Biol 697: Special Topics: An Introduction to Computational Data Analysis for Biology

Weekly Schedule: Tuesday & Thursday 11-12:30

Overview: This course will cover the basic statistical knowledge necessary for a graduate student to design, execute, and analyze a basic research project. The course aims to have students focus on thinking about the biological processes that they are studying in their research and how to translate them into statistical models. The course will take a hands-on computational approach, teaching students the statistical programming language R. In addition to teaching the fundamentals of data analysis, we will emphasize several key concepts of efficient computer programming that students can use in a variety of other areas outside of data analysis.

We will emphasize the underlying principle behind modern statistical analysis – that nearly every biological system can be described with a simple series of linear or nonlinear relationships with some meaningful error distribution around them. Additionally, we will emphasize thinking about whole biological systems, causality, and the limits of inference that can be drawn from observational versus experimental studies.

The course will build through a series of topics. We will begin by thinking about the basics of what is data, how do we curate it, and how do we efficiently visualize it. We will move on to thinking about natural systems and sampling design to derive inferences about the a single property within a system, such as the distribution of bird beak lengths or levels of gene expression. We will move on and think about how we describe causal processes within a system. We will discuss the different techniques used to fit models that describe these causal processes. From there, we will move on to an exploration of the role of experiments in deriving inferences about our study systems. We will move on to topics concerning how to construct and evaluate statistical models of complex systems from either experimental or observational data; we will then end with a discussion of the comparison of multiple alternative hypotheses.

Objectives:

1) To learn how to think about your study system and research question of interest in a systematic way in order to design an efficient sampling and experimental research program.

2) To understand how to analyze collected data to derive the most information possible about your research questions.

3) Provide the grounding needed to effectively collaborate with statistical experts.

4) Allow students to feel sufficiently comfortable with the basic principles of statistical analysis so that they can learn and implement techniques outside of the purview of this course.

Prerequisites: I will assume a basic knowledge of algebra and introductory calculus (although no calculus will be used). Undergraduate courses in probability theory and computer science are useful, but not required. Students who are new to programming should skim chapter 1 of Adler before beginning the course.

Required Texts:

Adler, J. (2009) **R in a Nutshell: A Desktop Quick Reference.** O'Reilly Media.

Vickers, A. (2009) **What is a p-value anyway? 34 Stories to Help You Actually Understand Statistics.** Addison Wesley.

Whitlock, W.C. and Schluter, D. (2008) **The Analysis of Biological Data**. Roberts and Company Publishers.

Recommended Texts:

I will be drawing on examples and materials from a few other sources. They include wonderful examples of R code in the context of data analysis. You are not required to have these, but you will either find them useful in this course or in future endeavors.

Bolker, B. (2009) **Ecological Models and Data in R.** Princeton University Press.

Matloff, N. (2011) **The Art of R Programming: A Tour of Statistical Software Design.** No Starch Press. http://nostarch.com/artofr.htm

Song, S. Qian (2009) **Environmental and Ecological Statistics with R**. Chapman and Hall/CRC Press, London.

Content and teaching approach: The course will be a mixture of lecture and hands-on data analysis lab. Students will be expected to have a computer available during the course so that they can follow examples and attempt in-class problems.

Grading: Your grade will be determined by a combination of weekly homework, a midterm, and a final exam. Homework will consist of a problem set and a short response to a chapter from Vickers. Homework will be worth 50% of your course grade. All exams will be take-home. The midterm will be worth 20% and the final will be worth 30%. *Additionally, students may earn extra credit for a statistical write-up of their own research data to be turned in during the finals period.*

Course Content:

Week	Topic
1	Data & Data Management
2	Biological Processes & Statistical Distributions
3	Data Visualization
4	Simulation & Basic Hypothesis Testing
5	Sampling Design
6	Fitting Linear Models: Least Squares
7	Fitting Linear Models: Likelihood
8	Generalized Linear Models

9	Experiments & the Linear Model (ANOVA)
10	Multiple Continuous Predictors
11	What should I sample? Simpson's Paradox
12	Interactions & Nonlinearities
13	Bootstrapping
14	Model Comparison

Things you need: A large amount of computer programming will be necessary to successfully complete the course, so students will need easy access to computers running R (or with administrative access to download R), which is free, open-source software and some form of spreadsheet software (Microsoft Excel, Open Office, etc.). We will learn how to load R and R packages in the class. Ideally, students will start the class with a general idea their project system or an ecosystem of interest (e.g., studying insects in salt marshes, experimentally driven levels of gene expression, patterns of biodiversity across a bathymetric gradient, yeast reproductive rates, etc.) as there will be opportunities for students to use their own data for course credit.

Office Hours: Prof. Byrnes will hold office hours Wednesday from 2:00-3:30.

Course notes: Slides and code for each lecture will be available on the course website before each lecture.

Software

- R - http://www.r-project.org/
- R Studio, a fantastic cross-platform interface for R - http://www.rstudio.org/

Useful Online References for R

R-Bloggers. *Read this daily*. http://www.r-bloggers.com/

John Verzani, "simpleR", in PDF

Patrick Burns, The R Inferno. "If you are using R and you think you're in hell, this is a map for you."

Thomas Lumley, "R Fundamentals and Programming Techniques" (large PDF)

A list of tutorials in R from universities around the world http://pairach.com/2012/02/26/r-tutorials-from-universities-around-the-world/

Additional Books About R and Statistical Computing

A fairly comprehensive list can be found at http://www.r-project.org/doc/bib/R-books.html. Below, I highlight a few of my favorites that overlap and extend the material in this course:

Benjamin M. Bolker. *Ecological Models and Data in R*. Princeton University Press, 2008. ISBN 978-0-691-12522-0. [Publisher Info | http://www.zoology.ufl.edu/bolker/emdbook/]

Julian J. Faraway. *Extending Linear Models with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC, Boca Raton, FL, 2006. ISBN 1-584-88424-X. [bib | Discount Info | Publisher Info |http://www.maths.bath.ac.uk/~jjf23/ELM/]

John Fox and Sanford Weisberg. *An R Companion to Applied Regression.*Sage Publications, Thousand Oaks, CA, USA, second edition, 2011. ISBN 978-1-4129-7514-8. [http://socserv.socsci.mcmaster.ca/jfox/Books/Companion/index.html]

M. Henry H. Stevens. *A Primer of Ecology with R*. Use R. Springer, 2009. ISBN 978-0-387-89881-0. [Discount Info | Publisher Info]

Paul Teetor. *R Cookbook*. O'Reilly, first edition, 2011. ISBN 978-0-596-80915-7. [http://oreilly.com/catalog/9780596809157]

John Verzani. *Using R for Introductory Statistics*. Chapman & Hall/CRC, Boca Raton, FL, 2005. ISBN 1-584-88450-9. [Discount Info | Publisher Info | http://wiener.math.csi.cuny.edu/UsingR/]

Hadley Wickham. *ggplot: Elegant Graphics for Data Analysis*. Use R. Springer, 2009. ISBN 978-0-98140-6. [Discount Info | Publisher Info]

Journals to Keep an Eye On

The Journal of Statistical Software. http://www.jstatsoft.org/

Methods in Ecology and Evolution.
http://www.methodsinecologyandevolution.org/

The R Journal. http://journal.r-project.org/current.html