National Unit Specification

General information

Unit title: Data Science: Statistics (SCQF level 5)

Unit code: J2G8 45

Superclass: RB

Publication date: July 2019

Source: Scottish Qualifications Authority

Version: 02, August 2019

Unit purpose

The purpose of this unit is to introduce learners to the fundamental statistical concepts required in the field of data science.

This **specialist** unit is intended for learners with a vocational or academic interest in STEM, particularly computer science and data science. The unit is not a general introduction to statistics; it is an introduction to the statistics required in data science. No previous knowledge of statistics is required but learners must be sufficiently numerate before attempting the outcomes, which might be demonstrated by possession of the Core Skill in *Numeracy* at SCQF level 5.

The unit introduces basic statistical methods that are fundamental to data science, and applies that knowledge using simple data analysis tools. Although the focus is statistics as it relates to data science, general statistical techniques are introduced when these underpin more specialist knowledge. Learners will also gain practical skills in applying their knowledge to realistic problems using contemporary data analysis tools.

At the completion of this unit, learners will understand the statistical foundations of data science. Learners may progress to more advanced units in this field such as *Data Science: Statistics* at SCQF level 6.

Outcomes

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

1 Describe statistical methods as they relate to data science.
2 Describe dataset features and tools for analysing datasets.
3 Carry out statistical calculations on datasets using data analysis tools.

National Unit Specification: General information (cont)

Unit title: Data Science: Statistics (SCQF level 5)

Credit points and level

1 National Unit credit at SCQF level 5: (6 SCQF credit points at SCQF level 5)

Recommended entry to the unit

Learners will require numeracy skills before attempting this unit, which may be evidenced by possession of the Core Skill unit in *Numeracy* at SCQF level 5 or equivalent. No previous knowledge of statistics or data science is required.

Core Skills

Achievement of this Unit gives automatic certification of the following:

Complete Core Skill	Numeracy at SCQF level 5

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

Core Skill components	Providing/Creating Information at SCQF level 4 Critical Thinking at SCQF level 4

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. For example, if this unit is delivered as part of the National Progression Award in Data Science at SCQF level 5 there is overlap with other units within this award (particularly J2G3 45 *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 5)

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

Describe statistical methods as they relate to data science.

Performance criteria

(a) Distinguish between descriptive and inferential statistics.
(b) Describe sampling methods and the significance of sampling in data science.
(c) Describe probability and the significance of probability in data science.
(d) Describe the characteristics of the normal distribution.
(e) Describe common descriptive statistical features of the normal distribution including standard deviation.
(f) Describes methods of calculating trends in datasets.

Outcome 2

Describe dataset features and tools for analysing datasets.

Performance criteria

(a) Describe quantitative and qualitative data.
(b) Describe continuous and discrete data, and structured and unstructured data.
(c) Describe the coefficient of correlation between datasets.
(d) Describe the tools used to analyse datasets.
(e) Describe the selection of visualisation for specific types of data.

Outcome 3

Carry out statistical calculations on datasets using data analysis tools.

Performance criteria

(a) Derive a range of descriptive statistics from the datasets.
(b) Derive correlation co-efficient for the datasets.
(c) Derive visualisations of the datasets to show patterns and trends.
(d) Compare datasets in terms of their descriptive statistics and correlation co-efficient.

National Unit Specification: Statement of standards (cont)

Unit title: Data Science: Statistics (SCQF level 5)

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 **two** forms.

1 Knowledge evidence.
2 Product evidence.

The **knowledge evidence** will relate to Outcome 1 and Outcome 2. The knowledge evidence may be written or oral or a combination of these. The amount of evidence may be the minimum required to infer competence across both outcomes. For example, in Outcome 1, only the most common sampling methods need be described (Performance Criterion (b)); in Outcome 2, only the main tools used to analyse large datasets need be described (Performance Criterion (d)). The descriptions (for both outcomes) must include examples (for every performance criterion). In the case of descriptive statistics (Outcome 1, Performance Criterion (e)) at least one worked example of each statistic must be provided. The descriptive statistics must include standard deviation.

The knowledge evidence may be sampled when testing is used. Testing must be carried out under supervised conditions and it must be controlled in terms of location and time. Access to reference material is not permitted. The sampling frame, on all occasions, must include Outcome 1 and Outcome 2 (but not every performance criterion within each outcome). The sampling frame must always include Outcome 1, Performance Criteria (e); competence in this performance criterion may be inferred from correct statistical calculations (without providing separate descriptions). The testing of Performance Criterion (e) must always include standard deviation.

The product evidence will relate to Outcome 3. The product evidence will take the form of a completed statistical analysis of at least two datasets using data analysis tools. The datasets will be supplied to the learner, and must comprise at least 500 data items (each). The analysis must involve the comparison of two datasets in terms of their descriptive statistics and correlation co-efficient. At least three visualisations must be created, at least one of which must compare the datasets.

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 5) 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 5)

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 basic statistics that underpin data science. No previous knowledge of statistics is required but learners are presumed to be numerically competent.

This unit may be undertaken alone or as part of the National Progression Award in Data Science at SCQF level 5, in which case it builds on the statistics contained within J2G3 45 *Data Science* at SCQF level 5.

This unit has three outcomes. Outcome 1 relates to basic statistics; Outcome 2 relates to datasets; and Outcome 3 involves using data analysis tools, such as Microsoft Excel™, to carry out calculations.

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 covers basic statistics but links each statistical technique to data science. For example, when discussing sampling methods (Performance Criterion (b)), it should be emphasised how data science permits very large samples to be collected and analysed. The depth of treatment of each topic will be light. For example, the treatment of probability (Performance Criterion (c)) will cover basic probability — not conditional probability, which is covered at level 6. The performance criteria are self-explanatory and require no further explanation.

Outcome 2: This outcome introduces datasets to learners. The performance criteria are relatively self-explanatory. Learners may be aware of some tools (performance criterion d) that can be used to analyse large datasets (such as Microsoft Excel™) but not more specialised tools. An important part of this outcome is the types of visualisations (Performance Criterion (e)). Learners should be introduced to a wide range of contemporary visualisations than can be used to summarise different datasets.

Outcome 3: This outcome applies the knowledge and skills acquired during Outcome 1 and Outcome 2 to using data analysis tools to carry out statistical calculations. For example, learners could use Microsoft Excel™ to work out a range of descriptive statistics, evaluate correlations and create visualisations to compare two datasets.

National Unit Support Notes (cont)

Unit title: Data Science: Statistics (SCQF level 5)

Guidance on approaches to delivery of this unit

There are three outcomes in this unit. It is recommended that the outcomes are taught in sequence (Outcome 1, Outcome 2 and Outcome 3). A possible distribution of time is:

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

The delivery of Outcome 2 and Outcome 3 could be combined, providing 25 hours of teaching and learning, since they are closely related (Outcome 2 relates to datasets and tools, and Outcome 3 applies this knowledge).

Teacher exposition will be required in Outcome 1, when a range of statistical methods are introduced. It is recommended that learners practice statistical calculations once each method has been taught. It is important that methods are not taught in isolation, and that they are related to real life. For example, when teaching standard deviation, it is important that learners can not only calculate standard deviation but also appreciate the real-world significance of the statistic (68-95-99.7 rule).

Outcome 2 and Outcome 3 involve the use of software tools to analyse datasets. 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 Google Sheets™. Learners should practice with (relatively) large datasets (datasets with at least 500 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.

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, this could commence early in the unit and be carried out throughout the life of the unit.

A wide range of instruments of assessment could be used to satisfy the evidence requirements.

A traditional approach to assessment could involve the use of a selected response test for knowledge evidence and a practical assignment for product evidence. The selected response test could comprise a multiple choice test of learners' knowledge of Outcome 1 and Outcome 2. The test would sample from the knowledge domain (Outcome 1 and Outcome 2). An appropriate pass mark would be set. The practical assignment would require learners to use software to work out a range of statistics, and create associated visualisations, for two or more datasets. A checklist could be used to assess the calculations.

More contemporary approaches to assessment include the creation of a web log or portfolio. The web log (blog) would record learning over the life of the unit. The blog would record, on a daily or weekly basis, the learning and activities undertaken by each learner. Practical work could be captured in the blog by linking specific post(s) to examples of statistics calculated by the learner. The completed blog would have to satisfy all performance criteria. Alternatively, a portfolio could be used as a repository for the descriptions, calculations and comparisons required in all three outcomes. The completed portfolio would have to satisfy all performance criteria.

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.

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 Outcome 1 and Outcome 2, when learners are introduced to descriptive statistics and visualisations.

This Unit has the Core Skill of Numeracy at SCQF level 5 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 4 is embedded in this unit and the Critical Thinking component of Problem Solving at SCQF level 4 is embedded in this unit. When a learner achieves these units, their Core Skills profile will also be updated to include these components.

History of changes to unit

Version	Description of change	Date
02	Core Skill of Numeracy at SCQF level 5 embedded. Core Skills Component Providing/Creating Information at SCQF level 4 embedded and the Core Skills Component Critical Thinking at SCQF level 4 embedded.	16/08/19

General information for learners

Unit title: Data Science: Statistics (SCQF level 5)

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 provide a basic introduction to the statistics used in data science. No previous knowledge or experience of statistics is required but you are presumed to possess numeracy skills before attempting this unit.

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 basic statistics that underpin data science. There are three parts to this unit.

1 Basic statistics.
2 Introduction to datasets.
3 Using software to calculate statistics.

The basic statistics introduces you to the fundamental statistics involved in data science such as measures of central tendency and measures of dispersion. You will also learn how to identify patterns and trends in data.

The introduction to datasets looks at the characteristics of data and the tools that you can use to analyse data. These tools include familiar software products such as Excel™ and less familiar ones such as Tableau™.

The final part of the unit looks at how to use software to calculate statistics. You will use software to work out statistics, find correlations between datasets and create data visualisations.

The assessment of this unit might involve a test of your knowledge and a practical assignment. Most of your time will be spent learning about statistics. Assessment will not take much time.

When you complete this unit you could learn more about statistics in data science by doing more advanced units in this subject area such as *Data Science: Statistics* at SCQF level 6.

This Unit has the Core Skill of Numeracy at SCQF level 5 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 4 is embedded in this unit and the Critical Thinking component of Problem Solving at SCQF level 4 is embedded in this unit. When a learner achieves these units, their Core Skills profile will also be updated to include these components.