# Higher National Unit Specification

## General information

## Unit title: Statistics for Data (SCQF level 9)

## Unit code: J4YA 36

**Superclass:** CA

**Publication date:** September 2020

**Source:** Scottish Qualifications Authority

**Version:** 01

## Unit purpose

The purpose of this unit is to allow learners build upon their previous statistical knowledge and how it relates to data science. This unit is intended for learners who possess knowledge of statistics for data (for example after having completed J4YA 35 *Statistics for Data* at SCQF level 8) and would like a deeper understanding of the statistical tools that can be used to extract insights from data.

This is a **specialist** unit, suitable for learners with an interest in using advanced statistical techniques to analyse data. Learners should have comfortable working knowledge of chosen data analysis software, data manipulation skills, and an understanding of core statistical concepts.

This unit covers a range of statistical methods and techniques as they relate to data science, including: multivariate and logistic regression, classification techniques, discrete probability distributions and Bayes' Theorem. Learners will also have opportunities to carry out regression analyses using statistical software.

At the completion of this unit, learners may progress to J4YD 36 *Machine Learning* at SCQF level 9, where the statistical learning techniques gained in this unit can be applied to machine learning.

## Higher National Unit Specification: General information (cont)

**Unit title:** Statistics for Data (SCQF level 9)

## Outcomes

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

1 Describe the characteristics of multivariate regression models.
2 Explain fundamental probability and statistical concepts for advanced data analysis.
3 Perform multivariate regression analysis and classifications, including manual model development and model analysis.

## Credit points and level

1 Higher National Unit credit at Scottish Credit and Qualifications Framework (SCQF) level 9: (8 SCQF credit points at SCQF level 9)

## Recommended entry to the unit

While entry is at the discretion of the centre, learners should possess statistical knowledge before attempting this unit. A familiarity with descriptive statistics, probability and statistical software is assumed. This may be evidenced by completion of *Statistics for Data* (SCQF level 8).

## Core Skills

Opportunities to develop aspects of Core Skills are highlighted in the support notes for this unit specification.

There is no automatic certification of Core Skills or Core Skill components in this unit.

## 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.

The context for this unit is applicable statistics for data science, rather than the broader context of general statistics. Although the unit will teach the underlying principles, it is expected that statistical calculations will be done using appropriate software, which may include generic software, such as Excel™ (with appropriate add-ins), and/or dedicated software, such as SPSS.

## 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 9)

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 the characteristics of multivariate regression models.

### Knowledge and/or skills

- Multivariate linear, logistic and polynomial regression
- Evaluation of regression model in terms of fit
- Knowledge of applicable function for given data characteristics

## Outcome 2

Explain fundamental probability and statistical concepts for advanced data analysis.

### Knowledge and/or skills

- Discrete probability distributions including Bernoulli, Binomial and Poisson distributions
- Probability distribution functions including PMF, PDF and CDF
- Difference between Bayesian and probabilistic inference
- Re-sampling methods including bootstrapping
- Classification using tree-based methods
- Statistical foundations of machine learning

## Outcome 3

Perform multivariate regression analysis and classifications, including manual model development and model analysis.

### Knowledge and/or skills

- Carrying out regression analyses
- Performing classifications
- Selection of applicable function for given data
- Analysis of manually developed model, including relative importance of predictors and interpreting effects.
- Analysis of logistic regression model, including development of confusion matrix, ROC curve development and analysis, and application of cross validation

## Unit title: Statistics for Data (SCQF level 9)

### 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 test.

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 use statistical software to choose the applicable function for given data and manually build **at least one** multivariate regression model for each of the following types:

- linear
- logistic
- polynomial.

Each of the multivariate regression models built by the learner must demonstrate that the learner can:

- perform **at least one** method of analysis for each of the three model types. Example methods of analysis are: relative importance of predictors, interpreting effects, development of confusion matrix, ROC curve development and analysis, application of cross validation, etc (as applicable).
- build **at least one** classification tree model.

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 9)

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 utilising advanced statistical tools for data analysis, in addition to an understanding of the underlying probabilistic and statistical concepts. Learners should already have 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, and they will gain even more competence through completion of this unit.

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 various types of multivariate regression analyses and their uses. The most straightforward progression would be to review simple linear regression, build upon it by adding additional predictors, explore manual model building, and then introduce logistic and polynomial regression. A review of exponentials and logarithms may be necessary before introducing logistic regression. Some key learning points are:

- using plots to interpret and investigate the effects of interactions between predictors
- demonstrating manual regression model development through an iterative process
- investigating real-world applications of the regression types
- explore regression model evaluation techniques

## Higher National Unit Support Notes (cont)

## Unit title: Statistics for Data (SCQF level 9)

**Outcome 2:** The focus in this outcome is to introduce the learners to the fundamental statistical and probabilistic concepts underpinning advanced data analysis techniques. At the time of writing, probabilistic inference is more common than Bayesian in the field of data analysis. However, Bayesian models are used quite often for machine learning, and it is important to have a good understanding of both.

**Outcome 3:** The goal for this outcome is to allow learners the opportunity to exercise the tools and methods outlined in outcomes 1 and 2. It is particularly important for learners to experience the process of evaluating their models.

## 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

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 focuses on the manual development of models utilising the methods learned in earlier outcomes, 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.

## Higher National Unit Support Notes (cont)

## Unit title: Statistics for Data (SCQF level 9)

## 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 appropriate use cases for different methods and models. 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 the majority of the knowledge and skill statements (including at least one question in every instance).

The practical tasks could be carried out over an extended period of time. They would allow the learner to demonstrate competence in applying the manual modelling techniques and analysis tools learned in outcomes 1 and 2 to problem datasets. The set of practical tasks must cover all of the practical competences set out in outcome 3.

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 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**.

## Opportunities for developing Core and other essential skills

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

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

Learners are expected to explain fundamental probability and statistical concepts for advanced data analysis and describe the characteristics of multivariate regression models. In so doing, they will be covering aspects of the *Communication* Core Skill.

Furthermore, 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 manual model development and model analysis.

Learners will have the opportunity to demonstrate aspects of the *Numeracy* Core Skill through performing statistical calculations while undertaking this unit.

## History of changes to unit

| Version | Description of change | Date |
|---|---|---|
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |



Unit template: June 2017

## General information for learners

## Unit title: Statistics for Data (SCQF level 9)

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 build upon your previous statistical knowledge to provide a deeper understanding of the statistical tools that can be used to extract insights from data. You should have familiarity with descriptive statistics, probability, and a chosen statistical software through the successful completion of J4Y8 35 *Statistics for Data* at SCQF level 8 or equivalent.

Although the unit will teach the underlying principles, it is expected that you will perform statistical calculations using appropriate software. You will be introduced to multivariate linear, logistic, and polynomial regression, and the tools needed to analyse these regressions. You will gain an understanding of advanced statistical and probabilistic concepts needed for data analysis including:

- Discrete probability distributions
- Probability distribution functions
- Bayesian and probabilistic inference
- Re-sampling methods
- Bootstrapping
- Tree-based classification

You will practice these methods by producing regression analyses and classifications on sample datasets *via* the chosen statistical software. An important part of these analyses is also being able to measure their performance, and you will practise using the diagnostic tools required to do so.

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

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

At the completion of this unit, you may progress to J4YD 36 *Machine Learning* at SCQF level 9, where the statistical learning techniques gained in this unit can be applied to machine learning.