Data Analysis and Interpretation

Assessment Rubric

Definition

Courses in this competency develop students' abilities to process, represent, assess, and interpret quantitative data sourced from the natural world, and to use it for the purpose of formulating and testing scientific hypotheses.

Framing Language

Data analysis and interpretation refer to the use of statistical and graphical methods to inspect, clean, transform, and model quantitative data in order to discover patterns and trends, develop models, and test hypotheses. Twenty-first century society requires that people be able to engage with data that is presented to them from a variety of sources. Additionally, people need to be able to construct their own sets of data effectively. This competency targets students' ability to process, represent, assess, and interpret quantitative empirical data, and to use it for the purpose of formulating and testing scientific hypotheses. Although such skills can be encountered and developed in a broad variety of disciplinary contexts, the competency situates them in the natural sciences because that provides a consistent, convenient, and efficient setting for teaching the fundamentals in synergy with their exposure to the natural sciences. Laboratory sections of science courses can be particularly effective settings for this. Students should re-encounter these ideas in disciplinarily appropriate forms throughout their degree programs.

A course carrying the Data Analysis and Interpretation in the Natural Sciences designation should, as a central aspect of its design:

1. Teach students the methods (basic concepts, skills, and procedures) of quantitative data analysis (QDA), including:
    a. the preparation and processing of raw quantitative empirical data,
    b. the application of statistical methods to characterize and assess data, and
    c. the use of mathematical concepts, language, and equations to describe ("model") relationships between empirical quantities — both as hypotheses to be tested empirically and as characterizations of the patterns revealed or suggested by empirical data.
2. Develop students' fluency with a variety of representations of scientific data (e.g., tables, charts, and plots), including the abilities to:

    a. critically interpret such representations, and
    b. choose and construct appropriate representations.
3. Instruct students in the interpretation and use of empirical data by appropriately using data analysis methods and representations to:
    a. reveal, characterize, and quantify patterns within and between empirical quantities, and
    b. test hypotheses against empirical evidence.

While the various rubric dimensions of the Competency's three SLOs necessarily overlap, for purposes of implementing the rubric:

- SLO #1 foregrounds the student's ability to choose and execute the procedural and primarily numerical elements of quantitative data analysis;
- SLO #2 foregrounds the student's ability to interpret the information provided by visual representations of data, and to choose how to effectively communicate data via such representations; and
- SLO #3 foregrounds the student's ability to *use* the skills of the prior two SLOs in the service of "doing science": discovering and characterizing patterns among empirical observables, and testing hypotheses against empirical data.

Glossary

Aspect – “Aspects” of a dataset refers to the various possible raw, derivative, reduced, or summarizing numerical values that could be presented (e.g., category counts, averages, standard deviations, transformed values, or best-fit parameters and residuals).

Characterize – To "characterize" a pattern or relationship is to quantify or precisely specify it in whatever way is appropriate to the pattern, which is highly dependent upon the nature of the pattern and the disciplinary and course context.

Communication Objective – A representation's “communication objective” means the specific feature(s), aspect(s), trend(s), or quality(ies) of a dataset that it is intended to communicate. For example, the choice of whether to use a pie chart, grouped or stacked bar graph, box-and-whisker plot, or other representation, and the choice of specific representation features details, will depend on whether one's goal is to convey a qualitative sense of the relative number of counts in different categories, to communicate a trend in how category counts relate to categories, to reveal whether the differences in category counts are statistically significant, etc.

Features and details – The “features and details” of a representation refers to all the choices that must be made to construct a specific example of a generic representation type, such as: what columns to include in a table, whether and how to subdivide or group table rows and/or columns, how to scale plot axes, whether and what trendlines to superimpose, whether and how to represent uncertainties or spread, what labeling and legends to add, what colors and fonts to use, etc.

Hypothesis – In this context, a "hypothesis" is an assertion about the behavior or characteristics of a natural system, such as the expected value of an empirically meaningful quantity, the identity or nature of an unknown sample, the causal dependence (or not) of one variable on another, the functional form of a correlation between quantities, or the presence or absence of a possible behavior or quality of a natural system. A hypothesis may be a prediction made by a theoretical model, a phenomenological claim or inference drawn from prior experience, or someone else's experimental result to be reproduced or refuted.

Methods – Quantitative data analysis “methods” refers to the elemental concepts, skills, and procedures involved in performing data analysis, including recording empirical data with appropriate precision and documentation, collecting data into a usable and well-labeled form (e.g., a spreadsheet or lab notebook table), calculating derivative quantities from raw measured values, estimating uncertainties, applying statistical reductions (e.g., calculating means and standard deviations), fitting trendlines and extracting their parameters, determining whether the difference between two uncertain values is statistically significant, etc. (The particular methods students are expected to demonstrate will vary according to discipline and course level, and are best determined by disciplinary curriculum experts.)

Pattern – A “pattern or relationship within data", for brevity simply called a "pattern" in the rubric, refers to whatever latent numerical features of the dataset as a whole carry significance regarding the natural system being described: the best estimate and uncertainty of a single physical quantity, the existence of a correlation between two variables, the approximate functional form of a correlation between two variables, the statistical significance of a particular difference or of the values of a variable for different groups, the presence or absence of a signal hidden within noisy data, the presence of specific meaningful features within a dataset, etc.

Representation – “Representations” of scientific data refers to data tables, histograms, simple or stacked bar charts, pie charts, 2D scatterplots, plots with error bars, plots with logarithmic scales, plot arrays, contour plots, heatmaps, bubble plots, dendrograms, network graphs, Sankey plots, chord diagrams, etc.

Robustness – The "robustness" of a pattern or relationship refers to its strength or clarity within the data, and to the interpreter's degree of confidence that it represents a meaningful feature of the natural system being investigated rather than a statistical fluke, a systematic error or contamination effect, or a figment of the interpreter's imagination. Depending on the nature of the dataset and pattern and the disciplinary and course context, robustness can be characterized by a qualitative description or by a quantitative measure of uncertainty, statistical significance, correlation coefficient, etc.

Sequence - A “sequence” of quantitative data analysis methods refers to the set of individual methods that must be executed to complete an analysis, such as recording data, reviewing and cleaning it, applying transformations and reductions, and extracting final results (specific numerical values, judgments, etc.) with some qualification of their accuracy or reliability.

Rubric Key

Dimension – A dimension expresses a fundamental aspect of a given Student Learning Outcome.

Level – The levels of learning describe progressive achievement, moving from Developing (Level 1) to Sophisticated (Level 4). Basic Competence in this MAC competency is achieved at Level 2. Students should be given opportunities to develop further levels of achievement in their upper-level, program-specific courses, after their initial introductory-level exposure to its fundamentals in a MAC-designated course.

SLO – A Student Learning Outcome (SLO) expresses the core learning goals of a curriculum. This rubric presents the SLOs for this MAC competency. Any course designated to deliver this competency is expected to state these SLOs verbatim in the course syllabus and to foreground them in its design and delivery.

SLO #1: Apply quantitative analysis to understand the natural world.				
Dimensions	**Sophisticated competence** (Level 4)	**Enhanced competence** (Level 3)	**MAC (Basic) competence** (Level 2)	**Developing competence** (Level 1)
Apply Quantitative Data Analysis Methods	Correctly executes an appropriate sequence of QDA methods by choosing, connecting, and adapting methods and sequence segments based on a sound understanding of the elements' purposes, principles, and limitations, on a dataset with substantially novel characteristics compared to prior examples.	Correctly selects and executes an appropriate sequence of QDA methods by choosing and connecting methods and sequence segments seen in prior examples, on a dataset differing from previously-encountered ones in ways that are structural and affect how methods are applied but not which apply or how results are interpreted.	Correctly executes a specified sequence of QDA methods by following prior examples, on a dataset differing from previously-encountered ones only in details.	Applies individual QDA ideas, skills, and procedures, following prior examples or step-by-step instructions, on a dataset differing from previously-encountered ones only in details (such as the number and values of data points and the specific empirical quantities they represent). Application may be partial or incomplete.
The stages of this rubric dimension describe students' path to mastery of QDA methods as a progression from reproducing individual skills, to reproducing multi-method analysis sequences, to constructing and adapting analysis sequences by choosing from among known methods when faced with a dataset that requires some adaptation, to flexibly drawing from a toolkit of known methods on the				

basis of deep understanding.				
SLO #2: Construct and interpret data tables, charts, graphs, or other representations of scientific data.				
Dimensions	**Sophisticated competence** (Level 4)	**Enhanced competence** (Level 3)	**MAC (Basic) competence** (Level 2)	**Developing competence** (Level 1)
Interpret Representations of Scientific Data	Correctly interrelates and synthesizes information from multiple representations of one or more related data sets to articulate a thorough, nuanced, and critical understanding of the natural system they describe.	Correctly and critically evaluates what can be plausibly or strongly inferred from a given representation, and identifies what it might omit, obscure, distort, or misrepresent.	Correctly summarizes the overall characteristics and trends implicitly portrayed by specific representation examples, and identifies anomalies and outliers revealed in the dataset.	Describes in words the information explicitly conveyed by specific representation examples.
The stages of this rubric dimension describe students' path to mastery of interpreting data representations as a progression from interpreting their explicit information literally, to noticing and describing overall trends and anomalies, to critically evaluating what a representation clearly communicates and what it obscures or distorts, to synthesizing information from multiple related representations.				
Construct Representations of Scientific Data	Selects, constructs, and describes (e.g., with captions or narrative text) a set of representations to effectively communicate multiple features and characteristics of a complex dataset, choosing appropriate communication objectives for each and for the overall set.	Chooses the aspects of a dataset to represent, selects an appropriate representation type to communicate them, and constructs a correct and effective representation to meet a specified communication objective.	Correctly constructs a specified type of representation for a given dataset, appropriately adapting the representation's specific features and details to the dataset and a specified communication objective.	Constructs a specified type of representation for a given dataset, following a similar example and with the specific features and details explicitly specified or illustrated.
The stages of this rubric dimension describe students' path to mastery of constructing data representations as a progression from constructing a representation of given data with the type, features, and details specified, to making choices about the features and details of a representation that best suit a particular case, to choosing what aspects of a data set to represent and what representation				

type best fits those, to creating an interrelated set of representations portraying multiple aspects of a complex dataset.				
SLO #3: Analyze and interpret quantitative data to evaluate and test hypotheses about the natural world.				
Dimensions	**Sophisticated competence** (Level 4)	**Enhanced competence** (Level 3)	**MAC (Basic) competence** (Level 2)	**Developing competence** (Level 1)
Identify Patterns and Relationships within Data	Chooses and explores various analytic techniques to discover, precisely quantify, and rigorously assess the robustness of multiple patterns within a complex and novel dataset, including unfamiliar kinds of patterns.	Appropriately applies analytic methods and constructs representations to correctly reveal, identify, characterize, and assess the robustness of meaningful patterns and relationships within a dataset of familiar structure, having prior experience with the kinds of patterns to consider.	Constructs appropriate representations of ready-to-represent data and uses them to correctly identify meaningful patterns and relationships present within it from a set of known and previously seen patterns and relationships, approximately or crudely characterizing them and their robustness.	Identifies a specified pattern or relationship in a provided representation of data.
The stages of this rubric dimension describe students' path to mastery of identifying patterns and relationships within data as a progression from identifying the presence or absence of a specific possible pattern or relationship in provided representations; to constructing specified representations and using them to determine which patterns and relationships from a specified set of possibilities are present; to choosing and applying analytic methods and representations to identify, quantify, and assess the robustness of various patterns and relationships from a specified set; to resourcefully applying an arsenal of analytic techniques to explore a novel dataset, uncover and precisely characterize patterns and relationships latent within it, and rigorously asses their robustness.				
Test Hypotheses with Empirical Data	Formulates an original, testable hypothesis about the natural world, identifies and obtains a relevant dataset (experimentally or second-hand), analyzes the data to test the hypothesis, and presents the claim,	Correctly applies QDA methods to determine whether a given raw dataset supports or contradicts a given hypothesis, and articulates an argument for their conclusion, constructing	Given an empirical dataset and accompanying already-complete analysis results and representations, identifies relevant characteristics and uses them as evidence to argue how the data support	Attempts to explain why and how a particular dataset, together with its analysis and results (all provided or described), supports or contradicts a specific hypothesis.

	results, and justification clearly and compellingly.	appropriate representations to communicate the key empirical findings.	or contradict a specified hypothesis.	
The stages of this rubric dimension describe students' path to mastery of testing hypotheses with empirical data as a progression from understanding an explanation of why and how a given empirical dataset supports or contradicts a given hypothesis; to finding and articulating evidence within a provided set of analysis results for or against a given hypothesis; to applying QDA methods on a dataset to test a given hypothesis and present a clear case for its support or contradiction; to generating one's own hypotheses, determining the relevant dataset for testing it, carrying out the appropriate analyses, and presenting the conclusions clearly and convincingly.				

Revised version. Approved by the General Education Council, April 2023.