
- the population standard deviation (σ)
- the confidence level

In our example...

- {1,3,5,2} stored in List 2
- 1.4
- 90%

How you do it...

Notes	You should press	You will see
To get to the correct menu	DEL 2 (STAT) F4 (INTR) F1 (Z) F1 (1-S)	
Remember that the confidence level must be input as a decimal	▼ 0 . 9 EXE	
You will automatically move to input σ	1 . 4 EXE	
Select which list your data are stored in	F1 (LIST) 2 EXE	
You can then say if the frequencies are stored in another list, or if the frequency of each item is 1	▼ F1 (1)	
You can then tell the calculator to find the interval	▼ ▼ EXE	

What to write down...

Using normal distribution:

$$\bar{x} = 2.75$$

$$1.60 < \mu < 3.90 \text{ (3SF from GDC)}$$

*These instructions were written based on the CASIO model fx9860G SD and might not be true for other models. If in doubt, refer to your calculator's manual.

G Confidence interval for the mean with known variance (from data)

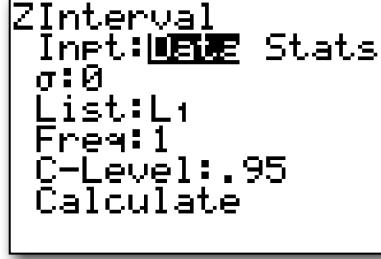
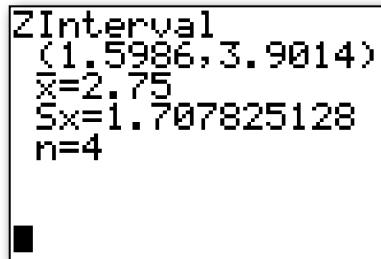
You will need...

- The sample stored in a list
- The population standard deviation (σ)
- The confidence level

In our example...

- {1,3,5,2} stored in List 2
- 1.4
- 90%

How you do it...

Notes	You should press	You will see
To get to the correct menu	STAT (Stat) ▶ ▶ (Test) 7 (Z-Int)	
The default setting is to input data. Move down to set the standard deviation	▼ (σ) 1 □ 4 ENTER	
You may need to change the list being used. If the frequencies are stored in another list you can change that too. By default the frequency of each item is 1	2nd 2 (L ₂) ENTER	
Put in the confidence interval as a decimal	▼ ▼ (C-Level) 0 □ 9 ENTER ENTER	

What to write down...

Using normal distribution:

$$\bar{x} = 2.75$$

$$1.60 < \mu < 3.90 \text{ (3SF from GDC)}$$

* These instructions were written based on the TEXAS model TI-84 Plus Silver Edition and might not be true for other models. If in doubt, refer to your calculator's manual.

H Confidence interval for the mean with known variance (from stats)

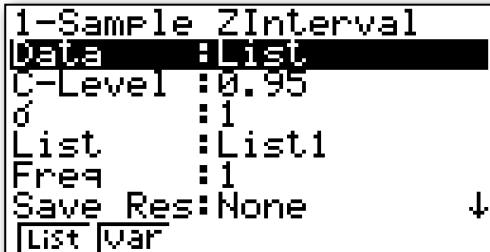
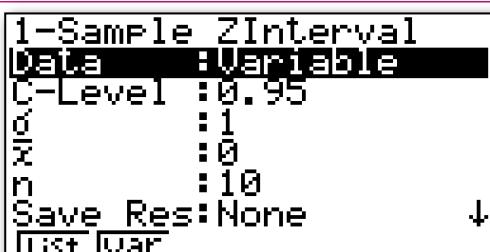
You will need...

- the sample mean (\bar{x})
- the population standard deviation (σ)
- the number of data items (n)
- the confidence level

In our example...

- 2.75
- 1.4
- 4
- 90%

How you do it...

Notes	You should press	You will see
To get to the correct menu	[MENU] [2] (STAT) [F4] (INTR) [F1] (Z) [F1] (1-S)	
Set input mode to variables	[F2] (Var)	
Remember that the confidence level must be input as a decimal	[▼] [0] [.] [9] [EXE]	
Enter σ	[1] [.] [4] [EXE]	
Enter \bar{x}	[2] [.] [7] [5] [EXE]	
Enter n	[4] [EXE]	

instructions continue on next page →

* These instructions were written based on the CASIO model fx9860G SD and might not be true for other models. If in doubt, refer to your calculator's manual.

continued ...

You can then tell
the calculator to
find the interval



EXE

1-Sample ZInterval
Left = 1.59860246
Right = 3.90139754
 \bar{x} = 2.75
n = 4

What to write down...

Using normal distribution:

$$1.60 < \mu < 3.90 \text{ (3SF from GDC)}$$

* These instructions were written based on the CASIO model fx9860G SD and might not be true for other models. If in doubt, refer to your calculator's manual.

H Confidence interval for the mean with known variance (from stats)

You will need...

- The sample mean (\bar{x})
- The population standard deviation (σ)
- The number of data items (n)
- The confidence level

In our example...

- 2.75
- 1.4
- 4
- 90%

How you do it...

Notes	You should press	You will see
To get to the correct menu	STAT (Stat) ▶ ▶ (Test) 7 (Z-Int)	ZInterval Inpt: Data Stats σ :0 List:L1 Freq:1 C-Level:.95 Calculate
Change the setting to input summary statistics	▶ (Stats) ENTER	ZInterval Inpt: Data Sum σ :1 \bar{x} :0 n :10 C-Level:.95 Calculate
Move down to set the standard deviation	▼ (σ) 1 . 4 ENTER	
Enter the sample mean (\bar{x})	2 . 7 5 ENTER	
Enter the number of data items (n)	4 ENTER	

instructions continue on next page →

* These instructions were written based on the TEXAS model TI-84 Plus Silver Edition and might not be true for other models. If in doubt, refer to your calculator's manual.

continued ...

Enter the confidence interval as a decimal (C-Level)

0 9 [ENTER] [ENTER]

ZInterval
 $(1.5986, 3.9014)$
 $\bar{x}=2.75$
 $S_x=1.707825128$
 $n=4$

What to write down...

Using normal distribution:

$1.60 < \mu < 3.90$ (3SF from GDC)

* These instructions were written based on the TEXAS model TI-84 Plus Silver Edition and might not be true for other models. If in doubt, refer to your calculator's manual.

I Hypothesis test for the mean with known variance (from stats)

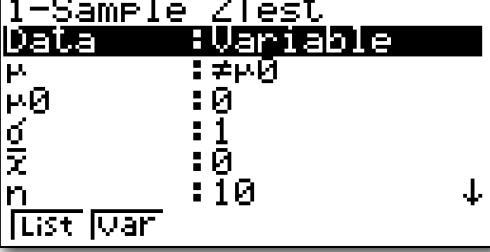
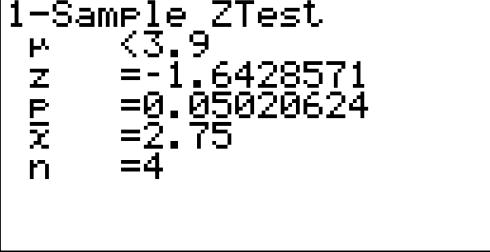
You will need...

- the sample mean (\bar{x})
- the population standard deviation (σ)
- the number of data items (n)
- the mean according to the null hypothesis (μ_0)
- the alternative hypothesis

In our example...

- 2.75
- 1.4
- 4
- 3.9
- $\mu < \mu_0$

How you do it...

Notes	You should press	You will see
To get to the correct menu	MENU 2 (STAT) F3 (TEST) F1 (Z) F1 (1-S) F2 (Var)	
Set the direction of the alternative hypothesis	▼ F2 (<)	
Enter the value of the mean under the null hypothesis	▼ 3 . 9 EXE	
Enter σ	1 . 4 EXE	
Enter \bar{x}	2 . 7 5 EXE	
Enter n	4 EXE	
You can then tell the calculator to perform the test	▼ EXE	

What to write down...

Under H_0 , $Z = \frac{\bar{x} - 3.9}{\sqrt{1.4/4}} \sim N(0,1)$

$$Z = -1.64$$

$$p\text{-value} = 0.0502$$

* These instructions were written based on the CASIO model fx9860G SD and might not be true for other models. If in doubt, refer to your calculator's manual.

I Hypothesis test for the mean with known variance (from stats)

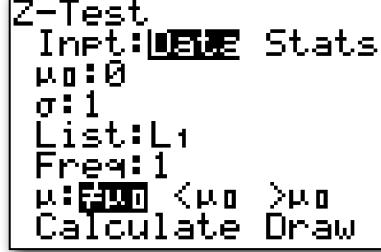
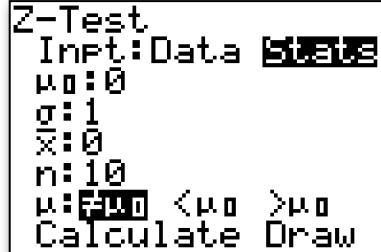
You will need...

- The sample mean
- The population standard deviation (σ)
- The number of data items (n)
- The mean under the null hypothesis (μ_0)
- The alternative hypothesis

In our example...

- 2.75
- 1.4
- 4
- 3.9
- $\mu < \mu_0$

How you do it...

Notes	You should press	You will see
To get to the correct menu	[STAT] (Stat) [▼] [▼] (Test) [1] (Z-Test)	
Change the setting to input summary statistics	[▼] (Stats) [ENTER]	
The default setting is to input data. Move down to enter the mean under the null hypothesis	[▼] [3] [▼] [9] [ENTER]	
Enter the standard deviation	[1] [▼] [4] [ENTER]	
Enter the sample mean (\bar{x})	[2] [▼] [7] [5] [ENTER]	
Enter the number of data items (n)	[4] [ENTER]	

instructions continue on next page →

* These instructions were written based on the TEXAS model TI-84 Plus Silver Edition and might not be true for other models. If in doubt, refer to your calculator's manual.

continued ...

Select which alternative hypothesis you wish to test

- $(< \mu_0)$
- ENTER
- (Calculate)
- ENTER

Z-Test
 $\mu < 3.9$
 $z = -1.642857143$
 $P = .0502062319$
 $\bar{x} = 2.75$
 $n = 4$

What to write down...

$$\text{Under } H_0, Z = \frac{\bar{x} - 3.9}{\sqrt{1.4/4}} \sim N(0,1)$$

$$Z = -1.64$$

$$p\text{-value} = 0.0502$$

* These instructions were written based on the TEXAS model TI-84 Plus Silver Edition and might not be true for other models. If in doubt, refer to your calculator's manual.

J Finding the correlation coefficient and the equation of the regression line

You will need...

- the x -data stored in list 1
- the y -data stored in list 2

In our example...

- {3, 3, 10}
- {12, 10, -4}

How you do it...

Notes	You should press	You will see
Go to the statistics menu then the calculation submenu	[MENU] [2] [STAT] [F2] [CALC]	
Select the regression option then x	[F3] (REG) [F1] (x)	

What to write down...

From GDC $r = -0.993$ and $y = -2.14x + 17.4$

* These instructions were written based on the CASIO model fx9860G SD and might not be true for other models. If in doubt, refer to your calculator's manual.

J Finding the correlation coefficient and the equation of the regression line

You will need...

- The x -data stored in list 1
- The y -data stored in list 2

In our example...

- {3, 3, 10}
- {12, 10, -4}

How you do it...

Notes	You should press	You will see
Ensure that the calculator is set to 'Diagnostics On'	[2nd] [0] (Catalog) [▼] ... [▼] (Diagnostics on) [ENTER] [ENTER]	CATALOG DependAsk DependAuto det< DiagnosticOff ►DiagnosticOn dim< Disp
Use the linear regression function	[STAT] [►] (Calc) [4] (LinReg(ax+b)) [ENTER]	LinReg $y = ax + b$ $a = -2.142857143$ $b = 17.42857143$ $r^2 = .9868421053$ $r = -.9933992678$

What to write down...

From GDC $r = -0.993$ and $y = -2.14x + 17.4$

* These instructions were written based on the TEXAS model TI-84 Plus Silver Edition and might not be true for other models. If in doubt, refer to your calculator's manual.

Glossary

Words that appear in **bold** in the definitions of other terms, are also defined in this glossary. The abstract nature of this option means that some defined terms can realistically only be explained in terms of other, more simple concepts.

Term	Definition	Example
acceptance region	The values of the test statistic for which there is no sufficient evidence to reject the null hypothesis	If we are testing $H_0: \mu = 5$ against $H_1: \mu > 5$ for a normal distribution $H \sim N(\mu, 3^2)$, using a single observation and a 5% significance level , then the acceptance region is $]-\infty, 9.93[$ because if $H \sim N(5, 3^2)$ then $P(X \geq 9.93) = 0.05$
alternative hypothesis	The statement that opposes the null hypothesis	To test whether a die is biased towards 6s we would use the alternative hypothesis $H_1: P(\text{roll a } 6) > \frac{1}{6}$
biased estimator	A statistic used to estimate an unknown parameter whose expectation is not equal to the actual value of the parameter	The sample variance s_n^2 is a biased estimator of the population variance σ^2
Central Limit Theorem	The result stating that the sample sum and the sample mean of a large sample approximately follow a normal distribution.	The sample mean has distribution $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, where $\mu = E(X)$ and $\sigma^2 = \text{Var}(X)$
confidence interval	An interval calculated from the sample, which contains the true value of a population parameter with pre-determined probability	The sample 1, 2, 3, 4 gives [0.446, 4.55] as the 95% confidence interval for the population mean
confidence level	The probability that the confidence interval contains the true value of the population parameter	A 90% confidence interval is wider than a 95% confidence interval
correlation coefficient	A numerical measure of linear relationship between two random variables; related to covariance	Negative correlation implies that if one variable increases the other decreases
covariance	A numerical measure of linear relationship between two random variables	Covariance is related to the correlation coefficient : $\text{Cov}(X, Y) = \rho \sigma_X \sigma_Y$
critical region	The values of the test statistic for which the null hypothesis is rejected; same as rejection region	If we are testing $H_0: \mu = 5$ against $H_1: \mu > 5$ for a normal distribution $X \sim N(\mu, 3^2)$, using a single observation and a 5% significance level , then the critical region is $[9.93, \infty[$ because if $X \sim N(5, 3^2)$ then $P(X \geq 9.93) = 0.05$

Term	Definition	Example								
critical value	The value on the boundary between rejection and acceptance regions	If we are testing $H_0: \mu = 5$ against $H_1: \mu > 5$ for a normal distribution $X \sim N(\mu, 3^2)$, using a single observation and a 5% significance level , then the critical value is 9.93 because if $X \sim N(5, 3^2)$ then $P(X \geq 9.93) = 0.05$								
degrees of freedom	(for a <i>t</i> -distribution) A parameter which determines the shape of a particular <i>t</i>-distribution	For a sample of size 6 from a population with mean 11, the statistic $\frac{\bar{X} - 11}{s_{n-1}/\sqrt{6}}$ follows <i>t</i> ₅ distribution with 5 degrees of freedom								
efficiency	(of an estimator) The variance of the estimator	More efficient estimator has smaller variance								
hypothesis test	A statistical procedure used to decide whether there is significant evidence that the value of a population parameter has changed or is different from expected	To test whether a coin is biased we could use a hypothesis test in which the null hypothesis is $P(\text{heads}) = \frac{1}{2}$ and the alternative hypothesis is $P(\text{heads}) \neq \frac{1}{2}$								
line of best fit	A straight line that best represents linear relationship between two random variables; same as regression line	The line of best fit always passes through the point (\bar{x}, \bar{y})								
null hypothesis	The default statement which is accepted unless there is significant evidence against it	To test whether a die is biased towards 6s we would use the null hypothesis $H_0: P(\text{roll a } 6) = \frac{1}{6}$								
one-tailed test	A hypothesis test where the alternative hypothesis is of the form $p > p_0$ or $p < p_0$	To test whether a coin is biased towards heads we would use a one-tailed test with the alternative hypothesis $H_1: P(\text{get a head}) > \frac{1}{2}$								
point estimate	A single value calculated from the sample and used to estimate a population parameter	The sample 1, 2, 3, 4 gives a point estimate 2.5 for the population mean								
Probability Generating Function	A polynomial in which the coefficient of t^k is the probability of the random variable taking the value k	Random variable X with probability distribution <table border="1" style="margin-left: auto; margin-right: auto;"> <tr> <td>k</td><td>0</td><td>1</td><td>3</td></tr> <tr> <td>$P(X = k)$</td><td>0.2</td><td>0.5</td><td>0.3</td></tr> </table> Has generating function $G(t) = 0.2 + 0.5t + 0.3t^3$	k	0	1	3	$P(X = k)$	0.2	0.5	0.3
k	0	1	3							
$P(X = k)$	0.2	0.5	0.3							
<i>p</i>th percentage point	The value of a random variable X such that the probability of X taking this value or lower is $p\%$	The 85 th percentage point of $N(0, 1)$ is 1.04 because if $X \sim N(0, 1)$ then $P(X \leq 1.04) = 0.85$								

Term	Definition	Example
p-value	(of a hypothesis test) The probability of the observed outcome, or more extreme, being observed when the null hypothesis is true	If we are testing $H_0 : P(\text{heads}) = \frac{1}{2}$ against $H_1 : P(\text{heads}) < \frac{1}{2}$ and we use a sample of 10 coin tosses, then the p-value corresponding to the outcome '3 heads' is $P(X \leq 3) = 0.172$
regression line	A straight line that best represents linear relationship between two random variables; same as line of best fit	The regression line always passes through the point (\bar{x}, \bar{y})
rejection region	The values of the test statistic for which the null hypothesis is rejected; same as critical region	If we are testing $H_0 : \mu = 5$ against $H_1 : \mu > 5$ for a normal distribution $X \sim N(\mu, 3^2)$, using a single observation and a 5% significance level , then the rejection region is $[9.93, \infty[$ because if $X \sim N(5, 3^2)$ then $P(X \geq 9.93) = 0.05$
sample mean	A statistic found by calculating the mean of the <u>sample values</u> . It is denoted by \bar{X}	For a sample of size 3, the sample mean is $\bar{X} = \frac{X_1 + X_2 + X_3}{3}$, where X_1, X_2, X_3 are random variables representing three independent observations of the random variable X
significance level	(of a hypothesis test) A numerical measure of how likely an outcome must be in order to reject the null hypothesis	If we are testing $H_0 : P(\text{heads}) = \frac{1}{2}$ against $P(\text{heads}) > \frac{1}{2}$ using a sample of 10 coin tosses, and we choose to reject the null hypothesis if $X \geq 8$, the significance level of this test is 5.47%
test statistic	A random variable whose value can be calculated from a sample, used in a hypothesis test	If we are testing $H_0 : \mu = 5$ against $H_1 : \mu < 5$ for a normal distribution with unknown variance, using a sample of size 3, the test statistic could be $T = \frac{\bar{X} - 5}{s_{n-1}/\sqrt{2}}$, where $\bar{X} = \frac{X_1 + X_2 + X_3}{3}$ and $s_{n-1}^2 = \frac{X_1^2 + X_2^2 + X_3^2}{3} - \bar{X}^2$
two-tailed test	A hypothesis test where the alternative hypothesis is of the form $p \neq p_0$	To test whether a coin is biased, without wanting to know whether heads or tails are more likely, we would use a two-tailed test with the alternative hypothesis $H_1 : P(\text{get a head}) \neq \frac{1}{2}$

Term	Definition	Example
type I error	An incorrect conclusion to a hypothesis test where a correct null hypothesis is falsely rejected	The probability of type I error is equal to the significance level of the test
type II error	An incorrect conclusion to a hypothesis test where an incorrect null hypothesis is falsely accepted	To find the probability of type II error we need an alternative value for the population parameter
t-distribution	The distribution of $\frac{\bar{X} - \mu}{s_{n-1} / \sqrt{n}}$ where the standard deviation s_{n-1} has been estimated from a sample	We need to use the t-distribution with confidence intervals and hypothesis test whenever the population variance is unknown
t-test	A hypothesis test for the population mean in which the test statistic follows a t-distribution	We use a t-test when \bar{X} follows a normal distribution but the population variance is unknown
unbiased estimator	A statistic used to estimate an unknown parameter whose expectation is equal to the actual value of the parameter	The sample mean \bar{X} is an unbiased estimator of the population mean μ
Z-test	A hypothesis test for the population mean in which the test statistic follows a normal distribution	We use a Z-test when the population variance is known and X follows a normal distribution

Index

- acceptance region, 72, 89, 90
definition, 159
exercises, 77, 79–80
- alternative hypothesis, 71, 93
definition, 159
exercises, 76, 77, 84, 86–87, 91–92
- average of a sample *see* sample mean
- biased estimator, 48, 67
definition, 159
worked example, 52–53
- binomial distribution, negative, 24–27
- bivariate distributions, 97–99
covariance and correlation, 100–7
linear regression, 107–11
mixed exam practice, 113–14
summary, 111–12
- calculator skills, 131
confidence interval for mean with known variance, 148–53
confidence interval for mean with unknown variance, 136–40
correlation coefficients, 157–58
hypothesis test for mean with known variance, 154–56
hypothesis test for mean with unknown variance, 141–47
probabilities in *t*-distribution, 132–33
t-scores given probabilities, 134–35
- central limit theorem (CLT), 16–17, 19, 160
- confidence interval, 55, 68, 75
for a mean with unknown variance, 63–67
calculator skills, 136–40, 148–53
definition, 159
for the population mean (variance known), 55–59
- confidence level, 55, 68, 159
- continuous random variables
distributions of functions of, 43–45
expectation of a function of, 4, 18
- correlation coefficient, 101–7
calculator skills, 157–58
definition, 159
formula, 102, 111
- covariance, 100–1, 111, 159
- critical region, 72, 79, 94
definition, 159
worked examples, 74–76
see also rejection region
- critical region method, 72–73
- critical value, 72, 75, 79, 160
- cumulative distribution functions, 38–46
- cumulative probability function, 38–43
- degrees of freedom, 61, 160
- discrete bivariate distributions, 97–99
- discrete random variables
distribution of sum of, 32–34
expectation of a function of, 4, 18
and probability generating function, 27–31
- distributions, 22
of functions of continuous random variables, 43–45
- geometric, 22–24
- mixed exam practice, 35–37
- negative binomial, 24–27
- probability generating functions, 27–34
of sample averages and sums, 16–18
summary, 34
- dummy variable, 28
- efficiency, 53, 68, 160
- errors in hypothesis testing, 88–90, 94
- estimators
biased, 48, 67, 159
unbiased, 48–49, 51–53, 162
- expectation
function of continuous random variables, 4, 18
sample mean and sample sum, 9–12
- extreme values, 5, 7
- geometric distributions, 22–24
- hypothesis testing, 71–73
for a mean with known variance, 78–81
for a mean with unknown variance, 81–84
calculator skills, 141–47, 154–56
errors in, 88–93
examination practice, 95–96
paired samples, 85–87
summary, 93–94
- hypothesis tests
definition, 160
one-tailed test, 71
p-value of, 72
significance level of, 72
t-test, 81–82
two-tailed test, 71
type I and type II errors, 88
Z-test, 78–79
- independent (random) variables, mean and variance
effect of adding/multiplying, 5–7
from same population, 6–7
- introductory problem, 1, 115
- inverse normal distribution, 79, 94
- law of diminishing returns, 10
- least squares regression, 107–11, 112
- line of best fit, 107–8, 112
definition, 160
see also regression line
- linear combinations of normal variables, 12–15, 19
- linear regression, 107–11, 112
- mean of a sample *see* sample mean
- mean with known variance
confidence interval for, 55–59
hypothesis testing, 78–81

mean with unknown variance
 confidence interval for, 63–67
 hypothesis testing, 81–84
 median, finding, 39–41
 mixed exam practice, 117–22
 bivariate distributions, 113–14
 combining random variables, 20–21
 cumulative distribution functions, 47
 hypothesis testing, 95–96
 normal distribution, 35–37
 statistical distributions, 35–37
 unbiased estimators and confidence intervals, 69–70
 mode, finding, 40–41
 negative binomial distributions, 24–27
 probability generating function, 33
 negative correlation, 100, 104
 normal distribution, 2
 central limit theorem, 16–18, 19
 linear combination of normal variables, 12–15
 mixed exam practice, 35–37
 null hypothesis, 71, 93, 160
 and errors in hypothesis testing, 88–89
 one-tailed test, 71, 93
 definition, 160
 worked examples, 74, 75–76
p-value, 72, 73, 74
 definition, 93, 161
 paired samples, 85, 94
 Pearson's product moment correlation coefficient, 102
 percentiles, finding, 39
 point estimate, 55, 68, 160
 Poisson distribution, 13
 population correlation coefficient, 111
 population mean
 confidence interval for, 55–59
 unbiased estimates, 48–51
 population variance, unbiased estimates, 48–51
 positive correlation, 100–1, 104
 probability density function, 38–45
 probability distribution *see* distributions
 probability generating function, 27–28, 34, 160
 to find distribution of sum of discrete random variables, 32–33
 worked examples & exercises, 29–32, 33–34
 probability mass function, 22, 25, 31, 34, 38, 46
*p*th percentage point, 39, 61, 160
 quartiles, finding, 39
 random variables, combining, 2
 distribution resulting from, 16–18
 effect of a constant, 2–5
 effect of adding variables, 5–9
 examination practice, 20–21
 expectation and variance of sample mean and sample sum, 9–12
 linear combinations of normal variables, 12–15
 summary, 18–19
 range, 48
 regression line, 107–9
 definition, 161
 exercises, 109–11
 finding equation of, 157–58

rejection region, 72
 definition, 161
 and errors, 88, 89
 exam hint, 76
 worked example, 79
 sample mean, 9, 67
 confidence interval, 55–56
 definition, 161
 distribution of, 16–18
 expectation and variance of, 9–12, 18
 unbiased estimator of population mean, 48, 51
 sample product moment correlation coefficient, 103, 111
 sample size for applying central limit theorem, 16
 sample sum
 distribution of, 16–18
 expectation and variance of, 10, 19
 sample variance, 48, 52, 67
 scatter diagrams, correlation, 100
 significance level, 72
 definition, 161
 exercises, 77
 worked examples, 73–76
see also hypothesis testing
 standard deviation, 5–6, 10, 48
 estimate of, *t*-distribution, 60, 63
 and paired samples, 85
 sum of a sample *see* sample sum
t-distribution, 60–62
 calculator skills, 132–33
 definition, 162
 use of, 63–64
T-score, 68, 81, 94
 calculator skills, 134–35
 formula, 60
t-test, 81–82, 162
 test statistic, 72, 78, 81, 88, 94, 161
 two-tailed test, 71, 93
 definition, 161
 worked examples, 73–75
 type I error, 88–89, 94
 definition, 162
 exercises, 91–93
 type II error, 88–90, 94, 162
 unbiased estimator, 48–49
 definition, 67, 162
 exercises, 54–55
 theory of, 51–53
 variance
 biased estimator of standard deviation squared, 48
 effect of adding a constant, 2–3
 and independent random variables, 5–7
 of negative binomial distribution, 25, 27
 of sample mean, 9–10
 of sample sum, 10
 unbiased estimates, 49
 using probability generating functions, 30
see also covariance; hypothesis testing

Z-score, 56, 57, 78, 94
 formula, 60
Z-test, 78–79, 94, 162