GG 413 Introduction to Statistics & Data Analysis [3 cr]

- Exploratory Data Analysis
- Error Propagation
- Probability Theory and Statistics
- Statistical Tests
- Curve Fitting and Regression
- Sequence Analysis
- Spectral Analysis
- Analysis of Directional Data

Prerequisites:
Math 242 and GG250 (or the instructor’s consent). Basic linear algebra will be covered in class.

Course Purpose and Content:
Quantitative skills are extremely important in the natural sciences. With the rapid development of remote sensing from satellites and remotely operated vehicles and on-line access via the Internet, the amount of data an engineer and ocean, Earth, or environmental scientist must process and interpret is overwhelming; being able to analyze data on a computer becomes a necessity and often a job requirement. This course is an introductory class on how to quantitatively analyze data in the sciences. Its purpose is to present the fundamentals of exploratory data analysis, elementary probability theory and statistics, interpolation, curve fitting, regression, time-series (sequence and spectral) analysis, and analysis of directional data (e.g., fault strikes, wind directions). Examples on how to use the various techniques on real data sets (including the presence of "outliers" or bad data) will be stressed. Although we will concentrate on ocean, Earth and environmental applications it should be noted that these techniques are general and applicable to all the Sciences. Homework will be assigned, covering both theoretical as well as practical, computer-oriented problems. Computations may be carried out on the GG department's computers using suitable software (e.g., Matlab, Python, R), or on students’ own laptops. There will be one mid-term and one final exam in addition to the homework problems.

Text:
Recommended (optional) text: John C. Davis, Statistics and Data Analysis in Geology, 3rd edition, John Wiley. However, the instructor’s extensive notes will supplement the optional text. These will be available as an e-book by fall.

GG Student Learning Objectives (SLOs):
GG department has defined 5 learning objectives for the undergraduate degree program related to Relevance of Geology and Geophysics, Technical knowledge, Scientific method, Oral and written skills, and Evaluating Phenomena. This course incorporates content relevant to 2 of those:

- SLO2: Students can apply technical knowledge of relevant computer applications, laboratory methods, field methods, and the supporting disciplines (math, physics, chemistry, biology) to solve real-world problems in geology and geophysics.
- SLO3: Students use the scientific method to define, critically analyze, and solve a problem in earth science.

Course Goals:
This course will enable students to perform basic data analysis of typical natural science data sets using suitable computer software. The course stresses the importance of knowing uncertainties, how they affect the significance of results, and how to assign confidence limits to computed results. Specific student learning outcomes for the course are to:

- Understand how to apply exploratory data analysis techniques to discover structure in data.
- Learn how measurement errors propagate in calculation of derived quantities.
- Understand concepts of sample, population, distributions, and the central limit theorem.
- Gain experience in performing hypothesis testing for significance.
- Gain familiarity with linear algebra and least squares methods for curve fitting and regression.
- Explore various ways to examine sequential data.
- Understand the principles of spectral analysis and the key concepts aliasing and leakage.
- Become acquainted with statistical estimates and hypothesis testing for directional data.

Students will reach the course SLOs by successfully completing weekly homework and demonstrating factual knowledge at the mid-semester exam and at the final exam. Weaknesses in mathematical preparation will be determined by an early quiz, giving students an opportunity to address such weaknesses as the course progresses.

Assessment and Grading:

Data analysis requires hands-on practice. Accordingly, the course requires weekly problem sets that require a mix of mathematical and computational exercises. The homework problem set is a key component of the class and counts for 60% of the course grade. Homework must be handed in by the due date in order for a student to get full credit, unless a student has a valid excuse and has made arrangements with the instructor to hand it in late. Late homework will receive 50% credit only. If a student anticipates a conflict for exams, he or she must re-schedule the exam prior to the scheduled date. The final grade will be a weighted average of grades for homework (60%), the mid term (20%), and the final exam (20%). Grading will account for overall class performance, individual effort in completing the assignments, and attendance. If a student is concerned about his or her grade they may ask the instructor at any time about their standing.

Class Format:

This is a lecture course with no lab. Students are encouraged to actively ask questions in class, particularly if they do not understand the material being discussed. Most of the important material will be discussed in class, including problem sets similar to those in the homework. Students will get the most out of the lectures if they keep up with the reading and homework assignments. Students are expected to complete homework assignments using Matlab, Octave, Python or R either in the department’s computer room or on their computer. Supplemental Matlab functions needed for the assignments will be provided; note this is not a programming course.

Lecture Notes:

The required reading is from online lecture notes. These notes are supplemented by the recommended text. Problem sets will be distribute via Laulima and students will be required to submit homework via the Laulima course website.