QBIO 310: Statistical Thinking for Quantitative Biology Syllabus

General Information

Lecture time: MW 2:00-3:20
Lecture location: THH 118
Section times (2 sections, go to only one): Tues 12:30-1:20 and Tues 2:00-2:50
Section location: RRI 421

Instructors: Peter Calabrese (he/him) and Michael "Doc" Edge (he/him)
Instructor email: petercal@usc.edu, edgem@usc.edu
Calabrese office hours: TBD
Edge office hours: TBD

Teaching Assistant: Obadiah Mulder (he/him)
TA email: omulder@usc.edu
TA office hour: TBD

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

Course Description

This is an upper-division course designed to introduce computational 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 remainder 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. You can access an electronic copy of the book via the USC library. If you prefer to have a physical copy, you can buy one from Oxford University Press or your favorite bookseller.

The book also has a github repository (https://github.com/mdedge/stfs) with supplementary material, including all the code used in the book and solutions to all exercises.

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.

Lecture slides will be posted. 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.

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 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. However, the way in which we approach these questions is quite different from in, for example, AP statistics. 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

35% Homework
5% Participation
20% Term Paper
20% exam 1
20% exam 2

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

Participation

The primary ingredient of the participation grade is section check-ins. Most sections will include a "check-in" question or two, with credit earned for answering the questions. In the event that you need to miss some sections, make-up assignments will be offered.

Homework

There will be approximately 10 homework assignments during the semester, due every 1-2 weeks. Doing the homework will be your most important method for learning the material. Homework 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, demonstrating both full understanding and unusual insight. All "2"s would give you a perfect homework score. Scores of "3" will not happen often and are considered bonus. Bonus points apply only to the homework grade (i.e. homework scores above 100% count as 100%).

Homework will be submitted on blackboard. Assignments will typically be due at 11:59 pm on Mondays (not every Monday, however). Assignments that are up to one week late will receive half credit. You are encouraged to work collaboratively on the homework, but please write your own solutions. We will drop your lowest homework score.

Questions

There is a general forum for questions and discussions on blackboard (under Tools > Discussion Board). Please ask questions about course content and general logistics here—if you have a question, someone else in the class likely has the same question, and answering it publicly will benefit everyone. If you have a logistics question related to your personal circumstances, please email an instructor or TA.

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/. We recommend you also use RStudio, an interactive development environment designed for use with R. (The instructors will be using it.) RStudio is also free.

(Download RStudio Desktop from https://www.rstudio.com/products/rstudio.) RStudio requires an active R installation.

Course Schedule (Subject to change)

Jan 8 (Edge): Intro, course policies. The least-squares line as a motivating problem.
Reading: *STFS* Prelude and chapter 1, chapter 3.

Unit 1: Mathematical and computational tools for statistics

Lectures in weeks 1-2 will be flipped. Please watch taped lectures before class; we will work on programming in groups during class time.

Jan 10 (flipped lecture, Calabrese): The statistical programming language R, part 1: R markdown and RStudio, interacting with R, exploratory data analysis.
Reading: *STFS* chapter 2. **Note that we are reading chapter 3 before chapter 2.**

Week 2
Jan 15: Martin Luther King, Jr. Day, no class
Jan 17 (flipped lecture, Calabrese): R, part 2 - Functions and Loops, data input/output
Reading: *STFS* Appendix B.

Week 3
Jan 22 (Calabrese): Probability 1 (Foundations, Axioms, independence, conditional probability, Bayes' Theorem)
Reading: *STFS* chapter 4 (through the end of section 4.2 / Box 4-2, pp 38-48)
Jan 24 (Calabrese): Probability 2. (Discrete and continuous random variables, pdfs and cdfs, distribution families)
Reading: *STFS* chapter 4 (sections 4.4-4.8, pp 48-58)

Week 4
Jan 29 (Calabrese): Probability 3. (Expectation, Variance, and the law of large numbers)
Reading: *STFS* chapter 5 (through the end of section 5.2, pp 60-68)
Jan 31 (Calabrese): Probability 4. (Correlation and covariance; The central limit theorem)
Reading: *STFS* chapter 5 (sections 5.3 and 5.5)

Week 5
Feb 5 (Calabrese): Probability 5. (conditional distributions; a model for linear regression)
Reading: *STFS* chapter 5 (sections 5.4, 5.6-5.7)
Feb 7 (Calabrese): Probability 6.
Reading:

Unit 2: Basic statistical theory

Week 6
Feb 12 (Calabrese): Properties of Estimators: Bias, Variance, Mean Squared Error, and Consistency.
Reading: *STFS* interlude; chapter 6 through the end of section 6.4.

Feb 14 (Calabrese): Properties of Estimators: Efficiency and Robustness. Decision Theory.
Reading : *STFS* chapter 6, sections 6.5-6.10.

Week 7
Feb 19: President's Day, no class
Feb 21: (Calabrese): The importance of data quality + exam review
Reading: None

Week 8
Feb 26 (Calabrese): **Exam 1 (in class)**
Reading: *STFS* chapter 7 through the end of section 7.2.
Feb 28 (Edge): Standard errors and confidence intervals
Reading: *STFS* chapter 7 through the end of section 7.2.

Week 9
March 4 (Edge): p values and hypothesis tests
Reading: *STFS* chapter 7, sections 7.3-7.4
March 6 (Calabrese): Power and effect size, multiple testing. Criticisms of NHST
Reading: *STFS* chapter 7, sections 7.6-7.9 (skip optional section 7.5)

March 11-15: Spring Break, no class

Unit 3: Three major approaches to estimation and inference

Week 10
March 18 (Edge): Plug-in estimators, the method of moments, and the bootstrap.
Reading: *STFS* chapter 8 through the end of section 8.2.
March 20 (Edge): Permutation tests.
Reading: *STFS* chapter 8, sections 8.3-8.5.

Week 11
March 25 (Edge): Maximum-likelihood estimation.
Reading: *STFS* chapter 9, through section 9.2; skip optional section 9.2.2.
March 27 (Edge): Wald test, score test, and likelihood-ratio test.
Reading: *STFS* chapter 9, sections 9.3-9.5.

Week 12
April 1 (Edge): The Bayesian Alternative: Priors and posteriors.
Reading: *STFS* chapter 10.
April 3 (Edge): Causal inference
Reading: None

Unit 4: Models for data analysis

Lectures during weeks 13 and 14 will be flipped. You will be expected to watch pre-taped lectures before class and to complete hands-on data analysis and simulation exercises in class.

Week 13
April 8 (Flipped lecture, Edge): Assessing linear regression assumptions, multiple linear regression
Reading: *STFS* Postlude, through the end of section Post.2.1
April 10 (Flipped lecture, Edge): Special cases and relatives of linear regression 1
Reading: Course notes (will be posted on blackboard)

Week 14
April 15 (Flipped lecture, Edge): Special cases and relatives of linear regression 2
Reading: Course notes posted on blackboard
April 17 (Flipped lecture, Edge): Generalized linear models
Reading: *STFS* sections Post.2.2-end

Week 15
April 22 (Edge): Statistics in Society: Eugenics as a cautionary tale
Reading: None
April 24 (Calabrese): Neural networks
Reading: None
Final project due April 26th, 11:59 pm

Final exam due Monday, May 6th

Term paper

For most students, the term paper will consist of a write-up of an analysis of data from the Framingham heart study.

If you are conducting research, you may have your own dataset that you would like to analyze, which you can pursue with special permission. You may also get special permission to perform 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.

We will release more details on the term paper later in the semester.

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.