National Unit Specification

General information

Unit title: Data Science: Statistics (SCQF level 6)

Unit code: J2G8 46

Superclass:	RB
Publication date:	July 2019
Source:	Scottish Qualifications Authority
Version:	02 (August 2019)

Unit purpose

The purpose of this unit is to develop learners' knowledge of statistics as they relate to data science. The unit builds on the statistical concepts introduced in *Data Science: Statistics* at SCQF level 5.

This **specialist unit** is intended for learners with a vocational or academic interest in STEM, particularly computer science and data science. Learners should possess some knowledge of statistics before commencing this unit, which may be evidenced by possession of J2G8 45 *Data Science: Statistics* at SCQF level 5 or equivalent.

The unit explains statistical concepts and theorems that are important in data science including hypothesis testing and Bayes' Theorem. It prepares learners for carrying out a statistical study and then shows learners how to carry out the study using contemporary data analysis tools.

At the completion of this unit, learners will be appreciate statistical concepts that are important in data science, and be able to apply these concepts to a practical statistical study. Learners may wish to complete other units in this field, such as J2G7 46 *Machine Learning* at SCQF level 6.

Outcomes

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

1 Explain statistical methods and theorems as they relate to data science.
2 Explain the factors contributing to a statistical study within the framework of a data science project.
3 Carry out a statistical study with the aim of contributing to a data science project.

National Unit Specification: General information (cont)

Unit title: Data Science: Statistics (SCQF level 6)

Credit points and level

1 National Unit credit at SCQF level 6: (6 SCQF credit points at SCQF level 6)

Recommended entry to the unit

Learners will require numeracy skills before attempting this unit, which could be evidenced by possession of the Core Skill in *Numeracy* at SCQF level 6. It is desirable, but not required, that they also have knowledge of statistical methods at SCQF level 5 including a broad familiarity with descriptive statistics.

Core Skills

Achievement of this Unit gives automatic certification of the following:

Complete Core Skill	Numeracy at SCQF level 6

Achievement of this Unit gives automatic certification of the following Core Skills components:

Core Skill components	Providing/Creating Information at SCQF level 5 Critical Thinking at SCQF level 5

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.

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. For example, if this unit is delivered as part of the National Progression Award in Data Science at SCQF level 6, there is overlap with other units within this award (particularly J2G4 46 *Data Science*) and there will be opportunities to contextualise and integrate teaching, learning and assessment across component units.

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**.

National Unit Specification: Statement of standards

Unit title: Data Science: Statistics (SCQF level 6)

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 methods and theorems as they relate to data science.

Performance criteria

(a) Explain sampling theory and sampling methods.
(b) Explain probability, including conditional probability, and the significance of probability in data science.
(c) Explain the normal and the standard normal distributions and a range of descriptive statistics relating to these distributions.
(d) Explain the central limit theorem and its application in data science.
(e) Explain linear regression and Bayes' Theorum and their application to data science.
(f) Explain hypothesis testing and its relevance to data science.

Outcome 2

Explain the factors contributing to a statistical study within the framework of a data science project.

Performance criteria

(a) Explain statistical study planning.
(b) Explain common flaws in statistical study design.
(c) Explain selection of sampling methods.
(d) Explain selection of statistical methods for analysing datasets.
(e) Explain selection of statistical methods for comparing two datasets.
(f) Explain the selection of data visualisations for different types of datasets.

Outcome 3

Carry out a statistical study with the aim of contributing to a data science project.

Performance criteria

(a) Justify the sampling frame.
(b) Justify the selection of descriptive and inferential statistics.
(c) Justify the selection of data visualisations.
(d) Derive descriptive and inferential statistics.
(e) Create data visualisations of study result.

National Unit Specification: Statement of standards (cont)

Unit title: Data Science: Statistics (SCQF level 6)

Evidence requirements for this unit

Learners will need to provide evidence to demonstrate the performance criteria across all outcomes. The evidence requirements for this unit will take **one** form: product evidence.

The **product evidence** will relate to all outcomes. It will take the form of a statistical study. The study must involve the analysis of two datasets, comprising at least 10,000 data items (each). It must use at least two of the methods defined in Outcome 1. For example, the study may involve descriptive statistics and linear regression.

The following evidence must be produced by each learner.

1 Study plan including sample method(s).
2 Potential weaknesses in the plan.
3 Statistical analysis including descriptive and inferential statistics.
4 Comparison between datasets.
5 Visualisations to compare and summarise datasets.
6 Justifications of sampling frame, descriptive and inferential statistics, and visualisations.

The evidence must be produced by the learner, without assistance. The analysis may be done in lightly controlled conditions, over an extended period of time, at times and places at the discretion of the learner.

The SCQF level of this unit (level 6) 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.

National Unit Support Notes

Unit title: Data Science: Statistics (SCQF level 6)

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 purpose of this unit is to introduce learners to the statistics that underpin data science.

It is recommended that learners possess knowledge of basic descriptive statistics before commencing this unit.

This unit may be undertaken alone or as part of the National Progression Award in Data Science at SCQF level 6, in which case it builds on the statistics contained within J2G4 46 *Data Science* at SCQF level 6.

This unit has three outcomes. Outcome 1 relates to statistical methods and theorums; Outcome 2 relates to statistical studies; and Outcome 3 involves using data analysis tools, such as Microsoft Excel™, to carry out a statistical study.

Please note that the following guidance does not seek to explain each performance criterion. This section 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 delivery. As such, it is not representative of the actual time spent teaching or learning specific competences or the relative importance of each competence.

Outcome 1: This outcome relates to statistical methods and theorems that are important in data science such as linear regression and Bayes' Theorum.

The performance criteria are self-explanatory. Note that it is expected that learners can calculate the various statistics in this outcome as part of their explanations. However, the level of treatment of the more complex techniques (such as Bayes' Theorum and hypothesis testing) may be relatively light.

Outcome 2: This outcome explains how to carry out a statistical study. The performance criteria are relatively self-explanatory. An important part of this outcome is learner's knowledge of visualisations (Performance Criterion (f)) and being able to select appropriate visualisations for a variety of different datasets.

Outcome 3: This outcome requires learners to carry out a statistical study. The performance criteria are self-explanatory. The study should be relatively large and comprise a significant statistical analysis, involving a range of descriptive and inferential statistics, including a number of visualisations.

National Unit Support Notes (cont)

Unit title: Data Science: Statistics (SCQF level 6)

Guidance on approaches to delivery of this unit

There are three outcomes in this unit. It is recommended that Outcome 1 is taught before Outcome 2 and Outcome 3, which should be taught together (since they both relate to statistical studies). A possible distribution of time is:

- Outcome 1: 20 hours
- Outcome 2 and 3: 20 hours

Teacher exposition will be required in Outcome 1, when a range of statistical methods and theorems are introduced. There is a lot of content in Outcome 1. It is important that each statistical method and theorem is related to data science and not explained in isolation. For example, Big Data has had a significant impact on sampling because of the availability of very large datasets, which has significant implications for sampling methods.

Outcome 2 and Outcome 3 involve the use of software tools to carry out a statistical study. This is, obviously, best done through hands-on practice with software products. A range of software products could be used such as Microsoft Excel™ and Tableau™. Learners should practice with (relatively) large datasets (datasets with at least 10,000 records/examples). Ideally, this should be real data, relating to topics of interest to learners.

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 for this unit could take the form of a **practical assignment**, which would require learners to carry out a statistical study. The study could be supplied to learners or they could be permitted to select their own study. The study must use at least two of the statistical methods defined in Outcome 1, which includes:

- Sampling
- Probability
- Descriptive statistics
- Linear regression
- Bayes' Theorum
- Hypothesis testing

Ideally the study should involve several of these statistical techniques. For example, a study might involve sampling, descriptive statistics (for historical purposes) and linear regression (for prediction purposes).

Unit title: Data Science: Statistics (SCQF level 6)

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.

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**.

Opportunities for developing Core and other essential skills

There are opportunities in this unit to develop Core Skills.

This Unit has the Core Skill of Numeracy at SCQF level 6 embedded. When a learner achieves the unit, their Core Skills profile will also be updated to include this Core Skill.

The Providing/Creating Information component of Information and Communication Technology at SCQF level 5 and the Critical Thinking component of Problem Solving at SCQF level 5 is embedded in this unit. When a learner achieves these units, their Core Skills profile will also be updated to include these components.

The unit is particularly well suited to developing the Core Skills of *Numeracy* and *Information and Communication Technology (ICT)*. *ICT* skills will be used throughout the unit, particularly Outcome 3. *Numeracy* skills will be developed in all outcomes.

History of changes to unit

Version	Description of change	Date
03	Coding change due to hierarchy	23/08/19
02	Core Skill of Numeracy at SCQF level 6 embedded. Core Skills Component Providing/Creating Information at SCQF level 5 embedded and the Core Skills Component Critical Thinking at SCQF level 5 embedded.	16/08/19

Additional copies of this unit specification can be purchased from the Scottish Qualifications Authority. Please contact the Business Development and Customer Support team, telephone 0303 333 0330.

General information for learners

Unit title: Data Science: Statistics (SCQF level 6)

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.

This unit will develop your knowledge of statistics and give you experience in carrying out a large statistical study. You should have some experience of statistics before attempting this unit or, at least, well developed numeracy skills.

Data science is becoming very important. Data science is the process of exploring large amounts of data to identify patterns and trends and make predictions. For example, data science is used to discover cancers, find new planets and predict crime. Statistics is a fundamental part of data science.

This unit introduces you to the statistics that underpin data science. There are three parts to this unit.

1 Statistical methods and theorems that are important in data science.
2 How to carry out a statistical study.
3 Carrying out a statistical study in a data science context.

You will be introduced to statistical techniques such as conditional probability, linear regression and Bayes Theorem, all of which are important in data science.

You will be shown how to design a statistical study, including what to avoid, and how to select samples and carry out an analysis on datasets, including the creation of data visualisations.

Lastly, you will have the opportunity to apply this knowledge to a real statistical study using data analysis software such as Microsoft Excel™ and Tableau™.

The assessment of this unit will involve carrying out a study by yourself, using a variety of data analysis tools.

When you complete this unit you could learn more about statistics in data science by doing more advanced units in this area at SCQF level 7 and above.

This Unit has the Core Skill of Numeracy at SCQF level 6 embedded. When a learner achieves the unit, their Core Skills profile will also be updated to include this Core Skill.

The Providing/Creating Information component of Information and Communication Technology at SCQF level 5 and the Critical Thinking component of Problem Solving at SCQF level 5 is embedded in this unit. When a learner achieves these units, their Core Skills profile will also be updated to include these components.

The unit is particularly well suited to developing the Core Skills of *Numeracy* and *Information and Communication Technology (ICT)*. *ICT* skills will be used throughout the unit, particularly Outcome 3. *Numeracy* skills will be developed in all outcomes