Course/Unit Support Notes

Advanced Higher Statistics Course/Unit Support Notes

Please refer to the note of changes at the end of this document for details of changes from previous version (where applicable).

August 2017, version 2.3

Contents

Introduction

These support notes are not mandatory. They provide advice and guidance on approaches to delivering and assessing the Advanced Higher Statistics Course. They are intended for teachers and lecturers who are delivering the Course and its Units.

These support notes cover both the Advanced Higher Course and the Units in it.

The Advanced Higher Course/Unit Support Notes should be read in conjunction with the relevant:

Mandatory Information:

- Course Specification
- Course Assessment Specification
- Unit Specifications

Assessment Support:

- Specimen and Exemplar Question Papers and Marking Instructions
- Exemplar Question Paper Guidance
- Guidance on the use of past paper questions
- Unit Assessment Support*

Related information

Advanced Higher Course Comparison

Further information on the Course/Units for Advanced Higher Statistics

This information begins on page 11 and both teachers and learners may find it helpful.

General guidance on the Course/Units

Aims

The aims of the Course are to enable learners to:

- understand the appropriateness of different methods of data collection, particularly ways of sampling from a population
- select and use appropriate statistical models to assist with the analysis of data
- consider and evaluate assumptions required for chosen models
- understand the notion of probability
- interpret results in context, evaluating the strength and limitations of their models
- develop skills in effectively communicating conclusions reached on the basis of statistical analysis

Progression

In order to do this Course, learners should have achieved the Higher Mathematics Course.

Learners who have achieved this Advanced Higher Course may progress to further study, employment and/or training. Opportunities for progression include:

- Progression to other SQA qualifications
  - Progression to other qualifications at the same level of the Course, eg Mathematics, or Mathematics of Mechanics, Professional Development Awards (PDAs) or Higher National Certificates (HNCs)
- Progression to further/higher education
  - For many learners a key transition point will be to further or higher education, for example to Higher National Certificates (HNCs) or Higher National Diplomas (HNDs) or degree programmes.
  - Advanced Higher Courses provide good preparation for learners progressing to further and higher education as learners doing Advanced Higher Courses must be able to work with more independence and less supervision. This eases their transition to further/higher education. Advanced Higher Courses may also allow 'advanced standing' or partial credit towards the first year of study of a degree programme.
  - Advanced Higher Courses are challenging and testing qualifications — learners who have achieved multiple Advanced Higher Courses are regarded as having a proven level of ability which attests to their readiness for education in higher education institutions (HEIs) in other parts of the UK as well as in Scotland.
- Progression to employment
  - For many learners progression will be directly to employment or work-based training programmes.

This Advanced Higher could be part of the Scottish Baccalaureate in Science. The Scottish Baccalaureates in Expressive Arts, Languages, Science and Social Sciences consist of coherent groups of subjects at Higher and Advanced Higher level. Each award consists of two Advanced Highers, one Higher and an Interdisciplinary Project, which adds breadth and value and helps learners to develop generic skills, attitudes and confidence that will help them make the transition into higher education or employment.

Hierarchies

Hierarchy is the term used to describe Courses and Units which form a structured progression involving two or more SCQF levels.

This Advanced Higher Course is not in a hierarchy with the Higher Mathematics Course or its Units.

Skills, knowledge and understanding covered in this Course

This section provides further advice and guidance about skills, knowledge and understanding that could be included in the Course.

Teachers and lecturers should refer to the *Course Assessment Specification* for mandatory information about the skills, knowledge and understanding to be covered in this Course.

The development of subject-specific and generic skills is central to the Course. Learners should be made aware of the skills they are developing and of the transferability of them. It is the transferability that will help learners with further study and enhance their personal effectiveness.

The skills, knowledge and understanding that will be developed in the Advanced Higher Statistics Course are:

- knowledge and understanding of a range of complex statistical concepts
- the ability to identify and use appropriate statistical models
- the ability to apply more advanced operational skills in statistical contexts
- the ability to use mathematical reasoning skills to extract and interpret information, think logically and solve problems
- the ability to communicate conclusions, exhibiting appreciation of their limitations
- the ability to think analytically about the consequences of methodological choices

Approaches to learning and teaching

Advanced Higher Courses place more demands on learners as there will be a higher proportion of independent study and less direct supervision. Some of the approaches to learning and teaching suggested for other levels (in particular, Higher) may also apply at Advanced Higher level but there will be a stronger emphasis on independent learning.

For Advanced Higher Courses, a significant amount of learning may be self-directed and require learners to demonstrate a more mature approach to learning and the ability to work on their own initiative. This can be very challenging for some learners, who may feel isolated at times, and teachers and lecturers should have strategies for addressing this. These could include, for example, planning time for regular feedback sessions/discussions on a one-to-one basis and on a group basis led by the teacher or lecturer (where appropriate).

Teachers and lecturers should encourage learners to use an enquiring, critical and problem-solving approach to their learning. Learners should also be given the opportunity to practise and develop research and investigation skills and higher order evaluation and analytical skills. The use of information and communications technology (ICT) can make a significant contribution to the development of these higher order skills as research and investigation activities become more sophisticated.

Learners will engage in a variety of learning activities as appropriate to the subject, for example:

- researching information for their subject rather than receiving information from their teacher or lecturer
- using active and open-ended learning activities such as research, case studies, project-based tasks and presentation tasks
- making use of the internet to draw conclusions about specific issues
- engaging in wide-ranging independent reading
- recording, in a systematic way, the results of research and independent investigation from different sources
- communicating findings/conclusions of research and investigation activities in a presentation
- participating in group work with peers and using collaborative learning opportunities to develop teamworking
- a mix of collaborative, co-operative or independent tasks which engage learners
- using materials available from service providers and authorities
- problem solving and critical thinking
- explaining thinking and presenting strategies and solutions to others
- effective use of questioning and discussion to engage learners in explaining their thinking and checking their understanding of fundamental concepts

- making links in themes which cut across the curriculum to encourage transferability of skills, knowledge and understanding — including with technology, geography, sciences, social subjects and health and wellbeing
- participating in informed debate and discussion with peers where they can demonstrate skills in constructing and sustaining lines of argument to provide challenge and enjoyment, breadth, and depth, to learning
- drawing conclusions from complex information
- using sophisticated written and/or oral communication and presentation skills to present information
- using appropriate technological resources (eg web-based resources)
- using appropriate media resources (eg video clips)
- using real-life contexts and experiences familiar and relevant to young people to meaningfully hone and exemplify skills, knowledge and understanding

Teachers and lecturers should support learners by having regular discussions with them and giving regular feedback. Some learning and teaching activities may be carried out on a group basis and, where this applies, learners could also receive feedback from their peers.

Teachers and lecturers should, where possible, provide opportunities to personalise learning and enable learners to have choices in approaches to learning and teaching. The flexibility in Advanced Higher Courses and the independence with which learners carry out the work lend themselves to this. Teachers and lecturers should also create opportunities for, and use, inclusive approaches to learning and teaching. This can be achieved by encouraging the use of a variety of learning and teaching strategies which suit the needs of all learners. Innovative and creative ways of using technology can also be valuable in creating inclusive learning and teaching approaches.

Centres are free to sequence the teaching of the Outcomes, Units and/or Course in any order they wish.

- Each Unit could be delivered separately in any sequence.

And/or:

- All Units may be delivered in a combined way as part of the Course. If this approach is used, the Outcomes within Units may either be partially or fully combined.

There may be opportunities to contextualise approaches to learning and teaching to Scottish contexts in this Course. This could be done through mini-projects or case studies.

Developing skills for learning, skills for life and skills for work

The following skills for learning, skills for life and skills for work should be developed in this Course.

2 Numeracy

2.1 Number processes
2.2 Money, time and measurement
2.3 Information handling

5 Thinking skills

5.3 Applying
5.4 Analysing and evaluating

Teachers and lecturers should ensure that learners have opportunities to develop these skills as an integral part of their learning experience.

It is important that learners are aware of the skills for learning, skills for life and skills for work that they are developing in the Course and the activities they are involved in that provide realistic opportunities to practise and/or improve them.

At Advanced Higher level it is expected that learners will be using a range of higher order thinking skills. They will also develop skills in independent and autonomous learning.

Approaches to assessment

Assessment in Advanced Higher Courses will generally reflect the investigative nature of Courses at this level, together with high-level problem-solving and critical thinking skills and skills of analysis and synthesis.

This emphasis on higher order skills, together with the more independent learning approaches that learners will use, distinguishes the added value at Advanced Higher level from the added value at other levels.

There are different approaches to assessment, and teachers and lecturers should use their professional judgement, subject knowledge and experience, as well as understanding of their learners and their varying needs, to determine the most appropriate ones and, where necessary, to consider workable alternatives.

Assessments must be fit for purpose and should allow for consistent judgements to be made by all teachers and lecturers. They should also be conducted in a supervised manner to ensure that the evidence provided is valid and reliable.

Unit assessment

Units will be assessed on a pass/fail basis. All Units are internally assessed against the requirements shown in the *Unit Specification*. Each Unit can be assessed on an individual Outcome-by-Outcome basis or via the use of combined assessment for some or all Outcomes.

Assessments must ensure that the evidence generated demonstrates, at the least, the minimum level of competence for each Unit. Teachers and lecturers preparing assessment methods should be clear about what that evidence will look like.

Sources of evidence likely to be suitable for Advanced Higher Units could include:

- presentation of information to other groups and/or recorded oral evidence
- exemplification of concepts using (for example) a diagram
- interpretation of numerical data
- investigations
- answers to multiple choice questions
- short written responses
- case studies

Evidence should include the use of appropriate subject-specific terminology as well as the use of real-life examples where appropriate.

Flexibility in the method of assessment provides opportunities for learners to demonstrate attainment in a variety of ways and so reduce barriers to attainment.

The structure of an assessment used by a centre can take a variety of forms, for example:

- individual pieces of work could be collected in a folio as evidence for Outcomes and Assessment Standards
- assessment of each complete Outcome
- assessment that combines the Outcomes of one or more Units
- assessment that requires more than the minimum competence, which would allow learners to prepare for the Course assessment

Teachers and lecturers should note that learners' day-to-day work may produce evidence which satisfies assessment requirements of a Unit, or Units, either in full or partially. Such naturally-occurring evidence may be used as a contribution towards Unit assessment. However, such naturally-occurring evidence must still be recorded and evidence such as written reports, recording forms, PowerPoint slides, drawings/graphs, video footage or observational checklists provided.

Combining assessment across Units

A combined approach to assessment will enrich the assessment process for the learner, avoid duplication of tasks and allow more emphasis on learning and teaching. Evidence could be drawn from a range of activities for a combined assessment. Care must be taken to ensure that combined assessments provide appropriate evidence for all the Outcomes that they claim to assess.

Combining assessment will also give centres more time to manage the assessment process more efficiently. When combining assessments across Units, teachers/lecturers should use e-assessment wherever possible. Learners can easily update portfolios, electronic or written diaries and recording sheets.

For some Advanced Higher Courses, it may be that a strand of work which contributes to a Course assessment method is started when a Unit is being delivered and is completed in the Course assessment. In these cases, it is important that the evidence for the Unit assessment is clearly distinguishable from that required for the Course assessment.

Preparation for Course assessment

Each Course has additional time which may be used at the discretion of the teacher or lecturer to enable learners to prepare for Course assessment. This time may be used near the start of the Course and at various points throughout the Course for consolidation and support. It may also be used for preparation for Unit assessment, and, towards the end of the Course, for further integration, revision and preparation and/or gathering evidence for Course assessment.

For this Advanced Higher Course, the assessment method for Course assessment is a question paper. Learners should be given opportunities to practise this method and prepare for it.

Authenticity

In terms of authenticity, there are a number of techniques and strategies to ensure that learners present work that is their own. Teachers and lecturers should put in place mechanisms to authenticate learner evidence.

In Advanced Higher Courses, because learners will take greater responsibility for their own learning and work more independently, teachers and lecturers need to have measures in place to ensure that work produced is the learner's own work.

For example:

- regular checkpoint/progress meetings with learners
- short spot-check personal interviews
- checklists which record activity/progress
- photographs, films or audio records

Group work approaches are acceptable as part of the preparation for assessment and also for formal assessment. However, there must be clear evidence for each learner to show that the learner has met the evidence requirements.

For more information, please refer to SQA's *Guide to Assessment*.

Added value

Advanced Higher Courses include assessment of added value which is assessed in the Course assessment.

Information given in the *Course Specification* and the *Course Assessment Specification* about the assessment of added value is mandatory.

In Advanced Higher Courses, added value involves the assessment of higher order skills such as high-level and more sophisticated investigation and research skills, critical thinking skills and skills of analysis and synthesis. Learners may be required to analyse and reflect upon their assessment activity by commenting on it and/or drawing conclusions with commentary/justification. These skills contribute to the uniqueness of Advanced Higher Courses and to the overall higher level of performance expected at this level.

In this Course, added value will be assessed by means of a question paper. This is used to assess whether the learner can retain and consolidate the knowledge and skills gained in individual Units. It assesses knowledge and understanding and the various different applications of knowledge such as reasoning, analysing, evaluating and solving problems.

Equality and inclusion

It is recognised that centres have their own duties under equality and other legislation and policy initiatives. The guidance given in these *Course/Unit Support Notes* is designed to sit alongside these duties but is specific to the delivery and assessment of the Course.

It is important that centres are aware of and understand SQA's assessment arrangements for disabled learners, and those with additional support needs, when making requests for adjustments to published assessment arrangements. Centres will find more guidance on this in the series of publications on Assessment Arrangements on SQA's website: www.sqa.org.uk/sqa/14977.html.

The greater flexibility and choice in Advanced Higher Courses provide opportunities to meet a range of learners' needs and may remove the need for learners to have assessment arrangements. However, where a disabled learner needs a reasonable adjustment/assessment arrangements to be made, you should refer to the guidance given in the above link.

Further information on Course/Units

The first column refers to sub-skills associated with each Assessment Standard.

The second column is the mandatory skills, knowledge and understanding given in the *Course Assessment Specification*. This includes a description of the Unit standard and the added value for the Course assessment. Skills which could be sampled to confirm that learners meet the minimum competence of the Assessment Standards are indicated by a diamond bullet point (♦). Those skills marked by an arrow bullet point (➢) can be assessed as part of the added value for the Course assessment.

For Unit assessment, to assess any sub-skill it would be sufficient for assessors to assess any one ♦ associated with that sub-skill except for the following:

Data Analysis and Modelling

Assessment Standard	Sub-skill	Unit assessment requirement
1.3	Modelling a discrete random variable	Both ♦s should be assessed.
1.4	Using discrete probability distributions	Both ♦s should be assessed.

Statistical Inference

Assessment Standard	Sub-skill	Unit assessment requirement
1.1	Working with the distribution of sample means and sample proportions	♦3 and ♦5 should be assessed. ♦1, ♦2 and ♦4 are optional.
1.3	Fitting a linear model to bivariate data	Both ♦s should be assessed.
1.3	Assessing the linear association between two variables	♦3 and ♦1 or ♦2 should be assessed.

In the first column there is advice given in brackets on the minimum requirements to assess each sub-skill at Unit assessment level.

The third column gives suggested learning and teaching contexts to exemplify possible approaches to learning and teaching. These also provide examples of where the skills could be used in activities.

Learners who are planning to go on to further studies in Statistics may find it beneficial to have used a statistical calculator during this Course.

Statistics: Data Analysis and Modelling (Advanced Higher)

1.1 Applying skills to data presentation and interpretation

Sub-skill	Description of Unit standard and added value	Learning and teaching contexts
Interpreting the Exploratory Data Analysis (EDA) of univariate data (It would be sufficient to assess only one ♦.)	♦ Present and interpret sample data in an appropriate form using a table, dotplot, stem-and-leaf diagram and boxplot. Appreciate that there are different methods of data collection and the difference between discrete and continuous data	The ability to categorise data as discrete or continuous must be established. Application of techniques in EDA should be applied to all data to establish reasonable assumptions about the underlying population.
	♦ Identify possible outliers and suggest possible action to be taken	Learners are expected to be able to communicate reasons for judgements and suggestions. Outliers are often identified by using fences within a data set. Values which lie beyond these fences are considered to be possible outliers. A commonly used definition of a lower fence is $Q_1 - 1{\cdot}5 \times IQR$ and that of an upper fence $Q_3 + 1{\cdot}5 \times IQR$, where IQR represents the inter quartile range.

1.2 Applying skills to probability theory

Sub-skill	Description of Unit standard and added value	Learning and teaching contexts
Working with theoretical and experimental probabilities (It would be sufficient to assess only one ♦.)	♦ Appreciate the necessary conditions for, and use of, the addition and multiplication laws of probability	The use of well-constructed experiments is key to an appropriate understanding of concepts of probability and underlines the importance of dealing with uncertainty as a central tenet of statistics
	♦ Calculate probabilities for events which are not mutually exclusive	Simple, practical examples of mutually exclusive events and non-mutually exclusive events should be explored to underline that the calculation of probabilities in each case are different and can give significantly different results depending on the exclusivity, or not, of two or more events.
	➢ Compare calculated theoretical probabilities with those obtained experimentally, or by simulation using appropriate technology	When comparing calculated theoretical probabilities with those obtained experimentally, or by simulation, it is important to understand the accuracy or otherwise of simulations in predicting further events. Similarly, how closely theoretical probabilities will be reflected by actual events should be clearly understood.

Calculating conditional probabilities	♦ Calculate simple conditional probabilities ➢ Calculate conditional probabilities requiring the use of Bayes' Theorem or equivalent methods	It is acceptable, preferable even, for learners to establish conditional probabilities intuitively and to formalise calculations using appropriate notation. Bayes theorem is used to 'reverse the condition', so that if the probability of F given that E has occurred is known, then the probability of E given that F has occurred can be found. Equivalent methods to Bayes' Theorem would include tree diagrams, Venn diagrams, tabulated probabilities and set notation. A full algebraic treatment of Bayes theorem is not a requirement. The concept of combining several events into two, E and not-E, should be understood and practised.
1.3 Applying skills to discrete random variables		
Sub-skill	**Description of Unit standard and added value**	**Learning and teaching contexts**
Modelling a discrete random variable (Both♦s should be assessed.)	♦ Construct the probability distribution of a discrete random variable ➢ Generate values of discrete data by simulation or experiment and compare their distribution to theoretical models ♦ Calculate the mean and standard deviation of a discrete random variable	It is instructive for learners to meet repeated simulations to allow comparison between a mathematical model and what happens when an experiment/simulation is undertaken. An appreciation of a model's strengths in accurately predicting long term results, and its limitations in terms of immediate prediction, will underpin much of more advanced statistical thinking.
Using the laws of expectation and variance	♦ Use the laws of expectation and variance: $\mathrm{E}(aX+b)=a\,\mathrm{E}(X)+b$ $\mathrm{E}(X\pm Y)=\mathrm{E}(X)\pm\mathrm{E}(Y)$ $\mathrm{E}(aX\pm bY)=a\,\mathrm{E}(X)\pm b\,\mathrm{E}(Y)$ $\mathrm{V}(aX+b)=a^2\,\mathrm{V}(X)$ $\mathrm{V}(X\pm Y)=\mathrm{V}(X)+\mathrm{V}(Y)$, where X and Y are independent	In considering the laws of expectation and variance, demonstration of simple examples will enable learners to establish the general results. In particular, extra care should be exercised in emphasising the difference between variance and standard deviation and that the laws should be applied to variance and converted to/from standard deviation where necessary. In practice standard deviation is more used by statisticians. A graphical approach is particularly useful in highlighting $\mathrm{V}(-X)=\mathrm{V}(X)$ and $\mathrm{V}(X+b)=\mathrm{V}(X)$.

	➢ Calculate $\mathrm{SD}(aX \pm bY)$ where X and Y are independent	

1.4 Applying skills to particular probability distributions		
Sub-skill	**Description of Unit standard and added value**	**Learning and teaching contexts**
Using discrete probability distributions (Both ♦s should be assessed.)	♦ Calculate uniform, binomial and Poisson probabilities	Selection of an appropriate distribution to model data from a given context is expected. Use an EDA and/or other techniques (such as chi-squared goodness of fit test, covered elsewhere in the Course) to confirm or establish a possible distribution of a data set. Use of ${}^{n}C_{r}$ is required and can be introduced in a variety of practical contexts. Reference should be made to Pascal's triangle. A useful link can be made with the Binomial Theorem and how combinations are relevant in some algebraic expansions. Assumptions required for a particular model should be introduced and discussed. Unnecessarily laborious calculations can be avoided by use of cumulative probability tables or functions on calculators and spreadsheets. These calculating aids can also be useful in comparing a Poisson approximation to a binomial distribution. (This approximation is not a requirement). This might be a suitable point at which to introduce the idea of hypothesis testing in an informal way using a binomial distribution. In such a situation the only interpretation required is that of whether or not a given result is probable, given the assumed binomial model.
	♦ Use standard results for the mean and variance of these distributions	Learners are expected to use these standard results: $\mathrm{E}(U) = \frac{k+1}{2} \quad \mathrm{V}(U) = \frac{k^2-1}{12}$ $\mathrm{E}(B) = np \quad \mathrm{V}(B) = npq$ $\mathrm{E}(P_o) = \lambda \quad \mathrm{V}(P_o) = \lambda$

	➢ Simulate these distributions using appropriate technology and compare them to probability distribution models	Spreadsheets can be useful in generating both simulated data and modelled results for chosen parameters, and then displaying a comparison of the two in the manner of an EDA. A more sophisticated approach will be met later in the Course with the chi-squared goodness-of-fit test. Learners are expected to be able to communicate reasons for decisions and be able to make suggestions for improving the quality of the given model and also why they have selected the model.
Using continuous probability distributions (It would be sufficient to assess only one ♦.)	♦ Calculate rectangular (continuous uniform) probabilities and use standard results for the mean and variance of this distribution	Learners are also expected to use: $\mathrm{E}(U)=\frac{a+b}{2} \quad \mathrm{V}(U)=\frac{(b-a)^2}{12}$
	♦ Calculate normal probabilities	
	➢ Calculate probabilities in problems involving the sum or difference of two independent normal random variables	Learners should be familiar with the use of the laws of expectation and variance and their application in the context of combining independent normal distributions.
Using the normal approximation to discrete probability distributions	♦ Demonstrate an understanding of appropriate conditions for a normal approximation to a binomial or Poisson distribution, together with the parameters of the approximate distribution	Consideration should be given to the reason (ease of calculation without significant loss of accuracy) for approximating a binomial or Poisson distribution, together with an understanding of the limitations of the process, and the conditions necessary for such approximations to be sufficiently accurate. The rules of thumb adopted for this Course are: Use the normal approximation to a binomial distribution when *np* and *nq* are both > 5. Use the normal approximation to a Poisson distribution when λ > 10. The validity of these rules can be investigated using calculators or spreadsheets.
	➢ Demonstrate the use of a continuity correction when applying a normal approximation to the binomial and Poisson distributions	Where the distribution of the discrete variable is known, exact theoretical probabilities can be calculated without resort to the normal approximation. However, a familiarity in the use of approximations in establishing an understanding of the underlying data in the more usable context of the normal distribution means that the ability to apply a normal approximation may still be tested. Learners should be able to compare probabilities calculated under a normal approximation with those using the binomial or Poisson distribution directly and so establish the accuracy of any such approximations.

Statistics (Advanced Higher) Statistical Inference		
1.1 Applying skills to sampling and the central limit theorem		
Sub-skill	**Description of Unit standard and added value**	**Learning and teaching contexts**
Identifying and using appropriate random sampling methods (It would be sufficient to assess only one ♦.)	♦ Appreciate that there are different methods of data collection, and be able to generate a simple random sample from a population	Understand the difference between random and non-random sampling.
	♦ Describe and distinguish between simple random, systematic, stratified and cluster sampling	The difference between a sample and a census should be stressed. The advantages and disadvantages of each type of sampling should be discussed, noting that quota/convenience is not an example of random sampling.
	➢ Appreciate that non-random sampling methods such as quota or convenience sampling could lead to an unrepresentative sample and biased conclusions	Rather than just discussing theoretical concepts, it may be more instructive to collect some data for analysis and in so doing to encounter both difficulties in achieving random sampling and also practical solutions to these problems. For instance, possible strategies for sampling the lengths of words and sentences used could be considered when trying to distinguish between the works of two different authors.
Working with the distribution of sample means and sample proportions (It would be sufficient to assess ♦3 **and** ♦5.)	♦ Demonstrate an understanding that the sampling distribution of the sample mean from a parent population that is normal is itself normal	Many online simulations are easily accessed and can provide very clear illustration of the distribution of sample means from a variety of parent populations, eg Rice Virtual Lab in Statistics (RVLS).
	♦ Demonstrate an understanding that the sampling distribution of the sample mean from a parent population, which is not normal, is approximately normal, by invoking the Central Limit Theorem when the sample is large enough	The Central Limit Theorem states that for sufficiently large n (in this Course $n \geq 20$) the distribution of the sample mean is approximately normal, irrespective of the distribution of the parent population. Learners are required to quote and use, although not prove that the distribution of the sample mean is approximately $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
	♦ Describe the sampling distribution of the sample mean and use the appropriate standard error in calculations involving this distribution	The quantity $\frac{\sigma}{\sqrt{n}}$ is often referred to as the standard error of the sample mean.

	➢ Describe the sampling distribution of the sample proportion and use the appropriate standard error in calculations involving this distribution	The Central Limit Theorem also states that for sufficiently large n (in this Course $n \geq 20$) and $np > 5$ $nq > 5$ the distribution of a sample proportion is approximately normal. Learners are required to quote and use, although not prove that the distribution of a sample proportion is approximately $$\hat{p} \sim N\left(p, \frac{pq}{n}\right)$$ It is not the parameters that are approximate but the distribution. It might be instructive to derive the formulae for these parameters. The quantity $\sqrt{\frac{pq}{n}}$ is often referred to as the standard error of the sample proportion.
	♦ Use the sample mean as a best estimate of the population mean	In elementary sampling theory the population mean/proportion is estimated by the sample mean/proportion and the population variance by the sample variance given by $$s^2 = \frac{\sum\left(x - \bar{x}\right)^2}{n-1}$$
	♦ Use the sample variance as an estimate of the population variance	The reasoning behind the $n-1$ denominator can be demonstrated in a classroom, with learners generating their own sample mean distributions and investigating denominators of n and $n-1$ for different sample sizes. Spreadsheets can be useful in this context, particularly the random number generation functions. It may be instructive to prove, using expectation algebra from *Data Analysis and Modelling 1.3*, that the best estimate of the population variance is obtained using the $n-1$ denominator. When working with the distribution of sample means or proportions, care should be taken to clarify precisely which statistics are calculated from the sample and which parameters are assumed for the parent population.

1.2 Applying skills to intervals and estimation		
Sub-skill	**Description of Unit standard and added value**	**Learning and teaching contexts**
Obtaining confidence intervals (It would be sufficient to assess only one ♦.)	♦ Calculate a *z*-interval for the population mean	This might be an appropriate point to undertake some practical sampling and then to discuss what can be inferred about the population under consideration, particularly with reference to whether or not the population variance is known. Common statistical contexts in which population variance is assumed can be encountered, although this may well not be the case in classroom situations. If σ is known a 95% confidence interval for the population mean is given by $$\bar{x} \pm 1{\cdot}96\frac{\sigma}{\sqrt{n}}$$ It is also common practice to use a 99% confidence interval where 1·96 is replaced by 2·58.
	♦ Appreciate the need to use Student's *t*-distribution when the population variance is unknown	When the population variance has to be estimated by the sample variance, another source of variability is introduced and the *t*-distribution should be used in place of the normal distribution. The shape of the *t*-distribution depends on the number of degrees of freedom, $\upsilon = n-1$. As $\upsilon \to \infty$ the t-distribution tends to the normal distribution and as υ decreases the distribution becomes more spread out, with a marked difference for $\upsilon \leq 20$. Use of online simulations can illustrate this clearly. The t-distribution was derived by W. S. Gossett in Dublin in 1908 while conducting tests on the average strength of Guinness beer. An employee was not permitted to publish under their own name so Gossett used the pseudonym Student and hence the name of the distribution. If σ is unknown, an approximate 95% confidence interval for the population mean is given by $$\bar{x} \pm t_{n-1,0{\cdot}975}\frac{s}{\sqrt{n}}$$ This can be interpreted as saying that the best estimate of the population mean μ is the sample mean $\bar{x}$ but that it is recognised that this is only an estimate of the population mean and that the true value of the population mean lies within a range of possible values, dependent on the size of the sample.

		A useful definition of a 95% confidence interval is that if a large number of samples are taken and a confidence interval computed for each, then 95% of these intervals would be expected to contain the population mean.
	➢ Calculate an approximate confidence interval for the population proportion	This is a particularly challenging concept and may best be tackled with real data. An approximate confidence interval for the population proportion p is given by $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$ for $n \geq 20$, $np > 5$ and $nq > 5$, where $\hat{q} = 1 - \hat{p}$ and $\hat{p}$ is the sample proportion. While a confidence interval for the population mean may often be calculated with an assumed population variance, this is not the case with the population proportion, and the $\sqrt{\frac{\hat{p}\hat{q}}{n}}$ term is calculated using the estimates of the population proportion p.

Using control charts	♦ Construct and interpret a Shewhart control chart for the sample mean or proportion	Shewhart charts attempt to clearly illustrate and quantify the variation inherent in natural processes and to allow some statistical determination of whether or not such processes are proceeding in an expected manner. For example, the automated fill volume of a soft drink bottle can be sampled at suitable time intervals to check whether or not it is consistently at least the stated volume. A 3-sigma Shewhart chart has control limits drawn 3 standard deviations either side of the expected value. The process may be out of statistical control if any single data point is outside the 3σ limits. The control limits for the sample mean are $\mu \pm 3\frac{\sigma}{\sqrt{n}}$ The control limits for the sample proportion are $p \pm 3\sqrt{\frac{pq}{n}}$. Note the similarity between control limits for the sample proportion and confidence interval for the population proportion. The use of a normal approximation to a binomial distribution can be of concern in this type of chart as the approximation is poor when $np < 5$, as can be the case. Learners will not be expected to work with a control chart when $np < 5$.
	➢ Use the Western Electric rules to recognise when a process may be out of statistical control and influenced by common or special causes	One way of determining when a process may be out of statistical control is to use the Western Electric rules, which are: Any single data point falls outside the 3σ limits. Two out of three consecutive points fall beyond the same 2σ limit. Four out of five consecutive points fall beyond the same 1σ limit. Eight consecutive points fall on the same side of the centre line. The calculation of the probabilities of each of these four events should be considered.
	➢ Estimate chart parameter(s) from a sample when population data is unavailable	When a population mean is unavailable, the sample mean is used as the best estimate of the population mean. Estimation of the population variance is beyond the scope of this course. The population proportion can be estimated from the sample proportion and, in this case, the population variance can also be estimated.

1.3 Applying skills to bivariate analysis		
Sub-skill	**Description of Unit standard and added value**	**Learning and teaching contexts**
Fitting a linear model to bivariate data (Both ♦s should be assessed.)	♦ Interpret a scatterplot observing whether or not a linear model is appropriate ♦ Calculate the least squares regression line of y on x	A scatterplot can give a quick impression of the degree of correlation (for a linear relationship) or association (for a non-linear relationship) between the variables. The linear model used is $Y_i = \alpha + \beta x_i + \varepsilon_i$, where Y_i is the expected value for a given x_i, α and β are the population y-intercept and gradient respectively and ε_i is the error term, which may arise from an error in measurement or from natural variation. The model assumes that: • ε_i are independent • $\mathrm{E}(\varepsilon_i) = 0$ • $\mathrm{V}(\varepsilon_i) = \sigma^2$ (a constant for all x_i) Estimates for α and β, a and b, which give the fitted line $y = a + bx$, are calculated using $b = \frac{S_{xy}}{S_{xx}}$ and $a = \bar{y} - b\bar{x}$, where $S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$ $S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$ (used in next section) $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$ Derivation of these alternative formulae might be interesting for some learners, but will not be examined. There is an opportunity to combine several aspects of mathematics in the derivation of the formulae used for b and a. The minimum value of $\sum \varepsilon_i^2$ is to be found. This can be done by completing the square or using differential calculus. The latter method relies upon

		partial differentiation, so may be a little unsatisfactory, but does provide a nice illustration of the use of differentiation in an applied context.
	➢ Appreciate the difference between regressing y on x and x on y	The regression of y on x gives a formula for y in terms of x and is used to predict a y value for a given x. The regression of x on y gives a formula for x in terms of y and is used to predict an x value for a given y.
Assessing the linear association between two variables (It would be sufficient to assess ♦3 **and (**♦1 **or** ♦2**)**.)	♦ Calculate and interpret the product moment correlation coefficient	The product moment correlation coefficient (pmcc) r is calculated using $r = \frac{S_{xy}}{\sqrt{S_{xx}\ S_{yy}}}$ and measures the strength of linear association between two variables. The idea that association is not the same as causation should be appreciated.
	♦ Calculate and interpret the coefficient of determination	The coefficient of determination R^2 is given by $R^2 = \frac{S^2_{xy}}{S_{xx}\ S_{yy}}$ and is the square of the pmcc. R^2 gives the proportion of the total variation in the response variable that is explained by the linear model and in some spreadsheets is the value given when lines of best fit are calculated. A small value indicates that the line is of little use for prediction.
	♦ Calculate a fitted value and its residual	The 'hat' notation is used to specify a fitted value (one obtained from the line of best fit), as opposed to a data value. Use the calculated (or given) line of best fit $\hat{Y}_i = a + bx_i$ to obtain a fitted value for a given x_i. For data point (x_i, y_i) the fitted value at x_i is $\hat{Y}_i$ and the residual e_i is given by $y_i - \hat{Y}_i$

	➢ Interpret a residual plot	A residual plot is used to check the model assumptions that • $\mathrm{E}(\varepsilon_i) = 0$ • $\mathrm{V}(\varepsilon_i) = \sigma^2$ (a constant for all x_i) Ideally the plot of residuals against fitted values should show a random scatter centred on zero. If this is not the case (systematic pattern or variance of residuals not constant), then the model may be inappropriate (perhaps non-linear) or the data may require to be transformed to restore constant variance.
	➢ Comment on given simple transformations to obtain improved models	If data transformation is appropriate and successful, any assumptions previously not satisfied should now be clear in a second residual plot.
Estimating with bivariate data	♦ Assess the reliability of prediction based on fitted values, considering the effects of correlation, interpolation and extrapolation	At this point, a qualitative approach is taken recognising that, with even fairly strong correlation, referring to a fitted value as a definite outcome is not appropriate. Data about men's and women's gold medal winning times in the Olympic 100-metres, published in *Nature*, gives an interesting focus for discussion about interpolation and extrapolation. *Nature* **431**, 525 (30 September 2004) | doi:10.1038/431525a; Published online 29 September 2004.
	➢ Calculate the appropriate statistics required for bivariate intervals	The sum of the squared residuals (SSR) is given by $\sum\left(y_i - \hat{Y}_i\right)^2$ It can be shown that SSR = $S_{yy} - \frac{S^2_{xy}}{S_{xx}}$. Derivation of this alternative formula might be interesting for some learners, but will not be examined. The variance of the residuals σ^2 is estimated by $s^2 = \frac{SSR}{n-2}$.

	➢ Construct a prediction interval for an individual response ➢ Construct a confidence interval for a mean response	A further assumption that $\varepsilon_i \sim N(0, \sigma^2)$ permits the construction of a $100(1-\alpha)\%$ prediction interval for an individual response Y_i/x_i which is given by $\hat{Y}_i \pm t_{n-2,1-\alpha/2}\, s\sqrt{1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}}$ ii) a $100(1-\alpha)\%$ confidence interval for a mean response $E(Y_i/x_i)$ which is given by $\hat{Y}_i \pm t_{n-2,1-\alpha/2}\, s\sqrt{\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}}$ In a prediction or confidence interval, the reliability of the estimate depends on the sample size, the variability in the sample and the value of x_i

Statistics: Hypothesis Testing (Advanced Higher)

1.1 Applying skills to parametric tests

Sub-skill	Description of Unit standard and added value	Learning and teaching contexts
Identifying and performing an appropriate one sample test for the population mean and proportion (It would be sufficient to assess only one ♦.)	♦ Perform a specified test for the population mean, for the cases i) σ^2 known (z-test) ii) σ^2 unknown but a large sample (z-test) iii) σ^2 unknown with a small sample (t-test)	The following terms should be understood: null hypothesis H_0, alternative hypothesis H_1, level of significance, one/two-tail test, distribution under H_0, test statistic, critical value/region, p-value, reject/accept H_0. A formal approach to hypothesis testing is expected: • State the hypotheses, the level of significance and whether a 1 or 2-tail test. • Compute, under H_0, the test statistic and/or p-value. • Accept or reject H_0. • Communicate the conclusion in context. Both the z and t-tests are concerned with making inferences about population means and have the underlying assumption that populations are distributed normally, with known or unknown variance respectively, and that sample values are independent.
	♦ Perform a z-test for the population proportion	Learners could be introduced to a continuity correction of $\pm\frac{1}{2n}$ to make the test comparable to that for using the normal approximation to the binomial distribution, although this is not a requirement.
	➢ Select and justify the choice of an appropriate test, together with its underlying assumptions	No formulae will be given for these one sample tests.
Identifying and performing an appropriate two sample test (independent or paired data) for comparing population means and proportions (it would be sufficient to	♦ Use a t-test to assess evidence about the population mean difference in a paired data experiment	With paired data it may be possible to work with the difference between pair values and hence a single distribution paired sample t-test, even for small n, where $T_{n-1} = \frac{\bar{X}_d - \mu_d}{\frac{S_d}{\sqrt{n}}}$ where $\mu_d = 0$

assess only one ♦)	♦ Test the hypothesis that two populations have the same mean for cases where the population variances are i) known (*z*-test) ii) unknown but samples are large (*z*-test) iii) unknown and samples are small (*t*-test)	If not paired, then the comparison of two samples, each with their own mean and variance, from two populations is a more complex undertaking than that with a single sample from a given population in the previous outcome. If the two populations can be assumed to be normal and independent with either known variances or both sample sizes at least 20, then, in this Course, a two-sample *z*-test may be used where $Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ Using the laws of variance from *Data Analysis and Modelling 1.3* may be useful here. Otherwise, assuming that the population variances are equal, we may use a two-sample *t*-test where $T_{n_1+n_2-2} = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where $s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$
	♦ Test the hypothesis that two populations have the same proportion, for only the case where both samples are large	With large samples $(n_i \geq 20$, $n_i p_i > 5$ and $n_i q_i > 5)$, the population proportion test uses $Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where p is the pooled proportion $\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ and $q = 1 - p$.
	➢ Select and justify the choice of an appropriate test, together with its underlying assumptions	Formulae for *z* and *t* will be given for all two population tests.

1.2 Applying skills to non-parametric tests		
Non parametric tests make no assumptions about the distributional form of populations, eg normality. As a result the hypotheses are often framed in terms of medians rather than means. The use of ranks may help to reduce the influence of outliers in the data. The use of a continuity correction is expected if a normal approximation is employed. Formulae for the mean and variance of the test statistic will be given.		
Sub-skill	**Description of Unit standard and added value**	**Learning and teaching contexts**
Identifying and performing an appropriate test for population median/s (It would be sufficient to assess only one ♦.)	♦ Use a Wilcoxon Signed-Rank test to assess evidence about the population median from a simple random sample and about the population distributions from paired data	The Wilcoxon tests assume that the populations are symmetrical (not necessarily normal) and thus the means and medians are equal so that the null hypothesis can refer to either of these. Learners may be required to use a table of critical values (appreciating that the smaller sum of ranks is the test statistic) or to employ a normal approximation for sample sizes of at least 20.
	♦ Use a Mann-Whitney test to assess evidence about the medians of two populations using independent samples	The Mann-Whitney test assumes that the two populations have the same shape and variability. The table of critical values will be provided in the data booklet for sample sizes up to 20. A normal approximation should be used for larger sample sizes. It would be appropriate to encourage learners to find a *p*-value from first principles, leading to an understanding of a table of combinations.
	➢ Use a normal approximation, when required, in any calculation of a test statistic or *p*-value	
	➢ Select and justify the choice of an appropriate test, together with its underlying assumptions	
Identifying and performing an appropriate chi-squared test (It would be sufficient to assess only one ♦.)	♦ Perform a chi-squared test for goodness-of-fit to a discrete distribution	Hypotheses need to be carefully stated and learners should be aware that this is a one-tail test.
	♦ Perform a chi-squared test for association in a contingency table	The appropriate number of degrees of freedom, k-1-m for a goodness-of-fit test, or $(r-1)(c-1)$ for a test of association, needs to be known, where m is the number of parameters which need to be estimated in order to find the expected frequencies (eg 1 for a Poisson distribution if the mean is not stated), r is the number of rows in a contingency table and c the number of columns.

	➢ Deal with small expected frequencies	When using a chi-squared statistic, approximating a discrete distribution with a continuous one, the approximation is not reliable if expected frequencies are too small. When working with small expected frequencies, learners should recognise that, for a reliable test: • at least 80% of expected frequencies should be ≥ 5 • none should be < 1 To ensure that these criteria are met, categories/frequencies may have to be combined with a resultant loss in the number of degrees of freedom. Yates' correction for a 2×2 contingency table is not required but all expected frequencies should be ≥ 5.
1.3 Applying skills to bivariate tests		
Sub-skill	**Description of Unit standard and added value**	**Learning and teaching contexts**
Identifying and performing an appropriate hypothesis test on bivariate data (It would be sufficient to assess only one ♦.)	♦ Test the hypothesis that the slope parameter in a linear model is zero	The β–test (for the slope parameter being zero) makes the assumption that ε_i are independent and identically distributed N(0, σ^2) and uses the *t*-statistic $t = \frac{b\sqrt{S_{xx}}}{s}$ The conclusion should comment on the evidence for the model being useful for prediction.
	♦ Test the hypothesis that the population correlation coefficient is zero	The ρ–test (for the pmcc being zero) makes the assumptions that the variables are independent and follow approximately a bivariate normal distribution and uses the *t*-statistic $t = \frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}}$ The conclusion should comment on the evidence for a linear association between the variables. Some learners may find an interesting mathematical exercise in proving that the two *t*-statistics above are in fact equivalent.
	➢ Communicate appropriate assumptions	

Further exemplification

The following two exemplars illustrate a combined approach to the learning and teaching of the *Statistical Inference* Unit and the *Hypothesis Testing* Unit.

Both exemplars are for illustration purposes only and must **not** be used as assessment instruments.

Exemplar 1: Statistical Inference and Hypothesis Testing — authorship

This exemplar demonstrates how the Assessment Standards could be covered for both of the above Units.

Specifying the problem and planning the investigation [2.1],

Stylometry is the application of the study of linguistic style and is often used to determine the authorship of anonymous or disputed documents. Historically it has had legal, academic, literary and political applications. Lorenzo Valla, an Italian Catholic priest, is credited with one of the first examples of such techniques in 1439 in his proof that the Donation of Constantine[1] was a forgery and, more recently, academics have confirmed the view that William Shakespeare collaborated with some his contemporaries, including Christopher Marlowe, in some of his work.

The development of modern computers has enabled researchers and scholars to hone their methods as large quantities of data can be analysed in a fraction of the time previously required; indeed it is possible for entire texts to be analysed rather than just a sample.

This investigation, however, is simpler and is interested in whether the authorship of two different novels can be determined from samples of the novels.

I decided to compare two authors who are considered among the best of the 19th Century, Charles Dickens and Oscar Wilde. I chose to compare *Great Expectations*, one of Dickens' fifteen novels, and Wilde's only novel, *The Picture of Dorian Gray*.

Selecting relevant data [2.2]

I decided to sample 50 words and 50 sentences from each of the novels and record the number of letters in each of the words and the number of words in each of the sentences.

The samples were taken using a two-stage cluster sampling method. Each page on the book was considered as a cluster and 50 pages were chosen at random

[1] The Donation of Constantine is a forged document supposedly written by the Roman Emperor Constantine (285–337 AD), giving the Catholic Church ownership of huge parts of the western Roman Empire.

using a random number generator. Different random numbers were generated for each of the two books. A line on the page was then randomly chosen and the sixth word and the first full sentence starting on that line selected. It was important that I did not simply pick the first word of a page or line as new paragraphs/chapters invariably start with a shorter word.

The following issues were considered when sampling:

- If a page was selected at random more than once, I generated a new randomly chosen page number.
- Any words which had a hyphen were treated as one word.

The word lengths were compared first. The following frequency table shows the number of letters in the words from the samples of 50 words from the two books:

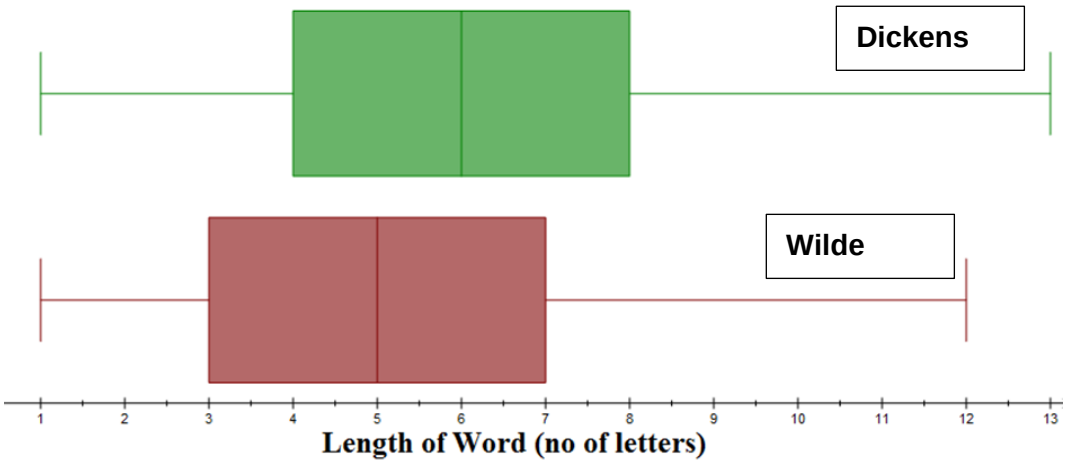
Presenting and analysing the data [2.3]

		Number of letters												
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Frequency	**Wilde**	1	6	10	5	5	6	5	5	5	1	0	1	0
	Dickens	2	2	4	6	7	5	8	5	2	3	5	0	1

In these samples, the modal word length was 3 for Wilde and 7 for Dickens, while the range was almost identical.

I calculated 5-figure summaries for the two samples so that I could draw boxplots which clearly show the distributions of the samples.

5-figure summary	Wilde	Dickens
Lowest value	1	1
Lower quartile	3	4
Median	5	6
Upper quartile	7	8
Highest value	12	13



While the boxplots are quite similar, each of the values in the 5-figure summary for Wilde is one less than that of Dickens (apart from the lowest values) which could suggest that Wilde tends to use slightly shorter words than Dickens. However, the distributions are so similar that I think it would be difficult to determine the authorship of a sample of words from each of the novels.

The samples for sentences can be compared in the same way.

The range of values was much greater for the number of words in the sentences sampled, so I drew a back-to-back stem-and-leaf diagram.

Sentence length

888776655543333	0	334568889
8544432221000000	1	000112377779
9877753220	2	033456
54310	3	14667
95	4	6678
	5	001234
81	6	556
	7	08
	8	8
	9	2
	10	
	11	2

Wilde **Dickens**

n = 50 3 | 4 = 34

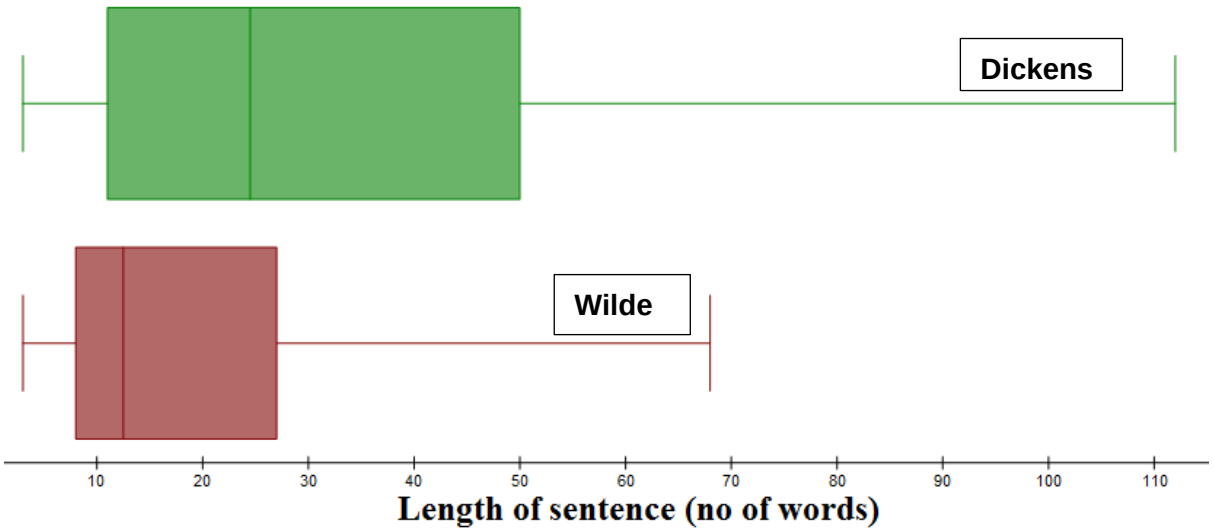
It can be seen from the stem-and-leaf diagram that two of the sentences from *The Picture of Dorian Gray* and one from *Great Expectations* were considerably larger than the others in the samples.

An outlier can be defined as a value which is more than 1·5 times the interquartile range above the upper quartile (this boundary is known as an upper fence).

In the sample from *Great Expectations*, the interquartile range is 39, so the upper fence is 50 + 1·5 × 39 = 108·5, which means that a sentence length of 112 is an outlier. Similarly, the upper fence in the other sample is 55·5 which means that the two largest sentence lengths are also outliers. The presence of the outliers in the samples is important and they cannot be removed or ignored.

The diagram suggests that Wilde's sentence lengths tend to be shorter, with the mode being 10 versus a modal length of 17 for Dickens. There is further evidence of this from the 5-figure summary and boxplots.

5-figure summary	Wilde	Dickens
Lowest value	3	3
Lower quartile	8	11
Median	12·5	24·5
Upper quartile	27	50
Highest value	68	112



These boxplots are quite different. The median value in the sample of words from Wilde's novel is only slightly larger than the lower quartile from Dickens' sample and the median from this sample is almost twice as big as Wilde's, which suggests that there is a difference in the sentence lengths of the two novelists and that Wilde tends to write shorter sentences than Dickens. I think it would be easier to determine the authorship of a sample of sentences from each of the novels.

While boxplots can give a useful visual comparison of the two samples, a more robust approach is to apply a hypothesis test to the samples.

As the data is discrete numerical data, I chose to use the Mann-Whitney test. This is a powerful non-parametric test which can be used to test whether or not two independent samples come from identical probability distributions. In this investigation it can be used to test whether or not the two authors use the same length of words or sentences in their novels. The test assumes that the distributions of the words and sentences used by Wilde and Dickens have the same shape and variance so that any difference in the populations of words and sentences comes from a difference in their location, which is indicated by their medians. The test will involve the two samples for the words, then the sentences, to be combined, ranked from 1 to 100 and a rank sum calculated for each of the two authors. The use of the Mann-Whitney test also removes any concern about the outliers as their size relative to the rest of the data does not impact on the test result.

Using a spreadsheet, I ranked the samples and obtained the following rank sums for the samples:

	Words	**Sentences**
Wilde	2261·5	2093·5
Dickens	2788·5	2956·5

As the sample size of both of the samples used is 50, the normal approximation to the Mann-Whitney test is required. Let W be the smaller rank sum, then $W \sim \mathrm{N}(\mu, \sigma^2)$ with

$\mu = \mathrm{E}(W) = \frac{1}{2}\mathrm{n}(n+m+1)$ and $\sigma^2 = \mathrm{Var}(W) = \frac{1}{12}nm(n+m+1)$, where n and m are the sizes of the two samples.

In this test $n = m = 50$, which gives:

$\mu = \mathrm{E}(W) = \frac{1}{2}50\times(50+50+1) = 2525$ and

$\sigma^2 = \mathrm{Var}(W) = \frac{1}{12}50\times50\times(50+50+1) = 21041\frac{2}{3}$ so $W \sim \mathrm{N}(2525,\ 21041\frac{2}{3})$.

A continuity correction of $\pm\frac{1}{2}$ is applied when calculating the p-value since a normal approximation is being used.

Two-tail hypotheses tests at the 5% significance level were carried out; testing the null hypothesis of no difference between the median word lengths/sentence lengths against the alternative hypothesis that the median word length/sentence

length for Wilde is different to that of Dickens. Let η_W denote the median for Wilde and η_D denote the median for Dickens.

Mann-Whitney test for words

H_0: $\eta_W = \eta_D$ and H_1: $\eta_W \neq \eta_D$

2-tail test, $\alpha = 0{\cdot}05$

$p = 2 \times P(W \leq 2261{\cdot}5)$

$$p = 2 \times P(Z \leq \frac{2262 - 2525}{\sqrt{21041\frac{2}{3}}})$$

$p = 2 \times P(Z \leq -1{\cdot}81)$

$p = 2 \times 0{\cdot}035$

$p = 0{\cdot}07$

Since 0·07 > 0·05, there is insufficient evidence at the 5% significance level to reject H_0; the median word length in Wilde's novel is not different from the median word length in Dickens' novel.

This result was not surprising as the boxplot for the word length suggested that the distributions of the word lengths are similar.

Mann-Whitney test for sentences

H_0: $\eta_W = \eta_D$ and H_1: $\eta_W \neq \eta_D$

2-tail test, $\alpha = 0{\cdot}05$

$p = 2 \times P(W \leq 2093{\cdot}5)$

$$p = 2 \times P(Z \leq \frac{2094 - 2525}{\sqrt{21041\frac{2}{3}}})$$

$p = 2 \times P(Z \leq -2{\cdot}97)$

$p = 2 \times 0{\cdot}0015$

$p = 0{\cdot}003$

Since 0·003 << 0·05, H_0 is rejected, the result is highly significant, suggesting there is considerable difference between the length of sentences in the two novels with Wilde's sentences being shorter than Dickens'.

Communicating the conclusion [2.4]

In conclusion, the lengths of words in the samples from the two authors seemed to be similar, and this was confirmed by the result of the hypothesis test which suggested that it would be difficult to determine the authorship of a small sample of words from their novels. If given two samples of sentence lengths, however, it should be possible to determine which sample belonged to which author as Wilde tended to use shorter sentences than Dickens.

Exemplar 2: Statistical Inference and Hypothesis Testing — social factors and academic attainment

This exemplar demonstrates how the Assessment Standards could be covered for both of the above Units.

Do social factors have an effect on academic attainment?

Specifying the problem and planning the investigation [2.1]

It can be argued that one of the main aims of democratic governments is to try and provide equal life chances to all its citizens. If some sections of society do not have access to all resources, then inequality is likely to increase and the more unequal a society, the less stable and civilised it is.

The Scottish Government uses the Scottish Index of Multiple Deprivation (SIMD) to investigate equality of provision and subsequently plan policy to try and help those most in need.

The SIMD considers 38 indicators within the 7 domains of employment, income, health, education, access, crime and housing, each with different weighting. As an example, employment contributes 28% and access 9% to the overall index. An example of one of the three employment indicators is 'working age unemployment claimant count averaged over 12 months'. The SIMD is calculated for each of 6506 datazones within Scotland and each datazone has on average 800 people living in it. Datazones are closely aligned with post-code areas. The datazones are ranked in order and given an overall relative ranking with 1 being most deprived and 6505 being the least deprived. For my own post-code of EH10 6LT, in 2012, the SIMD rank is 6487, and while the 7 domain ranks are reasonably consistent, the crime domain rank is significantly different at 5117 (obtained from the Scottish Neighbourhood Statistics website).

One key area of public spending is education, with government, local authorities and schools all looking to target funding towards policies that will have the greatest impact for all pupils. If it is decided to spend more on sports' facilities to hopefully go some way to tackling the problem of obesity in Scotland, it would make sense to direct more funding to areas which currently have poor provision. While educational outcomes are very difficult to measure, one thing that can be measured is exam results and these are used in many contexts to make educational judgements. The Scottish Government currently gives all schools feedback relating exam performance to SIMD.

In this study, the connection between SIMD and exam performance for S4 pupils will be investigated, with average tariff score taken as the measure of exam performance. All qualifications are given a numerical tariff, depending on level of difficulty and this is totalled for each pupil and then averaged for each datazone.

Average tariff score data for a datazone can be obtained from the Scottish Neighbourhood Statistics website. The SIMD rank for a datazone is provided in an Excel file (Data zone SIMD data Mar 15).

As a possible relationship between two variables is being investigated, a bivariate analysis is appropriate. Once data has been acquired, a scatterplot will give a clearer picture of the dataset, followed by a more formal analysis to find a regression equation, the strength of any correlation and any evidence of a linear relationship between the two variables.

Selecting relevant data [2.2]

Sampling strategy

Data is available for the population of all 6505 datazones from the Scottish Neighbourhood Statistics website, but the number of requests is limited so a sample of 50 datazones will be used. A simple random sample could be taken, but it is also possible to group the datazones according to SIMD and then conduct stratified sampling. If this is not done, it would be possible to select a random sample that did not sample all groups. Thus stratified sampling would be expected to give a more representative random sample.

Sampling methods

Using the supplied file (Data zone SIMD data Mar 15), the datazones can be ordered according to SIMD. The 6505 datazones will be put into ten strata of size approximately 650 and five datazones randomly sampled form each stratum. Selection of datazones will be performed using the RANDBETWEEN function in Excel.

		Chosen rank within stratum				
Initial rank	**Final rank**	**1**	**2**	**3**	**4**	**5**
1	650	451	112	621	47	267
651	1300	1138	1280	1286	731	1198
1301	1950	1939	1869	1370	1746	1484
1951	2600	2560	2249	2506	1960	2458
2601	3250	3171	2844	3043	2919	2814
3251	3900	3707	3539	3431	3689	3712
3901	4550	4393	4415	4084	4432	4165
4551	5200	5126	5118	4944	4876	5084
5201	5850	5388	5651	5682	5791	5522
5851	6505	5960	5957	6183	6187	6204

Chosen datazones

S01001957	S01001245	S01004045	S01006063	S01003644
S01004800	S01003833	S01005987	S01005892	S01004027
S01005649	S01005750	S01004653	S01005702	S01004155
S01005838	S01000623	S01001125	S01003189	S01000916
S01002669	S01004943	S01003491	S01002725	S01003533
S01003917	S01003939	S01003676	S01003321	S01002551
S01005473	S01001211	S01000719	S01003532	S01002896
S01005681	S01005349	S01003888	S01000341	S01005317
S01006293	S01004900	S01003163	S01005835	S01005051
S01001688	S01001989	S01002336	S01006116	S01002123

This is an example from the datatzone selection part of the accompanying spreadsheet (Data zone vs SIMD rank sampling Mar 15, Sample selection tab). At each recalculation, the chosen ranks change, so the numbers here do not correspond to those of the chosen sample.

The first stratum is for ranks between 1 and 650, and the ranks chosen by the RANDBETWEEN(1, 650) function are 451, 112, 621, 47 and 267. A LOOKUP function was used to identify that the datazone with SIMD of 451 was S01001957.

When using the LOOKUP function, it was necessary to order the datazones in SIMD order, as opposed to the given datazone order, to allow the function to perform as intended.

From the Scottish Neighbourhood Statistics website the average tariff score for S4 pupils in 2012/13 for the chosen datazones was obtained and results are shown in Table 1 below.

Table 1: raw sample data

Datazone	**SIMD (x)**	**S4 ave tariff score (y)**	**x^2**	**y^2**	**xy**
S01003572	167	194	27889	37636	32398
S01001874	415	201	172225	40401	83415
S01003287	423	132	178929	17424	55836
S01002778	564	111	318096	12321	62604
S01000074	571	132	326041	17424	75372
S01004816	939	195	881721	38025	183105
S01001345	945	156	893025	24336	147420
S01002721	984	178	968256	31684	175152
S01004642	1233	186	1520289	34596	229338
S01001035	1289	201	1661521	40401	259089
S01005819	1321	142	1745041	20164	187582
S01005949	1634	139	2669956	19321	227126
S01001028	1764	188	3111696	35344	331632
S01002104	1852	137	3429904	18769	253724
S01004801	1855	115	3441025	13225	213325
S01004487	2031	204	4124961	41616	414324
S01002253	2173	159	4721929	25281	345507
S01000653	2174	156	4726276	24336	339144
S01004958	2253	196	5076009	38416	441588
S01002434	2307	223	5322249	49729	514461
S01004153	2646	155	7001316	24025	410130
S01005486	2654	178	7043716	31684	472412
S01002374	2902	184	8421604	33856	533968
S01005669	2923	274	8543929	75076	800902
S01005339	3120	176	9734400	30976	549120
S01005105	3286	187	10797796	34969	614482
S01000726	3485	268	12145225	71824	933980
S01001985	3559	245	12666481	60025	871955
S01002006	3674	154	13498276	23716	565796

S01003425	3733		287	13935289	82369	1071371
S01003626	4182		207	17489124	42849	865674
S01003745	4191		351	17564481	123201	1471041
S01005472	4216		214	17774656	45796	902224
S01006149	4282		212	18335524	44944	907784
S01005313	4333		233	18774889	54289	1009589
S01000418	4601		169	21169201	28561	777569
S01004998	4680		195	21902400	38025	912600
S01006055	4794		190	22982436	36100	910860
S01005371	5116		211	26173456	44521	1079476
S01006029	5297		209	28058209	43681	1107073
S01005200	5307		232	28164249	53824	1231224
S01005636	5356		240	28686736	57600	1285440
S01001091	5429		168	29474041	28224	912072
S01000369	5505		138	30305025	19044	759690
S01004897	5788		219	33500944	47961	1267572
S01001980	5978		254	35736484	64516	1518412
S01000702	6043		135	36517849	18225	815805
S01001916	6268		231	39287824	53361	1447908
S01001974	6482		206	42016324	42436	1335292
S01001735	6485		299	42055225	89401	1939015
S01000144		no result				
Totals	**163209**		**9766**	**705074147**	**2025528**	**33852578**

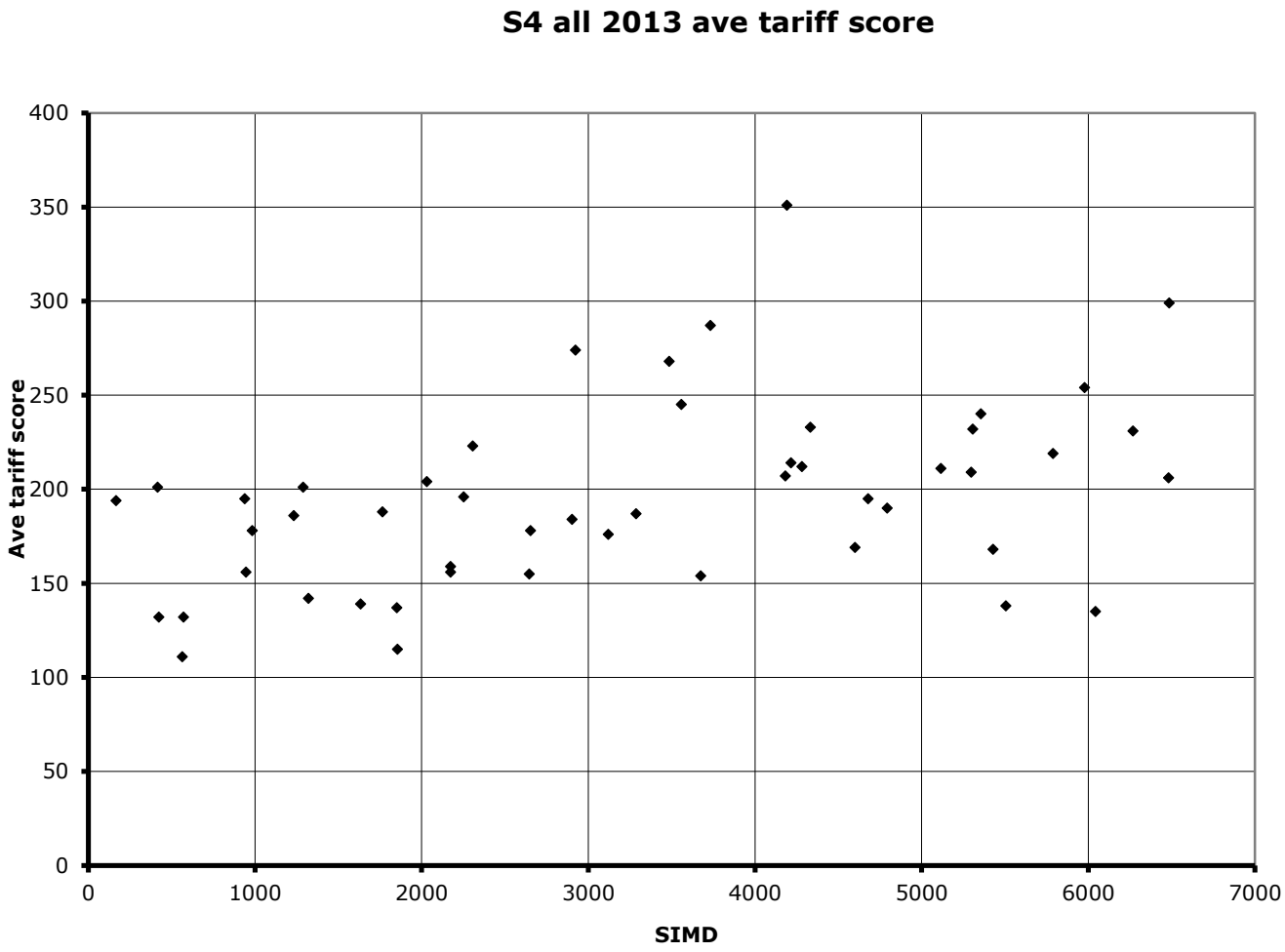
During data gathering, no result of tariff score was available for datazone S01000144, so it was determined which stratum this datazone was from and another random SIMD generated in the same stratum.

Presenting and analysing the data [2.3]

The raw data gives the scatterplot in figure 1 below.

Figure 1: Scatterplot of raw data

It can be observed that there might be weak positive correlation between SIMD and average tariff score.



The regression line of y on x and the coefficient of determination were calculated on the spreadsheet as shown in table 2 below.

Table 2 calculations

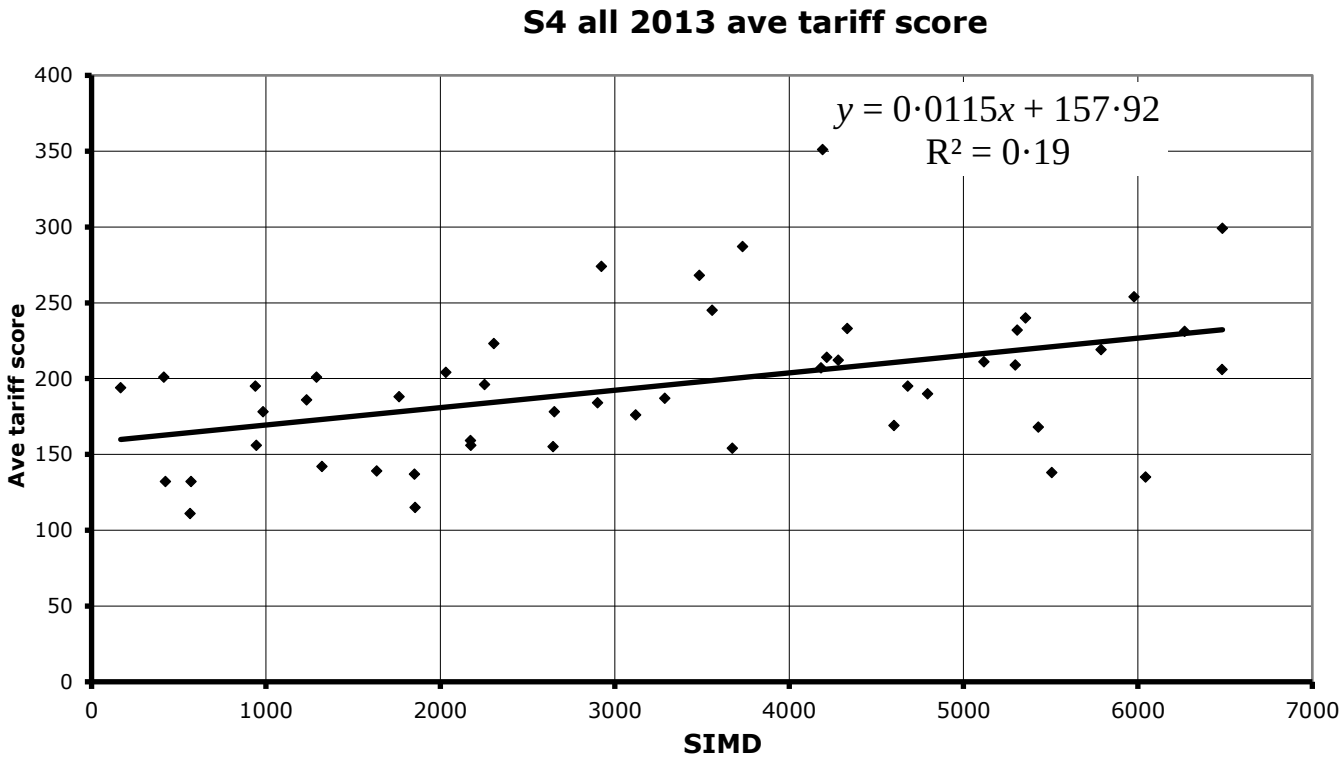
Sxx	172330593.4
Syy	118032.88
Sxy	1974596.12
n	50
x bar	3264.18
y bar	195.32
b	0.011458187
a	157.918416
R^2	0.191686343
t 50, 0.975	2.009
SSR	95407.58892
S	44.5831594
beta test	3.373853696

X	**Fitted**	**(x-x bar)^2**	**inside root**	**lower**	**upper**	
5500	220.9384431	4998891.07	1.04900757	129.202382	312.674504	**Prediction**
5500	220.9384431	4998891.07	1.02900757	130.081094	311.795792	**Confidence**

Linear regression model	$y = 157{\cdot}9 + 0{\cdot}115x$
Coefficient of determination	0·1917

These results were added to the scatterplot in figure 2 using the Excel line fitting options.

Figure 2: Scatterplot with line of best fit



The significance of these results was investigated using a *ß*-test for the slope parameter.

H_0: $ß = 0$ H_1: $ß \neq 0$
Two-tail test with $\alpha = 0{\cdot}05$

Assumption: residuals are independent and identically distributed.

$$t = \frac{b\sqrt{S_{xx}}}{s} = 3{\cdot}374 \text{ (from Table 2)}$$

$t_{48,\ 0{\cdot}975} \approx 2$ (critical value for $\nu = 48$ not available)

$3{\cdot}374 > 2$ so we can reject H_0 at the 5% level

and we have evidence that the population slope parameter is non-zero and that the model is useful for prediction.

Note that had we chosen to carry out a ρ-test instead, which yields the same value of *t*, we could have rejected the hypothesis that the population product moment correlation coefficient is zero and concluded also that there is evidence of a linear relationship between Ave Tariff Score and SIMD.

Communicating the conclusion [2.4]

The coefficient of determination of 0·1917 indicates that about 19% of the variation in average tariff score is due to SIMD rank, but that 81% of the variation is not explained by SIMD rank.

It would be most unlikely that such a complex outcome as educational performance could be explained so simply by one variable and this seems a result that fits intuition to some extent and so is not a particularly insightful conclusion.

With such weak correlation, it is unwise to use the regression equation for individual pupils. To illustrate this a 95% prediction interval is calculated for an SIMD of 5500. (See Table 2)

x	Fitted value	Lower PI	Upper PI
5500	221	129	313

So for an individual living in an area with an SIMD of 5500 they can be expected to achieve an average tariff score of between 129 and 313.

With this large sample size, the confidence interval for x = 5500 is very similar to the prediction interval as the missing term of 1/50 contributes little to the calculation (See Table 2).

With such a wide interval, the statistics offer very little that is useful numerically.

Schools are given similar data for their pupils and are expected to plan improvements from their conclusions. It is not easy to know what to do, just because pupils from more deprived backgrounds do less well in examinations.

A blanket conclusion is not useful, as illustrated by the wide prediction interval, and schools would need to consider support strategies that would be useful for individual pupils. Waiting until these results are published would seem too late, and it would be a good idea to identify pupils and strategies as soon as possible after they join a school.

Summary

The scatter diagrams and calculated statistics all point to the fact that social background (in this case multiple deprivation) has some bearing on educational outcomes but it is only one of many factors that need to be considered in planning to support pupils through their schooling.

Appendix

All data, graphs and calculations are in the Excel spreadsheet ‘Data zone vs SIMD rank sampling Mar 15’.

Document reference	Spreadsheet tab
Datazone selection	Sample selection
Table 1 Raw sample data	Sample data
Figure 1 Scattergraph of raw data	Sample graph no line
Table 2 Calculations	Sample calculations
Figure 2 Scattergraph with line of best fit	Sample graph fit

Appendix 1: Reference documents

The following reference documents will provide useful information and background.

- Assessment Arrangements (for disabled candidates and/or those with additional support needs) — various publications are available on SQA's website at: www.sqa.org.uk/sqa//14977.html.
- *Building the Curriculum 4: Skills for learning, skills for life and skills for work*
- *Building the Curriculum 5: A framework for assessment*
- Course Specification
- Design Principles for National Courses
- *Guide to Assessment*
- Principles and practice papers for curriculum areas
- *SCQF Handbook: User Guide* and SCQF level descriptors
- *SQA Skills Framework: Skills for Learning, Skills for Life and Skills for Work*
- *Skills for Learning, Skills for Life and Skills for Work: Using the Curriculum Tool*
- *Coursework Authenticity: A Guide for Teachers and Lecturers*

Administrative information

Published: August 2017 (version 2.3)

History of changes to Advanced Higher Course/Unit Support Notes

Version	Description of change	Authorised by	Date
2.0	Extensive changes to 'Further information on Course/Units' section. Exemplars included.	Qualifications Development Manager	May 2015
2.1	'Further exemplification' section: information about the purpose and function of the exemplars added.	Qualifications Development Manager	September 2015
2.2	'Further information on Course/Units' section: minor changes to wording relating to second sub-skill for Assessment Standards 1.1 and 1.3 for the Statistical Inference Unit.	Qualifications Manager	May 2016
2.3	Corrections and clarifications to the tables of skills: ♦ page 13 — formula corrected ♦ pages 14 and 15 — wording changed from 'know these standard results' to 'use these standard results'; formatting of formula corrected. ♦ page 25 —information about continuity corrections added to the learning and teaching contexts column	Qualifications Manager	August 2017