Biology 313: Scientific Computing for Biologists, Fall 2011

Instructor: Paul M. Magwene **Office**: FFSC 4103
Phone: 613-8159 **Email**: paul.magwene@duke.edu

Description

The focus of this course is statistical computing for the biological sciences with an emphasis on common multivariate statistical methods and techniques for exploratory data analysis. A major goal of the course is to help graduate students in the biological sciences develop practical insights into methods that they are likely to encounter in their own research, and the potential advantages and pitfalls that come with their use.

Prerequisites

Enrollment is limited to graduate students or undergraduates with permission of instructor. No previous programming experience is required, but familiarity with basic statistical concepts (equivalent of STA 213) is assumed.

Grading

Grading is based on weekly homework assignments. These homework assignments will typically consist of statistical problem solving exercises and/or programming tasks.

Course Website

https://github.com/pmagwene/Bio313

Texts

[1] Janert, P. K. 2010. Data Analysis with Open Source Tools. O'Reilly, Cambridge.

[2] A. B. Downey, J. Elkner and C. Meyers. How to think like a computer scientist: learning with Python. Available in HTML and PDF form under an open source license.

[3] Wickens, T. D. 1995. The geometry of multivariate statistics. Lawrence Earlbaum Associates, New Jersey.

Other Recommended Texts

[6] Hamilton, A. G. 1989. Linear algebra: An introduction with concurrent examples. Cambridge University Press, Cambridge.

A well organized and readable introduction to linear algebra. If your linear algebra skills are shakey (or non-existent) I'd recommend working through some of the exercises in each chapter.

[7] Krzanowski, W. J. 2000. Principles of multivariate analysis. Oxford University Press, New York.

I would have made this a required text but it's become unreasonably expensive (I bought my copy in 2003 for $25 used on Amazon, it's now over $100 used). Nonetheless, if you plan on having just on book on multivariate statistics on your bookshelf this is the one I'd recommend.

[8] Sokal, R. R. and F. J. Rohlf. 1995. Biometry. W. H. Freeman, New York.

Another good text to have on your bookshelf. A readable and well organized basic statistics book with examples drawn from the biological literature.

Syllabus

Date	Topic
August 30	Introduction; Getting Acquainted with R and Python, Literate Programming
September 6	Data as Vectors and Exploratory Data Analysis; vector operations, dot product, correlation, regression as projection, univariate visualizations
September 13	Linear Algebra Review I; Descriptive statistics as matrix operations, multivariate visualizations
September 20	Linear Algebra Review II; Regression models
September 27	Eigenvectors and Eigenvalues; Principal Components Analysis
October 4	Singular Value Decomposition, Biplots, and Correspondence Analysis
October 11	Fall Break
October 18	Discriminant analysis and Canonical Variate Analysis
November 25	Analyses based on Similarity/Distance I; Hierarchical and K-means clustering
November 1	Analyses based on Similarity/Distance II; Multidimensional scaling
November 8	Randomization and Monte Carlo Methods; Jackknife, Bootstrap
November 15	Building Bioinformatics Pipelines I; Pipes, redirection, subprocesses
November 22	Building Bioinformatics Pipelines II; Putting the concepts to work
November 29	Building Bioinformatics Pipelines III; Polishing the interface and generating publication quality graphics

Graduate Classes End December 2