Next Generation Higher National Unit Specification

Probability and Statistics for Data Science (SCQF level 8)

Unit code: J6CK 48

SCQF level: 8 (24 SCQF credit points)

Valid from: session 2023–24

Prototype unit specification for use in pilot delivery only (version 1.0) September 2023

This unit specification provides detailed information about the unit to ensure consistent and transparent assessment year on year.

This unit specification is for teachers and lecturers and contains all the mandatory information required to deliver and assess the unit.

This edition: September 2023 (version 1.0)

Unit purpose

This specialist unit is for learners who want to develop their existing knowledge of statistics. It focuses on the specific statistics required in data science.

Learners should have previous knowledge and experience of statistics before starting this unit and could have completed the Statistics for Data unit at SCQF level 7 or 8.

The unit teaches the necessary background of how to quantify uncertainty that is essential in data science and machine learning, covering various statistically related topics and theories. In particular, it covers:

- probability theory preliminaries: discrete and continuous distributions and their properties, maximum likelihood estimations of parameters
- marginal, joint, and conditional probabilities
- linear regression and binary classification
- advanced probabilistic algorithms: Bayes classifier, Markov chains, Viterbi, expectation-maximisation algorithms
- parametric and non-parametric hypothesis testing: type-I and type-II errors, p-value, confidence intervals
- data correlation
- data visualisation, data sampling, and information theory: entropy, probability distribution divergences, decision trees and random forest

The unit applies probability and statistics to well-known pattern recognition applications, such as classification and regression. Data visualisation not only discusses how data are represented, but also covers the psychology behind plots, unveiling details on how people perceive data representations.

On completing this unit, learners can progress to other machine learning-related units, such as the Artificial Intelligence unit at SCQF level 8 and the Big Data unit at SCQF level 8.

Unit outcomes

Learners who complete this unit can:

1 apply probability theory, distribution functions, and parameters estimation to data
2 fit pattern recognition algorithms to data
3 apply parametric and non-parametric hypothesis testing to data
4 determine the similarities between distributions and data
5 apply data sampling, data visualisation, and information theory on theoretical and practical contexts

Evidence requirements

Learners must provide knowledge and product evidence.

Knowledge evidence

Evidence must demonstrate all the knowledge defined in the 'Knowledge and skills' section and should be the minimum required to infer competence.

Learners can produce this evidence over an extended period of time, under lightly-controlled conditions, or holistically generated in conjunction with other units in a group award. You can sample evidence if using testing, however, the test must be carried out in timed and controlled conditions.

Product evidence

Evidence must demonstrate competence to analyse a real-world dataset. Such analysis must:

- demonstrate competence in the training of at least two different (but task-related) algorithms
- demonstrate at least one sampling technique
- demonstrate at least one hypothesis testing technique
- demonstrate critical analysis with the use of appropriate graphs and plots

Learners must carry out several tasks on a real-case dataset. To achieve all the learning outcomes, they should apply most of the statistical skills and pattern recognition algorithms covered in the unit to solve the given problems.

They must submit two items of evidence:

- Source code: to show how they performed the data analysis — this assesses their skills.
- Documentation summarising the findings of data exploration and the decisions they made to develop their computer program. Images, such as data visualisation plots, should be included, if appropriate.

You should emphasise to learners that all their code should be reproduceable on any computer.

The ‘Additional guidance’ section provides further information on assessment standards and suggestions on how you should assess learners.

Authentication is required when the evidence is produced in lightly-controlled conditions.

The standard of evidence should be consistent with the SCQF level of this unit.

You should use appropriate level descriptors when making judgements about the evidence.

Knowledge and skills

The following table shows the knowledge and skills covered by the unit outcomes:

Knowledge	Skills
Learners should understand: ♦ random variables and probability density functions: — function integration — discrete: uniform, Bernoulli, Poisson — continuous: uniform, gaussian and normal distribution — expected value and variance — mode ♦ maximum likelihood estimation — mean — standard deviation ♦ marginal, joint, and conditional probabilities: — Bayes rule ♦ regression: — metrics: mean absolute error, mean squared error, coefficient of determination — linear regression — ridge regression — lasso — elastic net ♦ binary classification: — logistic regression — perceptron — SVM — kernel trick — non-linear SVM	Learners can: ♦ recognise different probability density functions: — calculate expected value and variance — identify the mode of a distribution ♦ apply maximum likelihood estimation (MLE) to estimate the parameters in a probability density function ♦ apply Bayes rule ♦ fit a regression model (linear, ridge, lasso, and elastic net) to a given dataset: — closed-form solutions — iterative approach using gradient descent ♦ fit a binary classification model to a given dataset, such as logistic regression, perceptron, and support vector machines (SVM) ♦ fit advanced probabilistic algorithms algorithm to data: — Bayes classifier — Markov chains — Viterbi algorithm — expectation-maximisation algorithm (Gaussian mixture model) ♦ perform parametric and non-parametric hypothesis testing: — calculate p-value — determine confidence intervals — determine if the null hypothesis should be accepted or rejected ♦ perform data correlation

Knowledge	Skills
Learners should understand: ♦ advanced probabilistic algorithms: — Bayes classifier — Markov chains — Viterbi algorithm — expectation-maximisation algorithm (Gaussian mixture model) ♦ hypothesis testing: — type-I and type-II errors — p-value — confidence intervals — one- and two-tailed hypothesis testing — parametric: t-test — non-parametric: Mann-Whitney — ANOVA ♦ data correlation: — Pearson correlation coefficient — Kendall correlation coefficient — Spearman correlation coefficient — autocorrelation ♦ data visualisation: — type of plots or charts — psychology of data visualisation — retinal variables ♦ data sampling ♦ information theory: — Kullback-Leibler divergence — Jensen-Shannon divergence — entropy — cross-entropy — mutual information — information gain — Gini-index — data processing inequality ♦ decision trees and random forest	Learners can: ♦ generate plots for data visualisation: — determine which type of plot better visualises and represents a given dataset — use of retinal variables ♦ determine which sampling method to use given a dataset: — random sampling — systematic sampling — stratified sampling — cluster sampling ♦ calculate information theory measures for data: — calculate the Kullback-Leibler and Jensen-Shannon divergence between two distributions — entropy — cross-entropy — mutual information — information gain — Gini-index ♦ fit decision trees: — with the use of boosting — apply decision trees in random forest

Meta-skills

Throughout the unit, learners develop meta-skills to enhance their employability in the data science sector.

Self-management

This meta-skill includes:

- focusing: sorting, filtering
- initiative: decision making

Social intelligence

This meta-skill includes:

- communicating: receiving information, giving information, storytelling
- leading: influencing

Innovation

This meta-skill includes:

- curiosity: observation, questioning, information sourcing, problem recognition
- creativity: imagination, idea generation, visualising, maker mentality
- sense-making: pattern recognition, holistic thinking, synthesis, opportunity recognition, analysis
- critical thinking: deconstruction, logical thinking, judgement, computational thinking

Literacies

Throughout this unit, learners have opportunities to develop their literacy skills.

Numeracy

This is implicit in all areas of the unit and heavily used in some. Learners enhance their numeracy skills while studying the topics included in this unit.

Communication

Learners develop communication skills through data visualisation and hypothesis testing. Part of the assessment includes producing a formal report to communicate findings.

Digital

This unit shows learners how to implement statistically relevant pattern recognition algorithms in a computer program.

Delivery of unit

We recommend you deliver this unit as follows:

- lectures (40 per cent)
- tutorials (20 per cent)
- practical sessions (40 per cent)

Lectures should provide theoretical knowledge, while tutorials and practical sessions provide the skills detailed in the 'Knowledge and skills' section. Tutorials and practical sessions should enhance numeracy and digital literacy.

We suggest the following distribution of time:

- Lectures (48 hours):
  - probability theory preliminaries: 6 hours
  - marginal, joint, and conditional probabilities: 4 hours
  - linear regression and binary classification: 8 hours
  - advanced probabilistic algorithms: 8 hours
  - hypothesis testing: 10 hours
  - data correlation: 2 hours
  - data visualisation, data sampling, and information theory: 10 hours

- Practical sessions (48 hours):
  - probability theory preliminaries: 6 hours
  - marginal, joint, and conditional probabilities: 4 hours
  - linear regression and binary classification: 8 hours
  - advanced probabilistic algorithms: 8 hours
  - hypothesis testing: 10 hours
  - data correlation: 2 hours
  - data visualisation, data sampling, and information theory: 10 hours

- Tutorials (24 hours):
  - probability theory preliminaries: 3 hours
  - marginal, joint, and conditional probabilities: 2 hours
  - linear regression and binary classification: 3 hours
  - advanced probabilistic algorithms: 4 hours
  - hypothesis testing: 4 hours
  - data correlation: 2 hours
  - data visualisation, data sampling, and information theory: 6 hours

Additional guidance

The guidance in this section is not mandatory.

We recommend that you deliver this unit after the Mathematics for Data unit at SCQF level 8 and before or with machine learning-related units in the Higher National Diploma (HND) Data Science programme. You should recap some of the calculus topics, such as function integration, for learners in their first lecture.

Although lectures and tutorials would normally be held in a lecture theatre, you should organise practical sessions in a laboratory that has workstations, with the software for this unit already installed. We recommend you use open-source programming languages for the practical exercises, such as Python or R. MATLAB is also a suitable alternative, although it is not available for free. You should encourage group working, where appropriate, particularly during tutorial and practical sessions.

For lectures, you should introduce the main theorems concerning probability theory, such as the central limit theorem and the law of large numbers (although proofs may be omitted). When teaching statistics, we recommend that you use real-world datasets.

Approaches to assessment

You can assess knowledge evidence using a supervised 2-hour assessment. This assessment must be closed-book, however you can provide learners with a formula sheet. You should design the assessment to assess all outcomes, with respect to the ‘Knowledge and skills’ section.

We recommend a threshold score of 60 per cent. However, you may consider a threshold score of between 50 per cent and 60 per cent, if this ensures proficiency in knowledge and competency of skills for this unit.

Learners should submit two items of evidence:

- Source code (40 per cent of the mark): this shows how they performed data analysis and assesses their skills.
- Report (60 per cent of the mark): a 1500 word (approximately) structured document summarising the findings from their data exploration and the decisions made to develop their computer program. Reports should also include images (such as data visualisation plots), where appropriate.

You can decide the type of tasks the assessment should include. For example, if you use a dataset for a classification task, learners could train (at least) two classification models, such as SVM or decision trees. Alternatively, learners can perform regression analysis in a different scenario. In this case, an example could be to train two regression models (such as ridge regression and linear regression). Learners could demonstrate knowledge and skills in sampling strategies to deal with unbalanced datasets. They could perform hypothesis testing to test whether the performance of two (or more) trained models are statistically different (given the significance level).

We recommend that you provide the assessment brief to learners 4 to 5 weeks before the submission deadline. We suggest that you set a threshold score of 50 per cent for each item of evidence.

They can submit the assessment electronically, including the code and the formal report.

Alternatively, learners could create a blog to record their learning journey through the unit. The blog would contain all the defined knowledge and skills, and collectively meet the evidence requirements.

Equality and inclusion

This unit is designed to be as fair and as accessible as possible with no unnecessary barriers to learning or assessment.

You should take into account the needs of individual learners when planning learning experiences, selecting assessment methods or considering alternative evidence.

Guidance on assessment arrangements for disabled learners and/or those with additional support needs is available on the assessment arrangements web page.

Information for learners

Probability and Statistics for Data Science (SCQF level 8)

This information explains:

- what the unit is about
- what you should know or be able to do before you start
- what you need to do during the unit
- opportunities for further learning and employment

Unit information

This unit deepens your mathematical background for data science. It introduces an intriguing branch of mathematics: probability and statistics and shows you how to quantify the uncertainty.

You learn the basics of probability theory and probability density functions, and how to calculate certain information from them, such as expected value and variance. From there, you learn how to estimate a probability distribution from real data. This supports you in data analysis and pattern recognition by using algorithms and hypothesis testing; a statistical tool that allows you to prove (or disprove) assumptions that someone can make on data. Once you gain confidence with data, you are introduced to information theory. Data carries information and, as such, it can be quantified. Information theory is relevant because of how decision trees (and random forest) operate, and they are covered in the last part of the unit.

The unit is both theoretical and practical. Although probability theory is mostly mathematics, statistics is a discipline that goes hand-in-hand with data analysis and visualisation. All the topics covered in lectures have a tangible practical counterpart.

By studying this unit, you enhance a broad range of skills, such as thinking logically, independently, and holistically, and develop them through theory and practical sessions applied to data analysis using statistical tools. You also strengthen your computational thinking and imagination, as practical sessions are aimed not only at digital literacy, but also meta-skills (such as visualising, observation, idea generation, and maker mentality). You improve your problem recognition and deconstruction skills, as you solve complex problems by dividing them into smaller (and more manageable) problems.

You enhance your opportunity and pattern recognition skills, as probability theory and statistical tools are at the foundations of machine learning algorithms. The strengthening of analysis, judgement, and synthesis skills gives you the confidence to carry out the final assessment. As statistics revolves around data analysis and interpretation, this unit also improves the following meta-skills: receiving and giving information, as well as questioning and information sourcing.

To assess your knowledge and skills, you must perform data analysis tasks using all the statistical tools you have covered. These tasks could be solved by writing a suitable computer program, and by documenting your investigation and findings in a structured report.

Administrative information

Published: September 2023 (version 1.0)

Superclass: RB

History of changes

Version	Description of change	Date

Note: please check SQA's website to ensure you are using the most up-to-date version of this document.