QBIO 310: Statistical Thinking for Quantitative Biology Syllabus

General Information

Time: MW 9:30-10:50
Location: Online
Instructor: Michael “Doc” Edge (Pronouns: he/him) and Peter Calabrese (Pronouns: he/him)
Instructor email: edgem@usc.edu, petercal@usc.edu
Instructor office hours: TBD
Teaching Assistant: N/A
Note: Please bring a laptop and scratch paper to class.

Welcome! We are looking forward to working with you this semester.

Course Description

This is an upper-division course designed to introduce biologists to statistical theory for data analysis. Students will also learn basic programming skills in the statistical programming language R. The course is more mathematically demanding and more focused on general theory than BISC 305. At the same time, it is gentler and more targeted at biological data than courses that cover similar material in the math department, such as MATH 407 and 408. We will spend approximately 2/3 of the semester exploring simple linear regression, taking time to learn some statistical theory, to view linear regression from non-parametric/semi-parametric, likelihood-based, and Bayesian perspectives, and to implement methods in R. The remaining 1/3 of the semester will be a tour of some important techniques useful for describing, visualizing, and modeling different types of data, including from studies with multiple independent variables or dichotomous outcomes.

Textbook

We will be using *Statistical Thinking from Scratch: A Primer for Scientists*, by M.D. Edge, Oxford University Press, 2019. It is available from the campus bookstore and online booksellers.

Course Notes

In this course, we will take the time to learn one statistical method deeply first, and then we will add breadth at the end. This involves some mathematics and computer programming. Some of you may not have had math classes for a while and may have little experience programming. That will make the course a bit harder, but it is still possible to succeed with hard work and a good attitude. The grading system (see below) is designed to reward effort.

We will flip the classroom for some of the material, meaning that you will be expected to watch a taped lecture before class, and we will spend the class time learning actively. Slides for both traditional and flipped lectures will be posted.

Learning Goals

By course's end, our aim is that you will be able to:

- Discuss the philosophy involved in typical statistical estimation and inference, in which models are posited as data-generating processes with unknown parameters.
- Read and understand mathematical descriptions of simple statistical models.
- Explain the assumptions involved in justifying various views of the least-squares line, including a minimal "exploratory data analysis" view and views arising from semiparametric, parametric, and Bayesian models.
- Understand probabilistic and statistical concepts including expectation, variance, covariance, correlation, the law of large numbers, the central limit theorem, bias, consistency, efficiency, confidence intervals, *p* values, power, bootstrapping, permutation tests, likelihood, prior distributions, and posterior distributions.
- Design legible and informative data displays.
- Learn new methods for data analysis, such as linear regression, ANOVA, generalized linear models including logistic regression, principal component analysis, and linear mixed models, identifying principled reasons for choosing analysis methods.
- Explore the properties of statistical procedures using simulation and elementary probability calculations.
- Use R to analyze and plot data, as well as write code to implement basic versions of procedures like bootstrapping and permutation testing.

Prerequisites

There are no specific requirements to enroll. The main requirement is that you have an interest in learning about using mathematics and computation to support scientific claims with data. Beyond that, comfort with algebra is very helpful. Some familiarity with the ways in which statistical analyses are used in research is helpful—if the words "mean," "median," "mode," "scatterplot," "standard deviation," "t-test," "confidence interval," are at least vaguely familiar, you are covered on this dimension. We will use some basic calculus, and we will be programming. Courses in these areas will likely help you feel comfortable initially but are not required.

If you have taken MATH 407 and 408 or equivalent courses, then the material in this class would be repetitive for you, and you are urged to take a different course.

Grading Policy

Your final grade will be calculated on the basis of a weighted average, with the weights

40% Homework
10% Participation
20% Term Paper
20% Final Exam (take-home)
10% Midterm (take-home)

We will ask you to affirm that you have followed the rules for each exam.

Participation

Most lectures will include an activity that's worth a participation point. In case you need to miss some class sessions, we will make some substitute activities available that you can submit later.

Homework

There will be 8-12 homework assignments during the semester, due every 1-2 weeks. Doing the homework will be your most important method for learning the material. Homeworks will be graded on a 0-3 scale, where a 0 indicates that a homework is missing or less than 50% complete, a 2 counts for full credit and represents a good effort on all problems, though some results may be wrong; and a 3 represents an exceptional effort. All "2"s would give you a perfect homework score. Scores of "3" will not happen often and are considered bonus.

You are encouraged to work collaboratively on the homeworks, but please write your own solutions. We will drop your lowest homework score.

Software

We will use R, a programming language designed for statistical computing. R is available free online from the R Project website, https://www.r-project.org/. I recommend you also use RStudio, and interactive development environment designed for use with R. RStudio is also free. (Download the open source version of RStudio Desktop from https://www.rstudio.com/products/rstudio.) RStudio requires an active R installation.

Course Schedule (Subject to change)

Intro, course policies. The regression line as a motivating problem.
Reading: *STFS* Prelude and Chapter 1.

Unit 1: Mathematical and computational tools for statistics

Week 2
A conceptual tour of calculus: Optimizing and Summing.
Reading: *STFS* Appendix A.
Interacting with R and Exploratory Data Analysis
Reading: *STFS* chapter 2.

Week 3
R, part 2: Functions and Loops.
Reading: *STFS* Appendix B.
The least-squares line.
Reading: *STFS* chapter 3.

Week 4
Probability 1. (Foundations, Axioms, independence, conditional probability, Bayes' Theorem)
Reading: *STFS* chapter 4 (through Bayes' Theorem section)
Probability 2. (Discrete and continuous random variables, pdfs and cdfs, distribution families)
Reading: *STFS* chapter 4 (Discrete random variables through chapter end)

Week 5
Probability 3. (Expectation, Variance, the law of large numbers, the Central Limit Theorem)
Reading: *STFS* chapter 5 (intro, Expectation, Variance, and Central Limit Theorem sections)
Probability 4. (Correlation, covariance, conditional distribution, and a model for regression)
Reading: *STFS* chapter 5 (remaining sections)

Unit 2: Basic statistical theory

Week 6
Properties of Estimators: Bias, Variance, Mean Squared Error, and Consistency.
Reading: *STFS* chapter 6.
Properties of Estimators: Efficiency and Robustness. Decision Theory.
Reading : *STFS* chapter 6.

Week 7
Confidence Intervals and null hypothesis significance testing.
Reading: *STFS* chapter 7.
Power and effect size. Criticisms of NHST.
Reading: *STFS* chapter 7

Midterm around this time

Unit 3: Three major approaches to estimation and inference

Week 8
Plug-in estimators, the method of moments, and the bootstrap.
Reading: *STFS* chapter 8.
Permutation tests.
Reading: *STFS* chapter 8.

Week 9
Maximum-likelihood estimation.
Reading: *STFS* chapter 9.
Wald test, score test, and likelihood-ratio test.
Reading: *STFS* chapter 9.

Week 10
The Bayesian Alternative: Priors and posteriors.
Reading: *STFS* chapter 10.
Using the posterior distribution.
Reading: *STFS* chapter 10.

Unit 4: Models for data analysis

Week 11
Assessing linear regression assumptions, multiple linear regression and causal inference.
Special cases of linear regression (t-tests, regression analysis, one- and two-way ANOVA, etc.)
Reading: *STFS* Postlude

Week 12
Generalized linear models (logistic regression)
Random-effects models
Reading: *STFS* Postlude

Week 13
The expectation maximization (EM) algorithm
Markov-Chain Monte Carlo and Approximate Bayesian Computation
Reading: TBD

Week 14
Principal component analysis
LASSO and Ridge Regression
Reading: TBD

Term paper due the last day of regular instruction by 5 pm.

Term paper guidelines

For most of you, the term paper will involve modeling and analyzing a data set of your choice. Some of you may be writing honors theses, and analyzing the dataset that will form the basis of your thesis would be a great way to take advantage of this project. If you do not have a thesis dataset, you can consider i) another dataset from your thesis lab, ii) a dataset from one of your classmates, iii) a public dataset, such as one included in an R package or posted on the R directory “Free data sets” page: https://r-dir.com/reference/datasets.html. Your analysis must be original; you cannot repeat an existing analysis of a public dataset.

A second option for the term paper is to run a simulation study, wherein you study the properties of some statistical procedure(s) when applied to hundreds or thousands of simulated datasets with known properties. You could use such a study to compare the properties of different statistical procedures, such as bootstrap-based vs. normal theory confidence intervals or linear mixed models vs. generalized estimating equations. If you are interested in completing a simulation study rather than an analysis of existing data, please check in with an instructor, and we will help you make a plan.

The below outline describes a term paper involving a data analysis. Compared with a standard research paper in biology, your term paper will have abbreviated introduction, methods, and discussion, but it will have an expanded results section.

Term Paper Outline. (The approximate rubric below is out of 90. There will also be 10 points awarded for mechanics and style. If there are mechanics/style problems that make any of the below sections difficult to comprehend in certain places, then the relevant section scores may suffer as well.)

Introduction. State the question/hypothesis that motivates your examination of the dataset, with a brief theoretical framework. Finish the introduction with a statement that says how this data set addresses the question. This introduction should be brief—something like 1-3 paragraphs—and include 2-5 references. The goal here is *not* to provide a thorough literature review but just to tell the reader why you're analyzing this dataset and what question you'll be trying to answer (10 points).

Methods. Describe how the data were collected. One short paragraph. The idea is to highlight the key features of the data that will help the reader interpret the results, not necessarily to provide all the details necessary for replicating the study.

Exploratory data analysis. Create figures and tables to illustrate the major patterns in the data. The figures should have appropriate axis labels and, if applicable, either a legend or labels directly on the plot. Each figure will also need a caption. (Captions are written in your word processing software and/or LaTeX, not produced in R.) I expect 2 to 5 figures. Embed the figures in your text, commenting on the main things you can learn from the summaries. (10 points, including the R code, which will be in the Appendix)

Model Description. State what your competing models are both in English and with equations. The competing models can be just the null model and a particular model that corresponds to your hypothesis as in a classical *hypothesis testing* procedure. But you can also decide that it is more relevant to compare different complex models. (10 points).

Model estimation and analysis. We will cover three major perspectives for data analysis this quarter: nonparametric/semiparametric, parametric frequentist (maximum likelihood), and Bayesian. Unless you get a special exception for your project, you will need to analyze your data using techniques from *two* of these three families of methods. That is, you will complete two parallel sets of analysis in two different frameworks.

For example, if you use linear regression, you might estimate the parameters using least squares and then compute standard errors *both* using bootstrapping (non/semiparametric) and using typical normal-theory standard errors (parametric frequentist), then continue to build confidence intervals and compute p values using each set of standard error estimates. Or perhaps you are using a linear mixed model, and you compute point and interval estimates using maximum likelihood (maximum likelihood) and also by maximum a posteriori methods (Bayesian), basing prior distributions for the parameters on past work. Many other combinations are possible. Whatever you do, give and interpret complete results using both analysis methods, and compare and contrast the relative advantages and disadvantages of the analysis approaches you use. (35 points)

Analysis of model appropriateness. Assess the appropriateness of the model(s) you used. Do your data seem to violate the model's main assumptions? Use data displays and/or hypothesis tests to make your case, and discuss the implications. (10 points)

Conclusion. Write a brief conclusion of your modeling efforts, the statistical analysis and the implications for the question that was raised in the introduction. One or two paragraphs. (10 points)

Appendix. R code.

Statement on Academic Conduct and Support Systems

Academic Conduct:

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in SCampus in Part B, Section 11, "Behavior Violating University Standards" policy.usc.edu/scampus-part-b. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, policy.usc.edu/scientific-misconduct.

Support Systems:

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
studenthealth.usc.edu/counseling
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

National Suicide Prevention Lifeline - 1 (800) 273-8255 – 24/7 on call
suicidepreventionlifeline.org
Free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week.

Relationship and Sexual Violence Prevention and Services (RSVP) - (213) 740-9355(WELL), press "0" after hours – 24/7 on call
studenthealth.usc.edu/sexual-assault
Free and confidential therapy services, workshops, and training for situations related to gender-based harm.

Office of Equity and Diversity (OED)- (213) 740-5086 | Title IX – (213) 821-8298
equity.usc.edu, titleix.usc.edu
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants. The university prohibits discrimination or harassment based on the following *protected characteristics*: race, color, national origin, ancestry, religion, sex, gender, gender identity, gender expression, sexual orientation, age, physical disability, medical condition, mental disability, marital status, pregnancy, veteran status, genetic information, and any other characteristic which may be specified in applicable laws and governmental regulations. The university also prohibits sexual assault, non-consensual sexual contact, sexual misconduct, intimate partner violence, stalking, malicious dissuasion, retaliation, and violation of interim measures.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
usc-advocate.symplicity.com/care_report
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office of Equity and Diversity |Title IX for appropriate investigation, supportive measures, and response.

The Office of Disability Services and Programs - (213) 740-0776
dsp.usc.edu
Support and accommodations for students with disabilities. Services include assistance in providing readers/notetakers/interpreters, special accommodations for test taking needs, assistance with architectural barriers, assistive technology, and support for individual needs.

USC Support and Advocacy - (213) 821-4710
uscsa.usc.edu
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity at USC - (213) 740-2101
diversity.usc.edu
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
dps.usc.edu, emergency.usc.edu
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-120 – 24/7 on call
dps.usc.edu
Non-emergency assistance or information.