Higher National Unit Specification

General information

Unit title: Statistics for Data (SCQF level 8)

Unit code: J4Y8 35

Superclass:	CA
Publication date:	December 2020
Source:	Scottish Qualifications Authority
Version:	02

Unit purpose

This unit is intended for learners who wish to develop their knowledge of statistics as it applies to data science. This is a **specialist** unit, suitable for learners with an interest in using statistics for data analysis.

Well-developed numeracy skills and some previous knowledge of basic statistics is assumed. Learners should also possess previous experience of using numerical software, such as spreadsheet software.

This unit covers a range of descriptive and inferential statistics, explores statistical distributions including the normal distribution, and explains how to carry out hypothesis testing. The unit also looks at how linear regression can be used for forecasting. Learners will gain knowledge on the underlying principles and will learn to perform statistical calculations using appropriate software, which may include generic and/or dedicated statistical software.

After completing this unit, learners may progress to J4YA 36 *Statistics for Data* at SCQF level 9.

Outcomes

On successful completion of the unit the learner will be able to:

1   Explain statistical and probability concepts relevant to data analysis.
2   Describe statistical methods and their applications.
3   Perform statistical analyses including linear regression and hypothesis testing.

Higher National Unit Specification: General information (cont)

Unit title: Statistics for Data (SCQF level 8)

Credit points and level

1 Higher National Unit credit at Scottish Credit and Qualifications Framework (SCQF) level 8: (8 SCQF credit points at SCQF level 8)

Recommended entry to the unit

While entry is at the discretion of the centre, learners are expected to possess well-developed numeracy skills before undertaking this unit. This may be evidenced by possession of the Core Skills unit in *Numeracy* (F3GF 12) at SCQF level 6. Some knowledge of basic statistics is desirable, but not essential.

Core Skills

Achievement of this Unit gives automatic certification of the following:

Complete Core Skill   Information and Communication Technology at SCQF level 6

There are also opportunities to develop aspects of Core Skills which are highlighted in the support notes of this unit specification.

Context for delivery

If this unit is delivered as part of a group award, it is recommended that it should be taught and assessed within the subject area of the group award to which it contributes.

This unit is a suitable progression from J476 35 *Working with Data* at SCQF level 8. It builds on the analyses introduced in that unit and provides learners with a deeper understanding of statistical methods as they apply to data analysis.

Equality and inclusion

This unit specification has been designed to ensure that there are no unnecessary barriers to learning or assessment. The individual needs of learners should be taken into account when planning learning experiences, selecting assessment methods or considering alternative evidence.

Further advice can be found on our website **www.sqa.org.uk/assessmentarrangements**.

Higher National Unit Specification: Statement of standards

Unit title: Statistics for Data (SCQF level 8)

Acceptable performance in this unit will be the satisfactory achievement of the standards set out in this part of the unit specification. All sections of the statement of standards are mandatory and cannot be altered without reference to SQA.

Where evidence for outcomes is assessed on a sample basis, the whole of the content listed in the knowledge and/or skills section must be taught and available for assessment. Learners should not know in advance the items on which they will be assessed and different items should be sampled on each assessment occasion.

Outcome 1

Explain statistical and probability concepts relevant to data analysis.

Knowledge and/or skills

♦   Difference between descriptive and inferential statistics
♦   Set theory including laws of set operations
♦   Probability theory including counting rules, event combination, and conditional probability
♦   Probability distributions including uniform, normal, exponential, binomial and Poisson
♦   Distribution analysis techniques
♦   Difference between populations and samples, and parameters and statistics
♦   Impact of data science on traditional statistical approaches

Outcome 2

Describe statistical methods and their applications.

Knowledge and/or skills

♦   Sampling methods and sampling bias including the impact of small and large datasets on sampling
♦   Under- and over-sampling to balance datasets
♦   Descriptive statistics including moving average and standard deviation
♦   Correlation co-efficient and spurious correlations
♦   Graphs and charts including scatter diagrams and box plots
♦   Measures of statistical significance (confidence intervals)
♦   Computational and classical hypothesis testing
♦   Selection of statistical tests including T-tests, Z-tests and ANOVA
♦   Linear regression

Higher National Unit Specification: Statement of standards

Unit title: Statistics for Data (SCQF level 8)

Outcome 3

Perform statistical analyses including linear regression and hypothesis testing.

Knowledge and/or skills

♦ Use of statistical software
♦ Calculation of summary statistics
♦ Application of sampling techniques
♦ Creating probability distributions
♦ Analysing distribution types and features
♦ Performing a linear regression analysis
♦ Performing computational and classical hypothesis tests
♦ Illustrating relationships using graphs and charts

Evidence requirements for this unit

Learners will need to provide evidence to demonstrate their knowledge and/or skills across all outcomes. The evidence requirements for this unit will take **two** forms.

1 Knowledge evidence
2 Product evidence

Knowledge evidence relates to outcomes 1 and 2. Evidence is required for all knowledge and/or skills statements within these outcomes. The amount of evidence may be the minimum required to infer competence. The evidence may be produced over an extended period of time in lightly controlled conditions.

Knowledge evidence may be sampled when testing is used. In this case, the evidence must be produced under controlled conditions in terms of location, timing and access to reference materials. The sampling frame must cover all outcomes (1–2) but not all knowledge/skills statements; however, the majority of the knowledge/skills should be sampled (at least once) in every instance.

The knowledge evidence may be written or oral or a combination of these. Evidence may be captured, stored and presented in a range of media (including audio and video) and formats (analogue and digital).

The **product evidence** will relate to outcome 3. It will demonstrate that the learner has the competence to carry out **at least one** statistical analysis on a given dataset with the following characteristics:

♦ Demonstrates competent use of statistical software
♦ Demonstrates two or more summary statistics
♦ Demonstrates at least one sampling technique
♦ Demonstrates at least one probability distribution with an analysis of key features
♦ Demonstrates at least one linear regression analysis
♦ Demonstrates at least one hypothesis test, either computational or classical
♦ Demonstrates at least two relationships using graphs and charts

Higher National Unit Specification: Statement of standards (cont)

Unit title: Statistics for Data (SCQF level 8)

This evidence may be produced over the life of the unit, under loosely controlled conditions (including access to reference materials). Authentication will be necessary (see below).

The SCQF level of this unit (level 8) provides additional context on the nature of the required evidence and the associated standards. Appropriate level descriptors should be used when making judgements about the evidence.

When evidence is produced in loosely controlled conditions it must be authenticated. The *Guide to assessment* provides further advice on methods of authentication.

The support notes section of this specification provides specific examples of instruments of assessment that will generate the required evidence.

Higher National Unit Support Notes

Unit title: Statistics for Data (SCQF level 8)

Unit support notes are offered as guidance and are not mandatory.

While the exact time allocated to this unit is at the discretion of the centre, the notional design length is 40 hours.

Guidance on the content and context for this unit

The first part of this guidance relates to all outcomes. Subsequent parts relate to specific outcomes.

This unit is intended to give learners confidence and competence in performing descriptive statistical analysis, in addition to an understanding of the underlying probabilistic concepts. They will develop a familiarity with a chosen software for performing statistical analysis, which may include generic software, such as Excel™ (with appropriate add-ins), and/or dedicated software, such as SPSS or R.

Please note that the following guidance, relating to specific outcomes, does not seek to explain each knowledge/skills statement, which is left to the professionalism of the teacher. It seeks to clarify the statement of standards where it is potentially ambiguous. It also focuses on non-apparent teaching and learning issues that may be over-looked, or not emphasised, during unit delivery. As such, it is not representative of the relative importance of each knowledge/skill.

It is to be noted that these outcomes are not intended to be delivered as separate elements of the unit (see *Guidance on approaches to delivery*).

Outcome 1: This outcome introduces the learners to the underpinning probabilistic and statistical concepts required for all statistical data analysis and is intended to provide a solid foundation of basic ideas to build from. Statistical software is not necessarily required to understand these fundamentals. The main learning points are:

♦ understanding the individual definitions of probability vs statistics
♦ the maths behind probabilities
♦ an introduction to counting rules and how they pertain to probability theory
♦ the basics of event combination, like an event's complement, 'AND' and 'OR' combinations of events, and event exclusivity
♦ an introduction to descriptive vs inferential statistics and populations vs samples

The primary takeaway is how basic probability concepts lead to more advanced statistical concepts like inferential statistics and sampling.

Higher National Unit Support Notes (cont)

Unit title: Statistics for Data (SCQF level 8)

Outcome 2: The focus in this outcome is to introduce the learners to a variety of statistical methods that can be applied to a dataset. Tackling this outcome will introduce learners to a statistical software of choice, which will allow them to see how the calculations for the knowledge/skills are carried out and how the relevant graphs and charts are formed. Some clarification of the individual knowledge/skills are:

♦ When exploring the difference between sampling error and bias, real-world examples are helpful. There is no shortage of news articles on the problem of sampling bias in important studies.
♦ When introducing the correlation coefficient calculation, an emphasis that correlation does not imply causation is essential. Real-life examples of spurious statistics can be very helpful.
♦ When exploring distribution analysis, an emphasis on the importance of the normal distribution with real-life examples will help to solidify its usefulness.
♦ When introducing the most common graphs and charts used in statistical analysis, box plots are both particularly important and particularly unintuitive. It is important to ensure learners grasp the basics so that they learn to quickly review data for potential sources of difficulty like outliers and skew.
♦ At the time of writing, computational hypothesis testing is the most commonly used in the field of data analysis. However, the theory behind classical methods is important, and a good understanding of both, computational and classical methods, is essential.

Outcome 3: The goal for this outcome is to allow learners the chance to exercise the concepts learned in outcomes 1 and 2. This is where learners will gain extensive familiarity with the statistical software of choice as they apply statistical methods to given datasets.

Guidance on approaches to delivery of this unit

The popularity of data science has resulted in a wide range of resources for those wishing to learn more about statistics as it pertains to data analysis. Useful online resources can be found at:

♦ https://towardsdatascience.com
♦ https://www.datacamp.com
♦ https://www.datasciencecentral.com

Datasets for practical work can be found on Kaggle:

♦ https://www.kaggle.com/datasets

A suggested distribution of time, across the outcomes, is:

Outcome 1: 10 hours
Outcome 2: 15 hours
Outcome 3: 15 hours

Higher National Unit Support Notes (cont)

Unit title: Statistics for Data (SCQF level 8)

The learning in this unit should be treated in a holistic manner to develop the learners' understanding and skills in tandem. In particular, the knowledge and skills in outcomes 1 and 2 should be acquired by the study of particular situations and datasets, and the implementation of methods on those datasets. Outcome 3 deals with the ultimate production of statistical analysis, but can be explored in the introduction of each concept and method.

Summative assessment may be carried out at any time. However, when testing is used (see evidence requirements) it is recommended that this is carried out towards the end of the unit (but with sufficient time for remediation and re-assessment). When continuous assessment is used (such as the use of a web log), this could commence early in the life of the unit and be carried out throughout the duration of the unit.

There are opportunities to carry out formative assessment at various stages in the unit. For example, formative assessment could be carried out on the completion of each outcome to ensure that learners have grasped the knowledge contained within it. This would provide assessors with an opportunity to diagnose misconceptions and intervene to remedy them before progressing to the next outcome.

Further advice on instruments of assessment is included in the section below.

Guidance on approaches to assessment of this unit

Evidence can be generated using different types of assessment. The following are suggestions only. There may be other methods that would be more suitable to learners.

Centres are reminded that prior verification of centre-devised assessments would help to ensure that the national standard is being met. Where learners experience a range of assessment methods, this helps them to develop different skills that should be transferable to work or further and higher education.

Assessment could be carried out using:

♦ a selected response test that covers the knowledge and understanding for outcomes 1 and 2
♦ a set of practical tasks that cover the practical competence and understanding for outcome 3

Each selected response question could be structured as four options (one key) with a pass mark of 60% for the whole test. Use should be made of situational questions to assess the learner's competency in distinguishing populations and samples, combination types, descriptive and inferential statistics, probability properties, and appropriate method selection. The test could consist of a relatively high number of questions (30 or 40, for example), lasting an hour, which would cover outcomes 1 and 2 and sample all of the knowledge and skill statements (including at least one question for each statement).

The practical tasks could be carried out over an extended period of time. They would allow the learner to demonstrate competence in applying the analytical methods taught to problem datasets. Competence in displaying the resulting analysis through graphs and charts would also be evidenced as part of the practical tasks. The set of practical tasks must cover all of the practical competences set out in outcome 3.

Higher National Unit Support Notes (cont)

Unit title: Statistics for Data (SCQF level 8)

A more contemporary approach to assessment would involve the use of a web log (blog) to record learning (and the associated activities) throughout the life of the unit. The blog could provide the knowledge evidence (in the descriptions and explanations). The blog should be assessed using defined criteria to permit a correct judgement about the quality of the digital evidence. In this approach to assessment, every knowledge and skill must be evidenced; sampling would not be appropriate.

Formative assessment could be used to assess learners' knowledge at various stages throughout the life of the unit. An ideal time to gauge their knowledge would be at the end of each outcome. This assessment could be delivered through an item bank of selected response questions, providing diagnostic feedback to learners (when appropriate).

If a blog is used for summative assessment, it would also facilitate formative assessment since learning (including misconceptions) would be apparent from the blog, and intervention could take place to correct misunderstandings on an on-going basis.

It is important to ensure that work submitted by a learner is their own. The risk of malpractice is greater when you do not have the opportunity to observe learners carrying out assessment activities. There are various web-based services that can detect plagiarism, but the following strategies can also be effective in authenticating learners' work:

♦   questioning
♦   write-ups under controlled conditions
♦   witness testimony
♦   use of personal logs
♦   personal statements produced by your learners

The use of case studies which require learners to include information from their own experience can also help to reduce plagiarism. You should ensure that learners are clear about how to access resources, especially from the internet, how to reference the material they use, and the extent to which they may confer with others or seek support.

Opportunities for e-assessment

E-assessment may be appropriate for some assessments in this unit. By e-assessment we mean assessment which is supported by Information and Communication Technology (ICT), such as e-testing or the use of e-portfolios or social software. Centres which wish to use e-assessment must ensure that the national standard is applied to all learner evidence and that conditions of assessment as specified in the evidence requirements are met, regardless of the mode of gathering evidence. The most up-to-date guidance on the use of e-assessment to support SQA's qualifications is available at **www.sqa.org.uk/e-assessment**.

Higher National Unit Support Notes (cont)

Unit title: Statistics for Data (SCQF level 8)

Opportunities for developing Core and other essential skills

The unit provides opportunities to develop some of the following Core Skills:

♦ *Information and Communication Technology (ICT)* at SCQF level 6
♦ *Problem Solving* at SCQF level 6
♦ *Numeracy* at SCQF level 6

Learners are expected to make use of software for performing statistical analysis, which may include generic software, such as Excel™ (with appropriate add-ins), and/or dedicated software, such as SPSS or R. They will address several components of the Core Skill in *Information and Communication Technology (ICT)* in so doing.

They will also cover components of the *Problem Solving* Core Skill while carrying out the statistical analysis.

Learners will have the opportunity to develop aspects of the *Numeracy* Core Skill through learning the maths behind probability and counting rules related to probability theory, among other concepts.

The Core Skill of Information and Communication Technology at SCQF level 6 is embedded in this unit. When a learner achieves the unit, their Core Skills profile will also be updated to include this Core Skill.

History of changes to unit

Version	Description of change	Date
02	Core Skill Information and Communication Technology at SCQF level 6 embedded.	09/12/20

General information for learners

Unit title: Statistics for Data (SCQF level 8)

This section will help you decide whether this is the unit for you by explaining what the unit is about, what you should know or be able to do before you start, what you will need to do during the unit and opportunities for further learning and employment.

The purpose of this unit is to introduce you to the probabilistic and statistical foundations behind statistics-driven data analysis and to develop your competence in using software to perform statistical analysis. You should possess well-developed numeracy skills before undertaking this unit.

You will gain an understanding of probability core definitions, how to make statistical inferences, and how to determine when to make those inferences. The statistical methods you will learn include:

♦ Sampling methods
♦ Descriptive statistics
♦ Correlations
♦ Distributions
♦ Statistically relevant graphs and charts
♦ Measures of statistical significance
♦ Hypothesis testing
♦ Linear regression

You will practise these methods by producing statistical analysis on sample datasets. In order to do this, you will develop competence in the use of a selected software for performing statistical analysis.

The assessments for this unit will give you the opportunity to evidence your knowledge and understanding of the various statistical and probabilistic methods and concepts. There will also be practical assignments that allow you to evidence your competence in developing statistical analysis.

While undertaking this unit, you will have opportunities to develop aspects of the SCQF level 6 *Numeracy, Problem Solving* and *Information and Communication Technology (ICT)* Core Skills.

Once you have completed this unit, you may progress to J4YA 36 *Statistics for Data* at SCQF level 9.

The Core Skill of Information and Communication Technology SCQF at level 6 is embedded in this unit. When a learner achieves the unit, their Core Skills profile will also be updated to include this Core Skill.