Earth 125/225: Statistics and Data Analysis in the Geosciences
Winter 2018

Course goals

Geoscience is becoming an increasingly data-driven discipline and it is more important than ever to know how to analyze and interpret the significance of your data. The goal of this course is to introduce you to a variety of statistical and analytical methods that you can use in research, graduate school, or during your future career.

At the end of this course, you will know which test to use to answer particular types of questions and you will be able to interpret the results and their meaning. This will require you to understand your data and recognize the requirements, assumptions, and limitations of different statistical tests and analytical methods.

This course will also introduce you to coding, an extremely practical skill for research and something that can set you apart with employers. You will learn to use R, a powerful open-source programming language designed for data science and widely used in both academia and business.

At the end of this course, you will be able to manipulate data, perform statistical tests, program functions for data analysis, and plot and interpret results using the R programming language.

Course structure and approach

The best way for you to succeed is by spending hands-on time programming in R. Coding will come more easily to some people, while the learning curve will be steeper for others, so this course is designed so that you can work at your own pace. Instead of a set schedule of topics and deadlines, you will instead work through modules that will build up your coding skills and statistical knowledge.

Evaluation and grading uses a competency-based framework, which assigns grades based on demonstration of mastery and progress through the series of modules, rather than from summative assessments like problem sets or a final exam. You can spend as much time as you need on each module, and you can repeat the exercises as often as is necessary to demonstrate mastery. This means that even if the learning curve is steep, everyone will have the chance to succeed without having to worry about falling behind.

Modules are grouped into levels based on the sophistication of the statistical method and the complexity of the R coding required. Each level has a final assessment; like the individual modules exercises, the final assessment can be resubmitted as often as necessary to demonstrate your mastery of the methods.

Module exercises will be available and can be submitted at any time before 5 PM on Friday March 23rd.

Outline of modules

Foundation

Descriptive statistics module: Central tendency (mean and median), dispersion (variance, standard deviation, interquartile range, coefficient of variation), standard error and confidence intervals

Level 1

Univariate tests module: T test, F test, analysis of variance (ANOVA) and Tukey HSD test

Non-parametric tests module: Kolmogorov-Smirnov test, Mann-Whitney U test, Kruskal-Wallis test, Levene's test

Categorical tests module: Goodness-of-fit (exact binomial test, exact multinomial test), independence (chi-squared test, Fisher's exact test)

Correlation and regression module 1: Parametric and non-parametric correlation, linear regression (ordinary least squares)

Level 2

Multivariate tests module: Hotelling T^2 test (including Mahalanobis distance)

Correlation and regression module 2: Logistic regression, multiple regression, partial correlation

Ordination module: Principal component analysis (PCA), non-metric multidimensional scaling (NMDS)

Level 3

Resampling methods module: Randomization and bootstrapping

Correlation and regression module 3: Generalized least squares, quantile regression, multilevel models

Maximum likelihood estimation module: Model selection with AIC

Above and Beyond

Correlation and regression module 4: Generalized linear mixed models

Machine learning classification module: decision trees, random forests

Contact info/Office hours

Email: mclapham@ucsc.edu, office: A208, phone: 459-1276

Office Hours: Will be held in the VisLab (EMS building A170) to provide space for more people. The room is booked and I will be there from 10-11:30 on Tuesdays and 3-4:30 on Wednesdays. You can also find me at my office at other times if you have more questions.

Grading

Students who complete all level 1 modules will achieve a passing (C) grade in the class, students who complete up to the end of the level 2 modules will achieve a B, and students who complete all modules to the end of level 3 will achieve an A. Intermediate (plus and minus) grades can be achieved with partial completion of the subsequent level, so for example completion of part of level 3 would correspond to a B+ or A- grade. Students who complete some or all of the "above and beyond" modules will gain the satisfaction of knowing they have exceptional statistical skills. As this is the first year for this style of grading, I retain the option of giving people higher grades than specified here – we’ll see how things go!

Tips for success

1. Set aside time to work through the modules outside of the regular class meetings. There aren’t deadlines for homework, papers, or exams, so you will need to be self-directed to make progress. It may help to schedule set times on your calendar, and to take advantage of times when workload in your other classes is lighter.
2. Stick with it! Coding can be inherently frustrating because you spend most time trying to figure out why it isn’t working, especially as you tackle more complicated tasks. You may feel confused (or even angry!) at times, but it will all start to come together.
3. When debugging your code, run it in the smallest chunks possible to diagnose which part might be causing the problem. If you’ve found the problem but don’t know what to do, or if you can’t find the problem, don’t hesitate to ask. Like any language, learning to code is hard but I aim to provide a supportive environment where everyone can succeed.