Pearson
Edexcel Level 3 Advanced GCE in Statistics (9ST0)

Two-year Scheme of Work

For first teaching from September 2017

Version 3.5 – updated November 2025

Contents

INTRODUCTION

Content of this document is based on the accredited version of Pearson Edexcel Level 3 Advanced GCE in Statistics (9ST0) and includes detailed scheme of work for the content of A Level Statistics.

This scheme of work is based upon a two-year delivery model for A Level Statistics.

The scheme of work is broken up into units and sub-units, so that there is greater flexibility for moving topics around to meet planning needs.

Each unit contains:

- Specification references
- Prior knowledge (from GCSE or other units in the scheme of work)
- Keywords
- Unit summary (the purpose of the unit, the rationale to its place in the scheme of work, links with other topics, opportunities for the use of technology)

Each sub-unit contains:

- Recommended teaching time, though of course this is adaptable according to individual teaching needs
- Objectives for students at the end of the sub-unit
- Teaching points
- Opportunities for embedding the Statistical Enquiry Cycle (SEC)
- Common and possible mistakes
- Other Notes (inc. off specification extension material)

Note: the estimated teaching hours are approximate and should be used as a guideline only.

Our free support for the A Level Statistics specifications can be found on the Pearson Edexcel A Level Statistics (2017) website (https://qualifications.pearson.com/en/qualifications/edexcel-a-levels/statistics-2017.html) and on the Maths Emporium (https://mathsemporium.com/).

AMENDMENTS

A summary of changes made between version 2 and version 3.4 of this scheme of work are listed below.

There is a detailed list of every change in the change log which is accompanied by a scheme of work with changes tracked in red.

Summary of changes made between version 2 and version 3	**Page number**
Exemplars structured with “Exemplar” titles	Throughout
All instances of “Accept H_0” replaced with “Do not reject H_0”	Throughout
All exemplars where multiple valid methods can be used amended to exemplify these methods	Throughout
Page numbering fixed	Throughout
General formatting issues	Throughout
Some phrasing/mistakes/typographical errors addressed	Throughout
“Common Mistakes” sections formatted as bullet points	Throughout
Teaching times for Units 0a, 0c, 0d, 1b, 2c changed	
Unit 0b – Has been removed. Unit 0c now labelled as Unit 0b, Unit 0d now labelled as Unit 0c	Throughout
Unit 0c – Note about new Unit 26 (Statistical project) added	Page 29
Unit 1a – More clarity over notation and use of calculators	Page 35
Unit 1c – Teaching points about anomalies and data cleaning added	Page 40
Unit 1d – Advantages and Disadvantages about numerical measures updated and new exemplar	Page 43, 46
Unit 2a – Definitions of discrete and continuous updated to match specification	Page 49
Unit 2b/2d – Notes about appropriateness of certain data visualisations updated to reflect modern practice	Pages 47-66
Unit 2b – Notes about trends/variation updated	Page 56
Unit 2c – “Technology and Software” entire sub-unit rewritten	Pages 60-63
Unit 2d – Notes about answering exam questions added	Page 66
Unit 4 – More clarity over what is assessed in the exams added	Pages 79-85
Unit 5b – New exemplar about inverse binomial questions added	Page 92-93
Unit 6c – More clarity over what students are expected to know about regression lines added	Pages 105-107
Unit 7a – More information regarding the continuous uniform distribution added	Pages 110-113

Unit 7a – New exemplar added	Page 111
Unit 7c – Reference and exemplar for finding both unknown parameters removed (in version 3.2)	Page 119-122
Unit 8b – Exemplars (about cluster and snowball sampling) updated	Pages 131-132
Unit 8c – Advantages and disadvantages of sampling methods updated	Page 135-136
Unit 9a – Definitions of parameter and statistic updated to match specification	Page 139
Unit 9c – More clarity over conditions for normal approximation	Page 149
Unit 9c – New exemplar using inverse binomial added	Page 152
Unit 10 – Note about ρ_s and hypotheses about Spearman's Rank tests added	Pages 155-168
Unit 10b – More clarity over the bivariate normal distribution	Page 162
Unit 11a – Clarity over *p*-values added	Pages 176
Unit 11b – New exemplar for proportion tests with normal approximation added	Pages 182-184
Unit 13c – Clarity over Wilcoxon Rank-Sum hypotheses added	Page 212
Unit 14b/c – "Quantified differences" added as a possible assessment, with methods exemplified	Pages 215-222
Unit 15b – New exemplar about Inverse Poisson added	Page 231
Unit 17 – Entire unit rewritten to include topics about Hypergeometric tree diagrams and conditional probability distributions (in addition to Bayes' Theorem)	Pages 246-256
Unit 18 – Clarity over how to interpret confidence intervals in relation to a claimed mean and the relationship between two confidence intervals	Page 260
Unit 18c – Properties of *t*-confidence intervals added	Page 274
Unit 18c – Common mistakes and notes updated	Page 275-276
Unit 19c – New exemplar about finding P(Type I error) for binomial proportion tests	Page 287
Unit 19c – Table of advantages and disadvantages for critical regions and p-values added	Page 288
Unit 19d – Ways to reduce Type I and Type II errors and consequences added	Page 290
Unit 19d – New exemplar	Page 291
Unit 19d – List of hypothesis tests that P(Type II Errors) and Power can be calculated for added	Page 296
Unit 20 – Diagrams of the exponential distribution added throughout	Pages 297-305

Unit 20a – Table of mean/variance/sd for Poisson and Exponential distributions added	Page 300
Unit 20b – Extra note about conditional probability added	Page 302
Unit 21a – New exemplar for P(Type II Error) added	Page 313-314
Unit 22 – Hypotheses updated throughout	Pages 327-342
Unit 22a – New exemplar using Continuous Uniform Distribution added	Page 334
Unit 22b – Table of how to estimate parameters for each distribution added	Page 338
Unit 24c – Exemplars updated	Page 369
Unit 25 – Clarity on how to interpret Cohen's *d* added	Page 375
Unit 26 – Statistical Project guidance added	Page 378
Appendix: Some Desmos links updated	Page 379

OPPORTUNITIES FOR PROMOTING EQUALITY AND DIVERSITY

This scheme of work is designed to promote equality and diversity as much as possible. Examples given are constructed in a way which celebrate, and increases awareness of, the protected characteristics as defined by The Equality Act 2010, or the use of statistics in other cultures. Teachers should embed equality and diversity into their own lesson plans and may use these examples as a guide.

TECHNOLOGY

In addition to the recommended calculator, activities in Desmos (http://www.desmos.com) are available for use as both teaching aids and student-led activities. They are listed in the appendix and referred to in the unit summaries and teaching points of sub-units.

LAYOUT

Example questions are provided throughout the scheme of work. Questions are in a **bold typeface** and the corresponding model answers are in *italic typeface*.

A Level Statistics	
Paper 1: Data and Probability 33⅓%, 2 hours, 80 marks	Specification topics 1-7,11-13, 18 (not 7.2) may be assessed in this paper
Paper 2: Statistical Inference 33⅓%, 2 hours, 80 marks	Specification topics 7.2, 8-10, 13-17, 19-21 may be assessed in this paper
Paper 3: Statistics in Practice 33⅓%, 2 hours, 80 marks	All topics in the specification may be assessed in this paper

Assessment Objectives		
Code	**Objective**	**Weighting**
AO1	Demonstrate knowledge and understanding, using appropriate terminology and notation, of standard statistical techniques used – • to collect and represent data • to calculate summary statistics and probabilities • in relation to hypotheses and inference.	**55%**
AO2	Interpret statistical information and results in context and reason statistically to make predictions, construct arguments, make decisions and draw conclusions	**25%**
AO3	Critically assess the reliability and validity of statistical methodologies and the conclusions drawn through the application of the statistical enquiry cycle.	**20%**

For further information regarding the assessment objectives, visit
https://www.gov.uk/government/publications/gce-subject-level-guidance-for-statistics

Two-year Scheme of Work overview

Year 1 content

Unit		Title	Estimated hours
0		**Foundations of Statistics**	
	a	Measures of Central Tendency (Mean, median, mode, range)	**2**
	b	Introduction to Probability (Set Theory Notation, Probability Diagrams)	**2**
	c	The Statistical Enquiry Cycle	**Minimum 1**
1		**Numerical Measures**	
	a	Quartiles, Percentiles and the Interquartile Range	**1**
	b	Variance and Standard Deviation	**2**
	c	Outliers and Scaling	**2**
	d	Suitability of Numerical Measures (Advantages and Disadvantages)	**1**
2		**Data Representation and Interpretation**	
	a	The Language of Statistics (Terminology)	**1**
	b	Reading and Interpreting Univariate Statistical Diagrams	**2**
	c	The Use of Software (Spreadsheets and Databases)	**2**
	d	Misrepresentation and the Suitability of Statistical Diagrams	**1**
3		**Probability Theory**	
	a	Concepts in Probability (Terminology and the Addition Rule)	**1**
	b	Conditional Probability (including the Multiplication Rule)	**2**
	c	Mutually Exclusive and Independent Events	**1**
4		**Discrete Random Variables**	
	a	Introduction to Random Variables (Terminology, Notation, Tabulated Probabilities, Uniform Distribution)	**1**
	b	Expectation and Variance	**2**
5		**The Binomial Distribution**	
	a	Conditions for a Binomial Distribution	**1**
	b	Finding probabilities from a Binomial Distribution	**1**
	c	Mean and Variance	**1**
			27 hours (minimum)
6		**Bivariate Data**	
	a	Terminology	**1**

Unit		Title	Estimated hours
	b	Correlation Coefficients	**2**
	c	The Least Squares Regression Line	**1**
7		**The Normal Distribution**	
	a	Continuous Random Variables (Rectangular distribution, Normal Distribution, Properties of the Normal Distribution, Terminology)	**1**
	b	The Normal Distribution (Finding probabilities, Finding *z*-values using the Inverse Normal Function)	**3**
	c	Finding Unknown Parameters	**2**
8		**Data Collection**	
	a	Terminology and Random Sampling (inc. the use of random numbers)	**1**
	b	Sampling Methods	**1**
	c	Suitability of Sampling Methods (Advantages and Disadvantages)	**2**
9		**Estimation and Approximation**	
	a	Terminology	**1**
	b	The Sampling Distribution of the Mean of a Normal Distribution (inc. applying in context)	**2**
	c	The Normal Approximation to the Binomial (Conditions, Continuity Corrections, finding approximate probabilities and applying in context)	**2**
10		**Introduction to Hypothesis Testing**	
	a	Terminology	**1**
	b	Hypothesis testing about the population PMCC (including suitability)	**2**
	c	Hypothesis testing about the Spearman's Rank Correlation Coefficient (including suitability)	**2**
11		**Methods of Hypothesis Testing**	
	a	Hypothesis tests about a sample mean from a Normal Distribution with known variance	**2**
	b	Hypothesis tests about a proportion	**2**
12		**Contingency Tables**	
	a	Introduction to Contingency Tables	**1**
	b	Hypothesis tests for association between two variables	**2**
	c	Context	**1**
			32 hours
13		**Non-Parametric Hypothesis Tests**	
	a	One-sample Sign Test	**1**

Unit		Title	Estimated hours
	b	One-sample Wilcoxon Signed-Rank Test	**1**
	c	Wilcoxon Rank-Sum Test	**1**
14		**Experimental Design**	
	a	Terminology	**1**
	b	Paired Sign Test	**1**
	c	Paired Wilcoxon Signed-Rank Test	**1**
15		**The Poisson Distribution**	
	a	Conditions for a Poisson Distribution	**1**
	b	Finding probabilities from a Poisson Distribution	**1**
	c	Mean and Variance	**1**
16		**Combinations of Independent Random Variables**	
	a	Expectation and Variance of a linear combination of independent random variables	**1**
	b	The sum of two Poisson variables	**1**
	c	Linear combinations of Normal variables	**1**
			12 hours

Year 2 content

Unit		Title	Estimated hours
17		**Further Probability Theory:**	
	a	Bayes' Theorem	**1**
	b	Hypergeometric Tree Diagrams	**1**
	c	Conditional Probability Distributions	**1**
18		**Confidence Intervals and the Central Limit Theorem**	
	a	Confidence Intervals for the mean of a Normal Distribution with known variance	**2**
	b	The Central Limit Theorem	**1**
	c	Confidence Intervals for the mean of a Normal Distribution with unknown variance (the *t*-distribution)	**2**
19		**Concepts in Hypothesis Testing**	
	a	Hypothesis tests for a mean using a large sample	**1**
	b	Hypothesis tests for a sample mean of a normal distribution with unknown variance	**2**
	c	Significance levels, critical regions and *p*-values	**1**
	d	The Power of a Hypothesis Test	**1**
20		**The Exponential Distribution**	
	a	Conditions for the Exponential Distribution	**1**
	b	Finding probabilities from an Exponential Distribution	**1**
21		**Hypothesis Tests between Two Parameters**	
	a	Hypothesis tests for the difference between two population means with known variances	**2**
	b	Hypothesis tests for the difference between two population means with unknown, but equal, variances	**2**
	c	Hypothesis tests for the difference between two proportions	**2**
22		**Goodness of Fit**	
	a	Calculating Expected Frequencies for a Goodness of Fit test (Binomial, Poisson, Normal, Exponential, Rectangular and specified discrete distributions)	**2**
	b	Hypothesis tests for the Goodness of Fit	**4**
23		**Further Experimental Design**	
	a	Terminology	**1**
	b	Hypothesis tests for the difference between two population means, using paired samples (paired-sample *t*-test)	**2**
			30 hours

Unit		Title	Estimated hours
24		**Analysis of Variance (ANOVA)**	
	a	Conditions for Analysis of Variance	**1**
	b	One-factor Analysis of Variance	**3**
	c	Two-factor Analysis of Variance	**2**
25		**Effect Size: Cohen's d**	**2**
26		**The Statistical Project**	**10**
			18 hours

Statistical Enquiry Cycle (SEC)

The Statistical Enquiry Cycle (SEC) underpins the study of Statistics. Students need to be able to apply the knowledge and techniques outlined in this section within the framework of the SEC. The cycle covers five stages:

- initial planning
- data collection
- data processing and presentation
- interpretation of results
- evaluation and review.

The detail of the SEC is provided below. During their learning students should develop their understanding of the SEC through a variety of authentic contexts. Practical experience of the cycle is integral to their understanding of the principles of the SEC. SEC identified questions on the exam papers will be designed to cover a minimum of 3 elements of SEC as detailed below.

A Initial planning

Subject content	
Students must understand the importance of initial planning when designing a line of enquiry or investigation including:	
1	identifying factors that may be related to the problem under investigation
2	defining a question or hypothesis (or hypotheses) to investigate
3	deciding what data to collect, and how to collect and record it, giving reasons
4	engaging in exploratory data analysis in order to investigate the situation
5	developing a strategy for how to process and represent the data giving reasons
6	justifying the proposed plan with regards ensuring a lack of bias.

B Data collection

Subject content	
Students must recognise the constraints involved in sourcing data including:	
1	when designing unbiased collection methods for primary sample data
2	when researching sources of secondary data, including from reference publications, the internet and the media
3	the importance of declaring the data collection methodology, including appreciating the importance of acknowledging sources
4	appreciating the inherent bias that may be incorporated through the use of leading questions either by accident or through agenda-driven design.

C Data processing and presentation

Subject content	
Students must understand a range of techniques in order to process, represent and discuss data including:	
1	organising and processing data, including an understanding of how technology can be used
2	make inferences about the population using appropriately chosen diagrams and summary measures to represent data including an understanding of outputs generated by appropriate technology
3	appreciating how to avoid misrepresentation of data.

D Interpretation of results

Subject content	
Students must appreciate the need to consider the context of the problem when interpreting results:	
1	analysing/interpreting diagrams and calculations/measures
2	drawing together conclusions that relate to the questions and hypotheses addressed
3	using appropriate tests to determine the statistical significance of the findings
4	discussing the reliability of findings.
5	Students must show an understanding of the importance of the clear and concise communication of findings and key ideas, and awareness of target audience.

E Evaluation and review

Subject content	
Students must be able to understand the importance of evaluating statistical work including:	
1	identifying weaknesses in approaches used to collect or display data
2	recognising the limitations of findings by considering sample size and sampling technique
3	suggesting improvements to statistical processes or presentation
4	refining processes to elicit further clarification of the initial hypothesis.

UNIT 0: Foundations of Statistics

Return to overview

SPECIFICATION REFERENCES

1.4 Compare different data sets, using appropriate diagrams or calculated measures of central tendency and spread: mean, median, mode, range, interquartile range, percentiles, variance and standard deviation.

1.5 Calculate measures using calculators and manual calculation as appropriate.

2.1 Know and use language and symbols associated with set theory in the context of probability.

2.2 Represent and interpret probabilities using tree diagrams, Venn diagrams and two-way tables.

PRIOR KNOWLEDGE

It is desirable that students have a familiarity with the following:

GCSE (9-1) in Mathematics at Higher Tier

A2 Substitute numerical values into formulae and expressions.

A19 Solve linear simultaneous equations.

P3 Relate relative expected frequencies to theoretical probability, using appropriate language and the 0-1 probability scale.

P4 Apply the property that the probabilities of an exhaustive set of outcomes sum to one

P6 Use tables, grids, Venn diagrams and tree diagrams.

S2 Interpret frequency tables.

S4 Median, mean, mode, range; interpretation, analysis and comparison of data sets through the aforementioned.

S5 Apply statistics to describe a population.

KEYWORDS

average, bias, central tendency, chance, data, equation, estimate, formula, frequency, frequency table, grouped frequency table, leading questions, mean, measure, median, mode, probability, probability tree, random, range, simultaneous equations, spread, substitute, two-way table, Venn diagram

UNIT SUMMARY

This unit is to “bridge the gap” between GCSE Mathematics and A Level Statistics. The only topic in this unit that is not covered on the GCSE specification is the Statistical Enquiry Cycle but is included at this early stage in order to prepare students to be aware of this throughout all units.

The rest of the A Level is built on these concepts and so it is advisable it is taught as early as possible. There is a degree of flexibility in the teaching time, depending on the ability and prior knowledge of the students. It is recommended that, at the very least, one hour is spent on the SEC and its importance in the world of statistics.

Spreadsheets can be utilised here to construct frequency tables and two-way tables. Students who have access to this software may check their answers with inbuilt spreadsheet functions, calculators and by hand. Most importantly, students should all be familiar with how to use the statistics mode on their calculators, including inputting data both from an unsorted list and from a frequency table, and finding the appropriate numerical measures. The equation solver will be useful in Unit 7c.

The Office for National Statistics (www.ons.gov.uk) has freely available datasets which can be used, adapted, or form the basis of a scenario requiring a student to calculate and analyse appropriate numerical measures in context.

World Bank **http://data.worldbank.org**, OECD **https://data.oecd.org/**

New Zealand Census at School **http://new.censusatschool.org.nz**

It will also help when writing questions which promote equality and diversity.

0a. Foundations of Statistics: Measures of Central Tendency (1.4) (1.5)	Teaching time 2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Calculate the mean from an unsorted list of data or a frequency table, or the estimated mean from a grouped frequency table, using the formula or a calculator.
- Calculate the mode, median and range from an unsorted list of data or a frequency table (not a histogram) using a formula or a calculator.

TEACHING POINTS

Recap the skills taught at GCSE Higher Tier (9-1).

The concepts of mean (add up the values, divide by the total frequency), median (the middle number) and mode (the most common) could be revised from GCSE.

The notation and formula of the sample mean, if not already seen, could be covered $\left(\bar{x} = \frac{\Sigma x}{n}\right)$. Using the statistics mode on the calculators will allow them to check their answer for the mean.

From a discrete sample of size n in ascending order, the median is the $\left(\frac{n+1}{2}\right)\text{th}$ indexed value if n is odd, or the mean of the $\left(\frac{n}{2}\right)\text{th}$ and $\left(\frac{n}{2}+1\right)\text{th}$ indexed value if n is even.

When given a frequency table, the mean $\left(\bar{x} = \frac{\Sigma fx}{\Sigma f}\right)$ could be calculated both by hand and checked on the calculator. The median and mode can be calculated as before. For a grouped frequency table, an estimate of the mean can be found as above, but using x as the midpoints of class intervals. Again, the estimated mean could be found both by hand and using the calculator.

Although the manual calculation is important in learning the work, emphasise that the use of the calculator is the expected method in the exam.

The range of a set of data (the difference between the largest value and smallest value) is revision from GCSE, for both an unsorted list of data or from a frequency table.

Plenty of practice in a variety of contexts would be beneficial to students and allow them to check their answers using the formulae, the calculator and any available software.

Students need to be able to formulate conclusions and interpretations through the comparison of numerical measures of two data sets.

Exemplar

The median UK household disposable income was £26,300 in the financial year ending 2016 and £25,700 in the financial year ending 2015. It is claimed that most households had a higher disposable income in 2016 than in 2015. Comment on this claim, justifying your answer.

Source: Office for National Statistics – Household disposable income and inequality in the UK: financial year ending 2016

The claim may be supported. Since the numerical measures considered are medians, this means that half of the UK households had a disposable income of at least £26,300 in the financial year ending 2016, and half of the UK households had a disposable income of at most £25,700 in the financial year ending 2015. However, the data does not tell us about the change of individual households - the upper and lower halves may have swapped between the years, and the median may still increase.

Lots of practice at formulating grammatically correct sentences would be beneficial. Care should be given to the wording of answer, since a colloquial phrasing may result in an incorrect statement.

OPPORTUNITIES FOR EMBEDDING THE SEC

C1 Calculating numerical measures by hand and using technology. Exposure to software such as spreadsheets and calculators will further enhance the students' experience of real-world data analysis.

D1 Interpreting numerical measures.

Exemplar

In 2023, a survey was conducted to ascertain how many people in Wales consider themselves to be "Welsh". The following is a list of numbers, in thousands, of people living in an area of Wales who consider themselves to be "Welsh":

40.4 86.3 50.3 50.7 57.1

Find the mean number of people per area of Wales in this sample who consider themselves to be "Welsh" in 2023.

(Source: https://statswales.gov.wales/Catalogue/Equality-and-Diversity/National-Identity/nationalidentity-by-area-identity)

Using the formula : $\frac{40.4+86.3+50.3+50.7+57.1}{5} = 56.96$ *thousand people, so* 57 thousand (or 570 000 *people, to 3 significant figures.*

COMMON AND POSSIBLE MISCONCEPTIONS

Common errors include:

- incorrect data entry into a calculator/spreadsheet;
- when calculating the mean from a frequency table, dividing by the number of categories rather than the total frequency (the same applies to the estimated mean of a grouped frequency table);
- incorrectly finding the midpoint of a grouped category.

Most of these errors can be avoided through the effective use of the statistics mode of the calculator; encourage students to check their answers this way.

When using the calculator functions, errors often occur due to students selecting the incorrect mode, or not setting up a frequency column.

NOTES

Q_2 is a standard notation for sample median and may be used to achieve full marks (in line with the mark scheme).

0b. Foundations of Statistics: Introduction to Probability (2.1) (2.2)	Teaching time 2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand the concept and scale of probability.
- Illustrate probability using probability trees, Venn diagrams and two-way tables and use them to determine probabilities of basic events.
- Understand and use set theory notation.

TEACHING POINTS

Recap the skills taught at GCSE Higher Tier (9-1).

Set theory notation is revision from GCSE, specifically $A \cup B$ and $A \cap B$ as the union and intersection of the sets *A* and *B* respectively. The complement of *A* will be denoted by A'.

The concept of probability may need to be revised, including basic problems.

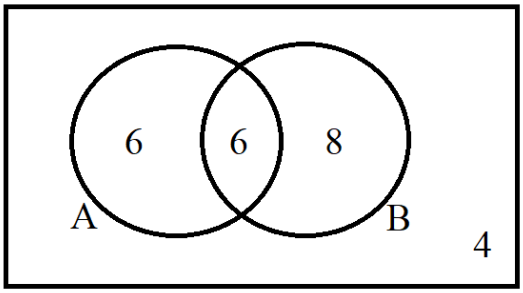
For example:
There are 4 green, 6 blue and 8 orange sweets in a bag. What is the probability of randomly choosing a green sweet from the bag?

The notation of P(*A*) can be introduced here, linking to the set theory notation above. It is important that students get familiar with this notation, and it is equally important that questions are correctly phrased. The omission of the word “randomly” in the above example changes the nature of the question, and it is important that students see correct phrasing from the start.

Venn diagrams is revision from GCSE and it is advisable that probabilities of two or three events that can be represented on Venn diagrams are practised. **At A level, it is common for 3 events to appear.** Practice labelling the Venn diagrams with either frequencies or probabilities, depending on the context.

Exemplar

Let A be the event "a person chosen at random from the class has brown hair" and B be the event "a person chosen at random from the class has blue eyes". Out of a class of 24 students, 6 people have both brown hair and blue eyes, 12 people have brown hair and 14 people have blue eyes. Draw a Venn diagram and find $\mathrm{P}(A' \cap B')$.



So $\mathrm{P}(A' \cap B') = \frac{4}{24} = \frac{1}{6}$

Two-way tables should be revised from GCSE, and probabilities of two events that can be represented in a two-way table should be practised (the example above works well here too).

Exemplar

Find $\mathrm{P}(A' \cap B')$

	A	*A'*	*Total*
B	***6***	*8*	***14***
B'	*6*	*4*	*10*
Total	***12***	*12*	***24***

So $\mathrm{P}(A' \cap B') = \frac{4}{24} = \frac{1}{6}$

Probability trees is revision from GCSE and it is advisable that probabilities that can be represented on a tree diagram are practised.

For example

The probability of a train being late is 0.4. The probability of being late to work if the train is late is 0.8 and the probability of being late to work if the train is on time is 0.3. Draw a probability tree and find the probability of being late to work.

When faced with questions like the example above, encourage students to begin each solution with "Let A be the event...". This not only encourages good practice but will raise the quality and readability of solutions.

Note: Problems may well extend to 3 events as in Paper 1 SAM

Exemplar

A survey of students taking a politics module at university were asked which, if any, of three influential political leaders, *N*, *A* and *F*, they admired.

Of these students:

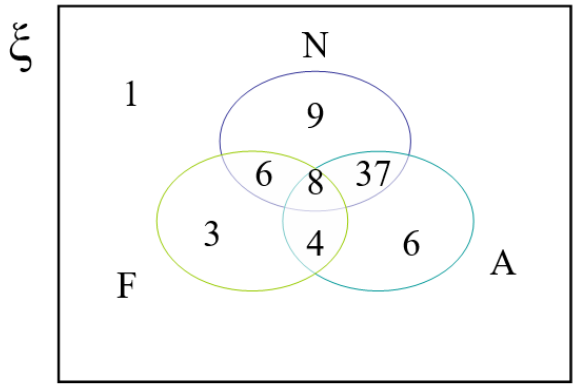
60 admired *N*
55 admired *A*
21 admired *F*
45 admired *N* and *A*
12 admired *A* and *F*
14 admired *N* and *F*
8 admired all three leaders and 1 admired none of the leaders.

For a randomly selected student:

Event N is defined as 'student admires *N*''
Event A is defined as 'student admires *A*'
Event F is defined as 'student admires *F*'

Draw a fully labelled Venn diagram to illustrate this information.

Solution



OPPORTUNITIES FOR EMBEDDING THE SEC

D1 Interpreting probabilities in context.

At this moment, there is little in the way of direct application to the SEC. However, the skills in this sub-unit will be vital for the application of **Stage D** of the SEC in later units.

Exemplar

In a survey of the UK labour market during October to December 2016, there were 3,793 thousand people (to the nearest thousand) identifying their race as "ethnic minority" (this category includes Chinese, Pakistani, Bangladeshi etc.). According to the survey, 74 thousand people identified as a "Chinese male", 2,105 thousand people identified as an "ethnic minority male" and 91 thousand people identified as a "Chinese female". Let *M* be the event "a randomly chosen person identifies as male" and let C be the event "a randomly chosen person identifies as Chinese".

Find $\mathrm{P}(M' \cap C')$ and explain what this means in this context.

(Source: Office for National Statistics – A09: Labour market status by ethnic group)

Using a two-way table:

	M	M'	Total
C	74	91	165
C'	2031	1597	3628
Total	2105	1688	3793

$\mathrm{P}(M' \cap C') = \frac{1597}{3793}$ *is the probability that a randomly chosen person from the ethnic minority category in survey has identified as neither male nor Chinese.*

COMMON AND POSSIBLE MISTAKES

Common errors include:

- confusing the notation for union ∪ and intersection ∩;
- misreading questions involving the words "or" or "and".

NOTES

Probability theory will be seen properly in Unit 3, but it is important for students to revise the basics prior to this.

0c. Foundations of Statistics: The Statistical Enquiry Cycle (SEC)	Teaching time Minimum 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Appreciate the basic idea of the Statistical Enquiry Cycle.
- Appreciate the bias introduced through leading questions or non-random sampling.

TEACHING POINTS

The Statistical Enquiry Cycle could be introduced to students at this early stage, together with some examples. Group work, discussion and case studies are ideal here to allow students to appreciate the real-world contexts of statistics. Although not on the A level specification, questionnaire design from the legacy GCSE specification is an ideal starting point, in particular allowing students to critique biased questions.

Exemplar

Pierre is going to carry out a survey using a questionnaire.

He wants to find out how often people play sport.

Design a suitable question for Pierre to use on his questionnaire.

(Source: GCSE 1MA0 Foundation 2F November 2016)

How many hours do you play sport per week?

Less than 1 hour	*At least 1 hour but less than 2 hours*	*At least 2 hour but less than 4 hours*	*4 hours or more*

This can then move on to biased sampling methods, allowing students to form their own opinions and discuss them with peers.

Group work activities could include: given a particular question, discuss how you would go about collecting data in order to answer this question; how would you use these data to answer your question; how would you present your data to explain to others; how could you do things differently?

The answers to these questions need not necessarily involve unbiased sampling techniques or in-depth statistical analysis knowledge outside of what they have learnt at GCSE. As long as the process has been followed with justification, students will begin to appreciate the nature of the SEC.

For example:
You are working in a music retail store and you are in charge of creating the main shop window display for the month. You want to create a display that appeals to the majority of people who are likely to buy music. Discuss in your groups how you would collect data in order to answer this question; how you would use these data; how you would present your data to justify to the manager and what you would do as a result of your analysis. Also discuss the disadvantages of your process.

There are many possible ways to tackle such a process; lower ability students may come up with something as basic as:

Ask some friends what music is popular. Find out the most popular artist. Make a bar chart of the data and create a display for that artist's music. My friends may not like the same music as other people.

As basic as this answer is, it satisfies points **A1**, **A3**, **B4**, **C1**, **C2**, **D1**, **D2** and **E2** of the SEC, showing a basic appreciation of the process.

Case study activities could include: providing an attempt at data collection and analysis in order to answer a question, discuss the advantages and disadvantages of the attempt. How would you improve the method?

For example:
To govern the vast Inca Kingdom (circa. 1500, Peru), it was necessary to collect statistical information like population figures, the character of the soil, crops etc. Specifically trained officials were entrusted with this task. Knotted strings (called the Quipu String) enabled the officials to register and remember the data they had collected. Discuss the advantages and disadvantages to this method of data collection.

(Source: National Museum of Denmark)

If the group activity suggested earlier is employed, attempts made by students can be presented to other groups for discussion.

For students to appreciate the importance behind effective sampling, the exit polls for the UK general elections 1997, 2001, 2005, 2010, 2015 and 2017 are good examples of when the samples reflected the true population. The exit poll for the UK general election 1992 is a good example of when the sample does not reflect the true population.

OPPORTUNITIES FOR EMBEDDING THE SEC

This entire sub-unit is dedicated to the SEC.

COMMON AND POSSIBLE MISTAKES

This is a topic which has not been examined in great depth until now. There are parallels to the modelling cycle seen in GCE A level Mathematics, so some common mistakes from this topic will be given, as it is anticipated that these mistakes will arise again:

Students can generally correctly state assumptions, but they need to make sure that any assumptions or statements about the model relate directly to the context they are considering.

NOTES

This topic will be embedded throughout the scheme of work and it is important that students have an awareness of the SEC as early as possible. There is an opportunity to put the SEC into practice by completing a statistical project using the material learnt in this course. Topics could be suggested that interlink with other A Level subjects beyond statistics. Such a project gives an opportunity for using statistics in practice across mathematics, business, law, economics, sports, biosciences, earth sciences and a range of other subjects. This will be detailed at the end of the Scheme in Unit 26.

UNIT 1: Numerical Measures

Return to overview

SPECIFICATION REFERENCES

1.4 Compare different data sets, using appropriate diagrams or calculated measures of central tendency and spread: mean, median, mode, range, interquartile range, percentiles, variance and standard deviation.

1.5 Calculate measures using calculators and manual calculation as appropriate.

1.6 Identify outliers by inspection and using appropriate calculations.

1.7 Determine the nature of outliers in reference to the population and original data collection process.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics

Numerical measures covered so far: median, mean, mode, range (Unit 0a).

Frequency tables and grouped frequency tables (Unit 0a).

The Statistical Enquiry Cycle (Unit 0c).

GCSE (9-1) in Mathematics at Higher Tier

A2 Substitute numerical values into formulae and expressions.

S4 Quartiles, Interquartile range, outliers; interpretation, analysis and comparison of data sets through the aforementioned.

S5 Apply statistics to describe a population.

KEYWORDS

average, bias, central tendency, chance, data, deviation, equation, estimate, formula, frequency, frequency table, grouped frequency table, interquartile range, leading questions, linear scaling, mean, measure, median, mode, outliers, population, quartiles, probability, probability tree, random, range, sample, skew, simultaneous equations, scaling, spread, standard deviation, substitute, symmetrical, two-way table, variance, Venn diagram,

UNIT SUMMARY

This unit extends the concepts of numerical measures learnt at GCSE, and also lets students appreciate that each measure has its place in statistics. Knowledge of when they are to be used is important. Introducing the correct notation and terminology is crucial at this stage.

The concepts of variance, standard deviation and outliers will be revisited many times throughout the A level, and the suitability of numerical measures in given contexts provides opportunities to embed the SEC throughout.

Effective use of the statistics mode on a calculator is essential. A spreadsheet (or the spreadsheet mode on the calculator) can be used, together with the in-built functions for numerical measures (where appropriate). When locating the standard deviation on the calculator, it is important that students know which refers to the sample standard deviation, and which refers to the population standard deviation. Most calculators use s_x for the sample standard deviation ((n – 1) divisor) and σ_x (n divisor), for the population standard deviation.

It will be desirable for students to know when the use of s_x and σ_x is appropriate when they see *t-test*s in Unit 19.

1a. Numerical Measures: Quartiles, Percentiles and the Interquartile Range (1.4) (1.5)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Determine the median, lower quartile, upper quartile and the interquartile range from a discrete set of data.

TEACHING POINTS

Recap the skills taught at GCSE Higher Tier (9-1).

The median is revision from Unit 0. There have been ambiguities and alternative methods for calculating the quartiles from a discrete set of ascendingly ordered data:

Method 1: Lower half and upper half.

A discrete data set can be split into a lower half and upper half.

- Locate the sample median,
- All values less than the median form the lower half and all values above the median form the upper half. The median itself is not part of either half (so is ignored for these purposes).
- The lower quartile (LQ) is the median of the lower half and the upper quartile (UQ) is the median of the upper half.

Exemplar

Consider the set of data 1, 3, 4, 4, 6, 7, 8. Find the lower quartile and upper quartile.

Using Method 1: *The median of the data set is 4. This splits the data set into the lower half (1,3,4) and the upper half (6,7,8). Note that the median is not a member of either half. The lower quartile is the median of the lower half, in this case 3. The upper quartile is the median of the upper half, in this case 7.*

Method 2: Round-up method.

Let *n* be the size of the data set.

- Calculate $\frac{n}{4}$.
- If $\frac{n}{4}$ is a non-integer, round UP to the nearest integer and the value with that index is the lower quartile (LQ).
- If $\frac{n}{4}$ is an integer, the mean of the $\frac{n}{4}$ and $\left(\frac{n}{4}+1\right)$ indexed values is the lower quartile (LQ).

  The upper quartile is calculated in the same way, except using $\frac{3}{4}n$.

Consider the previous example, and using Method 2:

$n=7$, so for the lower quartile we consider $\frac{7}{4}=1.75$ which is not an integer. So rounding up, we want the 2nd number, which is 3. For the upper quartile we consider $\frac{3}{4}(7)=5.25$ which is also not an integer. Rounding up, we want the 6th number, which is 7.

Also, students can use a calculator directly to obtain results as this is a GCSE topic. All valid answers will gain marks.

Note that the two methods are not equivalent, for example the case where $n=9$ yields two different answers, but both will be accepted. Any quartile value obtained from the statistics mode on a calculator will also be accepted. In all cases, the interquartile range (IQR) is calculated by UQ – LQ.

OPPORTUNITIES FOR EMBEDDING THE SEC

C1 Calculating numerical measures.

D1 Interpreting numerical measures.

Most of point **C1** is covered already in this sub-unit. Exposure to software such as spreadsheets and calculators will further enhance the students' experience of real-world data analysis.

Exemplar

In a UK survey between April 2013 and December 2013 of people with disabilities aged 16-24, the number of people in employment was recorded every three months. The following is a table showing these figures:

Survey Period	Number of people with disabilities in employment (thousands)
Apr-Jun 2013	2,841
Jul-Sep 2013	2,812
Oct-Dec 2013	2,907
Jan-Mar 2014	2,891
Apr-Jun 2014	2,953
Jul-Sep 2014	3,073
Oct-Dec 2014	3,052
Jan-Mar 2015	3,144
Apr-Jun 2015	3,184
Jul-Sep 2015	3,153
Oct-Dec 2015	3,202
Jan-Mar 2016	3,259
Apr-Jun 2016	3,320
Jul-Sep 2016	3,393
Oct-Dec 2016	3,489

It is claimed that the interquartile range of the number of people with disabilities in employment is 350 thousand. Comment on this claim.

Source: Office for National Statistics – A08 Economic activity of people with disabilities aged 16-64, UK

Put the numbers in order:

2812, 2841, 2891, 2907, 2953, 3052, 3073, 3144, 3153, 3184, 3202, 3259, 3320, 3393, 3489

The median is 3144. The median of the lower half is 2907 and the median of the upper half is 3259. So the lower quartile is 2907 and the upper quartile is 3259 (NOTE: these values can also be obtained from the calculator). The interquartile range is 352, so the claim may be accurate, since the interquartile range for the period recorded is 352 thousand, similar to the claim of 350 thousand. However, this is correct only for the period recorded and does not take into account data from before or after this period.

COMMON AND POSSIBLE MISTAKES

Common errors include:

- including the median in the lower half and/or upper half when employing Method 1;
- rounding down when employing Method 2;
- not considering integer and non-integer cases in Method 2;
- inputting data into the calculator incorrectly.

NOTES

Extension Exemplar

In order to locate the *k*th percentile of a discrete data set, Method 2 is employed by using $\frac{kn}{100}$. Here, the lower and upper quartiles of Method 2 are equivalent to locating the 25th and 75th percentiles respectively. This is not in the specification.

Find the 23rd percentile of the data set.

$n = 7$, so consider $\frac{23}{100}(7) = 1.61$ which is not an integer. Rounding up, we want the 2nd number, which is 3.

Inter-percentile ranges, deciles and quintiles are used in real-world statistics, and may be assessed in A level Statistics - this will be defined in exam questions.

The notation Q_1 and Q_3 are standard notation to denote the lower and upper quartiles respectively (See notes in Unit 0a). Students may use these symbols and gain access to full marks (in line with the mark scheme).

1b. Numerical Measures: Variance and Standard Deviation (1.4) (1.5)	Teaching time 2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Use a calculator to determine the variance and standard deviation from an unsorted list of data, a frequency table or grouped frequency table.
- Calculate the variance and standard deviation using a formula and summary statistics.

TEACHING POINTS

Recap the skills taught at GCSE Higher Tier (9-1).

The formula for the estimated population variance from a sample $s_x^2 = \frac{1}{n-1}\sum\left(\bar{x}-x\right)^2$ is in the formula book. For technical reasons, it is advisable to omit the explanation of $n-1$ instead of n at this point, but students need to know that the divisor, $n-1$, is for the estimated population variance from a sample and the divisor, *n,* is for the variance when calculated from a whole population.

The standard deviation is the square root of the variance; this can be explained by the fact the deviations from the mean were squared when calculating the variance.

The alternative formula for the estimated population variance from a sample $\left(\frac{1}{n-1}\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)\right)$ is in the formula book, and could be derived depending on the ability of the students. Both formulae for the sample variance are in the formula book.

When given a frequency table, the use of $s_x^2 = \frac{1}{\sum f - 1}\sum f\left(\bar{x}-x\right)^2 = \frac{1}{\sum f - 1}\left(\sum f x^2 - \frac{\left(\sum f x\right)^2}{\sum f}\right)$ may be used. When given a grouped frequency table, estimates of the variance and standard deviation can be found by using x as the midpoint of the class. Students must understand why these are only estimates.

Questions can include finding the (estimated) variance and standard deviation from: a list of numbers, a (grouped) frequency table, or summary statistics (given $\sum\left(\bar{x}-x\right)^2$ and *n*, or $\sum x^2$, $\sum x$ and *n*).

Utilising the statistics mode on the calculator is very important here – students could be encouraged to input the data into their calculator given the raw data and avoid using the formula unless the summary statistics are given. In the event that both the data and the

summary statistics are given, encourage students to use both methods and check their answers.

OPPORTUNITIES FOR EMBEDDING THE SEC

C1 Calculating numerical measures.

Most of point **C1** is covered already in this sub-unit. Exposure to software such as spreadsheets and calculators will further enhance the students' experience of real-world data analysis.

Exemplar

The table below shows the recorded number (in thousands) of people in employment who have declared themselves as mixed race:

Survey Period	Number of people (thousands)
Jan-Mar 2015	329
Apr-Jun 2015	302
Jul-Sep 2015	305
Oct-Dec 2015	287
Jan-Mar 2016	312
Apr-Jun 2016	332
Jul-Sep 2016	365
Oct-Dec 2016	372

Find the standard deviation of these data.

(Source: Office for National Statistics – A09: Labour market status by ethnic group)

Using the calculator: $s_x = 30.3$ *thousand (3 sig. fig.).*

COMMON AND POSSIBLE MISTAKES

Common errors include:

- substituting the incorrect values into the formulae;
- substituting the correct values into the incorrect places in the formulae;
- misinterpreting $\sum x^2$ as $\left(\sum x\right)^2$ and vice versa;
- dividing by *n* instead of $n-1$;
- rounding too early in the calculations;
- forgetting to square root the variance when asked for the standard deviation.

These are addressed through plenty of practice.

Other common errors include:

- dividing by the number of groups rather than the total frequency;
- finding $f\sum x^2$ instead of $\sum fx^2$;
- using σ instead of *s*;
- using σ^2 instead of s^2;
- using *s* instead of s^2 (and vice versa).

These can be addressed by utilising the statistics mode on the calculator.

Incorrect data entry on a calculator is another common error.

1c. Numerical Methods: Outliers and Scaling (1.6) (1.7)	Teaching time 2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Know that a value may be considered an outlier if it lies outside the intervals $(\text{LQ} - 1.5 \times \text{IQR}, \text{UQ} + 1.5 \times \text{IQR})$ (may be given in question) or $\left(\bar{x} - 3s_x, \bar{x} + 3s_x\right)$.
- Identify possible reasons for outliers in a given context.
- Find numerical measures from a linear scaling.

TEACHING POINTS

Recap the skills taught at GCSE Higher Tier (9-1).

The concept of outliers will have been touched upon briefly at GCSE, and revisited here.

It is suggested that plenty of questions involving the calculation of outliers are given. When faced with grouped frequency tables, estimates of outlier boundaries can be calculated from estimates of the mean and standard deviation. Students must be able to identify if there are

- definitely no outliers (the outlier boundary lies outside the class boundaries)
- definitely some outliers (the outlier boundary lies below an entire class), or
- possibly some outliers (the outlier boundary lies inside a class).

The formulae LQ – 1.5IQR and UQ + 1.5IQR are referred to in the literature and practice as **Tukey Fences**, named after John Tukey (1915-2000), the American statistician. Whilst the term “Tukey fence” will not be examined, it is good practice to prepare students for terminology they would likely come across in employment.

The two formulae for calculating the outlier boundaries are in the objectives above.

Exemplar

In a survey, a sample of 44 people attend a support group for people with mental health issues. The ages (x years) are summarised in the grouped frequency table.

Age (x years)	$20 < x \leq 23$	$23 < x \leq 25$	$25 < x \leq 27$	$27 < x \leq 30$	$30 < x \leq 40$
Frequency	3	15	18	6	2

Given that the mean of these data is $\bar{x} = 25.8$ and the standard deviation is $s_x = 2.72$ (both given to 3 sig. fig.), determine whether there are any outliers.

Since $\bar{x} - 3s_x = 25.8 - 3 \times 2.72 = 17.6$, there are definitely no outliers at the lower end. Since $\bar{x} + 3s_x = 25.8 + 3 \times 2.72 = 34.0$, then there may be outliers at the upper end, but we do not know for sure.

Explain to students that a value lying outside the outlier boundaries may be considered an outlier and possible reasons may be suggested as to the source of the outlier. However, these values should not be omitted from the sample unless there is a genuine reason relating to the recording or circumstances of the value. Such values are called **anomalies** and emphasise that **anomalies** and **outliers** are different. Explain that although there is a statistical method to determine outliers, anomalies cannot be determined statistically and must be identified through a thorough investigation of the experiment. The practice of removing anomalies from the data is called **cleaning** the data.

Exemplar

Some French beans were picked and the beans were removed from the pods, ready for cooking. The number of beans in each pod was recorded. The majority of beans in each pod ranged from 2 and 7. One pod did not contain any beans, and one pod contained 15 beans. It is claimed that these two pods are outliers and should be removed from the sample. Suggest possible explanations as to the origin of these pods and comment on the suggestion that they should be removed from the sample.

The pod containing no beans may have come from a weaker plant and the pod containing 15 beans may have come from a very strong yielding plant. However, without further evidence that the pods picked were from a different population of French beans, or that the data were recorded erroneously, the data should not be removed from the sample.

The follow-up question below could also be included in the above exemplar:

Exemplar

It was later discovered that the pod containing 15 beans came from a different species of bean.

Since there is evidence that this pod originated from a different population from that of the rest of the sample, this value is an anomaly and can justifiably be cleaned from the data.

Students must be aware of the effects of linear scaling. The vertical line chart or the scatter graph activities on Desmos may be of assistance here. Support students to appreciate that if a constant is added to or subtracted from the data, then it is added to or subtracted from the mean. If the data are multiplied by a constant then the mean is also multiplied by the constant. So if $y = ax + b$ then $\bar{y} = a\bar{x} + b$.
Another key area of understanding is that the standard deviation is unchanged if a number is added to or subtracted from the data, but if the data are multiplied by a constant then the standard deviation is also multiplied by the constant. So if $y = ax + b$ then $s_y = as_x$ (for $a > 0$). Also, ensure students are aware that $s_y^2 = a^2 s_x^2$.

The classic example is the conversion from °C to °F:

Exemplar

During a science experiment, the temperature (x °C) of a solvent is recorded and the experiment is replicated 5 times. The mean temperature is $\bar{x} = 78$ °C and the standard deviation is $s_x = 3.42$ °C. The scientist wishes to have all the data recorded in °F, so uses the conversion formula
$y = \frac{9}{5}x + 32$ where x is measured in °C and y is measured in °F.
Find the mean temperature and standard deviation in °F.

The mean temperature is $\bar{y} = \frac{9}{5}\bar{x} + 32 = \frac{9}{5} \times 78 + 32 = 172.4$ *°F.*

The standard deviation is $s_y = \frac{9}{5}s_x = \frac{9}{5}(3.42) = 6.156$ *°F.*

OPPORTUNITIES FOR EMBEDDING THE SEC

C3 Determining whether or not data should be removed.

D1 Interpreting numerical measures in context and making suitable decisions.

COMMON AND POSSIBLE MISTAKES

- Using outlier boundaries $\bar{x} \pm 1.5s_x$ (or LQ – 3 ×IQR, UQ + 3 ×IQR)
- Using the notation μ and σ instead of $\bar{x}$ and s_x.
- Forgetting to square-root the variance when using the standard deviation.
- Automatically assuming that an outlier should be removed from data analysis.
- Assuming that there are definitely outliers when the outlier boundary lies within a class of a grouped data set.
- Using the formula $s_y = as_x + b$, (NOT $s_y = as_x$), when using a linear scaling.

NOTES

In the literature and in practice, outlier boundaries of $\bar{x} \pm 2s_x$ can be used. Under a normal distribution, the intervals $(\mu - 3\sigma, \mu + 3\sigma)$ and $(\text{LQ} - 1.5 \times \text{IQR}, \text{UQ} + 1.5 \times \text{IQR})$ are approximately equivalent.

Linear scaling will be seen again in Units 4, 6 and 16.

1d. Numerical Measures: Suitability of Numerical Measures (1.4)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Appreciate the advantages and disadvantages over using the mean/standard deviation or the median/interquartile range.
- Select appropriate numerical measures for a given context.

TEACHING POINTS

The advantages, disadvantages and suitability of each numerical measure depend on the context, and provide an opportunity for discussion, case study and application.
When students are asked to choose and justify a measure then any answer for which adequate justification can be provided will gain credit. Comments should be made in the form of bullet points.
Advantages and disadvantages include, but are not limited to:

	Advantages	**Disadvantages**
Mean	Influenced equally by each value in the data. More appropriate for large samples or symmetrical data.	In a small sample, it may be significantly affected by one very unusual value. Requires a complete dataset.
Mode	Easy to find.	There may be more than one mode, and it may not even be useful.
Median	Not all values are needed to find the median. Less affected by skew.	Can be difficult to identify if the data set is very large (and no available technology) Ignores some data in the dataset and may not be a “true” measure of average
Range	Easy to calculate. Only needs the biggest and smallest values.	Is influenced by a very large or very small value
Interquartile Range	Not influenced by very large or very small values. Less affected by skew/non-symmetrical distributions. Very easy for small samples without technology.	Difficult when there are a lot of values unless suitable technology is available. Ignores the top and bottom 25% so information about the full distribution is lost.
Standard Deviation	Influenced equally by each piece of data. More appropriate for large samples or symmetrical data. Very useful in statistical modelling.	More complicated than other measures if there is no technology available. Difficult when there are a lot of values unless suitable technology is available. Requires a complete dataset.

Use of tarsia puzzles or matching activities is particularly useful here.

Decisions about numerical measures can be made from the context alone, without knowledge of the calculation of above numerical measures (*a priori* decisions).

The mean and standard deviation are appropriate numerical measures for the heights of people in the United Kingdom, since the data set is large and symmetrical.

Decisions about numerical measures can also be made given a choice of numerical measures after calculation, and applying it to the context (*a posteriori* decisions).

From a sample of eight year old children from a primary school, the mean height is 150 cm and the median height is 120 cm. The median would be an appropriate measure because 150 cm is quite big for an eight year old child and the value is likely to have been influenced by an outlier.

OPPORTUNITIES FOR EMBEDDING THE SEC

Plenty of contextual problems which are designed for students to determine the most suitable measure of central tendency and spread to use, giving reasons.

A5 Making *a priori* decisions (decisions prior to the evaluation of measures and based on the context), giving reasons.

C3 Making *a posteriori* decisions (decisions after the evaluation of measures) and identifying whether the data has been misrepresented.

D1 Making *a posteriori* decisions (decisions after the evaluation of measures) and interpreting the calculated numerical measures in context.

For example, a series of different contexts can be given to groups of students for them to form their own opinions over which is the most suitable numerical method. After discussion with other students, giving their reasons, they will gain a better understanding of the SEC and develop an appreciation of how to represent their findings clearly.

Exemplar

The table below shows the number of Filipino people living in England and Wales and their age groups, according to the 2021 census.

Age	Frequency
0-15	24 320
16-24	20 410
25-34	30 435
35-49	46 415
50-64	30 015
65+	10 520

(Source: Office for National Statistics – Detailed ethnic group by age and sex in England and Wales, Census 2021)

Without any calculation, choose a measure of central tendency and a measure of spread to analyse these data, justifying your answer.

I would use the median and interquartile range.
We cannot find the mean or standard deviation as there is no upper class boundary for the 65+ class. Since the frequency of this class is less than a quarter of the total frequency, it would not be included in the interquartile range, and I can still find the median by removing the bottom 38 ages.

Or

I would use the mean and standard deviation.
Although we may know which class the median and quartiles lie in, we wouldn't know exact values.
Choosing a reasonable upper class boundary for the last class (for example, 90 years old), I could find an estimate for the mean and standard deviation.

COMMON AND POSSIBLE MISTAKES

The suitability of the numerical measures must be presented in a form suitable to the context of the question, and not a generic "bookwork" answer.

Exemplar

To test the lifetime of a brand of household furniture, a random sample of 30 armchairs are subjected to repeated compressions by a hydraulic press to simulate a human weight. After each compression, the density of the cushion is calculated. The number of compressions required for the density of the cushion to reach a set benchmark is recorded. The armchair is then recycled. Give two reasons why the median and interquartile range would be better numerical measures to use over the mean and standard deviation.

Acceptable response structure

- *The median does not require all 30 armchairs to be recycled, only 16 need to reach the benchmark density to calculate the median.*
- *The interquartile range is unaffected by armchairs which, for example, break the armchair straight away or have cushions which never reach the benchmark density.*

Unacceptable response structure

- *The median does not require a complete dataset*
- *The interquartile range is unaffected by extreme values*.

Note that in the example above, the “unacceptable response” may not gain full marks (in line with the mark scheme) because there is no context from the original question displayed in the answer.

UNIT 2: Data Representation and Interpretation

Return to overview

SPECIFICATION REFERENCES

1.1 Interpret statistical diagrams including bar charts, stem and leaf diagrams, box and whisker plots, cumulative frequency diagrams, histograms (with either equal or unequal class intervals), time series and scatter diagrams.

1.2 Know the features needed to ensure an appropriate representation of data using the above diagrams, and how misrepresentation may occur.

1.3 Justify appropriate graphical representation and comment on those published.

1.8 Appreciate that data can be misrepresented when used out of context or through misleading visualisation.

4.1 Know and use terms for variables: random, discrete, continuous, dependent and independent.

4.5 Interpret graphical representations or tabulated probabilities of characteristic discrete random variables.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics

Mean, median, mode and basic probability (Unit 0a) will be useful here, but not essential – knowledge from GCSE (see below) should be sufficient.

Knowledge of variance and standard deviation will allow scope for a larger scope of questions to be explored (Unit 1a)

GCSE (9-1) in Mathematics at Higher Tier

G18 Angles of sectors of circles (for Pie Charts).

P1 Describe and analyse the frequency of outcomes.

P2 Randomness and fairness.

S1 Infer properties of populations from a sample.

S2 Interpret tables, charts and diagrams, including frequency tables, bar charts, pie charts, vertical line charts, time series and know their appropriate use.

S3 Interpret histograms and cumulative frequency graphs and know their appropriate use.

S4 Interpret, analyse and compare the distributions of data sets from univariate empirical distributions through appropriate graphical representation (inc. box plots) and measures of central tendency and spread.

KEYWORDS

angle, bar charts, box and whisker diagrams, box plots, class, class boundary, class width, continuous, cumulative frequency, discrete, frequency, frequency density, histograms, interquartile range, line charts, mean, median, misrepresentation, mode, qualitative, quantitative, quartile, percentiles, pie charts, random, range, sector, stem and leaf, time series, trend, variables,

UNIT SUMMARY

This unit is a recap from GCSE, albeit with more analysis and evaluation encouraged. The SEC provides lots of starting, and development, points on how to interpret data and avoid misrepresentation.

It is at this point we introduce the formal language of statistics, and it is contained in this unit prior to analysing published statistical diagrams. It is important students use and know the correct terminology thus are able to justify the suitability of statistical diagrams to the contexts of the problem.

The use of spreadsheets and databases (and their associated in-built functions) is included as a separate sub-unit – this is new to the A level in Statistics (as it is to the GCE AS/A level in Mathematics and GCSE (9-1) Mathematics). Although there is no specific reference to software terminology in the specification, it is included in Appendix 4: Use of calculators and other technology of the specification (page 42, 9ST0 Specification).

There are plenty of interactive Desmos activities that have been created to help teachers and students in these topics. Bar charts, dual bar charts, stacked bar charts, pie charts, vertical line charts and histograms are all available to use.

Note that comparative pie charts will not be specifically assessed on this course but an appreciation that area represents frequency is expected.

On the Maths Emporium (www.mathsemporium.com), you can find two end-of-topic tests on this unit:

- Data Visualisation
- Technology and Software

(both can be found here).

2a. Data Representation and Interpretation: The Language of Statistics (1.1) (1.3) (4.1)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Know and understand the terms: random, variable, random variable, qualitative, quantitative, discrete, and continuous.
- Know and understand the terms: class, class width, class interval, upper class boundary, lower class boundary.
- Identify the difference between qualitative and quantitative variables in a variety of contexts.
- Identify the difference between discrete quantitative and continuous quantitative variables in a variety of contexts.
- Identify the type of variable given a univariate statistical diagram.
- Identify the class width and boundaries given a grouped frequency table or histogram.

TEACHING POINTS

It is highly important that students are introduced to the language of statistics as early as possible. It will underpin the quality of their calculations, conclusions and justifications presented. Moreover, it is important to be able to identify the correct type of variable given a statistical diagram. It was decided to formalise the topic in this early sub-unit, and it will be revisited in future units as more terminology arises.

Prior to revising histograms, students will need to know the correct terminology for classes. This section is a logical point to do this.

The keywords for this section, with definitions are:

Random (in the context of random variable)	Cannot be predicted with certainty
Variable	Something which is able to take any of a range of values
Random Variable	A variable whose values cannot be predicted with certainty
Qualitative	Categorical
Quantitative	Numerical
Discrete / Discrete Variable	Takes set/pre-described values / a variable whose values can be counted
Continuous / Continuous Variable	Takes any number in a range of values/ A variable that must be measured
Class	A group of values of a random variable e.g. $10 \leq x < 15$
Class Width	The range of the class e.g. for $10 \leq x < 15$, the class width is 5.

Class Interval	The interval of the class e.g. $10 \leq x < 15$
Upper Class Boundary	The upper limit of the class e.g. for $10 \leq x < 15$, the upper class boundary is 15.
Lower Class Boundary	The lower limit of the class e.g. for $10 \leq x < 15$, the lower class boundary is 10.

Matching activities, dominoes or tarsia puzzles can be used. Although the definitions are important, it is more important that students can identify when to use them in context.

Exemplar

During the 2017 General Election, a sample of 1,000 constituents eligible to vote in Hereford were surveyed at random and asked which political party they were voting for. The results were recorded and displayed using a pie chart.

a) **What does "at random" mean in this context?**
The eligible constituents in the population each had an equal probability of being chosen to be surveyed in the sample.

b) **What is the variable being recorded?**
The variable is the political party being voted for.

c) **Describe this type of variable.**
The variable is a qualitative random variable.

Other questions could be as simple as presenting students with a fully labelled pie chart and asking them to describe the type of variable.

Remind students that bar charts should have gaps.

OPPORTUNITIES FOR EMBEDDING THE SEC

Plenty of real-world contexts should be provided to students.

A3 Determining what data in the scenario to collect (either from context or an initial hypothesis).

A5 Identifying the type of variable in context.

Exemplar

The table below shows the recorded number (in thousands) of people in employment who have declared themselves as mixed race:

Survey Period	Number of people
Jan-Mar 2015	329
Apr-Jun 2015	302
Jul-Sep 2015	305
Oct-Dec 2015	287
Jan-Mar 2016	312
Apr-Jun 2016	332
Jul-Sep 2016	365
Oct-Dec 2016	372

Determine and describe the random variable in this context.

(Source: Office for National Statistics – A09: Labour market status by ethnic group)

The random variable is the recorded number of people in employment who declare as mixed race. It is a discrete quantitative random variable.

Questions like this example are good for differentiating those with good language comprehension.

COMMON AND POSSIBLE MISTAKES

The spelling of “qualitative” and “quantitative” is often incorrect. Although an incorrect spelling may still gain full marks (provided the words are clearly correct and in line with the mark scheme) in examinations, it should be corrected at every opportunity. The majority of mistakes here are in language comprehension. In the above example, some students are unsure whether the variable is the time period, what ethnicity people have declared, or the number recorded. A distinction should be made between the words “discrete” and “discreet” and corrected when appropriate.

NOTES

Some ambiguous contexts involving quantitative data may present themselves; it is important that students justify themselves when deciding between discrete and continuous.

For example, the age of a person (if recorded in years) may be discrete or continuous, depending on how the data are recorded. The most likely answer will be discrete, but a student who describes the variable as continuous could be awarded full marks (in line with the mark scheme) provided they justify why it is continuous (e.g. *time is continuous*).

There are links to Unit 2d where students will be required to identify the most suitable statistical diagram to represent data.

2b. Data Representation and Interpretation: Reading and Interpreting Univariate Statistical Diagrams (1.1) (4.5)	Teaching time 2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Identify and interpret information presented in a simple bar chart, dual bar chart, stacked bar chart, vertical line charts and stem and leaf diagrams.
- Identify and interpret information presented in a pie chart.
- Identify and interpret information presented in a grouped frequency table, cumulative frequency curve or histogram (of equal and unequal class intervals).
- Identify and interpret information presented in a box and whisker plot.
- Identify and interpret information presented as a time-series.
- Calculate the mean, median, mode, quartiles, percentiles, range and interquartile range from vertical line charts, stem and leaf diagrams, cumulative frequency curves and histograms.
- Appreciate the use of vertical line charts to represent probability distributions.

TEACHING POINTS

Recap the skills taught at GCSE Higher Tier (9-1), and encourage the use of correct terminology.

It is advisable that basic questions involving extracting data from the statistical diagram and providing a conclusion in the context of the data are practised and encouraged.

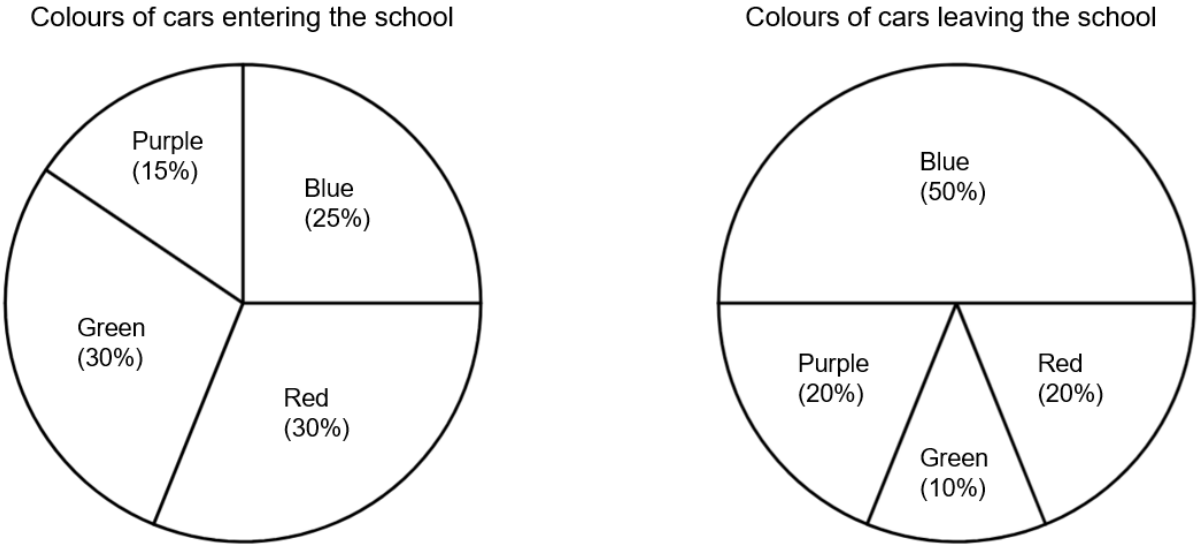
Students are not required to construct any of these diagrams, only interpret them. However, encourage students to practise constructing these diagrams using appropriate software (and definitely not by hand), which will help them appreciate the SEC.

The rationale for this is the availability of technology. Real-world statistics is not performed through manual construction of statistical diagrams, and software is used extensively – emphasise this to students. Use the Desmos activities as either a teaching aid or as a student-led activity.

Qualitative Data

An ideal way to approach bar charts and pie charts is through comparisons of two data sets. The skills of extracting data can then be combined with contextual reasoning.

Exemplar



The data above show a survey taken on a Monday morning recording the colour of cars that enter the school, and a survey taken on a Friday evening recording the colour of cars that leave the school.

In total, 80 cars were recorded entering the school grounds on Monday morning.

a) **How many green cars entered the school grounds on Monday morning?**
30% of 80 is 24, so there were 24 green cars which entered the school grounds on Monday morning.

It is claimed that more blue cars left the school on Friday evening than entered on Monday morning.

b) **Comment on this claim.**
50% of cars leaving the school grounds on Friday evening were blue.
This is a higher proportion than those arriving on Monday morning.
However, we do not know how many cars left the school grounds on Friday evening.
So we cannot claim that more blue cars left the school on Friday evening than on Monday morning.

Discrete Data

Converting between bar charts, frequency tables and stem and leaf diagrams is one way of linking the topics together; finding measures of central tendency (or spread, if Unit 1b has been taught), together with a comparison between two or more data sets will help consolidate reasoning skills.

Box and whisker plots (also called box plots) can be used for discrete data, and outliers will be represented on a box plot by a cross (×) or asterisk (*) or dot (·). The "whisker" terminals may either be the highest/lowest non-outlier values (if known) or the Tukey fences (outlier boundaries) (only to be seen here if Unit 1c has been taught first).

Continuous data

Finding percentiles (including quartiles) from cumulative frequency curves (for continuous data, the
m^{th} percentile is found by finding m% of n, the sample size). Cumulative frequency curves together with grouped frequency tables allow calculations of the estimated mean (and standard deviation, if Unit 1b has been taught) as well as percentiles, interquartile range and range. Again, comparisons of two or more data sets will help reasoning skills.

Box plots can be used for continuous data. Properties of a box plot are detailed earlier (in discrete data).

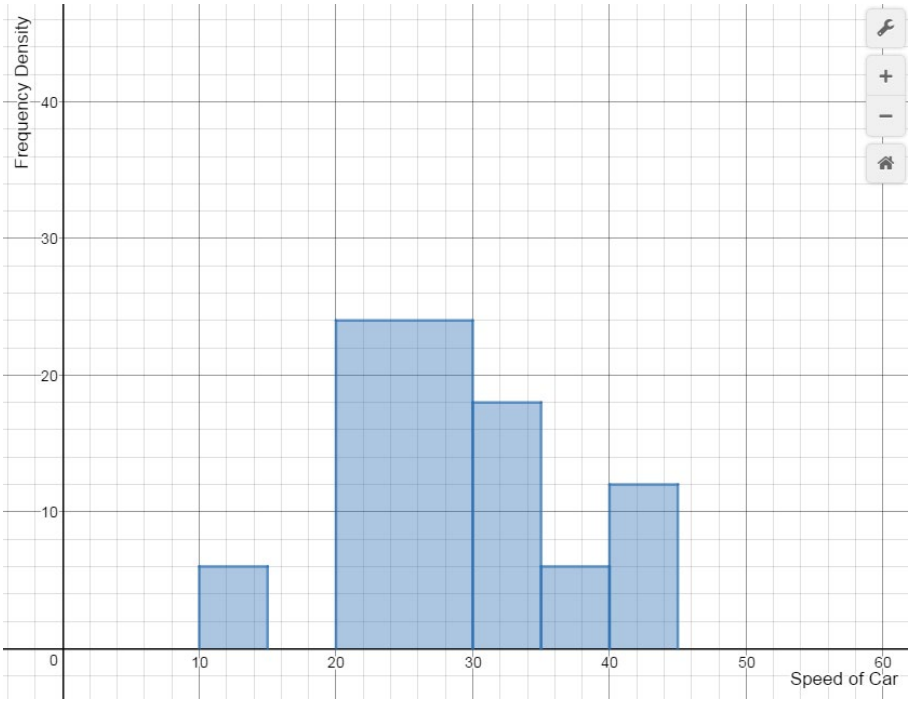
Histograms are traditionally one of the harder topics at GCSE. Students must be aware that the area of each bar represents the frequency. The height of each bar represents the frequency density. An area scale will have a meaningful interpretation, unlike frequency density.

Intermediate questions can introduce a missing scale for frequency density with contextualised information in the question allowing students to calculate the scale, or converting from a histogram back to a grouped frequency table (thus allowing calculation of the mean/standard deviation).

Extension for more able students (not in the specification)

Harder histograms questions can involve determining the median.

Extension



Key: 1 small square represents 4 cars.

A policeman records the speed of the traffic on a busy road with a 30 mph speed limit.
He records the speeds of a sample of 450 cars. The histogram represents the results.
Find the median speed of car.

(Source: Edexcel GCE Mathematics/Statistics 1 (6683/01), June 2012)

By adding the squares, the number of cars represented by each bar are 30 cars, 240 cars, 90 cars, 30 cars and 60 cars respectively.
The total number of cars is 450, so half of this is 225 cars.
This puts the median in the second bar.
After removing the number cars in the first bar (195 cars), we have (median – 20) × 24 = 195. So by rearranging the equation, the estimate of the median is 28.125 mph.

Time Series

Students will be required to comment on trend lines. To support this, students need to be familiar with identifying the direction of trend lines (upward/downward) and shape of trend lines (linear/non-linear). Particular shapes of trend lines (e.g. quadratic, logarithmic, exponential) will not be expected. Students also need to be familiar with variation (how the data are distributed around the trend line), specifically seasonal (repeating) and random variation.

Discrete Probability Distributions

These can be represented by a bar chart, and links between relative frequency and probability could be made.

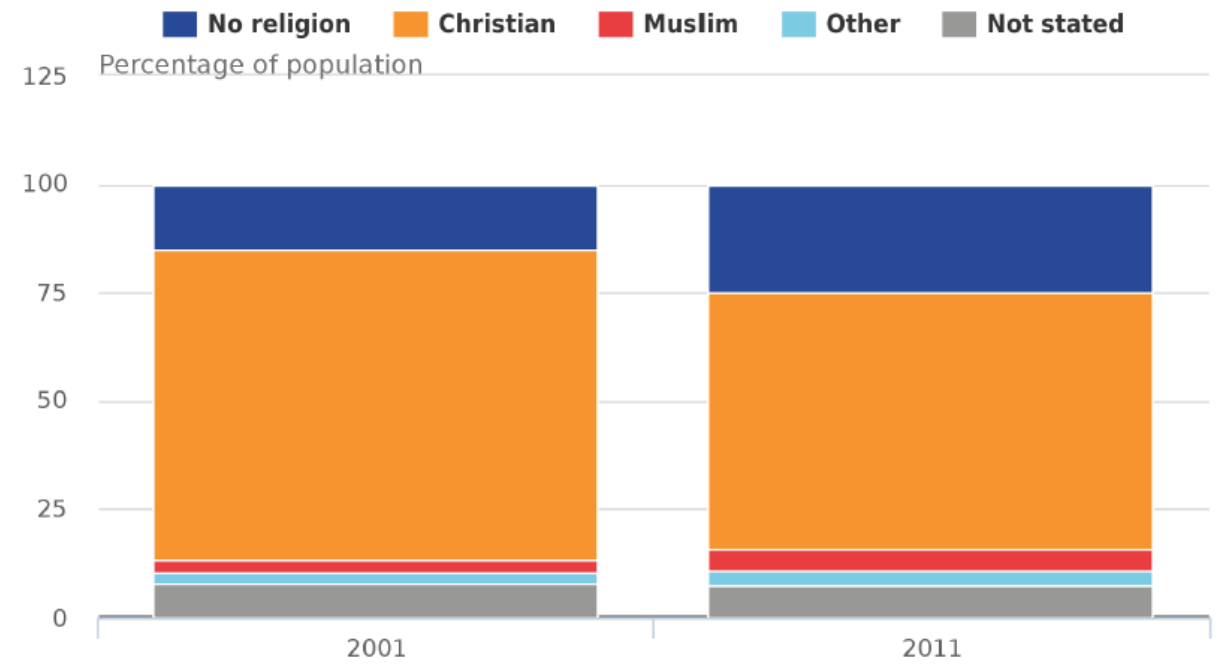
OPPORTUNITIES FOR EMBEDDING THE SEC

D1 Interpreting statistical diagrams in context.

D5 Drawing conclusions from calculated numerical measures, or making contextual comparisons between data sets, using language appropriate to a given target audience; providing students with instructions about presenting their answer in the form appropriate to a specified target audience; identifying the target audience from the statistical diagram or context of the question.

Exemplar

Figure 3: Change in religious affiliation, 2001-2011, England and Wales



Source: Census - Office for National Statistics

Source of diagram: Office for National Statistics - Religion in England and Wales, 2011

It is claimed that there are more Muslim people in the UK in 2011 than there were in 2001.
Give two reasons why this information may not be correct.
The chart does not state the total number of the population between the two years.
The percentage of the population is greater in 2011 than in 2001.
However, we do not know the population sizes so cannot say for certain that the total number is greater.
Another reason is that the chart is only for people in England and Wales, but the claim is for the whole UK (including Scotland and Northern Ireland.)

COMMON AND POSSIBLE MISTAKES

- Assuming a graph displaying percentages/proportions (e.g. pie charts) refer to the frequencies (and vice versa);
- misreading scales on (e.g.) time series;
- interpreting the height of a histogram bar as the frequency as opposed to the area;
- not reading the context surrounding the data visualisation.

NOTES

This section covers a wide range of statistical univariate diagrams. It is recommended that this topic is regularly reviewed throughout the two years since this is a fundamental skill in the statistical enquiry cycle.

2c. Data Representation and Interpretation: The Use of Software (Appendix 4)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Read and interpret data presented in a spreadsheet or database.
- Understand and use the terms: filter, sort, query, table, list, link, order, field, record, join, group, cell.
- Describe methods for extracting specific information from a spreadsheet or database using the correct terms described above.

TEACHING POINTS

This topic is new to A Level Statistics, and indeed most new qualifications in mathematics. There is a growing shift towards the use of technology in mathematics and statistics, due to its increased use in real-world modelling and applied mathematics. It is recommended this topic is divided into one hour for spreadsheets and one hour for databases.

Students need to be able to identify the purposes of the spreadsheet.

Spreadsheets organise data in a grid of **cells**. The columns are generally labelled A, B, C etc. and the rows are generally labelled numerically. Each cell has a reference e.g. A1 = column A, row 1. Spreadsheets are user friendly and visually intuitive, but all data is accessed when the spreadsheet is opened (this can lead to an inefficient use of time or a limited amount of space). To interpret the data in a spreadsheet, the data need to be correctly labelled to facilitate this. The recommended calculators also have a spreadsheet mode. Allow students to get familiar with this mode, including inputting simple formulae relating to different cells.

Data can be **sort**ed into an order (ascending/descending if numerical, alphabetical/reverse alphabetical if alphanumeric). Data can also be **filter**ed where a condition on data is imposed and only data which satisfies this condition will be displayed. Standard arithmetic on cells can be conducted and formulae must begin with "=". Symbols representing addition, subtraction, multiplication and division are +, -, * and / respectively. Some standard formulae must be known; examples include (but not limited to) =AVERAGE() (finds the mean of the cells specified), =COUNT() (counts the number of cells specified containing numbers), =STDEV.S() (finds the sample standard deviation of the cells specified). Detailed functions like =VLOOKUP are useful, but not essential, to know.

Students must be aware of standard computer functions such as copy and paste in a spreadsheet. Data may be copied into new spreadsheets from other spreadsheets to, for example, compare or clean data.

Exemplar

	A	B	C
1	**Name**	**Age**	**Vegetable**
2	Ali	32	Peas
3	Rich	33	Peas
4	James	32	Sweetcorn
5	Elen	28	Carrots
6	Dilip	32	Broccoli
7	Chi-Wai	32	Kale

A sample of people, with their names and favourite vegetables, was recorded in a spreadsheet. The first 7 records in the list are shown above.

a) **Explain how you would obtain a list of people aged 32, in alphabetical order.**
Filter *the Age field to match the criteria of "Age = 32", and then* ***sort*** *the Name field into alphabetical order.*

b) **The final row with data in this spreadsheet is row 156. Explain how the spreadsheet can calculate the average age of people recorded in this spreadsheet.**
Use a ***formula*** *to calculate the* ***average*** *of* ***cells B2 to B156****.*
OR
=AVERAGE(B2:B156)

Databases organise data in **tables** of **fields** and **records**. A **field** is a trait/characteristic to be recorded. A **record** is a collection of field values. Databases are more technical but can compartmentalise data so not all data is accessed when the database is opened. This means more data can be stored in a database than in a spreadsheet.

To extract information from a database, a **query** must be produced. A query can sort, filter, group or join tables together. The idea of **sort** and **filter** is the same as with spreadsheets. A **group** will return a list of field values with no duplication (this is useful if you are using a database constructed by others and it is unknown how data are recorded). A **join** (known in computer science as an *inner-join*) will join two tables together by using a common field as a reference to match records together.

Students need to be able to identify fields and records. They must be aware that a query which either selects, sorts, filters or groups data or joins tables of data together must be produced for databases. The database language SQL (Structured Query Language) is taught in most computer science courses and may be taught here. Students producing an answer to a database question in SQL would have access to full marks (in line with the mark scheme). Database questions may be displayed as a schema (a description of how the database is structured).

Exemplar

Patient data are recorded in a database. The first 6 records of this database are shown below.

Patient_ID	First_Name	Surname	Age	Blood_Type	BMI
0135	Yusef	Almeida	34	O+	27
0152	Yu Xin	Lim	45	A-	42
0182	Alexia	Yurchenko	19	A+	32
0211	Aoife	O'Clare	66	AB+	29
0244	Bethan	Rogers	34	O+	35
0256	Babatunde	Obolo	39	B-	35

Explain how you would use the database to:

a) **obtain a list of patients over the age of 25,**
*Produce a **query** which **filters** the **Age** field to **over 25**.*
OR (e.g. in SQL)
SELECT First_Name, Surname FROM Database
WHERE Age > 25;

b) **obtain a list of patients with a BMI under 40 in alphabetical order by first name,**
*Produce a **query** which **filters** the **BMI** field to **under 40** and **sorts First_name** in **alphabetical order***
OR (e.g. in SQL)
SELECT First_Name, Surname FROM Database
WHERE BMI < 40
ORDER BY First_Name ASC;

c) **obtain a list of all possible blood types with no duplication.**
*Produce a **query** which **groups** by **Blood_type***
OR (e.g. in SQL)
SELECT Blood_type FROM Database
GROUP BY Blood_type;

Another table of data is contained in this database containing the following fields:

Field	Data Type	Description
Patient_ID	Numerical	The unique identification number of a patient
Date_of_Birth	Numerical	The date of birth of the patient
Home_Address	String	The home address of the patient
Medication	String	The medication the patient is currently on
Medical History	String	The full medical history (medication, referrals, surgeries etc.) of the patient

d) **Explain how to obtain a list of patients names and their home addresses.**
*Produce a **query** which **joins** the two tables together using the **common field***

Patient_ID
OR (e.g. in SQL)
SELECT First_Name, Surname, Home_Address FROM Database1 INNER JOIN Database2
WHERE Database1.Patient_ID = Database2.Patient_ID;

OPPORTUNITIES FOR EMBEDDING THE SEC

The entire sub-unit is related to technology and software.

C1 Processing data, including awareness of how technology can be used to aid the statistical processes.

COMMON AND POSSIBLE MISTAKES

- Mixing up or not using the correct terminology relating to software;
- reading the incorrect values from the spreadsheet;
- providing a spreadsheet method in a database context or vice versa (e.g. producing a query for a spreadsheet);
- not using given field names in the question (e.g. “Blood Type” instead of “Blood_type”).

NOTES

There is an end-of-topic test on technology and software on the Maths Emporium (www.mathsemporium.com) which you can find here.

2d. Data Representation and Interpretation: Misrepresentation and the Suitability of Statistical Diagrams (1.2) (1.3) (1.8)	Teaching time 1 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Select suitable diagrams to represent qualitative, discrete quantitative and continuous quantitative variables.
- Identify ambiguity and misleading scales in published univariate statistical diagrams.
- Appreciate how a statistical diagram can misrepresent data and suggest ways to remedy this.

TEACHING POINTS

It is important that students are aware of the most appropriate statistical diagram for the random variable in question. Unit 2 reaches a crescendo by uniting the previous sub-units together: in Unit 2a, students define and identify the type of variable from a contextual scenario or statistical diagram. In Unit 2b, students are exposed to more statistical diagrams and consolidation of these types of random variable. Here, students are to put the two together and determine the suitability of each diagram.

Students must know that:

- Bar charts and pie charts are the most appropriate diagrams to represent qualitative data (if two categories have similar frequencies and one wishes to highlight that one is larger than the other, a bar chart would be more appropriate).
- Vertical line charts **(assumed knowledge from GCSE),** stem and leaf diagrams are the most appropriate diagrams to represent discrete quantitative data. Bar charts and pie charts may be used if the quantitative data are treated as qualitative (e.g. age groups as opposed to ages)
- Vertical line charts, bar charts or histograms may be used to represent a discrete probability distribution.
- Cumulative frequency curves and histograms are the most appropriate diagrams to represent continuous quantitative data.
- Time series are the most appropriate diagrams to represent quantitative data indexed by discrete time points.
- Box and whisker plots can be used for any quantitative data.

There are plenty of studies of published statistical diagrams showing incomplete, unlabelled, misleading or otherwise unsuitable diagrams that you could show students. Examples could include, but are not limited to, bar charts with unlabelled logarithmic scales, pie charts with incorrectly calculated angles, time series with no labels or scale, histograms where the height represents the frequency (and the class width is not labelled as 1).

Students will need to give examples as evidence in the context of the question.

Exemplar

The diagram in Figure 1 is taken from HM Treasury's 'National Infrastructure Plan 2013'. It shows the value of future projects (projects 'in the pipeline') according to sector.

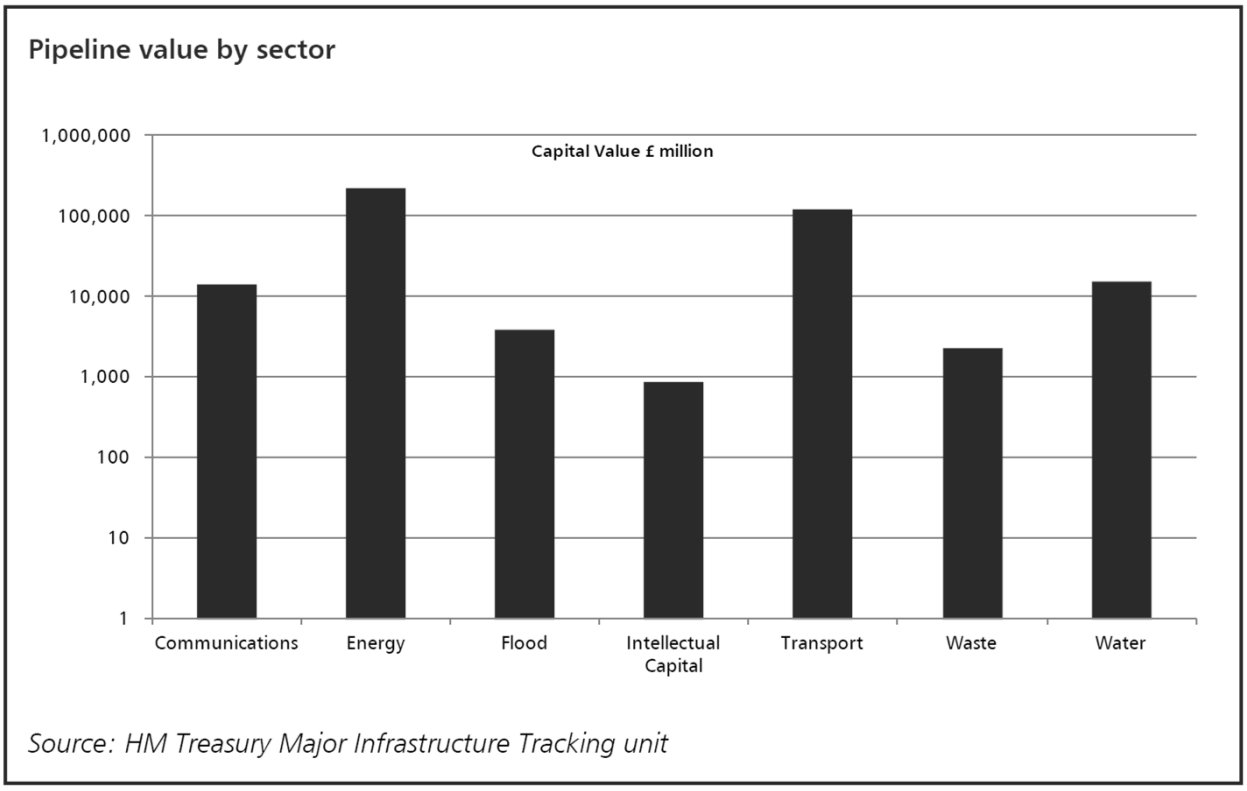


Figure 1

Explain, in the context of the question, how this may be misleading to a reader.

(Source: Sample Assessment Material)

The scale on the bar chart is not a linear scale. 1, 10,100,1000 etc
This gives the impression that Intellectual Capital has a pipeline value of just below £1,000 million
This gives the impression also that Waste has a pipeline value of just over £1,000 million.
In reality there could be a difference of £1,000 million between the two projects.

OPPORTUNITIES FOR EMBEDDING THE SEC

A5 Identifying appropriate statistical diagrams together with justification from a contextual scenario.

C2 Given published statistical diagrams, extracting information and applying it to the context in which the data were collected.

C3 Identifying misrepresentation.

E1 Identifying inappropriate statistical diagrams for the data type given.

E3 Suggesting correct alternatives statistical diagrams.

COMMON AND POSSIBLE MISTAKES

Students may be presented with published diagrams and be expected to comment on them. If the question asks to “comment on” the data, then answers are expected to relate to the data being represented on the graph. If the question asks to “criticise” the graph, then answers are expected to give (usually negative) criticisms of the graph and its intended use. A common mistake is students “criticise” when they should “comment” and vice versa.

A question may specify a number of points to make in a response (e.g. “make three comments”). Students are expected to adhere to this limit. If a student provides more than the requested number of points then they may not achieve full marks (in line with the mark scheme) if any of the points are incorrect.

UNIT 3: Probability Theory

Return to overview

SPECIFICATION REFERENCES

2.1 Know and use language and symbols associated with set theory in the context of probability.

2.2 Represent and interpret probabilities using tree diagrams, Venn diagrams and two-way tables.

2.3 Calculate and compare probabilities: single, independent, mutually exclusive and conditional probabilities.

2.4 Use and apply the laws of probability to include conditional probability.

2.5 Determine if two events are statistically independent.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics

The basic concepts of probability, the use of tree diagrams, Venn diagrams, two way tables and set theory notation (Unit 0b).

GCSE (9-1) in Mathematics at Higher Tier

P1 Record, describe and analyse the frequency of outcomes of probability experiments using tables and frequency tables.

P2 Apply ideas of randomness, fairness and equally likely events to calculate expected outcomes of multiple future experiments.

P3 Relate relative expected frequencies to theoretical probability, using appropriate language and the 0-1 probability scale.

P4 Apply the property that the probabilities of an exhaustive set of outcomes sum to one; apply the property that the probabilities of an exhaustive set of mutually exclusive events sum to one.

P6 Use tables, grids, Venn diagrams and tree diagrams.

P7 Construct theoretical sample spaces for single and combined experiments with equally likely outcomes and use these to calculate theoretical probabilities.

P8 Calculate the probability of independent and dependent combined events, including using tree diagrams and other representations, and know the underlying assumptions.

S2 Interpret frequency tables.

KEYWORDS

addition law, at least, at most, biased, complement, conditional, event, exceeds, experiment, frequency, given, independent, intersection, mutually exclusive, outcome, possibility, probability, random, relative frequency, sample space, tree diagram, two-way table, unbiased, union, Venn diagram,

UNIT SUMMARY

This unit builds on and develops the notions of probability seen at GCSE. This unit is very similar to the probability unit in the A level Mathematics with the focus primarily on numerical contexts and their interpretation, rather than algebraic exercises.

This unit is extended further in Unit 17 and the idea of independent events is a recurring theme throughout the course. Conditional probabilities can arise in virtually any unit where probabilities are to be calculated. They are especially focussed on in Unit 20.

Activities in Desmos that may help: Vertical Line Charts.

3a. Probability Theory: Concepts in Probability (2.1) (2.2) (2.4)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Know and understand the terms: experiment, outcome, event, sample space, complement, mutually exclusive, independent.
- Represent probabilities using a grid, two-way table, Venn diagram or probability tree.
- Identify the complement of an event in context.
- Apply the addition rule for probability: $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.

TEACHING POINTS

During this sub-unit, introduce students to, and encourage them to use, the correct terminology for probability:

- An experiment is a process which yields an outcome.
- An event is a set of outcomes.
- A sample space is the set of all possible outcomes.
- The complement of an event, A, is the set of outcomes not contained in the event (this is denoted by A′).
- Two events are mutually exclusive if they share no common outcome.
- Two events are independent if the occurrence of one event does not affect the probability of the other occurring.

Matching activities, tarsia puzzles or exercises to practise the terminology in context are useful activities here.

Revise two-way tables, Venn diagrams and probability trees. Also introduce the idea of a grid to represent a sample space, for example the possible outcomes of rolling two standard six-sided dice. Small examples are best (larger examples tend to be more time-consuming than useful).

The probability of the complement of the event is “1 – the probability of an event”. Although most students will find this concept simple, many will find it difficult to apply it in context. Start with easy examples:

Exemplars

Given that a coin lands on either heads or tails, what is the complement of the event "The coin lands on heads"?
The coin lands on tails

Given that the upmost facing number on a die can take the values 1, 2, 3, 4, 5 or 6, what is the complement of the event "the die shows an even number"?
The die shows an odd number

One example, which is not that easy for students to identify without guidance:

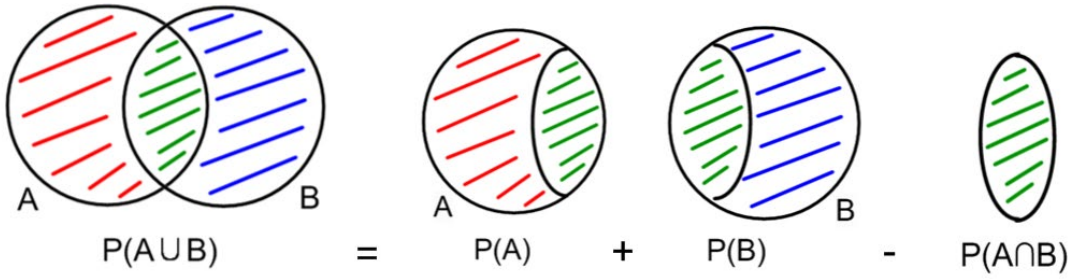
Exemplar

I have to catch a sequence of trains to reach my destination. What is the complement of the event "I do not catch any of my trains on-time" ?

I catch at least one of my trains on-time. (see mistakes below).

Students may be asked questions involving probability trees where they must find the probability of "at least one".
Students who calculate this probability constructively (adding up all possibilities) will have a harder time than those who identify that the complement of "at least one" is "none", and hence $\mathrm{P}(\text{at least one}) = 1 - \mathrm{P}(\text{none})$.

The addition rule/law of probability can be explained using a Venn diagram or a two-way table. It is advisable for students to see this justification, although it is not examinable.



Plenty of examples without context could be attempted before introducing context.

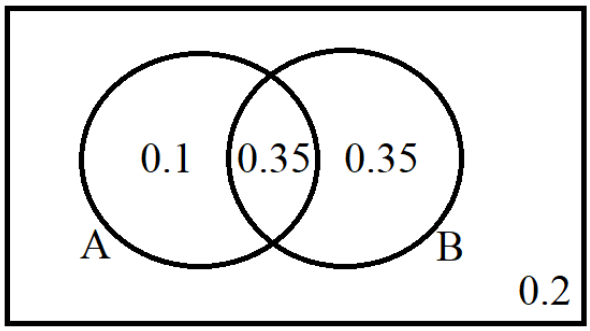
A valid alternative to using the addition rule would be to use a two-way table.

Exemplar

$\mathrm{P}(A) = 0.45$, $\mathrm{P}(B) = 0.7$ and $\mathrm{P}(A \cup B) = 0.8$. Draw a Venn diagram to represent these probabilities and calculate $\mathrm{P}(A \cap B')$.

Method 1: *Using the addition rule*

$0.8 = 0.45 + 0.7 - \mathrm{P}(A \cap B)$, *so* $\mathrm{P}(A \cap B) = 0.45 + 0.7 - 0.8 = 0.35$



So $\mathrm{P}(A \cap B') = 0.1$.

Method 2*: Using a two-way table*

Since $\mathrm{P}(A \cup B) = 0.8$ *then* $\mathrm{P}(A' \cap B') = 0.2$

	A	*A'*	*Total*
B			*0.7*
B'		*0.2*	
Total	*0.45*		*1*

The table can now be filled:

	A	*A'*	*Total*
B	*0.35*	*0.35*	*0.7*
B'	*0.1*	*0.2*	*0.3*
Total	*0.45*	*0.55*	*1*

So $\mathrm{P}(A \cap B') = 0.1$.

Gradually introduce context:

Exemplar

Based on previous experience, Oli passes 7 out of 10 maths papers and passes half of biology papers.
There is a 80% chance of his passing one of the exams. Find the probability he will pass both exams.
Let B be the event "Oli passes a biology paper" and let M be the event "Oli passes a maths paper". From the question, $\mathrm{P}(B) = 0.5$ *and* $\mathrm{P}(M) = 0.7$ *and* $\mathrm{P}(B \cup M) = 0.8$.

Method 1: Using the addition rule,

$$\mathrm{P}(B \cup M) = \mathrm{P}(B) + \mathrm{P}(M) - \mathrm{P}(B \cap M),$$

we have $0.8 = 0.5 + 0.7 - \mathrm{P}(B \cap M)$, *so* $\mathrm{P}(B \cap M) = 0.5 + 0.7 - 0.8 = 0.4$.

Method 2: Using a two-way table,
$\mathrm{P}(B \cup M) = 0.8$ *so* $\mathrm{P}(B' \cap M') = 0.2$

	B	*B'*	*Total*
M	*0.4*	*0.3*	*0.7*
M'	*0.1*	*0.2*	*0.3*
Total	*0.5*	*0.5*	*1*

So the probability Oli will pass both exams is 0.4.

OPPORTUNITIES FOR EMBEDDING THE SEC

D2 Questions involving calculating probabilities, and referring to these probabilities when investigating claims made in the question.

Exemplar

A popular street food vendor claims that half of her customers are male under the age of 45. Market research on similar stalls shows that the probability of a male customer buying food is 0.7 and the probability of a customer under the age of 45 buying food is 0.6. The probability that a person who isn't male and is over the age of 45 buys food from the stall is 0.15.
Find the probability of a male customer under the age of 45 buying food from the stall, and comment on the vendor's claim.

Let M be the event "A male customer buys food from the stall".
Let A be the event "A customer under the age of 45 buys food from the stall".
$\mathrm{P}(M' \cap A') = 0.15$ *since this is the probability of a person* ***over 45*** *who* ***isn't male*** *buying food from the stall.*
Method 1: Using the addition rule

The complement of $M' \cap A'$ is $M \cup A$, so $\mathrm{P}(M \cup A) = 0.85$.
Using the addition rule: $\mathrm{P}(M \cap A) = \mathrm{P}(M) + \mathrm{P}(A) - \mathrm{P}(M \cup A) = 0.45$.

Method 2: Using a two-way table

	M	*M'*	*Total*
A	*0.45*	*0.15*	*0.6*
A'	*0.25*	*0.15*	*0.4*
Total	*0.7*	*0.3*	*1*

So $P(M \cap A) = 0.45$

The street vendor is not completely wrong, just under half of his customers are males under the age of 45.

COMMON AND POSSIBLE MISTAKES

- Assuming the complement of the event "none" is the event "all";
- Assuming the probability of the intersection of two events is the sum of the probability of each event;
- not including the intersection of two events when calculating the probability of one event (from a Venn diagram);
- mixing up intersections and unions.

3b. Probability Theory: Conditional Probability (2.3) (2.4)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand that the probabilities of events can change if another event occurs.
- Use and interpret the correct notation for conditional probability: $\mathrm{P}(A|B)$.
- Find conditional probabilities from two-way tables, Venn diagrams and probability trees.
- Apply the multiplication law for probability: $\mathrm{P}(B)\mathrm{P}(A|B) = \mathrm{P}(A \cap B)$.

TEACHING POINTS

Students must be aware that events can affect other events, and so the probability of these events can change.
For example, the probability that I take my umbrella to work is 0.5, but if it is raining when I leave home, the probability that I take my umbrella to work is 0.95.

The notation used is $\mathrm{P}(A|B)$ meaning "the probability the event A occurs, given the event B has already occurred". Two-way tables are a good starting point for introducing the idea of conditional probability:

Exemplar

The table below shows the number of students studying for a science degree course. The data are stratified by gender and the type of science degree.

	Physics	Chemistry	Biology	Total
Male	45	64	78	**187**
Female	38	57	64	**159**
Total	**83**	**121**	**142**	**346**

a) What is the probability that a randomly chosen person, studying the science degree course, is a female student studying chemistry?

$\frac{57}{346}$

b) Given that a chemistry student has been randomly chosen from those studying the science degree course, what is the probability that the student is female?

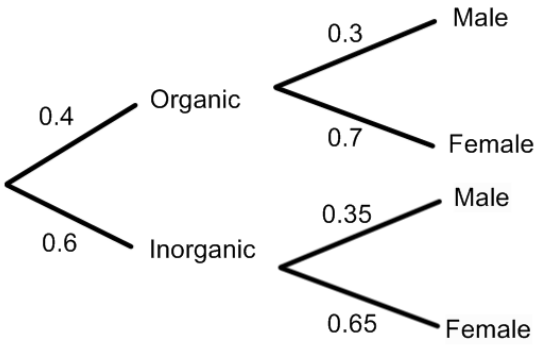
$\frac{57}{121}$

After two-way tables, applying the notation of $\mathrm{P}(A|B)$ to the second branches of a two-stage probability tree will help students understand the notion behind conditional probability, as well as allowing them to use probability trees to solve conditional probability questions.

Exemplar

Within the chemistry department, students either study organic or inorganic chemistry.
The probability that a student studies organic chemistry is 0.4.
The probability that a student studying organic chemistry is female is 0.7.
The probability that a student studying inorganic chemistry is male is 0.35.

Draw a tree diagram to represent these probabilities and hence find the probability that a randomly chosen student is female.



The probability that a randomly chosen student is female is $0.4 \times 0.7 + 0.6 \times 0.65 = 0.67$.

Venn diagrams with two events can be represented on a 2×2 two-way table, allowing students to make the link between them. Then the Venn diagrams can be generalised to three-events once the two-event case has been mastered.

The multiplication law for probability can be explained using a two-way table or a probability tree. Conditional probabilities can also be found using diagrams by restricting the sample space.
Either method could gain full marks in an exam (in line with the mark scheme).

Exemplar

The probability of testing positive for a rare disease is 0.06. The probability of having the disease and the test being correct is 0.4. Given that Joe has tested positive for the disease, what is the probability he has the disease?

Let D be the event "tests positive for the disease" and let H be the event "has the disease".

So $\mathrm{P}(D) = 0.06$ *and* $\mathrm{P}(\mathrm{H} \cap \mathrm{D}) = 0.04$.

Method 1: Using the multiplication law
Hence $\mathrm{P}(H|D) = \frac{\mathrm{P}(H \cap D)}{\mathrm{P}(D)} = \frac{0.04}{0.06} = \frac{2}{3}$.

Method 2: Using a two-way table

	D	*D'*	*Total*
H	***0.04***		
H'	***0.02***		
Total	***0.06***	*0.94*	*1*

$P(H|D) = \frac{0.04}{0.06} = \frac{2}{3}$ *(out of D, what is the probability of H)*

Harder questions can involve probability trees where the probabilities of Event A followed by Event B are known, and the student is asked to find the probability of Event B given Event A – this is Bayes' Theorem, which will be seen later in the course. However, it is a good extension exercise provided enough scaffolding is given in the question. The removal of scaffolding could be done in Unit 17.

OPPORTUNITIES FOR EMBEDDING THE SEC

C2 Using appropriate diagrams to represent conditional probability (tree diagrams, two-way tables).

D1 Interpreting conditional probabilities in context (including the correct interpretation of "given" in context).

D2 Using conditional probabilities as supporting evidence to support or refute a claim.

D5 Reaching conclusions using language appropriate for a given target audience.

COMMON AND POSSIBLE MISTAKES

- Interpreting $\mathrm{P}(A|B)$ as $\mathrm{P}(A \cap B)$ - this is down to language and comprehension.
- Using $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$ for any two events, not just independent ones.
- Interpreting $\mathrm{P}(A|B)$ as $\mathrm{P}(B|A)$.

3c. Probability Theory: Mutually Exclusive and Independent Events (2.3) (2.5)

Teaching time
1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand that if events A and B are mutually exclusive, then $\mathrm{P}(A \cap B) = 0$.
- Understand that if events A and B are statistically independent, then $\mathrm{P}(A|B) = \mathrm{P}(A)$ and $\mathrm{P}(B|A) = \mathrm{P}(B)$.
- Use the modified addition formula for mutually exclusive events: $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$.
- Use the modified multiplication formula for independent events: $\mathrm{P}(A \cap B) = \mathrm{P}(A)(B)$.
- Identify mutually exclusive events.
- Identify statistically independent events.

TEACHING POINTS

Define the terms “mutually exclusive” and “independent” and translate these into formulae. Venn diagrams are useful to explain the link for mutually exclusive events. Tree diagrams are useful to explain independent events.

A simple algebraic exercise in modifying the addition and multiplication rule for probability could be shown, either by example or an exploration activity for the more mathematically able.

It is suggested that the identification of mutually exclusive and independent events is carried out with justification referencing appropriate formulae.

Exemplar

From an orchestra of 36 musicians, a musician is picked at random to present a bunch of flowers to the conductor. The orchestra is made up as follows:

	Male	**Female**
Wind Section	10	8
String Section	8	7
Percussion Section	2	1

Let M be the event “the chosen musician is male”, W be the event “the chosen musician is in the wind section” and S be the event “the chosen musician is in the string section”.

Which two of M, W and S are statistically independent?
Which two are mutually exclusive?

$\mathrm{P}(M) = \frac{20}{36}$, $\mathrm{P}(W) = \frac{18}{36}$ *and* $\mathrm{P}(S) = \frac{15}{36}$

$\mathrm{P}(M \cap W) = \frac{10}{36}$, $\mathrm{P}(M \cap S) = \frac{8}{36}$, $\mathrm{P}(W \cap S) = 0$

$\mathrm{P}(M)\mathrm{P}(W) = \frac{10}{36} = \mathrm{P}(M \cap W)$ *so* M *and* W *are statistically independent.*

[Alt: $\mathrm{P}(M|W) = \frac{10}{18} = \frac{20}{36} = \mathrm{P}(M)$ *so* M *and* W *are statistically independent.*]

$\mathrm{P}(W \cap S) = 0$ *so* W *and* S *are mutually exclusive*

[Alt: $\mathrm{P}(W \cup S) = \frac{18+15}{36} = \frac{33}{36}$ *and* $\mathrm{P}(W) + \mathrm{P}(S) = \frac{18}{36} + \frac{15}{36} = \frac{33}{36}$ *so* W *and* S *are mutually exclusive.*]

OPPORTUNITIES FOR EMBEDDING THE SEC

B1 With examples such as the above, assessment of key terms such as "random" in the context of the question will help reinforce the idea of unbiased sampling methods.

D1 Interpreting probabilities in order to determine independence or mutual exclusivity of events.

COMMON AND POSSIBLE MISTAKES

- Confusing the terms mutually exclusive and independent;
- using the formulae $\mathrm{P}(A) + \mathrm{P}(B) = \mathrm{P}(A \cap B)$ and $\mathrm{P}(A)\mathrm{P}(B) = \mathrm{P}(A \cup B)$;
- (As stated in the notes of Unit 3b) using the formulae $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$ for any two events, not just independent ones.

The biggest mistake usually made by students is a failure to refer to appropriate calculations and formulae when identifying mutually exclusive or independent events.

UNIT 4: Discrete Random Variables

Return to overview

SPECIFICATION REFERENCES

4.2 Calculate probabilities and determine expected values, variances and standard deviations for discrete distributions.

4.3 Use discrete random variables to model real-world situations.

4.5 Interpret graphical representations or tabulated probabilities of characteristic discrete random variables.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics

Numerical Measures (Unit 1)

Basic probability (Unit 0b)

Vertical line charts, bar charts and histograms (Unit 2b)

GCSE (9-1) in Mathematics at Higher Tier

A2 Substitute numerical values into formulae and expressions.

P1 Record, describe and analyse the frequency of outcomes of probability experiments using tables and frequency tables.

P2 Apply ideas of randomness, fairness and equally likely events to calculate expected outcomes of multiple future experiments.

P3 Relate relative expected frequencies to theoretical probability, using appropriate language and the 0-1 probability scale.

P4 Apply the property that the probabilities of an exhaustive set of outcomes sum to one.

P7 Construct theoretical sample spaces for single and combined experiments with equally likely outcomes and use these to calculate theoretical probabilities.

S2 Interpret frequency tables and vertical line charts.

KEYWORDS

at least, at most, discrete, discrete random variable, exceeds, expectation, frequency, mean, probability, random, random variable, relative frequency, standard deviation, sum, variable, variance,

UNIT SUMMARY

This unit introduces students to the idea of probability distributions at a basic level. Discrete random variables are familiar to students who have studied GCSE mathematics (e.g. values of the scores on a die or spinner), although not referred to by name.

It is advised that this unit is covered after Unit 1, because the ideas of variance and standard deviation are a fundamental topic within this unit. The concepts of probability and the interpretation of vertical line charts to represent the probability distributions of discrete random variables need only be of GCSE level, but this unit may be more accessible to those who have seen Units 2 and 3 prior to this unit.

Depending on your teaching philosophy (either general to specific, or specific to general), this unit could be taught before or after the Binomial distribution, since it is a special case of a discrete random variable. This scheme of work places the unit before the Binomial distribution due to author preference.

The notions of expectation and variance of a random variable will be seen throughout all probability distributions covered in this course. The notation $\mathrm{P}(X = x)$ and associated inequalities will be introduced in this chapter and continually used when calculating probabilities from both discrete and continuous probability distributions. Calculators should be used wherever possible.

Activities on Desmos that may be useful: Vertical Line Charts.

4a. Discrete Random Variables: Introduction to Random Variables (4.5)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand and use the notation $\mathrm{P}(X = x)$, $\mathrm{P}(X < x)$, $\mathrm{P}(X > x)$, $\mathrm{P}(a < X < b)$ and the corresponding weak inequalities, $\mathrm{P}(X \leq x)$, $\mathrm{P}(X \geq x)$, $\mathrm{P}(a \leq X \leq b)$
- Calculate the probability of an event given a tabulated probability distribution.
- Identify a discrete random variable and its associated probabilities from a variety of contexts.
- Recognise a discrete uniform distribution.

TEACHING POINTS

It is suggested that students start by seeing the tabulated form of a discrete random variable with the associated probabilities. These will have been seen at GCSE, although not by name. An appropriate context will help students identify the concepts with the situations when a tabulated discrete random variable would be used. Students must appreciate that a tabulated probability distribution of a discrete random variable is the theoretical analogue of a frequency table.

GCSE questions on the topic usually ask students to determine the missing value from a table, using the property that the sum of probabilities of all possible outcomes equals 1.

The notation will be new to students. Plenty of practice involving the different notations, including the differences between $\mathrm{P}(X < x)$ and $\mathrm{P}(X \leq x)$ is advised. Encourage students to show their working, including listing the values they are using in their sum. Tarsia activities offering practice at linking words to inequalities may help.

When introducing context, it is suggested that students practise identifying keywords in the question and writing down the mathematical meaning using the correct notation; for example “at least” means ≥, “at most” means ≤ and “exceeds” means >. Students must define their variables (or subscripts) at the start of each answer: “*Let X be…*” – this is good practice and identifies the dependent variable from the context of the question. Students may not be awarded full marks (in line with the mark scheme) if variables or subscripts are not clearly defined. Remind students of finding complementary events and that $\mathrm{P}(A') = 1 - \mathrm{P}(A)$.

Students will also need to identify from a context when the outcomes of a discrete random variable are all equally likely. Although the Discrete Uniform Distribution by name and notation is not on the specification, there is no reason why students cannot be examined on it as a tabulated probability distribution. By formalising the concept by name, students may be able to identify the properties of such a distribution more easily.

OPPORTUNITIES FOR EMBEDDING THE SEC

D1 Interpreting probabilities in the context of the question.

D2 Using interpretation of their probabilities as evidence to support or refute a claim.

Exemplar

The manager of a factory has set up a new production process and is monitoring the proportion of defective items produced. The number of defective items, X, in a sample of size 5, can be modelled by the following distribution:

x	0	1	2	3	4	5
$\mathrm{P}(X = x)$	0.66	0.25	0.05	0.02	0.01	0.01

a) **Find the probability that there is a defective item in the sample.**
$\mathrm{P}(X \geq 1) = 1 - \mathrm{P}(X = 0) = 1 - 0.66 = 0.34$
Alternatively: $\mathrm{P}(X \geq 1) = 0.25 + 0.05 + 0.02 + 0.01 + 0.01 = 0.34$

It was established over many years that under the old process the probability that an item was faulty was 0.1, independently of whether other items were defective. If *Y* is the number of defective items in a sample of 5, under the old process, the probability distribution is:

y	0	1	2	3	4	5
$\mathrm{P}(Y = y)$	0.59	0.33	0.07	0.01	0	0

b) **Find the probability that there was a defective item in the sample.**
$\mathrm{P}(Y \geq 1) = 1 - \mathrm{P}(Y = 0) = 1 - 0.59 = 0.41$
Alternatively: $\mathrm{P}(Y \geq 1) = 0.33 + 0.07 + 0.01 + 0 + 0 = 0.41$

c) **Compare the results of your calculations in (a) and (b). Comment briefly on whether the new process seems to be an improvement, referring to relevant calculations.**
The probability of having a defective item in the sample is larger under the old process. This suggests that the new process produces fewer defective items. However, there is a higher risk of having 4 or more defective items in the sample under the new process.

COMMON AND POSSIBLE MISTAKES

Students find the notation $\mathrm{P}(X = x)$ and associated inequalities difficult to grasp initially. Translating the mathematical symbols into English sentences is the best way for them to understand this notation: "*The probability that the random variable X takes the value x*". Introducing context will also help consolidate this concept, as well as develop their comprehension. For example:

"*Let X be the number of Sikhs employed by a pharmaceutical company.*
$\mathrm{P}(X \geq 7)$ *is the probability that the number of Sikhs employed by a pharmaceutical company is at least 7"*

NOTES

The equivalent unit in the GCE AS and A level Mathematics course puts a greater emphasis on mathematics. For example, it will be common to see questions on discrete random variables with two missing probabilities and enough information given in the question in order to calculate these probabilities through simultaneous equations.

4b. Discrete Random Variables: Expectation, Variance and Linear Scaling (4.2)	Teaching time 2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Appreciate the concept of expectation/expected value.
- Find the expectation, variance and standard deviation of a discrete random variable given a tabulated probability distribution.
- Find the expectation, variance and standard deviation of a linearly scaled discrete random variable (that could arise in a context)

TEACHING POINTS

The expected value, $\mathrm{E}(X)$, can be considered the theoretical mean of a probability distribution. Using the idea of calculating the mean from a frequency table, the formula for the expectation of a discrete random variable can be derived in a similar fashion. This formula is considered "assumed knowledge" and must be remembered. Students may be asked to "show that" E(*X*) is a particular value and will be expected to use this formula.

The expectation and variance may be obtained by using the statistics mode on the calculator.

Encourage students to input the tabulated probability distribution onto the calculator, using the frequency column for the probabilities. In these circumstances, emphasise the following to students:

- The statistics s_x and s_x^2 will not be defined on the calculator, and an error message will be generated if they are requested.
- The variance, Var (*X*), and standard deviation will be given by σ_x^2 and σ_x respectively.
- On a calculator, the mean μ or E(*X*) will be given by $\bar{x}$.

Practice at calculating the expectation could be done first, allowing students to revise and develop their skills at substituting values into a formula. Once this has been mastered, the variance and standard deviation practice can follow. The skills portion of the sub-unit could finish with calculating both from a tabulated probability distribution. Check all answers using the statistics mode on the calculator. Unless summary statistics are given in a question, students could use the calculator when determining the expectation and variance from a tabulated probability distribution.

When introducing context, the correct terminology should be used as well as phrases such as "the expected value".

It is advisable to revise linear scaling from Unit 1.

If $y = ax + b$ then $\bar{y} = a\bar{x} + b$ and $s_y^2 = a^2 s_x^2$ or $s_y = as_x$.

Students must appreciate that the same will be true for the theoretical equivalents:
If $Y = aX + b$ then $\mathrm{E}(Y) = a\mathrm{E}(X) + b$ and $\mathrm{Var}(Y) = a^2\mathrm{Var}(Y)$.

Alternatively, students may "linearly scale" the *x* values in a probability distribution table and use the calculator to find the mean and variance of the linearly scaled variable.

OPPORTUNITIES FOR EMBEDDING THE SEC

D1 Interpreting the expected value and standard deviation in context.

D2 Reaching conclusions based on interpretation of these numerical measures, and using these conclusions as a basis of comparison or evidence for refutation.

Exemplar

The head coach of a wheelchair basketball team training for the next Paralympic games wants to analyse the team's previous performances. The number of points scored per game over many four-year training cycles is recorded as a discrete random variable. In order to have a chance at qualifying, the team need to consistently average at least 8 points per game. After a statistician is consulted, the calculated expected value is found to be 9.4 and the standard deviation is 3.1. The coach thinks that his team will qualify for the next Paralympic games. Comment on the coach's claim.

The team scored an average of 9.4 points per game in previous years, but the standard deviation of 3.1 indicates they could have scored as low as 6.3 points per game. Although they averaged over 8 points per game, it is not likely that they consistently scored this. The coach may therefore be incorrect.

COMMON AND POSSIBLE MISTAKES

- Forgetting to square-root the variance to obtain the standard deviation.
- Using $\mathrm{Var}(aX + b) = a\mathrm{Var}(X) + b$ or $a\mathrm{Var}(X)$ instead of $a^2Var(X)$.

NOTES

As in the previous sub-unit, the use of simultaneous equations to answer questions on discrete random variables will not be assessed on this qualification, but can be used as an extension exercise for the more mathematically able.

UNIT 5: The Binomial Distribution

Return to overview

SPECIFICATION REFERENCES

5.1 Know when a binomial model is appropriate (in real world situations including modelling assumptions).

5.2 Know methods to evaluate or read probabilities using formula and tables.

5.3 Calculate and interpret the mean and variance.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics

Discrete Random Variables (Unit 4, or GCSE list below)

Probability (Unit 3, or P8 below)

GCSE (9-1) in Mathematics at Higher Tier

A2 Substitute numerical values into formulae and expressions.

P1 Record, describe and analyse the frequency of outcomes of probability experiments using tables and frequency tables.

P2 Apply ideas of randomness, fairness and equally likely events to calculate expected outcomes of multiple future experiments.

P3 Relate relative expected frequencies to theoretical probability, using appropriate language and the 0-1 probability scale.

P4 Apply the property that the probabilities of an exhaustive set of outcomes sum to one.

P7 Construct theoretical sample spaces for single and combined experiments with equally likely outcomes and use these to calculate theoretical probabilities.

P8 Calculate the probability of independent combined events.

S2 Interpret frequency tables and vertical line charts.

KEYWORDS

at least, at most, Bernoulli, binary, binomial, conditions, discrete, discrete random variable, distribution, event, exceeds, expectation, failure, independent, mean, probability, outcome, random, random variable, standard deviation, success, trial, variable, variance,

UNIT SUMMARY

This topic is the only discrete probability distribution referred to by name in Year 1 of the A level course and follows directly on from Unit 4. However, it is possible to teach this topic prior to the general topic on discrete random variables. Expectation and variance

will need to be taught in relation to the binomial distribution at some point (Unit 4b). It is placed after the general concepts of discrete random variables in this scheme of work due to author preference.

Bernoulli trials are a fundamental concept in probability and statistics and, although not referred to by name, form the basis of the concepts behind a binomial distribution. Contextual situations which can be modelled by a binomial distribution should be seen and justified throughout. It is also the first time in this scheme of work where the formal notation of a probability distribution is used, namely $\mathrm{B}(n, p)$. It is highly important that students are defining their variables correctly, and thus using the notation $\sim$ to mean "has the distribution".

This topic has links with the normal distribution (Unit 7, 9c) and the Poisson distribution (Unit 15). The binomial distribution will also be used in hypothesis testing for a proportion (Unit 11b), the Sign test (Unit 13a, 14b) and the difference between two proportions (Unit 21c).

Unlike in previous specifications, the table of probabilities from a binomial distribution (which are provided in the formula book) is, in fact, obsolete. It is now a requirement of the specification that every student uses a calculator which can calculate binomial probabilities. This shortens the teaching time of this topic from its legacy counterparts.

As a historical footnote, the origins of the Swiss mathematician Jacob Bernoulli (1655-1705), whose name is lent to the trials used in a binomial distribution, can be discussed.

Activities on Desmos that will be helpful for this unit: Binomial Distribution.

5a. The Binomial Distribution: Conditions for a Binomial Distribution (5.1)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Recognise when to use the binomial distribution and appreciate situations where the binomial distribution may be an appropriate model
- State any assumptions necessary in order to use the binomial distribution
- Understand and use the notation $\sim \mathrm{B}(n, p)$ where n is the number of trials and p is the probability of success

TEACHING POINTS

The binomial distribution is a special case of a discrete random variable. The conditions that students must know are:

- There are exactly two outcomes to each trial: “success” and “failure”.
- The trials are independent of each other.
- The number of trials in the experiment is fixed.
- The probability of success in a trial is constant.

A random variable following a binomial distribution is defined as the number of “successes”. One method of remembering these conditions is using TINS: Two [outcomes], Independent [trials], Number [of trials is fixed], Same [probability of success]. Alternative methods involve flashcards, continual identification through contextual scenarios, or songs.

A basic, initial example to allow students to meet the binomial distribution is: **A fair coin is tossed 5 times and will either land on heads or tails**. This allows students to relate the conditions of a binomial distribution to the scenario. It is very important that students are in the habit of defining their variables, since it is important how the students have defined “success” in the context of the situation. For example: *Let X be the number of times the coin lands on heads*.

Identifying the number of trials and the probability of success should then be expressed as: $X \sim \mathrm{B}(5,0.5)$.

When introducing context, it is suggested that students are given a wide range of examples to allow them to appreciate how often a binomial distribution can model a real-world situation. Similarly, it is advisable that contexts which may not be modelled by a binomial distribution are seen.

Exemplar

A students' union wants to take a survey about how effective a university's financial assistance policy is. This particular university claims that 9% of their students are from families who earn less than £15,000 a year (low-income students). The union begins by conducting exploratory data analysis on a random sample of 10 chemistry students and asking whether they are low-income students.
Comment on the suitability of a binomial distribution in this case.

- *A binomial distribution may be suitable here with each "trial" a person in the sample and "success" is that person is a low-income student.*
- *The number of people asked is fixed at 10*
- *The probability that a student classifies as "low-income" is 0.09. (provided the university's claim is correct across all subjects)*
- *The probability that one chemistry student in the sample is a low-income student does not affect the probability that another student in the sample is also a low-income student, so the trials are independent.*
- *Finally, a student is either a low-income student or not, so there are exactly two outcomes for each trial.*

Exemplar

At airport security, the home office is conducting a survey of the ethnicity of British people passing through border control. Previous research has indicated that 23% of people passing through border control identify as British Asian. The home office decide to take a census of all the people passing through border control on a particular day, and recording whether they are British Asian.
Comment on whether a binomial distribution is suitable in this case.

- *A binomial distribution would **not** be suitable here.*
- *Each trial is "a person passing through border control". A success is "person is British Asian".*
- *The probability of each trial may be constant (if previous research is correct) and there are exactly two outcomes.*
- *However, the number of trials is **not fixed** since it is not known how many people pass through border control on that particular day.*
- *Also, it is likely that family groups may pass through border control: children may have the same ethnicity as their parents so it is likely that trials may not be independent from one another*.

OPPORTUNITIES FOR EMBEDDING THE SEC

A1 Identifying the factors in a scenario which are important in order to be able to identify a binomial distribution.

A3 Identifying what data to collect, and identifying "success" and "failure".

A4 Appreciating the importance of exploratory data analysis as a starting point prior to a full investigation into a problem (as seen in the first example above).

D2 Considering the context of the scenario when concluding whether or not a binomial distribution is appropriate.

E1 Identifying when a binomial distribution is not appropriate can lead to a discussion of how the data collection method can be altered in order to fit a binomial distribution. For example, the second example can be amended such that *a random sample of 100 people was taken* – this would address the issues of the number of trials and independence.

COMMON AND POSSIBLE MISTAKES

- Students often find it difficult to contextualise independence, or identify whether two events can be independent.
- Students are often reluctant to state uncertainty in their answers e.g. "the events might not be independent" – stating uncertainty is important and it demonstrates an appreciation of the limitations of statistics.

A solution presenting information in clear bullet points is advised.

5b. The Binomial Distribution: Finding Probabilities from a Binomial Distribution (5.2)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Appreciate the derivation of the formula $\mathrm{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$.
- Use their calculator to find $\mathrm{P}(X = x)$.
- Use their calculator to find $\mathrm{P}(X \leq x)$.
- Calculate other probabilities e.g. $\mathrm{P}(X \geq x)$, $\mathrm{P}(a < X \leq b)$ etc.
- Use trial and error / the "Inverse Binomial" mode on a calculator to find values given a probability.

TEACHING POINTS

Continuing from the previous sub-unit with the example: **A fair coin is tossed 3 times, and will either land on heads or tails**. In order to appreciate where these probabilities come from, the links between probability trees and tabulated probability distributions can be seen.

Let X be the number of times the coin lands on heads. Then $X \sim \mathrm{B}(3,0.5)$.

Listing the possible combinations of heads (0 to 3), and using probability rules for independent events will allow students to appreciate that $\mathrm{P}(X = x) = \binom{n}{x} p^x (1-p)^x$ where *p* is the probability of success and $\binom{n}{x}$, sometimes written ${}^{n}\mathrm{C}_{x}$, is the total number of combinations of achieving *x* successes from *n*. The formula for $\binom{n}{x}$ is not needed or assessed. However, the underlying concept could be discussed.

Once all the probabilities ($x = 0,1,2,3$) have been calculated, they could be displayed in a tabulated probability distribution. Then the concepts of $\mathrm{P}(X < x)$, $\mathrm{P}(X > x)$ and the corresponding weak inequalities, $\mathrm{P}(X \leq x)$, $\mathrm{P}(X \geq x)$, can be revised and related to a binomial distribution context.

The biggest change in the specification is the use of technology: it is now a requirement of the specification that students use a calculator which has the capacity to calculate binomial probabilities. The recommended calculator can calculate $\mathrm{P}(X = x)$ and $\mathrm{P}(X \leq x)$, given inputs of *n* and *p*. Giving students plenty of practice with their calculators will allow them to familiarise themselves with potentially new technology, or an unfamiliar mode on their calculator. These calculators can also generate the tabulated probability distributions, and can be used to check their answers.

These recommended calculators may not calculate other binomial probabilities other than the two described above, so conventional methods for calculating other

probabilities still need to be taught as before (note more advanced graphical calculators will evaluate other probabilities). For example:

$$\mathrm{P}(X < x) = \mathrm{P}(X \leq x - 1)$$

$$\mathrm{P}(X > x) = 1 - \mathrm{P}(X \leq x)$$

$$\mathrm{P}(a \leq x \leq b) = \mathrm{P}(x \leq b) - \mathrm{P}(x < a)$$

The use of a visual number line can be incredibly helpful here.

Questions without context could be seen first:

Exemplar

X is a random variable and $X \sim \mathrm{B}(13, 0.23)$. Find $\mathrm{P}(X > 4)$.

Using a number line from 0 to 13:

0	*1*	*…*	*3*	*4*	*5*	*…*	*13*

$\mathrm{P}(X > 4) = 1 - \mathrm{P}(X \leq 4) = 1 - 0.8414 = 0.1586.$

Calculating two-sided inequalities is done in much the same way, but emphasise that some calculators always gives $\mathrm{P}(X \leq x)$, and students should know the difference between $\mathrm{P}(X < x)$ and $\mathrm{P}(X \leq x)$. Using the calculator to generate the fully tabulated probability distribution, and then summing the appropriate values is an alternative option, albeit highly inefficient.

When introducing context, it is helpful to use genuine scenarios where a binomial distribution can be used for modelling. It is highly advisable that words and phrases such as “at most” and “exceeds” are used – these will have been seen in Unit 4. It is important that students have good literacy skills in order to deal with questions such as these.

Students may also be required to find the number of successes required to satisfy a probability (Inverse Binomial). Some calculators have an “inverse binomial” mode but care should be taken whether the value produced by the calculator is the correct value. Trial and error is a valid method for Inverse Binomial.

Exemplar

It is known that 20% of components are faulty. During a quality control check, a random sample of 30 components are taken and the number of defective components counted. The batch will be accepted if at most *k* faulty components are found, otherwise the batch is rejected. The quality control officer wants the probability of finding at most *k* faulty components to be less than 12%. Find the value of *k*.

Let X be the number of faulty components.

Since a component is either faulty or not, there are two possible outcomes, making binomial distribution a possibility. We assume that the sample is random and that the components are independent of each other. So, since there is a fixed number of 30 components and the chance a component is faulty is 20%, X~B(30,0.2).

We want P(X ≤ k) < 12%

0	..	*k-1*	*k*	*k+1*	…	*30*
		< 0.12				

Using trial and error on the calculator:
P(X ≤ 2) = 0.0442 < 0.12

P(X≤ 3) = 0.123 > 0.12

So k = 2

OPPORTUNITIES FOR EMBEDDING THE SEC

C1 Using the calculators in order to calculate binomial probabilities, and binomial cumulative probabilities. The tables of binomial probabilities could be seen by all students – this will allow them to appreciate the usefulness of technology in the world of statistics.

C2 Vertical line charts can be used to demonstrate a binomial distribution.

D1 Calculating probabilities and interpreting them in the context of the question.

D2 Relating the calculated probabilities to a particular claim made in the question.

COMMON AND POSSIBLE MISTAKES

Many mistakes occur during the reading of contextual questions, e.g. at least 3 is interpreted as > 3. Misinterpreting $\mathrm{P}(X < a)$ as $\mathrm{P}(X \leq a)$. Calculating $\mathrm{P}(a \leq X \leq b) = \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a)$.

On most modern calculators, there are two modes: Binomial PD and Binomial CD. Binomial PD is used to calculate $\mathrm{P}(X = a)$. Binomial CD is used to calculate $\mathrm{P}(X \leq a)$. Students may use the incorrect binomial mode on the calculator.

Previous mistakes have been to do with the use of the formula and calculating the binomial coefficient. Due to the binomial mode on the calculator, these mistakes are unlikely to occur.

NOTES

An extension activity in the form of determining the binomial probabilities using the formula can be used for the more mathematically able.

5c. The Binomial Distribution: Mean and Variance (5.3)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand the mean and variance of $\mathrm{B}(n,p)$ are np and $np(1-p)$ respectively.
- Use the mean and variance to assess the suitability of a binomial distribution.

TEACHING POINTS

Revise the concepts of expectation/mean, variance and standard deviation of a discrete probability distribution.

The mean of a binomial distribution *np* is easily explained through the context of a binomial distribution: **I toss a coin 100 times, and there is a 0.5 probability of the coin landing on heads. What is the expected number of heads?**

or

6% of employees at a company have identified as having suffered with mental illness at some point in their life. Out of a sample of 200 employees, what is the expected number of people who have suffered with mental illness?

The variance $np(1-p)$ is harder to explain through the context of the question. However, the result is a direct application of the formula for the variance and can be explained as such – the proof of both the mean and variance of a binomial distribution will not be assessed.

Questions without context could be seen first: $\boldsymbol{X \sim \mathrm{B}(34, 0.23)}$**, find the mean and standard deviation of *X*.**

Questions on mean, variance and standard deviation can be applied to questions from Unit 5b: **What is the expected number of low-income students from the sample?**

The unit could finish with comments on the suitability of a binomial distribution, referencing the mean and variance of a sample.

Exemplar

At airport security, the home office is conducting a survey of the ethnicity of British people passing through border control.
Previous research has indicated that 23% of people passing through border control identify as British Asian.
The home office decide to take a random sample of 200 from all of the people who has passed through border control during a period of 2 weeks, and record whether they are British Asian or not.
The mean number of British Asians passing through passport control was 46.1, with a standard deviation of 10.2.
Comment on the suitability of the binomial distribution in this case.

Let X be the number of British Asians passing through border control in a sample of 200.
Then $X \sim \mathrm{B}(200,0.23)$*.*
If the binomial model were suitable, then $\mathrm{E}(X) = 200 \times 0.23 = 46$ and $\mathrm{Var}(X) = 200 \times 0.23 \times (1 - 0.23) = 35.42$*, giving a standard deviation of 5.95.*
The binomial distribution may not be a suitable model - as although the mean value is close to the expected value, the observed standard deviation is quite a bit higher than the expected standard deviation.

OPPORTUNITIES FOR EMBEDDING THE SEC

C2 Calculating the mean and variance and applying it to the context of the question, making inferences about a population.

D1 Calculating numerical measures and interpreting them in the context of the question.

D2 Using the calculation of the mean and variance to assess the suitability of the binomial model.

COMMON AND POSSIBLE MISTAKES

Forgetting to square-root the variance in order to find the standard deviation. This is especially common in contextual questions where the objective is to compare the expected variance/standard deviation with an observed one.

NOTES

Students will not be expected to calculate parameters of a binomial distribution given the theoretical mean and variance.

UNIT 6: Bivariate Data

Return to overview

SPECIFICATION REFERENCES

1.1 Interpret statistical diagrams including … scatter diagrams.

1.6 Identify outliers by inspection and using appropriate calculations.

1.7 Determine the nature of outliers in reference to the population and original data collection process.

4.1 Know and use terms for variables: random, discrete, continuous, dependent and independent.

7.1 Calculate (only using appropriate technology – calculator) and interpret association using Spearman's rank correlation coefficient or Pearson's product moment correlation coefficient.

7.4 Calculate (only using appropriate technology – calculator) and interpret the coefficients for a least squares regression line in context; interpolation and extrapolation, and use of residuals to evaluate the model and identify outliers.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics

Numerical Measures (Unit 1)

GCSE (9-1) in Mathematics at Higher Tier

R14 interpret the gradient of a straight line graph as a rate of change.

S6 use and interpret scatter graphs of bivariate data; recognise correlation and know that it does not indicate causation; draw estimated lines of best fit; make predictions; interpolate and extrapolate apparent trends while knowing the dangers of so doing.

KEYWORDS

bivariate, coefficient, correlation, domain of validity, extrapolation, gradient, interpolation, least squares regression line, linear scaling, mean, outlier, product moment correlation coefficient, rank, regression, residual, scatter, Spearman's rank correlation coefficient,

UNIT SUMMARY

This unit introduces students to the idea of bivariate data. Although not referred to by name, they will have seen this at GCSE. The concepts of correlation and lines of best fit are extended and given a more formalised tone in this unit. Although requiring no other pre-requisites than those skills learnt at GCSE, it has been left until this point so students can master skills with univariate data before moving on to bivariate data.

This unit begins formalising correlation, giving a numerical value to describe the strength of the correlation. Pearson's PMCC and Spearman's rank correlation coefficient are used to describe the correlation, and students must contextualise these values and relate them to the context of the question. It is essential that students appreciate that correlation is not causation, and it is advisable that plenty of examples are seen to demonstrate this idea. At this stage, the idea of a bivariate normal distribution need not be addressed but students will need to be aware of it by the time they reach hypothesis testing about correlation in Unit 10b.

The least squares regression line is a formalised version of the line of best fit. Encourage predicting using both the equation of the regression line and a graphical representation of the regression line to show the link. The use of spreadsheets to generate scatter diagrams and regression lines will allow students to appreciate the role of technology in statistics, as well as help them check their numerical answers. The Scatter Graph activity in Desmos may be of use.

Identifying outliers by inspection is further extended by the use of residuals of a scatter graph. It would be helpful to have seen outliers in Unit 1c, but is not necessary to be teaching this unit.

This topic is used when testing for the correlation or association of two variables (Unit 10).

Scenarios involving a regression line may involve two random variables or one independent and one dependent variable.

6a. Bivariate Data: Terminology (4.1)

Teaching time
1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand the definition of “bivariate”.
- Identify, and understand the difference between, independent and dependent variables.
- Identify bivariate contextual situations.
- Read and interpret a scatter diagram.
- Identify correlation from a scatter graph: positive, negative or none.
- Interpret correlation in context.
- Make predictions using a line of best fit.
- Appreciate the use of a line of best fit and the relationship between the sign of the gradient of the line and the correlation.

TEACHING POINTS

Emphasise that bivariate means two-variable. It is important for students to identify bivariate data and the independent and dependent variables in suitable contexts. A scatter diagram is the most appropriate way of graphically representing these data.

Correlation

Revise the concepts of correlation from GCSE. This includes identifying correlation by eye (e.g. strong positive, weak negative) and describing the relationship in context.

Exemplar

An environmental conservationist is trying to determine whether there is a relationship between the temperature of water and the amount of bacteria present in the water.
She takes 20 samples of lake water during different times of the day at different temperatures, and then records the amount of bacteria in the water.
Suggest a suitable diagram to represent these data, justifying your answer.
A scatter diagram is best to display these data, because the data she is collecting is bivariate/there are two variables involved. She is also attempting to investigate a relationship between the two variables.

Regression

Variables may be classified as independent (or explanatory) and dependent (response). The variable that is being changed/controlled by the data collector is the independent variable; the variable that is being recorded is the dependent variable. The independent variable is recorded on the horizontal axis. A line of best fit can be used to make predictions or estimates.

Exemplar

The environmental conservationist is now trying to recreate the lake conditions in the laboratory. Using 20 samples of lake water and heating them to different temperatures, she records the amount of bacteria in the water. Identify the independent and dependent variables in this context.
The independent variable is the temperature and the dependent variable is the amount of bacteria in the water.

It is no longer a requirement to construct scatter diagrams by hand, but students could be encouraged to explore the use of spreadsheets or graphing software to generate scatter diagrams and lines of best fit from the software. Students would benefit from revising questions using the line of best fit e.g. **Estimate the amount of bacteria in the water at a temperature of $18\,^{\circ}\mathrm{C}$** s At this stage, it is advisable that only estimations using the line of best fit within the domain of validity (within the range of the data set / interpolation) are seen.

OPPORTUNITIES FOR EMBEDDING THE SEC

A1 Identifying the variables to record in order to answer a question.

A5 Identifying bivariate data from a context and using a scatter diagram to represent these data.

B3 Examples where it is difficult to determine from a context which variable is the independent variable and which is the dependent variable could be examples of poorly planned investigations.

An environmental conservationist is carrying out an experiment in the laboratory. She records the pollution levels of the water and the acidity levels of the water.

In this example, both variables could be random, or one of the variables could be non-random. Knowledge of the data collection process is essential in identifying which is which. The following can be added to the question:

In this experiment, the environmental conservationist uses a machine to inject pollutants at specified levels into her water source and records the level of acidity of the water.

The independent and dependent variables are more obvious. These examples and the potential ambiguity will help students appreciate the importance of recording data collection methods.

C1 The use of spreadsheet software to generate scatter graphs and lines of best fit can be used to help students who are less familiar with bivariate data.

C2 Identify the correlation from a published scatter graph and suggest what this might mean in the context of the population given in the question.

D2 Appreciating that correlation is not causation – students must understand that a correlation between two variables does not mean that one variable directly affects the other. Conclusions made must reflect this understanding.

COMMON AND POSSIBLE MISTAKES

- Confusing correlation with causation;
- confusing the independent and dependent variables;
- (after regression lines are seen) supposing that the gradient of the regression line is a measure of the strength of the correlation.

6b. Bivariate Data: Correlation Coefficients (7.1)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Use the calculator to determine Pearson's product moment correlation coefficient (PMCC)
- Appreciate the requirement for a linear trend in data for the PMCC to be the relevant coefficient
- Rank a set of data, including tied ranks
- Use the calculator to determine Spearman's Rank correlation coefficient (as PMCC of ranks)
- Identify the advantages and disadvantages of each correlation coefficient
- Interpret correlation coefficients in the context of the data.

TEACHING POINTS

It is no longer a requirement to calculate the product moment correlation coefficient using the formula, and the specification mentions, explicitly, that the PMCC is to be obtained from a calculator. It is therefore important that students have familiarised themselves with the statistics mode of the calculator and are aware of the calculations the calculator can do.

For the PMCC, practise data entry from a minimal contextualised question. Students need to be aware of the notation, *r*, for the PMCC and that $-1 \leq r \leq 1$. Students need to also be aware that a magnitude of r less than 0.3 indicates a weak correlation, a magnitude of r between 0.3 and 0.7 indicates a moderate correlation, a magnitude of r between 0.7 and 0.9 indicates a strong correlation and a magnitude of r between 0.9 and 1 indicates a very strong correlation.

When using Spearman's rank correlation coefficient, use examples where it would be appropriate to use Spearman's rank (see below) as opposed to artificial situations where it would be more appropriate to use the PMCC. It is no longer a requirement to calculate Spearman's rank correlation coefficient using the formula, so students are expected to use calculator methods here. Students need to first define a ranking on their data sets, for example: *The lowest values will be assigned a rank of 1*. This is important for consistency. If there are tied ranks, then the mean of the tied ranks are assigned to all tied values.

For example: if the data set were

	150	**130**	**145**	**130**	**145**	**130**	**170**	**110**
then the rankings would be	**7**	**3**	**5.5**	**3**	**5.5**	**3**	**8**	**1**

Once the data have been ranked, the ranked data can be inputted into the calculator and *r* can be obtained via the function on the calculator. The symbol for Spearman's rank correlation coefficient is r_s.

Students would benefit from seeing contextual questions involving the PMCC or Spearman's rank. Spearman's rank should be used if the variables recorded are already ranks/scores, if the data are subjective, or if there is a non-linear relationship between the variables (as seen from a scatter diagram). Otherwise, the PMCC can be used (*testing a PMCC requires that the underlying population can be regarded as having a bivariate normal distribution - this will be seen in Unit 10b).

Encourage interpreting the value of *r* or r_s in context. Try to discourage conclusions alluding to the direct effect of one variable on other another. Also encourage the contextual meanings of the strength of the correlation, and whether the relationship could be linear or not.

Exemplar

It is claimed that there may be a linear relationship between the number of vaccinations given in a particular month and the number of diagnoses of autism three months later.
The Spearman's rank correlation coefficient is calculated at 0.2458. Comment on the claim, giving two justifications for your answer.
The claim is probably not valid – Spearman's rank correlation does not indicate that the relationship is linear, even if the Spearman's rank correlation coefficient has high magnitude.
Also in this case, the correlation coefficient indicates a weak positive correlation. This shows that, although there may be some evidence that as the number of vaccinations increases, the number of diagnoses increase, the relationship is not strong.

Students must appreciate that if the data are scaled, the correlation coefficients are unaffected. A graphical illustration using graphing software will help explain this.

OPPORTUNITIES FOR EMBEDDING THE SEC

A1 Identifying the variables under consideration.

A5 Identifying bivariate data and using a scatter diagram to represent these data.

C1 Appreciating that technology is important in calculating the correlation coefficients.

D1 Interpreting the numerical values of the correlation coefficients correctly in the context of the question.

D2 Commenting on claims of association or a linear relationship between the variables, giving numerical values to justify their answer.

COMMON AND POSSIBLE MISTAKES

- Forgetting to find the mean of ranks when there are tied ranks;
- incorrect data entry into the calculator;
- interpreting correlation as causation during conclusions;
- using the PMCC when Spearman's rank is clearly more appropriate (this mistake is due to the omission of not ranking the data. The converse of this mistake does not tend to occur);
- misreading the calculator – mainly neglecting the negative sign on a correlation coefficient (this is especially a risk if the calculator display uses a very small font).

NOTES

A lot of the teaching from previous specifications, as well as the equivalent unit in the GCE AS and A level in Mathematics is omitted from this unit. It is no longer a requirement to use the formula to calculate these correlation coefficients, unless the student is undertaking GCE A Level Further Maths.

*Conditions for PMCC to be validly evaluated:
There is a linear relationship between the two variables
Also, but more technical at this level: outliers are either kept to a minimum, there is homoscedasticity of the data

6c. Bivariate Data: The Least Squares Regression Line (7.4)

Teaching time
1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand the properties of the least squares regression line
- Use the calculator to determine the equation of the least squares regression line
- Make predictions using the equation of the least squares regression line
- Use interpolation or extrapolation arguments to comment on the reliability of these predictions
- Understand the definition of a residual
- Identify outliers as values with large residuals
- Suggest possible explanations for outliers

TEACHING POINTS

It is suggested that the least squares regression line is introduced as a "mathematical" line of best fit. The equation of the regression line is given as $y = a + bx$. The two properties of the least squares regression line can be discussed:

- The sum of the residuals is zero,
- The sum of the squares of the residuals is as small as possible,

where a residual is the directed vertical distance from a point to the line: positive if the point is above the line, negative if the point is below the line. A residual is calculated by $y - \hat{y}$ where *y* is the observed *y* value at a point *x* and $\hat{y}$ is the value of *y* obtained from the regression line at the same value of *x*.

The zero sum can be explained by saying that the line goes through $(\bar{x}, \bar{y})$, and minimising the sum of squares ensures the line is as close to each point as it can be.

It is no longer a requirement to calculate the equation of the least squares regression line using a formula, and a calculator is expected to be used to obtain the gradient (*b)* and intercept (*a*) of the regression line. Students need to know that the intercept is the estimated value of the dependent variable when the independent variable is zero. Students need to know that the gradient represents the average increase/decrease in the dependent variable for each one unit increase of the independent variable. Students must be aware that any questions involving the interpretation of the *y*-intercept and gradient **must** be in context to achieve full marks (in line with the mark scheme).

Explain to students that the least squares regression line of *y* on *x* should only be used to predict *y* at given *x* values. This is because the residuals are minimised in the *y*-direction, but not necessarily in the *x*-direction. To predict *x* at a given *y* value, a regression line of *x* on *y* should be produced (this can be obtained by swapping the datasets on the calculator).

There is no reason why context cannot be introduced as early as possible here. At this stage students should be using the equation of the regression line in order to make predictions. Students need to also identify when the estimations made are inside the domain of validity (interpolation) or outside the domain of validity (extrapolation) and comment on the reliability of each prediction in the context of the question.

Explain to students that interpolation is generally reliable and extrapolation is generally not reliable. Reliability can also be assessed by considering the sizes of the residuals (relative to the magnitude of the data). A data set yielding large residuals indicates that the regression line may not be a good fit and, conversely, a data set yielding small residuals indicates that the regression line may be a good fit. Students can also consider magnitudes of the PMCC, *r*, to indicate information about the sizes of the residuals.

Exemplar

The coach of a T35 sprinter suffering from cerebral palsy is trying to determine whether the training regime is helping his sprinter improve her sprinting time. He records the length of time training (*x* hours) under his regime for that week, against a 100 m sprint (*y* seconds) at the end of the week.
When calculated, the equation of the least squares regression line is $y = 23.1 - 0.13x$.

a) **Explain what the 23.1 and 0.13 mean in context.**
The regression line indicates that if the sprinter doesn't do any of her coach's training that week, she will run the 100 m in 23.1 seconds.
For every extra hour of her coach's training, the regression line predicts her 100 m sprint time to decrease by 0.13 seconds.

b) **Estimate the time the taken for the sprinter to run 100 m, if she trains under her coach's regime for 90 hours that week.**
If $x = 90$, then $y = 23.1 - 0.13 \times 90 = 11.4$ seconds.

c) **Comment on the reliability of your estimate in part (b).**
The estimate probably involves extrapolation, since it is highly unlikely that an athlete would train for 90 hours in a week . This means the estimate is probably unreliable.

d) **On one particular week, the sprinter trained for 8 hours in the week and completed the 100 m sprint in 14.5 seconds. Calculate the residual for this particular sprint.**
The observed sprint time is $y = 14.5$ seconds. The regression line predicts that $23.1 - 0.13 \times 8 = 12.06$ seconds. The residual is $14.5 - 12.06 = 2.44$ seconds.

Outliers is revision from GCSE (or Unit 1). The identification of outliers, given a scatter diagram and a regression line, justified by reference to the size of the residual is advised. If possible, it is advisable this residual is calculated and compared with the magnitudes of other residuals. To link it with Unit 1c, outliers can be described as any datapoint which has a residual of more than 3 standard deviations of y (to link with the idea that an outlier is more than 3 standard deviations away than what is expected).

Students could also suggest possible explanations for outliers, but must appreciate that they should not be removed from a data set unless there is a genuine reason for doing so (see Unit 1c).

OPPORTUNITIES FOR EMBEDDING THE SEC

C1 Appreciating the role of technology when calculating the regression line. This can either be through spreadsheet software (for graphical representation) or the calculator (the equation).

D1 Interpreting the values of the gradient and intercept of the regression line in the context of the question.

D2 Interpreting the predictions made using the equation of the regression line in the context of the question.

D4 Deciding whether the estimates calculated are reliable or unreliable. Students also should be able to identify outliers, suggest reasons why they might have arisen, and discuss the possible effect on the regression line.

E3 Determining whether an outlier should be removed from the data set or not.

COMMON AND POSSIBLE MISTAKES

- Mixing up the values of the gradient and intercept from calculator to equation, and from equation to interpretation;
- assuming an extrapolated estimate is reliable;
- not including units in the interpretation of the gradient of the regression line.

NOTES

If the data were linearly scaled, the equation of the regression line will be affected. This is quite algebraic and will not be tested in detail. However, simple unit conversions and their affect on the values of a and b may be assessed. This can be explained by using the interpretations of a and b.

However, a scaling for a correlation coefficient (PMCC or Spearman's Rank) may be included in context (note: scaling has no effect on these correlation coefficients)

UNIT 7: The Normal Distribution

Return to overview

SPECIFICATION REFERENCES

4.4 Know the properties of a continuous distribution.

6.1 Know the specific properties of the normal distribution and know that data from such an underlying population would approximate to having these properties, with different samples showing variation.

6.2 Apply knowledge that approximately $\frac{2}{3}$ of observations lie within $\mu \pm \sigma$, and equivalent results for $\mu \pm 2\sigma$ and $\mu \pm 3\sigma$.

6.3 Determine probabilities and unknown parameters with a normal distribution.

6.4 Apply the normal distribution to model real-world situations.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics

Mathematical Techniques (Unit 0b, or GCSE list below)

Numerical Measures (Unit 1)

GCSE (9-1) in Mathematics at Higher Tier

A2 substitute numerical values into formulae and expressions, including scientific formulae.

A5 understand and use standard mathematical formulae; rearrange formulae to change the subject.

A17 solve linear equations in one unknown algebraically.

A19 solve two simultaneous equations in two variables.

P3 relate relative expected frequencies to theoretical probability, using appropriate language and the 0-1 probability scale.

P5 understand that empirical unbiased samples tend towards theoretical probability distributions, with increasing sample size.

S4 interpret, analyse and compare the distributions of data sets from univariate empirical distributions through appropriate graphical representation involving discrete, continuous and grouped data.

KEYWORDS

bell shape, continuous, continuous random variable, mean, normal distribution, parameter, population, random variable, random, rectangular, sample, standard deviation, standard normal distribution, statistic, symmetrical, variable, variance

UNIT SUMMARY

This unit is probably the most fundamental throughout the statistics course. It is used as the basis of many statistical tests, either directly or as an assumption. It is one of the most famous continuous probability distributions, and it is important that students get plenty of practice at calculating probabilities.

It is placed at this point during the scheme of work since the groundwork of numerical measures, probability, statistical diagrams and discrete random variables have already been laid. However, only knowledge of numerical measures and basic probability is required in order to learn this unit.

Due to the shift in emphasis of technology, calculators are now used to determine probabilities from a normal distribution (and values from an inverse normal distribution). Although tables of probabilities will be provided in the formula book, they are now obsolete (see Unit 5).

Activities on Desmos that will be helpful in this unit: The Normal Distribution, Model Fitting, Histograms

7a. The Normal Distribution: Continuous Random Variables (4.4) (6.1) (6.2)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand that the probability distributions of continuous random variables can be represented on a graph.
- Appreciate this curve is always above the horizontal axis, and the total area underneath the curve is 1.
- Identify a continuous uniform distribution.
- Know the properties of the Normal distribution.
- Use the notation $\mathrm{N}(\mu, \sigma^2)$.

TEACHING POINTS

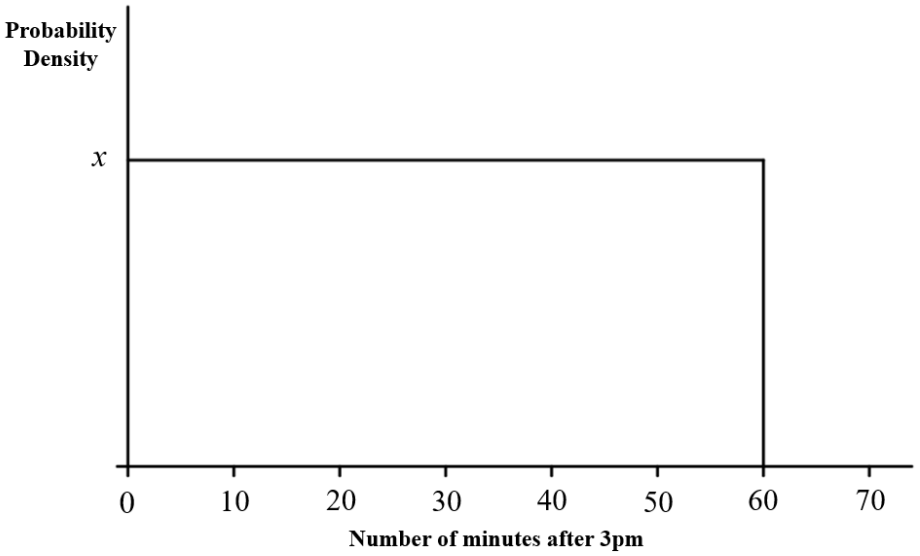
Revise the difference between discrete and continuous variables/distributions. Show the difference between the discrete uniform distribution and the continuous uniform (rectangular) distribution, by way of a diagram. This will be the students' first introduction to probability density curves, although they do not need to know them by name. What they do need to know is that the probability density curve is always above the horizontal axis, and the total area between the curve and the axis is equal to one. The continuous uniform distribution is a good way of easing students into this idea. Applying the notation of $\mathrm{P}(X < x)$ and other inequalities to the continuous uniform distribution and calculating probabilities will help prepare students for what is to come. The idea that $\mathrm{P}(X = x) = 0$ and $\mathrm{P}(X < x) = \mathrm{P}(X \leq x)$ can be explained with similar diagrams. Contrast these results with those of a discrete distribution, and explain that for a continuous random variable $\mathrm{P}(X = x) = 0$ does not necessarily mean "impossible".

Students need to be aware that the area underneath a continuous uniform distribution is 1 and that the area underneath the distribution represents the probability of a variable lying in that region. This also means that the probability density (height) of the curve is $\frac{1}{\text{width}}$. Students are also expected to understand that due to the symmetry of the rectangle shape, the mean is equal to the median and therefore the *x*-value of the midpoint of the rectangle. Students may be expected to calculate probabilities by using the diagram and calculating the appropriate areas.

Exemplar

A fire alarm is scheduled to be tested at a random point between 3pm and 4pm on a Wednesday afternoon. Matt decides to model this as a continuous uniform distribution over the interval $0 \leq X \leq 60$ where *X* is the number of minutes after 3pm.

He uses statistical software to generate a picture of the probability distribution.



a) **Write down the value of *x*.**
The width of the rectangle is 60, so $x = \frac{1}{60} = 0.0167$

b) **Write down the expected time the fire alarm will go off.**
The mean of X is in the middle, at 30 minutes. So the fire alarm is expected to go off at 3:30.

c) **Write down the probability that the fire alarm will go off at exactly 3:20.**
0

d) **Find the probability that the fire alarm will go off in the first 10 minutes.**
$P(0 \leq X \leq 10)$ *is the area of the rectangle section from 0 to 10. This has a width of 10 and a height of* $\frac{1}{60}$*. Hence* $P(0 \leq X \leq 10) = 10 \times \frac{1}{60} = \frac{1}{6} = 0.167$.

Introduce the normal distribution as another continuous random variable with the following properties:

- It is symmetrical about the mean
- It is bell shaped
- The total area underneath the curve is 1
- About $\frac{2}{3}$ of the area lies within one standard deviation of the mean, 95% of the area lies within two standard deviations of the mean, and 99.8% (as identified in the specification) of the area lies within three standard deviations of the mean.

Another use of a histogram can be used to link the normal distribution to familiar graphical representations, and as the width of each bar tends to zero, the normal distribution curve will present itself.

Students need to be aware that the normal distribution is dependent on the population mean μ and the variance σ^2, and a normal distribution can be represented with the notation $\mathrm{N}(\mu, \sigma^2)$.

The use of the calculator will help consolidate this: using the standard normal distribution, the calculator gives $\mathrm{P}(-1 \leq Z \leq 1) = 0.6827$, $\mathrm{P}(-2 \leq Z \leq 2) = 0.9545$ and $\mathrm{P}(-3 \leq Z \leq 3) = 0.9973$.

It is worth mentioning to students the anomaly on the calculator of the *NpD* button (that students will not be expected to use)

OPPORTUNITIES FOR EMBEDDING THE SEC

A1 Students could be introduced to the normal distribution using real-world examples where a normal distribution could be used.

C2 Sketching a normal distribution, clearly indicating the mean.

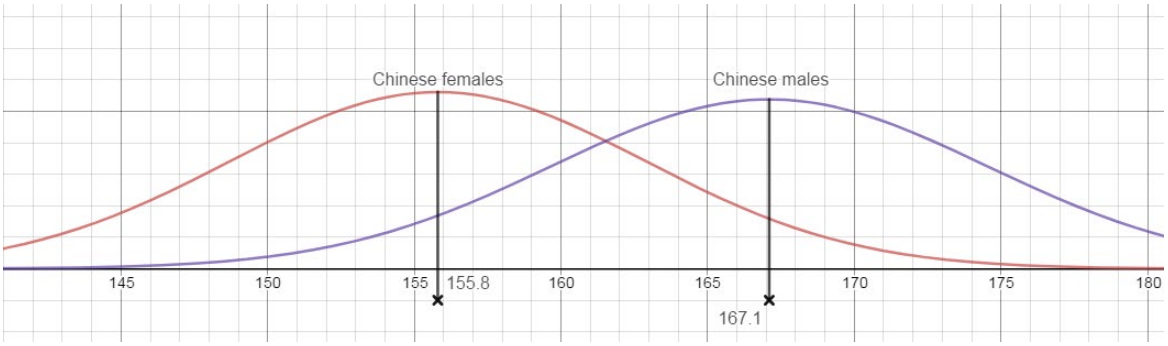
D1 Comparing the parameters of different normal distributions and represent these on sketches. They could also be able to interpret these measures in context.

Exemplar

It is known that the height of females of Chinese origin is normally distributed with a mean of
155.8 cm and a standard deviation of 7.11 cm.
It is also known that the height of males of Chinese origin is normally distributed with a mean of 167.1 cm and a standard deviation of 7.42.
(Source https://tall.life/height-percentile-calculator-age-country/)

a) **Compare the means and standard deviations of the two populations in context.**
On average, a Chinese male is 12 cm taller than a Chinese female.
The standard deviation of the heights of Chinese males is higher than that of Chinese female.
The heights of Chinese males are more varied than those of Chinese females.

b) **Use the properties of the normal distribution to sketch both normal distribution curves on the same axes.**
The mean of Chinese males is higher than that of Chinese females.
The curve for Chinese males will have a peak to the right of that of the curve for Chinese females.
The heights of Chinese males are more spread out, but the total area under the curve must total 1, so the peak for Chinese males will not be quite as high as that for Chinese females.



COMMON AND POSSIBLE MISTAKES

The percentages of areas corresponding to $(\mu - \sigma, \mu + \sigma)$ etc. are usually forgotten. However, this is now easily obtained by choosing appropriate values in the calculator.

NOTES

The use of graphing software (e.g. Desmos) can be used to demonstrate the links between histograms and the probability density function of the normal distribution. Model fitting (Distribution 3) is ideal here.

7b. The Normal Distribution: The Normal Distribution (6.3)	Teaching time 3 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Appreciate the standard normal distribution is denoted by Z and has mean 0 and variance 1.
- Use the calculator to find probabilities of a normal distribution.
- Use the calculator to find values of the inverse normal distribution $\mathrm{P}(X \leq x) = p$.
- Use symmetry to calculate values of the inverse normal distribution $\mathrm{P}(X \geq x) = p$.
- Use the inverse normal distribution to find limits for which the middle $n\%$ of observations lie.
- Interpret the normal distribution in context.

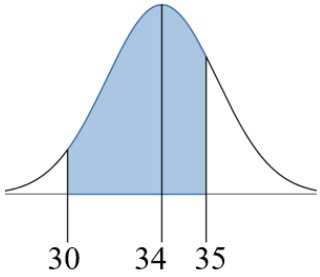
TEACHING POINTS

Define the standard normal distribution to be $Z \sim \mathrm{N}(0, 1^2)$. Emphasise that because the variance is the parameter, we must try to express it in this form. This will be good practice for later on (Unit 9). Prior to 2017, standardising a normal distribution to obtain z values was necessary due to the lack of a calculator to calculate probabilities for the normal distribution. Students are expected to utilise the calculator to find normal distribution probabilities and such questions will often be low tariff (1 or 2 marks).

Start with examples with no context, and allow students to practise using this new mode of the calculator. Calculators can now calculate probabilities from one-sided and two-sided inequalities, and when doing so encourage students to always draw a sketch, labelling the x-values and the areas of each region used. Ensure students see examples where either the variance or standard deviation is given. To calculate, for example, $\mathrm{P}(X < 45)$, input a 'theoretical' lower boundary into the calculator (say 5 or more standard deviations below the mean). A similar value is used for the reverse inequality.

Exemplar

Let $X \sim \mathrm{N}(34, 12.3^2)$. Find $\mathrm{P}(30 \leq X \leq 35)$.



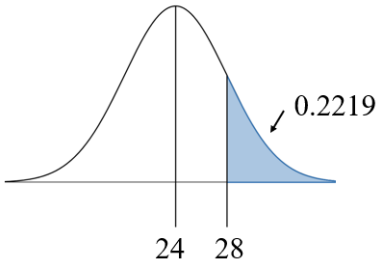
$\mathrm{P}(30 \leq X \leq 35) = 0.1599$.

Contextual questions can be then used, to allow students to appreciate when the normal distribution can be applied.

Exemplar

The plums from a particular variety of plum tree have masses which can be modelled by a normal distribution with mean 24 g and standard deviation 5 g. Plums weighing more than 28 g are graded as large.
What proportion of the plums is graded as large?

Let X be the mass of a plum. Then $X \sim \mathrm{N}(24, 5^2)$. A large plum has a mass larger than 28 g.
So $\mathrm{P}(X \geq 28) = 0.2119$.

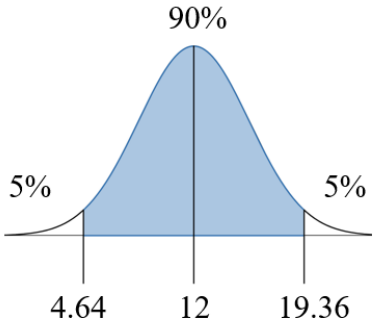


When using the inverse normal distribution, the calculator will only give cumulative areas (as in $\mathrm{P}(X \leq x) = p$ for a given p). Using the sketch, this will allow students to recognise whether they have identified the correct region. Emphasise that if $\mathrm{P}(X \geq x) = p$ is required, then $\mathrm{P}(X \leq x) = 1 - p$ (the sketch will help consolidate this concept).

By using a sketch, students can find the limits for which the middle $n\%$ of observations lie.

Exemplar

Let $X \sim \mathrm{N}(12, 20)$. Find the values which the middle 90% of the area lies.



From calculator, 5% of the area lies to the left of 4.64 and 5% of the area lies to the right (or 95% of the area lies to the left) of 19.36.
Therefore 90% of the area lies between 4.6440 and 19.36.

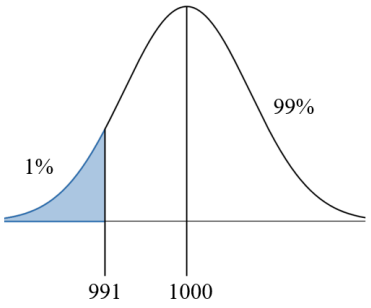
Contextual questions can be used once the calculator and basic skills have been mastered. As usual, encourage students to define their variables at the start.

Exemplar

Bags of sugar packed by a machine have masses which can be modelled by a normal distribution with mean 1000 g and standard deviation 4 g, if the machine is working correctly.

If the machine is working correctly, 1% of the bags are rejected because they are underweight. Calculate the minimum acceptable mass of a bag of sugar.

Let X be the mass of a bag of sugar. Then $X \sim \mathrm{N}(1000, 4^2)$.
Let a be the weight such that 1% of bags below this weight are rejected. So $\mathrm{P}(X \leq a) = 0.01$.



From calculator, $a = 991$ g.

It is advisable that students have plenty of practice at interpreting probabilities and x-values of a normal distribution in context. This can take time, but it is worth doing.

OPPORTUNITIES FOR EMBEDDING THE SEC

D1 Calculating probabilities or *x*-values of the normal distribution and then interpret them in the context of the question.

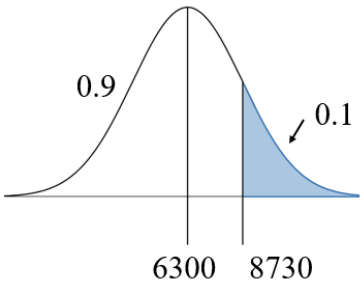
D2 Reaching conclusions or commenting on claims using the calculated probabilities or values as evidence.

D5 Presenting these conclusions in a language appropriate to a given target audience.

Exemplar

It has been found, from records taken over 80 years that the maximum flow at the time of the annual floods of a certain African river can be modelled by a normal distribution with mean 6300 m³/s and standard deviation 1900 m³/s.
The flow that will be exceeded on average once every ten years is called a 'ten year flood'.
Find the flow that will be exceeded during a ten-year flood.

Let X be the maximum flow at the time of the annual flood.
Let a be the flow that is exceeded 10% of the time.
So $X \sim \mathrm{N}(6300, 1900^2)$ *and* $\mathrm{P}(X \geq a) = 0.1$, *so* $\mathrm{P}(X \leq a) = 0.9$.



Using the calculator, $a = 8730$ m³/s.

COMMON AND POSSIBLE MISTAKES

- Not drawing a sketch – most errors occur through not drawing a sketch;
- putting a value greater than the mean on the left of the sketch (and vice versa);
- incorrectly converting between percentages and decimals (utilising the % function on the calculator can minimise this error);
- misreading the parameter in a normal distribution – if the variance is given and is inputted as the standard deviation and vice versa. Interpreting the probabilities or *x*-values of a normal distribution incorrectly;
- using the incorrect normal distribution mode on the calculator.

NOTES

This unit used to be much longer and more difficult for students. Tables of probabilities of the normal distribution are now obsolete and supplanted by the use of the calculator. The calculator not only can calculate cumulative probabilities, but any probabilities between any specified values, for any parameter specified. The inverse normal distribution can still only calculate cumulative probabilities. It is still important that students draw a sketch – this will be important in the next sub-unit.

7c. The Normal Distribution: Finding Unknown Parameters (6.3) (6.4)	Teaching time 2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Appreciate that a normal distribution can be transformed to the standard normal distribution, $Z \sim \mathrm{N}(0,1^2)$, using the transformation $z = \frac{x-\mu}{\sigma}$ or $x = \mu + z\sigma$.
- Appreciate that $\mathrm{P}(X \leq x) = \mathrm{P}\left(Z \leq \frac{x-\mu}{\sigma}\right)$.
- Find μ from a normal distribution with known variance.
- Find σ from a normal distribution with known mean.
- Interpret the findings in a variety of contexts.

TEACHING POINTS

Revise rearranging of equations (Unit 0).

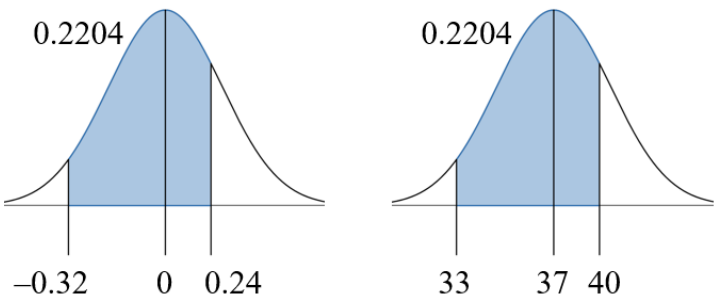
Recap the standard normal distribution $Z \sim \mathrm{N}(0,1^2)$. By drawing a suitable diagram for $X \sim \mathrm{N}(\mu, \sigma^2)$ (use specific examples if you wish), standardising can be explained by translating the graph so the mean is centred at 0 and then stretching the graph so the "spread" is correct. The Normal Distribution activity on Desmos may be of use here.

Several examples could be seen and verified by calculator that the transformation $z = \frac{x-\mu}{\sigma}$ does indeed work. The value of z can be explained as "the number of standard deviations above the mean".

Exemplar

Let $X \sim \mathrm{N}(37, 12.5^2)$.
Find the values a and b such that $\mathrm{P}(a \leq Z \leq b) = \mathrm{P}(33 \leq X \leq 40)$.

$a = \frac{33-37}{12.5} = -0.32$, $b = \frac{40-37}{12.5} = 0.24$.
Using the calculator, $\mathrm{P}(-0.32 \leq Z \leq 0.24) = 0.2204 = \mathrm{P}(33 \leq X \leq 40)$.



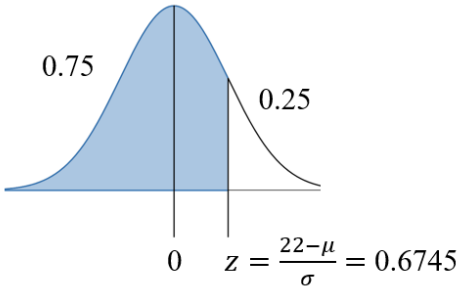
It is important that students verify both standard and non-standard probabilities using the calculator – this will help them appreciate the importance of standardising, as well as allowing calculator practice.

To find an unknown parameter of a normal distribution with a known parameter, start with questions with no context. Remember the 5 'S''s of the normal distribution: Set-up, Standardise, Sketch, Solve, Simultaneous. To begin, allow students to practise the first four.

Exemplar

Let $X \sim \mathrm{N}(\mu, 0.34^2)$ and it is known that $x = 22$ is the upper quartile.
Find μ.

Set-up: *If 22 is the upper quartile, then* $\mathrm{P}(X \leq 22) = 0.75$.
Standardise: $\mathrm{P}(X \leq 22) = \mathrm{P}\left(Z \leq \frac{22-\mu}{0.34}\right) = 0.75$ *OR* *Using* $x = \mu + z\sigma$ *gives* $22 = \mu + 0.34z$
Sketch Z~N(0.1):



Solve: The calculator gives $z = \frac{22-\mu}{0.34} = 0.6745$ *OR* $22 = \mu + 0.34 \times 0.6745$
Rearranging the equation: $\mu = 22 - 0.34(0.6745) = 21.8$ *(3 s f).*

A note about the previous example: students may use the equation solver on the calculator as an alternative to rearrangement of the question to access full marks (in line with the mark scheme). Students may also use trial and error as a valid alternative method for finding the parameter and could gain full marks (in line with the mark scheme) via this method.

When introducing students to context questions, use phrases such as "at most" or "exceeds" to reinforce literacy. As always, encourage students to define their variables at the start. It is advisable for students to practise questions finding μ and questions for finding σ.

OPPORTUNITIES FOR EMBEDDING THE SEC

D1 Calculating unknown parameters of the normal distribution and then interpret them in the context of the question.

D2 Reaching conclusions or commenting on claims using the calculated parameters as evidence.

D5 Presenting these conclusions in a language appropriate to a given target audience.

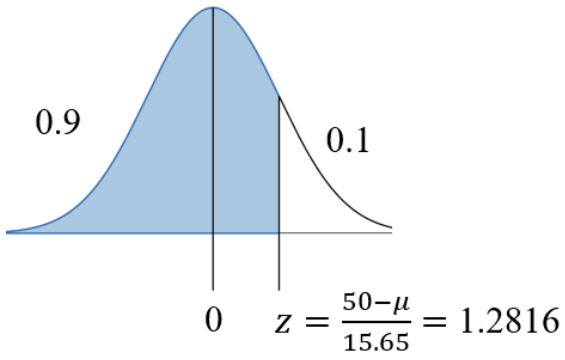
Exemplar

A company is under investigation by a national newspaper for age discrimination in the workplace. The company claims the mean age of their employees is 45.21 years old with a standard deviation of 15.65 years.
During the investigation, it was discovered that only 10% of their employees were over the age of 50.
Assuming the age of an employee is a continuous normally distributed random variable with a standard deviation of 15.65 years, comment on the company's claim, presenting your reasons as appropriate for the readership of the newspaper.

Let X be the age of an employee. Then $X \sim \mathrm{N}(\mu, 15.65^2)$.
It is known that $\mathrm{P}(X > 50) = 0.1$ *and therefore* $\mathrm{P}(X < 50) = 0.9$.

Standardising: Using $Z \sim \mathrm{N}(0,1^2)$, *we have* $\mathrm{P}\left(Z < \frac{50-\mu}{15.65}\right) = 0.9$ *(or* $50 = \mu + z \times 15.65$*)*

Sketch:



Solve: Using the calculator, $\frac{50-\mu}{15.65} = z = 1.2816$.
Rearranging, we have $50 = \mu + 1.2816 \times 15.65$. *Solving gives* $\mu = 29.94$ *years.*

The company's claim of the average age of 45.21 years old is much higher than the calculated value of 29.94 years old, suggesting the company is wrong about their mean age.

COMMON AND POSSIBLE MISTAKES

- Using any of the following incorrect standardisations: $\frac{x-\sigma}{\mu}$, $\frac{x-\mu}{\sigma^2}$, $\frac{\mu-x}{\sigma}$… (the list goes on);
- When finding the mean: rearranging the equation $x - \mu = z\sigma$ to $x - z\sigma = \mu$;
- many mistakes may occur during manual solving of a linear equation in either μ or σ. The equation solver on the calculator can minimise this. Alternatively, a method involving trial and improvement is also valid and may gain full marks (in line with the mark scheme).

NOTES

Always emphasise good practice by use of a sketch of the normal distribution. Situations where both μ and σ are unknown and simultaneous equations need to be used to find them (which may be seen on the 9MA0 A Level Mathematics Course) **will not be assessed**.

UNIT 8: Data Collection

Return to overview

SPECIFICATION REFERENCES

3.1 Know both simple (without replacement) and unrestricted (with replacement) random samples.

3.2 Know how to obtain a random sample using random numbers tables or random numbers generated on a calculator.

3.3 Evaluate the practical application of random and non-random sampling techniques: simple random, systematic, cluster, judgmental and snowball, including the use of stratification (in proportional and disproportional ratios) prior to sampling taking place.

3.4 Know the advantages and limitations of sampling methods.

3.5 Make reasoned choices with reference to the context in which the sampling is to take place. Examples include, but are not limited to: market research, exit polls, experiments and quality assurance.

3.6 Understand the practical constraints of collecting unbiased data.

4.1 Know and use terms for variables: random, discrete, continuous, dependent and independent.

PRIOR KNOWLEDGE

GCSE (9-1) in Mathematics at Higher Tier

R9 define percentage as ‘number of parts per hundred’; interpret percentages as a fraction or a decimal, and interpret these multiplicatively; express one quantity as a percentage of another.

S1 know the limitations of sampling.

KEYWORDS

cluster, data, judgemental, method, population, proportion, quota, random number, random sample, random, ratio, sample, simple, snowball, strata, stratified, unrestricted,

UNIT SUMMARY

This unit is linked closely to the SEC and forms a foundation for correct data collection. References to the SEC in this unit will be numerous, and the quality of communication should merit some teaching time.

Often, this unit has been placed at the start of the course (see, for example, the GCE AS and A Level in Mathematics – Statistics section). However, it is often the topic which students find tedious and disengage with. To allow students to gain familiarity with the more mathematical concepts and probability first, we place this unit here.

This unit will link to all future topics, especially when the SEC is embedded into the question. For example, a student could be asked to describe a method of collecting a cluster sample for the use of a contingency table prior to conducting a χ^2 test, or a student could be asked to identify whether or not the results of a hypothesis test would be valid if the method of data collection was via a judgemental sampling technique. The use of a random number generator on the calculators can be explored, together with how to use these numbers.

On the Maths Emporium (www.mathsemporium.com), you can find an end-of-topic test on this unit here.

8a. Data Collection: Terminology and Random Sampling (3.1) (3.2) (3.3) (4.1)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand the terms: random sample, simple random sample, unrestricted random sample.
- Understand how random numbers from a table or a calculator can be used in the sampling process.
- Describe a method of selecting a simple random sample in context.
- Describe a method of selecting an unrestricted random sample in context.

TEACHING POINTS

A random sample of size n has the following properties:

- Every member of the population has an equal chance of being selected for the sample,
- All subsets of the population of size n are possible,
- Every possible sample of size n has an equal chance of being selected.

[Please note that the definition of 'random sample' is not universally acknowledged in the greater statistical community, but that this is the definition used in this specification]

A simple random sample is obtained when members of the population are selected without replacement.

An unrestricted random sample is obtained when members of the population are selected with replacement.

It is advisable that students are familiar with the use of the random number generator on the calculator (the random number will be a number between 0 and 1, usually given to three decimal places). Alternatively, many calculators can generate a random number between two specified limits. There are many ways of using these random numbers, for example reading the first decimal place to obtain a one-digit random number.

To obtain a random sample (simple or unrestricted), students need to be aware that members of the population should be assigned a number. The random numbers are then used to select members of the population. In the case of a simple random sample, random numbers must be recorded so in the event of a duplicate arising, it can be identified and ignored. Students must then be able to apply this methodology in the context of a question. Students could be made aware of the use of censuses or records in a database in order to assist in the sampling process.

Exemplar

A television station wishes to identify the views of the general public about its latest drama. The head of drama wishes to take a random sample of 50 people. Describe how she can collect this sample.

- *The head of drama can use the electoral roll in order to obtain the names of the general public and assign each a number (enumerate) starting from 1 (or 0) upwards.*
- *She can then use the random number generator on the calculator to generate a number in the required range.*
- *Select the person who has been assigned that number.*
- *If the person has already been selected, discard the random number and generate a new number.*
- *Continue until 50 people have been selected.*

Questions may be presented to students in reverse, and students have to identify whether the sample has been chosen at random.

OPPORTUNITIES FOR EMBEDDING THE SEC

B1 Awareness of the dangers (e.g. privacy, data protection) and disadvantages (e.g. may not be able to get the details of every member of a population, numbering each member of the population takes time) of collecting a random sample.

B2 Identifying when a sample is not a random sample, from a source of secondary data.

B3 Appreciating that without the data collection methodology, how will they know if a sample is truly random?

B4 Appreciating that a non-random sampling method could be construed as biased.

Exemplar

A television station wishes to identify the views of the general public about its latest drama. The head of drama wishes to take a sample of 50 people using names from the electoral roll.

a) Give an advantage and a disadvantage of using the electoral roll.

- *An advantage of using the electoral roll is that the names of the majority of the population over the age of 18 will be available for sampling.*
- *A disadvantage of this is that nobody under the age of 18 will be listed, nor anybody who has opted be "private".*
  OR
  A disadvantage is that there are a lot of names on the roll, and it may be difficult to contact all of the selected people.

b) The head of drama picks 50 names listed on the electoral roll. Is this sampling method biased?

Since the sampling method has not been clearly described, it is unknown whether her sample was obtained randomly, or whether or not every possible sample of size 50 could be obtained from her method.

Therefore it is difficult to determine whether the sampling method is biased.

COMMON AND POSSIBLE MISTAKES

- Students often forget the second property of a random sample;
- When describing a data collection methodology, small details are often omitted (where did they get the population information from, how were the random numbers used etc.).
- Literacy and communication are often the biggest source of mistakes. Students need to remember that the sample itself cannot be biased – it is the sampling method which may be biased, meaning the sample is unrepresentative in a predictable way.

8b. Data Collection: Sampling Methods (3.3)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Describe a method of selecting a systematic sample in context.
- Describe a method of selecting a cluster sample in context.
- Describe a method of selecting a judgmental sample in context.
- Describe a method of selecting a snowball sample in context.
- Calculate quotas for a stratified sample from given information about the population.
- Describe a method of selecting a stratified sample in context.

TEACHING POINTS

This sub-unit consists of two parts: Understanding the methodology behind the sampling methods,
and applying the methodology to a given context.

Systematic Sampling

A systematic sample is obtained when members of the population are selected in a regular pattern. There are a few possible methods used for this sampling technique.

Exemplar

Describe how you would take a systematic sample of 30 people from a population of 1000.

Method 1
First, assign each of the 1000 people a number from 1 to 1000.
Since $\frac{1000}{30} = 33\frac{1}{3}$*, use random numbers to generate a number between 1 and 33, select the person assigned that number, and then select every 33rd person after that.*

Students need to be aware that this method is biased, as there would be 10 people at the end of the list that would never be considered.

Method 2

First, assign each of the 1000 people a number from 1 to 1000.

$\frac{1000}{30} = 33$ *remainder* 10*. Use random numbers to generate a number between 1 and 43 [33+10],*

select the person assigned that number, and then select every 33rd person after that.

This method is designed to remove the bias in **Method 1**. However, it introduces a new bias, as certain people are twice as likely to be selected as others. For example, person 34 would be selected if the random number was 1 or 34, but person 33 would only be selected if the random number was 33.

So, person 34 is twice as likely to be selected in the sample as person 33.

Method 3

First, assign each of the 1000 people a number from 1 to 1000.

Since $\frac{1000}{30} = 33\frac{1}{3}$*, use random numbers to generate a number between 1 and 34, select the person assigned that number, and then select every 34th person after that.*

If the final number is greater than 1000, wrap around to the start of the list again by subtracting 1000 from this number.

For example, if the final number is 1009, person number 9 should be selected.

This is a different way to remove the bias described in **Method 1**. However, it also introduces a similar bias to **Method 2**. For example, if the random number was 15, the final number for selection would be 1001, so person 1 would be selected. Person 1 would also be selected if the random number was 1. However, person 14 would only be selected if the random number was 14.

So, person 15 is twice as likely to be selected in the sample as person 14.

Method 4

First, assign each of the 1000 people a number from 1 to 1000.

$\frac{1000}{30} = 33\frac{1}{3}$

Use random numbers to generate a number between 1 and 1000, select the person assigned that number, and then select every 33rd person after that.

When the number is greater than 1000, wrap around to the start of the list again by subtracting 1000 from the number, and continue counting each 33rd person from here.

This is the best method for minimising bias.

Students need to be aware that, although the first number is selected at random, a systematic sample is not a random sample.

Cluster Sampling

A cluster sample is obtained when a random sample is taken, not from the whole population, but from selected clusters (either randomly or non-randomly) from the population. Students need to be able to identify appropriate clusters from the context of the question.

Exemplar

A cluster sample of 50 people elected to county councils in Wales is to be taken. There are 22 county councils in total, each with between 40 and 80 councillors.

a) **Describe appropriate clusters in this context.**
Appropriate clusters in this case would be individual county councils in Wales.

b) **Describe how you would take a cluster sample.**
- *Select 10 county councils in Wales that have different features (e.g. geographical location, rural/urban areas etc.) – this will give a good representation of Wales.*
- *For each of the selected county councils, assign each councillor a number from 1 to n, where n is the number of councillors in that council.*
- *Use a random number generator to generate 5 distinct random numbers between 1 and n.*
- *Select the five councillors assigned those numbers.*

A note about the above example: Clusters can be chosen at the sampler's discretion. The example above highlights that different county councils may have different features to provide a representation of Wales, the disadvantage being the lack of convenience. Clusters could equally be chosen as to maximise convenience (e.g. picking county councils that are close together) but this may decrease representation.

Students need to be aware that, although the clusters may be selected at random, and each member within each cluster is selected at random, a cluster sample is not a random sample.

Judgemental Sampling

A judgemental sample is obtained when members of the population are selected according to criteria based on the judgement of the sampler. Students must appreciate that this is a non-random sampling method, but must also be aware of occasions where it may be beneficial to be used (very specific situations).

Exemplar

A researcher is trying to determine whether the provisions for blind civil servants working in Scotland are acceptable. Explain why he might want to use a judgemental sample in this case.

Since "blind civil servants in Scotland" is a very specific phrasing, there may not be very many of them. To save time, the researcher may just pick civil servants in Scotland who are either blind, partially sighted, or have experience with blind people.

Make students aware that an understanding of the factors related to the question is important when taking a judgemental sample, and the method is best used when the criteria for population is less well-defined than in this example. If, in the previous example, the researcher only wanted to sample blind civil servants in Scotland, a random sample of the restricted population of blind civil servants in Scotland may be more appropriate. However, the question he is investigating need not necessarily require members of his sample to be blind themselves, but he would like them to have some knowledge or experience of issues relating to partial sight. In general, using this criterion is difficult to identify members of the population, so a judgemental sample would be more appropriate.

Snowball Sampling

A snowball sample is obtained when members of the sample recommend, or refer, new sample members into the sample. It is used when it is difficult to locate members of the population of interest. Students must appreciate that there are some issues where the question being investigated may be a sensitive topic, and information relating to the population of interest may not be easily available, for example rape victims, sufferers of FGM etc. If one member of the population can be identified, then they may know a friend, relative, member of the same support group etc. who could also volunteer for sampling.

Use professional judgement when utilising sensitive topics that may be triggering to students.

Snowball sampling is also present in modern day social media where a poll or survey is created and posted on social media. This can then be shared by friends/followers/subscribers amongst their contacts so the poll/survey can reach an audience not directly connected to the origin. In situations such as these, it isn't always possible to reach a set sample size so a termination step such as a timeframe is required.

Exemplar

A researcher wants to investigate the provision there is for Myeloma (a rare form of blood cancer) patients.
Due to the rarity of the disease, it is difficult for the researcher to find enough people to use in her sample of 30.
She has currently found 3 Myeloma patients.
Explain how the researcher could use a snowball sample in this situation.

The researcher could ask her three Myeloma patients if they attend a support group/clinic for people with the same disease.
Then ask if her patients know any other patients from this group/clinic who would volunteer to participate in the investigation.
If more patients participate in the investigation, she can ask the new patients to ask others until she has reached her sample of 30.

Proportional Stratified Sampling

Make students aware of the terms stratum and strata and be able to write the proportion represented by each of the strata as a fraction. Multiplying the fraction by the desired sample size and rounding appropriately (depending on the context) gives the desired sample size to represent each of the strata.

Examples with no context could be attempted first:

a) **Write down 50 as a fraction of 200.**
b) **Find ¼ of 8.**

This can then escalate to basic contexts:

Exemplar

A bag of 200 sweets contains chocolates, toffees or jelly beans.
In a particular bag of sweets, there are 50 chocolates.
A taste tester wants to use a stratified sample of 8 to test the quality of the sweets.
How many chocolates should be in the sample?

The fraction of chocolates in the bag is $\frac{50}{200} = \frac{1}{4}$*. So* $\frac{1}{4} \times 8 = 2$*. Hence there should be 2 chocolates in the sample.*

Students need to understand that the use of a simple random sample should be used to fill up each stratum. They must also be aware that not every sample of size *n* is possible from the population, so a stratified sample is not a random sample.

A disproportionate stratified sample may also be used when the population information is not available to hand, or when it is desirable to have a given proportion of strata in the

sample (e.g. wanting a sample to be half male, half female, regardless of the population proportions). The quotas for each strata should be determined through judgement (see below), but should be a minimum of 1.

OPPORTUNITIES FOR EMBEDDING THE SEC

A1 Identifying factors from the context to select an appropriate sampling method.

A3 Describing suitable data collection methods in detail.

A4 Exploratory data analysis may be required in order to know information about a population (e.g. strata proportions).

A6 Considering bias when selecting a sampling method.

B1 Recognising that judgemental and snowball sampling methods are easier to collect, but are non-random. Also recognising the constraints in collecting a random sample (time, money, difficulty etc.)

B3 Recognising that without the data collection methodology declared, readers of statistical research cannot take likely bias into consideration.

B4 For example, judgemental sampling has the potential to be very biased.

COMMON AND POSSIBLE MISTAKES

- Forgetting to detail how random numbers are used in the sampling process. Students also confuse a stratified sample with a systematic sample, since they both begin with ‘S’.
- Students thinking that a random sample must always be the ‘best’ way to sample without reference to the context or the practicalities.

8c. Data Collection: Suitability of Sampling Methods (3.4) (3.6)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Know the advantages and disadvantages of each sampling method.
- Select the most appropriate sampling method in context.
- Identify and describe the advantages and disadvantages of each sampling method in context.

TEACHING POINTS

Knowledge of the advantages and disadvantages of each sampling method is the first stage. Use of Tarsia puzzles, matching activities, group discussion etc. can be very effective here.

The advantages and disadvantages for each sampling method include (but not limited to):

Sampling method	Advantages	Disadvantages
Random sampling (Simple or Unrestricted)	Minimal bias. Most statistical theory assumes the use of a random sample	Potentially difficult to do. Can still have an unrepresentative sample. Potentially time-consuming or expensive. Population members can rarely be listed.
Systematic Sampling	The advantages are largely context-dependent and there are no generic advantages other than an element of randomness in selecting the first member.	The disadvantages are largely context-dependent and there are no generic disadvantages other than it is not a random sample.
Cluster Sampling	Useful if members of the population are grouped in clusters e.g. reduced travelling. More convenient than random or stratified sampling.	Less variation in smaller clusters than in the whole population
Proportional Stratified Sampling	Each stratum of interest is fairly represented in the sample. Minimal bias.	More complicated than random sampling. Information about the population needs be known to minimise bias. Population members can rarely be listed.

Disproportional Stratified Sampling	Each stratum of interest is represented in the sample. No information about the population is required.	Strata are not fairly represented.
Judgemental sampling	Easy and convenient to do.	Not random, potentially biased. Relies solely on the judgement of the sampler, who may or may not be correct themselves.
Snowball sampling	Sample elements are always of interest. Potentially a high yield of results for low effort	Not random, potentially biased. Relies on other people to volunteer information.

Once students have learnt the advantages and disadvantages of each sampling method, they can be exposed to contextual situations. Case studies and group work are particularly effective here. For example: split the class into groups of three. Each group receives a different scenario requiring students to decide what they think is the best sampling method, giving their reasons. Rotating the scenarios after a short time, for example, 5 minutes, and repeat the activity. At the end, you could use a class discussion of the best methods chosen for each scenario, together with justification. The class could then also critique the methods, identifying possible disadvantages of these decisions. It is important that students are aware that there are always disadvantages when choosing a sampling method.

OPPORTUNITIES FOR EMBEDDING THE SEC

A1 Identifying factors from the context to select the most appropriate sampling method.

A6 Ensuring that bias is taken into account when selecting a sampling method.

E1 Identifying disadvantages of the chosen sampling method.

E2 Recognising how the disadvantages of the chosen sampling method could affect the results or process of the investigation.

For example:

A researcher wants to investigate the provision there is for Myeloma (a rare form of blood cancer) sufferers. Due to the rarity of the disease, the researcher has found it difficult to find subjects to interview. She wants to take a sample of 30 people from different regions of the UK, and preferably an even split between male and female. Due to doctor-patient confidentiality, she was denied access to a hospital list of Myeloma patients.

This example has been constructed in such a way that a snowball, stratified or cluster sample could be considered. This example can generate discussion amongst the

students in order to appreciate that there isn't one "correct" answer, and all advantages and disadvantages must be considered before selecting a sampling method.
Whichever sampling method is chosen, the initial question must also be considered: 'What provision is there for Myeloma sufferers?' Regardless of the findings of her investigation, the sampling methodology must be taken into account before making conclusions. For example, if a cluster sample was chosen, is the provision for Myeloma suffers representative across the country, or just in the regions in which the clusters were chosen?

COMMON AND POSSIBLE MISTAKES

Apart from misremembering the advantages and disadvantages, students often forget to refer to the context of the question or write a long, confused piece of prose rather than clear bullet points.

UNIT 9: Estimation and Approximation

Return to overview

SPECIFICATION REFERENCES

6.5 Use the fact that the distribution of $\bar{X}$ has a normal distribution if X has a normal distribution.

6.6 Use the fact that the normal distribution can be used to approximate a binomial distribution under particular circumstances.

8.1 Use and demonstrate understanding of the terms parameter, statistic, unbiased and standard error.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics
Numerical Measures (Unit 1)
The Binomial Distribution (Unit 5)
The Normal Distribution (Unit 7)

KEYWORDS

biased, binomial, continuity correction, continuous, discrete, distribution, estimate, estimator, mean, normal, sample, sampling distribution, standard error, symmetrical, unbiased, variance,

UNIT SUMMARY

This unit further extends the normal distribution, and also links it to the binomial distribution. The sampling distribution of a mean of a normal distribution has previously been in legacy AS Mathematics Statistics module, but the normal approximation to the binomial used to be in legacy A2 Mathematics Statistics module. They are now grouped together under the same umbrella title of Estimation and Approximation, which are important concepts when analysing statistics and making inferences about a population.

Unbiased estimates are not assessed until Year 2, however knowledge of them here will allow students to appreciate the idea of the $\bar{X}$ distribution better. The concept that $\bar{X}$ is an unbiased estimator for μ, and there is a similar unbiased estimator for the variance is no longer on the specification. However, it may be useful for teachers to point this out as an extension activity.

Activities in Desmos that are helpful here: Sampling Distributions, Normal Distribution, Binomial Distribution with Normal Approximation.

9a. Estimation: Terminology (6.5) (8.1)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand the difference between a population parameter and a sample statistic.
- Understand that the sampling distribution of a statistic is the distribution of that statistic, when considered as a random variable, when derived from a random sample of size n.
- Appreciate that $\bar{X}$ is a random variable, with distributions called the sampling distribution of the mean.
- Know the definition of the terms: unbiased estimate, standard error.
- Appreciate that $\bar{x}$ is an unbiased estimate for μ and s^2 is an unbiased estimate for σ^2.

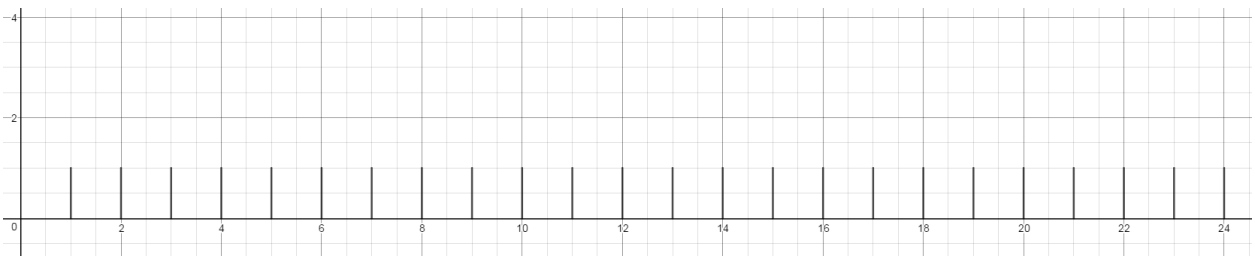
TEACHING POINTS

This unit is an interesting topic from both the point of view of statistics and the teaching of the statistics course.

Firstly, define the concept of a parameter: **a numerical property of the population**. Revise the parameters of the normal distribution and the binomial distribution.

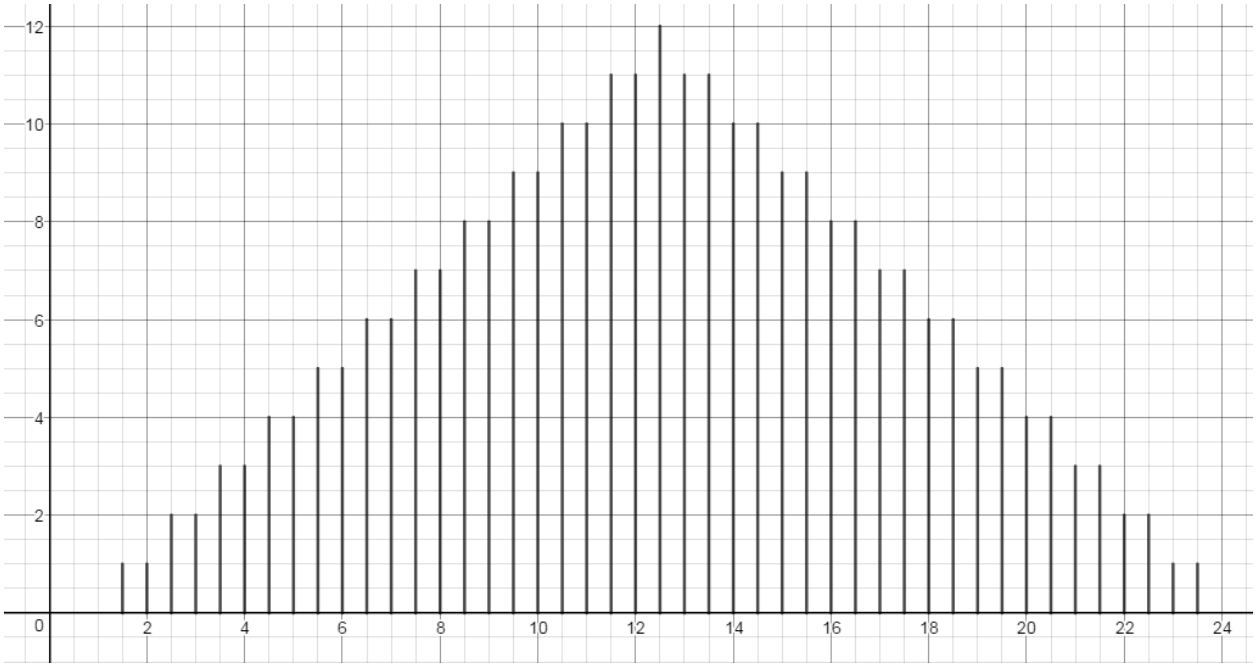
Then define the concept of a statistic: **a number calculated from a sample containing no unknown parameters**. Examples include the numerical measures calculated in Unit 1. Remind students of the difference between the sample mean $\bar{x}$, the population mean μ, the sample standard deviation *s*, and the population standard deviation σ. Identifying which are population parameters and which are sample statistics is a good way to help students make the distinction.

As a lead into the following sub-section, a group activity illustrating a sampling distribution could be used. For example, use this method of generating a sampling distribution of a mean from a sample size of two from a discrete uniform distribution: every member of the class is assigned a unique number. The frequency or probability of each number is plotted on a vertical line chart. The following screenshots are from the Sampling Distributions activity on Desmos (the number can be changed within the activity).



The above is an example vertical line chart for a class of 24 students (or 23 and the teacher should you wish to also get involved).

The class then pairs off, and the mean of each number is taken and recorded. The class then pick a different pairing and the process is repeated. Allow this process to repeat n times, each time the mean recorded (as many as practical). Plot these data on a vertical line chart, showing students a new probability distribution. To finish, show students the true sampling distribution of a mean of sample size 2, explaining that this is what they would get if they continued the activity.



The figure above is the complete sampling distribution of the means of samples of size 2, from the original example.

Explain to students that the full process would require every possible pairing to be recorded. Also explain that this activity could be generalised to all samples of size *n*, or any other statistic that can be calculated from a sample, e.g. the standard deviation.

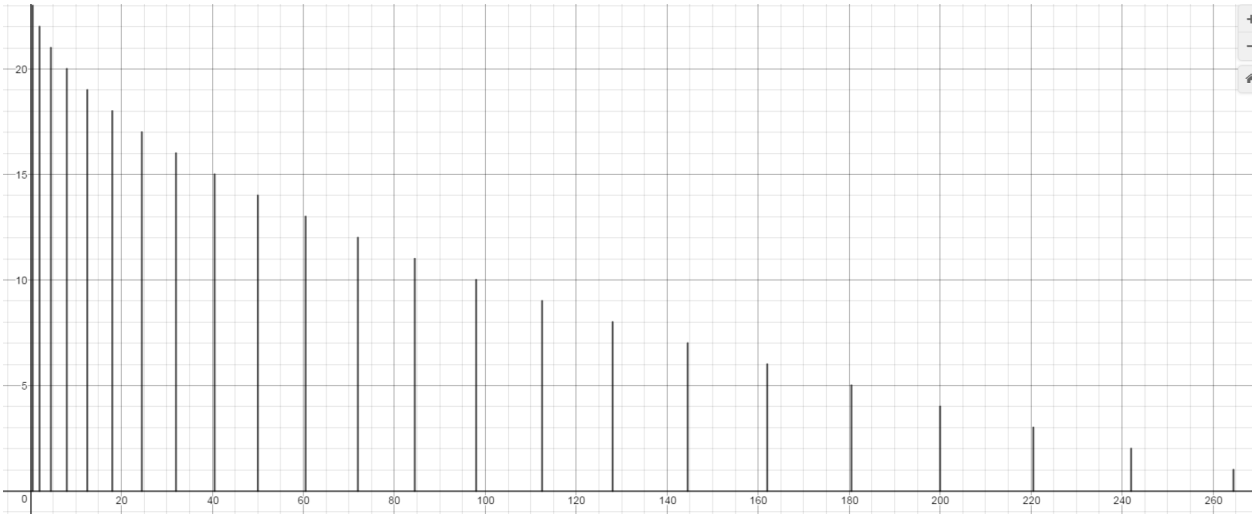
Emphasise to students that the sampling distribution is a different population from the one started out with (this is best seen with a uniform distribution), and emphasise the difference between the original distribution and the sampling distribution of the mean. In the example activity above the original random variable is *a number between 1 and n*, and the new population is *the mean of a sample of size 2 from the population of numbers between 1 and n*.

An unbiased estimate of a parameter is a statistic where the mean of the sampling distribution of that statistic is equal to the population parameter being estimated. This can be explained to students by way of example: the average of all possible sample means is the population mean. A recap of expectation (Unit 4) may be appropriate here, and a brief calculation using the example above will help here. Let the students see this example, illustrating that the mean of the sampling distribution is equal to the mean of

the population distribution (in the example of 24, the mean is 12 in both cases). Generalise that the sample mean is always an unbiased estimate for the mean (the algebraic proof can be seen in the following sub-unit).

The standard error of an estimate is the standard deviation of the sampling distribution of that statistic. Again, a recap of variance (Unit 1, 4) may be appropriate here together with a brief calculation using the above example.
The concept that $\bar{X}$ is an unbiased estimator for μ, and there is a similar unbiased estimator for the variance is no longer on the specification. It is worth pointing this out as an extension activity.

As an extension, students can investigate other sampling distributions, for example the sampling distribution of the variance.



The figure above is the sampling mean for the variance from the above example.

Finding the mean of this sampling distribution and the variance of the population distribution should show them to be equal (the algebraic proof of this is more complicated, but the most mathematically able of students should be able to follow the proof).

Emphasise that this doesn't imply that s is an unbiased estimate for the standard deviation – the example above will yield that the mean of the sampling distribution of the standard deviation is not equal to the population standard deviation.

OPPORTUNITIES FOR EMBEDDING THE SEC

B1 Appreciating that finding the sampling distribution of the mean, even from a small population, is highly impractical.

C1 Using software (spreadsheets etc.) to process data relating to the sampling distribution of the mean.

COMMON AND POSSIBLE MISTAKES

Students often confuse the original population distribution with its sampling distributions. This is down to poor comprehension.

NOTES

Students do not need to know much in the way of detail of unbiased estimates and standard errors. Indeed, they need only know that $\bar{x}$ is always an unbiased estimate for μ (or “the average of all sample means is the population mean”), s^2 (*n* -1 divisor) is an unbiased estimate for σ^2 (or “the average of all sample variances is the population variance”).

9b. Estimation: The Sampling Distribution of the Mean of a Normal Distribution (6.5)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand that the sampling distribution of the mean of a normal distribution, $\bar{X}$, is also normally distributed.
- Understand that $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
- Find probabilities and sample means using the normal distribution of $\bar{X}$.
- Distinguish between contexts involving X and $\bar{X}$.

TEACHING POINTS

Remind students that the normal distribution is an idyllic situation where nice things happen. Without proof, explain that if $X \sim \mathrm{N}(\mu, \sigma^2)$ then $\bar{X}$ also has a normal distribution.

One can prove that the population parameters of $\bar{X}$ are μ and $\frac{\sigma^2}{n}$ respectively:

$$\mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{n}\Sigma_{i=1}^{n} X_i\right) = \frac{1}{n}\Sigma_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\Sigma_{i=1}^{n}\mu = \frac{1}{n} \times n\mu = \mu$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\Sigma_{i=1}^{n} X_i\right) = \frac{1}{n^2}\Sigma_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}.$$

This uses the results from Unit 16, so you may wish to leave this until later.

Recap the normal distribution (Unit 7), using the calculator. Make sure you give students enough practice at finding both probabilities and x-values.

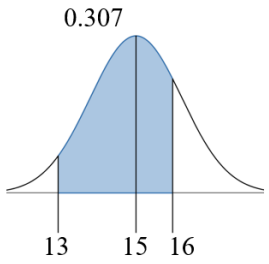
Then give students examples without context, asking them to find probabilities and values of both the population distribution and the sampling distribution.

Exemplar

Let X be a random variable such that $X \sim \mathrm{N}(15, 14.2)$.

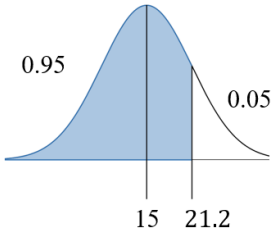
a) Find $\mathrm{P}(13 \leq X < 16)$.

Using the calculator: $\mathrm{P}(13 \leq X < 16) = 0.3068$.



b) Find the value of a such that $\mathrm{P}(X \geq a) = 0.05$.

$\mathrm{P}(X \leq a) = 0.95$. *Using the calculator, a =21.2*

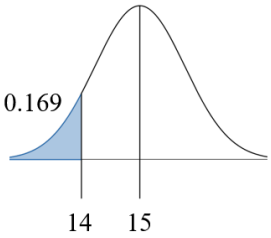


A sample of size 13 is taken at random from the population of X.

c) Find the probability that the mean of this sample is at most 14.

Since X is normally distributed, $\bar{X} \sim \mathrm{N}\left(15, \frac{14.2}{13}\right)$.

Using the calculator: $\mathrm{P}(\bar{X} \leq 14)$ =*0.1693*.



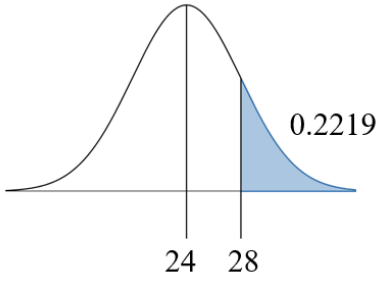
After students have done enough practice without context, start introducing context. This will allow students to appreciate there is a difference between the two distributions. This will not only recap skills learnt in Unit 7, but will also provide further meaning to the differences between X and $\bar{X}$. It is very important students define their variables correctly, which should be second nature to them by now. It is also important that they also define $\bar{X}$ together with the sample size they are using (especially if the question uses two different sample sizes in different parts).

Exemplar

The plums from a particular variety of plum tree have masses which can be modelled by a normal distribution with mean 24 g and standard deviation 5 g.

a) Plums weighing more than 28 g are graded as large. What proportion of the plums is graded as large?

Let X be the mass of a plum. So $X \sim \mathrm{N}(24, 5^2)$.

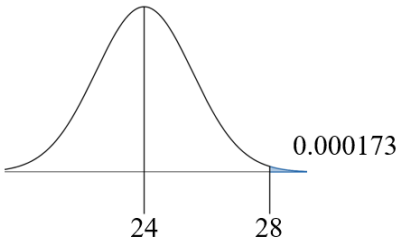


Probability of a large plum: $\mathrm{P}(X > 28) = 0.2119$

Proportion of plums graded large is 21.2%

b) 20 plums are selected at random.
What is the probability that the mean mass of these 20 plums exceeds 28 g?

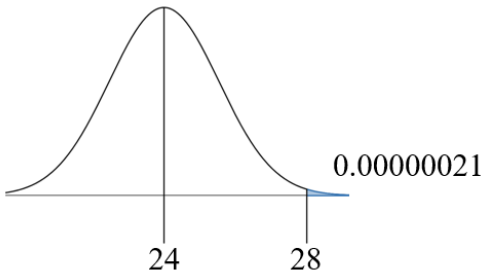
Let $\bar{X}$ be the mean mass of a plum from a sample of size 20. So $\bar{X} \sim \mathrm{N}\left(24, \frac{5^2}{20}\right)$.



Probability that the mean mass of a sample of 20 plums exceeds 28 g:
$\mathrm{P}(\bar{X} > 28) = 0.0002$.

c) This time 40 plums are selected at random.
What is the probability that the mean mass of these 20 plums exceeds 28 g?

This time, let $\bar{X}$ be the mean mass of a plum from a sample of size 40. So $\bar{X} \sim \mathrm{N}\left(24, \frac{5^2}{40}\right)$.



Probability that the mean mass of a plum exceeds 28 g : $\mathrm{P}(\bar{X} > 28) = 0.00000021$

Other questions that could be asked could be to find unknown parameters of the sampling distribution for the mean. This can also lead into interesting questions, for example to find the sample size.

Exemplar

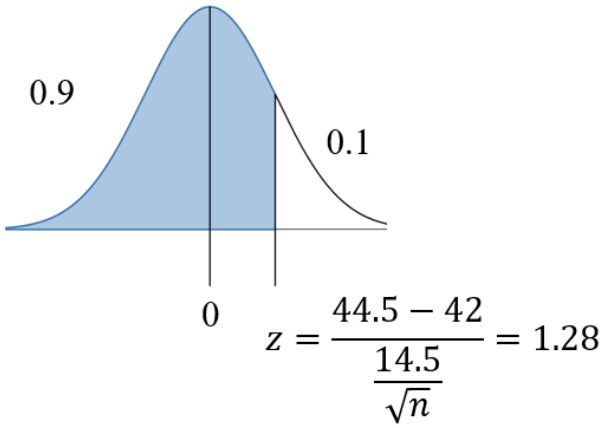
Let $X \sim \mathrm{N}(42, 14.5^2)$.
It is known that 10% of all samples of size n have a mean of at least 44.5. Find n.

Since X is normally distributed, $\bar{X} \sim \mathrm{N}\left(42, \frac{14.5^2}{n}\right)$.

Set-up: We know $\mathrm{P}(\bar{X} \geq 44.5) = 0.1$, *so* $\mathrm{P}(\bar{X} \leq 44.5) = 0.9$.

Standardising: $\mathrm{P}\left(Z \leq \frac{44.5-42}{\frac{14.5}{\sqrt{n}}}\right) = 0.9$ *(or* $44.5 = 42 + z \times \frac{14.5}{\sqrt{n}}$*),* *where* $Z \sim \mathrm{N}(0,1^2)$

Sketch:



Solve: Using the calculator, $z = \frac{44.5-42}{\frac{14.5}{\sqrt{n}}} = 1.2816$.

Rearranging: $\sqrt{n} = \frac{1.2816 \times 14.5}{44.5-42} = 7.43328$, *so* $n = 56$ *(rounded up)*
(Alternatively, this may be solved using the equation solver on the calculator)

OPPORTUNITIES FOR EMBEDDING THE SEC

D2 Interpreting the numerical measures calculated in the context of the question. It is especially important here, since there are two possible distributions to calculate from, and it must be clear which distribution they are applying the measure to in context.

D5 Awareness of what the sampling distribution means, so findings can be explained in a language appropriate to the given target audience.

For example, if the following question were added to the example above about plums:

Exemplar

d) **It is claimed that, based on the calculation made in part (c), there is no chance of a box of 40 plums containing a large plum.**
Comment on this claim.
The probability in part (c) shows that there it is almost impossible for the mean mass of 40 plums to be larger than 28 g.
This does not mean that there are no large plums in the box.
There could be a large plum in the box, with the other 39 plums having much smaller masses.

COMMON AND POSSIBLE MISTAKES

- When using both distributions, students may forget to use the standard error of the mean as the population parameter for the sampling distribution of the mean.
- Students may also use $\frac{\sigma}{n}$ instead of $\frac{\sigma}{\sqrt{n}}$ or σ^2 instead of σ. This is a problem when inputting population parameters into the calculator, or standardising.
- When standardising, many students often forget to use the sampling distribution for the mean. This needs to be addressed before Unit 11.
- Students must remember that a sample cannot have a normal distribution – it is the population which has the normal distribution.

NOTES

Students are expected to know that the variance of $\overline{X}$ is lower than X and low tariff questions may be asked about this.

9c. Estimation: The Normal Approximation to the Binomial (6.6)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Know the conditions when the normal distribution can be used to approximate the binomial distribution.
- Understand the use of, and apply, a continuity correction.
- Calculate approximations of binomial probabilities using the normal distribution.

TEACHING POINTS

Recap the binomial distribution. Use the Binomial Distribution with Normal Approximation on Desmos to show vertical line charts of the binomial probability distribution when n gets large and p is close to 0.5. Students should be able to recognise the vertical line charts resembling the bell-shape of the normal distribution.

Students must be aware that the normal distribution can be used to approximate the binomial distribution if:

- $n \geq 20$ and $p \approx 0.5$
- $np \geq 5$ and $n - np \geq 5$.
  (Note that $0.25 \leq p \leq 0.75$ is sufficient here since this is the smallest interval (if $n = 20$))

Recap the mean and variance of the binomial distribution.

Students need to understand that if the above conditions are satisfied, then if $X \sim \mathrm{B}(n, p)$ then $X \approx Y \sim \mathrm{N}\big(np, np(1-p)\big)$, or "we may assume $X \sim \mathrm{N}(np, np(1-p))$."

Recap the difference between discrete and continuous data. Remind students that the binomial distribution is discrete and the normal distribution is continuous. This is best explained through the vertical line chart used earlier, together with an overlay of the normal distribution curve.

The use of bars overlaid over the vertical line chart, such that the width of each bar is 1 (so the class interval for the bar representing $x = a$ is $a - 0.5 \leq x \leq a + 0.5$) could be seen, so students recognise that a vertical line chart can be transformed into a histogram. It would be beneficial for students to check that the areas of each bar corresponds to $\mathrm{P}(X = x)$, and this will help remind students that the probability is represented by the area.

The idea of a continuity correction can now be introduced, using the histogram produced earlier as a visual aid.

Students could begin by establishing bounds on discrete data.

Exemplar

Find the set of values which lie in the set $13 \leq x < 16$ to the nearest integer.

The integers in the set are 13, 14 and 15. So the set of values required is $12.5 \leq x \leq 15.5$.

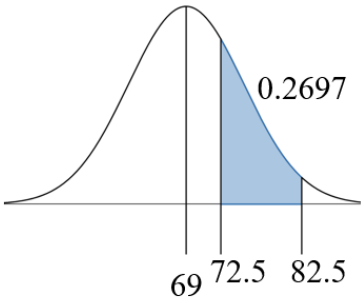
Students will find applying continuity corrections much easier when the inequalities are expressed in symbols, instead of words such as “at least” or “at most”. Afterwards, applying the normal approximation to the binomial with no context could be practised.

Exemplar

Let X be a random variable such that $X \sim \mathrm{B}(150, 0.46)$.
Find an approximation for $\mathrm{P}(72 < X < 83)$.

Since n *is large , and* $0.25 \leq p \leq 0.75$ *(or* $np = 150 \times 0.46 = 69$ *and* $n - np = 150 - 69 = 81$*)*
the normal approximation to the binomial can be used.
Hence, $X \approx Y \sim \mathrm{N}(69, 37.26)$

The integers in the set $72 < X < 83$ *are* $73\ to\ 82$. *So* $\mathrm{P}(72 < X < 83) \approx \mathrm{P}(72.5 \leq Y \leq 82.5)$.



So $\mathrm{P}(72 < X < 83) \approx 0.2697$ *using a normal approximation.*

Once these skills have been mastered, context can be introduced. Even more so than the previous sub-unit, students must be able to define their variables. Students will be always asked to use a **suitable approximation** in a question if the use of the normal approximation to the binomial is specifically required (this is due to the use of calculators, and their ability to calculate binomial probabilities for large n). Students may be asked to justify the use of the normal approximation, even without being asked to use this in the question.

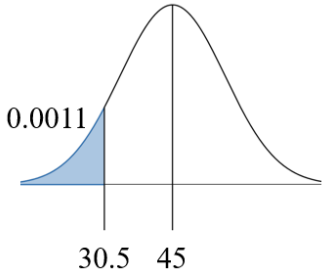
Exemplar

A certain variety of flower seed is sold in packets containing about 1000 seeds. The packets claim that 45% will bloom white and 55% red. This may be assumed to be accurate.

If 100 seeds are planted, use a suitable approximation to find the probability that at most 30 will bloom white.

Let X *be the number of seeds which bloom white. Then* $X \sim \mathrm{B}(100,0.45)$.
Since n *is large, and* $0.25 \leq p \leq 0.75$ *(or* $np = 100 \times 0.45 = 45$ *and* $n - np = 100 - 45 = 55$*),*
the normal approximation to the binomial can be used,
with $\mu = 45$ *and* $\sigma^2 = np(1-p) = 100 \times 0.45 \times 0.55 = 22.5$.
So $X \approx Y \sim \mathrm{N}(45, 22.5)$.

$\mathrm{P}(X \leq 30) is\ required$. *The largest integer in this set is 30. So* $\mathrm{P}(X \leq 30) \approx \mathrm{P}(Y \leq 30.5)$.



So $\mathrm{P}(X \leq 30) \approx 0.0011$.

If students are **not specifically requested to use an approximation** in this type of question, then a simple answer using the binomial probability function on the calculator directly is perfectly acceptable and sensible.

Students also need to be able to apply the normal approximation to the binomial distribution for inverse binomial questions.

Exemplar

It is known that 20% of components are faulty. During a quality control check, a random sample of 60 components are taken and the number of faulty components counted. The batch will be accepted if at most *k* faulty components are found, otherwise the batch is rejected. The quality control officer wants the probability of finding at most *k* faulty components to be less than 12%. Use the normal approximation to find the value of *k*.

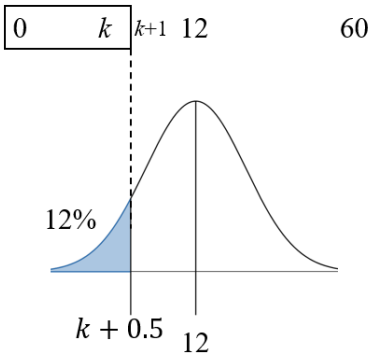
Let X be the number of faulty components.

Since a component is either faulty or not, the sample is random so the components are likely to be independent of each other, there is a fixed number of 60 components and the chance a component is faulty is 20%, X~B(60,0.2).

Since n *is large* $(n \geq 20)$ *and* $np = 60 \times 0.2 = 12 > 10$ *and* $n - np = 60 - 12 = 48 > 10$*, then normal approximation can be used.*
(Note: in this case p *is not between* 0.25 *and* 0.75 *but the normal approximation can still be used since* $np > 10$ *and* $n - np > 10$*)*

Since $\mu = np = 12$ *and* $\sigma^2 = np(1-p) = 60 \times 0.2 \times 0.8 = 9.6$*, then* $X \approx Y \sim N(12, 9.6)$

We want P(X ≤ k) < 12% so using the normal approximation and a continuity correction, we need $P(Y \leq k + 0.5) = 12\%$.



The calculator gives $k + 0.5 = 8.36$*, so* $k = 7.86$.

k *must be an integer so we must determine whether k = 7 or k = 8 for the initial requirement that* $P(X \leq k) < 12\%$*. Since* $P(X \leq 7) = 0.067 < 12\%$ *and* $P(X \leq 8) = 0.127 > 12\%$*, we must have* $k = 7$.

OPPORTUNITIES FOR EMBEDDING THE SEC

D4 Awareness that the conditions of the normal approximation to the binomial must be met before applying the normal approximation. Equally, students must be aware of the importance of a continuity correction.

D5 Communication of the reasons why conclusions made from an incorrect use of the normal approximation might be incorrect.

Exemplar

Over a long period, a company finds that 3% of its employees say they have a disability.
A random sample of 15 employees is taken from the employment roll.
Kevin decides to use a normal approximation to find that the probability that at least one employee is disabled.
He uses $X \sim \mathrm{N}(0.45, 0.4365)$ and calculates $\mathrm{P}(X \geq 1) = 0.203$.
He concludes that there is a 20.3% chance of having a disabled employee in his sample of 15.

a) **Give two different reasons why Kevin's method is not valid.**
*Since n is not large (or np is not bigger than 5), the conditions for when a normal approximation can be used for the binomial are **not** met, so the approximation shouldn't be used.*
Also, Kevin did not apply a continuity correction when calculating the probability.

b) **Explain, in context, why Kevin's conclusion is incorrect.**
The company finds that only 3 out of every 100 employees say they have a disability.
So the likelihood of finding a disabled employee in a sample of 15 randomly chosen employees is quite low, whereas Kevin is claiming a relatively high chance of finding a disabled employee in his sample.

COMMON AND POSSIBLE MISTAKES

- The biggest error students make is omitting continuity corrections. Literacy is again a big issue here, enhanced by the use of continuity corrections.
For example:
Let $X \sim \mathrm{B}(100, 0.5)$, so $X \approx Y \sim \mathrm{N}(50, 5^2)$.
Find the probability that X exceeds 55.
"Require $\mathrm{P}(X > 55) \approx \mathrm{P}(Y \geq 55.5)$" students may interpret as "require $\mathrm{P}(X \geq 55) \approx \mathrm{P}(Y \geq 54.5)$"
In this example, the student misinterpreted "exceeds" as "at least".
- Students often forget the conditions on p when using the normal approximation.

NOTES

A special case of the sampling distribution of the mean is the sampling distribution of the proportion. It is not in the Year 1 portion of the course but is used in Unit 21c when testing for the difference between two proportions. You could teach the following now. It is a combination of this sub-unit and the previous one, and some knowledge of Unit 4b is required.

Let $X \sim \mathrm{B}(n, p)$. If n is sufficiently large, then the normal approximation to the binomial can be used. So we can assume $X \sim \mathrm{N}(np, np(1-p))$. The sampling distribution of the proportion is defined as $\frac{X}{n}$ – this is essentially $\bar{X}$, where we treat X has a combination of n independent Bernoulli trials. Using the formula in Unit 4b, $\frac{X}{n}$ has mean $\mathrm{E}\left(\frac{X}{n}\right) = \frac{1}{n}\mathrm{E}(X) = \frac{1}{n}np = p$ and variance $\mathrm{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2}\mathrm{Var}(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$. So $\frac{X}{n} \sim \mathrm{N}\left(p, \frac{p(1-p)}{n}\right)$. Probabilities involving the proportion can now be calculated:

Extension Exemplar

A sample of 100 doves are selected from the population of doves. It is known that exactly half of the population are female. Find the probability that the proportion of the sample that is female exceeds 55%.

Let X be the number of doves who are female. So $X \sim \mathrm{B}(100, 0.5)$. Since n is large, the normal approximation to the binomial can be used, so the sampling distribution of the proportion can be used. So $\frac{X}{n} \sim \mathrm{N}\left(0.5, \frac{1}{400}\right)$. We want $\mathrm{P}\left(\frac{X}{n} \geq 0.545\right)$ (using the continuity correction) which equals 0.1841.

Although the use of continuity corrections when approximating to the binomial with the normal distribution is a standard topic, the use of continuity corrections in the sampling distribution for a binomial proportion is **not** included in the content.

UNIT 10: Introduction to Hypothesis Testing

Return to overview

SPECIFICATION REFERENCES

7.2 Use tables to test for significance of a correlation coefficient.

7.3 Know the appropriate conditions for the use of each of these methods of calculating correlation and determine an appropriate approach to assessing correlation in context.

8.1 Use and demonstrate understanding of the terms parameter, statistic, unbiased and standard error.

8.2 Know and use the language of statistical hypothesis testing: null hypothesis, alternative hypothesis, significance level, test statistic, 1- tail test, 2-tail test, critical value, critical region, and acceptance region and *p*-value.

8.3 Know that a sample is being used to make an inference about the population and appreciate the need for a random sample and of the necessary conditions.

8.4 Know that a sample is being used to make an inference about the population and appreciate the need for a random sample and of the necessary conditions.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics
Bivariate Data (Unit 6)
Normal Distribution (Unit 7)

KEYWORDS

acceptance region, alternative hypothesis, association, bivariate normal distribution, correlation, critical region, critical value, evidence, hypothesis, insufficient, negative, normal distribution, null hypothesis, one-tailed test, parameter, Pearson's product moment correlation coefficient, positive, *p*-value, rank, region, rho, significance level, significant, Spearman's rank correlation coefficient, statistic, sufficient, suggestion, tail, test statistic, two-tailed test,

UNIT SUMMARY

This unit is fundamental for almost a third of the course. Hypothesis testing is a core aspect of statistical analysis and should be carried out, interpreted and phrased correctly. The Statistical Enquiry Cycle has more of a role in this chapter due to the nature of hypothesis testing.

Legacy specifications have placed hypothesis testing for the sample mean of a normal distribution (a *z*-test) as the first hypothesis test seen. However, through previous

experience it is actually hypothesis tests about the correlation which students seem to understand more easily, and for this reason it is placed here, with *z*-tests in the following unit.

This will also be the first time students are required to use the table of values to determine the critical values for Pearson's PMCC and Spearman's rank correlation coefficients (the calculators are unable to calculate these values). It is suggested that *p*-values are left until Unit 11, since (unless certain calculators are used) students will be unable to calculate them for hypothesis tests about correlation, or non-parametric hypothesis tests (Units 13 and 14).

A note on the use of notation:
r denotes the sample PMCC,
r_s denotes the Spearman's rank correlation coefficient of the sample.
ρ denotes the population PMCC,
ρ_s denotes the Spearman's Rank correlation coefficient for the population

Note: The use of ρ_s is divided within statistical literature. One argument against the use of ρ_s is that r_s is not an unbiased estimate of ρ_s (unlike r is for ρ) and so the test itself (being a non-parametric test) is not actively testing about ρ_s. However, one argument for the use of ρ_s is that we are still inferring information about ρ_s by using r_s.

For the purposes of this qualification, the use of ρ_s could gain full marks (in line with the mark scheme) where used appropriately.

We use the standard notation for the null and alternative hypotheses, H_0 and H_1 respectively.

10a. Introduction to Hypothesis Testing: Terminology (8.1) (8.2) (8.3)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand the process of a hypothesis test.
- Understand the concepts of, and be able to formulate, a null hypothesis and an alternative hypothesis.
- Understand and use the concept of a significance level.
- Appreciate that hypotheses must be formulated prior to the evaluation of any statistics obtained from the sample.
- Identify the need to reach or state a clear conclusion about the results of a hypothesis test.

TEACHING POINTS

Students may find hypothesis tests confusing at first. As more hypothesis tests are introduced, students often fall into two categories: those who have understood the process and those who haven't. It is important to explain the fundamentals of hypothesis testing early to minimise this divide. Explain to students that they can infer information about the population from statistics obtained from samples. A hypothesis test is used to test the validity of these inferences.

Encourage students to follow a format for a hypothesis test. The notation used for the null and alternative hypotheses are H_0 and H_1 respectively. The following is standard:

- *Define the variable(s) being investigated.*
- *Formulate the null and alternative hypothesis.*
- *Declare the hypothesis test*
- *State the significance level* ***(Ofqual wish for this to be always 5% unless reasons for changing the level are part of the question.),***
- *State any chosen population parameters.*
- *Determine the critical value and/or region.*
- *Calculate the test statistic.*
- *Determine whether the test statistic is within the critical region.*
- *Decide whether the result is significant.*
- *Decide whether the null hypothesis is rejected or not.*
- *Interpret the findings of the hypothesis test in context.*

One particularly effective method is by linking the concept of a hypothesis test with a criminal trial.

- The null hypothesis is that there is no change or relationship (suspect is innocent until proven guilty).
- The alternative hypothesis is that there is a change or relationship (significant evidence to suggest the suspect is guilty).
- The significance level is the percentage chance of incorrectly rejecting the null hypothesis (the largest chance you will allow yourself to be wrong if you gave a guilty verdict).
- The test statistic is a statistic obtained from a sample in reference to the claim (the evidence).
- The critical value is the boundary of the critical region.
- The critical region is the range of values which result in rejecting the null hypothesis.
- The acceptance region is the range of values which result in not rejecting the null hypothesis.
- The p-value is the probability of the random variable taking the value equal to, or more extreme than, the test statistic.

Note that the "acceptance region" is a misnomer – a hypothesis test should **not** conclude to "accept H_0", but rather to "not reject H_0". Hypothesis tests should not be used to definitively conclude whether a particular hypothesis is true (which is what "accept H_0" is effectively saying). If one "does not reject H_0", we are concluding that "H_0 cannot be ruled out". It is good practice to use "Reject / Do not reject H_0" and mark schemes reflect this language too.

Make students aware of types of claim that may arise in the text of a question. For example, "there is a positive correlation in the population", "the population mean is smaller than 15.5", "the population proportion is not 0.46". This is important information for students to notice.
By use of a suitable diagram (the normal distribution curve is usually used here), it can be illustrated that the first two types of claim result in a critical region on one side of the curve (a one-tailed test). However, the latter claim would result in the critical region being split equally on both ends of the curve (a two-tailed test).

Laying the framework for a hypothesis test will allow students to refer back when carrying out many other hypothesis tests in future units. Contextual questions allowing students to identify the null and alternative hypotheses is usually a good starting point here.

Students must be aware that they shouldn't formulate hypotheses after the sample statistics have been analysed as this will lead to a biased investigation. Link this to the reality that a juror would always assume innocent until proven guilty, prior to seeing the evidence.

Explain to students that they are less likely to obtain a significant result with a lower significance level than with a higher one. Again, relating this to the courtroom analogy can help students understand why this is. The advantage of having a lower significance level is that there is a smaller chance of incorrectly rejecting the null hypothesis. The disadvantage of having a lower significance level is that there is a smaller chance of correctly rejecting the null hypothesis. Relating this to the courtroom analogy again helps (lower significance level means there is a smaller chance of giving a "guilty" verdict and being wrong, but a larger chance of giving a "not guilty" verdict and being wrong).

Finally, students need to be aware that any sample used to calculate the test statistic must be obtained randomly. A violation of this condition may result in the conclusions of the hypothesis test being invalid.

Exemplar

A researcher thinks that the heights of people are getting taller.
He collects a sample of 20 basketball players and carries out a hypothesis test about the mean at the 5% level.
He finds the results significant and concludes that the heights of the population are getting taller. Explain why his conclusion may not be valid.

The researcher has specifically chosen basketball players for his sample, and basketball players are usually a lot taller than average. Since the sample was not obtained from the population randomly, the results of the hypothesis test may not be valid.

OPPORTUNITIES FOR EMBEDDING THE SEC

A2 Formulating null and alternative hypotheses from the context of a situation.

A4 Awareness that it is not always obvious what question to investigate, so exploratory data analysis may be required.

A6 Awareness that bias can be introduced if the hypotheses are formulated after sample statistics have been analysed. Although it may seem that it won't make any difference since the data are already collected, it is bad practice and may influence the opinion of the person conducting the research.

D5 Interpreting the results of a hypothesis test in a language appropriate for a given target audience. Students must also appreciate that the general public do not understand statistical language, and so the results of any hypothesis test must be presented in a clear manner, in the context of the questions.

Exemplar

In an experiment on people's perception, each student in a sample of 100 university students was given a piece of paper which was blank except for a line 120 mm long. The students were asked to judge by eye the centre point of the line, and to mark it. Each student then measured the distance between the left hand end of the line and the mark made. If there were no bias in the students' perception of the centre of the line, the mean distance would be 60 mm.

a) **Define the null and alternative hypotheses.**
   $H_0: \mu = 60$ *mm*
   $H_1: \mu \neq 60$ *mm (no specific larger/smaller than 60mm claim)*
   where μ *is the population mean of the distance between the left of the line and the perceived centre of the line.*

According to previous studies, students tended to mark to the left of the centre more often than the right.

b) **From this information, define the null and alternative hypotheses.**
   $H_0: \mu = 60$ *mm*
   $H_1: \mu < 60$ *mm (claim students more tend to mark to the left - below 60mm)*
   where μ *is the population mean of the distance between the left of the line and the perceived centre of the line.*

From this current sample of 100 university students, it looked as though 95 of them marked more to the right of the centre than the left. One researcher suggests changing the alternative hypothesis to $H_1: \mu > 60$ mm.

c) **Explain what is wrong with this suggestion.**
   Bias is introduced by changing the alternative hypothesis to match the sample. We are now more likely to find a significant result, since we already know that the majority of the sample marked a cross to the right of the true centre. However, this may not be representative of the true population and so any significant result from this test should not be treated as significant.

COMMON AND POSSIBLE MISTAKES

There will be many mistakes made in hypothesis tests – these will be detailed in each corresponding sub-unit. The usual mistakes will be:

- formulating a null hypothesis as an inequality;
- not stating a conclusion in the context of the question;
- misidentifying a one or two-tailed test;
- not referring to the parameter in question (e.g. proportion, average);
- stating a definite conclusion (e.g. writing "there is no evidence…" as opposed to "there is insufficient / not significant evidence to suggest").

NOTES

Although Type I and Type II errors are not seen until Unit 19, it is important students see the advantages and disadvantages of using different significance levels.

10b. Introduction to Hypothesis Testing: Hypothesis Testing about the population PMCC (7.2)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Write, in detail, the process of a hypothesis test.
- Locate the critical values for the PMCC from the formula book.
- Carry out a hypothesis test about the PMCC.
- Understand the conditions for the results of a hypothesis test about the PMCC to be valid.
- Interpret the findings of a hypothesis test in context.

TEACHING POINTS

The previous sub-unit introduced students to the format of a hypothesis test, but not carry one out. This is the first opportunity for students to formally write out a hypothesis test and plenty of practice is key. It is advisable that many worked examples are seen before students attempt one solo.

It is also the first time the table of values in the formula book must be used – the recommended calculators cannot calculate critical values for the PMCC due to its difficulty (and students should appreciate this). Practise locating specific critical values in the tables prior to attempting a hypothesis test.

For example:

Find the critical value corresponding to a one-tailed hypothesis test at the 5% significance level on a sample of size 14.

Make students aware of the underlying assumptions of a hypothesis test about the PMCC: the underlying population must have a bivariate normal distribution. At this level, students are expected to know only basic properties of the bivariate normal distribution, namely:

- The independent variable (x) is normally distributed;
- The dependent variable (y) is normally distributed at each value of x;
- The relationship between x and y is “generally linear”
- The points on a scatter diagram exhibit an elliptical/oval shape.

Students must also appreciate that a violation of this condition can result in the conclusions of the hypothesis test being invalid.

Remind students of the definition of a null hypothesis, and appreciate that the null hypothesis is that there is no correlation (the population PMCC, ρ, is 0). The alternative hypothesis could either be positive correlation, negative correlation, or some correlation. Remind students that the population PMCC is denoted by ρ (rho) and should be defined

along with the hypotheses. The test statistic is the sample PMCC, r. Explain to students that the critical values are the level of correlation required to be "convinced" that there is a correlation within the population.

It is advisable that the first hypothesis test seen is testing for positive correlation, and students need to be able to understand that there are testing "one side", so "one-tailed".

Exemplar

A technician monitoring water purity believes that there is a relationship between the hardness of the water and its alkalinity.
Over a period of 10 days, the technician recorded the alkalinity and water hardness (in mg/l) and calculated $r = 0.9264$ to 4 decimal places.

a) **Test, at the 5% level the hypothesis that higher alkalinity is associated with higher water hardness, stating your conclusions clearly.**
$\mathrm{H}_0: \rho = 0$,
$\mathrm{H}_1: \rho > 0$, *(claim is 'higher water hardness')*
where ρ is the population PMCC between alkalinity and water hardness.

We will carry out a one-tailed hypothesis test about the PMCC at the 5% significance level with $n = 10$.

The test statistic is $r = 0.9264$
The critical region is anything bigger than ***or equal to*** 0.5494.

Since r is bigger than 0.5494, the result is significant.
We reject H_0.
There is significant evidence to suggest that there is a positive correlation between the alkalinity of the water and the hardness of the water.

b) **What assumption about the sample have you had to make in order to be able to carry out this hypothesis test?**
(see previous sub-unit) The sample must be random.

c) **What assumption about the population have you had to make in order to be able to carry out this hypothesis test?**
The populations of alkalinity levels of the water and hardness of the water follow a bivariate normal distribution.

In the above example, students may wish to write the statements of the hypothesis in words e.g.

H_0: *there is no correlation between the hardness of the water and its alkalinity,*
H_1: *there is a positive correlation between the hardness of the water and its alkalinity.*

Although the recommended calculators do not have this feature, some statistical software can calculate p-values for correlation and a p-value method may be used (see Unit 11).

When testing for negative correlation, ensure students realise that the absolute value of the critical value in the table is the same, but the sign is negative. Students could also see tests for "some correlation", and be able to identify that since it is unknown "which side" they are testing, they must test both sides and hence use a two-tailed test.

Some hypothesis tests could be seen where the test statistic isn't given, but the sample data allows students to calculate the test statistic.

Students need to also be aware that one advantage to testing for the population PMCC (compared with Spearman's rank) is that the actual data are used and it is more likely to detect correlation in the population (if any).

OPPORTUNITIES FOR EMBEDDING THE SEC

A2 Formulating a null and alternative hypothesis from context.

D2 Reaching conclusions from the results of a hypothesis test in the context of the question.

D3 Selecting an appropriate hypothesis test and determining if a result is statistically significant (this will become more relevant as the portfolio of hypothesis tests increases).

D5 Interpreting the results of a hypothesis test using language appropriate for a given target audience.

E2 Awareness that the underlying population must have a bivariate normal distribution, and that the sample must be obtained randomly.

At this early stage, allow students to familiarise themselves with hypothesis testing first. Embedding the SEC will become more frequent in subsequent units.

COMMON AND POSSIBLE MISTAKES

- Students may choose the null hypothesis as anything other than $\rho = 0$.
- Misidentifying a one-tailed test as a two-tailed test, or vice versa.
- Using the critical values from the Spearman's rank table.
- When stating the critical region, giving the acceptance region instead.
- Interpreting a result as significant when it isn't (or vice versa).
- Not writing a conclusion in the context of the question.
- Once the following sub-unit is taught, using the incorrect test or critical values from the table.

Students must remember that the test statistic and the critical value have the same sign in two-tailed tests.

NOTES

The use of "Accept H_1 or reject H_1" as a part of the conclusion is a divided issue. It is the opinion of the author that since the null hypothesis is the default opinion, you wouldn't "accept H_1", since it shouldn't have been a consideration in the first place. Equally, one wouldn't "reject H_1" since it was already "rejected" due to the definition of the null hypothesis.

The conclusion that is expected to be seen in examinations for full marks (in line with the mark scheme) are either "Reject H_0" or "Do not reject H_0"

Conclusions must **not be definite** and include 'there is significant evidence that …' or 'there is insufficient evidence to doubt …'

10c. Introduction to Hypothesis Testing: Hypothesis testing about the Spearman's Rank Correlation Coefficient (7.2)	Teaching time 2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Locate the critical values for Spearman's Rank correlation coefficients from the formula book.
- Carry out a hypothesis test about Spearman's rank correlation coefficient.
- Understand the conditions for the results of a hypothesis test about Spearman's rank correlation coefficient to be valid.
- Interpret the findings of a hypothesis test in context.

TEACHING POINTS

This sub-unit is very similar in style to the previous one. However, students will need to be redirected to the alternative table of values for Spearman's rank. Students need to be aware that, although the process is similar, the hypothesis tests are actually different and therefore require different critical values. Finding the values from both tables is one way to highlight the importance of reading:

For example:

Find the critical value for a two-tailed hypothesis test about Spearman's rank at the 5% level from a sample of size 12.

For example:

Find the critical value for a one-tailed hypothesis test about the PMCC at the 5% level from a sample of size 15.

The null hypothesis could either be "There is no association between…" or "$\rho_s = 0$". By "association", we are referring to a directional association (e.g. as one variable increases, the other variable increases) and not other (e.g. quadratic) associations. The alternative hypothesis may be "There is a positive association between…", "there is a negative association…" or "there is an association between…" (or the equivalent symbols involving ρ_s).

Remind students of the conditions for a hypothesis test about the PMCC to be valid, and made aware that this condition is not required for the hypothesis test about Spearman's rank (this is also an advantage).

Begin with hypothesis testing for a positive association, giving r_s in the question. Then introduce an example for negative association, reminding students that the absolute value of the critical values are the same except negative. Finally ensure students see a two-tailed test. It is advisable that students see every hypothesis in context, as this will

help with the SEC. To make things harder, give questions where students have to rank data and calculate r_s as part of the hypothesis test.

Exemplar

Seven cricketers were selected at random and their batting and bowling averages were recorded:

	Cricketer						
	A	**B**	**C**	**D**	**E**	**F**	**G**
Batting Average	22	51	51	38	60	5	8
Bowling Average	11	10	12	15	12	14	18

In cricket, a good batter attains a high batting average, the average number of runs per completed innings.
A good bowler attains a low bowling average, the average number of runs for each wicket taken over the season.

The captain of the local cricket team believed that good batter were good bowlers and that a team could be selected entirely on batting skill.
This belief was challenged, and the data above were collected in order to test the belief of the cricket team captain.
Test this belief at the 5% significance level using a Spearman's rank correlation coefficient.

H_0: *There is no association between batting average and bowling average (ranks).*
H_1*: There is a negative association between batting average and bowling average (ranks).*

We will carry out a one-tailed hypothesis test about the Spearman's rank correlation coefficient at the 5% level with $n = 7$.

The critical region is anything smaller than or equal to -0.7143.

Using a rank of 1 for the smallest values:

Cricketer	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Batting Average	*22*	*51*	*51*	*38*	*60*	*5*	*8*
Bowling Average	*11*	*10*	*12*	*15*	*12*	*14*	*18*
Batting Average Rank	*3*	*5.5*	*5.5*	*4*	*7*	*1*	*2*
Bowling Average Rank	*2*	*1*	*3.5*	*6*	*3.5*	*5*	*7*

The test statistic is $r_s = -0.5$.

Since r_s is not smaller than -0.7143, the result is not significant.
We do not reject H_0.
There isn't significant evidence to suggest that a good batter is necessarily a good bowler.

OPPORTUNITIES FOR EMBEDDING THE SEC

A2 Formulating a null and alternative hypothesis from the context.

A5 Ranking data for use in a hypothesis test.

D2 Reaching conclusions from the results of a hypothesis test in the context of the question.

D3 Selecting an appropriate hypothesis test and determining if a result is statistically significant. Now students have seen two hypothesis tests, they need to be able to select and justify their selection (e.g. underlying population is known to be bivariate normally distributed, association is known to be non-linear etc.)

D5 Interpreting the results of a hypothesis test in a language appropriate for a given target audience.

E2 Awareness that the underlying population does not have to have a bivariate normal distribution, but the sample must still be obtained randomly.

COMMON AND POSSIBLE MISTAKES

- Choosing the null hypothesis as anything other than "there is no association" or "$\rho_s = 0$".
- Misidentifying a one-tailed test as a two-tailed test, or vice versa.
- Using a hypothesis test for the PMCC instead of ranking the data and carrying out a hypothesis test for the Spearman's rank correlation coefficient.
- Using the critical values from the PMCC table.
- When stating the critical region, giving the acceptance region instead.
- Interpreting a result as significant when it isn't (or vice versa).
- Not writing a conclusion in the context of the question.
- Once the Wilcoxon rank-sum test (Unit 13c) is seen, students sometimes rank x and y values as a sample of size $2n$.

Students must ensure that the test statistic and the critical value have the **same sign** in **two-**tailed tests**.**

UNIT 11: Methods of Hypothesis Testing

Return to overview

SPECIFICATION REFERENCES

8.5 Conduct a statistical hypothesis test for the proportion in the binomial distribution and interpret the results in context using exact probabilities or, where appropriate, a normal approximation.

8.6 Conduct a statistical hypothesis test for the mean of a normal distribution with known or assumed variance, from a large sample, and interpret the results in context.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics
Binomial Distribution (Unit 5)
Normal Distribution (Unit 7)
Sampling Distribution of the Mean (Unit 9)
Normal Approximation to the Binomial (Unit 9)
Hypothesis testing (Unit 10)

KEYWORDS

acceptance region, alternative hypothesis, binomial distribution, critical region, critical value, evidence, hypothesis, insufficient, mean, normal approximation, normal distribution, null hypothesis, one-tailed test, parameter, proportion, *p*-value, region, significance level, significant, standard deviation, standard error, statistic, sufficient, suggestion, tail, test statistic, two-tailed test, variance,

UNIT SUMMARY

This unit continues the theme of hypothesis testing from the previous unit. It is presented as a separate unit for those wanting to teach these hypothesis tests prior to those for the correlation. In this case, Unit 10a could be inserted prior to Unit 11a.

Students could use the normal distribution mode on their calculators, exercising due diligence when inputting the population parameters. This should minimise the mistake of dividing by the standard deviation rather than the standard error when standardising.

Activities in Desmos that will help students understand the ideas behind these hypothesis tests: Binomial inference, *z*-test with critical regions.

The Statistical Enquiry Cycle, having been taught in bits in previous units, comes into its own here. Now students are getting more familiar with hypothesis testing, they can start

to appreciate the cycle and questions simulating a statistical investigation can be presented:

Exemplar

The keepers of a lighthouse were required to keep records of weather conditions.

The lighthouse is no longer manned, but a coastguard has a theory that the air has become clearer and so visibility has increased.
To test this theory the coastguard decides to record the visibility, in nautical miles, from the lighthouse around mid-day for a sample of 20 days during the following 3 months.

a) **State and describe a suitable sampling method the coastguard could use to collect the data.**
   - *A suitable sampling method could be taking a simple random sample.*
   - *Number the days from 1 to n, where n is the number of days in the three month period.*
   - *Use a random number generator to generate 20 random numbers in the range 1 to n, ignoring any repeats.*
   - *Select the days which are allocated the numbers generated.*
   - *Record the visibility around mid-day on the selected days.*

b) **Give an advantage and a disadvantage of your chosen sampling method.**
   - *An advantage is that the sample obtained is a random sample and the sampling method is unbiased.*
   - *A disadvantage is that it may not be practical to travel to the lighthouse during those particular days.*

(Note that for parts a) and b), any appropriate sampling technique can be used, and does not have to be random.)

c) **Other than a change in the sampling method, give one other way the coastguard could reduce bias in their sample.**
   The coastguard can ensure that the same person takes the readings, using the same method. at exactly mid-day during every day that the weather conditions are to be recorded.

After collecting his data, the results, in nautical miles, were as follows:

35	**21**	**12**	**7**	**2**	**14**	**18**	**20**	**16**	**11**
8	**8**	**5**	**11**	**28**	**35**	**16**	**35**	**9**	**17**

d) **Calculate the mean and standard deviation of this sample.**
   Using the calculator: $\bar{x} = 16.4$ *nautical miles and* $s_x = 10.0$ *nautical miles.*

Analysis of data over many years showed that the visibility at mid-day had a mean value of 14 nautical miles with standard deviation 9.4 nautical miles.

e) Assuming that the visibility can be modelled by a normal distribution with standard deviation 9.4 nautical miles, test the coastguard's theory at the 5% significance level.

Let X be the visibility, in nautical miles, from the lighthouse. So $X \sim N(\mu, 9.4^2)$.

$H_0: \mu = 14$,

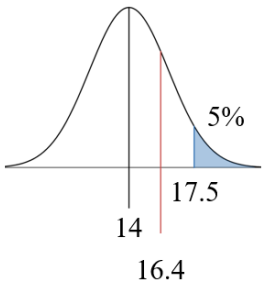
$H_1: \mu > 14$, *(theory is that visibility increased)*

where μ is the population mean of the visibility from the lighthouse in nautical miles.

We will carry out a one-tailed z-test at the 5% significance level.

Method 1: Using non-standardised critical regions

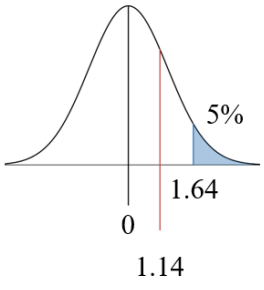
For a sample of 20, we have $\bar{X} \sim N\left(14, \frac{9.4^2}{20}\right)$.



Using the calculator, the critical region is $\bar{x} \geq 17.46$. *The test statistic is* 16.4.
Since $16.4 < 17.46$, *the result is not significant. We do not reject* H_0.

Method 2: Using standardised critical regions

Using $Z \sim N(0,1)$, *for a sample of* 20, *the test statistic is* $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{16.4-14}{\frac{9.4}{\sqrt{20}}} = 1.14$.



Using the calculator, the critical region is $z \geq 1.64$.
Since $1.14 < 1.64$, *the result is not significant. We do not reject* H_0.

Method 3: Using p-values

Using $\overline{X} \sim N\left(14, \frac{9.4^2}{20}\right)$, $P(\overline{X} > 16.4) = 0.126$

(or using $Z \sim N(0,1)$, *the test statistic is* $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{16.4-14}{\frac{9.4}{\sqrt{20}}} = 1.14. P(Z > 1.14) = 0.126$*)*

The one-tailed p-value is 0.126.
Since $0.126 > 0.05$, *the result is not significant. We do not reject* H_0.

There isn't sufficient evidence at the 5% significance level to suggest that the mean visibility has increased.

f) Explain whether or not the sampling method chosen in part (a) affects the findings of your hypothesis test in part (e).
Since the sample obtained in part (a) is a random sample, this shouldn't affect the conclusions of the hypothesis test.

Part f) is dependent on whether the sampling method chosen in part a) was random or not, and students need to be able to justify this.

A1 Identifying the factors (visibility) that relate to the problem (air clarity).

A2 Defining initial questions (has the air become clearer?), null and alternative hypotheses (part (e)) that can be investigated.

A3 Describing a suitable sampling method and data collection strategy (parts (a) and (b)).

A4 Using exploratory data analysis to determine population parameters (preamble prior to part (e)).

A5 Determining an appropriate distribution to model the underlying population in order to select an appropriate hypothesis test (not so much in this example).

A6 Justifying selections with reference to steps taken to reduce bias (part (c)).

B1 Identifying disadvantages to the data collection method chosen (part (b)).

B2 Identifying dangers when using secondary data sources (not in this example).

B3 Declaring the data collection methodology and appreciating why it has been declared.

B4 Identifying if data collection methods (e.g. questionnaires) are inherently biased (not in this example).

C1 Calculating numerical measures (part (d)) by hand or using technology.

C2 Calculating numerical measures using summary statistics or graphical representations generated by technology (not in this example).

C3 Identifying possible areas of misrepresentation (e.g. non-random sampling methods, outliers, bias).

D1 Interpreting numerical measures in context (part (e)).

D2 Reaching conclusions from the results of the hypothesis test (part (e)).

D3 Carrying out the selected hypothesis test and determining if a result is statistically significant (part (e)).

D4 Discussing the reliability, referring to sample methods and assumptions (part (f)).

D5 Making conclusions in a language appropriate to a given target audience (part (e)).

E1 Identifying weaknesses in the data collection methodology in relation to the method of data analysis (part (f)).

E2 Identifying the consequences of the weaknesses identified in **E1** in relation to the conclusions made in **D2** (part (f)).

E3 Making reasonable suggestions to overcome the weaknesses identified in **E1** to avoid the consequences in **E2** (not in this example, but could be added as a follow up).

E4 Carrying out the suggestions made in **E3** (again, not in this example but could be added as a follow up).

11a. Methods of Hypothesis Testing: Hypothesis Tests about a sample mean from a Normal Distribution with known variance (8.6)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Be able to carry out a hypothesis test about a sample mean from a normal distribution with known variance.
- Interpret the findings of a hypothesis test in context.
- Understand the use of a p-value.

TEACHING POINTS

Recap the normal distribution and the sampling mean of the normal distribution.

Recap the idea of a hypothesis test. In order to relate a critical value to the normal distribution, the use of the z-test with critical regions activity on Desmos may help here. Show students that the critical region can be represented on the graph of the normal distribution, and the area of the region should be equal to the significance level. Encourage students to use a sketch when carrying out this hypothesis test.

Due to the available technology, there are now multiple ways to carry out this test: using non-standardised critical regions (using the $\overline{X}$ distribution and the test statistic $\overline{x}$), using standardised critical regions (using the Z distribution and the test statistic $\frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$) or using *p*-values (using either distribution with the corresponding test statistics). Calculator methods are used throughout this sub-unit. Students may refer to this hypothesis test as a *z*-test.

The format of this hypothesis test should follow that of those previously seen. This time, the null hypothesis is "$\mu =$" where μ is the population mean. The alternative hypotheses could be μ greater than, less than or not equal to (the first two are one-tailed and the latter is two tailed). Emphasise to students that in a two-tailed test, the critical regions must still have a total area equalling the significance level – again a graph will help cement this idea.

Emphasise that since the random variable being tested is the sample mean, the sampling distribution of the mean should be used (either directly or via standardisation). Equally, students must appreciate that the population distribution must also be normal (since this is a condition for the sampling distribution of the mean to be normal).

The critical values may be found using a sketch and the inverse normal function on the calculator.

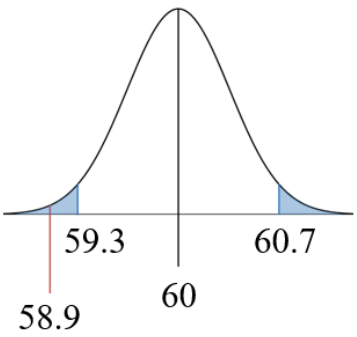
Exemplar

In an experiment on perception, each student in a sample of 100 university students was given a piece of paper which was blank except for a line 120 mm long.
The students were asked to judge by eye the centre point of the line, and to mark that centre point.
Each student then measured the distance between the left-hand end of the line and their mark. Their mean distance was 58.9 mm.
It is known from previous research that $\sigma = 3.71$ mm.

a) **If there were no bias in the students' perception of the centre of the line, the mean distance would be 60 mm.**
Determine whether there is significant evidence, at the 5% level, of any overall bias in the students' perception of the centre of a line.
State any assumptions you have made about the population.
Let X be the distance from the left hand end of the line to the mark made in mm.
We will assume that X is normally distributed.
$H_0: \mu = 60$ *mm,*
$H_1: \mu \neq 60$ *mm, (bias or not)*
where μ is the population mean distance from the left hand end of the line to the mark made.

We will carry out a two-tailed hypothesis test for the mean of a normal distribution at the 5% significance level.

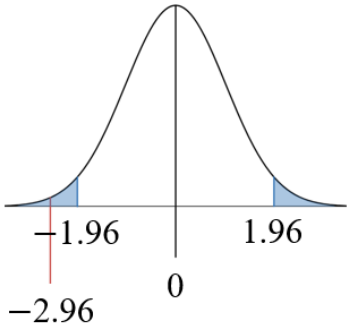
Method 1: Using non-standardised critical regions
Using $\bar{X} \sim \mathrm{N}\left(60, \frac{3.71^2}{100}\right)$*:*



The critical values are 59.3 *mm and* 60.7 *mm so the critical region is* $\bar{X} \leq 59.3$ *or* $\bar{X} \geq 60.7$*.*
The test statistic is 58.9, which is inside the critical region.

Method 2: Using standardised critical regions

Using $Z \sim \mathrm{N}(0,1)$*:*



The critical values are ± 1.96 *mm so the critical region is* $Z \leq -1.96$ *or* $Z \geq 1.96$*.*
The test statistic is $\frac{58.9-60}{\frac{3.71}{\sqrt{100}}} = -2.96$*, which is inside the critical region.*

The result is significant, so we reject H_0*. There is significant evidence to suggest that there is some overall bias in the students' perception of the centre of a line.*

b) Is your conclusion affected by whether or not the sample was taken at random? If so, how?

If the sample was not taken at random, the conclusion may not be valid.

Once enough practice has been given to students, explain to them that this is not the only way of carrying out a hypothesis test. Rather than using the critical values, p-values can be used instead. Explain to students that the probability of a value lying inside the critical region is (for a *z*-test) always the significance level (a graph will help here). For a one-tailed test, the probability of the corresponding region for the test statistic is the p-value (show them on a graph), and illustrate that if the test statistic is inside the critical region then the p-value will be smaller than the significance level (the area is smaller).

Note that the definition of a *p*-value is "the probability of observing the test statistic, or more extreme, if the null hypothesis is true". For a one-tailed test and where *TS* is the test statistic, the *p*-value is $\mathrm{P}(\overline{X} \geq TS)$ (for a lower tailed test) or $\mathrm{P}(\overline{X} \leq TS)$ (for an upper tailed test). For a two-tailed test, the *p*-value is **double** the value of the smaller of $\mathrm{P}(\overline{X} \leq TS)$ and $\mathrm{P}(\overline{X} \geq TS)$. Statistical software and calculators also report two-tailed *p*-values in this way.

The traditional method of comparing a probability with half the significance level for a two-tailed test is equivalent to doubling the probability (hence finding the *p*-value) and comparing with the full significance level and could still access full marks (in line with the mark scheme) **unless** the question specifically asks for "the *p*-value".

The following is a third method to answering the previous example:

Exemplar

Method 3 – Using p-values

Using $\overline{X} \sim \mathrm{N}\left(60, \frac{3.71^2}{100}\right) : \mathrm{P}(\overline{X} \leq 58.9) = 0.00151$

OR

Using $Z \sim \mathrm{N}(0,1)$*, the test statistic is* $\frac{58.9-60}{\frac{3.71}{\sqrt{100}}} = -2.96.\ \mathrm{P}(Z \leq -2.96) = 0.00151.$

The two-tailed p-value is $2 \times 0.00151 = 0.00303 < 0.05$

The result is significant, so we reject H_0*. There is significant evidence to suggest that there is some overall bias in the students' perception of the centre of a line.*

[Note: comparing 0.00151 with 0.025 is a valid method in this instance and could gain full marks (in line with the mark scheme).]

It is important that students see all methods, since the method of p-values will be more beneficial in the following sub-unit. Students will not be required to know the advantages and disadvantages of the p-value method until Unit 19. At this stage, allow students to choose whichever method they wish.

OPPORTUNITIES FOR EMBEDDING THE SEC

See the unit summary.

COMMON AND POSSIBLE MISTAKES

- Using the distribution for X as opposed to $\bar{X}$.
- Although students may not be standardising in this test, inputting the correct population parameters into the calculator is still likely to be an issue. Ensure plenty of practice and stating the use of $\bar{X}$ each time.
- Forgetting to square root the sample size when inputting the distribution into the calculator.
- Misidentifying a one-tailed test with a two-tailed test.
- Forgetting to double the probability to find a two-tailed *p*-value / halve the significance levels when applying a two-tailed test.
- Using a null hypothesis as anything other than "$\mu =$".
- Comparing a test statistic with a significance level, comparing a *p*-value with a critical value.

Students should remember that:

- When calculating probabilities for the *p*-value method in a hypothesis test, always use $X \leq$ or $X \geq$ and reject H_0 if the *p*-value is less than significance level;
- the test statistic and the critical value have the same sign;
- compare test statistic with critical value;
- compare *p*-value with significance level.

Students must also remember that

- A sample cannot have a normal distribution – it is the population which has the normal distribution.
- If asked to state necessary assumptions for the test to be valid, students should state that the population must be normal, in the **context** of the question.

11b. Methods of Hypothesis Testing: Hypothesis tests about a proportion (8.5) (8.6)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Carry out a hypothesis test about a proportion using a binomial distribution, using critical values or *p*-values.
- Use a Normal approximation to carry out a hypothesis test about a proportion.
- Interpret the findings of a hypothesis test in context.

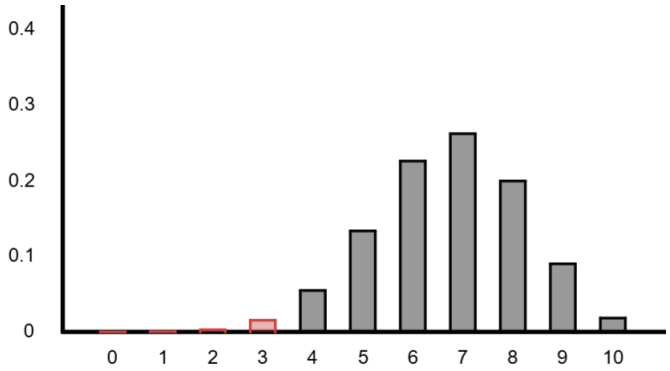
TEACHING POINTS

Recap the binomial distribution. Remind students that this is a discrete probability distribution. Encourage students to generate values for a binomial distribution and sketch corresponding vertical line charts.

To find critical values of the binomial distribution, encourage students to draw a sketch of the vertical line chart at the tails (no need to draw the whole distribution) and mark on when the cumulative probabilities reach the significance level. This mark should be between values, so students can see where the critical region is. The binomial inference activity on Desmos can be used and students can investigate using the software. The binomial cumulative distribution table function on the calculators is also very helpful but note that the Inverse Binomial function produces the value that gives a probability **closest to the significance level**, not necessarily that **below** (or equal to) the significance level as required.

Exemplar

Determine the critical region for $\mathbf{H_0}: X \sim \mathrm{B}(10, 0.67)$ for a lower one-tailed hypothesis test at the 5% significance level.

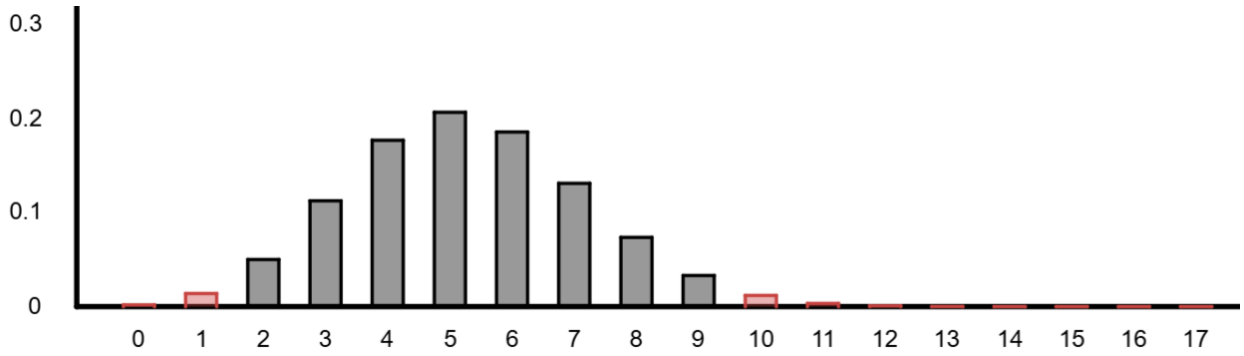


$\mathrm{P}(X \leq 3) = 0.0185$ *and* $\mathrm{P}(X \leq 4) = 0.0732$*, so the critical region is* $X \leq 3$ *since the probability*
$\mathrm{P}(X \leq 3) = 0.0185$ *is* ***below*** 0.05 *but* $\mathrm{P}(X \leq 4) = 0.0732$ *is* ***above*** 0.05.

*Note that is it **not possible** to obtain x such that* $P(X \leq x) = 0.05$ exactly because the binomial distribution is discrete.

Exemplar

Determine the critical region for $\mathbf{H_0}: X \sim \mathbf{B}(17, 0.31)$ for a two-tailed hypothesis test at the 5% significance level.



We want the probabilities of each part of the critical region to be 2.5% each.

For the lower tail, $\mathrm{P}(X \leq 1) = 0.0157$ *and* $\mathrm{P}(X \leq 2) = 0.0657$*, so the lower critical region is* $X \leq 1$ *since the probability* $\mathrm{P}(X \leq 1) = 0.0157$ *is **below*** 0.025 *but* $\mathrm{P}(X \leq 2) = 0.0657$ *is **above*** 0.025.

For the upper tail, $\mathrm{P}(X \geq 9) = 1 - \mathrm{P}(X \leq 8) = 1 - 0.9508 = 0.0492$ *and*

$$\mathrm{P}(X \geq 10) = 1 - \mathrm{P}(X \leq 9) = 1 - 0.9838 = 0.0162,$$

so the upper critical region is $X \geq 10$ *since the probability* $\mathrm{P}(X \geq 10) = 0.0162$ *is **below*** 0.025 *but* $\mathrm{P}(X \geq 10) = 0.0657$ *is **above*** 0.025.

As seen in this example, determining upper tail critical regions are harder for students. Students may, instead, prefer to consider when everything below the upper tail is at least 97.5% and find the upper tail critical value in this way.

By this time, it is advisable that students are familiar with the concept and process of a hypothesis test. This one will be more difficult due to the discrete nature of the distribution. However, a similar tack to the previous sub-unit would suffice.

The null hypothesis is "$\pi =$" (or "p=" will be condoned) where π (or p) is the population proportion being tested. This scheme of work will use π. The alternative hypotheses could be π is greater than, less than, or not equal to that value. Again, remind students that for a two-tailed test, the critical region must be split equally between each tail. Emphasise that it is good practice to define π in words so it is clear what is being tested.

The test statistic is the observed number of successes in the sample. A number line may be used to illustrate the critical region.

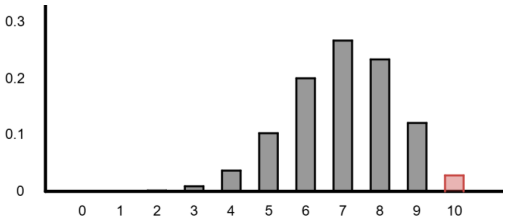
Exemplar

The probability that a certain type of seed germinates is 0.7. The seeds are stored at a different temperature, and when a packet of 10 seeds is tested, 9 germinate. Is this evidence at the 5% level of an increase in germination rate?

Let X be the number of seeds which germinate under the new temperature.

$\mathrm{H}_0: \pi = 0.7,$
$\mathrm{H}_1: \pi > 0.7,$
where π is the proportion of the population of seeds which germinate under the new temperature.

A one-tailed hypothesis test about the proportion at the 5% significance level, using $X \sim \mathrm{B}(10,0.7)$.



$\mathrm{P}(X \geq 9) = 1 - \mathrm{P}(X \leq 8) = 1 - 0.8506 = 0.1494$, *and*
$\mathrm{P}(X \geq 10) = 1 - \mathrm{P}(X \leq 9) = 1 - 0.9717 = 0.0283$
So the critical region is $X \geq 10$ *(since* 10 *is the largest value* X *can take, the critical region is* $X = 10$*).*

The test statistic is 9, which is not in the critical region. The result is not significant, so we do not reject H_0*. There is insufficient evidence to suggest that there is an increase in germination rate under this new temperature.*

Some students may find using the critical value method quite difficult, and so p-values can be used instead:

The significance level is 5%=0.05.

Since $\mathrm{P}(X \geq 9) = 0.1494 > 0.05$*, the result is not significant, so we do not reject* H_0*. There is insufficient evidence to suggest that there is an increase in germination rate under this new temperature.*

Again, students do not need to be aware of the advantages or disadvantages of the two methods until Unit 19, however if both methods are seen at this stage, most students will have already learnt through experience.

Normal approximation method

It is possible to apply a normal approximation to the binomial in order to carry out a hypothesis test about the proportion.

First recap the normal approximation to the binomial, and remind students about continuity corrections. Ensure students define their variables. Then emphasise that the approximation has a normal distribution, and the hypothesis test can now be carried out using the normal distribution instead.

There are multiple methods in carrying out a normal approximation within a hypothesis test, all of which could gain full marks (in line with the mark scheme) and are all exemplified below. One of the methods (using standardised proportions) uses the test statistic $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$ which may be found in the formula book.

Note that the lack of a continuity correction will be condoned and students who do not apply a continuity correction can still be awarded full marks (in line with the mark scheme). One reason for this is that, in practice, should the test statistic be so close to the critical value that the omission of a continuity correction would change the result of the test, the error could also be attributed to the use of the normal approximation itself. In practice, one would not apply the normal approximation and verify any results using an exact test.

Continuity corrections are generally not required when using methods involving proportions in hypothesis tests.

Exemplar

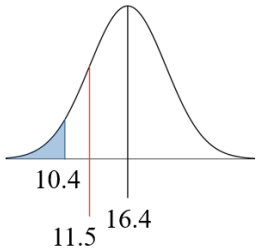
An insurance company sells life insurance and claims that at least 20% of their customers are vegan. An insurance broker looks over her week's sales and finds that only 11 out of 82 her sales were to vegans.
Using a normal approximation, test the insurance company's claim, using a 5% significance level.

Let X be the number of vegans who bought life insurance.

$\mathrm{H}_0: \pi = 0.2,$
$\mathrm{H}_1: \pi < 0.2,$
where π is the proportion of vegans who bought life insurance.

Carry out a one-tailed hypothesis test about the proportion at the 5% significance level, using $X \sim \mathrm{B}(82, 0.2)$.
Since n is large $n\pi \geq 5$ and $n - n\pi \geq 5$, (where π is the population proportion and *n* is the sample size) *a normal approximation can be used instead, so* $X \approx Y \sim \mathrm{N}(16.4, 13.12)$.

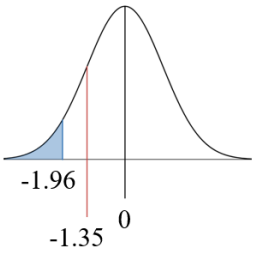
Method 1: Non-standardised critical regions with frequencies



The critical region is $Y \leq 10.4$

After applying a continuity correction, the test statistic is 11.5, (or test statistic = 11) which is not in the critical region.

Method 2: Standardised critical regions with frequencies



The critical region is $Y \leq -1.64$

After applying a continuity correction, the test statistic is $\frac{11.5-16.4}{\sqrt{13.12}} = -1.35$*, (or* -1.49 *with no CC) which is not in the critical region.*

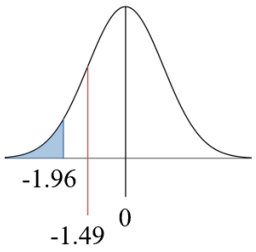
Method 3: p-values with frequencies

After applying a continuity correction, the test statistic is 11.5, (or test statistic = 11).

$\mathrm{P}(Y \leq 11.5)$ (or $\mathrm{P}(Z \leq -1.35)) = 0.088$), *so the one-tailed p-value is 0.088 > 0.05*

(Note: $P(Y \leq 11)$ *or* $P(X \leq -1.49) = 0.068$ *for no CC)*

Method 4: Standardised critical regions with proportions



The critical region is $Y \leq -1.64$

Since $\hat{p} = \frac{11}{82}$*, the test statistic is* $\frac{\frac{11}{82}-0.2}{\sqrt{\frac{(0.2)(1-0.2)}{82}}} = -1.49$*, which is not in the critical region.*

Method 5: p-values with proportions

Since $\hat{p} = \frac{11}{82}$*, the test statistic is* $\frac{\frac{11}{82}-0.2}{\sqrt{\frac{(0.2)(1-0.2)}{82}}} = -1.49.$

$\mathrm{P}(Z \leq -1.49) = 0.068$*, so the one-tailed p-value is 0.068 > 0.05*

The result is not significant, so we do not reject H_0*.*
There is insufficient evidence to suggest that the proportion of vegan customers is lower than 20%.

OPPORTUNITIES FOR EMBEDDING THE SEC

In addition to the overarching theme of the SEC as described in the unit summary:

E2 Instances where no critical region can be found due to the discrete nature of the binomial distribution is advised to be seen. Emphasise this as a limitation of the sample size, and a larger sample would be required to carry out a hypothesis test. Emphasise to students that in these situations, you cannot just simply change the level of significance since the sample has already been analysed.

Exemplar

During busy periods at a call centre, callers either get through to an operator immediately or are put on hold. A large survey revealed that 20% of callers were put on hold.

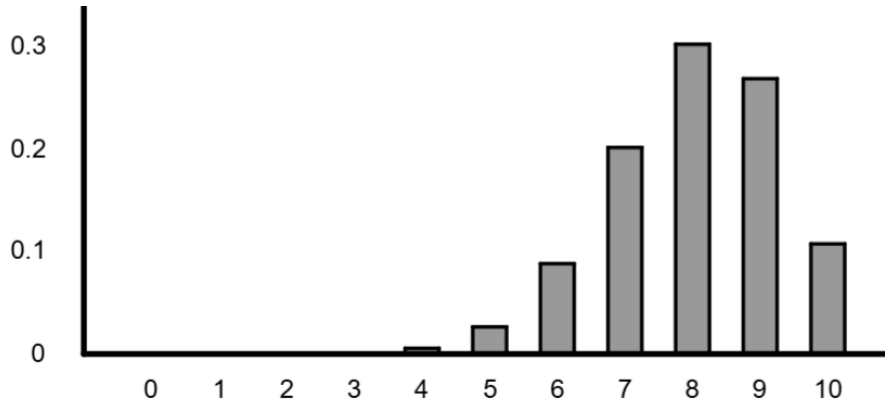
The call centre increases the number of operators with the intention of reducing the proportion of callers who are put on hold. A hypothesis test is carried out at the 10% level, to examine whether the centre has been successful in increasing the proportion of callers to get through immediately.

After the change, a random sample of 10 callers is taken and the number who get through immediately is recorded.
By considering the critical region, what conclusions can you draw and how would you suggest improving the hypothesis test?

Let X be the number of callers who get through immediately.

$\mathrm{H}_0: \pi = 0.8$,
$\mathrm{H}_1: \pi > 0.8$,
where π is the proportion of the population of callers who get through immediately.

We would carry out a one-tailed hypothesis test for the proportion at the 10% significance level, using $X \sim \mathrm{B}(10, 0.8)$.



Since $\mathrm{P}(X \geq 10) = \mathrm{P}(X = 10) = 0.1074 > 0.1$, there is no critical region. This means that the hypothesis test would always result in never rejecting H_0. This indicates that the sample taken was too small, and the hypothesis test should be carried out again using a larger sample.

COMMON AND POSSIBLE MISTAKES

In addition to all of the mistakes as stated in Unit 11b:

- Using a normal approximation when it is not valid.
- Misidentifying the upper tail critical values.
- When using p-values, concluding a result is significant if the p-value is larger than the significance level (especially for upper tail tests).

Students must remember that a sample cannot have a distribution – it is the population which has the probability distribution.

NOTES

Extension for information

It is possible to carry out this hypothesis test using the sampling distribution of a proportion (see Unit 9c for details) without standardisation.

Following the above example about vegan customers of life insurance:

Extension Exemplar

$\mathrm{H}_0: \pi = 0.2,$
$\mathrm{H}_1: \pi < 0.2,$
where p is the proportion of the population of vegan customers who bought life insurance.

We will carry out a one-tailed hypothesis test about the proportion at the 5% significance level, using
$\frac{X}{n} \sim \mathrm{N}\left(0.2, \frac{0.2(1-0.2)}{82}\right) = \mathrm{N}\left(0.2, \frac{2}{1025}\right)$*, since* n *is large.*

The critical region is $\frac{X}{n} \leq 0.1273$.

The test statistic is $\frac{11.5}{82} = 0.140$ ($\frac{11}{82} = 0.134$ *without continuity correction*) *which is not in the critical region. (either could gain full marks* (in line with the mark scheme)*)*

The result is not significant, so we do not reject H_0*. There is insufficient evidence to suggest that the proportion of vegan customers is lower than 20%.*

UNIT 12: Contingency Tables

Return to overview

SPECIFICATION REFERENCES

9.1 Construct contingency tables from observed data, combining data where appropriate, and interpret results in context.

9.2 Use a χ^2 test with the appropriate number of degrees of freedom to test for independence in a contingency table and interpret the results of such a test.

9.3 Know that expected frequencies must be greater than, or equal to, 5 for a χ^2 test to be carried out and understand the requirement for combining classes if that is not the case.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics

Hypothesis testing (Unit 10)

Probability - independent events (Unit 5)

GCSE (9-1) in Mathematics at Higher Tier

S2 Two way tables.

KEYWORDS

association, chi-squared, column, combine, contingency tables, contribution, critical region, critical value, expected, grand total, hypothesis, independent, observed, one-tailed, pool, row, source, strata, test statistic, total, upper tail,

UNIT SUMMARY

This topic introduces the idea of contingency tables and hypothesis testing for association between two qualitative random variables. These random variables are usually nominal (categories with no inherent order) but may be ordinal (categories with an order). This is the first time students are introduced to the χ^2 distribution, which is a distribution seen in other A level subjects such as Biology, Psychology or Geography. Students do not need to know the mathematics behind the χ^2 distribution other than how to use it within a hypothesis test. Even the more mathematically able would find it difficult to grasp, say, the probability density function of a χ^2 variable.

The use of the χ^2 graph on Desmos together with critical regions and the significance levels will help during teaching.

12a. Contingency Tables: Introduction to Contingency Tables (9.1) (9.2)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Appreciate the use of a two-way table in contingency table analysis.
- Calculate expected values of a contingency table.

TEACHING POINTS

Revise two-way tables from GCSE (or Unit 2). Also revise independent events and the multiplication rule for independent events from Unit 5.

Students need to know how to use the formula $\frac{\text{row total} \times \text{column total}}{\text{grand total}}$ to calculate expected frequencies of a contingency table. Students must be aware of what an expected frequency is: the frequency you would expect if the two random variables were independent of each other. This can be explained using the multiplication rule for probability: If two random variables X and Y are independent, then $\mathrm{P}(X)\mathrm{P}(Y) = \mathrm{P}(X \cap Y)$ for all values of X and Y. Use a contingency table to explain this.

		Y			
		Column 1	**Column 2**	**Column 3**	**Row Total**
X	**Row 1**				
	Row 2				
	Column Total				**(grand total)**

The probability of choosing a row is $\frac{\text{row total}}{\text{grand total}}$ and the probability of choosing a column is $\frac{\text{column total}}{\text{grand total}}$. If the random variables are independent, then the probability of choosing a cell in that row and that column would be $\frac{\text{row total} \times \text{column total}}{\text{grand total} \times \text{grand total}}$. However, this gives the probability – so to get the expected frequency, you multiply by the grand total, resulting in the required formula.

Some tables may be given in percentages – ensure students convert the percentages into frequencies. This is for later on when they carry out the hypothesis test. The use of the spreadsheet mode on the calculator will be useful, especially to check answers. Students need to be able to locate the “sum”, “:” and the letter symbols on their calculator in order to make use of this.

You could provide questions without context in order for students to practise calculating the expected frequencies. At this stage, do not worry about small expected frequencies.

Encourage students to calculate $(m-1)\times(n-1)$ cells from an $m\times n$ contingency table, and calculate the remaining cells using the row/column totals. This will help to explain degrees of freedom in the following sub-unit.

The layout of a contingency table varies due to teacher preference. Some have favoured the use of one contingency table showing both observed and expected frequencies:

	Column 1		**Column 2**		**Column 3**		**Total**
	O	E	O	E	O	E	
Row 1							**Total Row 1**
Row 2							**Total Row 2**
Total	**Total Col 1**		**Total Col 2**		**Total Col 3**		**Grand Total**

The alternative method of displaying the data is to have two two-way tables: one for the observed frequencies and one for the expected frequencies, both clearly labelled.

Observed	**Column 1**	**Column 2**	**Column 3**	**Total**
Row 1				**Total Row 1**
Row 2				**Total Row 2**
Total	**Total Col 1**	**Total Col 2**	**Total Col 3**	**Grand Total**

Expected	**Column 1**	**Column 2**	**Column 3**	**Total**
Row 1				**Total Row 1**
Row 2				**Total Row 2**
Total	**Total Col 1**	**Total Col 2**	**Total Col 3**	**Grand Total**

Show both methods and allow students to pick their preferred method.

Students need to be aware that a contingency table is used to display frequencies relating to the values of two random variables. Students may be required to construct a contingency table from given data or to identify data from a table to be extracted and put into a contingency table.

The SEC can be embedded into questions with context (see below).

OPPORTUNITIES FOR EMBEDDING THE SEC

In addition to describing possible sampling methods to obtain data:

A1 Identifying when the problem under investigation relates to two random variables and the frequencies at which they occur together.

A3 Identifying that a contingency table is an appropriate way of recording these data.

A4 A contingency table for a small sample could be used and the expected frequencies compared with the observed values in order to determine if there could be an association.

A5 Describing the process of analysing the data within a contingency table.

C1 Appreciating that a spreadsheet can be used to analyse a contingency table.

D1 Analysing the data within a contingency table.

D2 Appreciating and interpret the meaning of expected frequencies.

Exemplar

An investigation into colour-blindness and the sex of a person gave the following results:

	Colour-blind	**Not Colour-blind**
Male	36	964
Female	19	981

a) Describe a suitable sampling method that could have been used to obtain these data.

- *A stratified sample could be used here.*
- *Using a census, male members of a population can be assigned a number from 1 upwards.*
- *Using a random number generator, select the first 1000 people assigned the first 1000 random numbers generated (ignoring repeats).*
- *Repeat the process for the female members of the population.*
- *Once selected, ask each member whether they are colour-blind or not.*
- *Record these data together with the gender of the sample member.*

b) Explain why a contingency table can be used to represent these data.

We are trying to find out how many of each sex are colour-blind. Since these are frequencies relating to the values of two variables, a contingency table can be used.

c) Determine the expected values of each cell, assuming the gender and the colour blindness of a person are independent from each other.
(This information was obtained using the calculator)

Expected	Colour blind	Not Colour blind	Total
Male	27.5	972.5	**1000**
Female	27.5	972.5	**1000**
Total	**55**	**1945**	**2000**

d) By considering the expected frequency of colour-blind females, what conclusion might you make?
If the gender of the person and whether or not they are colour blind were independent of each other, the expected number of colour blind females would be 27.5, which is much higher than that observed (19).
This suggests that there may be an association between gender and colour blindness, and may suggest that males may be more likely to be colour-blind.

COMMON AND POSSIBLE MISTAKES

- If the calculator spreadsheet mode is used, inputting the incorrect cells when attempted to calculate expected values.
- Due to rounding errors, students may find that rows and columns do not add up to the correct total. If students have given answers to an appropriate degree of accuracy (e.g. 2 decimal places), they should be awarded full marks (in line with the mark scheme) if they are ± 0.02.

12b. Contingency Tables: Hypothesis tests for association between two variables (9.2)

Teaching time
2 hours

OBJECTIVES

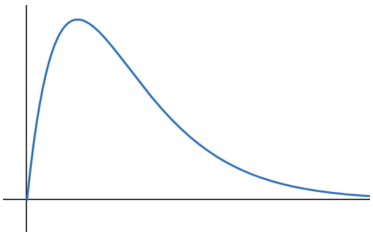
By the end of the sub-unit, students should be able to:

- Calculate $\frac{(O-E)^2}{E}$ for each cell of a contingency table and calculate the test statistic χ^2
- Determine the number of degrees of freedom from a contingency table
- Carry out a hypothesis test for association between two variables using a contingency table and the χ^2 distribution
- Understand the conditions when a hypothesis test using contingency tables can be used
- Combine cells of a contingency table in an appropriate way
- Identify, from given data, when use of a χ^2 test for association is appropriate

TEACHING POINTS

It is advisable that students appreciate that $\frac{(O-E)^2}{E}$ (where O is the observed frequency and E is the expected frequency) is the relative squared difference between the observed and expected values. It is also advisable that students appreciate that the total sum of these differences, $\chi^2 = \sum\frac{(O-E)^2}{E}$ approximately follows a special type of probability distribution, the χ^2 distribution. Use the χ^2 distribution graph on Desmos to illustrate this. It might help students if they are told that $\sum\frac{(O-E)^2}{E}$ is only approximately distributed with a χ^2 distribution, and approximation is not good if $E < 5$. So emphasise that this is only true for $E \geq 5$.

The χ^2 distribution shape is given in the formula book:



First, allow students to practise calculating the test statistic. Some people have used a combined table showing O, E and $\frac{(O-E)^2}{E}$ for each cell. However, some students find this confusing and have preferred the use of three separate tables. You may want to show them different layouts and allow students to choose. Questions without context can be used here.

The reason why it is important that students calculate only $(m-1)\times(n-1)$ expected values from an $m \times n$ contingency table, and using the totals to calculate the final cells

is because it is easier to explain the terms “degrees of freedom”. Explain to students that when the data are collected, the total number of people asked within each of the strata was already determined (see Unit 8). The distribution of frequencies within those strata can be chosen arbitrarily, until the last cell which is uniquely determined from the total. Students can then see that out of mn cells, only $(m-1)(n-1)$ were freely chosen and the final cells were already determined. Hence the number of degrees of freedom for a contingency table is $(m-1)(n-1)$.

Explain that a population parameter of the χ^2 distribution is the number of degrees of freedom, which is dependent on the sample size. Relate this to Unit 9 where the sampling distribution of the mean changed shape as the sample size changed (you may want to use the normal distribution activity on Desmos to revise this quickly). The χ^2 distribution graph on Desmos allows you to change the number of degrees of freedom and thus the shape of the graph. Explain that the test statistic shows the total relative squared error between the observed and expected frequencies, so a higher contribution would imply more association. Make this link with the fact that a χ^2 test is always one-tailed on the upper end. The standard symbol for the degrees of freedom is ν (nu).

It is beneficial for students to have some practice at locating the critical values of a χ^2 distribution from the formula book. Reiterate that the tail required is the upper tail and the test is always one-tailed. Also explain that the χ^2 distribution is difficult to deal with (like the sample PMCC distribution) and so p-values are only obtainable using graphical calculators or statistical software.

Now that all individual aspects of the χ^2 test have been practised, the full hypothesis test can be carried out. The null hypothesis is “*<variable 1> and <variable 2> are independent of each other”* or “*There is no association between…*”. The alternative hypothesis is the complement of these statements. Encourage students to use full sentences when defining their hypotheses, since there are no symbols to be used.

At this point, encourage students to find the expected frequencies before declaring their hypothesis test. This is to cover the scenarios where cells must be grouped in order to ensure the expected frequencies are large enough and the number of degrees of freedom could potentially change.

It is advisable that students have carried out enough hypothesis testing by now to be able to write out the test out formally. Begin with contingency tables that require no grouping of cells.

Once students have practised this, introduce them to a scenario where an expected frequency is smaller than 5. Allow students to discuss the most appropriate grouping, including the advantages and disadvantages behind the grouping.

OPPORTUNITIES FOR EMBEDDING THE SEC

The entirety of the SEC could be covered here, from collecting data through to analysing data and using the hypothesis test. Students need to be able to interpret the results of the hypothesis test.

Exemplar

The table below summarises the incidence of cerebral tumours in 129 neurosurgical patients.

		Type of tumour		
		benign	**malignant**	**Other non-malignant**
Site of tumour	**Frontal lobes**	23	9	6
	Temporal lobes	4	5	5
	elsewhere	34	24	19

(Other non-malignant tumours are ones which, while not malignant, still need to be removed eventually because if they were allowed to grow too much, they would interfere with normal brain function)

a) **Investigate, at the 5% significance level, whether the type of tumour is independent of the site within the brain.**
H_0: There is no association between the type of tumour and the site of the tumour.
H_1: There is an association between the type of tumour and the site of the tumour.

Expected		Type of tumour			
		benign	**malignant**	**Other non-malignant**	**Total**
Site of tumour	**Frontal lobes**	17.97	11.19	8.84	**38**
	Temporal lobes	6.62	4.12	3.26	**14**
	elsewhere	36.41	22.69	17.91	**77**
	Total	**61**	**38**	**30**	**129**

Because there are small expected frequencies, we must group some cells. Combine the "temporal lobes" row with "elsewhere" and form a new row "Elsewhere".

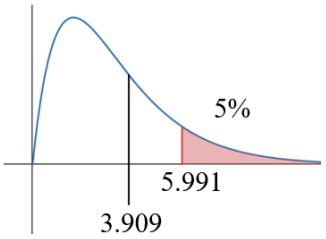
Observed		Type of tumour			
		benign	**malignant**	**Other non-malignant**	**Total**
Site of tumour	**Frontal lobes**	23	9	6	**38**
	Elsewhere	38	29	24	**91**
	Total	**61**	**38**	**30**	**129**

Expected		Type of tumour			
		benign	**malignant**	**Other non-malignant**	**Total**
Site of tumour	**Frontal lobes**	17.97	11.19	8.84	**38**
	Elsewhere	43.03	26.8	21.16	**91**
	Total	**61**	**38**	**30**	**129**

We will carry out a hypothesis test for association at the 5% significance level using the χ^2 distribution with $(3-1)(2-1)=2$ degrees of freedom.
The critical value is 5.991.

$\frac{(O-E)^2}{E}$		Type of tumour		
		benign	**malignant**	**Other non-malignant**
Site of tumour	**Frontal lobes**	1.418	0.429	0.912
	Elsewhere	0.588	0.181	0.381

The test statistic is $\chi^2 = 3.909$



Since $3.909 < 5.991$, the result is not significant. We do not reject H_0. There is insufficient evidence to suggest that there is any association between the type of tumour and the site of the tumour.

[Note: *p*-values obtained from a calculator can be utilised to complete this test and would access full marks (in line with the mark scheme)]

The investigation was extended to include data from a large number of hospitals in several European countries.
The regions of the brain were put into six categories (frontal lobes, temporal lobes and four others). A 3×6 contingency table was formed and the test statistic was calculated correctly as 21.7 with no grouping of cells being necessary.

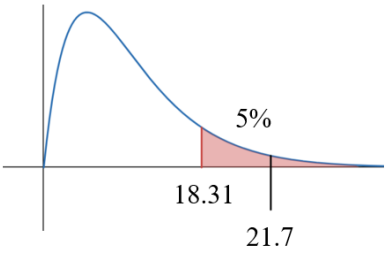
b) Investigate the hypothesis that the type of tumour is independent of the site

i. **using a 5% significance level**

H_0: *There is no association between the type of tumour and the site of the tumour.*
H_1: *There is an association between the type of tumour and the site of the tumour.*

We will carry out a hypothesis test for association at the 5% significance level using the χ^2 distribution with $(3-1)(6-1) = 10$ degrees of freedom.

The critical value is 18.31, the test statistic is 21.7.



Since $21.7 > 18.31$, the result is significant. We reject H_0.
There is significant evidence at the 5% level to suggest that there is an association between the type of tumour and the site where the tumour was located.

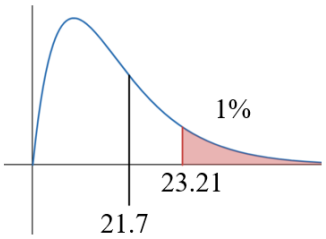
ii. using a 1% significance level

(as an example of a **different** significance level to the default 5%)

H_0: *There is no association between the type of tumour and the site of the tumour.*
H_1: *There is an association between the type of tumour and the site of the tumour.*

We will carry out a hypothesis test for association at the 1% significance level using the χ^2 distribution with $(3-1)(6-1) = 10$ degrees of freedom.

The critical value is 23.21, the test statistic is 21.7.



Since $21.7 < 23.21$*, the result is not significant. We do not reject* H_0*. There is insufficient evidence at the 1% level to suggest that there is an association between the type of tumour and the site the tumour was located.*

c) Compare and explain the conclusions you reached in parts (b) (i) and (ii).
If you are prepared to accept a risk of at most 5% of being incorrect, then you can conclude there is an association between the site of the tumour and the type of tumour. However, if you are prepared to accept a risk of at most 1% of incorrectly rejecting the H_0 *(this may be relevant as the consequences are great in a medical scenario) then you can conclude there is not an association between the site of the tumour and the type of tumour. Since the sample was already used for the hypothesis test at the 5% level, a new sample should be obtained for the hypothesis test at the 1%.*

Note: in part (a) above, students may combine “malignant” and “other non-malignant”, which is acceptable since all tumours in this grouping need to be treated.

COMMON AND POSSIBLE MISTAKES

- Not grouping cells together when there are small expected frequencies.
- Forgetting that the “bigger than 5” condition applies to expected frequencies, and not the observed frequencies or contributions to the test statistic.
- Grouping cells together in an inappropriate way. Using the critical value at the lower tail.
- Forgetting to square (O-E) in the formula.
- Not recording intermediate numbers to an appropriate degree of accuracy.
- Not using frequencies in the contingency table.

Remind students that in a hypothesis test they should always state the number of degrees of freedom.

In the past, other A level subjects which utilise the χ^2 test refer to it as a “two-tailed” test in the sense that it does not detect the direction of any association. Remind students that the test itself is one-tailed.

NOTES

Extension

For 2×2 contingency tables, Yates' correction $\sum \frac{(|O-E|-0.5)^2}{E}$ is used since it is a better approximation for χ^2 when there is only one degree of freedom. This can be explained using the χ^2 distribution graph on Desmos (the pdf for 1 degree of freedom has a different shape).

Yates' correction is **not** included in the specification and will not be assessed for 2×2 contingency tables and the test statistic of $\sum \frac{(O-E)^2}{E}$ is expected.

The spreadsheet mode on the calculator could be used, although spreadsheet formulas need not be used. Students utilising the cells to perform calculations will help them lay out their reasoning more clearly and minimise mistakes. In the past, students would have to calculate, write down the answer and repeat. The spreadsheet mode on the calculator will allow them to perform all calculations and have them stored in an easy-to-read layout, before transferring the results to paper.

12c. Contingency Tables: Context (9.1) (9.2) (9.3)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Identify possible sources of association from a frequency table.
- Interpret the contributions to the test statistic in context.

TEACHING POINTS

Students often find this final interpretation of contingency tables the most difficult, due to the level of communication and clarity required. A lot of the skill is “say what you see, say what it means”: students should treat each cell individually,

- stating the size of the contribution, and interpreting this value in context,
- comparing the expected and observed frequencies of the corresponding cells and interpreting this in context.

Students need only identify largest contributions to the test statistic. The results of the hypothesis test should be taken into account as well: if the result of the hypothesis test is not significant then any contribution to the test statistic that could have suggested a source of association should be treated as due to random error.

Start by practising the interpretation from completed contingency tables and hypothesis tests, to allow students to familiarise themselves with the level of language and interpretation required. Once this skill is mastered, students can start consolidating the entire unit (starting from a contingency table, through a hypothesis test and finishing with the interpretation).

Exemplar

During an investigation into the factors associated with socio-economic deprivation, it is claimed that there may be a link between smoking and deprived areas in the UK. The deprived areas are ranked according to the Index of Multiple Deprivation (IMD), and categorised as either Very Highly Deprived, Highly Deprived, Slightly Deprived and Least Deprived. A random sample of 500 people was taken across different areas and asked whether they were a smoker or not.

The observed results are shown in the table below:

Observed	Very Highly Deprived	Highly Deprived	Slightly Deprived	Least Deprived
Smoker	74	43	45	24
Non-Smoker	55	65	92	102

A statistician carries out a χ^2 test at the 5% significance level using the following two tables:

Expected	Very Highly Deprived	Highly Deprived	Slightly Deprived	Least Deprived
Smoker	47.988	40.176	50.964	46.872
Non-Smoker	81.012	67.824	86.036	79.128
$\frac{(O-E)^2}{E}$	**Very Highly Deprived**	**Highly Deprived**	**Slightly Deprived**	**Least Deprived**
Smoker	14.10	0.20	0.70	11.16
Non-Smoker	8.35	0.12	0.41	6.61

The statistician correctly concluded that the result was significant, and there was evidence to suggest the claim was correct.
By referring to the above tables, suggest possible sources of association between smoking and the level of deprivation.

- *There is a large contribution in the cells representing smokers in very highly deprived areas (14.1).*
- *The observed number of smokers (74) is much higher than the number expected (48) if there was no association between smoking and the level of deprivation.*
- *There is also a large contribution in the cell representing smokers in the least deprived areas.*
- *The observed number of smokers (24) is much lower than the number expected (47) if there was no association between smoking and the level of deprivation.*

OPPORTUNITIES FOR EMBEDDING THE SEC

Students need to be able to interpret the results of the hypothesis test, together with the appropriate numerical values calculated. This covers Stages **C** and **D** of the SEC. If a contingency table required a grouping of cells, resulting in the conclusion to be altered, students can suggest improvements to the process (for example, collecting a larger sample) which will also cover Stage E.

COMMON AND POSSIBLE MISTAKES

Students will often mix up the values they have spoken about. For example, a student may compare an observed frequency of 45 with an expected frequency of 55, but then say there were more than expected in that particular cell. This is rarely due to language, but more that they swapped the numbers around despite writing it down in the previous sentence. The usual mistake is to compare observed values with other observed values, rather than with expected values.

Students must be encouraged to quote numerical justification for their choice of association. Students who simply mention for each observed frequency whether it is greater/smaller than its corresponding expected frequency will not gain marks.

Students should **not** simply quote every difference between observed and expected values but should comment only on those with the largest contributions to the test statistic. Advise students to look at the mark allocation for these questions as a guide on how much to write.

UNIT 13: Non-Parametric Hypothesis Tests

Return to overview

SPECIFICATION REFERENCES

10.1 Use sign or Wilcoxon **signed-rank** tests to investigate population median in single sample tests and also to investigate for differences using a paired model.

10.2 Use the Wilcoxon **rank-sum** test to investigate for difference between independent samples.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics

Binomial Distribution (Unit 5)

Hypothesis testing (Unit 10a)

Normal distribution (Unit 7 - awareness only)

GCSE (9-1) in Mathematics at Higher Tier

A2 Substituting numbers into a formula.

KEYWORDS

alternative hypothesis, association, binomial, difference, distribution-free, hypothesis, independent, Mann-Whitney, median, non-parametric, null hypothesis, population, rank, sample, sign, significance level, Wilcoxon signed-rank, Wilcoxon rank-sum,

UNIT SUMMARY

This unit introduces non-parametric (distribution-free) hypothesis tests. This is where the normal distribution is not required in order to carry out the hypothesis tests. It is not necessary for students to have learnt the normal distribution by this point, although if they have it will give more motivation to why these tests are used (due to the restrictive nature of a population being normally distributed). In order to carry out the sign test, the binomial distribution will need to be used, with $p = 0.5$. The binomial inference graph in Desmos can help here.

The Wilcoxon Rank-sum test has been referred to as the Mann-Whitney *U* test in legacy specifications. The course will always refer to this test as the Wilcoxon Rank-sum test, but students can still be awarded full marks (in line with the mark scheme) if the name Mann-Whitney is used instead.

In legacy specifications, a generalisation of the Wilcoxon rank-sum test to three or more samples was used (the Kruskal-Wallis test, which is **not** included in the content). As an extension activity, students may want to learn how to carry out this useful hypothesis test. It will also give a non-parametric alternative to ANOVA in Unit 24, and will be a useful test to know in future careers.

13a. Non-Parametric Hypothesis Tests: One-Sample Sign Test (10.1)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand when a sign test can be used.
- Know the advantages and disadvantages of using a sign test.
- Carry out a sign test.
- Interpret the results of a sign test.

TEACHING POINTS

Revise the binomial distribution. Stick with $p = 0.5$.

Emphasise that sometimes the normal distribution cannot be assumed, or you already know the population is not normally distributed; in which case, the hypothesis test about the mean (Unit 11a) cannot be used. The sign test is not a test on a population mean, but rather on a **population median**.
The symbol η (eta) is used to denote the population median.

The sign test works by comparing sample members with the claimed median. If there is no change in the sample median, then you would expect there to be an equal number of sample values above the median as below. This is directly related to the binomial distribution, where a new random variable Y can be introduced where Y is the number of values above the median. If there is no change to the median, then the probability of a value being above the median is exactly 0.5. The number of trials is the sample size with any values equal to the claimed median removed.

The advantage of the sign test is that it is not necessary to assume a normal distribution, or any symmetry, from the underlying population. The disadvantage of the sign test is that it only tests about the population median and is not a very powerful test. As usual, the sample analysed must be obtained randomly in order for the results of the hypothesis test to be valid.

In this sub-unit, you may find it easier to teach the sign test through examples. A few worked examples will soon illustrate how the test works. You could use the binomial inference graph on Desmos to remind students how to carry out a hypothesis test about a proportion.

The null hypothesis is "$\eta =$" where η is the population median (the equivalent in words is also acceptable, but students must remember to refer to the **population** median). The alternative hypotheses could either be η is greater than, less than or not equal to. For two-tailed tests, students need to be reminded that the critical region areas must be split equally at either end. Using the binomial inference graph on Desmos will help remind students of this.

To calculate the test statistic, students need to count how many values are above the median. The easiest way is to assign each value a plus symbol (+) for all values above the median, and a minus symbol (−) for all values below the median. If the value is equal to the median, a 0 symbol can be assigned and not contribute to either count. The test statistic is the smaller of the number of plus signs or minus signs. Due to symmetry ($p = 0.5$) and the choice of the test statistic, only the critical region at the lower tail needs checking. This can be explained using the binomial inference graph on Desmos. Students can instead choose the larger of the number of plus or minus signs and consider the upper tail. Remind students that the probability should be doubled when using a two-tailed test with *p*-values.

Students can then be able to interpret the results of the sign test in context. Students need to be aware that it is the median that is being tested, and not the mean.

Although there is no motivation to do so, a sign test could be given with a large sample with a view to using the normal approximation to the binomial. In these cases, the test statistic could be calculated for students. This will be a precursor to the Central Limit Theorem (Unit 18c), although Year 1 students need not know this.

OPPORTUNITIES FOR EMBEDDING THE SEC

As with all hypothesis tests at this point, a question involving a full cycle of the SEC can be used.

Exemplar

A trading standards inspector visits a butcher who sells meat pies in her large shop.
The pies are supposed to contain at least 250 g of meat.
There have been complaints that the butcher does not put in enough meat (which is why the inspector went to her shop).

a) **Describe how the trading standards inspector could investigate whether or not the butcher puts in enough meat.**
   - *The inspector could number all of the pies in the shop from 1 to n.*
   - *Using a random number generator, select the pie corresponding to that random number.*
   - *Continue until a sample of an appropriate size (say 20) is selected, ignoring repeated random numbers.*
   - *The inspector could then weigh the meat in each pie in grams and carry out a hypothesis test to determine if the weight of the meat is lower than claimed.*

b) Give a reason why the inspector would take a sample, instead of investigating every pie.

There may be too many pies in the shop, and testing each one could take a long time.

Also, if the inspector tested every pie, the butcher would not have any to sell.

The inspector investigates the meat content of 12 randomly selected pies and records these figures, in grams:

234 256 231 234 251 249 243 216 249 232 250 253

The butcher says "I don't know what all the fuss is about. Some are bound to be a bit over and some a bit under."

c) Carry out a sign test, at the 5% significance level, to see whether the median is as high as 250 g.

$H_0: \eta = 250$ *g*

$H_1: \eta < 250$ *g*

Assigning +/- symbols to each pie (above/ below 250g):

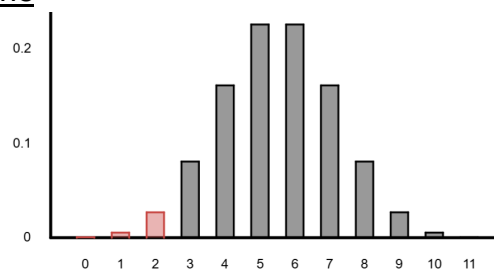
234	*256*	*231*	*234*	*251*	*249*	*243*	*216*	*249*	*232*	~~*250*~~	*253*
-	+	-	-	+	-	-	-	-	-	**0**	+

*The zero difference is **ignored** and that data point is removed.*

A one-tailed sign test at the 5% significance level, using $X \sim \mathrm{B}(\mathbf{11}, 0.5)$ *where X is the number of pies with meat content weighing less than 250 g is required.*

There are 3 + (or 7 –), so the test statistic is 3 (or 7).

Method 1: Critical regions



$P(X \leq 2) = 0.0327 < 0.05$ *and* $P(X \leq 3) = 0.113 > 0.05$ *so the critical region is* $X \leq 2$.

The test statistic of 3 is not in the critical region.

Method 2: p-values

$P(X \leq 3 +) \left(or\ P(X \geq 7 -)\right) = 0.1133$

So the one-tailed p-value is 0.1133 > 0.05.

Hence the result is not significant.
Do not reject H_0
There is insufficient evidence to suggest that the median amount of meat the butcher puts in the population of her pies is less than 250 g.

d) Give an advantage of using a sign test in this context.
There is no need to make any assumption regarding the distribution of the weights of meat in a pie.
It is also easy for the inspector to carry out.

COMMON AND POSSIBLE MISTAKES

- Students sometimes make conclusions about the mean as opposed to the median.
- Students often forget to remove values equal to the claimed median before using the binomial distribution.
- Students may forget to get the critical value appropriate for a two-tailed test.

Students must remember that:

- When calculating probabilities for the *p*-value method in a hypothesis test, always evaluate the probability that $X \leq$ or $X \geq$ (i.e. include the test statistic.)
- They must compare the test statistic with the critical value or compare the *p*-value with the significance level.

13b. Non-Parametric Hypothesis Tests: One-Sample Wilcoxon Signed-Rank Test (10.1)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand when a Wilcoxon signed-rank test can be used.
- Know the advantages and disadvantages of using a Wilcoxon signed-rank test.
- Carry out a Wilcoxon signed-rank test.
- Interpret the results of a Wilcoxon signed-rank test.

TEACHING POINTS

Again, emphasise to students that the assumption of the normal distribution for the underlying population is not required. There is one assumption about the underlying population in order to use the Wilcoxon signed-rank test: the population is distributed symmetrically about the median (which is also the mean). As usual, a sample must be obtained randomly for the results of the hypothesis test to be valid.

You may wish to practise calculating the test statistic and locating critical values from the Wilcoxon tables first. When given a sample $\{x_i\}$,

- subtract the median from each sample value $\{x_i - \eta\}$,
- rank the differences in order of **absolute** value/size (ignoring sign), with a lower rank denoting a smaller difference (tied ranks should be given the mean rank – see Unit 6c). Any differences of
  0 should be discarded as with the sign test.
- T_+ is the sum of the ranks with a positive difference. T_- is the sum of the ranks with a negative difference. T is the smaller of the two and this is the test statistic (students should check that the smaller value is the one that would be expected to be smaller under the alternative hypothesis). The number of values with a non-zero difference is *n*.

Using a table to show the differences and ranks will be helpful for students here. Encourage students to check that $T_+ + T_- = \frac{1}{2}n(n+1)$, since the right-hand side is the sum of the first *n* integers (students do not need to know this, but it would boost understanding if they did).

Extension

The critical values are more difficult to explain – students do not need to know where they come from but would gain a better appreciation of how the Wilcoxon signed-rank test works if they did.
An easy example is using $n = 3$ (you could use $n = 4$ or $n = 5$, but it would take longer). The possible distribution of ranks are (in columns):

1	-1	1	1	-1	-1	1	-1
2	2	-2	2	-2	2	-2	-2
3	3	3	-3	3	-3	-3	-3

By adding up the positive ranks in each column, the totals are:

6	5	4	3	3	2	1	0

The null hypothesis assumes that each outcome is equally likely so, for example, the probability that the sum of the positive ranks is 6 is $\frac{1}{8}$ and the probability that the sum of the positive ranks is 3 is $\frac{2}{8} = \frac{1}{4}$. By symmetry, the sum of the negative ranks follows the same probability distribution.

Showing students the tables for $n = 4$ and $n = 5$ illustrates how complicated these calculations can get. The critical values are given in the Formulae Book and are calculated by the integer k such that $P(T \leq k)$ is as **close to** the desired significance level (or $2 \times P(T \leq k)$ if the test is two-tailed). Students need to know that the critical values are given in the table and they should use these. The critical region is anything less than or equal to the critical value. Use the Wilcoxon Distribution with Critical Values activity on Desmos to help (it will display distributions up to $n = 10$).

The null hypothesis is "$\eta =$" where η is the population median. The alternative hypothesis could be either η is greater than, smaller than or not equal to. Students must remember to locate the correct critical value from the tables given the significance level and whether or not it is a one- or two-tailed test. The result is significant if the test statistic is smaller than or equal to the critical value (it is advisable that students are aware that because the smaller value was chosen, only the lower tail need to be considered).

OPPORTUNITIES FOR EMBEDDING THE SEC

As with all hypothesis tests at this point, a question involving a full cycle of the SEC can be used. The example seen in Unit 13a could easily be used as a Wilcoxon signed-rank test instead. The alterations to the end of the question are detailed below.

Exemplar

A trading standards inspector visits a butcher who sells meat pies. The pies are supposed to contain at least 250 g of meat but there have been complaints that the butcher does not put in enough meat (which is why the inspector went to her shop). The inspector investigates the meat content of 12 randomly selected pies and records these figures, in grams:

234 256 231 234 251 249 243 216 249 232 250 253

The butcher says "I don't know what all the fuss is about. Some are bound to be a bit over and some a bit under."

a) **Carry out a Wilcoxon Signed-Rank test at the 5% level to see whether the median is as high as 250 g.**

$H_0: \eta = 250\ g$

$H_1: \eta < 250\ g$

For the above data:

Data	234	256	231	234	251	249	243	216	249	232	~~250~~	253
Mean	250	250	250	250	250	250	250	250	250	250	250	250
Diff.	-16	6	-19	-16	1	-1	-7	-34	-1	-18	**0**	3
+ Rank		5			2							4
- Rank	7.5		10	7.5		2	6	11	2	9		

So $T_+ = 11$ *and* $T_- = 55$. *The test statistic is* $T = 11$.

Check: $11 + 55 = 66$ *and* $\frac{1}{2}(11)(12) = 66$.

We will ignore one value as the difference is zero.

A Wilcoxon signed-rank test at the 5% significance level, using $n = 11$ *is required.*

The critical value is 14 or less.

Since $11 < 14$, *the result is significant. We reject* H_0.

There is significant evidence to suggest that the population median amount of mean the butcher puts in her pies is smaller than 250 g.

b) **State any assumptions you have made about the underlying population.**

We have assumed that the population of the weights of meat in a pie is symmetrical about the median.

To further the embedding of the SEC, compare the results of the sign test (Unit 13a) and the Wilcoxon signed-rank tests to the same question. Notice that the results are not the same. This is because the Wilcoxon signed-rank tests take into account the sizes of the differences and therefore more likely to detect a difference, if one exists in the

population. Students who can explain this may be able to suggest a preference, covering Stage **E** of the SEC.

COMMON AND POSSIBLE MISTAKES

- Students often forget that although a normal distribution is not assumed, the underlying population still needs to be symmetrical about the median/mean.
- Students are likely to confuse this test with the one in the following sub-unit, due to the name.

The importance of the conditions for when the tests can be used is paramount here.

NOTES

Other exam boards and in the literature may use different critical values. These critical values are not the closest to the significance level, but the largest such integer k such that $\mathrm{P}(T \leq k) < \alpha$ where α is the significance level. Students are only required to have familiarity with the Wilcoxon signed-rank Table of critical values in the supplied Formulae Book.

Since the underlying distribution is assumed to be symmetrical, the mean and the median are assumed to be the same. In this test, the use of μ is condoned but η is expected and preferred, since the test utilises the properties of the median (as opposed to the mean) when carried out.

13c. Non-Parametric Hypothesis Tests: Wilcoxon Rank-Sum Test (10.2)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand when a Wilcoxon **rank-sum** test can be used
- Know the advantages and disadvantages of using a Wilcoxon rank-sum test
- Carry out a Wilcoxon rank-sum test
- Interpret the results of a Wilcoxon rank-sum test
- Understand and identify the type of data (2 independent samples) that is required for the test
- (Appreciate that the Wilcoxon rank-sum test was previously known as the Mann-Whitney test in the AQA legacy specification)

TEACHING POINTS

This is another example of a non-parametric test where the underlying population need not have a normal distribution. This is the first time students will carry out a hypothesis test on more than one sample, and will act as precursor to paired-sample tests (Unit 14) and ANOVA (Units 23 and 24). The population parameter being tested is the population median.

The assumptions of the Wilcoxon rank-sum test are that the observations are independent of each other and there are two samples obtained from populations with similar shaped distribution. This test is only for 2 independent sets of observations.

To begin, an explanation of how the Wilcoxon rank-sum test can be given (students do not need to know this but will gain a better appreciation if they did). An example of two independent samples of size two each, A and B. When ranked as a complete set, the possible orders of the samples are:

$$AABB, ABAB, ABBA, BAAB, BABA, BBAA$$

We then count, for each sample A element, how many sample B elements are smaller than that sample A element. We call this value U. For example: $AABB$, neither of the sample B elements are smaller than the sample A elements. This gives $U = 0$. However, for $BABA$, the first sample A element has only 1 sample B element smaller than it, but the second sample A element has 2 sample B elements smaller than it. This gives $U = 1 + 2 = 3$. The null hypothesis assumes each possible order is equally likely, resulting in a discrete probability distribution. Showing students the case when there are two independent samples of size 3 each will allow them to appreciate why the table of critical values exist.

For large samples, the above method is inefficient and the following method is an equivalent method (and should be followed for all cases).

The test statistic is calculated as follows:

- Both sets of observations are ranked **as one set,** with the smallest rank denoting the smallest value. Tied ranks will be given the mean of the given ranks.
- T_1 is the sum of the ranks of the first set of observations. T_2 is the sum of the ranks of the second set of observations. n_1 is the number of observations in the first set and n_2 is the number of observations in the second set.
- $U_1 = T_1 - \frac{1}{2}n_1(n_1+1)$ and $U_2 = T_2 - \frac{1}{2}n_2(n_2+1)$. The test statistic U is the smaller of the two (students should check that the smaller value is the one that would be expected to be smaller under the alternative hypothesis). This formula is given in the formula book.

As a check, students can check that $T_1 + T_2 = \frac{1}{2}(n_1+n_2)(n_1+n_2+1)$ although it is unlikely students will remember this. A different check that can be cross-referenced with the formula book is that $U_1 + U_2 = mn$.

Students also need to be introduced to the supplied Wilcoxon rank-sum tables to practise locating critical values. As in the Wilcoxon signed-rank test, the critical values are those closest to the significance level desired and hopefully students will appreciate why the tables exist. The result is significant if U is less than or equal to the critical value.

The null hypothesis can be written in symbols as "$\eta_X = \eta_Y$" or "$\eta_X - \eta_Y = 0$", with the subscripts of *X* and *Y* clearly defined. The alternative hypothesis could be "$\eta_X > \eta_Y$", "$\eta_X > \eta_Y$" or "$\eta_X \neq \eta_Y$" (or any equivalent rearrangement).

OPPORTUNITIES FOR EMBEDDING THE SEC

As with other hypothesis tests, a question involving a full cycle of the SEC can be used. The cycle of collecting a sample and identifying the advantages and disadvantages can be used as in Unit 13a. Note that in the below example, the hypothesis test is not stated so students will be required to identify the appropriate test (this covers point **A2** and **D4**).

Exemplar

The blood cholesterol levels of 11 men and 12 women were measured. These data are shown below.

Men **621, 237, 92, 745, 301, 550, 182, 723, 1301, 56, 104**

Women **208, 529, 104, 72, 377, 482, 50, 620, 1003, 162, 94, 391**

a) Is there evidence, at the 5% level, of a difference between the blood cholesterol levels of men and women?

Let X be the blood cholesterol levels of men and let Y be that for women.

H_0: $\eta_X - \eta_Y = 0$

H_1: $\eta_X - \eta_Y \neq 0$

Use a two-tailed Wilcoxon rank-sum test at the 5% significance level with $m = 11$ *and* $n = 12$*. Use a rank of 1 for the lowest value.*

X	621	237	92	745	301	550	182	723	1301	56	104	
Y	208	529	104	72	377	482	50	620	1003	162	94	391
Rank *X*	19	11	4	21	12	17	9	20	23	2	6.5	
Rank *Y*	10	16	6.5	3	13	15	1	18	22	8	5	14

$T_X = 144.5$ *and* $T_Y = 131.5$

So $U_X = 144.5 - \frac{1}{2}(11)(12) = 78.5$

and $U_Y = 131.5 - \frac{1}{2}(12)(13) = 53.5$

Check: $78.5 + 53.5 = 132 = 11 \times 12$.

Set $U = 53.5$ ($the\ lower\ available\ test\ statistic$)

The critical region is 34 or smaller.

Since $53.5 > 34$*, the result is not significant. We do not reject* H_0.

There is insufficient evidence at the 5% level to suggest that there is any difference in the distributions of the populations of blood cholesterol in men and women.

b) What assumptions have you had to make?

We have assumed that the samples of blood cholesterol levels for both men and women were obtained independently from each other and at random, and the distributions had a similar shape.

COMMON AND POSSIBLE MISTAKES

- Despite a summarised method in the formula book, students often forget the process of calculating the test statistic.
- The usual mistake of reading the incorrect critical value (significance level, one- or two- tailed).
- Only calculating the *U* value for the lower value of *T* (the lower value of *U* may arise from the higher value of *T*).
- Students are likely to confuse this test with the paired Wilcoxon signed-rank test. This will be even more relevant when the numbers of observations in each sample are equal.

Emphasise that students need to be able to identify when each test should be used.

Students **must** identify which of the *U* values is their test statistic.

NOTES

Other exam boards and the literature may use different critical values. These critical values are not the closest to the significance level, but the largest such integer k such that $\mathrm{P}(U \leq k) < \alpha$ where α is the significance level. This test is also known as the Mann-Whitney *U* test. Students who refer to this test using this name can still be awarded full marks (in line with the mark scheme).

There are actually two versions of a Wilcoxon Rank-Sum test: the first version (known as the Mann-Whitney U test) actually assumes nothing about the shapes of the underlying populations. This generalised the hypotheses of this test to

H_0: The samples are taken from populations with identical distributions
H_1: The samples are taken from populations with different distributions

This version is **not** being assessed on this specification.

The special case where one assumes the underlying populations to be of similar shapes (known as the Wilcoxon Rank-Sum test) as detailed earlier **is** assessed on this specification. However, a student who uses the hypotheses “The samples are taken from populations with identical distributions” can still be awarded full marks (in line with the mark scheme).

There is an extension to three or more independent samples known as the Kruskal-Wallis test. This is a non-parametric alternative to ANOVA (Unit 24). Some students may find it beneficial to explore this hypothesis test.

UNIT 14: Experimental Design

Return to overview

SPECIFICATION REFERENCES

10.1 Use sign or Wilcoxon signed-rank tests to investigate population median in single sample tests and also to investigate for differences using a paired model.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics
Binomial distribution (Unit 5)
Hypothesis testing (Unit 10a)
Sign test and Wilcoxon signed-rank tests (Unit 13)

KEYWORDS

alternative hypothesis, association, binomial, difference, distribution-free, experimental design, experimental error, hypothesis, independent, median, non-parametric, null hypothesis, paired, population, rank, sample, sign, significance level, unpaired, Wilcoxon signed-rank

UNIT SUMMARY

This unit finishes off the collection of non-parametric tests introduced in the previous sub-unit. These tests are presented in a separate unit because of its links to experimental design (as opposed to the links with their single sample analogues). Experimental design is explored in more depth in Unit 23 but students need to be aware of the motivation behind taking paired samples. This unit is primarily about hypothesis testing so use the binomial inference and Wilcoxon Distribution activities in Desmos.

14a. Experimental Design: Terminology (13.1)

Teaching time
1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand the term: experimental error
- Understand the difference between paired and unpaired data
- Appreciate the importance of experimental design

TEACHING POINTS

Students need to understand the difference between paired and unpaired data.

Students need to be able to appreciate that even if an experiment is planned carefully, and carried out under ideal, strict, consistent conditions, the outcome of the experiment may differ each time. Such variability of results is called experimental error.

One way of reducing experimental error is by using paired data. The example below illustrates how experimental error is reduced. Randomisation can be used to reduce bias.

It is important that students understand the motivation behind paired samples. Introducing the concept and importance of experimental design will not only motivate the following hypothesis tests, but also act as a precursor to experimental design in Unit 23 and cover stages **A**, **B** and **E** of the SEC.

OPPORTUNITIES FOR EMBEDDING THE SEC

A6 Students need to be able to identify in context how paired samples will reduce experimental error.

B1 Students need to be aware of the practicalities of randomisation and taking paired samples in the context of the question.

Stage E Students need to be able to suggest improvements to reduce experimental error given a context.

Exemplar

A group of children wanted to see whether the amount of air in their bicycle tyres made a difference to how easy it was to pedal their bicycles.
They decided to ride a particular route under two conditions: once with a tyre pressure of 40 psi and once with 65 psi.
The order in which they did this was decided by tossing a coin.

a) **Explain how experimental error is reduced in this context by using paired data.**
If one group of children used 40 psi and a different group of children used 65psi, any observed difference might be a difference between children or a difference between bikes, not between tyre pressures.
If there is a difference between pressures, it might be masked by variability between children.

b) **Explain how bias is reduced in this context by using randomisation.**
If they all did 65 psi first and 40 psi second, for example, they might have gone more slowly the second time because they were tired, or gone more quickly because they were used to the route, not because of using a different tyre pressure.

COMMON AND POSSIBLE MISTAKES

Students find it difficult to know the level of language, communication and clarity required for these questions.

Students often write verbose or unclear sentences which do not clearly demonstrate their understanding. Encourage students to write as concisely as possible, using bullet points. Account should be taken of the number of marks available, since this tends to indicate the number of points needed to be made to achieve full marks (in line with the mark scheme).

NOTES

Although the majority of experimental design is seen later in the specification, it can be used to embed the SEC into paired-comparison hypothesis tests, as well as providing a motivation behind the topic.

14b. Experimental Design: Paired Sign Test (10.1)

Teaching time
1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand when a paired sign test can be used
- Know the advantages and disadvantages of using a sign test
- Carry out a sign test.
- Interpret the results of a sign test

TEACHING POINTS

Revise the single sample sign test.

Other than the premise behind the hypothesis test, most of the objectives play out as in Unit 13a. Given paired samples, the sign test is a test for the median of the differences between pairs. The null hypothesis is always "$\eta_D = a$" where η_D is the median of the differences between the pairs and *a* is a claimed difference between the groups (usually this is 0, but need not be). The alternative hypotheses have greater than, less than or not equal to. *D* must be clearly defined.

When given paired samples $\{(x_i, y_i)\}$, find the differences between sample values and further subtract the claimed difference $\{x_i - y_i - a\}$. Proceed as in Unit 13a.

Exemplar

A group of children wanted to see whether the amount of air in their bicycle tyres made a difference to how easy it was to pedal their bicycles. They decided to ride a particular route under two conditions: once with a tyre pressure of 40 psi and once with 65 psi. The order in which they did this was decided by tossing a coin. The time it took (in minutes) for each circuit was:

Child	A	B	C	D	E	F	G	H	I	J
40 psi	34	54	23	67	46	35	49	51	31	27
65 psi	32	45	21	63	37	40	51	39	31	26

Using a sign test, test, at the 10 % significance level, whether the children are significantly faster with the higher pressure tyres.

Let X_{45} be the time taken to complete the route with a tyre pressure of 45 psi, and let X_{65} be the time taken to complete the route with a tyre pressure of 65 psi.
Let $D = X_{45} - X_{65}$.

$H_0: \eta_D = 0,$
$H_1: \eta_D > 0,$

where η_D is the population median of the difference between times taken to complete the route with a tyre pressure of 40 psi, and those of 65 psi.
*Note that differences found should be **consistent** with H_1*

Child	A	B	C	D	E	F	G	H	I	J
40 psi	34	54	23	67	46	35	49	51	~~31~~	27
65 psi	32	45	21	63	37	40	51	39	~~31~~	26
d	2	9	2	4	9	-5	-5	12	**0**	1
Sign	+	+	+	+	+	-	-	+		+

The zero is ignored and that data pair discarded.

A one-tailed paired-sample sign test at the 5% significance level using $D \sim \mathrm{B}(9, 0.5)$ is required.

There are 7 + and 2 −. The test statistic is 2.

Using the calculator, $P(D \leq 2) = 0.0898 > 0.05$, so the result is not significant. We do not reject H_0.
There is insufficient evidence at the 5% level to suggest that, on average, the children were significantly faster riding bicycles with 65 psi tyres rather than with 40 psi tyres.

OPPORTUNITIES FOR EMBEDDING THE SEC

A question involving a full cycle of the SEC can be used (see Unit 13a)

COMMON AND POSSIBLE MISTAKES

- Students sometimes make conclusions about the mean as opposed to the median.
- As can be seen from the above example, the language can be an issue (a faster time means a smaller time).
- A usual mistake is getting the critical value wrong for two-tailed tests.
- Forgetting to include *D* as a subscript on η in the hypotheses, and forgetting to clearly define *D*.
- Students often find signs inconsistent with one-tailed H_1 or fail to make their conclusion in the context of the question.
- In the cases when testing for a quantified median difference, e.g. H0: $\eta_D = 5$, then forgetting to subtract 5 from the difference $\{x_i - y_i\}$ (or not using 5 as a baseline).

14c. Experimental Design: Paired Wilcoxon Signed-Rank Test (10.1)

Teaching time
1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand when a Wilcoxon signed-rank test can be used
- Know the advantages and disadvantages of using a Wilcoxon signed-rank test
- Carry out a Wilcoxon signed-rank test
- Interpret the results of a Wilcoxon signed-rank test

TEACHING POINTS

Revise the unpaired Wilcoxon signed-rank test. The assumptions are the same as that in Unit 13b.

You may wish to practise calculating the test statistic and locating critical values from the Wilcoxon tables first. When given paired samples $\{(x_i, y_i)\}$,

- find the differences between sample values $\{x_i - y_i - a\}$, where *a* is a claimed median difference.
- rank the differences in order of absolute value/size (ignoring sign), with a lower rank denoting a smaller difference (tied ranks should be given the mean rank – see Unit 6c). Any differences of 0 should be disregarded.
- T_+ is the sum of the ranks with a positive difference. T_- is the sum of the ranks with a negative difference. T is the smaller of the two and this is the test statistic. The number of values with a non-zero difference is *n*.

Encourage students to check that $T_+ + T_- = \frac{1}{2}n(n+1)$, since the right-hand side is the sum of the first *n* integers (students do not need to know this, but it would boost understanding if they did).

The null hypothesis is "$\eta_D = a$" where η_D is the median difference between the pairs and *a* is the claimed median difference. The alternative hypothesis could have greater than, smaller than or not equal to.

The remainder of the hypothesis test is as in Unit 13b.

Exemplar

A group of children wanted to see whether the amount of air in their bicycle tyres made a difference to how easy it was to pedal their bicycles. They decided to ride a particular route under two conditions: once with a tyre pressure of 40 psi and once with 65 psi. The order in which they did this was decided by tossing a coin. The time it took (in minutes) for each circuit was:

Child	A	B	C	D	E	F	G	H	I	J
40 psi	34	54	23	67	46	35	49	51	31	27
65 psi	32	45	21	63	37	40	51	39	31	26

Using a paired-sample Wilcoxon Signed-Rank Test, test, at the 5% significance level, whether the children using the 65 psi tyres are more than 2 seconds faster than those using the 40 psi tyres.

Let X_{45} be the time taken to complete the route with a tyre pressure of 45 psi, and let X_{65} be the time taken to complete the route with a tyre pressure of 65 psi.
Let $D = X_{45} - X_{65}$.

$\mathrm{H}_0: \eta_D = 2,$
$\mathrm{H}_1: \eta_D > 2,$
where η_D is the population median for the difference between times taken to complete the route with a tyre pressure of 40 psi, and those of 65 psi. (Note that differences should be consistent with H_1)

Use rank 1 to denote the lowest difference.

Child	A	B	C	D	E	F	G	H	I	J
$\boldsymbol{X_{45}}$	34	54	23	67	46	35	49	51	31	27
$\boldsymbol{X_{65}}$	32	45	21	63	37	40	51	39	31	26
$\boldsymbol{D-2}$	**0**	7	**0**	2	7	-7	-4	10	-2	-1
Rank +		6		2.5	6			8		1
Rank -						6	4		2.5	

The zeroes are ignored and those data pairs discarded.

A one-tailed paired-sample Wilcoxon signed-rank test at the 5% significance level using $n = 8$ is required.

$T_+ = 23.5$ *and* $T_- = 12.5$.
Check: $23.5 + 12.5 = 36 = \frac{1}{2}(8)(9)$
The test statistic is $T = 12.5$.

The critical region is 6 *or smaller.*

Since $12.5 > 6$*, the result is not significant. We do not reject* H_0*.*
There is not significant evidence at the 5% level to suggest that, on average, the children using 65 psi tyres were more than 2 seconds faster than those using 40 psi tyres.

OPPORTUNITIES FOR EMBEDDING THE SEC

See Units 13a and 14a for examples. Note that **Stage E** can also be covered by comparing the results of a sign test and a Wilcoxon signed-rank test. The examples given show different conclusions: students need to be able to justify why this may be and explain whether or not assumptions are met.

COMMON AND POSSIBLE MISTAKES

- Students often forget that although a normal distribution is not assumed, the underlying populations still need to be symmetrical about the median/mean.
- Students are likely to confuse this test with the Wilcoxon rank-sum test.
- Students will make mistakes when dealing with tied-ranks (missing values, incorrect means etc.).
- Students may forget to include the subscript *D* on $\boldsymbol{\eta}$ or forget to define *D* completely.
- When dealing with quantified median differences, students may forget to subtract the quantified difference (or use the quantified difference as the baseline) or subtracting the incorrect way (e.g. giving a difference of the negative of what is required).
- Students may have obtained differences that are inconsistent with $\mathbf{H_1}$ and they may fail to identify which *T* is the test statistic.
- Students may conclude 'do not reject H_0' when the test statistic is equal to the critical value.

The importance of the conditions for when the tests can be used is paramount here.

As with the single-sample Wilcoxon Signed-Rank test, since the differences are assumed to be symmetrical, the mean difference and the median difference are assumed to be the same. In this test, the use of $\boldsymbol{\mu_D}$ is condoned but $\boldsymbol{\eta_D}$ is expected and preferred, since the test utilises the properties of the median (as opposed to the mean) when carried out.

UNIT 15: The Poisson Distribution

Return to overview

SPECIFICATION REFERENCES

18.1 Determine when a Poisson model is appropriate (in real world situations including modelling assumptions).

18.3 Evaluate probabilities for Poisson and exponential distributions and know the corresponding mean and variance.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics
Probability (Unit 3)
Discrete Random Variables (Unit 4)
Binomial Distribution (Unit 5)

KEYWORDS

at least, at most, binomial, conditions, constant, continuous, cumulative, discrete, distribution, event, exceeds, expectation, exponential, independent, mean, outcome, Poisson, probability, random variable, random, rate, standard deviation, sum, uniform, variable, variance

UNIT SUMMARY

It is very similar to the binomial distribution (Unit 5) in terms of mathematical and calculator skills. The major differences are the concept, conditions and interpretation.

This distribution has direct links with the Exponential distribution (Unit 20). It is advisable to see, and justify, contextual situations which can be modelled by a Poisson distribution throughout this unit. It will be beneficial to see situations which initially appear to be modelled by either a Binomial or Poisson distribution. This consolidates binomial distribution and enhances students' appreciation of the SEC (**Stage A**).

Unlike in previous specifications (and like the binomial distribution), the table of probabilities from a Poisson distribution (which are provided in the formula book) are obsolete but tables are still provided because the calculator requirements for GCE Statistics are the same as those for GCE Mathematics. The recommended calculator can calculate Poisson probabilities in the same way as it can binomial probabilities. This shortens the teaching time of this topic from its legacy counterparts.

As a historical footnote, the origins of the French mathematician Siméon Denis Poisson (1781-1840), whose name is lent to the distribution can be discussed.

The amount of material relating to the Poisson distribution has been vastly reduced since legacy specifications. The only application of the Poisson distribution outside of

this unit within the A level course is the Exponential distribution (Unit 20) and Goodness of Fit (Unit 22). However, a lot of concepts in many units can be applied to the Poisson distribution as extension material. Students who have career aspirations using applied statistics may find it beneficial to learn these techniques as extension material.

These topics include: The Poisson approximation to the binomial, the normal approximation to the Poisson, confidence intervals for λ.

Use the Poisson distribution activity in Desmos to help students with the probability distribution.

15a. The Poisson Distribution: Conditions for a Poisson Distribution (18.1)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Recognise when to use the Poisson distribution
- State any assumptions necessary in order to use the Poisson distribution
- Understand and use the notation $X \sim \text{Po}(\lambda)$ where λ is the mean rate of an event occurring

TEACHING POINTS

The Poisson distribution is another special case of a discrete random variable. The conditions that students must know are:

- Events occur randomly in a specified unit of space or time
- Events occur independently from each other
- The events occur at a constant average rate

Methods to remember these involve flashcards, tarsias, continual identification through contextual scenarios, or songs. The acronym ERIC is a good way of remembering the conditions for a Poisson distribution: E(vents occur), R(andomly), I(ndependently), C(onstant average rate). Students must also appreciate that, unlike the binomial distribution, there is no upper limit to the value of a Poisson variable.

A basic, initial example to allow students to meet the Poisson distribution is:
A student stands by the roadside and counts the number of cars that pass by on the road. On average, 3 cars pass by every minute.

This allows students to relate the conditions of a Poisson distribution to the scenario and discuss whether the conditions are satisfied, stating assumptions if necessary. For example, one may argue that the cars passing may not be independent if a convoy of cars pass by. Again, it is very important that students define their variables, this time ensuring that the time/space unit is stated: *Let X be the number of cars that pass by in a minute.*

Identifying the distribution and the mean rate should then be expressed as: $X \sim \text{Po}(3)$.

When introducing context, give students a wide range of examples to allow them to appreciate how common a Poisson distribution can be found in real-world contexts. Similarly, contexts which look as if they could be modelled by a binomial distribution can also be seen, in order for students to choose the appropriate distribution.

Exemplar

A students' union wants to take a survey about how effective a university's financial assistance policy is.
This particular university claim that 9% of their students are from families earning less than £15,000 a year (low-income students).
The union begin by conducting exploratory data analysis on a random sample of 10 chemistry students, and asking whether they are low-income students or not.

Suggest a suitable distribution in this case.

- *A binomial distribution may be suitable here, where each "trial" is a person in the sample and "success" is "person is a low-income student".*
- *The number of people asked is fixed at 10, and (provided the university's claim is correct and the same across all subjects) the probability that a student is a low-income student is 0.09.*
- *The probability that one chemistry student in the sample is a low-income student does not affect the probability that another student in the sample is a low-income student, so the trials are independent from each other.*
- *Finally, a student is either a low-income student or not, so there are exactly two outcomes for each trial.*

Exemplar

At airport security, the home office is conducting a survey of the ethnicity of British people passing through border control.
Previous research has indicated that, on average, 6400 people passing through border control every day are British Asian.
The home office decides to take record the ethnicity of all the people passing through a particular border control desk over a long period of time, and recording whether they are British Asian or not.

a) **Explain why a Poisson distribution could be suitable for this scenario, and state any assumptions that need to be made.**
    - *There is a fixed unit of time (one day).*
    - *There is a constant average rate.*
    - *They would be recording the number, X, of British Asians passing through the border desk every day.*
    - *If we assume that the people arriving at the border control desk are independent of each other, then $X \sim \text{Po}(6400)$.*

b) Comment on the suitability of your assumptions.
The assumption may not be suitable here since it is likely that family groups may pass through border control together: children may have the same ethnicity as their parents so it is likely that the people passing through may not be independent from one another.

OPPORTUNITIES FOR EMBEDDING THE SEC

A1 Identifying what factors in a scenario are important in order to be able to identify a Poisson distribution.

A3 Identifying what data to collect in order to determine whether the conditions are satisfied.

A4 As seen in the first example above, the importance of exploratory data analysis can be appreciated as a starting point prior to a full investigation into a problem.

D2 When concluding whether or not a Poisson distribution is appropriate, the context of the scenario must be considered.

E1 Identifying when a Poisson distribution is not appropriate can lead to a discussion of modelling assumptions.

Distinguishing between a binomial and a Poisson distribution is challenging for students (see below), so it will be beneficial to practise plenty of questions combining the two distributions. For example,
A student sits on the roadside and counts the number of cars that pass by on the road every minute. On average, 3 cars pass by in a minute *(Poisson)*.
The student replicates this experiment over 10 minutes and wants to know the expected number of minutes where more than 5 cars pass by *(binomial)*.

This example also uses the notion of replication, which is an aspect of experimental design seen in Unit 23.

COMMON AND POSSIBLE MISTAKES

- Students often find it difficult to identify and contextualise whether events occur at random or independent.
- Students also are reluctant to state uncertainty in their answers e.g. “the events might not be independent” – encourage stating uncertainty, as it demonstrates an appreciation of the limitations of statistics.
- Students often give generic answers when asked about the assumptions of the distributions e.g. “events are independent” instead of contextual answers.
- Students often do not relate the assumptions to the specific context being presented.
- Students may misidentify scenarios with having a Poisson distribution when actually they are binomial. The absence of an upper limit may help students recognise when a Poisson distribution is more appropriate than a binomial distribution.

Students should work with the interval for which observed data are available.

NOTES

Often in the literature, the assumption of “independence” is accompanied with the assumption that events must occur singly. Students who use the assumption of events occurring singly and independently could also be awarded full marks (in line with the mark scheme). Conversely, a violation of the “events occurring singly” assumption may be utilised to demonstrate a Poisson distribution is not suitable can be awarded full marks (in line with the mark scheme).

15b. The Poisson Distribution: Finding Probabilities from a Poisson Distribution (18.3)

Teaching time
1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Appreciate that there is a formula $\mathrm{P}(X = x) = \frac{e^{-\lambda}\lambda^r}{r!}$.
- Use a calculator to find $\mathrm{P}(X = x)$.
- Use a calculator to find $\mathrm{P}(X \leq x)$.
- Calculate other probabilities e.g. $\mathrm{P}(X \geq x)$, $\mathrm{P}(a < X \leq b)$ etc.
- Use trial and error to answer questions involve the Inverse Poisson distribution.

TEACHING POINTS

Remind students how to calculate binomial probabilities using the calculator. Practise finding probabilities $\mathrm{P}(X = x)$, $\mathrm{P}(X \leq x)$, $\mathrm{P}(X \geq x)$, $\mathrm{P}(a \leq X \leq b)$ and corresponding weak inequalities.

Students could be introduced to the number e at this point. Explain that it is a transcendental number, like π, whose decimal expansion is infinite with no repeating pattern, but appears in many aspects of mathematics. One area it appears in is the formula for $\mathrm{P}(X = x)$ in a Poisson distribution.

Note: Explaining the origins behind this formula will be too mathematically challenging for most students.

Continuing from the previous sub-unit with the example:

Exemplar

A student stands by the roadside and counts the number of cars that pass by on the road. On average, 3 cars pass by every minute.
Let X be the number of cars that pass by on the road every minute. Then $X \sim \mathrm{Po}(3)$.

Practise finding $\mathrm{P}(X = x)$ both using the formula and the Poisson distribution mode on their calculator. You may also want to use the Poisson distribution activity in Desmos to generate a vertical line representation of the probability distribution.

It is advisable that students have plenty of practice with their calculators, allowing them to familiarise themselves with this mode. These calculators can also generate the tabulated probability distributions, and these can be used to check their answers.

These calculators may not calculate Poisson probabilities other than the two described above, so conventional methods for calculating other probabilities still need to be taught as before, e.g.

$$\mathrm{P}(X < x) = \mathrm{P}(X \leq x - 1)$$

$$\mathrm{P}(X > x) = 1 - \mathrm{P}(X \leq x)$$

$$\mathrm{P}(a \leq x \leq b) = \mathrm{P}(x \leq b) - \mathrm{P}(x < a)$$

The use of a number line (from 0 to ∞ or "…") is extremely helpful to visualise the required outcomes.

Questions without context could be seen first:

Exemplar

X is a random variable and $X \sim \mathrm{Po}(0.23)$. Find $\mathrm{P}(X > 2)$.

0	*1*	*2*	*3* …

Using the calculator, $P(X > 2) = 1 - P(X \leq 2) = 1 - 0.9983 = 0.0017$.

Emphasise that the calculator has two Poisson distribution modes; one calculates $\mathrm{P}(X = x)$ and the other calculates $\mathrm{P}(X \leq x)$, and students must ensure they use the correct one (as with binomial probabilities).
Unlike the binomial distribution, using the calculator to generate the full tabulated probability distribution and summing the appropriate values may not be a viable alternative, due to the lack of an upper limit.

When introducing context, genuine scenarios where a Poisson distribution can be used for modelling can be seen. It is beneficial to use words and phrases such as "at most" and "exceeds" – these will have been seen in Unit 4. It is important that students have good literacy skills in order to deal with questions such as these.

Students may also be asked to use the inverse Poisson distribution to answer questions. Some calculators have "inverse Poisson" modes and some simple questions may be solved algebraically, but students are expected to use trial and error to answer these questions.

Exemplar

The number of computer faults at a business are believed to occur at random, independently of each other and at a constant average rate of 3.4 per day.

a) **Write down a suitable distribution for the number of computer faults in a day.**
Let X be the number of computer faults in a day. X~Po(3.4)

b) **Find the probability that more than 1 but at most 4 computer faults are reported in a day.**

0 1 | 2 3 4 | 5 …

Using the calculator, $\mathrm{P}(1 < X \leq 4) = \mathrm{P}(X \leq 4) - \mathrm{P}(X \leq 1) = 0.7442 - 0.1468 = 0.597$

The IT manager must do a full network diagnostic (costing the business additional money) if more than *k* computer faults are reported in a day. The IT manager wishes this to happen less than 3% of the time.

c) **Find the value of k**

0 … k-1 k | k+1 …

Using trial and error on the calculator:

$$\mathrm{P}(X > 6) = 1 - \mathrm{P}(X \leq 6) = 1 - 0.942 = 0.058 > 0.03$$

$$\mathrm{P}(X > 7) = 1 - \mathrm{P}(X \leq 7) = 1 - 0.977 = 0.023 < 0.03$$

So k = 7

OPPORTUNITIES FOR EMBEDDING THE SEC

C1 Students are expected to use the calculators in order to calculate Poisson probabilities and cumulative probabilities. The tables of Poisson probabilities may be seen by all students – this will allow them to appreciate the usefulness of technology in the world of statistics.

C2 Vertical line charts can be used to demonstrate a Poisson distribution.

D1 Calculating probabilities and interpreting them in the context of the question.

D2 Relating the calculated probabilities to a particular claim made in the question.

COMMON AND POSSIBLE MISTAKES

- Many mistakes occur during the reading of contextual questions, e.g. at least 3 is interpreted as >3.
- Misinterpreting $\mathrm{P}(X < 3)$ as $\mathrm{P}(X \leq 3)$.
- Calculating $\mathrm{P}(a \leq X \leq b) = \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a)$.
- Calculators with an “inverse Poisson” mode may report the value **closest** to a particular probability as opposed to less than/greater than.

NOTES

An extension activity in the form of determining the Poisson probabilities using the formula can be used for the more mathematically able. Questions about the binomial distribution can be practised alongside the Poisson distribution.

15c. The Poisson Distribution: Mean and Variance (18.3)

Teaching time
1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand the mean and variance of $\mathrm{Po}(\lambda)$ are both λ
- Appreciate how the mean and variance can be used to determine the suitability of a Poisson distribution
- Appreciate situations where the Poisson distribution may be an appropriate model

TEACHING POINTS

Revise the concepts of expectation (mean), variance and standard deviation of a discrete probability distribution, in particular the binomial distribution.

The mean of a Poisson distribution λ is easily explained by definition.

The variance λ is harder to explain, but the result is a direct application of the formula for the variance and can be explained as such – the proof of the mean and variance of a Poisson distribution will not be assessed.

Questions on mean, variance and standard deviation can be applied to questions from Unit 15b. Students must remember to square root the variance to obtain the standard deviation.

The unit could finish with comments on the suitability of a Poisson distribution, referring to the mean and variance.

Exemplar

At airport security, the home office is conducting a survey of the ethnicity of British people passing through border control.
Previous research has indicated that on average 6400 people passing through border control each day are British Asian.
The home office decide to sample every 5th person passing through border control, for a period of 2 weeks, and record whether they are British Asian or not.
After the data are recorded, the mean number of British Asians passing through passport control on a particular day was 6340, with a standard deviation of 84.2.
Comment on the suitability of the Poisson distribution in this case.

Let X be the number of British Asians passing through border control each day.
If the Poisson model were suitable, then $X \sim \mathrm{Po}(6400)$, $\mathrm{E}(X) = 6400$ *and* $\mathrm{Var}(X) = 6400$.
The Poisson distribution may not be a suitable model.
The mean value is close to the expected value but the standard deviation of 84.2 *gives a variance of* $84.2^2 = 7089.64$ *which is quite a bit higher than the expected variance for X.*

Note that the difference between the observed and expected variance is subjective – as long as the conclusion is justified and clear, students can be awarded full marks (in line with the mark scheme), although teacher's judgement should be exercised.

OPPORTUNITIES FOR EMBEDDING THE SEC

C2 Calculating the mean and variance and applying it to the context of the question, making inferences about a population.

D1 Calculating the sample mean and variance and interpreting them in the context of the question.

D2 Using the calculation of the expected mean and variance to assess the suitability of the Poisson model.

COMMON AND POSSIBLE MISTAKES

Forgetting to square root the variance in order to find the standard deviation. This is especially common in contextual questions where the objective is to compare the expected variance/standard deviation with an observed one.

NOTES

As an extension exercise, you can explain that if X is a Poisson variable, aX is not a Poisson variable if $a \neq 1$, and refer to the variance as a reason.

Extension Exemplar

A factory worker stands on the production line and counts the number of defective items that pass by.
On average, 3 defective items pass by every minute. It will cost the company 5p to replace each defective item.
If X is the number of defective items that pass by every minute, then it may be assumed that X follows a Poisson distribution.
Let Y be the cost of replacing defective items (in pence) every minute. Explain why Y is not a Poisson variable.

Since $X \sim \mathrm{Po}(3)$*, then* $\mathrm{E}(X) = \mathrm{Var}(X) = 3$.
However $Y = 5X$*, so* $\mathrm{E}(Y) = 5\mathrm{E}(X) = 15$ *and* $\mathrm{Var}(Y) = 5^2\mathrm{Var}(X) = 75$.
Since $\mathrm{E}(Y) \neq \mathrm{Var}(Y)$*, then* Y *cannot be modelled by a Poisson distribution.*

UNIT 16: Combinations of Independent Random Variables

Return to overview

SPECIFICATION REFERENCES

12.2 Evaluate the mean and variance of linear combinations of independent random variables through knowledge that if X_i are independently distributed $\left(\mu_i, \sigma_i^2\right)$ then $\sum a_i X_i$ is distributed $\left(\sum a_i \mu_i, \sum a_i^2 \sigma_i^2\right)$.

12.3 Evaluate probabilities for linear combinations of two or more independent normal distributions and apply this knowledge to practical situations.

18.1 Determine when a Poisson model is appropriate (in real world situations including modelling assumptions).

PRIOR KNOWLEDGE

Year 1 of A Level Statistics
Numerical Measures (Unit 1)
Normal Distribution (Unit 7)
Sampling distribution of the mean (Unit 9)

Year 2 of A Level Statistics
Poisson Distribution (Unit 15)

KEYWORDS

combinations, difference, distribution, independent, linear, mean, normal, Poisson, random variables, random, scaling, standard deviation, sum, variance,

UNIT SUMMARY

This unit extends the idea of linear scaling seen in Units 1, 4 and 6. Linear scaling links nicely to this unit where the concept and results of adding random variables are brought to the fore. It also brings together the notion of combining independent normal or Poisson variables. The linear combination of normal variables also explains the parameters in the sampling distribution of the mean of a normal distribution.

It is an extension of Units 1, 7 and 15.

There isn't much in the way of graphical representation in this unit, other than a visual aid for linear scaling.

16a. Combinations of Independent Random Variables: Expectation and Variance of a linear combination of independent random variables (12.2)

Teaching time
1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand that $\mathrm{E}(aX \pm b) = a\mathrm{E}(X) \pm b$ and $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.
- Understand that $\mathrm{E}(aX \pm bY + c) = a\mathrm{E}(X) \pm b\mathrm{E}(Y) + c$ and $\mathrm{Var}(aX \pm bY + c) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)$.

TEACHING POINTS

Revise linear scaling, both with observed mean and variance ($\bar{x}$ and s_x^2) and with theoretical mean and variance ($\mathrm{E}(X)$ and $\mathrm{Var}(X)$).

Begin introducing combinations of random variables. Emphasise that the random variables need to be independent for the following results to be true (see notes below). A standard example is: **Let X be the number showing uppermost on a fair six-sided die, and let Y be the number showing uppermost on a different fair dix-sided die**. Begin by writing the possible values of $X + Y$. Using a sample space grid, students can then write out the probability distribution of $X + Y$.

A second important example: **write down the possible values of $2X$**. This can illustrate to students the difference between the sum of two variables and the scalar multiple of a variable. It can also highlight why $X + X$ should not be used to describe two different variables. Again, the probability distribution should be written down.

As a revision exercise, students could find $\mathrm{E}(X)$, $\mathrm{E}(Y)$, $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$ and $\mathrm{E}(X + Y)$, $\mathrm{E}(2X)$, $\mathrm{Var}(X + Y)$ and $\mathrm{Var}(2X)$ using the formulae from Unit 4. Students should then be able to identify the link that $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$, $\mathrm{E}(2X) = 2\mathrm{E}(X)$ and $\mathrm{Var}(2X) = 4\mathrm{Var}(X)$. Combining the two results will produce the results specified in the objectives above.

You can either repeat the activity with $X - Y$ (obtaining $\mathrm{E}(X - Y) = \mathrm{E}(X) - \mathrm{E}(Y)$ and $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$) to show the desired results, or you can use algebra: $X - Y = X + (-Y)$ and using the fact that $(-1)^2 = 1$.

Introduce context when students are comfortable with the calculations. Students need to be able to define their variables and justify the use of the linear combinations in context.

OPPORTUNITIES FOR EMBEDDING THE SEC

A1 Identifying factors related to the problem under investigation in order to justify their use of a linear combination of random variables.

D1 Interpreting the expected value and standard deviation in the context of the questions.

D2 Reaching conclusions based on their interpretation of these numerical measures, and using these conclusions as a basis of comparison or evidence for refutation.

Linking together questions involving the binomial or Poisson distributions may also help consolidate material, and students may appreciate that the combination of binomial variables is not necessarily binomial, and a scalar multiple of a Poisson variable is not necessarily Poisson. This will also help students appreciate the interesting situation of the normal distribution.

Exemplar

The manager of an electronics factory is monitoring the proportion of defective transistors and capacitors produced.
The number of defective transistors, T, in a sample of size 5, can be modelled by the following distribution:

t	0	1	2	3	4	5
$\mathbf{P}(T = t)$	0.66	0.25	0.05	0.02	0.01	0.01

It was established over many years that the probability that a capacitor is faulty is 0.1, independently of whether other capacitors were defective.
A sample of 5 capacitors is taken and the number of defective capacitors, C, is recorded.

a) Write down the name of the probability distribution of C, giving its mean and variance.
C has a binomial distribution.
The mean is $5 \times 0.1 = 0.5$ *and the variance is* $5 \times 0.1 \times 0.9 = 0.45$.

b) Find the mean and variance of T.
Using the calculator, $\mathrm{E}(T) = 0.5$ *and* $\mathrm{Var}(T) = 0.79$.

The sales department of the company in charge of the factory decide to sell a bumper pack of 5 capacitors and 5 transistors.
Each defective transistor produced will cost the company 70p to replace.
Each defective capacitor produced will cost the company £1.20 to replace.

c) Calculate the expected cost from faulty items in the bumper pack.

$\mathrm{E}(0.70T + 1.20C) = 0.70\mathrm{E}(T) + 1.20\mathrm{E}(C) = 0.70(0.5) + 1.20(0.5) = 0.95.$
So it is expected to cost the company 95p from each pack.

d) Calculate the variance of the cost from faulty items from the bumper pack.

$\mathrm{Var}(0.70T + 1.20C) = 0.49\mathrm{Var}(T) + 1.44\mathrm{Var}(C) = 0.49(0.79) + 1.44(0.45) = 1.0351.$ *So the variance of the cost is 1.04.*

e) They decide to sell the pack for £1.50 each. Comment on this decision.

This is not a good decision.
The expected loss from each bumper pack is 95p which gives an expected profit of 65p per pack. However, the standard deviation of the loss is $\sqrt{1.0351} = 1.02$*, so some packs could result in an overall loss per pack.*

COMMON AND POSSIBLE MISTAKES

- Students sometimes use $\mathrm{Var}(aX - bY) = a\mathrm{Var}(X) - b\mathrm{Var}(Y)$, and other such mistakes.
- Students often find it difficult to distinguish between the sum of two random variables and the scalar multiple of one random variable. This is down to reading and comprehension which becomes more important as more than 2 variables are used.

NOTES

Students are only required to deal with cases where the random variables combined are independent of each other. However, the expectation formulae hold if the random variables are not independent of each other. The variance formulae require the random variables to be independent of each other.

16b. Combinations of Independent Random Variables: The Sum of Poisson Variables (18.1)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Appreciate that the sum of two independent Poisson variables also has a Poisson distribution.
- Understand that the mean of the sum of two independent Poisson variables is the sum of the two means.
- Understand that a Poisson variable can be sub-divided into the sum of two (or more) independent Poisson variables.
- Interpret the results in context.

TEACHING POINTS

Revise the Poisson distribution.

A standard example to begin:
A student sits on the roadside recording the number of cars passing her going north every 10 minutes. She also records, on a separate sheet, the number of cars passing her going south every 10 minutes.
This example should allow students to appreciate that (assuming independence of the cars passing her in either direction), both situations have a Poisson distribution. If the student were to consider the number of cars passing in either direction in a duration of 10 minutes, the conditions for a Poisson distribution still hold and is the sum of the two Poisson variables. Using the results in the previous sub-unit (or using the example), it is easy to explain the combined mean. Emphasise that the Poisson variables must be independent for this to be true.

This standard example can then be used to change the time-frame: if, instead of every 10 minutes, the student were to record the number of cars passing every hour, then this is a sum of six separate 10 minute durations. Encourage the use of subscripts to denote separate observations: *let* X *be the number of cars passing north in a 10 minute duration. Then the number of cars passing normal in an hour duration is* $X_1 + X_2 + X_3 + X_4 + X_5 + X_6$. Equally if, instead of every 10 minutes, the student were to record the number of cars passing every minute, then this is a sum of 10 separate 1 minute durations. Again use subscripts: *let X be the number of cars passing north in a 10 minute duration. Let Y be the number of cars passing north in a 1 minute duration. Then* $X = Y_1 + Y_2 + \cdots + Y_{10}$. Avoid the use of $6X$ or $10Y$ in these situations, since these mean something else entirely (indeed, if X has a Poisson distribution then aX does not, if $a \neq 1$).

Begin with non-contextual questions to allow students time to practise using the calculator. Then introduce context. Questions which also use the binomial distribution can later be combined with these Poisson questions, so students have to choose the appropriate distribution to use (see below).

OPPORTUNITIES FOR EMBEDDING THE SEC

A1 Identifying the factors related to the problem under investigation in order to justify their use of a linear combination of random variables.

D1 Interpreting the expected value and standard deviation in the context of the questions.

D2 Reaching conclusions based on their interpretation of these numerical measures, and using these conclusions as a basis of comparison or evidence for refutation.

Exemplar

An investigation by a government department is carried out in a particular county, dividing the county into regions and further dividing the regions into districts.
At the time of this investigation. a low-income household is defined as one which earns a total of less than £13,000 a year (before housing costs).
The investigation assumes that low-income households in a particular county occur independently, at random and at an average rate of 4.5 households per 1 km^2 district.
Find the probability that:

a) there are six or fewer low-income households in a randomly selected 1 km^2 district,

Let X be the number of low-income households in a 1 km^2 district. Then $X \sim \text{Po}(4.5)$.
Using the calculator, $\text{P}(X \leq 6) = 0.8311$.

b) in a randomly selected 2 km^2 district the number of low-income households is between two and ten (inclusive),

Let Y be the number of low-income households in a 2 km^2 district. Then $Y = X_1 + X_2 \sim \text{Po}(9)$.
$\text{P}(2 \leq Y \leq 10) = \text{P}(Y \leq 10) - \text{P}(Y \leq 1) = 0.7060 - 0.0012 = 0.7048$.

c) in a region of five 2 km^2 districts, there are at least 3 districts containing at least 7 low-income households.

With Y as in part (b), $\text{P}(Y \geq 7) = 1 - \text{P}(Y \leq 6) = 1 - 0.2068 = 0.7832$.
Let W be the number of 2 km^2 districts containing at least 7 low-income households.
Then $W \sim \text{B}(5, 0.7832)$.
$\text{P}(W \geq 3) = 1 - \text{P}(W \leq 2) = 1 - 0.0716 = 0.9284$.

The government considers a district containing more than 6 low-income households to be deprived. If over half of the districts in a region are deprived, then the region is considered deprived.

d) Comment on whether the region in part (c) is deprived or not.

Since there is a 93% chance that at least three of the five districts are considered deprived, the region is likely to be considered deprived.

COMMON AND POSSIBLE MISTAKES

- Students sometimes use $2X$ when they mean $X_1 + X_2$. Emphasise to students that while $X_1 + X_2$ has a Poisson distribution, $2X$ does not (the values of $2X$ will be even integers, whereas a Poisson variable can have a value of any non-negative integer).
- Students may confuse a binomial and a Poisson distribution in some contexts. In part (iii) in the above example, some students will attempt to find the probability of finding more than 21 low-income households over a region of 10 km^2 (a Poisson situation), which is different from determining the probability of at least 3 districts with at least 7 low-income households (a binomial situation).

NOTES

The difference of two Poisson variables is not Poisson, since $X - Y$ can be negative.

16c. Combinations of Independent Random Variables: Linear combinations of Normal Variables (12.3)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Appreciate that the sum or difference of two independent normal variables is also normally distributed.
- Appreciate that a scalar multiple of a normal variable is also normally distributed.
- Find and interpret probabilities of a linear combination of independent normal variables.

TEACHING POINTS

Revise the normal distribution.

Students need to know that any linear combination of independent normal variables is also normally distributed. Emphasise again that independence is a necessity to determine the variance.

You can also revise the sampling distribution of the mean, to help consolidate the idea that if $X \sim \mathrm{N}(\mu, \sigma^2)$, then $\bar{X} = \frac{\sum X_i}{n} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$. The proof of this can be found in Unit 9b.

Non-contextual questions could be given first.

Exemplar

Let $X \sim \mathrm{N}(45, 12^2)$ and $Y \sim \mathrm{N}(12, 16^2)$. Find the distribution of $2X + 4Y$.
Since $E(2X + 4Y) = 2E(X) + 4E(Y) = 2 \times 45 + 4 \times 12 = 138$ *and*
$Var(2X + 4Y) = 2^2 Var(X) + 4^2 Var(Y) = 4 \times 12^2 + 16 \times 16^2 = 4672$,
we have $2X + 4Y \sim N(138, 4672)$.

Contextual questions could then be attempted. This will take the most time, but you can use this opportunity to embed the SEC (see below).

OPPORTUNITIES FOR EMBEDDING THE SEC

Exemplar

The mass of a certain kind of biscuit can be modelled by a normal variable with mean 60 g and standard deviation 5 g.
A packet contains 12 randomly selected biscuits.
A factory produces, on average, 120 biscuits in a batch.

The mass of the packaging material can be modelled by a normal variable with mean 40 g and standard deviation 4 g and is independent of the masses of the biscuits.

a) **Find the probability that the total mass of a packet of biscuits is less than 750 g.**
Let X be the mass of a biscuit. Then $X \sim \mathrm{N}(60, 5^2)$.
Let Y be the mass of the packaging. Then $Y \sim \mathrm{N}(40, 4^2)$.
Let W be the mass of a packet of biscuits.
Then $W = X_1 + X_2 + \cdots + X_{12} + Y \sim \mathrm{N}(60 \times 12 + 40, 12 \times 5^2 + 4^2) = \mathrm{N}(760, 316)$.
So $\mathrm{P}(W \leq 750) = 0.2869$.

b) **(i) At what stage in the calculation did you need to use the fact that the sample of biscuits should be random?**
Adding the variances of the masses of 12 individual biscuits to give the variance of the mass of 12 biscuits is only valid if the masses of biscuits are independent of each other.
This will be true if the sample was obtained randomly.

(ii) At what stage in the calculation did you use the fact that the mass of the packaging is independent of the masses of the biscuits?
Adding the variance of the mass of the packaging to the variance of the mass of the 12 biscuits is only valid if the masses are independent of each other.

COMMON AND POSSIBLE MISTAKES

- Students sometimes use $\mathrm{Var}(aX - bY) = a\mathrm{Var}(X) - b\mathrm{Var}(Y)$ and other similar mistakes.
- Students may use the standard deviation instead of the variance when using the formula.
- Students may forget to square-root the correct variance to obtain the standard deviation.
- Students will mix up $2X$ with $X_1 + X_2$. Please emphasise to students that these mean two completely different things: $2X$ is doubling the value of an observation, whereas $X_1 + X_2$ is the sum of two independent observations.

Setting up the correct combination of random variables from context can be extremely challenging for students and will require a great deal of practice.

NOTES

Combining questions with the binomial and Poisson distributions will help consolidate the material and embed the SEC further.

UNIT 17: Further Probability

Return to overview

SPECIFICATION REFERENCES

2.2 Represent and interpret probabilities using tree diagrams, Venn diagrams and two-way tables.

2.4 Use and apply the laws of probability to include conditional probability.

4.2 Calculate probabilities and determine expected values, variances and standard deviations for discrete distributions.

4.6 Interpret rectilinear graphical representations of continuous distributions.

5.2 Know methods to evaluate or read probabilities using formula and tables.

6.3 Determine probabilities and unknown parameters with a normal distribution.

11.1 Calculate and use conditional probabilities to include Bayes' theorem for up to three events, including the use of tree diagrams.

12.1 Know the use and validity of distributions which could be appropriate in a particular real world situation: binomial, normal, Poisson and exponential.

18.3 Evaluate probabilities for Poisson and exponential distributions and know the corresponding mean and variance.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics
Probability (Unit 3)
Discrete Random Variables (Unit 4)
The Binomial distribution (Unit 5)
The continuous uniform distribution (Unit 7)
The normal distribution (Unit 7)
The Poisson distribution (Unit 15)

KEYWORDS

addition law, Bayes' Theorem, conditional, events, experiment, intersection, multiplication law, mutually exclusive, outcomes, probability, tree diagram, union, binomial, normal, continuous uniform, Poisson, hypergeometric

UNIT SUMMARY

This topic is an extension of multiple previous units on probability and probability distributions. In Unit 3c, Bayes' theorem was referred to and questions given enough scaffolding could be presented at the Year 1 portion of the course as an exercise in using the multiplication rule for probability. In this unit, we generalise this notion to three events (although the formula is true for k events – students will see at most three). The use of tree diagrams is essential here for students to appreciate how Bayes' Theorem works.

Also seen in this unit is the idea of a "hypergeometric" tree diagram. Although not referred to by name, students may see reference to it on some calculators. This extends the ideas of tree diagrams seen in Unit 3 and students will be required to deduce the correct number of combinations and understand that the probability of a sequence of events in a hypergeometric tree diagram is the same regardless of the order.

Finally, the unit revisits the binomial, normal, Poisson and continuous uniform probability distributions, as well as general discrete random variables, and applies the concept of conditional probability to them.

This topic revises many topics from earlier in the course and may be a good starting point to begin the second year. It is also beneficial to use this time to revise these topics and the underlying theory.

17a. Bayes' Theorem (11.1)

Teaching time
1 hour

OBJECTIVES

By the end of the unit, students should be able to:

- Calculate $\mathrm{P}(B)$ given $\mathrm{P}(B|A_i)$ using a tree diagram (for events A_1, A_2, A_3 and B).
- Use the formula for Bayes' Theorem: $\mathrm{P}(A_j|B) = \frac{\mathrm{P}(A_j)\mathrm{P}(B|A_j)}{\sum \mathrm{P}(A_i)\mathrm{P}(B|A_i)}$.

TEACHING POINTS

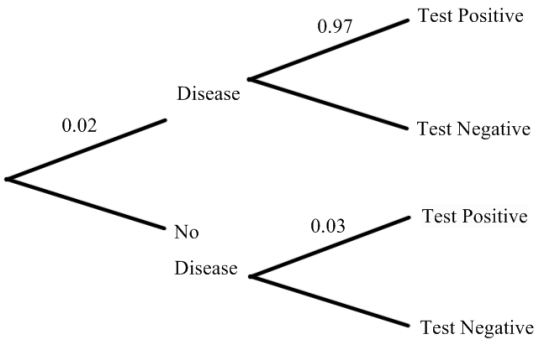
Revise probability (Unit 3). Pay special attention to tree diagrams, conditional probability and the multiplication rule $\mathrm{P}(B|A)\mathrm{P}(A) = \mathrm{P}(A \cap B)$.

To introduce Bayes' theorem, start with two events. The classic example is this:

A rare disease is present in 2% of the population.
A new test for the disease is submitted to the Health Research Authority, claiming to give a positive result 97% of the time if a patient has this disease.
However, the test returns a false positive (testing positive when the patient does not have the disease) 3% of the time.

A patient takes this test and tests positive for the disease. What is the probability of the patient having the disease?

It is best tackled using of a tree diagram:



Students need to identify the question: What is the probability of the patient having the disease, given they test positive? Ensure students define events early on: *Let D be the event "has the disease" and let T be the event "tests positive". We want* $\mathrm{P}(D|T)$.

Students may want to give the answer $\mathrm{P}(T|D)$ or $\mathrm{P}(D \cap T)$ – this is a good opportunity to remind them that they represent different things. Remind students that the multiplication rule can be used.

$$\mathrm{P}(D|T) = \frac{\mathrm{P}(D \cap T)}{\mathrm{P}(T)}$$

Using the tree diagram, students can easily calculate $\mathrm{P}(D \cap T) = 0.0194$.

Remind students that in order to calculate $\mathrm{P}(T)$, they must add up all possibilities in the tree diagram resulting in a positive test: This is "Disease and positive test" or "no disease and positive test". Mathematically: $\mathrm{P}(D \cap T) + \mathrm{P}(D' \cap T) = 0.02 \times 0.97 + 0.98 \times 0.03 = 0.0488$.

Hence $\mathrm{P}(D|T) = 0.0194 \div 0.0488 = 0.3975$. This is an example of Bayes' Theorem.

Draw attention to the fact that $\mathrm{P}(D \cap T) = \mathrm{P}(D)\mathrm{P}(T|D)$ and $\mathrm{P}(D' \cap T) = \mathrm{P}(D')\mathrm{P}(T|D')$. You can then show them the formula

$$\mathrm{P}(D|T) = \frac{\mathrm{P}(D \cap T)}{\mathrm{P}(D)} = \frac{\mathrm{P}(D)\mathrm{P}(T|D)}{\mathrm{P}(D)\mathrm{P}(T|D) + \mathrm{P}(D')\mathrm{P}(T|D')},$$

which looks horrific, but uses only values easily obtainable from the question, namely $\mathrm{P}(D)$, $\mathrm{P}(D')$, $\mathrm{P}(T|D)$ and $\mathrm{P}(T|D')$.

You can then explain this can generalise to more than two events, and the general result is:

$$\mathrm{P}(A_j|B) = \frac{\mathrm{P}(A_j)\mathrm{P}(B|A_j)}{\mathrm{P}(A_1)\mathrm{P}(B|A_1) + \mathrm{P}(A_2)\mathrm{P}(B|A_2) + \cdots + \mathrm{P}(A_n)\mathrm{P}(B|A_n)} = \frac{\mathrm{P}(A_j)\mathrm{P}(B|A_j)}{\sum \mathrm{P}(A_i)\mathrm{P}(B|A_i)}.$$

There will only be a maximum of 3 events involved for this specification. Students may present a solution using either the formula for Bayes' Theorem or using a tree diagram to be awarded full marks (in line with the mark scheme).

When practising the formula for Bayes' Theorem, start with non-contextualised examples, e.g. just giving the probabilities and substituting the numbers into the formulae. Then start introducing context.

Exemplar

Automated teller machines (ATMs) are sometimes found to be out of order and in need of repair.
There are three branches in a region where a particular bank's ATMs can be found.
The head office is responsible for managing repair requests.
Information found about the bank's ATMs and repair requests is given in the table.

Branch	Percentage of bank's ATMs in branch	Probability of ATM repair request
Town A	40%	0.03
Town B	35%	0.01
Town C	25%	0.02

Head office would like to answer the following question:
If a request for an ATM repair arrives, what is the probability that the request originated in the branch located in Town C?

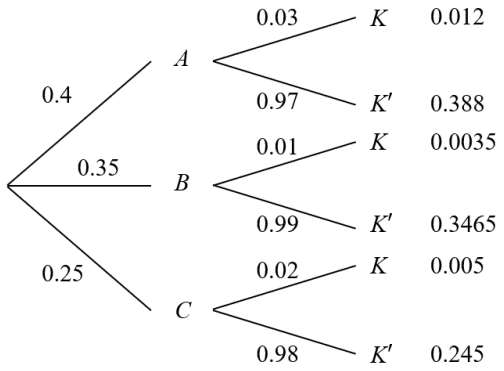
Let A be the event "the branch is in Town A", let B be the event "the branch is in Town B", let C be the event "the branch is in Town C"
Let K be the event "the ATM is broken".

So $\mathrm{P}(A) = 0.4$, $\mathrm{P}(B) = 0.35$, $\mathrm{P}(C) = 0.25$, $\mathrm{P}(K|A) = 0.03$, $\mathrm{P}(K|B) = 0.01$ *and* $\mathrm{P}(K|C) = 0.02$.

Method 1: Using Bayes' Theorem formula

$$\mathrm{P}(C|K) = \frac{\mathrm{P}(C)\mathrm{P}(K|C)}{\mathrm{P}(A)\mathrm{P}(K|A) + \mathrm{P}(B)\mathrm{P}(K|B) + \mathrm{P}(C)\mathrm{P}(K|C)}$$
$$= \frac{0.25 \times 0.02}{0.4 \times 0.03 + 0.35 \times 0.01 + 0.25 \times 0.02}$$
$$= 0.244$$

Method 2: Using a tree diagram



The probability that the ATM is broken is $P(K) = 0.012 + 0.0035 + 0.005 = 0.0205$.
The probability that the ATM is broken in Town C = $P(C \cap K) = 0.005$
So $P(C|K) = \frac{0.005}{0.0205} = 0.244$

OPPORTUNITIES FOR EMBEDDING THE SEC

Relating the probabilities calculated to the context of the question should be encouraged as much as possible. In the disease example above, a follow-up question may be added:

Exemplar

Comment on the effectiveness of the new test

There is a 40% chance of having the disease if the test comes back positive. This shows that the test is not effective and should be rejected by the Health Research Authority. However, this may often be as good as a test may get for an initial test. Patients may then have to undergo more detailed and expensive tests.

COMMON AND POSSIBLE MISTAKES

- Students may input the incorrect values into the formula - plenty of practice will be beneficial.
- Students may also misinterpret $\mathrm{P}(A|B)$ as $\mathrm{P}(A \cap B)$, as well as any mistakes identified in Unit 3.
- When students are asked to interpret the conditional probabilities, the incorrect probabilities are referenced.

17b. Hypergeometric Trees (2.2) (2.4) (11.1)

Teaching time
1 hour

OBJECTIVES

By the end of the unit, students should be able to:

- Model a hypergeometric situation by a tree diagram.
- Identify the possible number of combinations of a given set of unordered events occurring.
- Calculate probabilities from a hypergeometric situation either by tree diagram or by manual calculation.

TEACHING POINTS

This is an extension of Unit 3 tree diagrams but is a specific case of sampling without replacement (i.e. the denominator decreases by 1 each time). Students are not expected to know the phrase "hypergeometric" although some graphical calculators have this mode available.

Students are expected to appreciate that the probability of a given set of unordered events occurring is the same as the probability of a given set of ordered events occurring multiplied by the number of ways of ordering these events. The knowledge and use of $^{n}C_{r}$ is not required and students will be expected to manually calculate the number of combinations. Due to time restrictions in the examination, there will be a natural limit on the number of combinations that can be calculated.

Exemplar

Employees of a company can either be classified as factory workers or office workers. In one particular company, there are 35 office workers and 68 factory workers. A simple random sample of 4 employees is chosen from the company. Find the probability that two of the sample are office workers and the other two are factory workers.

Let X be "worker chosen is a factory worker" and Y be "worker chosen is an office worker". There are a total of 103 employees in the company.

Considering one particular combination (XXYY), the probability of choosing this combination is $\frac{35}{103} \times \frac{34}{102} \times \frac{68}{101} \times \frac{67}{100} = 0.05109$ ...

There are 6 ways of ordering this combination: XXYY, XYXY, XYYX, YXXY, YXYX, YYXX

So the probability of selecting 2 office workers and 2 factory workers in any order is $6 \times 0.05109 = 0.307$ *(3 s.f.)*

Exemplar

A nutritionist is investigating the effects of vitamins A, C and D on heart disease. A clinical study of 150 patients was carried out where each patient took one extra vitamin supplement to their usual diet. 35 patients were given extra supplements of vitamin A, 65 patients were given extra supplements of vitamin C and the rest were given extra supplements of vitamin D.

The nutritionist wishes to interview some patients for a follow-up study and takes a simple random sample of 3 patients.

What is the probability there is a patient from each vitamin supplement group in the sample?

Let A be "the patient chosen took vitamin A", C be "the patient chosen took vitamin C" and D be "the patient chosen took vitamin D".

$\mathbf{150 - 35 - 65 = 50}$ *patients took the extra supplement of vitamin D.*

Considering one particular combination (ACD), the probability of choosing this combination is $\frac{35}{150} \times \frac{65}{149} \times \frac{50}{148} = \mathbf{0.034388}$ ...

There are 6 ways of ordering this combination: ACD, ADC, CAD, CDA, DAC, DCA

So the probability of a patient of each vitamin supplement group being in the sample is $\mathbf{6 \times 0.034388 ... = 0.206}$ *(3.s.f)*

OPPORTUNITIES FOR EMBEDDING THE SEC

This topic links in with data collection methods, specifically simple random samples as well as proportional/disproportional stratified samples.

COMMON AND POSSIBLE MISTAKES

- Any common mistake from Unit 3.
- Adding probabilities instead of multiplying.
- Choosing a binomial distribution instead of identifying a hypergeometric situation.
- Failing to count all possible combinations.

NOTES

Students can answer this either by drawing a tree diagram or using the "shortcut" of "probability of one route × number of routes". Be aware that the answer space given in exams may not allow for the construction of a full tree diagram.

17c. Conditional Probability Distributions (2.4) (4.2) (4.6) (6.3) (12.1) (18.3)	Teaching time 1 hour

OBJECTIVES

By the end of the unit, students should be able to:

- Calculate conditional probabilities from a tabulated probability distribution of a discrete random variable, the normal distribution, the binomial distribution, the Poisson distribution, the continuous uniform distribution and the exponential distribution (see Unit 20b)

TEACHING POINTS

This topic combines the concepts of conditional probability with the probability distributions currently seen so far.

Students may need to identify conditional probabilities within questions either by symbols or in contextual examples. These conditional probabilities can be calculated either by use of the multiplication rule $\left(\mathbf{P}(\boldsymbol{A}|\boldsymbol{B}) = \frac{\mathbf{P}(\boldsymbol{A}\cap\boldsymbol{B})}{\mathbf{P}(\boldsymbol{B})}\right)$ or by using appropriate diagrams and restricting the sample space to the conditional event.

Begin with non-contextual examples first:

Exemplar

The discrete random variable X follows the probability distribution shown below:

x	0	5	10	20	40	100
$P(X = x)$	0.1	0.3	0.2	0.1	0.1	0.2

Find $\mathbf{P}(X \leq 40 \mid X \geq 10)$

Using the multiplication rule: $P(X \leq 40|X \geq 10) = \frac{P(10\leq X\leq 40)}{P(X\geq 10)} = \frac{0.2+0.1+0.1}{0.2+0.1+0.1+0.2} = \frac{2}{3}$

Alternatively: $P(X \geq 10) = 0.2 + 0.1 + 0.1 + 0.2 = 0.6$.
The probability that $X \leq 40$ *out of these options is* $0.2 + 0.1 + 0.1 = 0.4$.
So $P(X \leq 40|X \geq 10) = \frac{0.4}{0.6} = \frac{2}{3}$

Exemplar

Let $X \sim B(30, 0.24)$. Find $P(X \leq 12 \mid X > 6)$.

Using the multiplication rule:

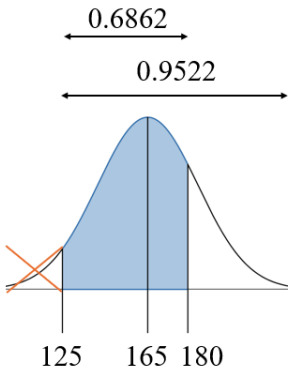
$$P(X \leq 12 \mid X > 6) = \frac{P(6 < X \leq 12)}{P(X > 6)} = \frac{P(X \leq 12) - P(X \leq 6)}{1 - P(X \leq 6)} = \frac{0.9845 - 0.3961}{1 - 0.3961}$$
$$= 0.974$$

Contextual examples can soon follow:

Exemplar

The heights of theme-park visitors are known to be normally distributed with a mean of 165 cm and a standard deviation of 24 cm. In order to ride the rollercoaster, a theme-park visitor must be at least 125 cm. Given that a theme park visitor is allowed to ride the roller-coaster, find the probability that they are shorter than 180 cm.

Let X be the height of a theme park visitor. $X \sim N(165, 24^2)$.
The probability required is $P(X < 180 \mid X \geq 125)$.
Using the multiplication rule: $P(X < 180 | X \geq 125) = \frac{P(125 \leq X < 180)}{P(X \geq 125)} = \frac{0.6862}{0.9522} = 0.721$



OPPORTUNITIES FOR EMBEDDING THE SEC

Relating the conditional probabilities to real-world scenarios should be encouraged as much as possible. Questions involving the support or violation of the assumptions required for a particular probability distribution is a good way of embedding the SEC.

COMMON AND POSSIBLE MISTAKES

- Failing to spot the conditional nature of the question.
- When using the calculator, inputting incorrect parameters when finding probabilities.

NOTES

Although not explicitly mentioned in the specification, this sub-unit is a combination of specification point 2.4 to all probability distributions. Only the exponential distribution is omitted here as it will be seen in Unit 20b (and in some cases, the memorylessness property may be used).

UNIT 18: Confidence Intervals and the Central Limit Theorem

Return to overview

SPECIFICATION REFERENCES

14.2 Know the use of the central limit theorem in the distribution of $\overline{X}$ where the initial distribution, X, is not normally distributed and the sample is large.

15.1 Use confidence intervals for the mean using z or t as appropriate, interpreting results in practical contexts.

15.2 Know that a change in sample size will affect the width of a confidence interval.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics
Normal Distribution (Unit 7)
Sampling Distribution of the Mean (Unit 9)

KEYWORDS

central limit theorem, confidence interval, estimate, mean, normal distribution, population, sample size, sample, sampling distribution of the mean, standard deviation, standard error, *t*-distribution, unbiased estimate, variance

UNIT SUMMARY

Some students find it difficult to find the confidence interval, let alone interpret it. The confidence intervals for normal distribution or confidence intervals for t-distribution activity in Desmos will help students visualise a confidence interval.

This is also the first time students will encounter the Student's *t*-distribution. The *t*-distribution activity in Desmos will show the differences between the *t*-distribution and the normal distribution, and show that the *t*-distribution tends towards the normal distribution as the number of degrees of freedom (or sample size) increases. As a historical footnote, a history of William Gosset (1876-1937) and why he penned under the pseudonym "Student" can be mentioned.

The Central Limit Theorem is the other concept in this unit. It is difficult to create a Desmos activity to help students see this but other software illustrating this theorem is easily available on the internet. The use of the Central Limit Theorem in this unit is purely to find confidence intervals using data from populations that may not have underlying normal distributions.

Activities in Desmos that can aid teachers and students are: Sampling distributions, Binomial distribution and normal approximation, Poisson distribution.

18a. Confidence Intervals and the Central Limit Theorem: Confidence Intervals for the mean of a Normal Distribution with known variance (15.1)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Appreciate the concept of a confidence interval
- Find a confidence interval for the mean of a Normal distribution with known variance
- Interpret a confidence interval in context

TEACHING POINTS

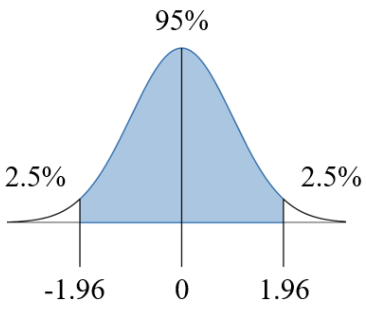
A $\alpha\%$ confidence interval is a range of values such that the probability that a randomly selected interval contains the population parameter is $\alpha\%$.

Revise the normal distribution. Pay special attention using the inverse standard normal distribution to find a where $\mathrm{P}(-a \leq X \leq a) = C$. Also revise standardising a normal distribution from $X \sim \mathrm{N}(\mu, \sigma^2)$ to $Z \sim \mathrm{N}(0,1^2)$ (Unit 7c). Although not necessary for finding a confidence interval in this sub-unit, it will be necessary in Unit 18c.

Remind students (from Unit 7) that the population mean is not always known but $\bar{x}$ is an unbiased estimate for μ. We can use the unbiased estimate $\bar{x}$ to find a confidence interval.

There are a few things students may be reminded of: as $\bar{x}$ changes over different samples, the resulting distribution is $\bar{X}$; if $X \sim \mathrm{N}(\mu, \sigma^2)$ then $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$; if $\bar{x}$ is being considered then $\bar{X}$ should be used; $\bar{x}$ can be transformed into the standard normal distribution using the transformation $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$.

Using the standard normal distribution, students are expected to use their calculators to determine a such that $\mathrm{P}(-a \leq Z \leq a) = 0.95$. A sketch and/or the Desmos activity may help here.



Emphasise that $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \pm 1.96$ in this case. Rearranging the formula gives $\mu = \bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$, and these are the upper and lower bounds on a 95% confidence interval. Allow students to find the similar numbers for 90% and 99% confidence intervals.

In general, for an α% confidence interval, confidence intervals are $\left(\bar{x} - z\frac{\sigma}{\sqrt{n}},\ \ \bar{x} + z\frac{\sigma}{\sqrt{n}}\right)$ where $2 \times \mathrm{P}(Z \leq -z) = \alpha\%$ and $Z \sim \mathrm{N}(0,1^2)$.

Begin with numerical, non-contextual examples. All examples given at this stage should also provide the variance or standard deviation, and assume a normal distribution.

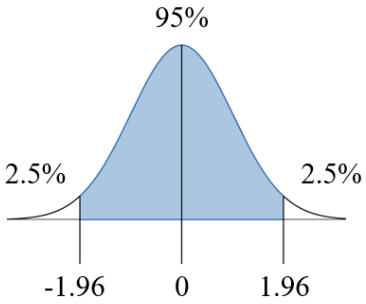
Exemplar

A sample of size 8 is taken from a population which is normally distributed with a standard deviation of 7.

481 455 468 457 469 463 469 458

Calculate a 95% confidence interval for the mean.

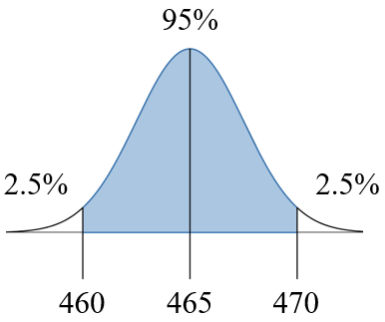
Let X be a random variable, $X \sim \mathrm{N}(\mu, 7^2)$, so $\bar{X} \sim \mathrm{N}\left(\mu, \frac{7^2}{8}\right)$. Using a calculator $\bar{x} = 465$.



A 95% confidence interval is $\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}$ *to* $\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$.

So the confidence interval is $\left(465 - 1.96 \times \frac{7}{\sqrt{8}}, 465 + 1.96 \times \frac{7}{\sqrt{8}}\right) = (460, 470)$

Encourage students to use the interval notation (a, b) to denote $a \leq \mu \leq b$.
Students can also check that these values are correct by finding c such that $\mathrm{P}(\mu - c \leq X \leq \mu + c) = 0.95$ from $\mathrm{N}\left(\bar{x}, \frac{\sigma^2}{n}\right)$.



At this point, encourage students to explore what happens to a confidence interval as n changes. The mathematically able will be able to identify this from the formula, but other students will need to experiment with examples. Students need to appreciate that the confidence interval gets narrower as the sample size increases, and wider if it decreases.

Two methods for calculating a confidence interval are presented here. Students will be able to access full marks (in line with the mark scheme) whichever method is presented.

Start introducing context into the questions – this will allow students to practise extracting information from the question. It will also give an opportunity for students to interpret the confidence in context. Students should always avoid giving stock answers such as "*The 95% confidence interval is the range of values in which we can be 95% confident that the true mean lies*" and always interpret the results in context.

Students will be expected to know that if a claimed mean lies inside the confidence interval, then there is insufficient evidence to suggest the claim is false. Conversely, if a claimed mean lies outside the confidence interval, then there is significant evidence to suggest the claim is false.

This can extend to overlapping confidence intervals. If two confidence intervals are constructed (one from each of two different populations) and overlap, then there is insufficient evidence to suggest the population means are different. Conversely, if the two confidence intervals do not overlap, then there is significant evidence to suggest the population means are different.

Such conclusions should be made in context.

Exemplar

A food processing factory produces large batches of jam.
In each batch the weight of jam in a jar can be modelled by a normal distribution with standard deviation 7 g.
The weights, in g, of the jam in a random sample of jars from a particular batch were:

481 455 468 457 469 463 469 458

a) Calculate a 95% confidence interval for the mean weight of jam in this batch of jars.

Let X be the weight of jam in a jar.
This question is a contextualised version of the previous example, so the confidence interval was calculated earlier as $(460,470)$ grams.

b) The factory claims that the weight of jam in each jar is 472 g.
Comment on this claim as it relates to this batch.

- *472 g is higher than the upper boundary of the 95% confidence interval for the population mean weight.*
- *This indicates that there is significant evidence to suggest the **mean** weight is not 472 g, as claimed.*
- *There are also 7 out of the 8 recorded weights in the sample below 472g.*

OPPORTUNITIES FOR EMBEDDING THE SEC

A1 Identifying the factors in the problem, leading to the use of the correct distribution (e.g. distinguishing between a confidence interval and the limits in which a single observation lies.)

A4 Appreciating that exploratory data analysis may needed to find the population variance.

C1 Calculating numerical measures from a sample.

D1 Interpreting confidence intervals in context.

D2 Interpreting confidence intervals in context and relating them to an initial question.

D4 Appreciating the level of uncertainty in a confidence interval.

D5 Reaching conclusions in context appropriate for a given target audience.

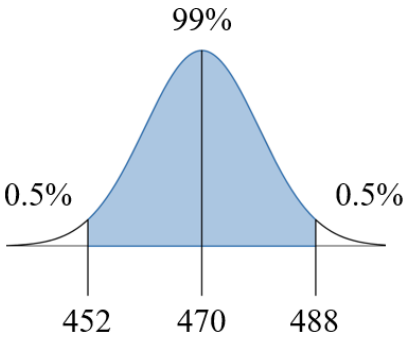
The above example can be modified to include the following question:

Exemplar

c) Assuming that the population mean weight of jam in a jar is in fact 470 g, calculate the limits within which 99% of weights of jam in these jars lie.

*Since we are looking at the weight of each individual jar (**not** a sample mean), we will use* $\mathrm{N}(470, 7^2)$.

We want a *and* b *such that* $\mathrm{P}(X \leq a) = 0.995$ *and* $\mathrm{P}(X \leq b) = 0.005$.



Using the calculator, $a = 488$ *and* $b = 452$*. So most individual jars will lie between 452 g and 488 g.*

d) Further comment on the factory's claim.

Although the mean weight of a jar of jam is highly likely to be lower than the claimed weight, it is still possible that individual jars of jam could weigh more than 472 g.

It is possible to find the smallest sample size needed for a given interval to be a $k\%$ confidence interval. This may involve rearranging equations and some algebra, although solutions obtained from the equation solver on the calculator or trial and error is equally valid and could gain full marks (in line with the mark scheme).

Exemplar

e) **How large a sample would be needed in order that the central 90% of sample means would lie in an interval of width at most 1 g?**

This time, as the sample size is unknown, $X \sim \mathrm{N}\left(\mu, \frac{7^2}{n}\right)$.

A 90% confidence interval is $\left(\bar{x} - 1.6449\frac{\sigma}{\sqrt{n}},\ \bar{x} + 1.6449\frac{\sigma}{\sqrt{n}}\right)$.

So the width of this confidence interval is $2 \times 1.6449 \times \frac{7}{\sqrt{n}}$.

We require that the width is at most 1.

Hence $2 \times 1.6449 \times \frac{7}{\sqrt{n}} \leq 1$.

Rearranging gives $2 \times 1.6449 \times 7 \leq \sqrt{n}$ *and so* $23.0286 \leq \sqrt{n}$.

Hence $530.3 \leq n$, *so the sample size must be at least 531.*

Alternatively, the value of 530.3 can be obtained from the equation solver on the calculator and the value of 531 could be obtained by trial and error.

COMMON AND POSSIBLE MISTAKES

- Any mistake listed in Units 7 and 9.
- Using σ instead of $\frac{\sigma}{\sqrt{n}}$
- Forgetting to square root the n
- Using σ^2 instead of σ.
- Misidentifying whether to use X or $\bar{X}$ (i.e. determining a confidence interval for μ, or finding limits between which k% of observations lie)
- Misreading or misinterpreting the question.

The biggest mistake is interpreting (for example) a 95% confidence as “the probability that μ lies in the interval is 0.95”. If 100 random samples of the same size were taken, and a 95% confidence interval constructed for each of them, then on average 95 out of the 100 intervals would contain the population mean. You, however, only have one such interval, and have no way of knowing whether your interval is one of the 95 which contain the mean. All you can say is that the probability that the procedure for the confidence interval will have a 95% chance of generating an interval that contains the mean.

NOTES

There are two ways to find the confidence interval: using $\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$ where z is the appropriate value from $\mathrm{N}(0,1^2)$, or using the normal distribution mode on the calculator with $\bar{x}$ and $\frac{\sigma}{\sqrt{n}}$. Both are valid methods, however the former will help when Unit 18c is taught.

18b. Confidence Intervals and the Central Limit Theorem: The Central Limit Theorem (14.2)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- State the Central Limit Theorem.
- Apply the Central Limit Theorem.
- Recognise when the Central Limit Theorem can be used in context.

TEACHING POINTS

The central limit theorem states that for any random variable X with any underlying distribution (discrete or continuous), mean μ and variance σ^2, if the sample size is sufficiently large then the sampling distribution of the mean $\bar{X}$ is approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$. Also, if the sample size is sufficiently large then the sample variance s_x^2 can be used as a best estimate for σ^2.

“Sufficiently large”, as a rule of thumb for this course, is $n \geq 30$.

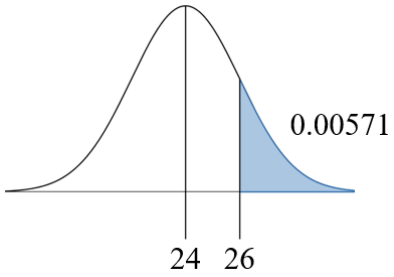
You can use the binomial distribution activity and the Poisson distribution activity to overlay the normal approximations over the top (although the latter is not in the defined content). Remind students of the conditions for a normal approximation to the binomial to illustrate the importance of the sample size. It is, in general, difficult to program in to Desmos the sampling distribution of the mean for any underlying distribution. However, there are plenty of articles and software on the internet one can use to illustrate the Central Limit Theorem.

Start by revising questions from Unit 9, but this time removing the condition that the underlying population being normally distributed.

Exemplar

The plums from a particular variety of plum tree are known to have mean mass 24 g and standard deviation 5 g.

a) **40 plums are selected at random.**
What is the probability that the mean mass of these 40 plums exceeds 26 g?
Let X be the mass of a plum. Let $\bar{X}$ be the mean mass of a plum from a sample of size 40.
So $\bar{X} \sim \mathrm{N}\left(24, \frac{5^2}{40}\right)$.



Probability that the mean mass of a plum exceeds 26 g : $\mathrm{P}(\bar{X} > 26) = 0.00571$.

b) **State any assumptions you have made about the distribution of the population of the masses of plum, giving reasons.**
No assumptions have been made about the underlying distribution of X.
By the Central Limit Theorem, since n is large (greater than 30), so we may assume $\bar{X}$ to be approximately normal distributed with mean 24 and variance $\frac{5^2}{40}$.

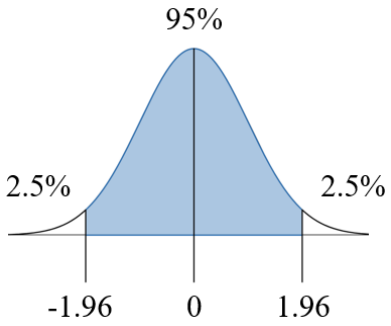
Ensure students get plenty of practice at these revised "Unit 9"-style questions. This will help in Unit 19. Students also need to see questions where the underlying populations are normally distributed and appreciate that in these cases, the Central Limit Theorem is not needed.

Once students have had enough practice at applying the Central Limit Theorem, return to confidence intervals. The premise remains exactly the same: if the sample size is sufficiently large, then $\bar{X}$ may be assumed to be normally distributed and so the normal distribution can be used to find confidence intervals (Unit 18a).

Exemplar

A Bhuna curry is a traditional dish from western Bangladesh and northeast India. The contents of a random sample of 80 tins of Bhuna curry sauce from a large batch were weighed. The sample mean content was 284.2 g and the standard deviation was found to be 4.9 g.

a) **Calculate a 95% confidence interval for the population mean.**
Let X be the content, in grams, of a tin of Bhuna curry sauce.
Since the sample size is large, the variance σ^2 can be estimated by 4.9^2 and we may assume that $\bar{X} \sim \mathrm{N}\left(\mu, \frac{4.9^2}{80}\right)$.
Using the standard normal distribution, the value z such that $\mathrm{P}(-z \leq Z \leq z) = 0.95$ is z=1.96.



So the 95% confidence interval for μ is
$\left(284.4 - 1.96 \times \frac{4.9}{\sqrt{80}}, 284.4 + 1.96 \times \frac{4.9}{\sqrt{80}}\right) = (283.3, 285.5)$

b) **How would your answer to (a) be affected if it were later discovered that the batch contained tins which had been filled by two different machines and the distribution of the weights was bimodal? Give a brief justification of your answer.**
The confidence interval would not be affected since the sample size is large, so the Central Limit Theorem says that $\bar{X}$ can be assumed to be normally distributed with a mean of μ and a variance of $\frac{4.9^2}{80}$, even if X is not normally distributed.

c) **How would your answer to (a) be affected if a random sample of 15 tins were taken instead of 80? Give a brief justification of your answer.**
The confidence interval would be affected/not valid since the sample size is not large, so we cannot assume a normal distribution for $\bar{X}$.

d) **How would your answer to (a) be affected if it were later discovered that the sample had not been taken at random?**
Any conclusions made about the confidence interval would be invalid since the sample must be obtained randomly.

Note that in part (c), any reference to using the t-distribution is incorrect since the t-distribution relies on the underlying population being normally distributed, but the variance not known – such comments are more likely to occur after Unit 18c.

OPPORTUNITIES FOR EMBEDDING THE SEC

A1 Identifying factors related to the problem to identify when the Central Limit Theorem can be applied.

B1 Identifying practical constraints when taking a large sample.

B4 Appreciating that a sample must be obtained randomly for conclusions to be valid.

C1 Calculating numerical measures about a sample.

C2 Calculating with summary statistics generated by technology.

D1 Interpreting numerical measures in context.

D2 Interpreting numerical measures and reaching conclusions about an initially defined question.

D4 Justifying the reliability of findings referring to any assumptions made.

D5 Reaching conclusions in a language appropriate for a given target audience.

E2 Recognising that the sample size must be sufficiently large to apply the Central Limit Theorem.

Questions incorporating sampling methods or data representation will further embed the SEC.

COMMON AND POSSIBLE MISTAKES

See the mistakes in Unit 18a. Also:

- Assuming a normal distribution for a population, and not using the Central Limit Theorem to determine whether the sampling distribution of the mean is normally distributed
- Using the Central Limit Theorem when the population is normally distributed, negating the need to use the Central Limit Theorem.
  e.g. Considering the example above relating to Bhuna curry sauce, stating $X \sim \mathrm{N}\left(284.4, \frac{4.9^2}{80}\right)$ when the population mean is unknown.

NOTES

Recall that the sample variance is an unbiased estimate for the population variance, meaning that on average the sample variance is equal to the population variance. This does not necessarily mean that a single sample variance is a “good” estimate for the population variance. However, if the sample is large, the error between a sample variance and a population variance is likely to be low which is why a simple sample variance may be used as a “good” estimate for the population variance. This is not necessarily true for the sample mean as an estimate for the population mean. Students are not expected to know this but may give a good justification for the inquisitive minded.

18c. Confidence Intervals and the Central Limit Theorem: Confidence Intervals for the mean of a Normal Distribution with unknown variance (15.1)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Appreciate that if the population distribution is normal and the variance is unknown, then the *t*-distribution should be used.
- Find a confidence interval for the mean of a Normal distribution with unknown variance (using the t-distribution).
- Interpret a confidence interval in context.
- Appreciate the advantages and disadvantages of using a t-distribution for a confidence interval.

TEACHING POINTS

In this sub-unit, the population variance is not known. Students need to know that if the population variance is unknown, then the sample variance is used as an estimate instead.

Remind students that as $\bar{x}$ changes over different samples, then the resulting distribution is $\bar{X}$. From Unit 9, remind students that the sampling distribution of the mean is one type of sampling distribution, and the sampling distribution of the standard deviation, S, also exists. Use the Sampling Distributions activity in Desmos to help. It can be explained to students that as the sample standard deviation s_x changes over different samples, the resulting distribution is the sampling distribution of the standard deviation, represented by the random variable S. Stress that if the underlying population is normally distributed, S is not normally distributed (unlike $\bar{X}$).

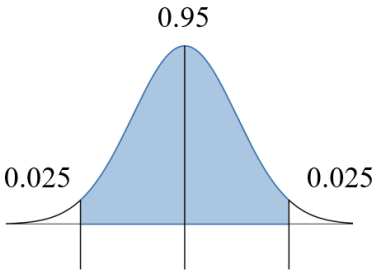
Remind students that as $\bar{x}$ changes over different samples, $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ follows a normal distribution of mean 0 and standard deviation 1. Emphasise that in order to use this, σ had to be known. If σ is unknown, then the sample standard deviation s_x can be used to estimate σ. Emphasise the importance of the division by $n-1$, since (from Unit 9) s_x^2 is an unbiased estimate for σ^2. Remind students that s_x is a biased estimate for σ.

It is here where the t-distribution is introduced. Students do not need to know the technical details behind the t-distribution, but a basic understanding would be beneficial. For an underlying normal distribution, as $\bar{x}$ and s_x change over different samples, $\frac{\bar{x}-\mu}{\frac{s_x}{\sqrt{n}}}$ doesn't necessarily follow a normal distribution because there are two things that are changing. The distribution it follows is called the Student's t-distribution. Use the *t*-distribution activity on Desmos to illustrate the similarities and differences. If the sample size is *n*, then the distribution has $n-1$ degrees of freedom – emphasise that the sample size changes the shape of the t-distribution. Show students that the *t*-

distribution tends towards the normal distribution as the sample size gets larger, and refer to the Central Limit Theorem.

Explain that although the *t*-distribution shares a lot of properties as the normal distribution, it is not the normal distribution: it is more complicated to work with and some calculators do not calculate the probabilities.

Finding the confidence intervals using the t-distribution is largely the same as the previous sub-unit, except the percentage points for the t-distribution should be used instead of those from the standardised normal distribution. In general, a confidence interval using the t-distribution is $\left(x - t\frac{s_x}{\sqrt{n}}\ ,\ x + t\frac{s_x}{\sqrt{n}}\right)$ where t is the appropriate standard critical value given in the tables.
Use the Confidence Intervals (t-distribution) activity on Desmos to help. Using the percentage points table, ensure students draw a sketch and refer to the sketch in the formula book: for a 95% confidence interval, the 0.975 column should be used.



Follow similar steps as in the last sub-unit: start with non-contextual examples to practise the skills followed by contextual examples later.

Exemplar

The lengths of components made by a machine are normally distributed. A random sample of components had lengths, in cm:

1.002 1.007 1.016 1.009 1.003 1.006

a) **Calculate a 95% confidence interval for the population mean.**
*Let X be the lengths of components made by a machine. Then $X \sim \mathrm{N}(\mu, \sigma^2)$ and $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Since the variance is **unknown**, we use the t-distribution.*
From the calculator: $\bar{x} = 1.007$ and $s_x = 0.00504$.
Since $n = 6,\ \nu = 5$.
From the tables, the standard critical value is 2.571.

So the confidence interval is
$\left(1.007 - 2.571 \times \frac{0.00504}{\sqrt{6}}, 1.007 + 2.571 \times \frac{0.00504}{\sqrt{6}}\right) = (1.002, 1.012)$
[Some calculators can calculate *t*-intervals: students who have this functionality may gain full marks in an exam (in line with the mark scheme).]

b) **State the probability that a randomly selected such interval does not contain the population mean.**
0.05

Allow students to compare confidence intervals using the normal distribution and the t-distribution, stating the relevant assumptions:

Exemplar

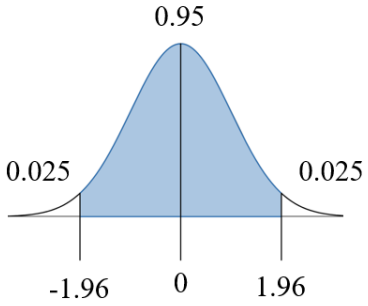
Sodium helps maintain blood pressure and regulates the body's fluid balance. The sodium level in the blood is measured in milliequivalents per litre (mEq/l). In both males and females, the blood sodium levels are known to be normally distributed.
The blood sodium levels in males are believed to have standard deviation 8.45 mEq/l.

A random sample of 12 females was obtained, and their blood sodium levels, measured to the nearest mEq/l, were as follows:

139 146 137 144 143 136 141 148 135 149 145 149

a) Find a 95% confidence interval for the mean blood sodium level of a female:

i. **assuming that the population standard deviation is 8.45 mEq/l (as for a male)**
Let X be the blood sodium levels in a female. Then $X \sim \mathrm{N}(\mu, \sigma^2)$ and $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
If $\sigma = 8.45$, then we can use the normal distribution to find the confidence interval. Using $\mathrm{N}(0,1^2)$, the value of z such that $\mathrm{P}(-z \leq Z \leq z) = 0.95$ is z= 1.96.



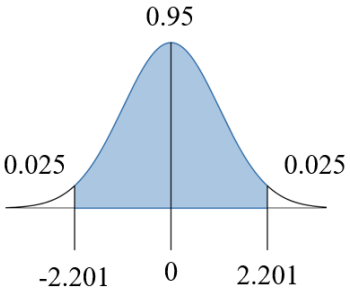
From the sample, $\bar{x} = 142.7$ and $n = 12$
So the 95% confidence interval is
$\left(142.7 - 1.96 \times \frac{8.45}{\sqrt{12}},\ 142.7 + 1.96 \times \frac{8.45}{\sqrt{12}}\right) = (137.9,\ 147.4).$

ii. making no assumption about the standard deviation for blood sodium levels in females.

If σ is unknown, then we use s_x as an estimate and use the t-distribution to find the confidence interval.

From the sample, $s_x = 5.033$, $n = 12$ and $\nu = 11$.

From the tables, the standard critical value for t is 2.201.



So the 95% confidence interval is

$$\left(142.7 - 2.201 \times \frac{5.033}{\sqrt{12}},\ \ 142.7 + 2.201 \times \frac{5.033}{\sqrt{12}}\right) = (139.5,\ 145.9)$$

b) Briefly give one argument in favour of using the confidence interval calculated in (a)(i) and one argument in favour of the confidence interval calculated in (a)(ii).

The confidence interval in part (i) uses the population standard deviation for males and, if the assumption that the standard deviation is the same for females is valid, gives a more dependable confidence interval for the mean.

The confidence interval in part (ii) does not make any assumption about the standard deviation, and so is valid even if the first assumption turned out to be false.

As n tends to infinity, the t-distribution tends towards the normal distribution. If the sample size is large enough, critical values from the normal distribution may be used instead of those from the t-distribution. "Large enough" in this context can mean "$n \geq 30$".

The properties of *t*-confidence intervals are slightly different to the *z*- counterparts. If a larger sample is taken, then the *z*-interval will become narrower / more precise. However, this may not be true for a *t*-confidence interval since a different sample may yield a different standard deviation which is large enough to counter the effect of taking a larger sample. This can be seen by the $t \times \frac{s_x}{\sqrt{n}}$ part of the formula.

OPPORTUNITIES FOR EMBEDDING THE SEC

A1 Identifying the factors in the question that lead to the use of the correct distribution.

A4 If the normal distribution is chosen instead of the t-distribution, awareness of the need for preliminary, exploratory data analysis first.

B1 Justifying the use of a t-distribution if practical issues arise when taking large samples.

C1 Calculating numerical measures from sample data.

D1 Interpreting the meaning of a confidence interval.

D2 Interpreting the meaning of a confidence interval in relation to an initial question.

D4 Discussing the reliability of findings, stating any assumptions made.

D5 Reaching conclusions in a language appropriate for a target audience.

E2 Justifying the use of a t-distribution in relation to the sample size and acknowledging the disadvantages.

E3 Suggesting ways of overcoming the disadvantages of using the t-distribution (e.g. take larger sample sizes, use exploratory data analysis to make assumptions about the population variance).

Questions that involve describing the sampling methods used to take a sample will further incorporate the SEC.

COMMON AND POSSIBLE MISTAKES

See the mistakes in Units 18a and 18b. Also:

- Using a normal distribution all the time.
- Using s_x to estimate σ and then using the normal distribution to find the critical value.
- Using the incorrect column of the tables to find the critical value (e.g. using the 0.95 column for the 95% confidence interval).
- Forgetting to subtract 1 from the sample size to calculate the number of degrees of freedom.
- Misunderstanding Context, language and communication.

The language regarding the width of the confidence interval must be clear. Students sometimes use “bigger” and “smaller” to refer to “wider” and “narrower” – discourage the use of these descriptors since “bigger” and “smaller” may also refer to the magnitude of the numbers contained in the confidence interval. Use descriptors such as “wider” or “less precise” and “narrower” or “more precise”.

NOTES

If the sample size is large ($n \geq 30$), there is a choice over whether ***z*** or ***t***-confidence intervals can be used. Both are equally valid as a selection. The conditions of the use of a ***t***-distribution is that the underlying population is normal and the variance is unknown. This is irrespective of the sample size. Hence, for example, a sample of size 100 may yield a valid ***t***-confidence interval using 99 degrees of freedom. Some calculators can calculate ***t***-values for 99 degrees of freedom but the tables in the formula book are limited, and so the use of CLT is a preferred method.

It is not true that the ***t***-distribution cannot be used for a large sample; so long as the population is normally distributed with an unknown variance then a ***t***-distribution is valid.

UNIT 19: Concepts in Hypothesis Testing

Return to overview

SPECIFICATION REFERENCES

14.2 Know the use of the central limit theorem in the distribution of $\bar{X}$ where the initial distribution, X, is not normally distributed and the sample is large.

15.3 Evaluate the strength of conclusions and misreporting of findings from hypothesis tests, including the calculation and importance of the power of a hypothesis test.

15.4 Know that sample size can be changed to potentially elicit appropriate evidence in a hypothesis test.

15.5 Interpret Type I and Type II errors, in hypothesis testing and know their practical meaning.

15.6 Calculate the risk of a Type II error.

15.7 Know the difference and advantages of using critical regions or *p*-values as appropriate in real life contexts in all tests in this subject content.

16.1 Know how to apply knowledge about carrying out hypothesis testing to conduct tests for the mean of a normal distribution with unknown variance using the *t* distribution

16.5 interpret results for 16.1 in context.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics

Binomial Distribution (Unit 5)

Normal Distribution (Unit 7)

Sampling Distribution of the Mean (Unit 9b)

Normal Approximation to the Binomial (Unit 9c)

Hypothesis Testing Terminology (Unit 10)

Hypothesis Testing about the mean (Unit 11)

Year 2 of A Level Statistics

The Central Limit Theorem (Unit 18b)

The *t*-distribution (Unit 18c)

KEYWORDS

alternative, approximation, binomial, central limit theorem, critical region, critical value, distribution, error, hypothesis, insufficient, mean, normal, null, population, power, probability, proportion, *p*-value, risk, sample, sampling distribution, significance, significant, standard deviation, standard error, sufficient, *t*-distribution, Type I error, Type II error, variance

UNIT SUMMARY

This unit extends the hypothesis tests about the mean seen in Unit 11. As a reminder, Unit 11 saw the use of the sampling distribution of the mean from a normal distribution with known variance in a hypothesis test (known as a *z*-test). Here we extend *z*-tests to when the underlying population does not have a normal distribution, but the sample size is large enough to invoke the Central Limit Theorem. We also modify the hypothesis test to situations when the underlying population does have a normal distribution, but the variance is unknown and the sample size is not large enough for the sample standard deviation to approximate the population standard deviation reliably (a *t-test*).

The main focus is the terminology and the concepts behind a hypothesis test. In Unit 10, students get a basic idea of the concept of a significance level. In Unit 11, students may have seen both the critical region method or the *p*-value method for conducting a hypothesis test. In this unit, more emphasis is placed on the advantages and disadvantages of *p*-values and critical regions, as well as introducing the concepts of Type I and Type II errors and what they mean in context. The unit finishes with the concept of the power of a hypothesis test.

Desmos activities that would be useful in this unit are: Normal distribution with critical values, *t* distribution, Type I and Type II errors.

19a. Concepts in Hypothesis Testing: Hypothesis tests for a mean using a large sample (14.2)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Carry out a hypothesis test for the mean using a large sample.
- Interpret the results of the hypothesis test in context.

TEACHING POINTS

This sub-unit is mainly revision from Unit 11a. Start with revising hypothesis tests for the mean when the underlying distribution is normal and the variance is known.

After enough practice, introduce the use of the Central Limit Theorem. It is straightforward to modify questions in context where the normal distribution is not assumed, the variance is known and the sample size is large (at least 30). Always encourage students to state whenever they are using the central limit theorem.

The example given in Unit 11a is an example of a question that can be used here. The only alteration to the solution would be instead of "*We will assume that X is normally distributed",* it could say (for example) "*Since the sample size is large, we may assume* $\bar{X} \sim \mathrm{N}\left(60, \frac{3.71^2}{100}\right)$ *by the Central Limit Theorem"*. The rest of the example is as before.

OPPORTUNITIES FOR EMBEDDING THE SEC

A1 Identifying that the underlying population is not normally distributed.

A2 Defining null and alternative hypotheses from the context.

A4 Using exploratory data analysis in order to determine the population standard deviation.

A6 Appreciating that the samples must be obtained randomly for any conclusions made to be valid.

B1 Appreciating that taking large random samples may not be practical

B3 Identifying whether a sample has been obtained randomly.

C1 Calculating numerical measures from a sample.

C2 Calculating numerical measures from summary statistics generated by technology.

D1 Analysing numerical measures in a hypothesis test.

D2 Reaching a conclusion from the hypothesis test.

D3 Identifying and conducting an appropriate hypothesis test to determine if a result is significant.

D4 Using the significance level and the sampling methods to justify the reliability of findings.

D5 Reaching conclusions in context in a language appropriate for a given target audience.

E1 Identifying if the sample was not obtained randomly and commenting on the validity of a conclusion.

E2 Identifying that the Central Limit Theorem cannot be used if the sample size is not large, so the hypothesis test would require an assumption of the normal distribution.

E3 Identifying that a large enough sample needs to be obtained in order to apply the Central Limit Theorem.

COMMON AND POSSIBLE MISTAKES

See Unit 11a or Unit 18b.

Students must remember that:

- When calculating probabilities for a *p*-value method in a hypothesis test, always use $X \leq$ or $X \geq$;
- reject H_0 if the *p*-value is less than significance level;
- the test statistic and the critical value have the same sign;
- compare test statistic with critical value;
- compare *p*-value with significance level.

19b. Concepts in Hypothesis Testing: Hypothesis tests for a sample mean of a normal distribution with unknown variance (16.1)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Carry out a hypothesis test for the mean of a normal distribution with unknown variance (a *t*-test)
- Interpret the results of a *t*-test in context
- Appreciate that, if the sample is large enough, the population variance may be estimated by the sample variance and a *z*-test can be used.

TEACHING POINTS

Revise the previous sub-unit and the concept of the *t*-distribution. Also revise confidence intervals using the *t*-distribution. The hypothesis test remains largely the same as before, except students must use the table of percentage points of a *t*-distribution.

Once students have practised finding critical values using the *t*-distribution, return to hypothesis testing. Ensure that the variance is not known and the sample sizes are small – this way you can emphasis that the *t*-distribution is sufficiently different from the normal distribution.

There are multiple ways to carry out a *t*-test.

- Using standardised critical regions: the test statistic is $\frac{\overline{x}-\mu}{\frac{s_x}{\sqrt{n}}}$ and the critical *t* values with $n-1$ degrees of freedom may be obtained from the tables or a calculator.
- Using non-standardised critical regions: the test statistic is $\overline{x}$ and the critical region is $\left(\mu \pm t \times \frac{s_x}{\sqrt{n}}\right)$ where t is the critical value with $n-1$ degrees of freedom which may be obtained from the tables or a calculator.
- Using *p*-values: some calculators can carry out a *t*-test and produce a *p*-value. These *p*-values should be compared with the significance level.

All three methods can gain full marks (in line with the mark scheme).

Exemplar

A trading standards inspector visits a butcher who sells meat pies.
The inspector investigates the meat content of 12 pies and records these figures, in grams:

234 256 240 234 251 251 243 216 251 232 260 240

The pies are supposed to contain 250 g of meat, but there have been complaints that the butcher does not put in enough meat (that is why the inspector went into her shop).

a) Test, at the 5% significance level, whether the complaints are justified.

Let X be the meat content, in grams, of a pie. We will assume $X \sim \mathrm{N}(\mu, \sigma^2)$.

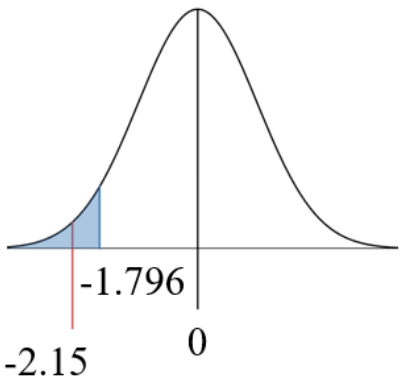
We will carry out a one-tailed t-test at the 5% significance level.
Since the variance is unknown, we will use the sample variance as an estimate and use the t-distribution with 11 degrees of freedom.

$H_0: \mu = 250$ *g,*
$H_1: \mu < 250$ *g,*
where μ is the population mean meat content of a pie.

The sample mean is 242.33 and the standard deviation is $s_x = 12.339$.
The sample size is 12, and the critical t-value from the table ($\upsilon = 11$) is -1.796.

Method 1: Using standardised critical regions

The test statistic is $t = \frac{\overline{x} - \mu}{\frac{s_x}{\sqrt{n}}} = \frac{242.33 - 250}{12.339/\sqrt{12}} = -2.15$

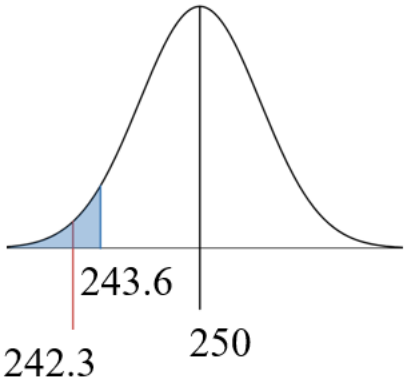


Since $-2.15 < -1.796$ is inside the critical region, the result is significant. We reject H_0
So there is significant evidence to suggest that the population mean meat content of the pies is lower than 250 g

Method 2: Using non-standardised critical regions

The test statistic is 242.3.

The critical value is $\mu + t \times \frac{s_x}{\sqrt{n}} = 250 - 1.796 \times \frac{12.339}{\sqrt{12}} = 243.6$*, so the critical region is* $X \leq 243.6$



Since 242.3 *is inside the critical region, the result is significant. We reject* H_0*. So there is significant evidence to suggest that the population mean meat content of the pies is lower than* 250 *g.*

Method 3: Using p-values from the calculator

The p-value is $0.0272 < 0.05$*, so the result is significant. We reject* H_0*. So there is significant evidence to suggest that the population mean meat content of the pies is lower than* 250 *g.*

b) State an assumption which you have had to make concerning the population from which the sample of 12 pies was drawn.

The underlying population of meat content in pies is normally distributed.

c) State an assumption which you have had to make concerning the sample of 12 pies.

The sample of pies is obtained randomly.

You could finish with examples where the sample size is large.
Since the percentage points of the *t*-distribution table lists degrees of freedom as high as 100, emphasise to students that if the sample size is large enough, they may use the normal distribution. "Large enough" in this context can mean "$n \geq 30$". As with confidence intervals, the *t*-distribution may still be used with large samples as long as the underlying population is normally distributed with an unknown variance.

OPPORTUNITIES FOR EMBEDDING THE SEC

A1 Identifying that the variance is not known and identifying the sample size to determine the appropriate test.

A2 Defining the null and alternative hypotheses from the context

A6 Appreciating that the sample must be obtained randomly for any conclusions made to be valid

B1 Identifying the practical constraints when obtaining large random samples

B3 Identifying whether a sample has been obtained randomly

C1 Calculating numerical measures from a sample

C2 Calculating numerical measures from summary statistics generated by technology

D1 Analysing numerical measures for use in a hypothesis test

D2 Reaching conclusions in relation to the hypothesis test

D3 Identifying and conducting an appropriate hypothesis test to determine if a result is statistically significant.

D4 Using the significance level and data collection methods to determine the reliability of the conclusions made

D5 Reaching conclusions in context in a language appropriate for a given target audience

E1 Identifying the consequences if a sample has not been obtained randomly

E3 Using a larger sample size in order to use a normal distribution.

Practising questions where the hypothesis test may be a *z*-test or a *t*-test will emphasise **D3**.

COMMON AND POSSIBLE MISTAKES

- Using the incorrect column for a *t*-distribution.
- Any usual mistake from a hypothesis test.
- Using a *z*-test instead of a *t*-test.
- Failing to state assumptions in the context of the question when asked to do so.

Students must remember:

- that the test statistic and the critical value have the same sign;
- to compare test statistic with critical value;
- to state the degrees of freedom.

19c. Concepts in Hypothesis Testing: Further concepts of Hypothesis testing (15.7)

Teaching time
1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Define a Type I and Type II error.
- Interpret a Type I and Type II error in context.
- State the probability of making a Type I error.
- Know the advantages and disadvantages of having a lower significance level.
- Appreciate when *p*-values should be used instead of critical regions.
- Appreciate when critical regions should be used instead of *p*-values.

TEACHING POINTS

A Type I error of a hypothesis test is rejecting the null hypothesis when it is actually true.

A Type II error of a hypothesis test is not rejecting the null hypothesis when it is actually false.

The following table is useful when teaching students the concept of Type I and Type II errors:

	H_0 is not rejected	**H_0 is rejected**
H_0 is true	Correct conclusion is made	Type I error
H_0 is not true	Type II error	Correct conclusion is made

One teaching method is by referring back to the courtroom analogy introduced in Unit 10a. A Type I error is the equivalent of convicting an innocent person. A Type II error is the equivalent of letting a guilty person walk free. Allow students to decide which scenario they believe to be preferable – the resulting discussion is a good and interesting activity.

Students need to appreciate that the probability that a Type I error is made is the size of the critical region (as a probability). For continuous distributions, this is the same as the significance level. For discrete distributions, the actual size of the critical region must be calculated. The notation for this is α (alpha). Utilise the Type I and Type II errors activity on Desmos (only the normal distribution is used on this activity, but it may be sufficient as a visual aid). Using the activity will help students appreciate that if the true mean is very different from the mean in the null hypothesis, then you are more likely to reject the null hypothesis. If the true mean is very similar to the mean in the null hypothesis, then you are less likely to reject the null hypothesis. This then translates directly to the probability of making Type II errors. The notation for this is β (beta).

The significance level can be changed on the Desmos activity (the appropriate critical value of the standard normal distribution is needed) and can illustrate that a smaller significance level results in a lower probability of a Type I error but a higher probability of a Type II error.

Students need to be aware that information about the population mean needs to be known in order to calculate the probability of making a Type II error (this will happen in the following sub-unit).

Students need to be able to interpret the meaning of Type I and Type II errors in the context of the question. This can either be done standalone (using the context and set-up of a hypothesis test) or as a follow-up to a full hypothesis test. Students need to be able to interpret the advantages and disadvantages of using different significance levels in context.

Exemplar

Macular degeneration is a condition which causes the loss of sight in one or both eyes. An investigation into the effects of smoking and the age of onset of macular degeneration is carried out. A researcher believes that the mean age of onset of macular degeneration of a smoker is lower than 60 (the mean age of onset for a non-smoker). She carries out a hypothesis test at the 5% significance level and finds the result to be significant.

a) **Explain what a Type I error is in this case (context is vital).**
A Type I error is concluding that the mean age of onset of macular degeneration for a smoker is lower than 60, when in reality it is 60.

b) **Explain what a Type II error is in this case (context is vital).**
A Type II error is concluding that the mean age of onset of macular degeneration for a smoker is 60, which in reality it is lower than 60.

c) **State the probability of making a Type I error in this case.**
0.05.

Exemplar

A particular drug originally used to treat cancer is used in halting the deterioration of macular degeneration. An investigation into the effectiveness of this drug is carried out. The existing treatment for macular degeneration is known to be 60% effective and it is suspected that the new drug will be more effective. A random sample of 20 patients is taken, the results recorded and a hypothesis test carried out at the 5% significance level.

a) **Find the critical region for this test.**
Using trial and error:
$\mathrm{P}(X \geq 16) = 1 - \mathrm{P}(X \leq 15) = 1 - 0.949 = 0.051 > 0.05$
$\mathrm{P}(X \geq 17) = 1 - \mathrm{P}(X \leq 16) = 1 - 0.984 = 0.016 < 0.05$
So the critical region is $X \geq 17$

b) **Find the probability of making a Type I error in this case.**
The probability of a Type I error is the size of the critical region (as a probability).
So $\mathrm{P}(X \geq 17) = \mathrm{P}(Type\ I\ Error) = 0.0160$

To finish, revise the critical region method and the *p*-value method (using a z-test). Students may appreciate that for discrete distributions (for example, the binomial distribution), *p*-values are easier to determine whether a result is statistically significant. Students may also appreciate that once the *p*-value is calculated, the level of significance needed to reject or not reject the null hypothesis for this sample will be known.
However, highly discourage students from calculating the *p*-value and then deciding on a level of significance – this is bad practice for a hypothesis test and students need to understand how this can lead to misrepresentation.

Students are also required to know the advantages and disadvantages of using critical regions and *p*-values. The following table gives (a non-exhaustive) list of possible advantages and disadvantages.

	Advantages	Disadvantages
Critical Regions	• Can easily replicate hypothesis tests without the need for further calculation • Needed if assessing the quality of the hypothesis test (i.e. calculating Type I and Type II errors) • Can give information about whether the sample size is large enough before data collection occurs (e.g. in a binomial proportion test where there is no critical region at a sensible significance level)	• Need to be a statistics specialist (or know enough about the statistical analysis) in order to fully understand where the critical regions come from. • Not widely used in employment / real-world statistical analyses
***p*-values**	• A standardised measure of "significance" across all hypothesis tests (i.e. the definition is the same regardless of the analysis) • Better used when communicating findings to a non-statistical specialist audience (but not the general public) • Most statistical software / employment utilise *p*-values	• Historical abuse and mis-interpretation (e.g. *p*-hacking) • Not mathematically well-defined for some tests (e.g. asymmetric/discrete/non-parametric). By convention, the definition for symmetrical continuous distributions is adopted as a standard. • Even in modern practices, the wider statistical community cannot agree on the best way to use *p*-values.

OPPORTUNITIES FOR EMBEDDING THE SEC

C3 Understanding that choosing a significance level based on the *p*-values calculated from a hypothesis test can lead to misrepresentation.

D1 Relating the significance level to the probability of making a Type I error (see mistakes below).

D5 Interpreting a Type I or Type II error in context, using language appropriate to a given target audience.

COMMON AND POSSIBLE MISTAKES

- Students often mix up the definitions of Type I and Type II errors.
- Giving generic, non-contextual answers when interpreting Type I and Type II errors.
- For a binomial proportion test, reporting the probability of a Type I error as the significance level (this may be true for continuous distributions but not necessarily for discrete distributions).
- Students may fall into the bad practice of calculating a *p*-value (e.g. 0.078) and then deciding to use a 10% significance level instead of a 5% significance level, just to show that a result is statistically significant. Discourage and avoid this practice and point out the misrepresentation that can occur using this method.

19d. Concepts in Hypothesis Testing: The Power of a Hypothesis Test (15.3) (15.6)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Calculate the probability of making a Type II error.
- Understand and calculate the power of a hypothesis test.
- Interpret the power of a hypothesis test in context.

TEACHING POINTS

The power of a hypothesis test is the probability of reaching the correct conclusion if the alternative hypothesis is true. Refer students to the table seen in the previous sub-unit.

Since a Type II error is reaching an incorrect conclusion when the alternative hypothesis is true, it will be easy to explain that the power of a hypothesis test is $1-\beta$, where β is the probability of making a Type II error.

Students need to be aware that when carrying out a hypothesis test, the two main aims are to minimise the probability of making a Type I error and to maximise the power (by referring to the table in the previous unit, the aim is to maximise the probability of making a correct conclusion). The maximum probability of making a Type I error is easy to set (the significance level).

The only way to decrease the probability of a Type I error is by reducing the significance level. However, this has the effect of increasing the probability of making a Type II error. There are two ways of decreasing the probability of making a Type II error:

- Increasing the significance level, which has the effect of increasing the probability of making a Type I error
- Increasing the sample size, which does not affect the probability of making a Type I error but it may not be practical/possible to take larger samples.

Students need to be able to calculate the power of a hypothesis test, so they need to be able to calculate β. Use the Type I and Type II error activity in Desmos as a visual aid. In order to calculate β, the significance level α, the critical value assuming the null hypothesis is true, and the true population mean is required.

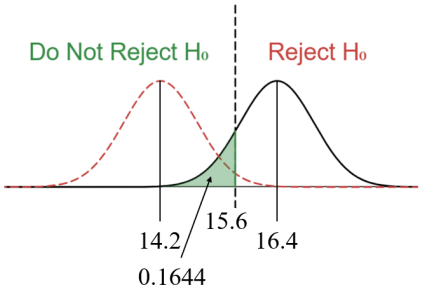
Exemplar

A one-tailed hypothesis test for the mean of a normal distribution is carried out. A sample size of 14 is taken from a normally distributed population with known variance 3.14. Researchers carry out the hypothesis test using the null hypothesis $\mathbf{H_0}: \boldsymbol{\mu} = \mathbf{14.2}$ and $\mathbf{H_1}: \boldsymbol{\mu} > \mathbf{14.2}$ at the 5% significance level.
If $\boldsymbol{\mu} = \mathbf{16.4}$, calculate the probability of making a Type II error.

The hypothesis test uses $\overline{X} \sim \mathrm{N}\left(14.2, \frac{3.14^2}{14}\right)$.

Method 1: Using $\overline{X}$

The critical value is 15.58 (or $14.2 + 1.645 \times \frac{3.14}{\sqrt{14}}$*).*
H_0 is not rejected if the test statistic is less than 15.58.

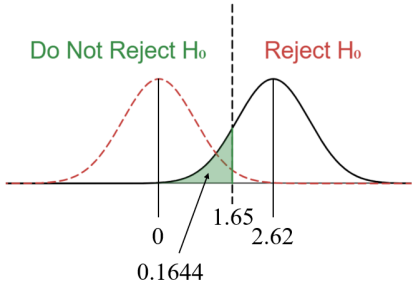


If the true mean is 16.4, then using $Y \sim \mathrm{N}\left(16.4, \frac{3.14^2}{14}\right)$ *we have*
β = P*(Type II error)* = P*(not rejecting* H_0 | H_0 *false)* = $\mathrm{P}(Y \leq 15.58 \mid \mu = 16.4) = 0.164$

Method 2: Using Z

Using $Z \sim N(0,1)$*, the critical value is 1.645. H_0 is not rejected if the test statistic is less than 1.645.*

Under H_0, the z-value for the true mean is $\frac{16.4-14.2}{\frac{3.14}{\sqrt{14}}} = 2.6215\ldots$



Using the z-value for the true mean, and using $Y \sim N(2.6215, 1)$ *we have*

β = P*(Type II error)* = P*(not rejecting* H_0 | H_0 *false)* = $\mathrm{P}(Y \leq 1.645 \mid \mu = 2.6215) = 0.164$

After enough non-contextual practice, introduce context to students.

Exemplar

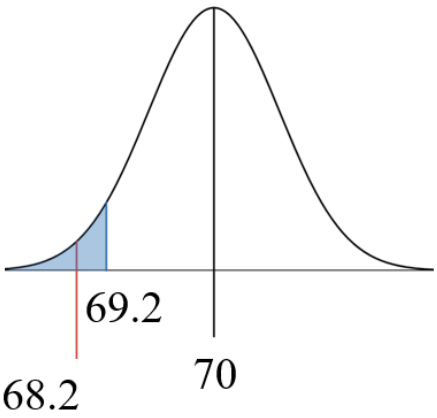
An inspector wished to determine whether eggs sold in the local farmers market as Size 1 have mean mass of at least 70.0 g, as she suspected that they may be underweight. She weighs a sample of 200 eggs and found the mean mass was 68.2 g with standard deviation 7.25 g.

a) **Test whether there is significant evidence, at the 5% level, that the eggs weigh less than 70.0 g on average.**
Let X be the mass of an egg. Since n is large, we may assume that $\sigma = 7.25$ and we may also assume that $\bar{X} \sim \mathrm{N}\left(\mu, \frac{7.25^2}{200}\right)$ by the Central Limit Theorem.

$H_0: \mu = 70$ g,
$H_1: \mu < 70$ g,
where μ is the population mean mass of an egg.

A one-tailed z-test at the 5% significance level using $\mu = 70$ is required.

<u>Method 1: Using non-standardised critical regions</u>
Using $\overline{X} \sim N\left(70, \frac{7.25^2}{200}\right)$
The test statistic is 68.2

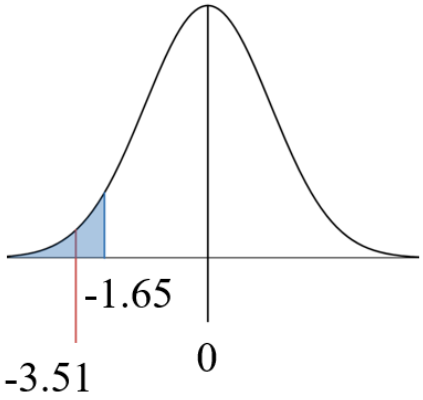


The calculator gives the critical region as 69.2 or less.
Since 68.2 < 69.2, the result is significant. We reject H_0.
There is significant evidence to suggest that the population mean mass of an egg is less than 70 g.

Method 2: Using standardised critical regions

Using $Z \sim N(0,1)$

The test statistic is $\frac{68.2-70}{7.25/\sqrt{200}} = -3.51$



The critical region is -1.65 *or less.*

Since $-3.51 < -1.65$*, the result is significant. We reject* H_0*.*

There is significant evidence to suggest that the population mean mass of an egg is less than 70 g.

Method 3: Using p-values

Using $\overline{X} \sim N\left(70, \frac{7.25^2}{200}\right)$ *(or* $Z \sim N(0,1)$*)*

$\mathrm{P}(\overline{X} \leq 68.2)$ $(or\ \mathrm{P}(Z \leq -3.51)) = 0.000223.$

The one-tailed p-value is 0.000223 < 0.05 so the result is significant. We reject H_0*.*

There is significant evidence to suggest that the population mean mass of an egg is less than 70 g.

b) Is your conclusion affected by whether or not the sample was taken at random?

If so, how?

The conclusion would be invalid if the sample was not obtained randomly.

c) Is your conclusion affected by whether or not the distribution of weights was normal?

Give a brief justification of your answer.

The underlying population need not be normally distributed, since n *is large and so we may assume that* $\bar{X}$ *is normally distributed, by the Central Limit Theorem.*

d) Explain, in the context of this question, the meaning of a Type I error.

A Type I error in this case would be when we concluded that the mean mass of an egg is smaller than 70 g, when in reality it is not smaller than 70g.

e) Explain, in the context of this question, the meaning of a Type II error.

A Type II error in this case would be when we concluded that the mean mass of an egg is 70 g, when in reality it is smaller than 70g.

f) Write down the probability of making a Type I error for this hypothesis test.

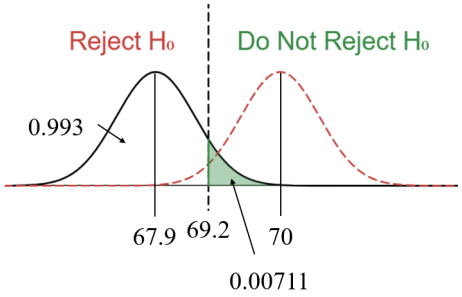
0.01

g) If, unknown to the inspector, the true mean mass was 67.9 g, calculate the power of this hypothesis test.

Method 1: Using $\overline{X} \sim \mathrm{N}\left(70, \frac{7.25^2}{200}\right)$

The critical value is 69.2 (or $70 - 1.645 \times \frac{7.25}{\sqrt{200}}$*).*

H_0 is not rejected if the test statistic is more than 69.2.



If the true mean is 67.9, then using $Y \sim \mathrm{N}\left(67.9, \frac{7.25^2}{200}\right)$ *we have*

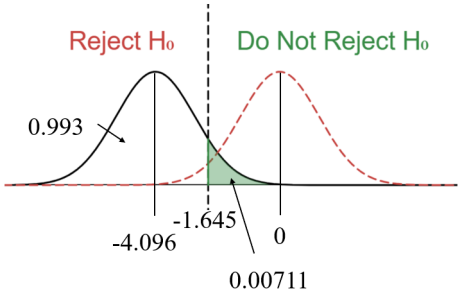
$\beta =$ P*(Type II error)* = P*(not rejecting H_0 | H_0 false)* = $\mathrm{P}(Y \geq 69.2 \,|\mu = 67.9) = 0.00711$

So the power is $1 - 0.00711 = 0.993$

Method 2: Using $Z \sim N(0,1)$

The critical value is –1.645. H_0 is not rejected if the test statistic is more than 1.645.

Under H_0*, the z-value for the true mean is* $\frac{67.9-70}{\frac{7.25}{\sqrt{200}}} = -4.096$



Using the z-value for the true mean, and using $Y \sim N(-4.096,1)$ *we have*

$\beta =$ *P(Type II error) = P(not rejecting H_0 | H_0 false) =*

$P(Y \geq -1.645 | \mu = -4.096) = 0.00771$

So the power is $1 - 0.00771 = 0.993$

Note that students may calculate power directly by using $\mathrm{P}(\text{rejecting } \mathrm{H}_0 | \mathrm{H}_0 \text{ false})$. This is valid and could gain full marks (in line with the mark scheme).

In order to interpret the power in context, show examples when the means are very similar and very different. Again, you may use the Type I and Type II error activity on Desmos to help. If the means are very similar, then the power of the hypothesis test will be quite low. If, in the above example, the inspector rejected the null hypothesis and the mean turned out to be 70.5, although the result is significant it wouldn't be "meaningful".

It is possible that a power calculation could be performed on a hypothesis test for the proportion:

Exemplar

A health centre claims that 95% of adults who have a flu jab in October do not catch flu during the next 6 months.
In order to test this, a random sample of 120 adults who had flu jabs in October was observed.
10 of the sample contracted flu during the next 6 months.
Test, at the 5% level, the claim made by this health centre and calculate the power of this hypothesis test if the true proportion of adults who do not catch flu was 98%.

Let X be the number of adults who contract flu after a flu jab
$H_0: \pi = 0.05$,
$H_1: \pi \neq 0.05$,
where π *is the proportion of the population of adults who contracted flu after the flu jab.*

A two-tailed test about the proportion at the 5% level using $X \sim \mathrm{B}(120, 0.05)$ *is required.*

Since $\mathrm{P}(X \leq 1) = 0.0155$ *which is below 0.05 and* $\mathrm{P}(X \leq 2) = 0.0575$ *which is above 0.05, the* ***lower critical region is*** $\boldsymbol{X \leq 1}$.

Since $\mathrm{P}(X \geq 10) = 1 - \mathrm{P}(X \leq 9) = 1 - 0.9214 = 0.0786$ *which is above 0.05 and* $\mathrm{P}(X \geq 11) = 1 - \mathrm{P}(X \leq 10) = 1 - 0.9616 = 0.0384$ *which is below 0.05, the* ***upper critical region is*** $\boldsymbol{X \geq 11}$.

The test statistic is 10 which is ***not*** *in the critical region, so the result is not significant. Do not reject* H_0.
There is insufficient evidence to suggest that the proportion of adults who contract the flu after the flu jab is any different to 95%.

If the true proportion were 98%, *then using* $X \sim \mathrm{B}(120, 0.02)$,
P(not rejecting H_0 *|* π *= 0.02)* = $\mathrm{P}(1 < X < 11 | \pi = 0.02) = \mathrm{P}(X \leq 10) - \mathrm{P}(X \leq 1) = 1 - 0.3054 = 0.6946$.
So if $\pi = 0.02$, *the power of this hypothesis test is 0.3054.*

In appropriate situations, the normal approximation to the binomial may also be assessed in conjunction with error / power calculations. These will, however, only be approximations to P(Type I) and P(Type II)/Power. To calculate them exactly, students must use the exact binomial.

Type II error probability and power questions can only be assessed on the following hypothesis tests:

- One-sample *z*-tests
- Binomial Proportion Tests
- Binomial Proportion tests with normal approximations
- Two-Sample *z*-tests (seen later)

This is due to the constraints of technology.

OPPORTUNITIES FOR EMBEDDING THE SEC

Calculating the power of a hypothesis test can be used as a follow-up question to any of the hypothesis tests mentioned above. The ways to embed the SEC can be found in these units, but in addition:

C1 Calculating the power of a hypothesis test or the probability of making a Type II error.

C3 Appreciating that if the null hypothesis is rejected in a low-power hypothesis test, a result may not be “scientifically meaningful” even if it is “statistically significant”.

D4 Using the power of a hypothesis test in relation to the reliability of the conclusion.

D5 Interpreting the power of a hypothesis test in context using language appropriate to a given target audience.

E2 Appreciating that a larger sample size can increase the power of a hypothesis test.

COMMON AND POSSIBLE MISTAKES

- Mixing up P(Type II) and Power;
- not altering the population parameter to the “true value”;
- leaving the question blank.

NOTES

This topic is also on the 9FM0 A Level Further Mathematics specification “Further Statistics 1”. However, there is more focus on the algebra and power functions than there is here. In 9ST0, students are only expected to calculate a probability of a Type II error or power for a particular value of the parameter.

UNIT 20: The Exponential Distribution

Return to overview

SPECIFICATION REFERENCES

18.2 Determine when an exponential distribution is appropriate (and its relationship to the Poisson distribution as a model of the times between randomly occurring Poisson events).

18.3 Evaluate probabilities for Poisson and exponential distributions and know the corresponding mean and variance.

PRIOR KNOWLEDGE

Year 2 of A Level Statistics
The Poisson Distribution (Unit 15)

KEYWORDS

conditional, continuous, discrete, distribution, e, expectation, exponential, interval, mean, parameter, Poisson, population, probability, sample, standard deviation, variance,

UNIT SUMMARY

This unit extends the work in Unit 15 on the Poisson distribution, introducing the exponential distribution. Just like Units 5, 7, and 15, this unit focusses primarily on finding probabilities and interpreting them in context.

The exponential distribution and Poisson distribution activities on Desmos can assist both teachers and students in visualising the probability density function.

20a. The Exponential Distribution: Conditions for the Exponential Distribution (18.2)

Teaching time
1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand the link between the Poisson distribution and the exponential distribution.
- Know the conditions when an exponential distribution may be appropriate.
- Know the mean and variance of the exponential distribution.

TEACHING POINTS

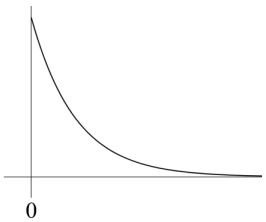
Introduce the exponential distribution as a new continuous distribution. Use the exponential distribution activity on Desmos to illustrate the shape. Also explain that the exponential distribution depends on a parameter, λ, which changes the shape of the exponential distribution. If a random variable X has an exponential distribution, then we write $X \sim \text{Exp}(\lambda)$.

Students need to be aware that if:

- events occur independently and randomly,
- events occur in a specified unit of space or time,
- events occur at a constant average rate,

then the space or time interval between events can be modelled by an exponential distribution. Students may immediately recognise this as the conditions for a Poisson distribution, except it is the space/time interval between events that is being modelled, as opposed to the number of events. The acronym ERIC may be revisited at this stage. Emphasise to students that a negative time or space between events doesn't make sense, so any random variable X following an exponential distribution must take positive values.

Show students the Poisson distribution and the exponential distribution together, to show how they change as λ changes. This can allow students to make the link that if more events occur in a fixed unit of space or time, the space or time interval will naturally decrease. The shape of the exponential distribution is below, demonstrating a lower limit of 0 and no upper limit.



The total area underneath the curve is 1.

To allow students to appreciate how the exponential distribution can be modelled, simple scenarios designed to allow students to apply the conditions can be given.

Exemplar

State, giving reasons, whether the exponential distribution is likely to provide an adequate model for the following situations:

a) **At airport security, the home office is conducting a survey of the ethnicity of British people passing through border control.**
The home office decide to record the times between British Asian people passing through a particular border control desk during 1 hour intervals.
The exponential distribution may be appropriate here if British Asians pass through border control randomly and independently in a fixed period of time.
The average rate at which these people pass through border control is likely to be constant.

b) **Following a cup semi-final victory, a football club ticket office receives a large number of telephone enquiries about tickets for the final.**
This resulted in the switchboard frequently being engaged.
The time between calls received during the first morning after the victory is recorded.
The exponential distribution may not be appropriate here.
When the switchboard is engaged, the calls are not received independently of other calls, because
some are held up and depend on the completion of the previous call.

Students need to appreciate that the mean of the exponential distribution is $\frac{1}{\lambda}$.
This is easy to explain: If an event occurs at a constant average rate of 3 times in a unit of space or time, then the average interval between successive occurrences is $\frac{1}{3}$ of that unit of space or time.
The variance of the exponential distribution is $\frac{1}{\lambda^2}$. Explaining the variance is harder and outside the realms of the A level in Statistics.

Students can also determine the suitability of the exponential distribution by comparing the mean and variance with observed results, or in situations where the mean and standard deviation of continuous data are the same.

The following table may help visualise the equivalence of numerical measures for the Poisson and the exponential distribution.

Distribution	Mean	Variance	Standard Deviation
Poisson	λ	λ	$\sqrt{\lambda}$
Exponential	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{1}{\lambda}$

Exemplar

A charity wants to build a well in a remote village in Africa to provide clean drinking water.
It is claimed that the number of cases of cholera diagnosed per week at this village can be modelled by a Poisson distribution with parameter 2.
Over many weeks, the charity records the time between diagnoses in a week-long period and discovered that the mean length of time between diagnoses is half a week and the variance is also half a week.
Comment on the suitability of a Poisson distribution model.

The Poisson distribution may not be suitable.
If a Poisson distribution with $\lambda = 2$ were suitable, then the time between diagnoses would follow an exponential distribution $\mathrm{Exp}(2)$.
The mean and variance of this distribution would be $\frac{1}{2}$ and $\frac{1}{4}$.
Although the mean time interval is half a week, the variance should be a quarter of a week, so the exponential distribution would not be suitable to model the time between intervals.
The Poisson distribution would therefore not be suitable to model the number of diagnoses per week.

OPPORTUNITIES FOR EMBEDDING THE SEC

A1 Identifying the conditions for when a Poisson or exponential distribution is appropriate.

A3 Understanding the difference between recording the number of events or the time between events.

A6 Justifying the use of a Poisson or Exponential distribution.

D2 Comparing the theoretical mean or variance of an exponential distribution with the observed mean or variance to determine suitability.

E3 Suggesting assumptions or improvements to the model that may be made in order to apply a Poisson or exponential distribution.

COMMON AND POSSIBLE MISTAKES

See Unit 15a. Also:

- Confusing the values of the variance $\frac{1}{\lambda^2}$ and the standard deviation $\frac{1}{\lambda}$.
- Mixing up the Poisson distribution with the Exponential distribution: emphasise that “the number of” is Poisson (discrete) and “time/space between” is exponential (continuous).
- Questions sometimes give the “mean time/space” as opposed to the “mean rate” and students often use $\lambda = \overline{x}$ as opposed to $\lambda = \frac{1}{\overline{x}}$

20b. The Exponential Distribution: Finding Probabilities from an Exponential Distribution (18.3)

Teaching time
1 hour

OBJECTIVES

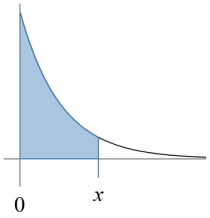
By the end of the sub-unit, students should be able to:

- Find probabilities from an exponential distribution.
- Interpret these probabilities in context.
- Appreciate that the exponential distribution has no memory.

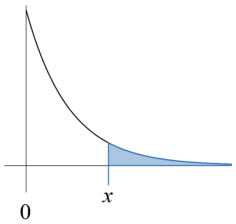
TEACHING POINTS

Revise the Poisson distribution, reminding students of the number $e \approx 2.71818$. The calculators can calculate Poisson probabilities so the use of the e button on the calculator may not have been seen (unless calculating Poisson probabilities using the formula has been practised). If they haven't done so already, ensure students are aware of where the e button is on the calculator and how to use it (i.e. using powers of e).

If $X \sim \text{Exp}(\lambda)$ then the cumulative probability $P(X \leq x) = 1 - e^{-\lambda x}$ (this is given in the formula book). This can be illustrated as a sketch:



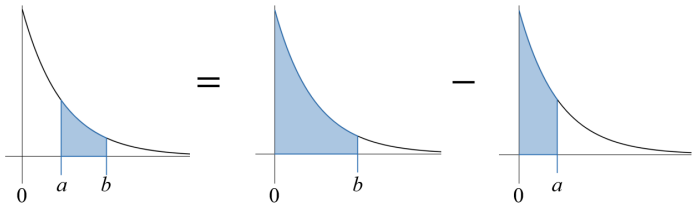
"Right-hand" probabilities can be calculated as $\text{P}(X \geq x) = 1 - \text{P}(X \leq x) = 1 - \left(1 - e^{-\lambda x}\right) = e^{-\lambda x}$. Note that this is not in the formula book and must be remembered. Using a sketch is helpful.



"Middle" probabilities can be calculated as

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = \left(1 - e^{-\lambda b}\right) - \left(1 - e^{-\lambda a}\right) = e^{-\lambda a} - e^{-\lambda b}$$

This is also not in the formula book. Sketches are useful to illustrate the formula.



Start with non-contextualised examples first to allow students time to practise using their calculator to calculate probabilities. The storage function on the calculator will help minimise errors with data entry. This distribution is the only one where students are required to calculate probabilities by hand rather than inputting them into a specific calculator mode.

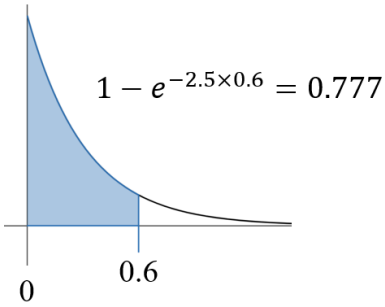
Introduce contextual scenarios later:

Exemplar

A radioactive element emits particles independently and at random.
The mean rate is 2.5 particles per second.

a) Calculate the probability that the interval between successive emissions is at most 0.6 seconds.

Let X be the time between successive emissions of particles. Then $X \sim \text{Exp}(2.5)$.



$\text{P}(X \leq 0.6) = 0.777$

Students need to be aware that the exponential distribution "has no memory".
This means that the time or space until the next event occurs does not depend on the time or space already elapsed since the last event.

This means that for $t_2 > t_1$, $\text{P}(X \leq t_2 | X \geq t_1) = \text{P}(X \leq t_2 - t_1)$ and
$\text{P}(X \geq t_2 | X \geq t_1) = \text{P}(X \geq t_2 - t_1)$.

The proof of this is not necessary and could only be shown to the mathematically able but is included as an **extension**:

$$\mathrm{P}(X \le t_2 | X \ge t_1) = \frac{\mathrm{P}(X \le t_2 \cap X \ge t_1)}{\mathrm{P}(X \ge t_1)} = \frac{\mathrm{P}(t_1 \le X \le t_2)}{\mathrm{P}(X \ge t_1)} = \frac{\mathrm{P}(t_1 \le X \le t_2)}{1 - \mathrm{P}(X \le t_1)}$$

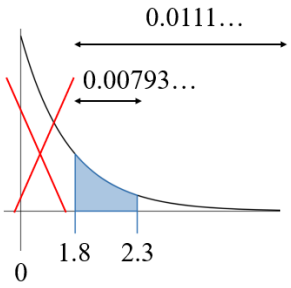
$$= \frac{e^{-\lambda t_1} - e^{-\lambda t_2}}{1 - (1 - e^{-\lambda t_1})} = \frac{e^{-\lambda t_1} - e^{-\lambda t_2}}{e^{-\lambda t_1}} = 1 - e^{-\lambda t_2 + \lambda t_1} = 1 - e^{-\lambda(t_2 - t_1)} = \mathrm{P}(X \le t_2 - t_1).$$

The proof of the second result is similar.
Rather than show the proof of this, it would be more beneficial for students to see examples of "no memory" at work.

As a follow up to the previous example:

Exemplar

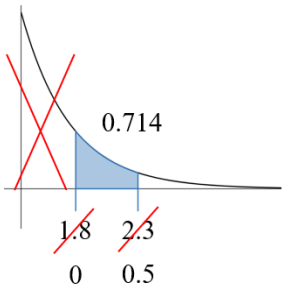
b) After the Geiger counter has been on for 1.8 seconds, the technician noticed there had been no emissions so far.
What is the probability that the first emission occurs 2.3 seconds after the counter has been turned on?
Using conditional probability:



$$\mathrm{P}(X \le 2.3 | X \ge 1.8) = \frac{\mathrm{P}(X \le 2.3 \cap X \ge 1.8)}{\mathrm{P}(X \ge 1.8)} = \frac{\mathrm{P}(1.8 \le X \le 2.3)}{1 - \mathrm{P}(X \le 1.8)}$$

$$= \frac{e^{-2.5 \times 1.8} - e^{-2.5 \times 2.3}}{e^{-2.5 \times 1.8}} = \frac{0.00793\ldots}{0.0111\ldots} = 0.7135$$

Or using "no memory":



$$\mathrm{P}(X \le 2.3 | X \ge 1.8) = \mathrm{P}(X \le 2.3 - 1.8) = \mathrm{P}(X \le 0.5) = 1 - e^{-2.5 \times 0.5} = 0.7135$$

Students may need to be reminded that if $X \leq 2.3$ and $X \geq 1.8$, then $1.8 \leq X \leq 2.3$.
Students may also need to be reminded that if $X \geq 2.3$ and $X \geq 1.8$, then $X \geq 2.3$.

Note that the memorylessness property of the exponential distribution can only be utilised if the condition is $X \geq a$. If a condition is given as $X \leq a$ (e.g. $P(X \geq 1.8|X \leq 2.3)$) then students will have to calculate this only by using conditional probability methods (such as in Unit 17c).

OPPORTUNITIES FOR EMBEDDING THE SEC

C1 Using probability theory to calculate probabilities ready for input into technology.

C2 Using technology to calculate probabilities.

D1 Interpreting probabilities.

D5 Interpreting probabilities in context, using language appropriate for a given target audience.

COMMON AND POSSIBLE MISTAKES

- Due to the higher amount of algebraic manipulation, there may well be algebraic mistake - practice helps.
- Students often read $\mathrm{P}(X \geq a)$ as $\mathrm{P}(X \leq a)$ and hence use "$1 -$" when they shouldn't. To minimise this, you could teach $\mathrm{P}(a \leq X \leq b) = e^{-\lambda a} - e^{-\lambda b}$ and if there isn't a "b" then the second term is zero.

It is highly advisable for students to utilise sketches of the distribution to minimise these mistakes.

NOTES

The probability density function of the exponential distribution is $\boldsymbol{\lambda e^{-\lambda x}}$. This is not given in the current formula book, nor is it on the specification. It should only be demonstrated to the mathematically able (e.g. students who also take A level Mathematics) who can easily distinguish between the equation of the probability density function and the area beneath it. Since integration is not on the course, these students will likely be in the minority.

UNIT 21: Hypothesis Tests for the Difference between Two Parameters

Return to overview

SPECIFICATION REFERENCES

16.2 Know how to apply knowledge about carrying out hypothesis testing to conduct tests for the difference of two means for two independent normal distributions with known variances

16.3 Know how to apply knowledge about carrying out hypothesis testing to conduct tests for the difference of two means for two independent normal distributions with unknown but equal variances

16.4 Know how to apply knowledge about carrying out hypothesis testing to conduct tests for the difference between two binomial proportions

16.5 Interpret results for 16.2-16.4 in context.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics

Binomial Distribution (Unit 5)

Normal Distribution (Unit 7)

Estimation and Approximation (Unit 9)

Hypothesis Testing: *z*-tests, tests for proportion (Units 10a and 11)

Year 2 of A Level Statistics

Linear Combinations of Random Variables (Unit 16)

Hypothesis Testing: *t*-tests (Unit 19b)

KEYWORDS

alternative hypothesis, binomial, Central Limit Theorem, continuous, critical region, critical value, difference, discrete, distribution, hypothesis, independent, insufficient, mean, normal, null hypothesis, pooled estimate, population, power, probability, proportion, risk, sample, sampling distribution, significance level, significant, standard deviation, sufficient, t-distribution, test statistic, Type I error, Type II error, variance,

UNIT SUMMARY

This unit returns to the idea of hypothesis testing, this time testing for the difference in parameters between two distributions. Two of the sub-units refer to the difference between two means and the final sub-unit refers to the difference between two proportions (where the normal approximation is used anyway). Utilise the *z*-test with critical regions, *t*-test with critical regions and the binomial distribution with normal approximation activities in Desmos to help.

The test statistics are listed in the formula book, but the tests themselves can be taught without the need to standardise the normal distribution, thanks to the technology on the more advanced graphical calculators. However, it is more important than ever to ensure that students state the distributions they are using. The formulae in the formula book can be used as a checking aid.

Noted that for Unit 21c it is advisable to refer to the notes in Unit 9c and Unit 11b. In order to carry out a hypothesis test for the difference between two proportions, knowledge of the sampling distribution of the proportion needs to be known. Although not on the Year 1 part of the course, it makes more sense to teach it in these respective units. You may move the notes of Unit 9c and 11b into the beginning of Unit 21c if the sampling distribution of the proportions have not yet been seen until this point.

The opportunities for embedding the SEC will be the same in each sub-unit, so is presented here as an overview:

A1 Identifying possible factors which may result in the population means or proportions being different.

A2 Defining an appropriate null and alternative hypothesis from the context of the question.

A3 Describing a suitable sampling method on collecting samples from two populations.

A4 Using exploratory data analysis to estimate the population variance.

A5 Identifying an appropriate hypothesis test to carry out to answer **A2**, and planning how to do it.

A6 Justifying the hypothesis test referring to relevant assumptions.

B1 Appreciating the practical constraints in taking large random samples.

B2 Appreciating the need for secondary data from multiple sources (in reference to **A4**).

B3 Recording **A3** and understanding the need for this to be declared.

B4 Understanding the context in which samples have been obtained and whether bias can be reduced through experimental design.

C1 Calculating numerical measures from a sample, and calculating probabilities including the power of a hypothesis test.

C2 Calculating numerical measures from summary statistics generated by technology.

C3 Understanding the dangers of defining a hypothesis or significance level after analysing the samples.

D1 Interpreting numerical measures in context.

D2 Interpreting the results of a hypothesis test in context.

D3 Carrying out an appropriate hypothesis test, and determining whether the result is significant.

D4 Referring to the significance level of a hypothesis test, the sampling methods used, the probability of a Type II error or the power of a hypothesis test.

D5 Reaching conclusions in context in a language appropriate to a given target audience.

E1 Understanding the disadvantages of using a certain sampling method, or referring to the probability of making a Type I or Type II error, or the power of a hypothesis test.

E2 Understanding that the conclusions of a hypothesis test may not be valid if the sampling technique is not random, or the sample size is insufficiently large.

E3 Suggesting alternative methods of sampling and/or hypothesis tests, together with any assumptions made or relaxed.

E4 Redoing a hypothesis test with a new sample incorporating the suggestions made in **E3**.

21a. Hypothesis tests for the Difference Between Two Parameters: two population means with known variances (16.2)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Write down the distribution defined by the difference of two independent sampling distributions of the means with known variances.
- Carry out a hypothesis test for the difference of two normally distributed population means with known variance.
- Carry out a hypothesis test for the difference of two population means with known variance, given large samples.
- Interpret the results of the hypothesis tests in context.

TEACHING POINTS

Revise the sampling distribution of the mean and linear combinations of independent normal variables (Units 9 and 16). Also revise hypothesis testing for the mean (Unit 11a).

Remind students that if X and Y are normally distributed, we write $X \sim \mathrm{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathrm{N}(\mu_y, \sigma_y^2)$ where the subscripts denote which population the mean and variance are from. Also, remind students that, if samples of size n_x are taken from X and samples of size n_y are taken from Y, then $\bar{X} \sim \mathrm{N}\left(\mu_x, \frac{\sigma_x^2}{n_x}\right)$ and $\bar{Y} \sim \mathrm{N}\left(\mu_y, \frac{\sigma_y^2}{n_y}\right)$. Emphasise that the sample sizes can be different here.

Finally, using Unit 16, remind students of the formulae for the mean and variance of linear combinations (both of which are in the formula book) so $\bar{X} - \bar{Y} \sim \mathrm{N}\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$.

Ensure students get practice at finding the distributions of $\bar{X}$ and $\bar{Y}$ given the distributions of X and Y and sample sizes for each:

Exemplar

It is known that X is normally distributed with a mean of 13.4 and a standard deviation of 5.4.
It is also known that Y is normally distributed with a mean of 18.3 and a variance of 20.1.
Given that samples of size 13 are taken from the population of X and samples of size 17 are taken from the population of Y, find the distribution of $\overline{X} - \overline{Y}$.

$X \sim \mathrm{N}(13.4, 5.4^2)$ *and* $Y \sim \mathrm{N}(18.3, 20.1)$.
So $\bar{X} \sim \mathrm{N}\left(13.4,\ \frac{5.4^2}{13}\right)$ *and* $\bar{Y} \sim \mathrm{N}\left(18.3,\ \frac{20.1}{17}\right)$.
Hence $\bar{X} - \bar{Y} \sim \mathrm{N}\left((13.4 - 18.3), \frac{5.4^2}{13} + \frac{20.1}{17}\right) = \mathrm{N}(-4.9,\ 3.4254)$.

You can then follow this up with probability questions as extra practice.

Students can then be introduced to the idea of a hypothesis test for the difference of two means. Given two samples, each taken from independent normal distributions, we want to determine if the population mean from one distribution is greater than, lower than or different from the population mean from the other distribution.

The choice of null hypotheses can be $H_0: \mu_X = \mu_Y + a$ or $H_0: \mu_X - \mu_y = a$ for a specified constant a. The alternative hypotheses can be $H_1: \mu_X > \mu_Y + a$ (or $H_1: \mu_X - \mu_Y > a$), or the corresponding hypotheses with $<$ or $\neq$.

Students must remember to state the test (difference of two means z-test or two-sample z-test), the tail (one or two tailed), the significance level, and (most importantly) the distribution they are using. There are three methods to be exemplified here, all of which could gain full marks (in line with the mark scheme) in an examination.

- Using non-standardised critical regions.
  The distribution to be used is $\overline{X} - \overline{Y} \sim \mathrm{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)$
  The test statistic is $\overline{x} - \overline{y}$
- Using standardised critical regions.
  The distribution to be used is $Z \sim \mathrm{N}(0,1)$
  The test statistic is $\frac{(\overline{x} - \overline{y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$ (this is given in the formula book)
- Using *p*-values
  Either distribution above can be used with the corresponding test statistic.

Students must remember to stay consistent with using either $\bar{x} - \bar{y}$ or $\bar{y} - \bar{x}$ and the critical value must be consistent with it (see mistakes below). Again, encourage students to draw a sketch.

Begin with examples where the null hypothesis is $\mu_X = \mu_Y$.

Exemplar

A lack of vegetable intake in food consumption is suspected of retarding the growth of muscle mass in athletes.

The following data are the results of an experiment to measure the percentage gain in muscle mass in 17 developing athletes given either a balanced diet (A) or a diet with no vegetables (B).
Athletes are allocated to diets at random.

Diet A	18.2	25.8	16.8	14.9	19.6	26.5	17.5			
Diet B	13.4	18.8	20.5	7.5	22.2	15.0	12.2	14.3	18.0	15.1

Assuming that the percentages for both Diet A and Diet B are normally distributed with standard deviations 4.5 and 4.9 respectively, investigate, at the 5% significance level, the claim that a lack of vegetable intake retards the mean percentage gain of the muscle mass in athletes.

Let X be the percentage increase of muscle mass in an athlete on Diet A.
Then $X \sim \mathrm{N}(\mu_X, 4.5^2)$.
Let Y be the percentage increase of muscle mass in an athlete on Diet B.
Then $Y \sim \mathrm{N}(\mu_Y, 4.9^2)$.
The sample from X has size 7, so $\bar{X} \sim \mathrm{N}\left(\mu_X, \frac{4.5^2}{7}\right)$.
The sample from Y has size 10, so $\bar{Y} \sim \mathrm{N}\left(\mu_Y, \frac{4.9^2}{10}\right)$.

$\mathrm{H}_0: \mu_X - \mu_Y = 0$,
$\mathrm{H}_1: \mu_X - \mu_Y > 0$,
where μ_X and μ_Y are the population mean percentage increases of muscle mass in athletes on diets A and B respectively.

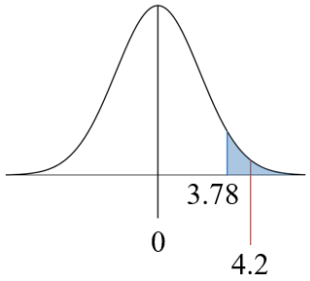
Carry out a one-tailed hypothesis test for the difference of two means at the 5% significance level.

From the samples, $\bar{x} = 19.9$ and $\bar{y} = 15.7$, so $\bar{x} - \bar{y} = 4.2$.

Method 1: Non-Standardised Critical Regions

Using $\bar{X} - \bar{Y} \sim \mathrm{N}\left(\mu_X - \mu_Y, \frac{4.5^2}{7} + \frac{4.9^2}{10}\right) = \mathrm{N}(0\,, 5.29)$*:*

The test statistic is 4.2.



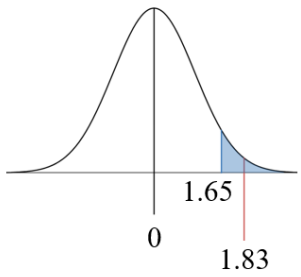
The critical value (from the calculator) is 3.78.

Since $4.2 > 3.78$*, the result is significant. We reject* H_0*.*

Method 2: Standardised Critical Regions

Using $Z \sim \mathrm{N}(0,1)$*:*

The test statistic, z = $\frac{(\bar{x}-\bar{y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}} = \frac{(4.2)-(0)}{\sqrt{\frac{4.5^2}{7}+\frac{4.9^2}{10}}} = 1.83.$



The critical value for z =1.65.

Since $1.83 > 1.65$*, the result is significant. We reject* H_0*.*

Method 3: p-values

Using $\overline{X} - \overline{Y} \sim \mathrm{N}\left(0, \frac{4.5^2}{7} + \frac{4.9^2}{10}\right)$ *(or* $Z \sim \mathrm{N}(0,1)$*)*

we have $P(\overline{X} - \overline{Y} > 4.2)$ *(*$= P(Z > 1.83)$*)* $= 0.034$*.*

Since $0.034 < 0.05$*, the result is significant. We reject* H_0*.*

There is significant evidence to suggest that the population mean percentage increase of muscle mass in an athlete on Diet A is higher than the population mean percentage increase of muscle mass in an athlete on Diet B.
This suggests that there is significant evidence to support the claim that a lack of vegetable intake retards the percentage increase in muscle mass on average.

As usual, it is important that students not only interpret the results in context but relate their conclusions back to the initial question.

Once this test technique has been mastered, whichever method is used, students could start to see situations where they are not testing whether one mean is bigger than another, but rather if one mean is greater than another by a certain amount, e.g. $\mathrm{H}_0: \mu_X - \mu_Y = 30$. These hypothesis tests play out in exactly the same way.

Finally, introduce questions where the two underlying populations are not normally distributed, but the sample sizes taken from each population are large ($n \geq 30$). Students can then invoke the Central Limit Theorem for both populations and continue with the hypothesis test. Ensure students make mention of the use of the Central Limit Theorem when writing out a hypothesis test.

As mentioned in Unit 19d, the probability of a Type II error (or the power) can be assessed for this test. An example of this as a follow up to the previous example is shown:

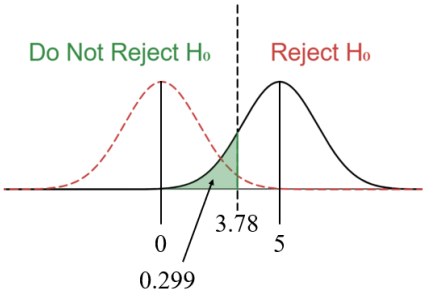
Exemplar

Find the probability of a Type II error if the mean percentage gain of the muscle mass in athletes were actually 5% higher under Diet B than under Diet A.

Method 1: Using $\overline{X} - \overline{Y} \sim \mathrm{N}\left(0, \frac{4.5^2}{7} + \frac{4.9^2}{10}\right)$

The critical value is 3.78 (or $0 + 1.645 \times \sqrt{\frac{4.5^2}{7} + \frac{4.9^2}{10}}$ *).*

H_0 is not rejected if the test statistic is less than 3.78.

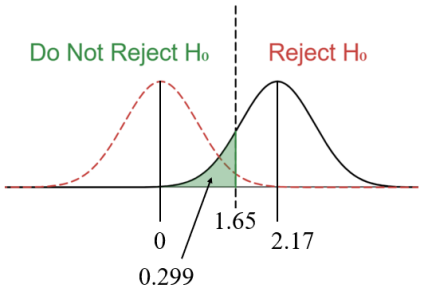


If the true mean is 5, then using $Y \sim \mathrm{N}\left(5, \frac{4.5^2}{7} + \frac{4.9^2}{10}\right)$ *we have*

$\beta =$ P*(Type II error)* = P*(not rejecting H_0 | H_0 false)* = $\mathrm{P}(Y \leq 3.78 \,|\mu = 5) = 0.299$

Method 2: Using $Z \sim \mathrm{N}(0,1)$

The critical value is 1.65. H_0 is not rejected if the test statistic is less than 1.65.
Under H_0*, the z-value for the true mean is* $\frac{5-0}{\sqrt{\frac{4.5^2}{7}+\frac{4.9^2}{10}}} = 2.17$



Using the z-value for the true mean, and using $Y \sim \mathrm{N}(2.17,1)$ *we have*

β = *P(Type II error) = P(not rejecting H_0 | H_0 false) =* $P(Y \leq -1.65 | \mu = 2.17) = 0.299$

OPPORTUNITIES FOR EMBEDDING THE SEC

An overview of how the SEC can be embedded is given in the unit summary. However, an example of **B4** and **C3** is presented as a follow up question to the above example:

Exemplar

After the experiment was performed, it was discovered that the athletes were not a homogeneous population. In fact, most of the athletes in the control group, eating Diet A, had been appreciably heavier than those in the experimental group at the start of the experiment. Discuss briefly the possible effect of the information on the validity of your analysis.

Percentage gain of muscle mass may be affected by the initial mass of an athlete. The observed difference might have been a difference between athletes with different initial weights, rather than a difference between the effects of different diets.

If no significant difference had been observed, it might be that there really was a difference in the population arising from different diets, but this difference was masked by the different between initial weights between two groups. The difference in initial weights is a confounding factor.

Here students need to appreciate that due to poor experimental design, the result is biased and so the reality is therefore misrepresented. Students can further use the SEC with **E3** by suggesting, say, paired samples to reduce experimental error.

COMMON AND POSSIBLE MISTAKES

See common and possible mistakes in Unit 11a, Unit 16c or Unit 18c.

Using the *t* distribution may not be appropriate here since the population variances may not be equal. This can be often identified by wildly different sample variances.

NOTES

Although the test statistic $\frac{(\bar{x}-\bar{y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}}$ is not explicitly stated as such in the formulae book, the distribution $\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}}$ is.

By replacing $\bar{X}$ with $\bar{x}$ and similarly with $\bar{Y}$, the test statistic can be obtained.

21b. Hypothesis Tests for the Difference Between Two Population Parameters: Two population means with unknown, but equal, variances (16.3)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Calculate the pooled estimate of common variance of two samples
- Carry out a hypothesis test for the difference of two normally distributed population means with unknown, but equal, variances
- Interpret the results of the hypothesis tests in context

TEACHING POINTS

Revise *t*-tests and working out critical values using the percentage points of the *t*-distribution.

If the variances for the two populations are unknown, but equal, then $\bar{X} - \bar{Y} \sim \mathrm{N}\left(\mu_X - \mu_Y, \frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y}\right)$. Students may need reminding from Units 18 and 19 that, if the variances are unknown, the *t*-distribution should be used. However, since there are two samples, which sample variance should be used to estimate σ^2?

The answer is both, but we calculate the pooled estimate of common variance (this is required since the difference may only be modelled by a *t*-distribution if the population variances are equal). This is a similar idea to calculating the pooled mean (which is seen in the statistics portion of the A level in Mathematics). It can be explained to students: Recall $s_x^2 = \frac{1}{n_x - 1}\sum(\bar{x} - x)^2$ and $s_y^2 = \frac{1}{n_y - 1}\sum(\bar{y} - y)^2$. So rearranging the formula, the sum of the squares of all sample elements (from samples) is

$$(n_x - 1)s_x^2 + (n_y - 1)s_y^2 = \sum(\bar{x} - x)^2 + \sum(\bar{y} - y)^2$$

This is a bit like combining the two samples into one sample (except using the two sample means, rather than recalculating the combined sample mean – the latter is not used since it is the deviations from the respective population means that are being considered). Remind students that when using the *t*-distribution, the number of degrees of freedom was $n_x - 1$ for the first population, and $n_y - 1$ for the second population. So for the pooled estimate, the number of degrees of freedom is $(n_x - 1) + (n_y - 1) = n_x + n_y - 2$. So the pooled estimate is

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}.$$

This formula is in the formula book. Allow students to calculate the pooled estimate of common variance from a wide selection of non-contextualised examples.

Start with:

Exemplar

Two samples are taken from two independent populations. The first sample has size 6 and has a variance of 14.7. The second sample has size 9 and has a standard deviation of 4.5. Find the pooled estimate of common variance.

We are told that $n_x = 6$, $n_y = 9$, $s_x^2 = 14.7$, $s_y^2 = 4.5^2 = 20.25$.

Hence $s_p^2 = \frac{5\times14.7+8\times20.25}{6+9-2} = 18.115\ldots$

and the pooled estimate of common variance is $s_p^2 = 18.1$ *(3 s.f.)*

For extra practice, allow students to calculate the pooled estimate of common variance from two sets of raw data. This will also give extra practice at using the calculator.

Once students have practised how to calculated the pooled variance, reintroduce the distribution

$$\bar{X} - \bar{Y} \sim \mathrm{N}\left(\mu_X - \mu_Y, \frac{\sigma^2}{n_x} + \frac{\sigma^2}{n_y}\right) = \mathrm{N}\left(\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)\right).$$

Remind students that the pooled variance is the best estimate for σ^2, so we may use

$$\bar{X} - \bar{Y} \sim \mathrm{N}\left(\mu_X - \mu_Y, s_p^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)\right).$$

Ensure students check in the formulae book – the standardised distribution is given as $\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{s_p^2\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}}$.

Since the variance is **unknown**, remind students that the *t*-distribution should be used. The number of degrees of freedom is $n_x + n_y - 2$ (this can also be checked in the formulae book in the denominator of s_p).

The hypothesis test then continues in a similar way to Unit 19b and Unit 21a. The test statistic is the difference between the sample means.

There are two methods exemplified here:

- Using standardised critical regions.
  The test statistic is $\frac{(\bar{x}-\bar{y})-(\mu_X-\mu_Y)}{\sqrt{s_p^2\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}}$ (this is given in the formula book).
  The critical value is given in the tables in the formula book (or can be obtained from the calculator).
- Using non-standardised critical regions.
  The test statistic is $\bar{x} - \bar{y}$
  The critical values can be obtained by using $(\mu_X - \mu_Y) + t\sqrt{s_p^2\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}$ where *t* is

the critical *t*-value given in the formula book or the calculator (this is not given in the formula book).

Some calculators can also calculate *p*-values for a two-sample *t*-test. These may be utilised to access full marks (in line with the mark scheme).

Exemplar

Industrial waste dumped in rivers reduces the amount of dissolved oxygen in the water. A factory was suspected of illegally dumping waste into a river. Specimens of water were taken from the river, six above the factory and eight below the factory, and the dissolved oxygen content (ppm) was as follows:

Above Factory	**4.9**	**5.1**	**4.7**	**5.0**	**5.3**	**4.6**		
Below Factory	**3.8**	**4.9**	**4.0**	**3.6**	**5.0**	**3.4**	**3.5**	**4.0**

Assume that the variances of the populations are equal.

a) **Investigate, at the 5% significance level, whether the mean of the dissolved oxygen content is more than 0.2 ppm higher above the factory than it is below the factory.**

Let X be the dissolved oxygen content in ppm from the river above the factory.
Let Y be the dissolved oxygen content in ppm from the river below the factory.
Since the variances are equal and unknown, the t distribution should be used.
The pooled estimate of common variance s_p^2 can be used to estimate σ^2:

$s_p^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}$.

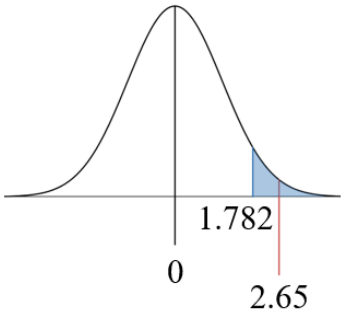
Since $n_x = 6$, $n_y = 8$ and from the samples, $s_x^2 = 0.0667$ and $s_y^2 = 0.3736$, then

$$s_p^2 = \frac{5 \times 0.0667 + 7 \times 0.3736}{6 + 8 - 2} = 0.2457.$$

$\mathrm{H}_0: \mu_X - \mu_Y = 0.2$ *ppm*
$\mathrm{H}_1: \mu_X - \mu_Y > 0.2$ *ppm,*
where μ_X and μ_Y are the population mean dissolved oxygen contents from the rivers above and below the factory respectively.
A one-tailed t-test at the 5% significance level with $6 + 8 - 2 = 12$ degrees of freedom is required.
The sample means are $\overline{x} = 4.9333$ and $\overline{y} = 4.025$, so $\overline{x} - \overline{y} = 0.908$.

Method 1: Using standardised critical regions

The test statistic is $\frac{0.908-0.2}{\sqrt{0.2457\left(\frac{1}{6}+\frac{1}{8}\right)}} = 2.65$

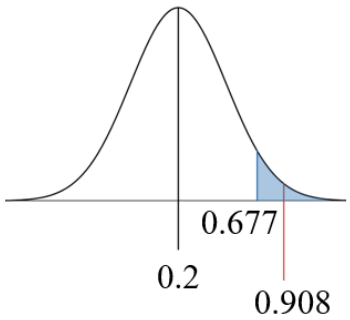


The critical region (from the tables or calculator) is 1.782 or more.
Since $2.65 > 1.782$*, the result is significant. We reject* H_0.

Method 2: Non-standardised critical regions

The test statistic is 0.908.

The standard critical t-value is 1.782, so the critical region is $0.2 + 1.782\sqrt{0.2457\left(\frac{1}{6}+\frac{1}{8}\right)} = 0.677$



Since $0.908 > 0.677$*, the result is significant. We reject* H_0.

There is significance evidence to suggest that the mean dissolved oxygen content in the river above the factory is more than 0.2 ppm higher than the mean dissolved oxygen content in the river below the factory.

b) What assumptions did you have to make in order for your test in (a) to be valid?

We had to assume that the population of dissolved oxygen content both above and below the factory were independent and normally distributed with equal variances.
We also must assume the two samples of water had been obtained independently and randomly.

OPPORTUNITIES FOR EMBEDDING THE SEC

In addition to the SEC guidance in the unit summary, the following question can further embed the SEC into questions:

Exemplar

One sample taken from below the factory gave a reading of 9.6. Investigation showed that this arose from an error in the measuring equipment, and it was decided to exclude this result from the data given above. Comment on this decision and state which assumptions might have been violated if the result had been included in the analysis.

- *It was wise to omit it, because it was known to be a wrong reading, so it did not come from the population of readings of oxygen content.*
- *If the reading 9.6 ppm had been included, the distribution of oxygen content below the factory would have had positive skew, so the assumption that both populations had normal distributions would not be valid.*
- *Also, the distribution of oxygen content below the factory would have had a larger variance, so the assumption that the populations had equal variances would be violated.*

This question addresses **B1** (checking measuring equipment), **C3** (misrepresentation), **D4** (reliability of findings), **E3** (explaining why the value should be omitted in order to improve the statistical process).

COMMON AND POSSIBLE MISTAKES

See common and possible mistakes in Unit 11a, Unit 16c or Unit 19b.

NOTES

Note that the assumption that the variances of both populations being equal is a strong one. Without extensive exploratory data analysis or a hypothesis test about the variance, this assumption cannot be justified. In this course, this assumption will always be made (or students must state this as an assumption). However, in A level Further Maths (Further Statistics 2), a hypothesis test for equality of variances is usually carried out in order to justify this assumption.

Students will see the F-distribution and carry out hypothesis tests using the F-distribution in Unit 24. However, as an extension exercise, it would beneficial for students who will use statistics in their future studies or career to see how a hypothesis test for variability is carried out.

For normally distributed populations X and Y with variances σ_X^2 and σ_Y^2 respectively, and samples taken from X and Y with sizes n_X and n_Y and sample variances s_X^2 and s_Y^2

respectively, then $\frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2}$ has an F-distribution with $n_X - 1$ degrees of freedom in the numerator and $n_Y - 1$ degrees of freedom in the denominator.

For a hypothesis test about whether or not two population variances are equal, the null hypothesis is $\sigma_X^2 = \sigma_Y^2 = \sigma$, which means that in the above ratio, $\frac{s_X^2}{s_Y^2}$ follows an F-distribution. The numerator is the larger of the sample variances. The critical values are upper tailed and can be found in the F-distribution tables. The test statistic is $\frac{s_X^2}{s_Y^2}$.

For the factory example earlier, normally the first investigation would be:

Extension Exemplar

Investigate, at the 5% significance level, whether the variability of the dissolved oxygen is the same above and below the factory.

Let X be the dissolved oxygen content in ppm from the river above the factory.
Let Y be the dissolved oxygen content in ppm from the river below the factory.

$\mathrm{H}_0: \sigma_X = \sigma_Y$,
$\mathrm{H}_1: \sigma_X \neq \sigma_Y$,
where σ_X and σ_Y are the population variances of dissolved oxygen content in ppm from the rivers above and below the factory respectively.

We will carry out a two-tailed F-test at the 5% significance level.

From the samples, $s_X^2 = 0.0667$ and $s_Y^2 = 0.3736$, so the test statistic is $F = \frac{0.3736}{0.0667} = 5.601$. Since s_Y^2 is in the numerator, $\nu_1 = 7$ and $\nu_2 = 5$.

From the tables, the critical value is 6.85, so the critical region is anything bigger than 6.85.

Since $5.601 < 6.85$, the result is not significant. We do not reject H_0, so there is insufficient evidence to suggest that the variances of the populations are not equal.

It is one of the easier hypothesis tests to carry out, and F-tests are on the specification (albeit only in Unit 24: ANOVA). In the case when the variances are not equal (or cannot be assumed equal), students may be asked to use (or may need to identify) a Wilcoxon rank-sum test (only for a test of whether difference between two medians is zero).

Note: this test is <u>not</u> in the defined content for the new specification so common variances must be assumed.

Outside of the A level course, if the variances are not equal and normal distributions may be assumed, then the Welch's *t*-test would be appropriate. The Welch's *t*-test is **<u>not</u>** in the new specification but is a standard test and well documented.

21c. Hypothesis Tests for the Difference Between Two Population Parameters: two proportions (16.4)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Calculate the pooled estimate of the proportion of two samples
- Carry out a hypothesis test for the difference of two normally distributed population means with unknown variance
- Interpret the results of the hypothesis tests in context

TEACHING POINTS

Students may need reminding that sample variances need to be "pooled" when dealing with two populations (see Unit 21b). The same is true for proportions – we are testing to see if the proportions are different, but under the null hypothesis they are equal so a pooled estimate of the proportion is required. This is best explained with an example:

If 50% of a sample of size 10 suffer from a particular disease, and 75% of a sample of size 20 suffer from this disease, what proportion of the total population suffer from this disease?

Since 50% of 10 is 5 and 75% of 20 is 15, when the samples are combined 20 out of the 30 have the disease, so the proportion is $\frac{2}{3}$.
Explain that the 5 came from multiplying the first sample size by the first sample proportion $(p_1 \times n_1)$ and the 15 was obtained likewise for the second sample $(p_2 \times n_2)$. The combined sample has size $n_1 + n_2$ and so the new proportion is $p = \frac{p_1 \times n_1 + p_2 \times n_2}{n_1 + n_2}$.
This is in the formula book.

Simple examples could be practised here, so students can calculate the pooled proportion.

At this point, if the sampling distribution of the proportion has not been seen, it could be seen now. Refer to the notes of Unit 9c and Unit 11b. In order to carry out a hypothesis test for the difference between two proportions, the sampling distribution of the proportion must be used. Students do not need to know the following, but will gain a greater appreciation of the hypothesis test if they do.

If we have two sampling distributions of the proportion, say $\frac{X}{n}$ and $\frac{Y}{n}$ with proportions p_X and p_Y respectively, then the difference $\frac{X}{n_X} - \frac{Y}{n_Y} \sim \mathrm{N}\left(p_X - p_Y, \ \frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}\right)$
where n_X and n_Y are the sample sizes of the respective sampling distributions. Since we are testing whether the proportions are equal or not, our null hypothesis is $\mathrm{H}_0: p_X = p_Y$, or $\mathrm{H}_0: p_X - p_Y = 0$.

Hence, under the null hypothesis, $\frac{X}{n_X} - \frac{Y}{n_Y} \sim \mathrm{N}\left(0, \ p(1-p)\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)\right)$, where

$p_X = p_Y = p$, the estimate of the population proportion using the pooled proportion of the two samples.

The remainder of the hypothesis test continues as in Unit 21a, with the test statistic being the difference in the two sample proportions. It must be stressed to students that the samples must be sufficiently large $(n \geq 30)$ for this test to be valid, as it uses the normal approximation to the binomial distribution (no continuity correction required)

There are three methods exemplified here:

- Using standardised critical regions.
  The critical values are found using the $Z \sim N(0,1)$ distribution.
  The test statistic is $\frac{p_1 - p_2}{\text{standard error}}$
  where standard error $= \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$
  and $p = \frac{p_1 \times n_1 + p_2 \times n_2}{n_1 + n_2}$
  This is given in the formula book.
- Using non-standardised critical regions.
  The critical values are found using the $\frac{X}{n_X} - \frac{Y}{n_Y} \sim \mathrm{N}\left(0, p(1-p)\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)\right)$
  distribution.
  The test statistic is $\frac{x}{n_x} - \frac{y}{n_y}$.
- Using *p*-values.
  These may be calculated using either of the above distributions with the corresponding test statistics.

Exemplar

A company that has been criticised for their lack of diversity amongst their employees decides to roll out a new recruitment initiative.
Prior to this initiative, the Chief Executive Officer (CEO) of this company takes a random sample of 36 employees who have voluntarily declared their ethnicity. She finds that 14 people in her sample have declared themselves as having an ethnicity other than "White". After the new initiative is rolled out, she takes a new random sample of 32 employees who have voluntarily declared their ethnicity. In this new sample, 18 have declared themselves as having an ethnicity other than "White".

Investigate, at the 5% significance level, whether the proportion of employees who have declared their ethnicity as other than "White" has changed since the new initiative was rolled out.

Let X be the number of employees who have voluntarily declared their ethnicity as other than "White" from before the new initiative.
Let Y be the number of employees who have voluntarily declared their ethnicity as other than "White" from after the new initiative. Then $X \sim \mathrm{B}(36, \pi_X)$ and $Y \sim \mathrm{B}(32, \pi_Y)$.

$H_0: \pi_X - \pi_Y = 0$,
$H_1: \pi_X - \pi_Y \neq 0$,
where π_X and π_Y are the proportion of the population of employees who have voluntarily declared their ethnicity as other than "White" from before and after the initiative was rolled out respectively.

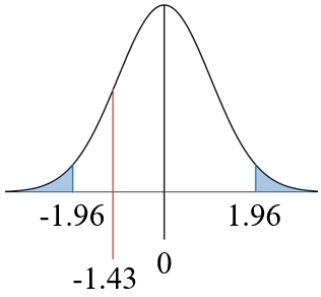
The pooled estimate of the proportion is $p = \frac{14+18}{36+32} = \frac{8}{17}$.
We also have $\left(\frac{8}{17}\right)\left(1 - \frac{8}{17}\right)\left(\frac{1}{36} + \frac{1}{32}\right) = \frac{1}{68}$.

A two-tailed difference of proportions test at the 5% significance level is required.

Method 1: Using standardised critical regions

Using $Z \sim N(0,1)$*, the standard error is* $\sqrt{\frac{1}{68}} = 0.121\ldots$

Therefore the test statistic is $\frac{\frac{14}{36}-\frac{18}{32}}{0.121\ldots} = -1.43$

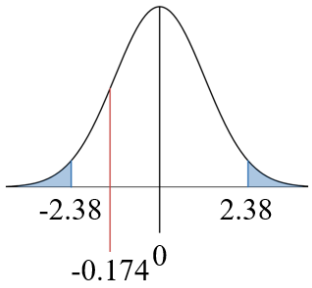


The critical values are ± 1.96.
Since $-1.96 \leq -1.43 \leq 1.96$*, the result is not significant. Do not reject* H_0.

Method 2: Using non-standardised critical regions

Using $\frac{X}{36} - \frac{Y}{32} \sim N\left(0, \frac{1}{68}\right)$

The test statistic is $\frac{14}{36} - \frac{18}{32} = -0.174$



Using the calculator, the critical values are ± 2.38.
Since $-2.38 \leq -0.174 \leq 2.38$*, the result is not significant. Do not reject* H_0.

Method 3: Using p-values

Using $\frac{X}{36} - \frac{Y}{32} \sim \mathrm{N}\left(0, \frac{1}{68}\right)$ *(or* $Z \sim \mathrm{N}(0,1)$*)*
We have $P\left(\frac{X}{36} - \frac{Y}{32} \leq -0.174\right)$ *(or* $P(Z \leq -1.43)$*) = 0.0761.*
So the two-tailed p-value is $2 \times 0.0761 = 0.152 > 0.05$ *so the result is not significant. Do not reject* H_0.

There is insufficient evidence to suggest that there is a difference in the proportion of employees who have declared their ethnicity as something other than "White" before and after the initiative was rolled out.

OPPORTUNITIES FOR EMBEDDING THE SEC

In addition to the guidance given in the unit summary, the opportunities for embedding the SEC are the same as those in Unit 21b.

In the example above, the following question may be added:

Exemplar

Explain why your conclusion may not be valid for the population of employees in the company.
The samples obtained are from the population of employees who have voluntarily declared their ethnicity. It is unknown whether the population of all employees have a similar proportion to the samples taken.

Specific to this sub-unit, the samples must be sufficiently large. This is a condition for the normal approximation to the binomial distribution. Point **E3** can be explored by referring to the sample size and how a small sample size would violate the use of the normal approximation.

COMMON AND POSSIBLE MISTAKES

The test statistic, $\boldsymbol{z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}}}$, is incorrectly evaluated, often because student fail to use the pooled estimate of the population proportion in the denominator.

See also common and possible mistakes in Unit 11a, Unit 16c or Unit 19b.

NOTES

As an extension activity, you could use the sampling distribution of the proportion to find the confidence intervals for *p*.

It is acceptable to use p for the population proportion provided subscripts are clearly defined (note that p is also the symbol for the pooled estimate of the proportion)

Continuity corrections are not required for the differences between two proportions.

UNIT 22: Goodness of Fit

Return to overview

SPECIFICATION REFERENCES

19.1 Conduct a statistical goodness of fit test for binomial, Poisson, normal and exponential distributions or for a specified discrete distribution using $\sum\frac{(O-E)^2}{E}$ as an approximate χ^2 statistic.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics
Discrete Random Variables (Unit 4)
Binomial Distribution (Unit 5)
Continuous Uniform Distribution (Unit 7)
Normal Distribution (Unit 7)
χ^2 tests and interpreting contributions to the test statistic (Unit 12)
Hypothesis Testing: terminology (Unit 10a)

Year 2 of A Level Statistics
Poisson Distribution (Unit 15)
Hypothesis Testing: concepts (Units 19c and 19d)
Exponential Distribution (Unit 20)

KEYWORDS

alternative hypothesis, binomial, Central Limit Theorem, chi-squared, continuous, critical region, critical value, difference, discrete, distribution, error, exponential, goodness of fit, hypothesis, independent, insufficient, mean, normal, null hypothesis, Poisson, pooled estimate, population, power, probability, proportion, reject, sample, sampling distribution, significance level, significant, standard deviation, sufficient, test statistic, Type I error, Type II error, variance,

UNIT SUMMARY

This unit continues the theme of hypothesis testing, but with more emphasis on modelling real-world contexts using theoretical distributions. We begin with the largest part of this unit: calculating expected frequencies from various distributions. It is highly advisable to give students plenty of practice here. The second sub-unit focusses on the hypothesis test itself.

You could use the Model Fitting activity on Desmos to help students visualise the idea of goodness of fit.

The opportunities for embedding the SEC will be the same in each sub-unit, so is presented here as an overview:

A1 Identifying possible factors which may lead to the identification of an appropriate distribution model.

A2 Defining an appropriate null and alternative hypothesis from the context of the question.

A3 Describing a suitable sampling method for collecting samples from a situation.

A4 Using exploratory data analysis to estimate the population parameters.

A5 Identifying an appropriate model to test for goodness of fit, and planning how to do the test.

A6 Justifying the hypothesis test referring to relevant assumptions.

B1 Appreciating the practical constraints in taking large random samples.

B2 Appreciating the need for secondary data from multiple sources (in reference to **A4**).

B3 Recording **A3** and understanding the need for this to be declared.

B4 Understanding the context in which samples have been obtained and whether bias might be introduced through poor experimental design.

C1 Calculating numerical measures from a sample, and calculating probabilities including the power of a hypothesis test.

C2 Calculating numerical measures from summary statistics generated by technology.

C3 Understanding the dangers of defining a hypothesis or significance level after analysing the samples.

D1 Interpreting numerical measures in context.

D2 Interpreting the results of a hypothesis test in context, including referring to the contributions to the test statistic to interpret a significant result.

D3 Carrying out an appropriate hypothesis test, and determining whether the result is significant.

D4 Referring to the significance level of a hypothesis test, the sampling methods used, the probability of a Type II error or the power of a hypothesis test.

D5 Reaching conclusions in context in a language appropriate to a given target audience.

E1 Understanding the disadvantages of using a certain sampling method, or referring to the probability of making a Type I or Type II error, or the power of a hypothesis test.

E2 Understanding that the conclusions of a hypothesis test may not be valid if the sampling technique is not random, or the sample size is insufficiently large.

E3 Suggesting alternative methods of sampling and/or hypothesis tests, together with any assumptions made or relaxed. Also identifying when to group categories in order to use a goodness of fit test.

E4 Redoing a hypothesis test with a new sample incorporating the suggestions made in **E3**.

22a. Goodness of Fit: Calculating Expected Frequencies for a Goodness of Fit Test (19.1)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Calculate observed frequencies from a table of percentages.
- Calculate expected frequencies using probabilities obtained from standard probability distributions.

TEACHING POINTS

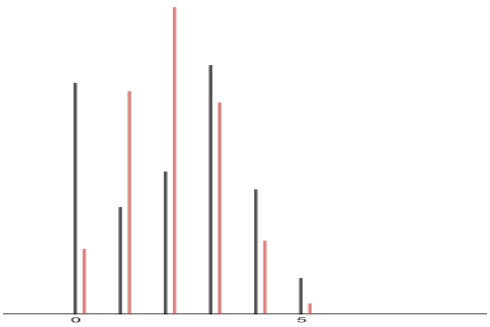
Revise the binomial, Poisson, normal and exponential distributions, specifically finding probabilities.

Starting with a simple example: flip a coin 5 times and record the number of heads. Ask students:

- to identify what proportion of heads they *expect* (emphasise the word *expect*),
- what distribution do they believe the number of heads should follow? (binomial)
- how could they test their belief that the proportion of heads is different from what they think? (a hypothesis test about a proportion)
- how could they test their belief that their chosen distribution is valid? (direct them to the conditions for a binomial distribution and allow students to identify what assumptions are made)

In the past, they may have compared the context to the conditions for a binomial distribution or compared the mean and variance of a binomial to the mean and variance of their collected data. Ask students why this isn't sufficient (e.g. subjective differences between observed and expected mean, can't be sure that events are independent etc.)

This can provide enough of an introduction to the Goodness of Fit test. Either repeat the experiment a few times, or get the students to each replicate the experiment twice, recording the number of heads each time. Use the Model Fitting activity on Desmos and input the data from the coin flipping experiment, and set it to binomial:



This could allow students to see whether their experiment fits a binomial distribution or not (in this screenshot, it does not). The red lines represent the binomial distribution and the black lines represent the observed data.

One suggested order: begin with qualitative random variables with given tabulated probabilities, then discrete random variables with tabulated probabilities, binomial, Poisson, continuous uniform, normal then finally exponential. Start with calculating the expected frequencies, given the population parameters. Expected frequencies are calculated by multiplying the probability by the sample size.

Qualitative random variables and discrete random variables with tabulated probabilities are a revision of GCSE:

Exemplar

A certain variety of plant produces white, pink or blue flowers. According to a genetic theory, the probabilities that the plants will be white, pink or blue-flowering are given below:

White	**Pink**	**Blue**
$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$

72 randomly chosen plants were studied. Calculate the expected numbers of white, pink and blue flowers from this sample.

	White	*Pink*	*Blue*
Probability	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$
Expected Frequency	$\frac{1}{6} \times 72 = 12$	$\frac{1}{3} \times 72 = 24$	$\frac{1}{2} \times 72 = 36$

Questions involving ratios could also arise. A modification to the above example could be: **According to a genetic theory, the plants will be white, pink or blue-flowering in the ratio 1:2:3**. This will remove the tabulated probability distribution and students will be required to calculate these probabilities before anything else.

You could embed the SEC here in a number of ways. The following requires students to identify the correct factor to be modelled in the question, as well as determining the probability of success in each occurrence. To further embed the SEC, you could omit part (a).

Exemplar

A coin is tossed to determine which cricket team should choose whether to bat first. The umpire of the local cricket team keeps 2 coins in his pocket so that he always has a coin available. A member of a cricket team claimed that the coins were biased. He borrowed the 2 coins and tossed them 40 times. He recorded the number of heads at each toss.

a) **Write down the name of the probability distribution which models the number of heads showing, if the coins were unbiased. State clearly the population parameters.**
Let X be the number of heads showing after each toss. Then $X \sim \mathrm{B}(2,0.5)$.

b) **Calculate the expected values if the coins were unbiased.**
Using the calculator:

x	***0***	***1***	***2***
$\mathrm{P}(X = x)$	*0.25*	*0.5*	*0.25*
Expected frequency	*10*	*20*	*10*

After enough practice has been given using the population parameters, start giving observed frequencies in order for students to estimate the population parameters.

Exemplar

The head of a large computing department keeps a record of the number of computers which break down on any particular day.

Number of computers which fail (x)	**0**	**1**	**2**	**3**	**4**	**5**
Frequency	**41**	**25**	**18**	**8**	**5**	**3**

In order to decide how many spare computers to keep, she wishes to model the situation using a Poisson distribution.

By estimating the appropriate population parameter, calculate the expected frequencies.

The population parameter of a Poisson distribution is λ which is the mean.
The mean of the observed frequencies is 1.2 (from the calculator).
The total frequency is 100.
Let X be the number of computers which fail. Assuming $X \sim \mathrm{Po}(1.2)$:

x	*0*	*1*	*2*	*3*	*4*	***5 or more****
$\mathrm{P}(X = x)$	*0.3012*	*0.3614*	*0.2168*	*0.0867*	*0.026*	*0.0079*
Expected Frequency $(\mathrm{P}(X = x) \times 100)$	*30.12*	*36.14*	*21.68*	*8.67*	*2.6*	*0.79*

*****Students must remember that, since the Poisson distribution has no upper limit, this must be taken into account when finding the final expected frequency.

A modification to this style of question could be to give observed percentages instead of frequencies. Students would then be required to calculate the observed frequencies prior to calculating the expected frequencies.

For continuous distributions, again start with calculating expected frequencies with given population parameters. Obviously, a grouped frequency table would be given in these cases.

Exemplar

A weaving mill produces pieces of fabric of nominal length 70 m and standard deviation 2.9 m. It is suspected that a <u>normal</u> distribution may be used to model the length of a piece of fabric.
The lengths of a random sample of 150 pieces were measured, and grouped into the following categories:

Length (m)	$\mathbf{65 - 67}$	$\mathbf{67 - 69}$	$\mathbf{69 - 71}$	$\mathbf{71 - 73}$	$\mathbf{73 - 75}$

Calculate the expected frequencies.

Let X be the length of a piece of fabric. Assuming $X \sim \mathrm{N}(70, 2.9^2)$ and using the calculator:

Length (m)	***65 or less***	$65 - 67$	$67 - 69$	$69 - 71$	$71 - 73$	$73 - 75$	***75 or more***
Probability	*0.0423*	*0.1081*	*0.2147*	*0.2698*	*0.2147*	*0.1081*	*0.0423*
Expected frequency	*6.345*	*16.215*	*32.205*	*40.47*	*32.205*	*16.215*	*6.345*

*****Since the normal distribution has no lower or upper limit, the lowest and highest classes must not specify limits. Extra classes,' *65 or less' and* '*75 or more',* are introduced here for this reason, and students need to understand that these classes would have observed frequencies of zero. You can then reintroduce observed

frequencies and finding estimates of the population parameters from the data given (or from summary statistics).

So far, all the probabilities can be calculated from the relevant distribution modes on the calculators. The continuous uniform distribution and the exponential distributions require manual calculation of the probabilities.

Exemplar

A test for reaction times is carried out by participants sitting in front of a screen with a button. In each test, a whistle is blown and a red light will flash on the screen at a random time between 0 and 10 seconds after the whistle is blown. The participants are instructed to press the button when they first see the red light. The time at which the red light flashes is assumed to be equally likely between 0 and 10 seconds after the whistle is blown.

The time at which the red light appears after the whistle is blown is recorded for 94 participants.

Time after the whistle is blown	**0 – 2**	**2 – 5**	**5 – 6**	**6 – 8**	**8 – 10**
Frequency	**7**	**6**	**19**	**25**	**37**

Calculate the expected frequencies for each category.

Let X be the time the red light flashes after the whistle is blown. Using the continuous uniform distribution over $0 \leq X \leq 10$*,*

Time after the whistle is blown	***0 – 2***	***2 – 5***	***5 – 6***	***6 – 8***	***8 – 10***
Probability	$2 \times \frac{1}{10} = 0.2$	$3 \times \frac{1}{10} = 0.3$	$1 \times \frac{1}{10} = 0.1$	$2 \times \frac{1}{10} = 0.2$	$2 \times \frac{1}{10} = 0.2$
Expected frequency	*18.8*	*28.2*	*9.4*	*18.8*	*18.8*

The exponential distribution will be the one most prone to error, due to the manual input involved.

Exemplar

The head of a large computing department keeps records from which she can calculate the time, in hours, between computer failures. She suspects the failures occur randomly and independently, and at a constant average rate. She calculated the times between 50 consecutive computer failures, and the percentage in each class was as follows:

Time (in hours)	$\mathbf{0-12}$	$\mathbf{12-24}$	$\mathbf{24-48}$	$\mathbf{48-72}$	**72 or more**
Percentage	$\mathbf{40\%}$	$\mathbf{28\%}$	$\mathbf{20\%}$	$\mathbf{6\%}$	$\mathbf{6\%}$

Using an appropriate probability distribution, calculate the expected frequencies.

Since the department head thinks the failures occur randomly, independently and at a constant rate, an ***exponential distribution*** *could be used to model the time between failures. Since 50 computer failures were recorded:*

Time (in hours)	$0-12$	$12-24$	$24-48$	$48-72$	$72-$
Observed Frequency	20	14	10	3	3

Using the calculator, the mean of this sample is 23.28*, so* $\lambda = \frac{1}{23.28} = \frac{25}{582}$*. Let* X *be the time, in hours, between computer failures. Assuming* $X \sim \text{Exp}(\lambda)$*:*

Time (in hours)	$0-12$	$12-24$	$24-48$	$48-72$	$72-$
Probability	$1-e^{-\lambda\times12}$	$e^{-\lambda\times12}$ $-e^{-\lambda\times24}$	$e^{-\lambda\times24}$ $-e^{-\lambda\times48}$	$e^{-\lambda\times48}$ $-e^{-\lambda\times72}$	$1-0.9546$
	0.4028	*0.2405*	*0.2295*	*0.0818*	*0.0454*
Expected Frequency	*20.14*	*12.025*	*11.475*	*4.09*	*2.27*

As mentioned in Unit 20, the use of the storage function on the calculator will make these calculations easier to input. Students need to remember that (like the Poisson and normal distributions) since there is no upper limit to an exponential distribution, the upper limit should be removed when calculating the expected frequencies. The final cell should then be calculated by either subtracting the sum of the other cells from 1 or using the result $P(X \geq a) = e^{-\lambda a}$.

OPPORTUNITIES FOR EMBEDDING THE SEC

The SEC has been embedded throughout the teaching points.

COMMON AND POSSIBLE MISTAKES

- Forgetting that percentages are not frequencies.
- Using the incorrect binomial or Poisson mode on their calculator (finding $P(X \leq x)$ instead of $P(X = x)$).
- Confusing the binomial parameter *n,* the total number of trials, with the total of the observed frequencies.
- Confusing the binomial parameter *n* , the total number of trials, with the total number of categories (e.g. categories 0 , 1, 2, 3, 4, 5 does not mean that *n* = 6)
- Forgetting that the final interval is unbounded for Poisson and exponential distributions.
- Forgetting that both the first and last intervals are unbounded for the normal distribution.
- Saying “constant rate” instead of “constant mean rate” (if the rate were constant, the variable would not be random).
- Coping with boundaries for continuous random variables of the form, for example, $40 - 44$, $45 - 49$ etc.

22b. Goodness of Fit: Hypothesis tests for the Goodness of Fit (19.1)

Teaching time
4 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Calculate the test statistic for a goodness of fit hypothesis test.
- Carry out a goodness of fit hypothesis test.
- Determine the number of degrees of freedom.
- Interpret the results of a goodness of fit hypothesis test.
- Know when to group classes or outcomes together in order to validate the goodness of fit test.
- Justify the reasons leading to the rejection of the null hypothesis.

TEACHING POINTS

Revise the main points of a χ^2 test for association with contingency tables, emphasising the test statistic $\sum \frac{(O-E)^2}{E}$ and that this statistic approximately follows a χ^2 distribution provided that the expected frequencies are at least 5. Again, use the chi-squared distribution activity on Desmos to help revise. Also remind students about the analysis of the contributions to the test statistic and how to interpret them in context.

Most of the work may have been done in the previous sub-unit. Calculating the test statistic and the hypothesis test (including the grouping of small expected frequencies) is the same as in Unit 12b. The only change is the determining the number of degrees of freedom.

The number of degrees of freedom is calculated as the number of constraints subtracted from the number of cells. Treating the table of expected frequencies as a $1 \times m$ contingency table, the number of degrees of freedom is $m - 1$ (this is because the total expected is equal to the total observed – students are not expected to know this). However, students need to appreciate that for each population parameter estimated, they should subtract another degree of freedom.

For example, for a random variable $X \sim \mathrm{B}(5, p)$, then the number of outcomes for X is 6 (the values 0 to 5). If *p* is known, then the number of degrees of freedom, ν, is $6 - 1 = 5$. However, if p is estimated from the sample data, then the number of degrees of freedom is $\nu = 6 - 1 - 1 = 4$.
Another example, a normally distributed random variable $Y \sim \mathrm{N}(\mu, \sigma^2)$ with a sample randomly obtained from Y grouped into 8 classes: if the mean and variance are known, then the number of degrees of freedom is $\nu = 8 - 1 = 7$. For each population parameter estimated, you would subtract another degree of freedom.

- The null hypothesis is H_0: The distribution is a suitable model for the observed data.
  The alternative hypothesis is H_1: The distribution is not a suitable model for the observed data.
  For specific distributions with given parameters, the null and alternative hypotheses can be stated in symbols, like $X \sim \mathrm{N}(5, 2.7^2)$.

Ensure that students state the name of the distribution that they are fitting.

- a qualitative or discrete quantitative random variable with given tabulated probabilities ($\nu =$ number of cells -1);
- a binomial distribution ($\nu =$ number of cells -1, if p is given, or
  $\nu =$ number of cells -2, if p is estimated);
- a Poisson distribution ($\nu =$ number of cells -1, if λ is given, or
  $\nu =$ number of cells -2, if λ is estimated);
- an Exponential distribution (ν as for Poisson);
- or a normal distribution ($\nu =$ number of cells -1 if both μ and σ are given,
  $\nu =$ number of cells -2 if exactly one of μ or σ are estimated, or
  $\nu =$ number of cells -3 if both μ and σ are estimated).

The easiest way to estimate the appropriate parameters is by using the sample mean of the observed data. The following table may be useful to summarise these methods.

Distribution	Parameter(s)	Estimate(s)	Reason
Binomial	p	$\frac{\overline{x}}{n}$	The mean of a binomial is np.
Poisson	λ	$\overline{x}$	The mean of a Poisson distribution is λ.
Exponential	λ	$\frac{1}{\overline{x}}$	The mean of an exponential distribution is $\frac{1}{\lambda}$.
Normal	μ	$\overline{x}$	The sample mean is an unbiased estimate for μ.
	σ^2	s_x^2	The sample variance is an unbiased estimate for σ^2.

Begin with examples where **none** of the classes needs to be grouped. To ease students into the hypothesis tests, begin with the expected values already calculated and the observed values given. Once students get more familiar with the hypothesis test, move the starting point further back, incorporating work from the previous sub-unit. The example below shows a "full" goodness of fit question.

Exemplar

The manager of a workshop has introduced a scheme to reduce the waste offcut of hardwood timber. The manager thinks that the distribution of lengths of waste timber can be modelled by a normal distribution. 100 lengths of waste timber are measured and the results are as follows:

Length l (mm)	$l < 20$	$20 \leq l < 40$	$40 \leq l < 60$	$60 \leq l < 80$	$80 \leq l < 100$	$100 \leq l < 120$
Frequency	**5**	**16**	**33**	**23**	**15**	**8**

Carry out a χ^2 test at the 5% significance level to see whether a <u>normal distribution</u> is a suitable model.

Let X be the length of waste timber. From the observed data, the sample mean is $\bar{x} = 60.2$ and the standard deviation is $s_x = 25.819$ (obtained from the calculator). These are the estimates for μ and σ.

H_0: *The normal distribution is a suitable model for the length of waste timber,*
H_1: *The normal distribution is not a suitable model for the length of waste timber.*

From the calculator:

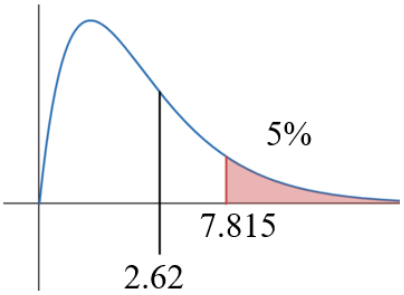
l	$l < 20$	$20 \leq l < 40$	$40 \leq l < 60$	$60 \leq l < 80$	$80 \leq l < 100$	$100 \leq l \leq 120$	$l \geq 120$
Probability	*0.0597*	*0.1573*	*0.2800*	*0.2815*	*0.1600*	*0.0513*	*0.0102*
Expected Frequencies	*5.97*	*15.73*	*28.00*	*28.15*	*16.00*	*5.13*	*1.02*

Since the final class has an expected frequency smaller than 5, we pool (combine) the last two classes:

l	$l < 20$	$20 \leq l < 40$	$40 \leq l < 60$	$60 \leq l < 80$	$80 \leq l < 100$	$100 \leq l$
Observed Frequencies	*5*	*16*	*33*	*23*	*15*	*8*
Expected Frequencies	*5.97*	*15.73*	*28.00*	*28.15*	*16.00*	*6.15*
$\frac{(O-E)^2}{E}$	*0.1576*	*0.0046*	*0.8929*	*0.9422*	*0.0625*	*0.5565*

Conduct a χ^2 test at the 5% significance level using $\nu = 6 - 1 - 2 = 3$ degrees of freedom.
(This is because there are 6 classes after pooling data, resulting in $6 - 1$ degrees of freedom. However, both μ and σ^2 were estimated from the data, so a further two degrees of freedom are subtracted.)

The test statistic is $\chi^2 = 2.62$. The critical region is $\chi^2 \geq 7.815$.



Since $2.62 \leq 7.815$*, the result is not significant. We do not reject* H_0*, as there is insufficient evidence to suggest that the normal distribution is an unsuitable model for the length of waste timber.*

Note that *p*-values obtained from the calculator (for calculators that can) may be utilised to access full marks (in line with the mark scheme).

For small expected frequencies, the classes should be grouped in an appropriate way, and only adjacent cells should be grouped. Remind students that if classes are grouped together and the parameters are estimated, the number of degrees of freedom may be calculated 0 or less. This means the test is not valid due to the sample size (see SEC below).

For examples which lead to the rejection of the null hypothesis, students could then be able to identify and interpret the largest contributions to the test statistic. This is done in an analogous way to Unit 12c.

Exemplar

The head of a large computing department keeps a record of the number of computers which break down on any particular day.

No. of computers which fail (x)	**0**	**1**	**2**	**3**	**4**	**5**
Frequency	41	25	18	8	5	3

In order to decide how many spare computers to have, he wishes to model the situation using a Poisson distribution. Carry out a χ^2 **test at the 5% significance level to see whether a Poisson distribution is not suitable.**
If a Poisson distribution is not suitable, suggest reasons why not by making references to the contributions to the value of χ^2**.**

Let X *be the number of computers which break down on any particular day.*
From the sample, $\bar{x} = 1.2$ *and the sample has size 100. The estimate for λ = 1.2*

H_0: *A Poisson distribution is a suitable model for the number of computers which break down on any particular day,*

H_1: *A Poisson distribution is not a suitable model for the number of computers which break down on any particular day.*

From the calculator:

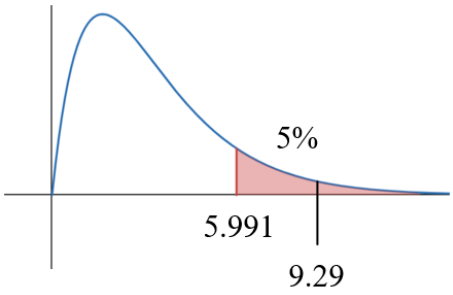
x	0	1	2	3	4	5 or more
$P(X = x)$	0.3012	0.3614	0.2169	0.0867	0.0260	0.0062
Expected Frequency	30.12	36.14	21.69	8.67	2.6	0.62

Since the expected frequencies for "$x = 4$" and "$x = 5$ or more" are less than 5, the cells are pooled:

x	0	1	2	3 or more
Observed Frequency	41	25	18	16
Expected Frequency	30.12	36.14	21.69	12.05
$\frac{(O-E)^2}{E}$	3.930	3.434	0.628	1.295

Conduct a χ^2 *test at the 5% significance level using* $\nu = 4 - 1 - 1 = 2$.
(This is because there are 4 classes after combining, and the value of λ *was estimated from the data)*

The test statistic is $\chi^2 = 9.29$. *The critical region is* $\chi^2 \geq 5.991$.



Since $9.29 \geq 5.991$, *the result is significant. We reject* H_0, *as there is significant evidence to suggest that a Poisson distribution may not be a suitable model.*
Since 3.93 *is a large contribution to* χ^2, *there are a lot more days than expected (41 as opposed to 30) where no computers fail.*
Also, 3.434 *is a large contribution to* χ^2, *there a lot fewer days than expected where only one computer failed (25 compared to 36).*

OPPORTUNITIES FOR EMBEDDING THE SEC

In addition to the unit summary and the previous sub-unit, points **E1**, **E3** and **E4** can be emphasised by acknowledging that if the population parameters are estimated and the data are not grouped into sufficiently divided classes, then the χ^2 test will not be valid (the number of degrees of freedom would be negative).

COMMON AND POSSIBLE MISTAKES

- Forgetting to group the cells.
- Forgetting to subtract one from the number of degrees of freedom for each parameter estimated.
- If the cells are grouped, using the number of degrees of freedom as if they weren't grouped.
- Writing H_0 incorrectly; it is important that students understand that data are given, so the data cannot "fit" a model – it is the model which fits the data.
- Not putting enough context in a conclusion.
- During the interpretation, comparing observed frequencies with other observed frequencies, rather than expected frequencies.
- Stating that a sample has a probability distribution (it is the population which may have a probability distribution).
- Students often make incorrect conclusions after a test of goodness of fit. If H_0 is rejected, there is significant evidence to suggest the distribution is not a suitable model. If H_0 is not rejected, there is insufficient evidence to suggest the distribution is not a suitable model.

Students must remember to state the degrees of freedom.

NOTES

Questions about identifying Type I and Type II errors could be asked, but students will not be required to calculate the risk of a Type II error or the power of this hypothesis test. This is due to the necessity of calculating p-values of a χ^2 distribution.

UNIT 23: Further Experimental Design

Return to overview

SPECIFICATION REFERENCES

13.1 Know and discuss issues involved in experimental design: experimental error, randomisation, replication, control and experimental groups, and blind and double blind trials.

13.2 Know the benefits of use of paired comparisons and blocking to reduce experimental error.

17.1 Use sign, Wilcoxon signed-rank or paired *t*-test, understanding appropriate test selection and interpreting the results in context.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics

Experimental design (Unit 14a)

Hypothesis tests for paired samples (Unit 14b and 14c)

Year 2 of A Level Statistics

Unpaired *z*-test and *t*-test (Unit 21)

KEYWORDS

alternative hypothesis, association, binomial, blind trials, blocking factor, blocking, control group, distribution-free, double blind trials, experimental design, experimental error, experimental group, hypothesis, independent, Mann-Whitney, median, non-parametric, null hypothesis, paired, population, randomisation, randomised block design, randomised design, rank, replication, sample, sign, significance level, *t*-distribution, unpaired, Wilcoxon signed-rank, Wilcoxon rank-sum,

UNIT SUMMARY

This unit builds on the introduction to experimental design (Unit 14). This unit is split into two focal points: the paired sample *t*-test and the theoretical concepts of experimental design together with their application. Students need to be able to justify their reasons for the design of experiments, referencing the ways of reducing bias and experimental error. A revision of the paired sign test and the paired Wilcoxon signed-rank test would be beneficial here to remind students of the purposes behind experimental design. Questions will often use the sign test and/or the Wilcoxon signed-rank test together with experimental design terminology.

This is a good unit to embed the SEC due to the focus on experimental design. The revision of the Sign test and the Wilcoxon Signed-rank test, together with the paired sample *t*-test will help complete the cycle.

A1 Identifying treatments, control groups, experimental groups etc. related to the factor under investigation.

A2 Defining the null and alternative hypotheses in relation to the context.

A3 Describing a data collection method, justifying the choice of method (referring to advantages, practicality and assumptions).

A4 Using exploratory data analysis to determine whether the underlying population satisfies the assumptions (e.g. symmetrical distribution, normal distribution).

A5 Understanding that the observations will be compared through analysis of the differences.

A6 Referring to experimental design concepts to reduce bias (randomisation) or experimental error (paired comparisons).

B1 Understanding the practicalities of taking a large random sample, identifying issues that cannot be overcome through experimental design.

B2 Using secondary source data and understanding the dangers of reporting findings (e.g. using a secondary source which is later retracted due to malpractice).

B3 Declaring and recording the methods described in **A3** and appreciating that the findings may not be accepted by the wider community.

B4 Showing an awareness of the need for blind or double blind trials to reduce the effect on outcomes.

C1 Calculating numerical measures and probabilities from collected data, using technology.

C2 Calculating numerical measures and probabilities from summary statistics generated by appropriate software.

C3 Appreciating that a lack of consideration of the sources of experimental error and bias can lead to misrepresentation.

D1 Analysing numerical measures and probabilities and comparing them to expected values.

D2 Interpreting the results of a hypothesis test relating to the context.

D3 Using an appropriate hypothesis test and determining the significance of the result.

D4 Referring to the data collection method, the underlying population and the external factors that may affect the dependent variable and discussing the effect on the findings.

D5 Reaching conclusions in context, using language appropriate to a given target audience.

E1 Identifying whether a sampling method described is suitable for the investigation, giving reasons.

E2 Appreciating that a conclusion may not be valid if the sample was not obtained randomly, or bias was not minimised, through a poor experimental design.

E3 Suggesting improvements to the process, referring to the underlying population or sampling technique.

E4 Describing ways of replicating the experiments to estimate the size of experimental error.

23a. Further Experimental Design: Terminology (13.1) (13.2) (13.3)	Teaching time 1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand the terms: experimental error, randomisation, replication, control and experimental groups, blind and double blind trials.
- Identify control groups and experimental groups in context.
- Identify situations where blind and double blind trials could be employed, with justification.
- Appreciate that replication can estimate the size experimental error and randomisation can reduce bias.

TEACHING POINTS

Revise Unit 14a, in particular the concepts of experimental error and paired comparisons.

A treatment is an aspect of the experiment that is to be compared (for example, in an experiment to compare the effect of vitamin supplements to improve mental health, the vitamin supplement is the treatment).

A control group receives no active treatment (or the standard treatment) and an experimental group receives an active treatment (or a new treatment). A placebo treatment is a treatment that looks like the real treatment, but actually has no effect on the dependent variable. In the above example, the control group would be the group of people who might take a placebo (e.g. an inert pill that looks like a vitamin supplement, but has no vitamin content) and the experimental group would be the group of people who take a vitamin supplement.

Experimental error is the variation in the dependent variable that arises from other sources other those taken into consideration (e.g. variation in people, plants, weather etc.).

Replication is repeating the experiment under apparently identical conditions. You can estimate the size of the experimental error through replication.

A paired comparison is a simple experimental design when two treatments are being compared. This reduces experimental error by minimising variation between sample elements.

Randomisation is the allocation of participants to the two groups in a random way. This could be through the use of random numbers. In the case of paired comparisons, this is the allocation of the different treatments to the participants or the order in which the experiments are carried out.

A blind trial is an experiment where the participants are unaware whether they are in the control group or the experimental group. A double blind trial is where neither the participants nor the researchers are aware of who is in the control group and who are in experimental groups. A blind trial is used to reduce the influence of expectations of the participants. A double blind trial will additionally reduce the influence of expectations of the staff looking after the participants.

For example, during an investigation into the efficacy of a drug, in a blind trial the patients would not be aware whether they are receiving the drug or a placebo. In a double blind trial the nurses or doctors are also unaware which patients are receiving the drug (apart from the research coordinator, who wouldn't be interacting with the patients). This ensures that the clinical outcomes are not affected by the expectations of the patients or the staff interacting with the patients.

Students will need some time and plenty of examples in order to understand and appreciate this terminology. Students must also be familiar enough with contexts to suggest why certain experimental designs are not appropriate, or even possible. Questions such as the following will help them understand not just the terminology but the purposes too:

Exemplar

As part of an investigation into the effects of caffeine on 16-19 year olds, a wellbeing charity identifies sets of three people who are closely matched in personal characteristics (age, race, socio-economic background and gender). For each set of three people, one person is given a placebo, one a low-strength caffeine tablet and one a high-strength caffeine tablet.

a) Identify the people who are part of the control group.
The control group are the people who were given a placebo.

b) **Explain what double blind trials are, and why they may be employed in this investigation.**
In double blind trials, neither the patients nor the people looking after the patients are aware which treatment has been allocated to whom. This is employed so the clinical outcomes are not affected by the expectations of patients or the staff who are looking after the patients.

OPPORTUNITIES FOR EMBEDDING THE SEC

Many areas of the initial planning and data collection stages can be embedded here. You could use the case study of Diederik Stapel, who was suspended from Tilburg University (Netherlands) for fabricated evidence, misrepresentation and omitting changes made between experiments. Other case studies can also be used for students to appreciate the importance behind declaring data collection methodologies and the ethics behind statistical studies.

COMMON AND POSSIBLE MISTAKES

- Confusing the definitions of “bias” and “experimental error”.
- Confusing which experimental designs reduce bias and which reduce experimental error.
- Giving definitions of experimental design terminology without reference to the context given. Although a student who can remember the formal definitions deserves merit, it is more beneficial for a student to apply the definitions to a wide range of contexts.
- Language and comprehension is another problem students struggle with.

NOTES

Completely randomised design will be seen in Unit 24b. Blocking and randomised block design will also be seen in Unit 24c. It is possible that questions on experimental design will be combined with hypothesis tests seen in Unit 14.

On the Maths Emporium (www.mathsemporium.com), you can find an end-of-topic test on this subunit here.

23b. Further Experimental Design: Hypothesis tests for the difference between two population means, using paired samples (17.1)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Carry out a paired sample *t*-test.
- Interpret the results of the hypothesis test.

TEACHING POINTS

This is a good opportunity to revise the paired versions of the sign test and the Wilcoxon Signed-Rank test from Unit 14, incorporating follow-up questions from the previous sub-unit. This will help to embed more stages of the SEC than was previous possible. Also revise the hypothesis test for the difference between two means with unknown variance (Unit 21b – the unpaired analogue of the *t*-test).

The null hypothesis is $H_0: \mu_D = a$ where μ_D is the difference in population means of the control and experimental groups respectively and a is a constant (usually zero, but not necessarily). The alternative hypotheses can have greater than, less than or not equal to. The assumptions required for a paired *t*-test are that the *differences* between the paired observations are normally distributed.

Suppose X_C and X_E are the observations from the control group and experimental group respectively such that each observation from one group has a corresponding pair in the other group. Let D be the differences between the pairs. If the assumption holds, then $D \sim \mathrm{N}(\mu_D, \sigma_D^2)$ where μ_D and σ_D are the population mean and standard deviation of the differences between the pairs.

There are two methods that will be exemplified here:

- Using standardised critical regions.
  The test statistic is $\frac{\overline{d}-\mu_D}{\frac{s_d}{\sqrt{n}}}$ and the critical values are the standard *t*-values given in the tables or the calculator. *(Note: Formula Book gives* $\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$ *and states 'Also used in matched pairs'.)*
- Using non-standardised critical regions.
  The test statistic is $\overline{d}$ and the critical values are obtained from $\mu_D + t \times \frac{s_d}{\sqrt{n}}$ where *t* is the standard *t*-value given in the tables or the calculator.

Note also that *p*-values obtained from calculators may be utilised to access full marks (in line with the mark scheme).

The remainder of the *t*-test is the same as the unpaired version (Unit 21b).

Exemplar

Two machines called analysers are used in a hospital laboratory to measure blood creatinine levels. To compare the performance of the two machines, a technician took seven specimens of blood and measured the creatinine level (in micromoles per litre) of each specimen using each machine. The results were as follows:

Specimen	1	2	3	4	5	6	7
Analyser A	119	173	100	99	77	124	73
Analyser B	106	153	83	100	69	123	67

a) Carry out a paired *t*-test at the 5% significance level to determine whether there is any difference between the analysers.

Let X_A be the reading of the creatinine level in micromoles per litre of the specimen using analyser A and let X_B be that for analyser B.
Let $D = X_A - X_B$.

$\mathrm{H}_0: \mu_D = 0,$
$\mathrm{H}_1: \mu_D \neq 0,$
where μ_D is the difference in mean in readings of the creatinine level of the specimen using analysers A and B respectively.
We will carry out a two-tailed paired t-test at the 5% significance level using $\nu = 7 - 1 = 6$

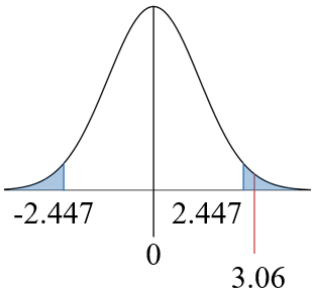
Specimen	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Analyser A	*119*	*173*	*100*	*99*	*77*	*124*	*73*
Analyser B	*106*	*153*	*83*	*100*	*69*	*123*	*67*
Difference	*13*	*20*	*17*	*-1*	*8*	*1*	*6*

From the calculator, $s_d = 7.90$ *and the test statistic* $\bar{d} = 9.14$.

Method 1: Using standardised critical regions

The test statistic in this case is $t = \frac{\bar{d} - \mu_D}{\frac{s_d}{\sqrt{n}}} = \frac{9.14 - 0}{7.90/\sqrt{7}} = 3.06$

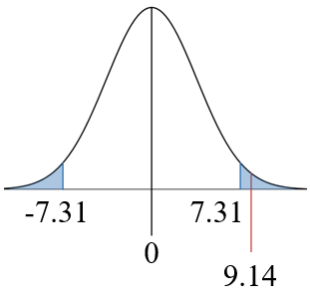
Standard critical value as given in the tables t with for $\nu = 6$ *is* ± 2.447



Since $3.06 > 2.447$ *the result is significant.*
We reject H_0*, so there is sufficient evidence to suggest that there is a difference between the mean readings of the two analysers.*

Method 2: Using non-standardised critical regions

The test statistic in this case is $\bar{d} = 9.14$
Standard critical value as given in the tables for t with $\nu = 6$ *is* ± 2.447*, hence the critical values are* $0 + 2.447 \times \frac{7.90}{\sqrt{7}} = 7.31$ *(and* -7.31*)*



Since 9.14 > 7.31, the result is significant.
We reject H_0*, so there is sufficient evidence to suggest that there is a difference between the mean readings of the two analysers.*

b) What assumptions have you had to make to be able to carry out this test?
We assumed that the differences between the readings of creatinine levels were normally distributed. We also assumed the specimens were taken at random.

The difference between the assumptions of the paired *t*-test and the Wilcoxon Signed-rank test is the assumption that the differences between the pairs are normally distributed. For the Wilcoxon Signed-rank test, the differences just need to be symmetrically distributed about the mean.

Note that if the sample size is sufficiently large ($n \geq 50$, say), the normal distribution can be used instead (this would be called a paired *z*-test, which is not within the scope of this course).

OPPORTUNITIES FOR EMBEDDING THE SEC

See the unit summary.

COMMON AND POSSIBLE MISTAKES

- Using either the unpaired *t*-test, the paired Wilcoxon signed-rank test or the Wilcoxon rank-sum test instead of the paired *t*-test.
- Consistency of the signs when using the critical value.
- Misreading the table of values for the *t*-distribution.

NOTES

It is highly possible that assessment questions for this unit will be combined with that of Unit 14. For example, students may be asked to carry out a paired sign test followed by a paired *t*-test and then compare the results at the end.

UNIT 24: Analysis of Variance (ANOVA)

Return to overview

SPECIFICATION REFERENCES

13.3 Use completely random and randomised block designs.

20.1 Conduct one-way analysis of variance, using a completely randomised design with appreciation of the underlying model with additive effects and experimental errors distributed as N(0, σ^2).

20.2 Conduct two-way analysis of variance without replicates, using a randomised block design with blocking.

20.3 Identify assumptions and interpretations in context.

PRIOR KNOWLEDGE

Year 1 of A Level Statistics
Numerical Measures (Unit 1)
Normal Distribution (Unit 7)
Experimental Design (Unit 14)

Year 2 of A Level Statistics
Unpaired *t*-test (Unit 21b)
Experimental Design and Paired *t*-test (Unit 23)

GCSE (9-1) in Mathematics at Higher Tier
A2 Substitute numerical values into formulae and expressions.

KEYWORDS

additive model, analysis, ANOVA, assumptions, between, blocking, columns, completely randomised design, degrees of freedom, effect, experimental error, *F* distribution, factor, independent, linear model, mean square, mean, one-factor, one-way, paired, pooled estimate of common variance, randomisation, randomised block design, rank-sum, rows, sum of squares, sum, *t*-distribution, total, two-factor, two-way, unpaired, variance, Wilcoxon, within,

UNIT SUMMARY

This unit extends the idea of the unpaired and paired *t*-tests seen in earlier units. It is important that students have seen randomisation, completely randomised designs and randomised block designs, in order to make the most of the SEC in this unit.

Although the examples provided in this unit work through the calculations from scratch, exam questions may involve a printout of summary statistics for students to work with. Ensure students see a mixture of both. Sometimes the ANOVA tables will be given and sometimes students will be required to make one themselves. Students will benefit from lots of practice here. It is also a very important topic due to its prevalence in modern day scientific, medical and clinical research. Students can hopefully, by the end of the unit, appreciate that multi-factor (three or more) ANOVA is possible, but it is outside the realms of this course.

The F-distribution makes its debut in this unit, unless the extension activity (mentioned in the notes of Unit 21b) of a hypothesis test for the difference between two variances has been taught. Use the F-distribution activity on Desmos to help illustrate how similar to the χ^2 distribution it is. It is also possible to easily calculate an effect size measure, η^2, from ANOVA. Although not on the specification, it does act as an introduction into the final unit.

The linear additive models of ANOVA are touched upon briefly. Students do not need to have in-depth knowledge of these models, nor do they need to know their relationship with regression. However, they do need to be aware of how the model works: the one factor ANOVA activity in Desmos (for 3 samples) will aid students in visualising the model.

The SEC can, as usual, be embedded by providing a synoptic element to questions:

A1 Identifying treatments, control groups, experimental groups etc. related to the factor or factors under investigation.

A2 Defining the null and alternative hypotheses in relation to the context.

A3 Describing a data collection method, justifying the choice of method (referring to advantages, practicality and assumptions).

A4 Using exploratory data analysis to determine if the underlying populations satisfies the assumptions (observations are normal distributed with equal variance).

A5 Understanding that the observations will be compared through one- or two-factor ANOVA, justifying the choice by making reference to the assumptions and advantages.

A6 Referring to experimental design concepts to reduce bias (completely randomised designs) or experimental error (randomised block designs).

B1 Understanding the practicalities of taking a large random sample.

B2 Using secondary source data and understanding the dangers of reporting findings (e.g. using a secondary source which is later retracted due to malpractice).

B3 Declaring and recording the methods described in **A3** and appreciating that the findings may not be accepted by the wider community.

B4 Showing an awareness of the need for randomisation to reduce bias and blind/double blind trials to reduce the influence of expectations of by patients and researchers.

C1 Calculating numerical measures from collected data, using technology.

C2 Calculating numerical measures from summary statistics generated by appropriate software.

C3 Appreciating that a lack of consideration of the sources of experimental error and bias can lead to misrepresentation.

D1 Analysing numerical measures from the samples obtained.

D2 Interpreting the results of a hypothesis test relating to the context.

D3 Using an appropriate hypothesis test and determining whether a result is statistically significant.

D4 Referring to the data collection method, the underlying population and the external factors that may affect the dependent variable, and discussing the effect on the findings.

D5 Reaching conclusions in context, using language appropriate to a given target audience.

E1 Identifying whether a sampling method described is suitable for the investigation, referring to the assumptions.

E2 Appreciating that a conclusion may not be valid if the sample was not obtained randomly, or bias was not minimised through poor experimental design.

E3 Suggesting improvements to the process, referring to the underlying population, sampling technique or blocking factor.

E4 Changing the blocking factor to elicit a more robust test.

24a. Analysis of Variance (ANOVA): Conditions for Analysis of Variance (20.3)

Teaching time
1 hour

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand when ANOVA can be used
- Have a basic idea of the theory behind ANOVA analysis

TEACHING POINTS

Revise the unpaired *t*-test from Unit 21b and the Wilcoxon rank-sum test from Unit 14c.

Analysis of Variance (commonly abbreviated to ANOVA) is a very common test in the world of applied statistics. Despite the name, the test itself is to determine whether or not a set of *n* independent normally distributed populations have equal means or not. One-factor (or one-way) ANOVA is an extension of the unpaired *t*-test from Unit 21b which tested whether or not a set of two independent normally distributed populations had equal means.

One of the conditions (as in Unit 21b) is that the set of n independent, normally distributed populations would have an unknown but common variance σ^2. Convey to students that one way of doing this without ANOVA is by using the unpaired *t*-test multiple times (the number of instances does not increase linearly, i.e. for a set of 3 populations, 3 unpaired *t*-tests must be carried out; for a set of 4 populations, 6 unpaired *t*-tests must be carried out). Since each hypothesis test comes with a probability of making a Type I error (the significance level), this error is compounded with each *t*-test making this method unreliable. The advantage of ANOVA is that all of the data are used in a single test (meaning better estimates of population parameters), it has more power than multiple t-tests, and it is easy to identify effects due to different factors.

ANOVA is best taught from a greatly simplified example: when the sample sizes are all the same. For instance:

It is suspected that the mean lifetimes of three brands of 60 watt lightbulbs are not the same. A sample of five lightbulbs is taken from the three brands:

Make	Lifetime above 1000 hours					Sample mean	Sample variance
A	16	15	13	21	15	16	9
B	18	22	20	16	24	20	10
C	26	31	24	30	24	27	11

Each sample gives an unbiased estimate of the population variance, but this is variance within each sample. From Unit 21b, the sample variances can be pooled together using s_p^2 (the formula for two variances is in the formula book).

The extension of s_p^2 is the sum of $(n_i - 1)s_i^2$ for each sample, divided by the sum of the sample sizes minus the number of samples. So for three samples, $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2+(n_3-1)s_3^2}{n_1+n_2+n_3-3}$ (compare this to the formula in the formula book).

In the example above, $s_p^2 = \frac{(5-1)\times 9+(5-1)\times 10+(5-1)\times 11}{5+5+5-3} = 10$. This is the "within samples" estimate of the variance, and call it s_w^2.

Recall from Unit 9a the sampling distribution of the mean, $\bar{X}$. We are assuming that the variances between the populations are equal. If the means are also equal, then these samples must originate from the same sampling distribution $\mathrm{N}\left(\mu, \frac{\sigma^2}{5}\right)$. So considering the sample means 16, 20 and 27 as a sample of size 3, this sample variance is 31, which would be an unbiased estimate for $\frac{\sigma^2}{5}$. So σ^2 can be estimated by $31 \times 5 = 155$. This is the "between samples" estimate of the variance and call it s_b^2.

If the means and the variances of all the populations are the same, then s_w^2 and s_b^2 should be estimates of the same variance σ^2. However, if the means are not the same, then s_b^2 will be an estimate of σ^2 plus some variability between the sample means (so $s_b^2 > s_w^2$).

This is why this method is called "analysis of variance", because we are analysing the variance to determine a result about the mean. Another way of explaining this would be: under the assumption that the means are the same, the variance estimates would be the same. So if the variance estimates are not the same, then means cannot be the same. In this example, the variance estimates are 10 and 155, which suggests that the means are not the same.

In the above example, if all three samples were combined to make one single sample of size 15, the combined sample variance would come out at $\frac{430}{14}$. Recall the formula for sample variance (in the formula book), so $\sum(x - \bar{x})^2 = (n-1) \times s_x^2$. Take the three sample variances calculated: $s_w^2 = 10$, $s_b^2 = 155$ and $s_t^2 = \frac{430}{14}$ (the total). By multiplying by the number of degrees of freedom in each case (for s_w^2 it was 12, for s_b^2 it was 2 and for s_t^2 it was 14) we would see $12 \times s_w^2 = 120$, $2 \times s_b^2 = 310$ and $14 \times s_t^2 = 430$, and students should be able to see that $120 + 310 = 430$. Emphasise to students that this always holds.

Introduce the case when the samples do not have the same size:

It is suspected that the mean lifetimes of three brands of 60 watt lightbulbs are not the same. Samples of lightbulbs are taken from the three brands:

Make	Lifetime above 1000 hours					Sample mean	Sample variance
A	15	15	13	21		16	12
B	18	22	20	16	24	20	10
C	26	31	24			27	13

Following the same procedure as before, we can calculate the "within samples" estimate of the variance:

$$s_w^2 = \frac{(4-1)\times 12 + (5-1)\times 10 + (3-1)\times 13}{4+5+3-3} = \frac{34}{3} = 11.33,$$

which has $4+5+3-3=9$ degrees of freedom (the sum of the sample sizes minus the number of samples).

However, students may notice now that the "between treatments" estimate is not mentioned since the sample sizes are different. However, we can find it by using the "total" property earlier. By combining all samples together to make a "total" sample, we can calculate the "total" variance $s_t^2 = \frac{3731}{132}$, which has $12-1=11$ degrees of freedom. Using the "total" property, we know that $9s_w^2 + 2s_b^2 = 11s_t^2$ (the two because it is one fewer than the number of samples). So $11s_t^2 - 9s_b^2 = 2s_b^2 = \frac{2507}{12}$. So $s_b^2 = \frac{2507}{24} = 104.5$ is the "between samples" estimate of the variance.

Although this method can be used to carry out ANOVA, none of these formulae are in the formula book. This method only illustrates how ANOVA works and students will gain a better appreciation of the method if they see it in action.
The method that is **best taught in practice** is detailed in the following sub-unit (ANOVA tables).

OPPORTUNITIES FOR EMBEDDING THE SEC

In addition to the unit summary, referring to the assumptions of independent, normal distributions with equal variances will add to the justification of a proposed method of processing data, as well as the discussion of the reliability of results and the evaluation of the process.

COMMON AND POSSIBLE MISTAKES

The majority of the mistakes will occur in the following sub-unit. The conditions of normality, independence and equal variance often get forgotten.

NOTES

If the normality assumption is violated, then the Kruskal-Wallis test is a non-parametric version which tests for the equality of medians. This is mentioned in the notes of Unit 13c.

The method detailed in this sub-section is a valid one. However, it is purely used as an illustration of how ANOVA works, and the method detailed in the following sub-sections can be described as equivalent to the one in this sub-section.

24b. Analysis of Variance (ANOVA): One-factor Analysis of Variance (20.1)

Teaching time
3 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand the term: completely randomised design.
- Appreciate how completely randomised design can be used in ANOVA.
- Complete a one-factor ANOVA table.
- Carry out a test for equality of means using ANOVA.
- Interpret the results of the hypothesis test in context.
- Understand and use the linear model for one-factor ANOVA.

TEACHING POINTS

Remind students that bias can be introduced from non-random sampling methods, and experimental error can be introduced from a poor experimental design, both of which can affect the reliability of conclusions.

Revise randomisation (Unit 23a). If there is more than one treatment (e.g. different concentrations of a drug to be administered to patients) then a completely randomised design involves each treatment being allocated at random to the participants. This is a good example for where ANOVA can be used (and embeds the SEC).

Recap the ideas of the "within samples" estimate of variance s_w^2, "between samples" estimate of variance s_b^2 and the "total" estimate of variance s_t^2. The formula for s_t^2 is $\frac{SS_T}{\nu_T}$ where SS_T is the total sum of squares:

$$SS_T = \sum_i \sum_j x_{ij}^2 - \frac{T^2}{n},$$

where $T = \sum_i \sum_j x_{ij}$ is the total sum of all observations and n is the total number of observations, and $\nu_T = n - 1$ is the number of degrees of freedom. The x_{ij} refers to the observation in the i^{th} row and the j^{th} column. The formula for SS_T is in the formula book, but the meanings of T and n are not. The formula for s_b^2 is $\frac{SS_B}{\nu_B}$ where SS_B is the "between groups" sum of squares:

$$SS_B = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{n},$$

Where $T = \sum_i \sum_j x_{ij}$ is the total sum of all observations, n is the total number of observations, T_i is the total sum of observations within a sample and n_i is the size of that sample. ν_B is one less than the number of samples. The formula for SS_B is in the formula book, but the definitions of T, T_i, n and n_i are not.

These are the only two formulae relating to a one-factor ANOVA table listed in the formula book.

From the previous subsection, SS_W is the "within groups" sum of squares and is calculated by $SS_T - SS_B$. The formula for $s_w^2 = \frac{SS_W}{\nu_W}$ where $\nu_W = \nu_T - \nu_B$ (students could check their answers by calculating the pooled variance of samples).

The value s_b^2 is also called the "mean square between groups" and is denoted by MS_B. The value s_w^2 is also called the "mean square within groups" or "residuals mean square" and is denoted by either MS_W or MS_E.

An ANOVA table looks like the following:

Source of Variation	**Sum of Squares (SS)**	**Degrees of Freedom (ν)**	**Mean Square (MS)**	F
Between groups				
Within Groups / Error				
Total				

The column of "Mean square" are the same values as calculated in the previous sub-unit for comparison.

Allow students to practise filling in the ANOVA table (without the F column). This will take a lot of time due to the complexities of the formulae, so plenty of worked examples and continuous checking is required.

Exemplar

It is suspected that the mean lifetimes of three brands of 60 watt lightbulbs are not the same. Samples of lightbulbs are taken from the three brands:

Brand	Lifetime above 1000 hours				
A	15	15	13	21	
B	18	22	20	16	24
C	26	31	24		

Complete the ANOVA table below.

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (ν)	Mean Square (MS)
Between brands			
Within brands			
Total			

Using the formula book: $SS_T = \sum_i \sum_j x_{ij}^2 - \frac{T^2}{n}$.

T *is the total sum of observations:* $T = 245$.
n *is the total number of observations:* $n = 12$.

The total sum of squares is $\sum_i \sum_j x_{ij}^2 = 5313$.

So $SS_T = 5313 - \frac{245^2}{12} = 310.917$.

Using the formula book: $SS_B = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{n}$.

$T_1 = 64$, $n_1 = 4$, $T_2 = 100$, $n_2 = 5$ *and* $T_3 = 81$, $n_3 = 3$.

So $\sum_i \frac{T^2}{n_i} = \frac{64^2}{4} + \frac{100^2}{5} + \frac{81^2}{3} = 5211$.

Hence $SS_B = 5211 - \frac{245^2}{12} = 208.917$.

$SS_W = SS_T - SS_B = 310.917 - 208.917 = 102$.

$\nu_B = 2$, $\nu_T = 11$ *so* $\nu_W = \nu_T - \nu_B = 11 - 2 = 9$.

Source of Variation	*Sum of Squares (SS)*	*Degrees of Freedom* (ν)	*Mean Square (MS)*
Between brands	208.917	2	104.458
Within brands	102	9	11.3333
Total	310.917	11	

This example is precisely the same example as in the previous sub-unit, and students may notice that the answers are the same. Note that some calculators can generate the above ANOVA table by inputting the data provided. Students who utilise this may still access full marks (in line with the mark scheme), although exam questions may require manual calculation if, for example, summary data are presented.

The F column is the ratio between the mean squares, $\frac{MS_B}{MS_W}$. As mentioned in the previous sub-unit, $MS_B > MS_W$ so ensure that MS_B is in the numerator. If the variances are equal then $F = 1$. The statistic $\frac{MS_B}{MS_W}$ follows an F-distribution, which is a ratio of two χ^2 distributions. Use the F-distribution activity on Desmos to illustrate to students how it similar to a χ^2 distribution. However, since there are two variables involved, there are two different values for the number of degrees of freedom. Show students the table of values from the F-distribution, highlighting that the number of degrees of freedom are fall into "numerator" and "denominator" categories.

For the above example, the question may continue:
Carry out a hypothesis test at the 5% significance level to investigate whether the mean lifetime of a bulb is different for the 3 brands.

The hypothesis test itself should be straightforward now.
The null hypothesis is $H_0: \mu_1 = \mu_2 = \cdots = \mu_n$ or $H_0: \mu_i = \mu$ for $i = 1,2, \ldots, n$
The alternative hypothesis is H_1: **at least two** means differ (from each other)
The alternative hypothesis may also be stated using appropriate notation as
$H_1: \mu_i \neq \mu_j$,for some i and j
It is not correct to state
$H_1: \mu_i \neq \mu$, for some i
since this is stating, effectively, that only one of the population means differs.
It is important that students appreciate that the test cannot pinpoint which population mean is different, it can only detect that there is a significant difference **between two of them** (largest and smallest means).
Variables and subscripts for each group must be clearly defined in the context of the question.

In 'real life' analysis, a 'post hoc' (after the event) additional test, for example **Tukey's** honestly significant difference (HSD) test, may be carried out in order to identify all significant differences between means. This is **not** in the specification.

In ANOVA, the F-test is always one-tailed and upper-tailed (like the χ^2 test) and the significance levels are clearly labelled in the tables.

The test statistic is the F value calculated in the ANOVA table.

To answer the example:

Exemplar

Carry out a hypothesis test at the 5% significance level to investigate whether the mean lifetime of a bulb is different for the 3 brands.

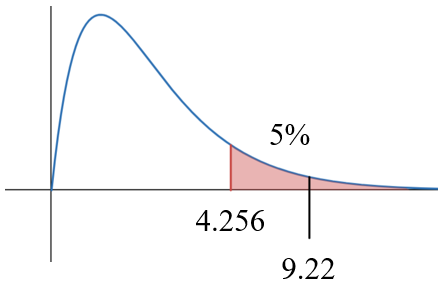
Let A be the lifetime of a bulb for lightbulb brand A, let B be that for brand B and let C be that for brand C.

$H_0: \mu_A = \mu_B = \mu_C$,
H_1: *At least two of the brands have a mean lifetime that differ from each other.*

We will carry out a one-factor ANOVA at the 5% significance level, using 2 *degrees of freedom in the numerator and 9 degrees of freedom in the denominator.*

$F = \frac{104.5}{11.33} = 9.22$ *is the test statistic.*

From the tables, the critical region is $F \geq 4.256$.



Since $9.22 \geq 4.256$, *the result is significant. We reject* H_0, *as there is significant evidence to suggest that at least two of the brands of lightbulb have different mean lifetimes.*

The linear one-factor model of ANOVA is $x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ where $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$. This is in the formula book. What isn't mentioned in the formula book is what the symbols mean: x_{ij} is the observation in the i^{th} row, j^{th} column; μ is the overall mean; α_i is the effect due to the i^{th} group; ε_{ij} is the inherent residual random variation. μ can be estimated by the grand mean $\left(\frac{\sum_i \sum_j x_{ij}}{n}\right)$, and α_i can be estimated by the difference between the i^{th} group sample mean and the grand mean. Simple rearranging of the equation enables the inherent residual random variations to be estimated. Questions could include:

Exemplar

It is suspected that the mean lifetimes of three brands of 60 watt lightbulbs are not the same. Samples of lightbulbs are taken from the three brands:

Make	Lifetime above 1000 hours				
A	15	15	13	21	
B	18	22	20	16	24
C	26	31	24		

A model for this design is $x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ where $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ where σ^2 is the common variance between all three brands of lightbulb. Estimate the overall mean μ, the effect due to Brand B of lightbulb α_B, and the residual error for the sample unit with a lifetime of 1020 hours, ε_{B2}.

The grand mean is $\frac{245}{12}$*, so* $\mu = 20.417$.

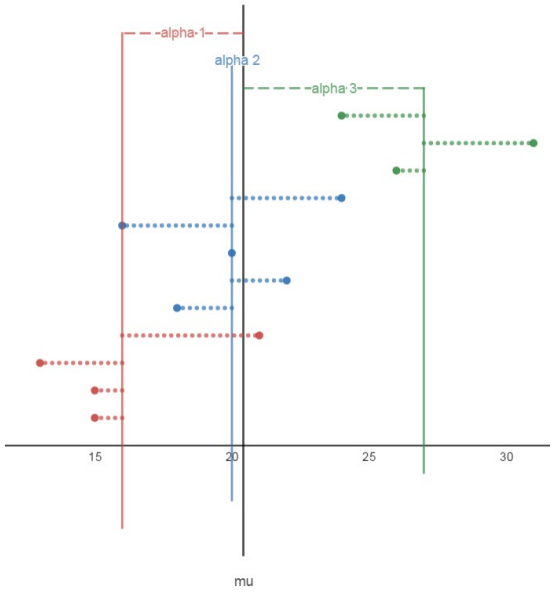
The mean of Brand B is $\frac{100}{5} = 20$. *So* $\alpha_B = 20 - 20.417 = -0.417$.

Finally, $22 - 20.417 + 0.417 = \varepsilon_{B2} = 2$.

This example determines the effects due to the make of lightbulb, together with the random variation of the bulb within the brand. In the above example, μ may be interpreted as "the average lifetime of a bulb above 1000 hours", without any reference to any particular group. α_B can be interpreted as "the effect of being a Brand B bulb". In the above example, the average lifetime of a Brand B bulb is -0.417 hours compared with the average bulb, so the effect of being a Brand B bulb is that they are expected to last 0.417 hours less than the average bulb.

ε_{B2} can be interpreted as the "inherent random error" of the second Brand B bulb in the sample. We are expecting a Brand B bulb to last 20 hours above 1000 hours, but this particular bulb lasted 2 hours longer.

Although the ANOVA test does not pinpoint the population mean which is different from the others, you can use the linear model to estimate and quantify the effects of the different groups. Students can decide which populations seem to have the highest and lowest means, but cannot comment on the in-between ones. Use the ANOVA Linear Model activity on Desmos to illustrate this model.



OPPORTUNITIES FOR EMBEDDING THE SEC

See the unit summary. References to completely randomised design and describing methods for how to reduce bias will help with **Stage A**. The use of the one-factor linear model will help students interpret the results of their hypothesis test in **Stage D**.

COMMON AND POSSIBLE MISTAKES

- Using a two-factor ANOVA instead of a one-factor ANOVA.
- Concluding that the population means (all) differ rather than there is a significant difference between at least two population means in the case that H_0 is rejected.

Students must remember to state the degrees of freedom.

NOTES

Although not on the specification, an effect size measure, $\eta^2 = \frac{SS_B}{SS_T}$ is easily calculated here. See the notes of Unit 25.

24c. Analysis of Variance (ANOVA): Two-factor Analysis of Variance (20.2)	Teaching time 2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand the terms: blocking, randomised block design.
- Appreciate that blocking can reduce experimental error.
- Complete a two-factor ANOVA table.
- Carry out a test for equality of means using two-factor ANOVA.
- Interpret the results of the hypothesis test in context.
- Understand and use the linear two-factor model for ANOVA.
- Appreciate how randomised block design can be used in ANOVA.

TEACHING POINTS

Students need to be aware that it is not always possible to eliminate variation arising from external factors from statistical experiments. For example, if the effect of an experimental drug on the white blood cell count of patients is being investigated, then the age/race/gender of the patient may have an effect on the dependent variable. It is not possible to remove these effects. Blocking is a method for dealing with such effects. Members of the experimental group are “blocked” together by an external factor in common (e.g. race, hair colour). The randomisation of treatments within each block is called a randomised block design.

The use of blocking ensures that if there is a difference between population means for the different levels of the factor of interest, such a difference is more likely to be detected.

Exemplar

In a comparison of the efficacy of three drugs, A, B and C, four people (Ranjit, Sally, Tallulah and Usain) are to be treated using these three drugs. After a week, samples of blood will be taken from these four people to be analysed. Three experimental designs are suggested:

Design 1		
A	B	C
Ranjit	Sally	Tallulah
Ranjit	Sally	Tallulah
Ranjit	Sally	Tallulah

Design 2		
A	B	C
Ranjit	Sally	Ranjit
Sally	Tallulah	Tallulah
Tallulah	Ranjit	Sally
Usain	Usain	Usain
Ranjit	Sally	Tallulah

Design 3		
A	B	C
Ranjit	Sally	Tallulah
Tallulah	Usain	Sally
Usain	Ranjit	Usain
Sally	Tallulah	Ranjit

For example, the first column of Design 2 says that drug A will be used to treat Ranjit for two weeks, and Sally, Tallulah and Usain for one week only. The written order within any column has no relevance.

a) **State two disadvantages of Design 1.**
Each drug is only used on one person, so the overall efficacy of the drug cannot be fairly judged. Usain does not receive a treatment and therefore is aware he is in the control group, influencing the clinical outcome through his expectation of no improvement. Any observed differences in outcome may be a difference between patients, not between treatments.

b) **Write down the name of Design 3.**
Randomised Block Design

c) **State one advantage of Design 3 over Design 2.**
Each person is treated by each drug in Design 3. In Design 2, the people are treated more than once by the same drug, but it is not the same drug for each person.

Revise the paired *t*-test from Unit 23b. Just as one-factor ANOVA could be thought of as an extension of the unpaired *t*-test, a two-factor ANOVA could be thought of as an extension of the paired *t*-test.

The assumptions for two-factor ANOVA are: the populations from which the samples are obtained are (approximately) normally distributed, there must be no interaction between the factors, the variances of all populations are equal. For two-factor ANOVA (just as in the paired *t*-test), the groups must have the same sample size.

Start with an example:

A pharmaceutical product is manufactured by a fermentation process. An experiment was run to compare three similar chemical salts, 1, 2 and 3, in the manufacture of the pharmaceutical product. There were four types of fermenter, A, B, C and D, available for use in the manufacture. Three fermentations were started in each type of fermenter, one containing salt 1, another salt 2 and the third salt 3. After two days, the same volume was taken from each fermenter and analysed. The results for amount of fermentation were as follows:

		Fermenter Type			
		A	**B**	**C**	**D**
Salt	**1**	**67**	**69**	**72**	**68**
	2	**78**	**73**	**80**	**69**
	3	**68**	**65**	**73**	**70**

This can demonstrate how blocking can be used in ANOVA: the fermenter type is the blocking factor and the amount of fermentation is the variable to be measured. The type of chemical salt is the factor of interest. A two-factor ANOVA table looks like:

Source of Variation	**Sum of Squares (SS)**	**Degrees of freedom (ν)**	**Mean square (MS)**	F
Between Rows	SS_R	ν_R	MS_R	$\frac{MS_R}{MS_E}$
Between Columns	SS_C	ν_C	MS_C	$\frac{MS_C}{MS_E}$
Error	SS_E	ν_E	MS_E	
Total	SS_T	ν_T		

It looks very similar to the one-factor ANOVA table, except there is an extra row for the second factor.

The total sum of squares $SS_T = \sum_i \sum_j x_{ij}^2 - \frac{T^2}{mn}$ is in the formula book. Here, x_{ij} is the observation in the i^{th} row and j^{th} column; T is the total sum of all observations $\left(\sum_i \sum_j x_{ij}\right)$; m is the number of rows and n is the number of columns. The definition of the symbols are not included in the formula book.

The "between rows" sum of squares $SS_R = \sum_i \frac{R_i^2}{n} - \frac{T^2}{mn}$ is in the formula book. Here, R_i is the i^{th} row total. Similarly, the "between columns" sum of squares $SS_C = \sum_j \frac{C_j^2}{m} - \frac{T^2}{mn}$ is in the formula book. Here, C_j is the j^{th} column total. Neither R_i or C_j are defined in the formula book.

Everything else is not in the formula book: The error sum of squares is $SS_T - SS_R - SS_C = SS_E$. The numbers of degrees of freedom are $\nu_R = m - 1$ (one fewer than the number of rows), $\nu_C = n - 1$ (one fewer than the number of columns), $\nu_T = mn - 1$ (one

fewer than the total number of observations) and $\nu_E = \nu_T - \nu_R$. The mean square values are the sum of squares divided by the corresponding number of degrees of freedom.

To complete the question for the above example:

Exemplar

Test, at the 5% level of significance, the hypothesis that the type of salt does not affect the mean amount of fermentation.

		Fermenter Type				
		A	***B***	***C***	***D***	***Row Total***
Salt	***1***	*67*	*69*	*72*	*68*	***276***
	2	*78*	*73*	*80*	*69*	***300***
	3	*68*	*65*	*73*	*70*	***276***
	Column Total	***213***	***207***	***225***	***207***	***852***

Using the calculator, $\sum\sum x^2 = 60710$.

$$SS_T = 60710 - \frac{852^2}{12} = 218$$
$$SS_R = \frac{276^2}{4} + \frac{300^2}{4} + \frac{276^2}{4} - \frac{852^2}{12} = 96$$
$$SS_C = \frac{213^2}{3} + \frac{207^2}{3} + \frac{225^2}{3} + \frac{207^2}{3} - \frac{852^2}{12} = 72$$

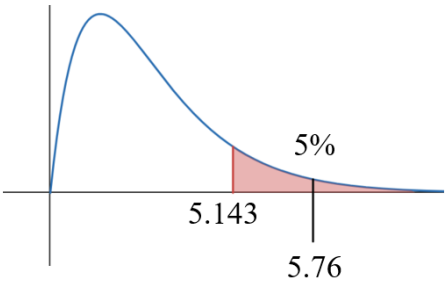
The two-factor ANOVA table looks like:

Source of Variation	***Sum of Squares*** (SS)	***Degrees of freedom*** (ν)	***Mean square*** (MS)	F
Between Salts	*96*	*2*	*48*	*5.76*
Between Fermenters	*72*	*3*	*24*	*2.88*
Error	*50*	*6*	*8.33*	
Total	*218*	*11*		

$H_0: \mu_1 = \mu_2 = \mu_3$,
H_1: ***at least two*** *of the population means of the amount of fermentation differ from each other,*
where μ_i *is the population mean of the amount of fermentation due to salt* i.

We will carry out a two-factor ANOVA between salts at the 5% significance level, using 2 degrees of freedom in the numerator and 6 degrees of freedom in the denominator.

From the tables, the critical region is $F \geq 5.143$. *The test statistic is* 5.76.



Since $5.76 \geq 5.143$*, the result is significant. We reject* H_0*, so there is significant evidence to suggest that the type of salt affects the mean amount of fermentation.*

In examples such as these, students can also determine how effective the blocking factor is:

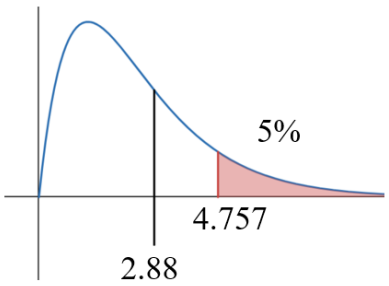
Exemplar

Determine whether the fermenter type is an effective blocking factor.

H_0: $\mu_A = \mu_B = \mu_C = \mu_D$,
H_1: ***at least two*** *of the population means of the amount of fermentation differ from each other.*

Conduct a two-factor ANOVA between fermenters at the 5% significance level, using 3 degrees of freedom in the numerator and 6 degrees of freedom in the denominator.

From the tables, the critical region is $F \geq 4.757$*. The test statistic is* 2.88*.*



Since $2.88 \leq 4.757$*, the result is not significant. We do not reject* H_0*, so there is insufficient evidence to suggest that the fermenter type affects the amount of fermentation. This suggests that the fermenter type is not an effective blocking factor.*

If there were significant differences between the fermenters, the differences in "fermenter effect" would introduce further variability in the amount of fermentation. Unless "choice of fermenter" is used as a blocking factor, this extra variability is added to the residual sum of squares. When there is a large amount of residual variation, any difference between the salts is less likely to be detected.

In the example above, however, there was no significant difference between the effects of the different fermenters, so the conclusion was not affected by whether the blocking factor was used or not.

The linear two-factor model for ANOVA is $x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ where $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$. This is listed in the formula book, but the definitions of the symbols are not. x_{ij}, μ and ε_{ij} are as in Unit 24b. This time α_i is the effect due to the row factor and β_j is the effect due to the column factor. As before, these can be estimated by the difference between the row/column means and the grand mean. Students can then decide which populations seem to have the highest and lowest means, but cannot comment on the in-between ones.

OPPORTUNITIES FOR EMBEDDING THE SEC

See the unit summary. In addition to this, references to the blocking factor and whether or not the blocking factor was effective or not can emphasise **Stage E** with analysis from **Stage D**.

COMMON AND POSSIBLE MISTAKES

- Dividing by the number of rows instead of the number of columns when calculating SS_R, and similarly dividing by the number of columns instead of the number of rows when calculating SS_C.
- Not labelling ANOVA tables in context, for example "between salts" and "between fermenters", which results in getting the degrees of freedom mixed up.
- Forgetting to state the degrees of freedom.

Students must remember to clearly define variables and subscripts for each factor level.

NOTES

It is worth noting that the assumption that "there is no interaction between the factors" is not strictly necessary for a two-way ANOVA. Students will only be assessed on this the special case where there is no interaction between factors. A full two-way ANOVA with replicates does not require this assumption, but to carry it out by hand would be too time consuming for an examination and, as such, is not on the course.

UNIT 25: Effect Size

Return to overview

SPECIFICATION REFERENCES

21.1 Know the notion of effect size as a complementary methodology to standard significance testing, and apply in authentic contexts

21.2 Know and use Cohen's *d* in simple situations

PRIOR KNOWLEDGE

Year 2 of A Level Statistics
Confidence Intervals (Unit 18)
Pooled Variance (Unit 21)
Hypothesis Testing

KEYWORDS

Cohen's *d*, confidence intervals, effect size, hypothesis testing, *p*-values, result, sample size, significance,

UNIT SUMMARY

The specification only requires students to be aware of the concept of effect size and carry out a rudimentary calculation to interpret effect size. It does not, however, do the topic justice and it must be emphasised to students that effect size can be misinterpreted and misrepresented if it is not understood properly.
Students will be required to calculate Cohen's *d* in simple cases and interpret the size of the effect and compare the information that the Cohen's *d* provides with that provided from the outcome of a *t*-test .

Students may also appreciate that usually (although not on the specification), Cohen's *d* is reported as a confidence interval in real world statistical publications.

25. Effect Size: Cohen's *d* (21.1)

Teaching time
2 hours

OBJECTIVES

By the end of the sub-unit, students should be able to:

- Understand the difference between statistical significance and effect size
- Appreciate that a *p*-value or the comparison between a test statistic and a critical region is not an indication of effect size
- Calculate and interpret the value of Cohen's *d*

TEACHING POINTS

Revise hypothesis testing using a *p*-value. One example:

The plums from a particular variety of plum tree have masses which can be modelled by a normal distribution with standard deviation 5 g. It is claimed that the mean mass of a plum is 24 g, but there have been reports that it could be higher.

a) 20 plums are selected at random and the mean mass of a plum from the sample is 28 g. Test at the 5% significance level the claim that the mean mass of a plum is 24 g.
b) This time 40 plums are selected at random and the mean mass of a plus from this sample is also 28 g. Test at the 5% significance level the claim that the mean mass of a plum is 24 g.

This is an example in Unit 9 – the *p*-value for part (a) is 0.000347 and the *p*-value for part (b) is 4.2×10^{-7}. For comparison, the critical region for part (a) is $X \geq 25.8$ and the critical region for part (b) is $X \geq 25.3$. Clearly this results in two significant results, leading to the rejection of the null hypothesis. However, the increase in sample size has clearly changed the *p* value. In both cases, the *p*-values are wildly different but the difference between the critical value and the test statistic is not. This example highlights that a hypothesis test does not measure the size of an effect.

Students need to be aware that a hypothesis test only detects if there is an effect (statistical significance) and possibly the direction of the effect (e.g. greater than, less than) but does not give any indication over how large the effect truly is (practical significance). As seen earlier, a larger sample size can result in a significant *p*-value (and a very low one). On the other hand, the non-rejection of the null hypothesis can occur more easily with a hypothesis test with low power (which is also dependent on sample size).

Effect size is not be directly influenced by sample size. Any change in the sample size does not directly translate to any meaningful change in the effect size. There are many measures of effect size (some people refer to them as measures of association), but only one will be assessed in the A level.

Cohen's *d* is used for measuring the effect size between two means and only for normally distributed variables. The formula for Cohen's *d* is given in the formula book:

$$d = \frac{\overline{x_1} - \overline{x_2}}{s} \text{ where } s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

where $\overline{x_i}$ and s_i^2 are the mean and variance taken from a sample of size n_i. Students may even recognise this as the standardised test statistic from Unit 21b, where the difference between two means is hypothesised to be zero and the sample sizes $\left(\frac{1}{n_x} + \frac{1}{n_y}\right)$ omitted. The value of *d* indicates the distance (measured in standard deviations) between the two means. As a guide, the specification lists the following:

- $0.2 < |d| \leq 0.5$ represents a small effect,
- $0.5 < |d| \leq 0.8$ represents a medium effect,
- $|d| > 0.8$ represents a large effect.

Note that these are only guidelines and statisticians should exercise caution when interpreting the results (as stated by Cohen himself).

Exemplar

A lack of vegetable intake in food consumption is suspected of retarding the growth of muscle mass in athletes.

The following data are the results of an experiment to measure the percentage gain in muscle mass in developing athletes given either a balanced diet (A) or a diet with no vegetables (B).

Diet A	**18.2**	**25.8**	**16.8**	**14.9**	**19.6**	**26.5**	**17.5**			
Diet B	**13.4**	**18.8**	**20.5**	**7.5**	**22.2**	**15.0**	**12.2**	**14.3**	**18.0**	**15.1**

Assuming that the percentages for both Diet A and Diet B are normally distributed, calculate Cohen's *d* and interpret what this means in context.

You may assume that the sample standard deviation for Diet A is 4.50 and the sample standard deviation for Diet B is 4.32, both correct to 3 significant figures.

The pooled sample variance is $s_p^2 = \frac{(7-1)\times 4.5^2 + (10-1)\times 4.32^2}{7+10-2} = 19.297$

The mean percentage increase from Diet A is 19.9 and the mean percentage increase from Diet B is 14.7. So

$$d = \frac{19.9 - 14.7}{\sqrt{19.297}} = 1.18.$$

This indicates a very large effect between the mean percentage increases between the two diets.

(note that the test statistic is $t = 1.94$ with t , $\upsilon = 15, 1\ tail$ critical value = 1.753 , so this leads to rejection of H_0 and the conclusion that Diet A leads to a greater muscle mass, on average)

In Unit 21a (where this example was last seen), it was discovered that the result was statistically significant (there was an effect). However, students can now also say that the result is practically significant (there was a large effect). Heavily emphasise, however, that effect size alone is not an effective reporting of results. The effect size should be accompanied with a rigorous hypothesis test.

It may be useful to distinguish each of the 4 possible cases:

Case 1 test result statistically significant with a large effect size;
Case 2 test result statistically significant with a small effect size;
Case 3 test result not statistically significant but a large effect size;
Case 4 test result not statistically significant but a small effect size.

Allow students to discuss the implications of each of these and to suggest ways to investigate further.

For example, a result which is not statistically significant, but has a large effect size may warrant further investigation with a larger study.

For example, a result which is statistically significant but has a small effect may not have any practical significance and care should be taken when interpreting the results.

It is worth noting that the sample size affects the *t*-test test statistic or *p*-value (larger samples more likely to determine a significant change), but does not affect Cohen's *d.*

OPPORTUNITIES FOR EMBEDDING THE SEC

As stated above, the effect size should be reported as a complementary result to a hypothesis test. It should also accompany a clear data collection methodology, design of experiment and evaluation of process. By this point in the course, students are advised to have practised many examples of a full SEC and this topic adds to **Stage D**.

COMMON AND POSSIBLE MISTAKES

- Confusing the Cohen's *d* thresholds with those of correlation.
- Interpreting Cohen's *d* in terms of correlation instead of effect.
- Confusing Cohen's *d* with a *p*-value.
- Using the standard deviations (as opposed to the variances) in the formula.

NOTES

Effect size is a relatively new topic in the world of statistics. However, it is rising to prominence in applied statistics to the point where journal editors will not accept a paper unless the effect size is reported. It is highly used in psychological and clinical science.

In practice, Cohen's *d* is usually reported in a confidence interval. This is not on the specification but students could be made aware of the current academic standard.

Other examples of effect size measures are the PMCC and SRCC. They provide a descriptive inferential interpretation about the population from the sample without being dependent on the sample size. However, unlike in the PMCC/Spearman's rank tests where the test statistics are also effect size measures, this is not true for tests about the mean.

Another effect size measure which is easy to calculate, but **not** on the specification, is the η^2 effect size measure for ANOVA. It is calculated as $\eta^2 = \frac{SS_B}{SS_T}$.

UNIT 26. The Statistical Project (SEC)

Return to overview

SUMMARY

Students completing a full run of the SEC and putting their A-Level knowledge into practice will consolidate and reinforce the importance of statistical techniques.

Using the SEC as a guide, in Stage A students can plan investigations by deciding what to investigate, what data they would need to collect and what steps are needed to reduce bias and experimental error. Ideally, students would have free reign over what area they wish to study and students with a particular interest in certain fields (e.g. geography, sport, computers, psychology etc) are given a chance to apply their knowledge to something they genuinely want to investigate.

In Stage B, students would collect their own data, either through primary or secondary (or both) methods and report their data collection methodologies. This also allows for an opportunity to experience first-hand the constraints involved when designing primary data collection methods or the reliability of secondary data.

In Stage C, students can use technology to organise and process their data. Discourage the use of organising data “by-hand” and encourage the use of software and technology. Spreadsheet and database software would be more fitting in with the specification but there is no reason why students cannot use other statistical software such as R, SPSS, Minitab etc. Students could also be able to select appropriate data visualisations for their data, and learn how to generate these visualisations using technology.

In Stage D, inferential statistical techniques (confidence intervals, hypothesis testing) could be used to investigate the question at hand. This may be limited to the amount of content currently taught at the time when this project is conducted. Communication of their findings to a target audience (for example, their peers who are also taking the qualification) is also a part of this stage. This may be via short presentations, academic posters or research reports.

Finally, in Stage E students could reflect on their project and examine the strengths and weaknesses in their approach, suggesting improvements and refinements (without doing them).

Anecdotally, the best time to conduct this project would be at the end of Year 12 (or their first year of study) since it allows students to consolidate their learning from that first year in a single project. Depending on the mode of communication of findings, employment skills such as verbal or visual presentation are developed.

The SEC may be used as a guide for assessment, should you wish this project to be assessed (using the points of the SEC as a list of objectives or criteria).

Appendix: Activities on Desmos

Return to overview

Activities on Desmos (www.desmos.com) have been created to aid teachers and students in exploring many of the probability distributions and statistical diagrams. These may be used freely.

- Bar Charts
- Dual Bar Charts
- Stacked Bar Charts
- Vertical Line Charts
- Pie Charts
- Histograms
- Binomial Distribution with Normal Approximation
- Scatter Graphs and regression lines, with residuals
- Normal Distribution
- Sampling Distributions
  To use when explaining the sampling distribution of the mean/variance/standard deviation
- *z* test with critical regions
  To use for hypothesis tests about the mean with known variance for large samples or an underlying Normal distribution
- Binomial inference
  To use for hypothesis tests about the proportion, or sign tests
- Chi-Squared Distribution
- Wilcoxon Signed-Rank with critical regions
- Poisson Distribution (with normal approximation)
- Confidence Intervals (Normal Distribution)
- Confidence Intervals (t distribution)
- t distribution with critical regions
  To use for hypothesis tests about the mean with unknown variance and an underlying normal distribution
- Type I and Type II errors
  To use when explaining Type I and Type II errors, and the power of a hypothesis test
- Exponential and Poisson Distributions
- Model Fitting
  To use for Goodness of fit
- F Distribution
- ANOVA Linear Model

Acknowledgments:

For more information on Edexcel and BTEC qualifications please visit our websites: www.edexcel.com and www.btec.co.uk