BIOL 4080 – Advanced Biostatistics
Spring 2023

Prof. Paul Nerenberg

Office Location: ASCB 121D
Phone: 323-343-2122
E-mail: pnerenb@calstatela.edu
Student Office Hours: Mo 2:30-3:30 pm, Tu 1-2 pm, We-Fr by appt.

Lectures: MoWe 11:00-11:50 am in SH 348
Labs: MoWe 12:15-1:30 pm in SH 348

Prerequisites: BIOL 3000 and (MATH 2045 or MATH 2110).

Course Overview

In the 21st century, being able to draw conclusions from and generate actionable insights using data are vital skills in the life sciences…and many other professional fields! This course builds on your introductory-level statistics knowledge to develop a more detailed understanding of statistical models and their theoretical foundations. We will specifically focus on generalized linear models – a family of statistical methods that includes many types of regression (linear, logistic, and Poisson), ANOVA, and survival analyses. The last part of the course will include brief overview of machine learning, with a focus on using neural networks to perform regression and classification.

Throughout the course we will use the statistical programming language R to perform exploratory data analysis and statistical inference on a variety of data sets drawn from examples in the life sciences. Likewise, there will be a constant and strong emphasis on making sense of our conclusions and explaining them clearly to a wide range of potential audiences/readers. In addition to being valuable in their own right, the skills and knowledge developed in this course will prepare you for other advanced courses in statistics and machine learning.

Course Learning Objectives

By the end of this course, students should be able to:

1. Describe the theory underlying advanced statistical methods.
2. State statistical models and their assumptions for advanced statistical methods.
3. Choose an appropriate type of statistical analysis for different experimental designs.
4. Perform a variety of statistical analyses using computer software.
5. Correctly interpret the results of advanced statistical procedures.
6. Understand the rationale behind scientific study designs and choices of statistical analyses that appear in recent scientific literature.

Course Materials and Resources

Textbook: *An Introduction to Generalized Linear Models*, 4th edition by Dobson and Barnett; ISBN: 9781138741515. This is an excellent resource for understanding the theoretical underpinnings of generalized linear models, written by two practicing biostatisticians.

Another reference book that you might find helpful is *Generalized Linear Models with Examples in R* by Dunn and Smyth; ISBN: 9781441901170.

Course website: Each student is expected to have access to Canvas through the MyCalStateLA portal. The course site on Canvas will be your primary resource for acquiring information, turning in problem sets and lab assignments, and accessing other required materials for this course. All course materials will be available on Canvas, and all course announcements will be posted there as well.

Software: If you have your own laptop, I would strongly encourage you to install R. It can be downloaded here: https://mirrors.nics.utk.edu/cran/. You should also install RStudio Desktop, a graphical interface and workspace for R, which can be obtained here: https://posit.co/download/rstudio-desktop/.

Lecture and Lab Expectations

Both portions of the course will be fairly interactive, and each day I will give short problems and longer group exercises for everyone to think about and work on collaboratively. You will benefit the most from these active learning activities if you give them your full attention and come to class having done the assigned reading. While I encourage you to bring your laptop to class so that you can do R programming on your own computer, please do not let it – or any other technology (e.g., your smartphone) – become a distraction.

Grading Policy, Course Activities, and Assignments

Your course grade is designed to reflect the level of your demonstrated proficiency with the course material, with a particular focus on being able to extract insights from data sets (through data analysis and statistical inference) and explain them in a way that is understandable to anyone with basic scientific knowledge. As such, your course grade will be determined by your performances on problem sets, exams, laboratory assignments, and a final project:

Activity	**Number**	**Percentage of course grade**
Problem sets	5	50%
Exams	2	30%
Final project	1	20%
Total		**100%**

Once your various grades are combined, the nominal ranges for your overall course grade will be:

A	= 92-100%	A-	= 88-91.9%	B+	= 84-87.9%	B	= 80-83.9%
B-	= 76-79.9%	C+	= 72-75.9%	C	= 68-71.9%	C-	= 64-67.9%
D+	= 60-63.9%	D	= 56-59.9%	D-	= 52-55.9%	F	= 0-51.9%

These ranges may be adjusted down (but never up!) at my discretion.

Problem sets: There will be five problem sets that both reinforce and expand on what we discuss during lectures. These problem sets will be a mixture of theoretical problems and data analysis/statistical inference exercises with R. All problem sets will be distributed via Canvas and should be uploaded there as well. *Late problem sets will not be accepted without my prior permission.*

Exams: There will be two in-class exams with a variety of question types that will emphasize (a) conceptual thinking based on the course content and (b) R programming as it applies to data analysis and statistical inference. *Make-up exams will be given only for absences with valid, documented excuses.*

Final project: An important aspect of this course is to become comfortable performing exploratory data analysis and carrying out statistical inference. As a mini-capstone project, you will select an article from original research literature that involves the use of generalized linear models for statistical inference. You will give a 15-minute oral presentation to the class about the article, with a particular focus on the statistical methodology employed and the insights obtained from this analysis. [I will provide plenty of feedback as you select your article and develop your presentation.] More details about this final project will be given as the semester progresses.

Getting Help

This is a challenging (and exciting!) upper-level course in biostatistics and will likely require 8-10 hours per week of outside effort in addition to normal lecture and lab hours. If you find that you need help:

(1) Please stop by either in-person or on Zoom to see me! I have two scheduled days/times (see page 1) for student office hours, and you should also feel free to e-mail to setup an appointment if you'd like to meet outside of those office hours either in-person or via Zoom.

(2) Send me a message on our class's Discord server. Discord is a free online chat tool that is accessible via web browser and also via apps for Windows, macOS, Android, iOS, etc. [The Slack privacy policy is accessible here: https://discord.com/privacy.] The invite link to join the class server is posted on Canvas.

(3) Get in touch with your classmates. For those on the receiving end of such requests, helping someone else learn the material is without a doubt the best way to tell if you have mastered it yourself. You may find that by explaining the material to someone else also helps deepen your understanding of it.

Communication Expectations

For the fastest response, please contact me via my Cal State LA e-mail address (i.e., not Canvas) or Discord. During the week, I will get back to you in no more than one working day (if not much sooner!). If you contact me over the weekend, then I will get back to you no later than Monday. Please note that I don't use electronic devices from Friday evening to Saturday evening.

University Policies

Student Handbook

Information on student rights and responsibilities, academic honesty, standards of conduct, etc., can be found online in the University Catalog (http://ecatalog.calstatela.edu) under Procedures and Regulations. Students are expected to abide by the University's Academic Honesty Policy, and those who violate this policy will be subject to disciplinary action, potentially receiving a failing grade in the course for a single violation. In particular, I would like to stress the following:

- *You are expected to do independent work on all exams.*
- Collaboration with proper attribution is encouraged on problem sets, but your write-ups should be done independently. *Simply copying another student's work is (always) considered plagiarism.*
- *Posting any part of an assignment to and using any solution from "study" websites (e.g., Chegg, Bartleby, and the like) is not allowed and will be considered a violation of the Academic Honesty Policy.*

Americans with Disabilities Act (ADA)

Reasonable accommodation will be provided to any student who is registered with the Office of Students with Disabilities (OSD) and requests needed accommodation. It is the responsibility of the student to initiate any request for accommodation in the course; OSD does *not* notify faculty unless the student requests it for that course. OSD can be reached in-person in the Student Services Building (Room 1320), on the web at https://www.calstatela.edu/osd, or by phone at 323-343-3140.

Mental and Physical Health

With both academic and personal challenges, college/graduate school can be stressful experience. Diminished mental and physical health, including significant stress, mood changes, excessive anxiety, or problems with eating or sleeping can interfere with academic performance. The source of symptoms may be strictly related to your coursework; if so, please speak with me. However, problems with relationships, family worries, loss, or a personal struggle or crisis can also contribute to decreased academic performance. Please remember that help is always available. Cal State LA provides resources for Counseling and Psychological Services (http://www.calstatela.edu/studenthealthcenter/caps) to support the academic success of our students.

Course Schedule

Note: This schedule is *approximate* and will likely evolve throughout the semester. Any changes to the schedule will be announced during lecture and posted online to Canvas; it is your responsibility to remain informed of any announced changes.

Week	Day/Date	Lecture Topic	Lab Topic
1	Mo – 1/23	Probability theory and calculus review	Intro to programming with R; intro to R Markdown
	We – 1/23	Probability theory and calculus review (cont.)	
2	Mo – 1/30	Probability distributions	Probability distributions in R; Problem Set 1 collaborative time
	We – 2/1	Probability distributions (cont.)	
3	Mo – 2/6	The model fitting process	Exploratory data analysis and model fitting
	We – 2/8	The model fitting process (cont.)	
4	Mo – 2/13	Exponential family of distributions	Problem Set 2 collaborative time
	We – 2/15	Generalized linear models	
5	Mo – 2/20	Simple linear regression	Linear Regression
	We – 2/22	Simple linear regression (cont.)	
6	Mo – 2/27	Multiple linear regression	Problem Set 3 collaborative time
	We – 3/1	Multiple linear regression (cont.)	
7	Mo – 3/6	Exam #1 review	**IN-CLASS EXAM #1**
	We – 3/8	**IN-CLASS EXAM #1**	
8	Mo – 3/13	ANOVA	One- and two-factor ANOVA; ANCOVA
	We – 3/15	ANOVA (cont.)	
9	Mo – 3/20	Repeated measures ANOVA	Repeated measures ANOVA; Problem Set 4 collaborative time
	We – 3/22	Logistic regression	
	Week of 3/27 and 3/29	***Spring Break***	***Spring Break***
10	Mo – 4/3	Logistic regression (cont.)	Logistic regression; Poisson and negative binomial regression
	We – 4/5	Poisson and negative binomial regression	
11	Mo – 4/10	Log-linear models	Log-linear models; Problem Set 5 collaborative time
	We – 4/12	Wrap-up of GLMs	
12	Mo – 4/17	Survival analysis	Survival analysis
	We – 4/19	Survival analysis (cont.)	
13	Mo – 4/24	Exam #2 review	**IN-CLASS EXAM #2**
	We – 4/26	**IN-CLASS EXAM #2**	
14	Mo – 5/1	Intro to machine learning	Neural networks
	We – 5/3	Intro to neural networks	
15	Mo – 5/8	Neural networks (cont.)	Final project group meetings
	We – 5/10	Neural networks (cont.)	
16	Mo – 5/15	**FINAL PROJECT PRESENTATIONS, 9:30-11:30 am (tentative)**	