Advanced Statistics in Ecology and Evolutionary Biology (EEB172/C202)

Winter 2021, 4 units

Professors:

Nandita Garud, ngarud@ucla.edu

Kirk Lohmueller, klohmueller@ucla.edu

Teaching assistant (EEB172): Madeline (Maddie) Cowen mcowen@ucla.edu

Zoom link for all classes and office hours: https://ucla.zoom.us/j/95099273025

	Time:	**Location:**
Lecture (EEB172/C202)	Tues 10-11:50pm	Zoom
Computer lab – grad students (C202)	Thurs 2-3:50pm	Zoom
Computer lab – undergrad students (172)	Fri 2-3:50pm	Zoom
Office hours – Dr. Garud	Wednesday 9-11am*	Zoom
Office hours – Dr. Lohmueller	Monday 1-2pm* Fridays 11-12pm*	Zoom
Office hours – Madeline Cowen	Mondays 3-5pm	Zoom

* During the respective weeks in which Dr. Garud and Dr. Lohmueller are teaching. Also available by appointment.

Introduction: Biological research relies strongly on statistical analysis tools to interpret experimental and observational data. Modern statistical approaches provide powerful tools for analyzing complex data sets such as interactions among system components. In this course we will overview and learn how to apply advanced statistical methods that go beyond linear models and mean comparison, including bootstrapping, permutations, basic probability theory, likelihood-based inference, Bayesian statistics, mixed effects linear models, supervised machine learning, and unsupervised machine learning. By the end of this course you will be able to explain which statistical approaches are appropriate for different types of research questions and critically evaluate their outputs.

All statistical analysis will be conducted in R.

Prerequisite: An understanding of basic statistics including descriptive statistics, regression, correlation, comparing means and proportions is required. Experience working with R is recommended.

Course Prerequisites: Stats 13, LS 40, or an equivalent course.

Course goals:

By the end of the course you will be able to:

1. Identify the appropriate statistical analysis for a given biological study.
2. Implement data analysis in the software R.

3. Interpret the results of a statistical analysis and report it visually.

<u>Contact and communication channels:</u> Course material and announcements will be made available through CCLE. You are welcome to seek help with lab assignment at office hours – times listed above or by appointment. Your best resource in this class are your class mates and grading in this class is NOT curved. All emails to instructors regarding this class should include: "EEBC202" or "EEB172" in their subject line. Please allow 24 hours for a response and do not except email responses during the weekend.

<u>Computer labs:</u> During computer labs you will use the programing language R to implement the statistical methods discussed in lecture. Lab work is meant to provide you with hands-on experience with the methods we learn. Please make sure that your computer has R and RStudio (or any other IDE) installed on it. All lab exercises will be available on CCLE at the beginning of each computer session. Computer labs will be held separately for the undergraduate section (EEB172) and the graduate section (EEBC202) and will have different assignments and different requirements, as detailed below.

<u>Assignments and grading</u>: To develop your statistical skills, practice techniques, and learn how to interpret statistical methods in the primary literature, there will be weekly readings and computer lab assignments. For lecture each week both undergraduate and graduate students (EEB172 and C202) will read 1-2 papers or chapters (provided as an online link) on a particular statistical technique. We will discuss these readings in class and you are expected to come prepared to these discussions.

Undergraduate students (EEB172): *<u>Computer lab:</u>* Participation in the computer lab is mandatory. Each week you will implement the statistical method we discuss in the programing language R. You will receive weekly lab assignments on Friday at the computer lab where you will begin working on it. If you do not finish working on the assignment during the lab session, you will be expected to continue working on it at home. You are required to upload your completed lab assignments to CCLE. Assignments are due by Friday at 2pm - before lecture begins. You are required to submit the code you developed for the lab assignment, along with graphs and any statistical outputs requested, using Markdown. Lab assignments should be handed in on time. Every late day will reduce one point from your homework grade. The first late day begins as soon as lecture starts. You are allowed to work with your peers as long as you state on your assignment who you worked with. There will be nine lab assignments to hand in and the lowest grade will be dropped so only eight assignments will count for the final grade. The last (10^{th}) lab section will be used to review the material from the entire course.

<u>Final exam:</u> The final exam will be a take home assignment that will be handed out at the end of the last computer lab section. In this final assignment you will be asked to analyze data we will provide, interpret the results and report them both visually and verbally, similar to the computer lab assignments.

<u>Grading:</u> Each lab assignment is worth 15 points (7.5%) and the assignment with the lowest grade will be dropped, leaving eight graded assignments for the final grade (8*7.5% = 60%), the final exam will account for 60 points (30%). Twenty points (10%) will go for participation in lecture and lab. The highest available grade is 200 (100%). Two bonus points will added for filling out course evaluations (1 for the lecture and 1 for the lab). You need 120 points (60%) to pass this course. Grading in this class is not curved and letter grading will be assigned according to the table below:

Letter grade	Percent:	Number of points
A	90-100%	180-200
B	80-90%	160-180
C	70-80%	140-160
D	60-70%	120-140
F	<60%	<120

Tips for success: In addition to attending and participating in lecture and the computer lab, you should expect to work on your assignments at home for an additional 1-4 hours each week. Solving all the lab assignments will prepare you for the final exam. If you do not finish your work on the assignment in lab, make sure you continue working on it early to allow enough time for seeking help from your peers or us, in case you need it.

Graduate students (EEB C202): *Computer lab:* Each week, we will implement the analysis of a research paper during the computer lab. The empirical paper will be presented by one or two students at the start of each lab session. Implementation of the paper results will be conducted in R and you will be provided with worksheets to guide your work. Working with other students is encouraged.

Guidelines for presenting the empirical paper: The presentation should focus on the biology and results. No need to go deeply into the statistics used – we will do that together. However, it will be helpful to present the kind of data collected. Make sure that in your presentation you:

(1) Introduce the study system and biological questions (hypotheses).
(2) Detail what the researchers measured.
(3) Present the results. You are encouraged to present figures and tables from the paper.
(4) Explain the biological conclusions.

If your paper addresses many questions and uses a large variety of approaches, please focus on the information relevant to the statistical test we are addressing that week.

Final project: Students will submit a final project in which they will analyze a dataset of their choosing using one of the methods we learnt during the quarter. Instructor approval of the dataset is required and using a dataset the students collected is highly recommended. Students must approve their dataset before the last day of the quarter and the final assignment is due on the last day of finals week. Students who do not have data they collected to analyze will be provided with data upon request.

Guidelines for final assignment: The final project write up should explain your biological question, your analysis and results. Include the following:

(1) Introduction to the biological question (0.5-1 page).
(2) Data collection methods that resulted in the dataset analyzed (0.5-1 page).
(3) Detailed explanation of the statistical methods (0.5-1 page).
(4) Results with figures and statistical outputs (length depends on number of figures). Do not forget to label your figure axes and provide illuminating figure captions!

(5) Brief discussion of results (0.5 page).

Grading: Presentation and performance in lab account for 40% of the grade. Attendance is mandatory and participation in lecture will count for 20% of the final grade. The final project will account for 40% of the grade. Grading is NOT curved!

Inclusive learning environment:

UCLA values diversity and inclusion. We expect everyone in this class to contribute to a respectful, welcoming, and inclusive environment to support the learning of all other members of the class. If there are aspects of the instruction or design of this course that result in barriers to your inclusion or accurate assessment or achievement, please notify us.

Students needing academic accommodations based on a disability should contact the Center for Accessible Education (CAE) at (310) 825-1501 or in person at Murphy Hall A255. CAE will assess all requested and communicate the adjustments to your professor and/or Teaching Assistant. Any students with CAE approval for proctoring arrangements during exams will need to inform the professor and/or TA prior to exams. When possible, students should contact CAE within the first two weeks of the quarter to allow reasonable time to coordinate accommodations. For more information, please visit the CAE website: http://www.cae.ucla.edu.

Resources are available to foster the well-being of all UCLA students as they pursue their academic goals. Any student who finds themselves in immediate distress, please call Counseling and Psychological Services (CAPS) to speak directly with a counselor 24/7 at (310) 825-0768, or please call 911. For more information, pleases visit the CAPS website: http://www.counseling.ucla.edu.

Academic Integrity:

It is our goal to help you all do super-well in this course. We also recognize that life is crazy in ways that it has never been before. Thus, if you are having problems keeping up with the material, please reach out to us. We will do our best to help you in this situation. Please do not cheat or engage in dishonest conduct. It is not worth it. Cheating has severe consequences. Please review the Dean of Student's Student Conduct Code (https://www.deanofstudents.ucla.edu/Individual-Student-Code).

You are NOT permitted to post any of the assignments or course materials to Course Hero or any other websites outside of CCLE.

Course schedule:

Week (dates)	***Topic***	***Readings*** ***Class:*** *read for Tuesday* ***(both EEB172 and C202)*** ***Lab:*** *read for Thursday* ***(C202 only)***
1 - Dr. Lohmueller	Overview of basic statistical concepts including, data	R for beginners

(1/5, 1/7)	summary descriptors (distribution, mean, median…), correlations, and mean comparisons.	**Lab:** Getting (re)acquainted with R and data
2 - Dr. Lohmueller (1/12, 1/14)	Resampling methods: bootstrapping and permutations - I (the basics)	**Class:** Crowley 1992, Resampling methods for computation-intensive data analysis in ecology and evolution. Annu. Rev. Ecol. Syst. **Lab:** Rennison et al 2019, Ecological factors and morphological traits are associated with repeated genomic differentiation between lake and stream stickleback. Proc Royal Soc B.
3 - Dr. Lohmueller (1/19, 1/21)	Resampling methods: bootstrapping and permutations - II (parametric bootstrap and simulation)	**Class:** Puth et al. 2015 On the variety of methods for calculating confidence intervals by bootstrapping. Journal of Animal Ecology. **AND** **Class:** https://medium.com/@aliaksei.mikhailiuk/a-note-on-parametric-and-non-parametric-bootstrap-resampling-72069b2be228 **Lab:** Hodgins-Davis et al 2019**,** Empirical measures of mutational effects define neutral models of regulatory evolution in *Saccharomyces cerevisiae*. PNAS
4 – Dr. Lohmueller (1/26, 1/28)	Basics of Probability and random variables	**Class:** https://www.analyticsvidhya.com/blog/2017/02/basic-probability-data-science-with-examples/ **AND** Viti et al 2015, A practical overview on probability distributions. J Thoracic Disease **AND** Rannala, Chapt 2 Modeling Biological Data (just pages 27-43). **Lab:** Problem set (no paper)

5 – Dr. Lohmueller (2/2, 2/4)	Likelihood-based inference	**Class:** Myung 2003**,** Tutorial on maximum likelihood estimation. Journal of Mathematical Psychology **AND** https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1 **Lab :** James & Eyre-Walker 2020, Mitochondrial DNA Sequence Diversity in Mammals: A Correlation between the Effective and Census Population Sizes. Genome Biol Evol.
6 - Dr. Garud (2/9, 2/11)	Bayesian inference	**Class:** https://towardsdatascience.com/introduction-to-bayesian-inference-18e55311a261 **AND** https://towardsdatascience.com/from-scratch-bayesian-inference-markov-chain-monte-carlo-and-metropolis-hastings-in-python-ef21a29e25a **AND** Beaumont and Rannala 2004,The Bayesian Revolution in Genetics **Lab:** Pritchard 2000, Inference of Population Structure Using Multilocus Genotype Data, Genetics
7 - Dr. Garud (2/16, 2/18)	Generalized linear models (GLM) and mixed models (GLMM)	**Class**: https://medium.com/analytics-vidhya/introduction-to-mixed-models-208f012aa865 **AND** https://online.stat.psu.edu/stat504/node/216/ **Lab:** Douhard et al, 2016, The influence of weather conditions during gestation on life histories in a wild Arctic ungulate, Proc. Roy. Soc. B.

8 - Dr. Garud (2/23, 2/25)	Model selection (AIC)	**Class:** https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced **AND** https://www.methodology.psu.edu/resources/AIC-vs-BIC/ **Lab:** Hirst and Forster, 2013, When growth models are not universal: evidence from marine invertebrates, Proc. Roy. Soc. B.
9 - Dr. Garud (3/2, 3/4)	Unsupervised machine learning *Dimensionality reduction methods such as PCA and k-means clustering*	**Class:** https://builtin.com/data-science/step-step-explanation-principal-component-analysis **AND** Novembre et al. 2008, Genes mirror geography within Europe. Nature. **Lab:** Backhed et al. 2015, Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life, Cell Host & Microbe.
10 - Dr. Garud (3/9, 3/11)	Supervised machine learning *Classification techniques including Random Forests, Logistic Regression, SVM, and Deep learning.*	**Class:** Valletta et al, 2017, Applications of machine learning in animal behaviour studies, Anim. Behav. **Lab:** Pasolli et al. 2015, Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLoS Comp. Bio.

Important dates:

EEB 172:

3/12 – Receive final assignment in Lab

3/19 – Last day to submit final

EEB C202:

1/7 – Last day to submit top 3 choices for paper presentation (C202) -- by 2pm

2/25 – Last day to approve dataset for final project for graduate students (C202).

3/19 – Last day to submit final project for graduate students (C202).