Endorsed for
Pearson Edexcel
Qualifications

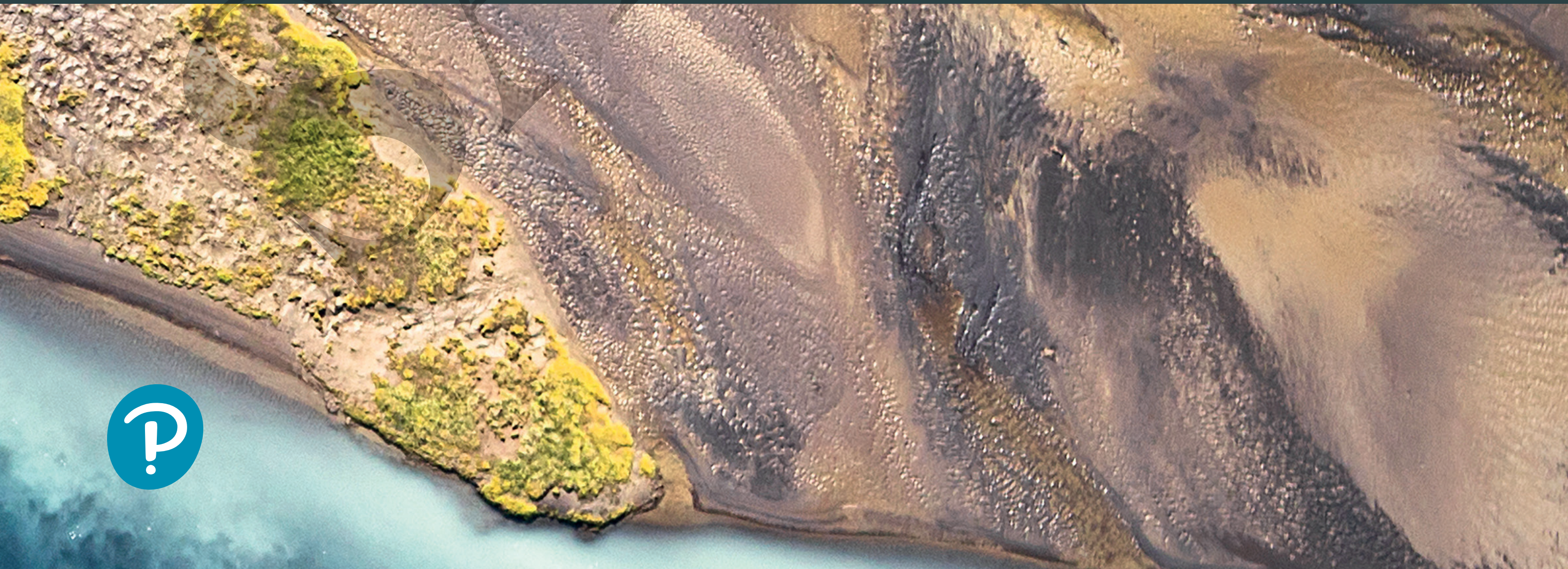
MATHEMATICS

eBook
included

PEARSON EDEXCEL INTERNATIONAL A LEVEL

STATISTICS 3

STUDENT BOOK



PEARSON EDEXCEL INTERNATIONAL A LEVEL

STATISTICS 3

Student Book

Series Editors: Joe Skrakowski and Harry Smith

Authors: Greg Attwood, Tom Begley, Ian Bettison, Alan Clegg, Martin Crozier, Gill Dyer, Jane Dyer, Keith Gallick, Susan Hooker, Michael Jennings, John Kinoulty, Guilherme Frederico Lima, Jean Littlewood, Bronwen Moran, James Nicholson, Su Nicholson, Laurence Pateman, Keith Pledger, Joe Skrakowski, Harry Smith

Published by Pearson Education Limited, 80 Strand, London, WC2R 0RL.

www.pearsonglobalschools.com

Copies of official specifications for all Pearson qualifications may be found on the website: https://qualifications.pearson.com

Text © Pearson Education Limited 2020
Edited by Eric Pradel
Typeset by Tech-Set Ltd, Gateshead, UK
Original illustrations © Pearson Education Limited 2020
Illustrated by © Tech-Set Ltd, Gateshead, UK
Cover design by © Pearson Education Limited 2020

The rights of Greg Attwood, Tom Begley, Ian Bettison, Alan Clegg, Martin Crozier, Gill Dyer, Jane Dyer, Keith Gallick, Susan Hooker, Michael Jennings, John Kinoulty, Guilherme Frederico Lima, Jean Littlewood, Bronwen Moran, James Nicholson, Su Nicholson, Laurence Pateman, Keith Pledger, Joe Skrakowski and Harry Smith to be identified as the authors of this work have been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

First published 2020

23 22 21 20
10 9 8 7 6 5 4 3 2 1

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 978 1 292245 18 8

Printed in Slovakia by Neografia

Picture Credits
The authors and publisher would like to thank the following individuals and organisations for permission to reproduce photographs:

Shutterstock.com: MPVAN 19, Vit Kovalcik 70, Leigh Prather 88; **Alamy Stock Photo:** Hero Images Inc. 1, PCN Photography 10; **Getty Images:** DuKai photographer 48

Cover images: *Front*: **Getty Images:** Werner Van Steen
Inside front cover: **Shutterstock.com:** Dmitry Lobanov

All other images © Pearson Education Limited 2020
All artwork © Pearson Education Limited 2020

Endorsement Statement
In order to ensure that this resource offers high-quality support for the associated Pearson qualification, it has been through a review process by the awarding body. This process confirms that this resource fully covers the teaching and learning content of the specification or part of a specification at which it is aimed. It also confirms that it demonstrates an appropriate balance between the development of subject skills, knowledge and understanding, in addition to preparation for assessment.

Endorsement does not cover any guidance on assessment activities or processes (e.g. practice questions or advice on how to answer assessment questions) included in the resource, nor does it prescribe any particular approach to the teaching or delivery of a related course.

While the publishers have made every attempt to ensure that advice on the qualification and its assessment is accurate, the official specification and associated assessment guidance materials are the only authoritative source of information and should always be referred to for definitive guidance.

Pearson examiners have not contributed to any sections in this resource relevant to examination papers for which they have responsibility.

Examiners will not use endorsed resources as a source of material for any assessment set by Pearson. Endorsement of a resource does not mean that the resource is required to achieve this Pearson qualification, nor does it mean that it is the only suitable material available to support the qualification, and any resource lists produced by the awarding body shall include this and other appropriate resources.

ABOUT THIS BOOK

The following three themes have been fully integrated throughout the Pearson Edexcel International Advanced Level in Mathematics series, so they can be applied alongside your learning.

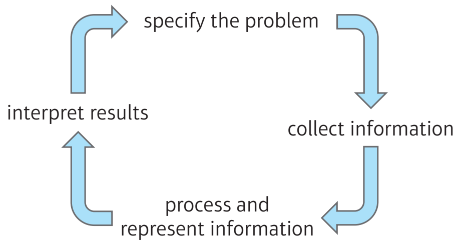
1. Mathematical argument, language and proof

- Rigorous and consistent approach throughout
- Notation boxes explain key mathematical language and symbols

2. Mathematical problem-solving

- Hundreds of problem-solving questions, fully integrated into the main exercises
- Problem-solving boxes provide tips and strategies
- Challenge questions provide extra stretch

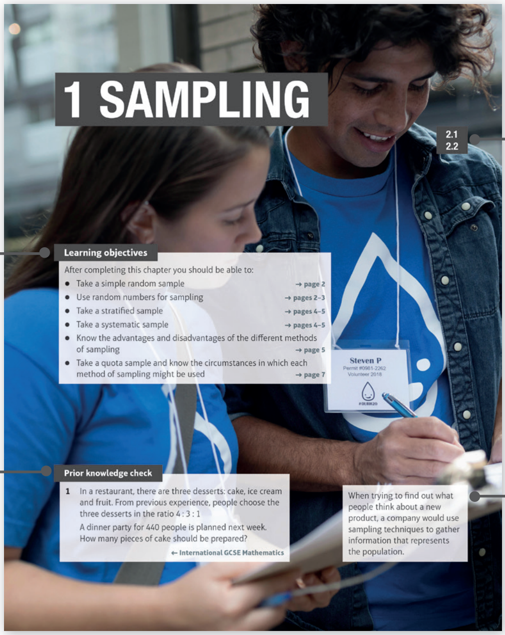
The Mathematical Problem-Solving Cycle



3. Transferable skills

- Transferable skills are embedded throughout this book, in the exercises and in some examples
- These skills are signposted to show students which skills they are using and developing

Finding your way around the book



1 SAMPLING

2.1
2.2

Learning objectives

After completing this chapter you should be able to:

- Take a simple random sample → page 2
- Use random numbers for sampling → pages 2–3
- Take a stratified sample → pages 4–5
- Take a systematic sample → pages 4–5
- Know the advantages and disadvantages of the different methods of sampling → page 5
- Take a quota sample and know the circumstances in which each method of sampling might be used → page 7

Prior knowledge check

1 In a restaurant, there are three desserts: cake, ice cream and fruit. From previous experience, people choose the three desserts in the ratio 4 : 3 : 1
A dinner party for 440 people is planned next week. How many pieces of cake should be prepared? ← International GCSE Mathematics

When trying to find out what people think about a new product, a company would use sampling techniques to gather information that represents the population.

Each chapter starts with a list of *Learning objectives*

The *Prior knowledge check* helps make sure you are ready to start the chapter

Glossary terms will be identified by bold blue text on their first appearance

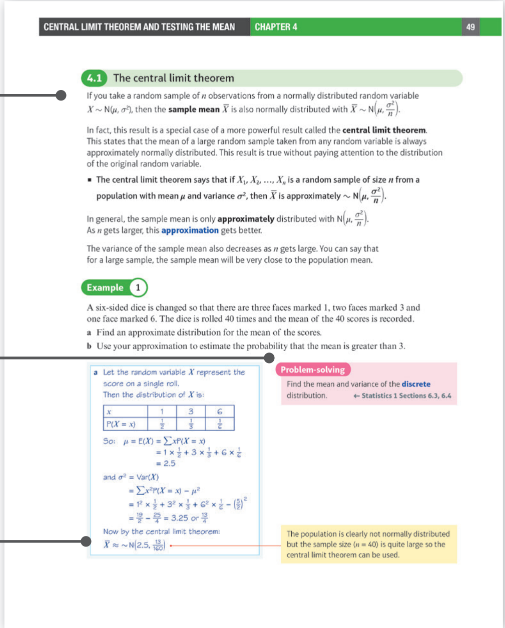
Each chapter is mapped to the specification content for easy reference

The real world applications of the maths you are about to learn are highlighted at the start of the chapter

Each section begins with an explanation and key learning points

Problem-solving boxes provide hints, tips and strategies, and Watch out boxes highlight areas where students often lose marks in their exams

Step-by-step worked examples focus on the key types of questions you'll need to tackle



CENTRAL LIMIT THEOREM AND TESTING THE MEAN CHAPTER 4 49

4.1 The central limit theorem

If you take a random sample of n observations from a normally distributed random variable $X \sim \mathrm{N}(\mu, \sigma^2)$, then the **sample mean** $\bar{X}$ is also normally distributed with $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

In fact, this result is a special case of a more powerful result called the **central limit theorem**. This states that the mean of a large random sample taken from any random variable is always approximately normally distributed. This result is true without paying attention to the distribution of the original random variable.

- **The central limit theorem says that if $X_1, X_2, \ldots, X_n$ is a random sample of size n from a population with mean μ and variance σ^2, then $\bar{X}$ is approximately $\sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$.**

In general, the sample mean is only **approximately** distributed with $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$. As n gets larger, this **approximation** gets better.

The variance of the sample mean also decreases as n gets large. You can say that for a large sample, the sample mean will be very close to the population mean.

Example 1

A six-sided dice is changed so that there are three faces marked 1, two faces marked 3 and one face marked 6. The dice is rolled 40 times and the mean of the 40 scores is recorded.

a Find an approximate distribution for the mean of the scores.

b Use your approximation to estimate the probability that the mean is greater than 3.

a Let the random variable X represent the score on a single roll.
Then the distribution of X is:

x	1	3	6
$\mathrm{P}(X = x)$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

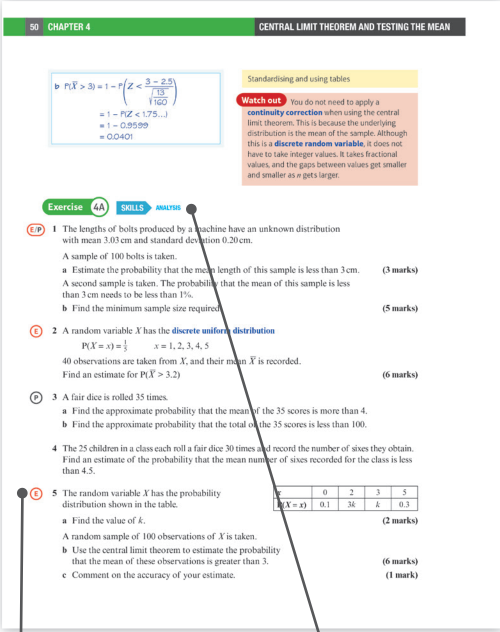
So: $\mu = \mathrm{E}(X) = \sum x\mathrm{P}(X = x)$
$= 1 \times \frac{1}{2} + 3 \times \frac{1}{3} + 6 \times \frac{1}{6}$
$= 2.5$

and $\sigma^2 = \mathrm{Var}(X)$
$= \sum x^2\mathrm{P}(X = x) - \mu^2$
$= 1^2 \times \frac{1}{2} + 3^2 \times \frac{1}{3} + 6^2 \times \frac{1}{6} - \left(\frac{5}{2}\right)^2$
$= \frac{19}{2} - \frac{25}{4} = 3.25$ or $\frac{13}{4}$

Now by the central limit theorem:
$\bar{X} \approx \sim \mathrm{N}\left(2.5, \frac{13}{160}\right)$

Problem-solving
Find the mean and variance of the **discrete** distribution. ← Statistics 1 Sections 6.3, 6.4

The population is clearly not normally distributed but the sample size ($n = 40$) is quite large so the central limit theorem can be used.



50 CHAPTER 4 CENTRAL LIMIT THEOREM AND TESTING THE MEAN

b $\mathrm{P}(\bar{X} > 3) = 1 - \mathrm{P}\left(Z < \frac{3 - 2.5}{\sqrt{\frac{13}{160}}}\right)$
$= 1 - \mathrm{P}(Z < 1.75\ldots)$
$= 1 - 0.9599$
$= 0.0401$

Standardising and using tables

Watch out You do not need to apply a **continuity correction** when using the central limit theorem. This is because the underlying distribution is the mean of the sample. Although this is a **discrete random variable**, it does not have to take integer values. It takes fractional values, and the gaps between values get smaller and smaller as n gets larger.

Exercise 4A SKILLS ANALYSIS

E/P **1** The lengths of bolts produced by a machine have an unknown distribution with mean 3.03 cm and standard deviation 0.20 cm.
A sample of 100 bolts is taken.
a Estimate the probability that the mean length of this sample is less than 3 cm. **(3 marks)**
A second sample is taken. The probability that the mean of this sample is less than 3 cm needs to be less than 1%.
b Find the minimum sample size required. **(5 marks)**

E **2** A random variable X has the **discrete uniform distribution**
$\mathrm{P}(X = x) = \frac{1}{5}$ $\quad x = 1, 2, 3, 4, 5$
40 observations are taken from X, and their mean $\bar{X}$ is recorded.
Find an estimate for $\mathrm{P}(\bar{X} > 3.2)$ **(6 marks)**

P **3** A fair dice is rolled 35 times.
a Find the approximate probability that the mean of the 35 scores is more than 4.
b Find the approximate probability that the total of the 35 scores is less than 100.

4 The 25 children in a class each roll a fair dice 30 times and record the number of sixes they obtain. Find an estimate of the probability that the mean number of sixes recorded for the class is less than 4.5.

E **5** The random variable X has the probability distribution shown in the table.

x	0	2	3	5
$\mathrm{P}(X = x)$	0.1	$3k$	k	0.3

a Find the value of k. **(2 marks)**
A random sample of 100 observations of X is taken.
b Use the central limit theorem to estimate the probability that the mean of these observations is greater than 3. **(6 marks)**
c Comment on the accuracy of your estimate. **(1 mark)**

Exercise questions are carefully graded so they increase in difficulty and gradually bring you up to exam standard

Exercises are packed with exam-style questions to ensure you are ready for the exams

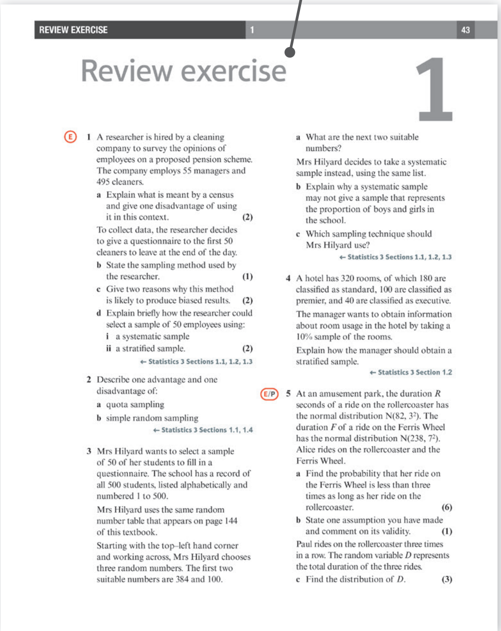
Transferable skills are signposted where they naturally occur in the exercises and examples

Exam-style questions are flagged with E

Problem-solving questions are flagged with P

Each chapter ends with a *Chapter review* and a *Summary of key points*

After every few chapters, a *Review exercise* helps you consolidate your learning with lots of exam-style questions



REVIEW EXERCISE 1 43

Review exercise 1

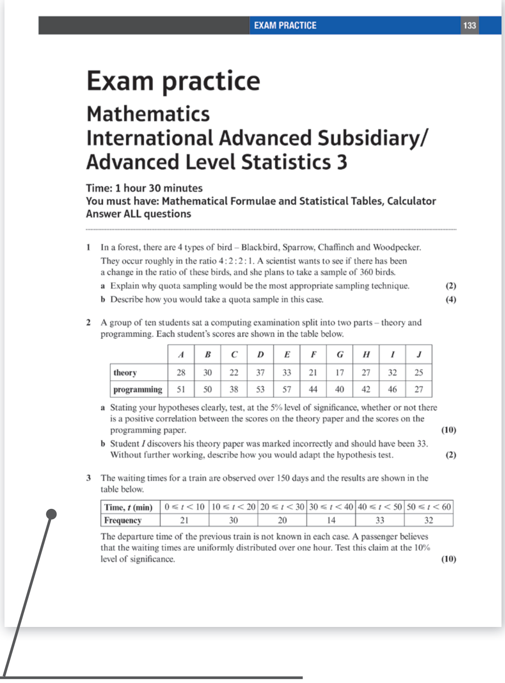
E **1** A researcher is hired by a cleaning company to survey the opinions of employees on a proposed pension scheme. The company employs 55 managers and 495 cleaners.
a Explain what is meant by a census and give one disadvantage of using it in this context. **(2)**
To collect data, the researcher decides to give a questionnaire to the first 50 cleaners to leave at the end of the day.
b State the sampling method used by the researcher. **(1)**
c Give two reasons why this method is likely to produce biased results. **(2)**
d Explain briefly how the researcher could select a sample of 50 employees using:
i a systematic sample
ii a stratified sample. **(2)**
← Statistics 3 Sections 1.1, 1.2, 1.3

2 Describe one advantage and one disadvantage of:
a quota sampling
b simple random sampling
← Statistics 3 Sections 1.1, 1.4

3 Mrs Hilyard wants to select a sample of 50 of her students to fill in a questionnaire. The school has a record of all 500 students, listed alphabetically and numbered 1 to 500.
Mrs Hilyard uses the same random number table that appears on page 144 of this textbook.
Starting with the top–left hand corner and working across, Mrs Hilyard chooses three random numbers. The first two suitable numbers are 384 and 100.
a What are the next two suitable numbers?
Mrs Hilyard decides to take a systematic sample instead, using the same list.
b Explain why a systematic sample may not give a sample that represents the proportion of boys and girls in the school.
c Which sampling technique should Mrs Hilyard use?
← Statistics 3 Sections 1.1, 1.2, 1.3

4 A hotel has 320 rooms, of which 180 are classified as standard, 100 are classified as premier, and 40 are classified as executive.
The manager wants to obtain information about room usage in the hotel by taking a 10% sample of the rooms.
Explain how the manager should obtain a stratified sample.
← Statistics 3 Section 1.2

E/P **5** At an amusement park, the duration R seconds of a ride on the rollercoaster has the normal distribution $\mathrm{N}(82, 3^2)$. The duration F of a ride on the Ferris Wheel has the normal distribution $\mathrm{N}(238, 7^2)$.
Alice rides on the rollercoaster and the Ferris Wheel.
a Find the probability that her ride on the Ferris Wheel is less than three times as long as her ride on the rollercoaster. **(6)**
b State one assumption you have made and comment on its validity. **(1)**
Paul rides on the rollercoaster three times in a row. The random variable D represents the total duration of the three rides.
c Find the distribution of D. **(3)**



EXAM PRACTICE 133

Exam practice

Mathematics
International Advanced Subsidiary/ Advanced Level Statistics 3

Time: 1 hour 30 minutes
You must have: Mathematical Formulae and Statistical Tables, Calculator
Answer ALL questions

1 In a forest, there are 4 types of bird – Blackbird, Sparrow, Chaffinch and Woodpecker.
They occur roughly in the ratio 4 : 2 : 2 : 1. A scientist wants to see if there has been a change in the ratio of these birds, and she plans to take a sample of 360 birds.
a Explain why quota sampling would be the most appropriate sampling technique. **(2)**
b Describe how you would take a quota sample in this case. **(4)**

2 A group of ten students sat a computing examination split into two parts – theory and programming. Each student's scores are shown in the table below.

	A	B	C	D	E	F	G	H	I	J
theory	28	30	22	37	33	21	17	27	32	25
programming	51	50	38	53	57	44	40	42	46	27

a Stating your hypotheses clearly, test, at the 5% level of significance, whether or not there is a positive correlation between the scores on the theory paper and the scores on the programming paper. **(10)**
b Student I discovers his theory paper was marked incorrectly and should have been 33. Without further working, describe how you would adapt the hypothesis test. **(2)**

3 The waiting times for a train are observed over 150 days and the results are shown in the table below.

Time, t (min)	$0 \leqslant t < 10$	$10 \leqslant t < 20$	$20 \leqslant t < 30$	$30 \leqslant t < 40$	$40 \leqslant t < 50$	$50 \leqslant t < 60$
Frequency	21	30	20	14	33	32

The departure time of the previous train is not known in each case. A passenger believes that the waiting times are uniformly distributed over one hour. Test this claim at the 10% level of significance. **(10)**

A full practice paper at the back of the book helps you prepare for the real thing

QUALIFICATION AND ASSESSMENT OVERVIEW

Qualification and content overview

Statistics 3 (S3) is an **optional** unit in the following qualifications:

International Advanced Subsidiary in Further Mathematics

International Advanced Level in Further Mathematics

Assessment overview

The following table gives an overview of the assessment for this unit.

We recommend that you study this information closely to help ensure that you are fully prepared for this course and know exactly what to expect in the assessment.

Unit	Percentage	Mark	Time	Availability
S3: Statistics 3 Paper code WST03/01	$33\frac{1}{3}$% of IAS $16\frac{2}{3}$% of IAL	75	1 hour 30 mins	June First assessment June 2020

IAS: International Advanced Subsidiary, IAL: International Advanced A Level.

Assessment objectives and weightings

		Minimum weighting in IAS and IAL
AO1	Recall, select and use their knowledge of mathematical facts, concepts and techniques in a variety of contexts.	30%
AO2	Construct rigorous mathematical arguments and proofs through use of precise statements, logical deduction and inference and by the manipulation of mathematical expressions, including the construction of extended arguments for handling substantial problems presented in unstructured form.	30%
AO3	Recall, select and use their knowledge of standard mathematical models to represent situations in the real world; recognise and understand given representations involving standard models; present and interpret results from such models in terms of the original situation, including discussion of the assumptions made and refinement of such models.	10%
AO4	Comprehend translations of common realistic contexts into mathematics; use the results of calculations to make predictions, or comment on the context; and, where appropriate, read critically and comprehend longer mathematical arguments or examples of applications.	5%
AO5	Use contemporary calculator technology and other permitted resources (such as formulae booklets or statistical tables) accurately and efficiently; understand when not to use such technology, and its limitations. Give answers to appropriate accuracy.	5%

Relationship of assessment objectives to units

	Assessment objective				
S3	AO1	AO2	AO3	AO4	AO5
Marks out of 75	25–30	20–25	10–15	5–10	5–10
%	$33\frac{1}{3}$–40	$26\frac{2}{3}$–$33\frac{1}{3}$	$13\frac{1}{3}$–20	$6\frac{2}{3}$–$13\frac{1}{3}$	$6\frac{2}{3}$–$13\frac{1}{3}$

Calculators

Students may use a calculator in assessments for these qualifications. Centres are responsible for making sure that calculators used by their students meet the requirements given in the table below.

Students are expected to have available a calculator with at least the following keys: +, –, ×, ÷, π, x^2, $\sqrt{x}$, $\frac{1}{x}$, x^y, ln x, e^x, $x!$, sine, cosine and tangent and their inverses in degrees and decimals of a degree, and in radians; memory.

Prohibitions

Calculators with any of the following facilities are prohibited in all examinations:

- databanks
- retrieval of text or formulae
- built-in symbolic algebra manipulations
- symbolic differentiation and/or integration
- language translators
- communication with other machines or the internet

Extra online content

Whenever you see an *Online* box, it means that there is extra online content available to support you.

SolutionBank

SolutionBank provides worked solutions for questions in the book.
Download the solutions as a PDF or quickly find the solution you need online.

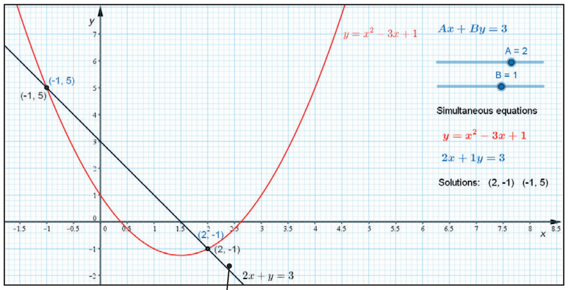
Use of technology

Explore topics in more detail, visualise problems and consolidate your understanding. Use pre-made GeoGebra activities or Casio resources for a graphic calculator.

Online Find the point of intersection graphically using technology.

GeoGebra

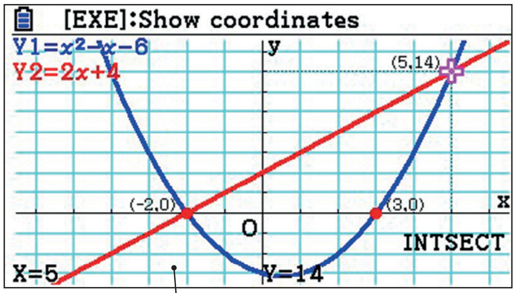
GeoGebra-powered interactives



Interact with the maths you are learning using GeoGebra's easy-to-use tools

CASIO®

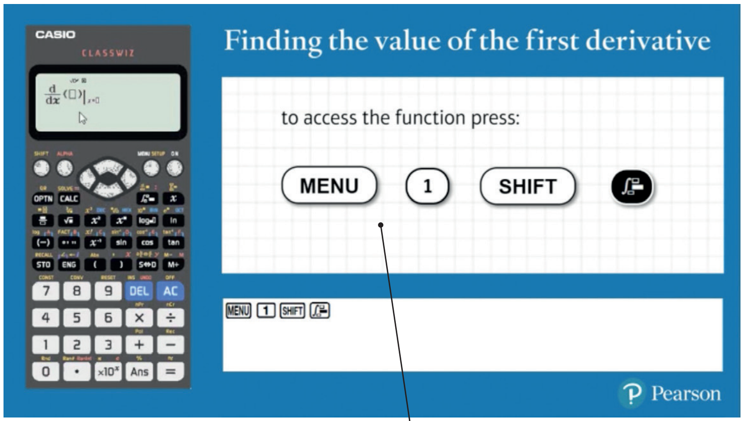
Graphic calculator interactives



Explore the maths you are learning and gain confidence in using a graphic calculator

Calculator tutorials

Our helpful video tutorials will guide you through how to use your calculator in the exams. They cover both Casio's scientific and colour graphic calculators.

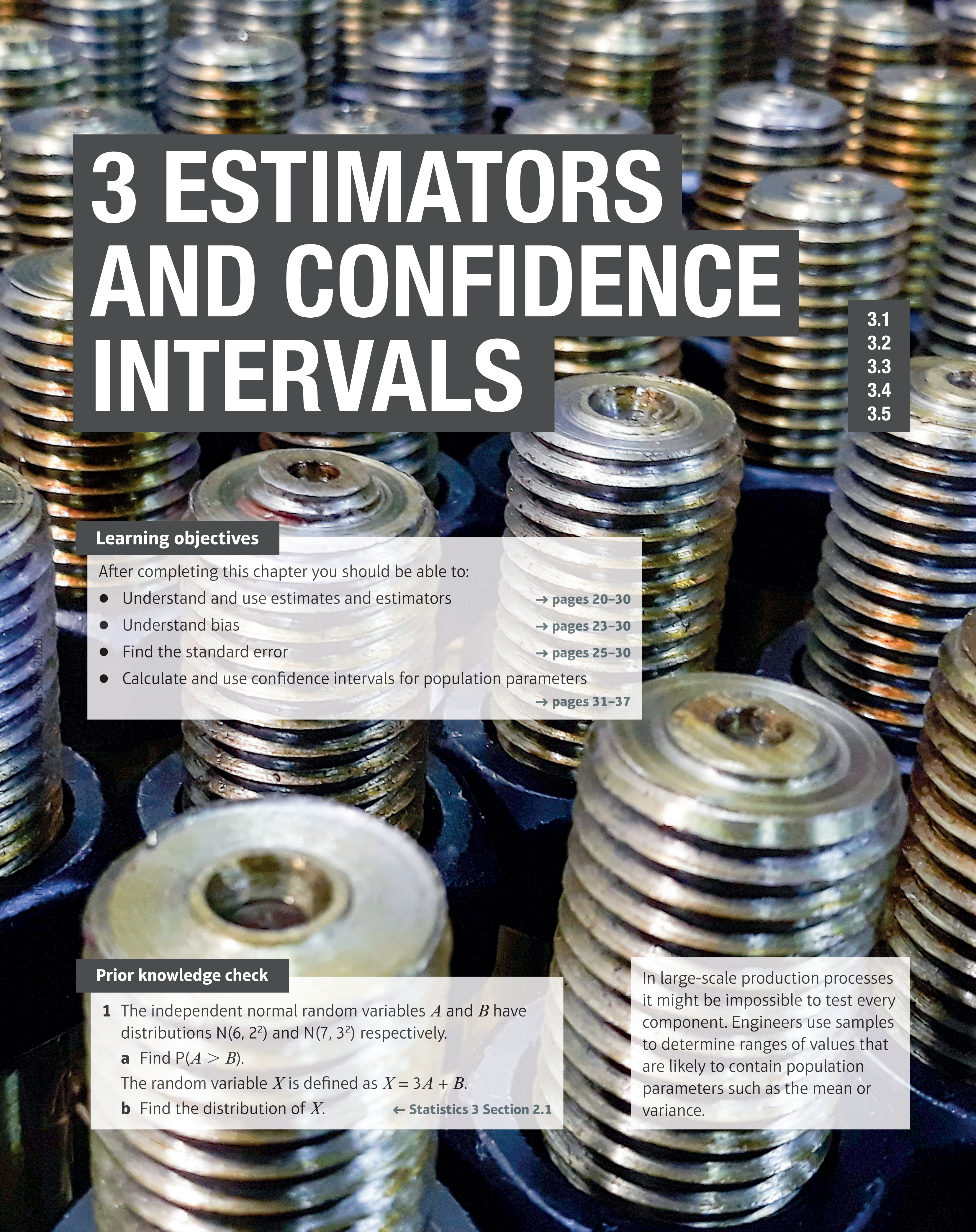


Finding the value of the first derivative

to access the function press:

Step-by-step guide with audio instructions on exactly which buttons to press and what should appear on your calculator's screen

Online Work out each coefficient quickly using the ${}^nC_r$ and power functions on your calculator.



3 ESTIMATORS AND CONFIDENCE INTERVALS

3.1
3.2
3.3
3.4
3.5

Learning objectives

After completing this chapter you should be able to:

- Understand and use estimates and estimators → pages 20–30
- Understand bias → pages 23–30
- Find the standard error → pages 25–30
- Calculate and use confidence intervals for population parameters → pages 31–37

Prior knowledge check

1 The independent normal random variables A and B have distributions $\mathrm{N}(6, 2^2)$ and $\mathrm{N}(7, 3^2)$ respectively.

a Find $\mathrm{P}(A > B)$.

The random variable X is defined as $X = 3A + B$.

b Find the distribution of X. ← Statistics 3 Section 2.1

In large-scale production processes it might be impossible to test every component. Engineers use samples to determine ranges of values that are likely to contain population parameters such as the mean or variance.

3.1 Estimators, bias and standard error

In a large population (e.g. the number of trees in a forest), it would take too long or cost too much money to carry out a census (e.g. to record the height of every tree). In cases like this, **population parameters** such as the mean μ or the standard deviation σ are likely to be unknown.

In Chapter 1, you looked at methods of sampling that allow you to take a representative sample to estimate various population parameters.

A common way of estimating population parameters is to take a **random sample** from the population.

Links A census observes every member of a population, whereas a sample is a selection of observations taken from a subset of the population. ← **Statistics 2 Section 6.1**

- **If X is a random variable, then a random sample of size n will consist of n observations of the random variable X. These are referred to as $X_1, X_2, X_3, \ldots, X_n$, where the X_i:**
  - **are independent random variables**
  - **each have the same distribution as X.**
- **A statistic T is defined as a function of the X_i that involves no other quantities, such as unknown population parameters.**

Notation X_i represents the ith observation of a sample. The value of the observation is denoted by x_i.

For example, $\overline{X}$, the sample mean, is a statistic, whereas $\sum_{i=1}^{n} \frac{X_i^2}{n} - \mu^2$ is not a statistic since it involves the unknown population parameter μ.

Example 1 SKILLS REASONING/ARGUMENTATION

A sample $X_1, X_2, \ldots, X_n$ is taken from a population with unknown population parameters μ and σ. State whether or not each of the following are statistics.

a $\frac{X_1 + X_3 + X_5}{3}$ **b** $\max(X_1, X_2, \ldots, X_n)$ **c** $\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$

a $\frac{X_1 + X_3 + X_5}{3}$ is a statistic.

It is only a function of the sample $X_1, X_2, \ldots, X_n$. A statistic need not involve all members of the sample.

b $\max(X_1, X_2, \ldots, X_n)$ is a statistic.

It is only a function of the sample $X_1, X_2, \ldots, X_n$.

c $\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$ is not a statistic.

The function contains μ and σ.

Since it is possible to repeat the process of taking a sample, the specific value of a statistic T will be different for each sample. If all possible samples are taken, these values will form a probability distribution called the **sampling distribution** of T.

- **The sampling distribution of a statistic T is the probability distribution of T.**

If the distribution of the population is known, then the sampling distribution of a statistic can sometimes be found.

Example 2

The masses, in grams, of boxes of apples are normally distributed with a mean μ and standard deviation 4. A random sample of size 25 is taken and the statistics R and T are calculated as follows:

$$R = X_{25} - X_1 \text{ and } T = X_1 + X_2 + \ldots + X_{25}$$

Find the distributions of R and T.

The sample will be $X_1, X_2, \ldots, X_{25}$ where each $X_i \sim \mathrm{N}(\mu, 4^2)$

State the distribution for each of the X_i.

Now $R = X_{25} - X_1 \Rightarrow R \sim \mathrm{N}(\mu - \mu, 4^2 + 4^2)$

that is $R \sim \mathrm{N}(0, (4\sqrt{2})^2)$

E$(X - Y)$ = E(X) − E(Y). Since each observation in a random sample is independent, you can also use Var$(X - Y)$ = Var(X) + Var(Y).

← Statistics 3 Section 2.1

$T = X_1 + X_2 + \ldots + X_{25}$

so $T \sim \mathrm{N}(25\mu, 25 \times 4^2)$

or $T \sim \mathrm{N}(25\mu, 20^2)$

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

← Statistics 3 Section 2.1

Example 3

SKILLS CREATIVITY

In a bag that contains a large number of counters, the number 0 is written on 60% of the counters, and the number 1 is written on the other 40%.

a Find the population mean μ and population variance σ^2 of the values shown on the counters.

A simple random sample of size 3 is taken from this population.

b List all the possible observations from this sample.

c Find the sampling distribution for the mean

$$\bar{X} = \frac{X_1 + X_2 + X_3}{3}$$

where X_1, X_2 and X_3 are the values shown on the three counters in the sample.

d Hence find E$(\bar{X})$ and Var$(\bar{X})$.

e Find the sampling distribution for the sample **mode** M.

f Hence find E(M) and Var(M).

a If X represents the value shown on a randomly chosen counter, then X has distribution:

x	0	1
P$(X = x)$	$\frac{3}{5}$	$\frac{2}{5}$

$\mu = \mathrm{E}(X) = \sum x\mathrm{P}(X = x) = 0 + \frac{2}{5} \Rightarrow \mu = \frac{2}{5}$

$\sigma^2 = \mathrm{Var}(X) = \sum x^2\mathrm{P}(X = x) - \mu^2 = 0 + 1^2 \times \frac{2}{5} - \frac{4}{25} \Rightarrow \sigma^2 = \frac{6}{25}$

b The possible observations are

(0, 0, 0)

(1, 0, 0) (0, 1, 0) (0, 0, 1)

(1, 1, 0) (1, 0, 1) (0, 1, 1)

(1, 1, 1)

List these systematically.

c $\mathrm{P}(\overline{X} = 0) = \left(\frac{3}{5}\right)^3 = \frac{27}{125}$ i.e. the (0, 0, 0) case

$\mathrm{P}\left(\overline{X} = \frac{1}{3}\right) = 3 \times \frac{2}{5} \times \left(\frac{3}{5}\right)^2 = \frac{54}{125}$ i.e. the (1, 0, 0), (0, 1, 0), (0, 0, 1) cases

$\mathrm{P}\left(\overline{X} = \frac{2}{3}\right) = 3 \times \left(\frac{2}{5}\right)^2 \times \frac{3}{5} = \frac{36}{125}$ i.e. the (1, 1, 0), (1, 0, 1), (0, 1, 1) cases

$\mathrm{P}(\overline{X} = 1) = \left(\frac{2}{5}\right)^3 = \frac{8}{125}$ i.e. the (1, 1, 1) case

So the distribution for $\overline{X}$ is

$\overline{x}$	0	$\frac{1}{3}$	$\frac{2}{3}$	1
$p(\overline{x})$	$\frac{27}{125}$	$\frac{54}{125}$	$\frac{36}{125}$	$\frac{8}{125}$

Since the sample is random, the observations are independent. So to find the probability of case (1, 0, 0) you can multiply the probabilities $\mathrm{P}(X_1 = 1) \times \mathrm{P}(X_2 = 0) \times \mathrm{P}(X_3 = 0)$. Remember that each X_i has the same distribution as X.

← **Statistics 2 Chapter 1**

d $\mathrm{E}(\overline{X}) = 0 + \frac{1}{3} \times \frac{54}{125} + \frac{2}{3} \times \frac{36}{125} + 1 \times \frac{8}{125} = \frac{18 + 24 + 8}{125} = \frac{2}{5}$

$\mathrm{Var}(\overline{X}) = 0 + \frac{1}{9} \times \frac{54}{125} + \frac{4}{9} \times \frac{36}{125} + 1 \times \frac{8}{125} - \frac{4}{25} = \frac{6 + 16 + 8}{125} - \frac{20}{125} = \frac{2}{25}$

e The sample mode can take values 0 or 1.

$\mathrm{P}(M = 0) = \frac{27}{125} + \frac{54}{125} = \frac{81}{125}$ i.e. cases (0, 0, 0), (1, 0, 0), (0, 1, 0), (0, 0, 1)

and $\mathrm{P}(M = 1) = \frac{44}{125}$ i.e. the other cases

so the distribution of M is

m	0	1
$p(m)$	$\frac{81}{125}$	$\frac{44}{125}$

f $\mathrm{E}(M) = 0 + 1 \times \frac{44}{125} = \frac{44}{125}$

and $\mathrm{Var}(M) = 0 + 1 \times \frac{44}{125} - \left(\frac{44}{125}\right)^2 = 0.228$

Notation Notice that $\mathrm{E}(\overline{X}) = \mu$ but $\mathrm{E}(M) \neq \mu$ and that neither $\mathrm{E}(\overline{X})$ nor $\mathrm{E}(M)$ is equal to the population mode, which is of course zero as 60% of the counters have a zero on them.

- A statistic that is used to estimate a population parameter is called an **estimator** and the particular value of the estimator generated from the sample taken is called an **estimate**.

You need to be able to determine how **reliable** these sample statistics are as estimators for the **corresponding** population parameters.

Since all the X_i are random variables having the same mean and variance as the population, you can sometimes find expected values of a statistic T, $\mathrm{E}(T)$. This will tell you what the 'average' value of the statistic should be.

Example 4

SKILLS ANALYSIS

A random sample $X_1, X_2, \ldots, X_n$ is taken from a population with $X \sim \mathrm{N}(\mu, \sigma^2)$.

Show that $\mathrm{E}(\overline{X}) = \mu$.

$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$$

$$\mathrm{E}(\bar{X}) = \frac{1}{n}\mathrm{E}(X_1 + \ldots + X_n)$$

Use $\mathrm{E}(aX) = a\mathrm{E}(X)$

$$= \frac{1}{n}(\mathrm{E}(X_1) + \ldots + \mathrm{E}(X_n))$$

$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$

$$= \frac{1}{n}(\mu + \ldots + \mu)$$

$$= \frac{n\mu}{n}$$

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

This example shows that if you use the sample mean as an estimator of the population mean, then 'on average' it will give the correct value.

This is an important property for an estimator to have. You say that $\bar{X}$ is an **unbiased estimator** of μ. A specific value of $\bar{x}$ will be an unbiased estimate for μ.

- **If a statistic T is used as an estimator for a population parameter θ and $\mathrm{E}(T) = \theta$, then T is an unbiased estimator for θ.**

In Example 3, you found two statistics based on samples of size 3 from a population of counters. The two statistics that you calculated were the sample mean $\bar{X}$ and the sample mode M. You could use either of them as estimators for μ, the population mean, but you saw that $\mathrm{E}(\bar{X}) = \mu$ and $\mathrm{E}(M) \neq \mu$. In this case, if you wanted an unbiased estimator for μ, you would choose the sample mean $\bar{X}$ rather than the sample mode M, which we would call a biased estimator. How about an estimator for the population mode? Neither of the statistics that you calculated had the property of being unbiased since $\mathrm{E}(\bar{X}) = \mu = \frac{2}{5}$ and $\mathrm{E}(M) = \frac{44}{125}$, whereas the population mode was 0.

The **bias** is the expected value of the estimator minus the **parameter** of the population it is estimating.

- **If a statistic T is used as an estimator for a population parameter θ, then the bias is $\mathrm{E}(T) - \theta$.**

In this case the bias is $\frac{44}{125}$

- **For an unbiased estimator, the bias is 0.**

In Example 4, the mean of a sample was an unbiased estimator for the population mean. If you take a sample X_1 of size 1 from a population with mean μ and variance σ^2, then the sample mean is $\bar{X} = X_1$, because there is only one value. So $\mathrm{E}(\bar{X}) = \mathrm{E}(X_1) = \mu$.

If you wanted to find an estimator for the **population variance**, you might try using the variance of the sample, $V = \frac{\sum(X_i - \bar{X})^2}{n}$

For our sample X_1 of size 1, the variance of the sample will be $\frac{(X_1 - \bar{X})^2}{1} = (X_1 - X_1)^2 = 0$

Hint In general, the variance of a sample will be an **underestimate** for the variance of the population. This is because the statistic $\frac{\sum(X_i - \bar{X})^2}{n}$ uses the sample mean $\bar{X}$ rather than the population mean μ, and on average the sample observations will be closer to $\bar{X}$ than to μ.

So for a sample of size 1, $\mathrm{E}(V) = 0 \neq \sigma^2$. This illustrates that the variance of the sample is not an unbiased estimator for the variance of the population.

You can use a slightly different statistic, called the **sample variance**, as an unbiased estimator for the population variance.

- **An unbiased estimator for σ^2 is given by the sample variance S^2 where:**

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

Notation S^2 is the estimator (a random variable), and s^2 is the estimate (an observation from this random variable).

There are several ways to calculate the value of s^2 for a particular sample:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \frac{S_{xx}}{n-1} \\ &= \frac{n}{n-1}\left(\frac{\sum x^2}{n} - \overline{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum x^2 - n\overline{x}^2\right) \end{aligned}$$

Links You can use the equivalence of these forms to show that s^2 is an unbiased estimate for σ^2.
→ **Exercise 3A Challenge**

The form that you use will depend on the information that you are given in the question.

Although a sample of size 1 can be used as an unbiased estimator of μ, a single observation from a population will not provide a useful estimate of the population mean. You need some way of distinguishing between the **quality** of different unbiased estimators.

Example 5 SKILLS ANALYSIS

A random sample $X_1, X_2, \ldots, X_n$ is taken from a population with $X \sim \text{N}(\mu, \sigma^2)$.

Show that $\text{Var}(\overline{X}) = \dfrac{\sigma^2}{n}$

$$\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$$

$$\text{Var}(\overline{X}) = \frac{1}{n^2}\text{Var}(X_1 + \ldots + X_n)$$

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

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

Use $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$

$$= \frac{1}{n^2}(\sigma^2 + \ldots + \sigma^2)$$

$$= \frac{n\sigma^2}{n^2}$$

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

One reason that the **sample mean** is used as an estimator for μ is that the variance of the estimator $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$ decreases as n increases. For larger values of n, the value of an estimate is more likely to be close to the population mean. So, a larger value of n will result in a **better** estimator.

- **The standard deviation of an estimator is called the standard error of the estimator.**

When you are using the sample mean $\overline{X}$ you can use the following result for the standard error.

- **Standard error of $\overline{X} = \frac{\sigma}{\sqrt{n}}$ or $\frac{s}{\sqrt{n}}$**

Watch out Although in general $\sigma \neq s$, you can use the second version of this standard error in situations where you **do not know** the population standard deviation.

Example 6

SKILLS ADAPTIVE LEARNING

The table below summarises the number of breakdowns X on a busy road on 30 randomly chosen days.

Number of breakdowns	2	3	4	5	6	7	8	9
Number of days	3	5	4	3	5	4	4	2

a Calculate unbiased estimates of the mean and variance of the number of breakdowns.

Twenty more days were randomly sampled, and this sample had $\bar{x} = 6.0$ days and $s^2 = 5.0$

b Treating the 50 results as a single sample, obtain further unbiased estimates of the population mean and variance.

c Find the standard error of this new estimate of the mean.

d Estimate the size of sample required to achieve a standard error of less than 0.25

a By calculator:

$\sum x = 160$ and $\sum x^2 = 990$

So $\hat{\mu} = \bar{x} = \frac{160}{30} = 5.33$

and $\hat{\sigma}^2 = s_x^2 = \frac{990 - 30\bar{x}^2}{29}$

$= 4.71$ (3 s.f.)

b New sample: $\bar{y} = 6.0 \Rightarrow \sum y = 20 \times 6.0 = 120$

$s_y^2 = 5.0 \Rightarrow \frac{\sum y^2 - 20 \times 6^2}{19} = 5$

So $\sum y^2 = 5 \times 19 + 20 \times 36$

$\Rightarrow \sum y^2 = 815$

So the combined sample (w) of size 50 has

$\sum w = 160 + 120 = 280$

$\sum w^2 = 990 + 815 = 1805$

Notation 'Hat' notation is used to describe an estimate of a parameter. For example: $\hat{\sigma}^2$ represents an estimate for the population variance σ^2. $\hat{\mu}$ represents an estimate for the population mean μ.

Use $s^2 = \frac{1}{n-1}\left(\sum x^2 - n\bar{x}^2\right)$ since you have values for $\sum x^2$ and $\bar{x}$.

Problem-solving First you need to use the formulae for $\bar{y}$ and s_y^2 to find $\sum y$ and $\sum y^2$.

Now combine with $\sum x$ and $\sum x^2$.
Let the combined variable be w.

Hint You can use your calculator to find unbiased estimates of the mean and variance but you should show your working in the exam.

Then the combined estimate of μ is

$$\overline{w} = \frac{280}{50} = 5.6$$

and the estimate for σ^2 is

$$s_w^2 = \frac{1805 - 50 \times 5.6^2}{49}$$

$$\Rightarrow s_w^2 = 4.8367\ldots = 4.84 \text{ (3 s.f.)}$$

c The best estimate of σ^2 will be s_w^2 since it is based on a larger sample than s_x^2 or s_y^2.

So the standard error is $\dfrac{s_w}{\sqrt{50}} = \sqrt{\dfrac{4.836\ldots}{50}}$

$= 0.311$ (3 s.f.)

Use the $\dfrac{s}{\sqrt{n}}$ formula for standard error.

d To achieve a standard error < 0.25 you require

$$\sqrt{\frac{4.836\ldots}{n}} < 0.25$$

You do not know the value for σ so you will have to use your best estimate of it, namely s_w.

$$\Rightarrow \quad \sqrt{n} > \frac{\sqrt{4.836\ldots}}{0.25}$$

$$\sqrt{n} > 8.797\ldots$$

$$\Rightarrow \quad n > 77.38\ldots$$

So we need a sample size of at least 78.

We have seen that for the independent observations $X_i \sim \mathrm{N}(\mu, \sigma^2)$, we can evaluate the statistic

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

In Example 4, we saw that

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

and in Example 5 that

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

Since X is normally distributed and each X_i is an independent observation, $\overline{X}$ must also be normally distributed, so we can create the distribution of the sample mean.

If $X_i \sim \mathrm{N}(\mu, \sigma^2)$ then $\overline{X} \sim \mathrm{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$, where $\dfrac{\sigma}{\sqrt{n}}$ is the standard error.

Example 7

Ten independent observations from $X \sim \mathrm{N}(15, 3^2)$ are taken.

a State the distribution of the sample mean.

b Find $\mathrm{P}(\overline{X} < 14)$

a Since X is normally distributed and the observations are independent,

$$\mathrm{E}(\overline{X}) = \mu = 15$$

and

$$\mathrm{Var}(\overline{X}) = \frac{3^2}{10} = 0.9$$

Therefore,

$$\overline{X} \sim \mathrm{N}(15, 0.9)$$

b Standardising:

$$\mathrm{P}(\overline{X} < 13) = \mathrm{P}\left(Z < \frac{14 - 15}{0.9}\right)$$

$$\mathrm{P}(Z < -1.11\ldots)$$

$$= 1 - \mathrm{P}(Z < 1.11\ldots)$$

$$= 1 - 0.8665$$

$$= 0.1335$$

We use $z = \dfrac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ to standardise.

Using the tables on page 135.

Exercise 3A

SKILLS PROBLEM-SOLVING

1 The lengths of nails produced by a certain machine are normally distributed with mean μ and standard deviation σ. A random sample of 10 nails is taken and their lengths $X_1, X_2, X_3, \ldots, X_{10}$ are measured.

SKILLS INTERPRETATION

i Write down the distributions of the following:

a $\displaystyle\sum_1^{10} X_i$ **b** $\dfrac{2X_1 + 3X_{10}}{5}$ **c** $\displaystyle\sum_1^{10}(X_i - \mu)$

d $\overline{X}$ **e** $\displaystyle\sum_1^{5} X_i - \sum_6^{10} X_i$ **f** $\displaystyle\sum_1^{10}\left(\frac{X_i - \mu}{\sigma}\right)$

ii State which of the above are statistics.

2 A large bag of coins contains 1 cent, 5 cent and 10 cent coins in the ratio 2 : 2 : 1

a Find the mean μ and the variance σ^2 for the value of coins in this population.

A random sample of two coins is taken and their values X_1 and X_2 are recorded.

b List all the possible observations from this sample.

c Find the sampling distribution for the mean $\overline{X} = \dfrac{X_1 + X_2}{2}$

d Hence show that $\mathrm{E}(\overline{X}) = \mu$ and $\mathrm{Var}(\overline{X}) = \dfrac{\sigma^2}{2}$

3 Find unbiased estimates of the mean and variance of the populations from which the following random samples have been taken.

a 21.3 19.6 18.5 22.3 17.4 16.3 18.9 17.6 18.7 16.5 19.3 21.8 20.1 22.0

b 1 2 5 1 6 4 1 3 2 8 5 6 2 4 3 1

c 120.4 230.6 356.1 129.8 185.6 147.6 258.3 329.7 249.3

d 0.862 0.754 0.459 0.473 0.493 0.681 0.743 0.469 0.538 0.361

4 Find unbiased estimates of the mean and the variance of the populations from which random samples with the following summaries have been taken.

a $n = 120$ $\sum x = 4368$ $\sum x^2 = 162\,466$

b $n = 30$ $\sum x = 270$ $\sum x^2 = 2546$

c $n = 1037$ $\sum x = 1140.7$ $\sum x^2 = 1278.08$

d $n = 15$ $\sum x = 168$ $\sum x^2 = 1913$

E **5** The concentrations, in mg per litre, of an element in 7 randomly chosen samples of water from a spring were:

240.8 237.3 236.7 236.6 234.2 233.9 232.5

a Explain what is meant by an unbiased estimator. **(1 mark)**

b Determine unbiased estimates of the mean and the variance of the concentration of the element per litre of water from the spring. **(4 marks)**

E **6** A sample of size 6 is taken from a population that is normally distributed with mean 10 and standard deviation 2.

a Find the probability that the sample mean is greater than 12. **(3 marks)**

b State, with a reason, if your answer is an approximation. **(1 mark)**

7 A machine fills cartons in such a way that the amount of drink in each carton is distributed normally with a mean of 40 cm^3 and a standard deviation of 1.5 cm^3.

A sample of four cartons is examined.

a Find the probability that the mean amount of drink is more than 40.5 cm^3.

A sample of 49 cartons is examined.

b Find the probability that the mean amount of drink is more than 40.5 cm^3 on this occasion.

E **8** Cartons of orange juice are filled by a machine. A sample of 10 cartons selected at random from the production line contained the following quantities of orange juice (in ml).

201.2 205.0 209.1 202.3 204.6 206.4 210.1 201.9 203.7 207.3

Calculate unbiased estimates of the mean and variance of the population from which this sample was taken. **(4 marks)**

9 A manufacturer of self-build furniture required bolts of two lengths, 5 cm and 10 cm, in the ratio 2 : 1 respectively.

a Find the mean μ and the variance σ^2 for the lengths of bolts in this population.

A random sample of three bolts is selected from a large box containing bolts in the required ratio.

b List all the possible observations from this sample.

c Find the sampling distribution for the mean $\overline{X}$.

d Hence find $\mathrm{E}(\overline{X})$ and $\mathrm{Var}(\overline{X})$.

e Find the sampling distribution for the mode M.

f Hence find $\mathrm{E}(M)$ and $\mathrm{Var}(M)$.

g Find the bias when M is used as an estimator of the population mode.

P **10** A biased six-sided dice has probability p of landing on a six.

SKILLS ADAPTIVE LEARNING

Every day, for a period of 25 days, the dice is rolled 10 times and the number of sixes X is recorded, giving rise to a sample $X_1, X_2, \ldots, X_{25}$.

a Write down $\mathrm{E}(X)$ in terms of p.

b Show that the sample mean $\overline{X}$ is a biased estimator of p and find the bias.

c Suggest a suitable unbiased estimator of p.

P **11** The random variable $X \sim \mathrm{U}[-\alpha, \alpha]$.

a Find $\mathrm{E}(X)$ and $\mathrm{E}(X^2)$.

A random sample X_1, X_2, X_3 is taken and the statistic $Y = X_1^2 + X_2^2 + X_3^2$ is calculated.

b Show that Y is an unbiased estimator of α^2.

E/P **12** Jiaqi and Mei Mei each independently took a random sample of students at their school and asked them how much money, in RMB, they earned last week. Jiaqi used his sample of size 20 to obtain unbiased estimates of the mean and variance of the amount earned by a student at their college last week. He obtained values of $\bar{x} = 15.5$ and $s_x^2 = 8.0$

SKILLS ADAPTIVE LEARNING

Mei Mei's sample of size 30 can be summarised as $\sum y = 486$ and $\sum y^2 = 8222$

a Use Mei Mei's sample to find unbiased estimates of μ and σ^2. **(2 marks)**

b Combine the samples and use all 50 observations to obtain further unbiased estimates of μ and σ^2. **(4 marks)**

c Explain what is meant by standard error. **(1 mark)**

d Find the standard error of the mean for each of these estimates of μ. **(2 marks)**

e Comment on which estimate of μ you would prefer to use. **(1 mark)**

E/P **13** A factory worker checks a random sample of 20 bottles from a production line in order to estimate the mean volume of bottles (in cm^3) from this production run. The 20 values can be summarised as $\sum x = 1300$ and $\sum x^2 = 84\,685$.

a Use this sample to find unbiased estimates of μ and σ^2. **(2 marks)**

A factory manager knows from experience that the standard deviation of volumes on this process, σ, should be 3 cm^3 and he wishes to have an estimate of μ that has a standard error of less than 0.5 cm^3.

b Recommend a sample size for the manager, showing working to support your recommendation. **(2 marks)**

c Does your recommended sample size guarantee a standard error of less than 0.5 cm^3? Give a reason for your answer. **(1 mark)**

The manager takes a further sample of size 16 and finds $\sum x = 1060$.

d Combine the two samples to obtain a revised estimate of μ. **(2 marks)**

E **14** After growing for 10 weeks in a greenhouse, the heights of certain plants have a standard deviation of 2.6 cm. Find the smallest sample that must be taken for the standard error of the mean to be less than 0.5 cm. **(3 marks)**

E **15** The hardness of a new type of material was determined by measuring the depth of the hole made by a heavy pointed device.

The following observations in tenths of a millimetre were obtained:

4.7 5.2 5.4 4.8 4.5 4.9 4.5 5.1 5.0 4.8

a Estimate the mean depth of hole for this material. **(1 mark)**

b Find the standard error for your estimate. **(2 marks)**

c Estimate the size of sample required so that in future the standard error of the mean should be just less than 0.05 **(3 marks)**

P **16** To work for a company, applicants need to complete a medical test. The probability of each applicant passing the test is p, independent of any other applicant. The medicals are carried out over two days and on the first day n applicants are seen, and on the next day $2n$ are seen. Let X_1 represent the number of applicants who pass the test on the first day and let X_2 represent the number who pass on the second day.

SKILLS
REASONING/ARGUMENTATION

a Write down $\mathrm{E}(X_1)$, $\mathrm{E}(X_2)$, $\mathrm{Var}(X_1)$ and $\mathrm{Var}(X_2)$.

b Show that $\frac{X_1}{n}$ and $\frac{X_2}{2n}$ are both unbiased estimates of p and state, giving a reason, which you would prefer to use.

c Show that $X = \frac{1}{2}\left(\frac{X_1}{n} + \frac{X_2}{2n}\right)$ is an unbiased estimator of p.

d Show that $Y = \left(\frac{X_1 + X_2}{3n}\right)$ is an unbiased estimator of p.

e Which of the statistics $\frac{X_1}{n}$, $\frac{X_2}{2n}$, X or Y is the best estimator of p?

The statistic $T = \left(\frac{2X_1 + X_2}{3n}\right)$ is proposed as an estimator of p.

f Find the bias.

P **17** In a bag that contains a large number of counters, the number 0 is written on 40% of the counters, the number 1 is written on 20% of the counters, and the number 2 is written on the remaining 40% of the counters.

SKILLS
REASONING/ARGUMENTATION

a Find the mean μ and the variance σ^2 for this population of counters.

A random sample of size 3 is taken from the bag.

b List all the possible observations from this sample.

c Find the sampling distribution for the mean $\overline{X}$.

d Find $\mathrm{E}(\overline{X})$ and $\mathrm{Var}(\overline{X})$.

e Find the sampling distribution for the **median** N.

f Hence, find $\mathrm{E}(N)$ and $\mathrm{Var}(N)$.

g Show that N is an unbiased estimator of μ.

h Explain which estimator, $\overline{X}$ or N, you would choose as an estimator of μ.

SKILLS ANALYSIS

Challenge

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

b Hence, or otherwise, show that s^2 is an unbiased estimate for the population variance σ^2.

3.2 Confidence intervals

The value of $\hat{\theta}$, which is an estimator of θ, is found from a sample. It is used as an unbiased estimate for the population parameter θ and is very unlikely to be exactly equal to θ.

There is no way of establishing, from the sample data only, how close the estimate is.

Instead, you can form a **confidence interval** for θ.

- **A confidence interval (C.I.) for a population parameter θ is a range of values defined so that there is a specific probability that the true value of the parameter lies within that range.**

For example, you could establish a 90% confidence interval, or a 95% confidence interval.

A 95% confidence interval is an interval such that there is a 0.95 probability that the interval contains θ.

Different samples will generate different confidence intervals since estimates for the parameter will change based on the data in the sample and the sample size.

Watch out The population parameter θ is fixed, so you cannot talk about its value in probabilistic terms.

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

Hence, if you know the population standard deviation, you can establish a confidence interval for the population mean μ using the standardised normal distribution.

Example 8

Given that X is normally distributed, show that a 95% confidence interval for μ, based on a sample of size n, is given by:

$$\left(\bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right)$$

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

and therefore

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathrm{N}(0, 1^2)$$

Using tables, you can see that for the N(0, 1^2) distribution:

$$\mathrm{P}(Z > 1.9600) = \mathrm{P}(Z < -1.9600) = 0.025$$

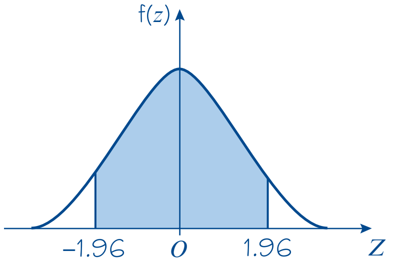
and so 95% of the distribution is between −1.9600 and 1.9600

Problem-solving

You will need to use the **standardised** normal distribution N(0, 1^2) to tackle problems like this.

If $X \sim \mathrm{N}(\mu, \sigma^2)$ then $Z = \frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1^2)$

← Statistics 1 Section 7.4



So $\quad \text{P}(-1.96 < Z < 1.96) = 0.95$

$$\Rightarrow \quad \text{P}\left(-1.96 < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96\right) = 0.95$$

Look at the inequality inside the probability statement:

$$-1.96 \times \frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < 1.96 \times \frac{\sigma}{\sqrt{n}}$$

Start to isolate μ.

$$\overline{X} + 1.96 \times \frac{\sigma}{\sqrt{n}} > \mu > \overline{X} - 1.96 \times \frac{\sigma}{\sqrt{n}}$$

Multiply by −1 and change the inequalities.

$$\overline{X} - 1.96 \times \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}$$

So the 95% confidence interval for μ is

$$\left(\overline{x} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right)$$

Notation The upper and lower values of a confidence interval are sometimes called the **confidence limits**.

- **The 95% confidence interval for μ is $\left(\overline{x} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right)$**

The value of 1.96 in the formula above is determined by the percentage points of the standardised normal distribution. By changing this value you can formulate confidence intervals with different levels of confidence.

For example, a 99% confidence interval would have 1.96 replaced by 2.5758, since that is the value of z such that $\text{P}(-z < Z < z) = 0.99$

Given one confidence interval, it is possible to calculate another.

Hint The choice of what confidence interval to use in a particular situation will depend on the problem involved but a value of 95% is commonly used if no other value is specified.

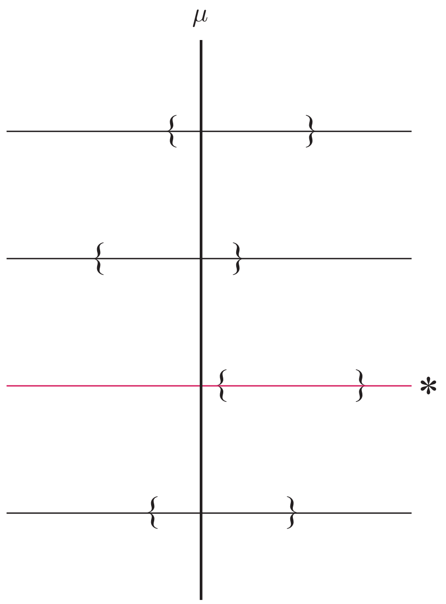
Interpreting confidence intervals

- First, it is important to remember that **μ is a fixed, but unknown**, number. Because of this, it cannot vary and does not have a distribution.
- Second, it is useful to remember that you base a 95% confidence interval on a probability statement about the normal distribution $Z \sim \text{N}(0, 1^2)$.
- Although you start by considering probabilities from the random variable Z, the final confidence interval does not tell you the probability that μ lies inside a fixed interval. Since μ is fixed, it is the confidence interval that varies (according to the value of $\overline{x}$).
- A 95% confidence interval tells you that the probability that the interval contains μ is 0.95

The diagram opposite shows the 95% confidence intervals calculated from different samples and also shows the position of μ. Suppose 20 samples of size 100 were taken, and 95% confidence intervals for μ were calculated for each sample. This would give 20 different confidence intervals, each based on one of the 20 different values of x.

If you assume that you know what the value of μ is, then you can plot each of these confidence intervals on a diagram similar to the one here; you would expect that 95% of these confidence intervals would contain the value μ, but about once in every 20 times you would get an interval which did not contain μ (marked * on the diagram).

The problem is that you never know whether the confidence interval calculated contains μ or not. However, 95% of the time (or 90%, or 99%, depending on the degree of confidence required), the interval will contain μ.



Example 9 SKILLS ADAPTIVE LEARNING

A 90% confidence interval is given by (32.1, 42.7)

Calculate a 95% confidence interval.

$\bar{x} + 1.6449 \times \frac{\sigma}{\sqrt{n}} = 42.7$

Find $\bar{x}$ and $\frac{\sigma}{\sqrt{n}}$ by solving simultaneously.

$\bar{x} - 1.6449 \times \frac{\sigma}{\sqrt{n}} = 32.1$

$2\bar{x} = 74.8$

$\bar{x} = 37.4$

$2 \times 1.6449 \times \frac{\sigma}{\sqrt{n}} = 10.6$

$\frac{\sigma}{\sqrt{n}} = \frac{10.6}{(2 \times 1.6499)} = 3.222\ldots$

A 95% confidence interval is:

$\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$

The lower limit:

$\bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}} = 37.4 - 1.96 \times 3.222\ldots = 31.08$

The upper limit:

$\bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}} = 37.4 + 1.96 \times 3.222\ldots = 43.72$

A 95% confidence interval is

(31.08, 43.72)

Example 10 SKILLS REASONING/ARGUMENTATION

The breaking strains of string produced at a certain factory are normally distributed with standard deviation 1.5 kg. A sample of 100 lengths of string from a certain batch was tested and the mean breaking strain was 5.30 kg.

a Find a 95% confidence interval for the mean breaking strain of string in this batch.

The manufacturer becomes concerned if the lower 95% confidence limit falls below 5 kg.
A sample of 80 lengths of string from another batch gave a mean breaking strain of 5.31 kg.

b Will the manufacturer be concerned?

a 95% confidence limits are:

$$\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}} = 5.30 \pm 1.96 \times \frac{1.5}{\sqrt{100}}$$

So a 95% confidence interval is (5.006, 5.594)

b Lower 95% confidence limit is:

$$\bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}} = 5.31 - \frac{1.96 \times 1.5}{\sqrt{80}}$$

$$= 4.98$$

so the manufacturer will be concerned.

Use the $\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$ formula.

Notation In your exam you can define a confidence interval:

- by giving the confidence limits, e.g. 5.30 ± 0.294
- using interval notation, e.g. (5.006, 5.594) or [5.006, 5.594]
- using inequalities, e.g. $5.006 < \mu < 5.594$

- **The width of a confidence interval is the difference between the upper confidence limit and the lower confidence limit. This is $2 \times z \times \frac{\sigma}{\sqrt{n}}$, where z is the relevant percentage point from the standardised normal distribution, for example 1.96, 1.6449, etc.**

The greater the width, the less information you have about the population mean. There are three factors that affect the width: the value of σ, the size of the sample n and the degree of confidence required. In a particular example where σ and n are determined, the only factor you can change to alter the width is the level of confidence. A high level of confidence (e.g. 99%) will give a greater width than a lower level of confidence (e.g. 90%), and the statistician has to weigh up the advantages of high confidence against greater width when calculating a confidence interval.

Example 11 SKILLS PROBLEM-SOLVING

A random sample of size 25 is taken from a normal distribution with standard deviation 2.5.
The mean of the sample is 17.8.

a Find a 99% C.I. for the population mean μ.

b What size sample is required to obtain a 99% C.I. with a width of at most 1.5?

c What confidence level would be associated with the interval based on the above sample of 25 but of width 1.5, i.e. (17.05, 18.55)?

a 99% confidence limits are:

$\bar{x} \pm 2.5758 \times \frac{\sigma}{\sqrt{n}} = 17.8 \pm 2.5758 \times \frac{2.5}{\sqrt{25}}$

So a 99% confidence interval is (16.51, 19.09)

Use the table on page 136 to find 2.5758

b Width of 99% C.I. is $2 \times 2.5758 \times \frac{2.5}{\sqrt{n}}$

so you require $1.5 > \frac{12.879\ldots}{\sqrt{n}}$

i.e. $n > 73.719\ldots$

so you need $n = 74$

Use the $2 \times z \times \frac{\sigma}{\sqrt{n}}$ formula or the definition for the width.

c A width of 1.5 ⇒ $1.5 = 2 \times z \times \frac{2.5}{\sqrt{25}}$

$z = 1.5$

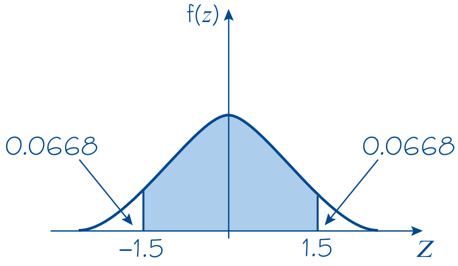
From the table on page 135 you find that

$\mathrm{P}(Z < 1.5) = 0.9332$

and so $\mathrm{P}(Z > 1.5) = \mathrm{P}(Z < -1.5)$

$= 1 - 0.9332$

$= 0.0668$



So the confidence level is

$100 \times (1 - 2 \times 0.0668) = 86.6\%$

The percentage of the confidence interval is given by the area between $z = \pm 1.5$

Exercise 3B

SKILLS PROBLEM-SOLVING

1 A random sample of size 9 is taken from a normal distribution with variance 36. The sample mean is 128.

a Find a 95% confidence interval for the mean μ of the distribution.

b Find a 99% confidence interval for the mean μ of the distribution.

2 A random sample of size 25 is taken from a normal distribution with standard deviation 4. The sample mean is 85.

a Find a 90% confidence interval for the mean μ of the distribution.

b Find a 95% confidence interval for the mean μ of the distribution.

3 A 95% confidence interval is given by (25.61, 27.19)
Calculate a 99% confidence interval.

4 A normal distribution has standard deviation 15. Estimate the sample size required if the following confidence intervals for the mean should have width of less than 2.

a 90% **b** 95% **c** 99%

E/P 5 A railway company is studying the number of seconds that express trains are late to arrive. Previous surveys have shown that the times are normally distributed and that the standard deviation is 50. A random sample of 200 trains was selected and gave rise to a mean of 310 seconds late.

a Find a 90% confidence interval for μ, the mean number of seconds that express trains are late. **(3 marks)**

Five different independent random samples of 200 trains are selected, and each sample is used to generate a different 90% confidence interval for μ.

SKILLS CRITICAL THINKING

b Find the probability that exactly three of these confidence intervals contain μ. **(2 marks)**

Hint Use a suitable **binomial distribution**.

E/P 6 Amy is investigating the total distance travelled by vans in current use. The standard deviation can be assumed to be 15 000 km. In a random sample, 80 vans were stopped and their mean distance travelled was found to be 75 872 km.

Amy suspects that the population is normally distributed, but claims that she can still use the normal distribution to find a confidence interval for μ. Find a 90% confidence interval for the mean distance travelled by vans in current use. **(3 marks)**

E/P 7 It is known that each year the standard deviation of the marks in a certain examination is 13.5 but the mean mark μ will fluctuate. An examiner wants to estimate the mean mark of all the candidates on the examination but he only has the marks of a sample of 250 candidates, which gives a sample mean of 68.4

a What assumption about the candidates must the examiner make in order to use this sample mean to calculate a confidence interval for μ? **(1 mark)**

b Assuming that the above assumption is justified, calculate a 95% confidence interval for μ. **(3 marks)**

Later, the examiner discovers that the actual value of μ was 65.3

c What conclusions might the examiner draw about his sample? **(2 marks)**

E/P 8 A student calculated 95% and 99% confidence intervals for the mean μ of a certain population but failed to label them. The two intervals were (22.7, 27.3) and (23.2, 26.8).

a State, with a reason, which interval is the 95% one. **(1 mark)**

b Estimate the standard error of the mean in this case. **(2 marks)**

c What was the student's unbiased estimate of the mean μ in this case? **(2 marks)**

E 9 The director of a company has asked for a survey to estimate the mean expenditure of customers on electrical appliances. In a random sample, 100 people were questioned and the research team presented the director with a 95% confidence interval of ($128.14, $141.86).

The director says that this interval is too wide and wants a confidence interval of total width $10.

a Using the same value of $\bar{x}$, find the confidence limits in this case. **(3 marks)**

b Find the level of confidence for the interval in part **a**. **(2 marks)**

The managing director is still not happy and now wishes to know how large a sample would be required to obtain a 95% confidence interval of total width no greater than $10.

c Find the smallest size of sample that will **satisfy** this request. **(3 marks)**

E **10** A factory produces steel sheets whose masses are known to be normally distributed with a standard deviation of 2.4 kg. A random sample of 36 sheets had a mean mass of 31.4 kg. Find 99% confidence limits for the population mean. **(3 marks)**

E **11** A machine is set up to pour liquid into cartons in such a way that the amount of liquid poured on each occasion is normally distributed with a standard deviation of 20 ml.

Find 99% confidence limits for the mean amount of liquid poured if a random sample of 40 cartons had an average content of 266 ml. **(3 marks)**

E/P **12** **a** The error made when a certain instrument is used to measure the body length of a butterfly of a particular species is known to be normally distributed with mean 0 and standard deviation 1 mm. Calculate, to 3 decimal places, the probability that the size of the error made when the instrument is used once is less than 0.4 mm. **(2 marks)**

b Given that the body length of a butterfly is measured 9 times with the instrument, calculate, to 3 decimal places, the probability that the mean of the 9 readings will be within 0.5 mm of the true length. **(3 marks)**

c Given that the mean of the 9 readings was 22.53 mm, determine a 98% confidence interval for the true body length of the butterfly. **(3 marks)**

Chapter review 3

E **1** The masses of bags of lentils, X kg, have a normal distribution with unknown mean μ kg and a known standard deviation σ kg. A random sample of 80 bags of lentils gave a 90% confidence interval for μ of (0.4533, 0.5227).

a Without carrying out any further calculations, use this confidence interval to test whether $\mu = 0.48$. State your hypotheses clearly and write down the **significance level** you have used. **(3 marks)**

A second random sample of 120 of these bags of lentils had a mean mass of 0.482 kg.

b Calculate a 95% confidence interval for μ based on this second sample. **(6 marks)**

E/P **2** The lengths of the tails of mice in a pet shop are assumed to have unknown mean μ and unknown standard deviation σ.

A random sample of 20 mice is taken and the length of their tails recorded.

The sample is represented by $X_1, X_2, \ldots, X_{20}$.

a State whether or not the following are statistics.

Give reasons for your answers.

i $\dfrac{2X_1 + X_{20}}{3}$ **ii** $\sum_1^{20}(X_i - \mu)^2$ **iii** $\dfrac{\sum_1^{20} X_i^2}{n}$ **(4 marks)**

b Find the mean and variance of $\dfrac{4X_1 - X_{20}}{3}$ **(3 marks)**

E **3** The breaking stresses of elastic bands are normally distributed.

A company uses bands with a mean breaking stress of 46.50 N.

A new supplier claims that they can supply bands that are stronger, and provides a sample of 100 bands for the company to test. The company checked the breaking stress X for each of these 100 bands and the results are summarised as follows:

$$n = 100 \qquad \sum x = 4715 \qquad \sum x^2 = 222\,910$$

a Find an **approximate** 95% confidence interval for the mean breaking stress of these new rubber bands. **(3 marks)**

b Do you agree with the new supplier, that they can supply bands that are stronger? **(2 marks)**

E/P **4** On each of 100 days, a scientist took a sample of 1 litre of water from a particular place along a river, and measured the amount, X mg, of chlorine in the sample. The results she obtained are shown in the table.

X	1	2	3	4	5	6	7	8	9
Number of days	4	8	20	22	16	13	10	6	1

a Estimate the mean amount of chlorine present per litre of water, and estimate, to 3 decimal places, the standard error of this estimate. **(3 marks)**

b Obtain approximate 98% confidence limits for the mean amount of chlorine present per litre of water. **(3 marks)**

Given that measurements at the same point under the same conditions are taken for a further 100 days,

c estimate, to 3 decimal places, the probability that the mean of these measurements will be greater than 4.6 mg per litre of water. **(3 marks)**

E **5** The amount, to the nearest mg, of a certain chemical in particles in the air at a weather station was measured each day for 300 days. The results are shown in the table.

Amount of chemical (mg)	12	13	14	15	16
Number of days	5	42	210	31	12

Estimate the mean amount of this chemical in the air, and find, to 2 decimal places, the standard error of this estimate. **(3 marks)**

E/P **6** Occasionally, a firm manufacturing furniture needs to check the mean distance between pairs of holes drilled by a machine in pieces of wood to ensure that no change has occurred. It is known from experience that the standard deviation of the distance is 0.43 mm. The firm intends to take a random sample of size n, and to calculate a 99% confidence interval for the mean of the population. The width of this interval must be no more than 0.60 mm.

Calculate the minimum value of n. **(4 marks)**

E **7** The times taken by five-year-old children to complete a certain task are normally distributed with a standard deviation of 8.0 s. In a random sample, 25 five-year-old children from school A were given this task and their mean time was 44.2 s.

a Find 95% confidence limits for the mean time taken by five-year-old children from school A to complete this task. **(3 marks)**

The mean time for a random sample of 20 five-year-old children from school B was 40.9 s. The headteacher of school B concluded that the overall mean for school B must be less than that of school A. Given that the two samples were independent,

b test the headteacher's conclusion using a 5% significance level. State your hypotheses clearly. **(6 marks)**

E/P SKILLS ADAPTIVE LEARNING

8 The random variable X is normally distributed with mean μ and variance σ^2.

a Write down the distribution of the sample mean $\overline{X}$ of a random sample of size n. **(1 mark)**

b State, with a reason, whether this distribution is exact or is an estimate. **(1 mark)**

An efficiency expert wishes to determine the mean time taken to drill a fixed number of holes in a metal sheet.

c Determine how large a random sample is needed so that the expert can be 95% certain that the sample mean time will differ from the true mean time by less than 15 seconds. Assume that it is known from previous studies that $\sigma = 40$ seconds. **(4 marks)**

E/P

9 A man regularly uses a train service which should arrive in Zurich at 09:31. He decided to test this stated arrival time. Each weekday for a period of 4 weeks, he recorded the number of minutes X that the train was late on arrival in Zurich. If the train arrived early then the value of X was negative. His results are summarised as follows:

$$n = 20 \qquad \sum x = 15.0 \qquad \sum x^2 = 103.21$$

a Calculate unbiased estimates of the mean and variance of the number of minutes late of this train service. **(3 marks)**

The random variable X represents the number of minutes that the train is late on arriving in Zurich. Records kept by the railway company show that over fairly short periods, the standard deviation of X is 2.5 minutes. The man made two assumptions about the distribution of X and the values obtained in the sample and went on to calculate a 95% confidence interval for the mean arrival time of this train service.

b State the two assumptions. **(2 marks)**

c Find the confidence interval. **(3 marks)**

d Given that the assumptions are reasonable, comment on the stated arrival time of the service. **(1 mark)**

E/P

10 The random variable X is normally distributed with mean μ and variance σ^2.

a Write down the distribution of the sample mean $\overline{X}$ of a random sample of size n. **(1 mark)**

b Explain what you understand by a 95% confidence interval. **(2 marks)**

A garage sells both leaded and unleaded fuel. The distribution of the values of sales for each type is normal. During 2010, the standard deviation of individual sales of each type of fuel was £3.25. The mean of the individual sales of leaded fuel during this time was £8.72. A random sample of 100 individual sales of unleaded fuel gave a mean of £9.71.

Calculate:

c an interval within which 90% of the sales of leaded fuel will lie **(3 marks)**

d a 95% confidence interval for the mean sales of unleaded fuel. **(3 marks)**

The mean of the sales of unleaded fuel for 2009 was £9.10.

e Using a 5% significance level, investigate whether there is sufficient evidence to conclude that the mean of all the 2010 unleaded sales was greater than the mean of the 2009 sales. **(6 marks)**

f Find the size of the sample that should be taken so that the garage owner can be 95% certain that the sample mean of sales of unleaded fuel during 2010 will differ from the true mean by less than £0.50. **(4 marks)**

E/P **11** **a** Explain what is meant by a 98% confidence interval for a population mean. **(2 marks)**

The lengths, in cm, of the leaves of oak trees are known to be normally distributed with variance 1.33 cm^2.

A sample of 40 oak tree leaves is found to have a mean of 10.20 cm.

b Estimate, giving your answer to 3 decimal places, the standard error of the mean. **(2 marks)**

c Use this value to estimate 95% confidence limits for the mean length of the population of oak tree leaves, giving your answer to 2 decimal places. **(3 marks)**

d Find the minimum size of the sample of leaves which must be taken if the width of the 98% confidence interval for the population mean is at most 1.50 cm. **(4 marks)**

E/P **12** **a** Write down the mean and the variance of the distribution of the means of all possible samples of size n taken from an infinite population having mean μ and variance σ^2. **(2 marks)**

b Describe the form of this distribution of sample means when:

i n is large

ii the distribution of the population is normal. **(2 marks)**

The standard deviation of all the till receipts of a supermarket during 2014 was £4.25.

c Given that the mean of a random sample of 100 of the till receipts is £18.50, obtain an approximate 95% confidence interval for the mean of all the till receipts during 2014. **(3 marks)**

d Find the size of sample that should be taken so that the management can be 95% confident that the sample mean will not differ from the true mean by more than £0.50. **(3 marks)**

e The mean of all the till receipts of the supermarket during 2013 was £19.40. Using a 5% significance level, investigate whether the sample in part **a** provides sufficient evidence to conclude that the mean of all the 2014 till receipts is different from that in 2013. **(6 marks)**

E/P **13** Records of the diameters of spherical metal balls produced on a certain machine show that the diameters are normally distributed with mean 0.824 cm and standard deviation 0.046 cm. Two hundred samples are randomly chosen, each consisting of 100 metal balls.

a Calculate the expected number of the 200 samples having a mean diameter less than 0.823 cm. **(2 marks)**

On a certain day, it was believed that the machine was faulty. It may be assumed that if the machine is faulty, it will change the mean of the diameters without changing their standard deviation. On that day, a random sample of 100 metal balls had mean diameter 0.834 cm.

b Determine a 98% confidence interval for the mean diameter of the metal balls being produced that day. **(3 marks)**

c Hence state whether or not you would conclude that the machine is faulty on that day given that the significance level is 2%. **(3 marks)**

E/P **14** A doctor claims that there is a higher mean heart rate in people who always drive to work compared to people who regularly walk to work. She measures the heart rates X of 30 people who always drive to work and 36 people who regularly walk to work. Her results are summarised in the table below.

	n	$\bar{x}$	s^2
Drive to work	30	52	60.2
Walk to work	36	47	55.8

a Test, at the 5% level of significance, the doctor's claim. State your hypotheses clearly. **(6 marks)**

b State any assumptions you have made in testing the doctor's claim. **(2 marks)**

The doctor decides to add another person who drives to work to her data.
She measures the person's heart rate and finds $X = 55$.

c Find an unbiased estimate of the variance for the sample of 31 people who drive to work. Give your answer to 3 significant figures. **(4 marks)**

Challenge

SKILLS ADAPTIVE LEARNING

Two independent random samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ are taken from a population with mean μ and variance σ^2. The unbiased estimators $\overline{X}$ and $\overline{Y}$ of μ are calculated. A new unbiased estimator T of μ is sought in the form $T = r\overline{X} + s\overline{Y}$.

a Show that, since T is unbiased, $r + s = 1$.

b By writing $T = r\overline{X} + (1 - r)\overline{Y}$, show that:

$$\text{Var}(T) = \left(\sigma^2\frac{r^2}{n} + \frac{1 - r^2}{m}\right)$$

c Show that the minimum variance of T is when $r = \dfrac{n}{n + m}$

d Find the best (in the sense of minimum variance) unbiased estimator of μ in the form $r\overline{X} + s\overline{Y}$.

Summary of key points

1 If X is a random variable, then a random sample of size n will consist of n observations of the random variable X, which are referred to as $X_1, X_2, X_3, \ldots, X_n$, where the X_i:

- are independent random variables
- each have the same distribution as X.

A statistic T is defined as a random variable consisting of any function of the X_i that involves no other quantities, such as unknown population parameters.

2 The **sampling distribution** of a statistic T is the probability distribution of T.

3 A statistic that is used to estimate a population parameter is called an **estimator** and the particular value of the estimator generated from the sample taken is called an **estimate**.

4 If a statistic T is used as an estimator for a population parameter θ and $\mathrm{E}(T) = \theta$, then T is an unbiased estimator for θ.

5 If a statistic T is used as an estimator for a population parameter θ, then the bias is $\mathrm{E}(T) - \theta$. For an unbiased estimator, the bias is 0.

6 An unbiased estimator for σ^2 is given by the **sample variance** S^2 where:

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

7 The standard deviation of an estimator is called the **standard error** of the estimator.

8 When using the sample mean $\overline{X}$, you can use the following result for the standard error:

Standard error of $\overline{X} = \frac{\sigma}{\sqrt{n}}$ or $\frac{s}{\sqrt{n}}$

9 A **confidence interval** for a population parameter θ is a range of values defined so that there is a specific probability that the true value of the parameter lies within that range.

10 A 95% confidence interval for the population mean μ is $\left(\overline{x} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right)$

11 The width of a confidence interval is the difference between the upper confidence limit and the lower confidence limit. This is $2 \times z \times \frac{\sigma}{\sqrt{n}}$, where z is the relevant percentage point from the standard normal distribution, for example 1.96, 1.6449, etc.