GG413: Introduction to Statistics and Data Analysis

Meetings: Tue/Thu 10:30-11:45, POST 708
Instructor/Office hours: Garrett Apuzen-Ito (gito@hawaii.edu); TBA, POST 810.
Prerequisites: Math242 (2nd semester calculus), GG250 (scientific programming using Matlab), or instructor consent
Textbook: Paul Wessels Lecture Notes. Recommended (optional) text: John C. Davis, Statistics and Data Analysis in Geology, 3rd Edition

OVERVIEW AND OBJECTIVES:
The main purpose of this course is to provide students with foundational understanding of the basic theory behind statistics and quantitative data analysis, and provide practical experience with real data sets using computer software (Matlab, Octave, or FreeMat). The course emphasizes solving problems and independent learning and inquiry. Students learn about the following topics...

- Learn how to apply exploratory data analysis techniques to characterize their data or discover structure within it
- Understand how to propagate errors in calculations of derived quantities
- Learn and apply concepts of samples, population, probability distributions, and the central limit theorem
- Doing formal hypothesis testing in interpreting data
- Be introduced to matrices, linear algebra, and least squares formalism for curve fitting and regression
- Explore various ways to examine sequential data
- Understand principals of spectral analysis and the key concepts of aliasing and leakage
- Analyze directional data

And more! Emphasis will be on techniques and data sets in the geosciences but the course is relevant to all fields of science.

STUDENT LEARNING OBJECTIVES
This course emphasizes three student learning objectives for undergraduate and graduate students:

- Students can apply technical knowledge of computer applications and mathematics and physics to solving real-world problems in geology and geophysics
• "Students use the scientific method to define, critically analyze, and solve a problem in earth science
• "Students can communicate scientific knowledge in both oral presentations and in writing

FORMAT AND WORKLOAD
Lectures are to be viewed outside of class on YouTube (links provided below). Class time is an interactive learning environment and largely dedicated to working problem sets. Problem sets will be assigned approximately weekly and will involve using computer software to apply and practice using the techniques covered. There will be a mid-term and a final exam.

GRADING
Data analysis is a very hands-on activity and there will be weekly problem sets that require a mix of mathematical and computational manipulations. Homework must be handed in at the beginning of class on the due date, unless you have made prior arrangements with me. Otherwise, unexcused late homework will receive 10% less credit for each day it is late. If you anticipate a conflict for exams, you must re-schedule the exam prior to the scheduled date. The final grade will be a weighted average of grades for homework (70%), mid term (15%), the final exam (15%).

WORKING COURSE SYLLABUS

1. Basic Statistical Concepts
Week 1: (Reading: Swan and Sandilands Handout and Wessel 1.1-1.3)
1.1 Classification of data (see video #1 on Data Types and Precision vs. accuracy)
1.2 Exploratory data analysis (see EDA_Lecture files)
1.3 Error Analysis
   video #2, Reporting uncertainties, significant figures, & errors of sums & differences
   video #3, Computing errors of products & quotients
   video #4, Computing errors of products & quotients for Gaussian distributions
Examples 1 and 2
Homework #1 and required datasets

Week 2: (HW #1 due Tuesday)
1.4 Probability Basics
Lecture Videos
#1: 1.4.1 Permutations
#2: 1.4.2 Combinations
#3: The Binomial probability distribution
#4: The Hypergeometric distribution, 1.4.3 Probability, 1.4.4 Some Rules of Probability
#5: 1.4.6 Additional rules, 1.4.7 Conditional Probability
#6: 1.4.7 Conditional Probability and Bayes Theorem
Examples: Binomial & Hypergeometric PDs (& Matlab scripts for examples 1 & 2), and Conditional Probability
Homework #2: Probability

Week 3: (HW #2 due)
1.5 The M&M’s of Statistics (Davis pages on Central Limit Theorem)
Lecture Videos:

#1: 1.5.1 Population and Samples, 1.5.2 Measure of central location (mean, median, mode)
#2: 1.5.3 Measure of variation
#2.5: 1.5.6 Covariance and Correlation
#3: 1.5.4 Robust Estimation (MAD)
#4: 1.5.5 Inference about the mean and Central Limits Theorem
#5: 1.5.8-1.5.10 Probability Distributions, Binomial and Normal Distributions
#6: 1.5.10 The Normal (Gaussian) Probability Density Function
#7: 1.5.10-1.5.11 Applications of the Normal Distribution & the Poisson’s Distribution

See example script for plotting the binomial and normal distributions.

HW3: Statistics and Probability Distributions and data for problems 2 & 3

3. Hypothesis Testing

Week 4: (HW #3 due)
Read the following sections:
1.5. Inferences about means of populations, Videos #1, #2, #3
2.1 Null Hypothesis, Videos #4
2.2. Parametric Tests (Students t, Chi-squared, F tests),
   #5: one and two-sample test of means
   #6: two-sample t-test of means
   #7: estimating the variance of a population
   #8: one-sample, chi-square test of variance
   #9: two sample test of F-test of variance

Tables: normal distribution, t-distribution, chi-squared, F-distribution

Hw4: Hypothesis Testing with Parametric Statistics

Week 5: (HW #4 due)
2.2 Parametric Tests, videos...
   #1: general aspects of Chi-squared
   #2: Chi-squared test of a pdf
   #3: Chi-squared test of a pdf, example
   #4: test of linear correlation
2.3 Non-Parametric Tests, see video
   #5 on Parametric vs. Non-Parametric tests
   #6: Sign test of central value

Hw5: Hypothesis Testing II: datasets: “quakedays.txt”, and “rho.txt”

Week 6: (HW #5 due)
2.3 Non Parametric Tests
2.3.2 videos #1 and #2: Mann-Whitney 2-sample U test of median
2.3.3 #3: Kolmogorov-Smirnov goodness of fit test (1 or 2 sample) to a pdf
2.3.4 #4: Non Parametric test for correlation

Tables: Mann-Whitney, K-S (1-sample), K-S (2-sample)

Hw6: Hypothesis Testing III, see Matlab script kolsmir.m

3. Linear (Matrix) Algebra and Least Squares Inversion for Model Fitting
Week 7: B (HW #6 due)
3.1-3.2 #1 Matrices: General concepts and definitions
3.3-3.5 #2 Matrix Addition, Dot Product, and Matrix Multiplication
3.6 #3 Determinant of a Matrix
3.8 #4 Matrix Division: the Inverse Matrix
3.10 #5 Simple Regression and #6 RMS Misfit
3.11 General Least Squares Regression: #8 Part I and #9 Part II

Hw7: Least Squares Regression: see dataset hf.txt

Week 8:
Review for Mid-term
>>MIDTERM (Covering material through HW #6) <<<<

Hw8: Least Square Regression II: see hawaii.txt, faultstep.txt, and heaviside.m

4. Analysis of Variance (ANOVA)

Week 10: (HW #8 due)

Video #1: Analysis of Variance (ANOVA) of Linear Regression
4.4 #2 One-way ANOVA
4.5 #3 Two-way ANOVA

See also Draper & Smith excerpt

Hw9: ANOVA, see Hw9_hf.txt, Hw9_Prob2_Chromium.txt, and Hw9_StudentPorosityMeasurements.txt

5. Sequences and Time Series Analysis

Week 11: (HW #9 due)

5.1 Markov Chains: videos #1 and #2

See detailed explanation of Example 5-1

Hw10: Markov Chains

Week 12: (HW #10 due)

5.5 Autocorrelation, Video #1
5.6 Cross-correlation, Video #2

Matlab script shown in videos, with data for auto- and cross-correlation

HW11: Autocorrelation and Cross-Correlation, data files: TEMPER.TXT, Chesapeake_salinity.txt

5.8 Spectral Analysis

Week 13: (HW #11 due)

Video #1: Introduction to spectral analysis
Video #2: Orthogonality of periodic functions
Video #3: Discrete Fourier series

5.9 The Periodogram or Discrete Power Spectrum, Video #4

Hw 12: Spectral Analysis. See data file honolulu_resampled.txt
**Week 14:**
6.0 Analysis of Directional Data
  - **Video #1:** Polar histogram, computing means and variance
  - **Video #2:** Confidence intervals, One-sample tests of means
  - **Video #3:** Two-sample F test of means
Read [Davis Hand out](#)

**Hw13:** Analysis of Directional Data
See data files [Iceland_West.txt](#) and [Iceland_East.txt](#), as well as Matlab script [polarhist.m](#)

Happy Thanksgiving!

**Week 15:** (HW #12)
6.0 Analysis of Directional Data

**Week 16:**
Review for Final Exam