Applied Statistical Modeling for Environmental Management

Spring 2023

M/W 10:15-11:30 am EST
Academic credit: 3 units
Course format: flipped (videos + discussion + lab)

Instructor's information

Dr. John Poulsen, Associate Professor

Office hours: See the class Sakai site
e-mail: john.poulsen@duke.edu
website: https://www.tropicalecology.us/

I am an ecologist, which means I use statistics to make sense of (usually messy) data to answer scientific questions. I took lots of statistics and modeling courses when I was in graduate school but learned even more by applying statistics to environmental problems. I won't (usually) write out theorem proofs or long equations on the board, and won't ask you to either. But you will learn what you need to know to start analyzing your own data and assessing the quality of other people's work.

What is this course about?

This course introduces statistical analysis and modeling for applied problems in the environmental sciences and resource management. The goals of the course are threefold: first, to make statistics intuitive and accessible so that you can understand and evaluate their use (both in daily and professional life); second, to provide you with a tool for conducting statistical analyses for your own research; and third, to make sure you know what you don't know. Statistics is like a loaded gun: helpful when used correctly and potentially dangerous in the wrong hands (Charles Wheelan, *Naked Statistics*). Think of the damage misused statistics could render on issues like climate change or global pandemics. This is an introductory course: the stunts that professional statisticians perform shouldn't be tried at home without supervision (or additional coursework).

So, what stunts will you learn in this class? We start with a brief refresher in probability and common statistical distributions, and move quickly to sampling, hypothesis testing, and statistical inference. Some of the principle statistical tests include one- and two-sample tests, resampling techniques, linear models (ANOVA, regression) and generalized linear models. That's enough statistical speak for one paragraph... the primary stunt you will learn is how to process data/information so that it is useful to you.

In the old days, scientists were mostly limited to software programs with drop-down menus like SPSS to analyze data; but now we have a flexible (and free) environment for statistical analysis and graphing – R. The difference between R and most other statistical packages is that it is open source software maintained by scientists for scientists. R is used within a command-line interface, which means you get to (read: must) write lines of code. This imposes a slightly steeper learning curve than other software, but also offers enormous freedom in the way you process data. (Believe me, there will come a time when pull-down menus feel like shackles.)

This course is designed for students with an introductory background in statistics. Terms like p-value, degrees-of-freedom, and confidence intervals should already be part of your vocabulary. This way we can focus on analysis of data using R, interpretation of results, and learning to report statistics in a lucid and

concise manner. Intended for both MEM and PhD students, the course does not require formal mathematical training, but does require knowledge of algebra and previous experience or willingness to write R code.

What background knowledge do I need before taking this course?

Students are expected to have a basic understanding of probability and statistics as covered in an undergraduate statistics course and should have passed the NSOE diagnostic exam. If you have not fulfilled the prerequisites, take the course at your own risk.

What will I learn in this course?

So, what should you be able to do at the end of this class?

1. Understand and articulate fundamental statistical concepts.
2. Interpret (either from a computer output or scientific papers) and communicate results from statistical analysis.
3. Look at data, formulate a hypothesis, and know which family of statistical tests can be appropriately applied.
4. Possess the skills necessary to manipulate datasets, analyze data, and produce graphs and figures using R.
5. Be thoughtful and critical consumers of statistical analyses.

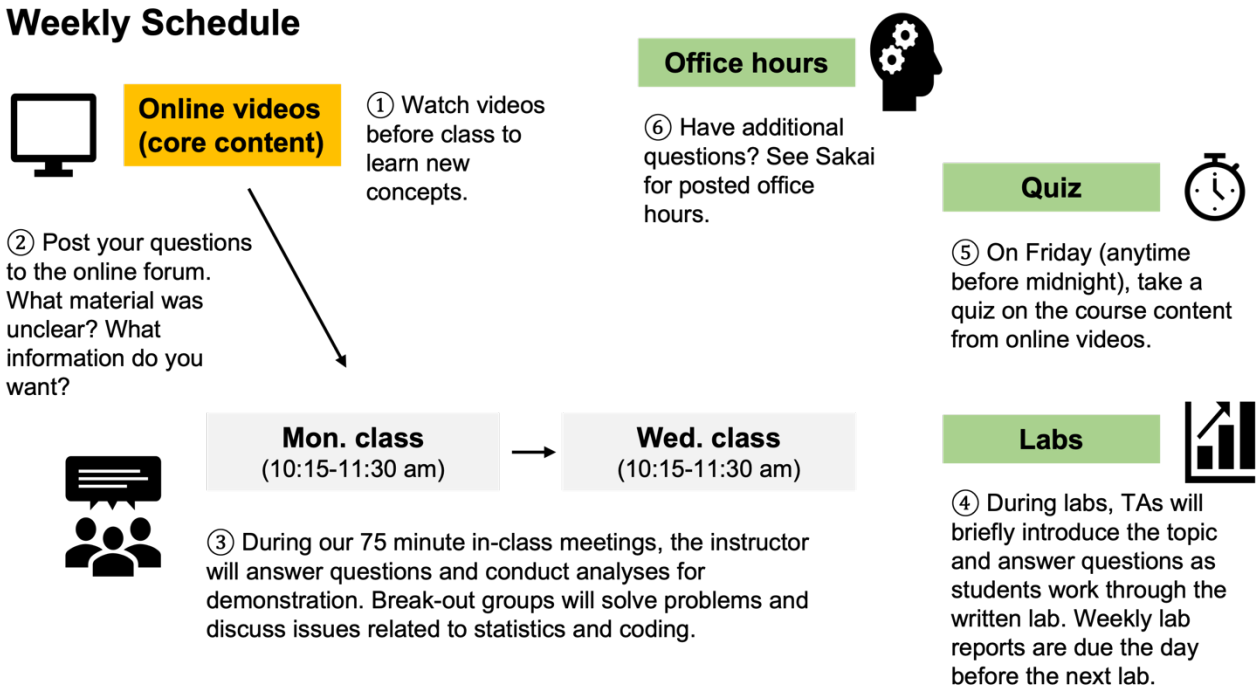
For those of you who like checklists, here you go…

- Probability and probability distributions
- Graphical and exploratory data analysis
- Inference using t-distributions
- Type I and II errors, power analysis
- Confidence intervals
- Hypothesis testing
- Randomization (permutation) techniques
- Experimental design
- Linear models: ANOVA, simple linear regression and multiple regression
- Ordinary least square and maximum likelihood estimation
- Generalized linear models
- Model testing and comparison
- Hierarchical models (mixed effect models)
- Use of a statistical computing software package (R)

What will I do in this course?

"A bad plan is better than no plan, and the most important quality of any plan is the flexibility to change."
– Anonymous from the internet

I have a plan… but reserve the right to adapt this plan if necessary. I ask for your patience and flexibility.



Each week, you will watch 2-3 short videos that provide the core information to be learned during the week. Sometimes there are mini-quizzes embedded in the videos. Please complete these online quizzes. They are not graded but will help assess your learning and the effectiveness of the videos (anonymous results are e-mailed to the instructor). If you have specific questions regarding the content from the videos, please post them on the online forum so they can be addressed during class meetings. Class time will be dedicated to answering questions and working in break-out groups for discussions and problem solving.

We will randomly assign students to break-out groups. The composition of break-out groups will be changed after the first exam (mid way through the semester) so that everyone has the opportunity to get to know each other and work together.

Each week, you will have a lab to complete by the day before your next lab meeting. TA's will lead lab sessions to discuss the lab topic, new R commands and code, and answer questions.

By midnight on Friday, you will take a quiz that covers the content from online videos. Quizzes can include solving problems, writing or interpreting R code, or answering questions regarding statistical theory and knowledge. The point of the quizzes is to assess understanding and help students keep up with the class.

If you have additional questions, please attend office hours. Both the instructor and TA's will have office hours during the week. Come to office hours with specific questions.

See more details for each of these activities below in "How will my grade be determined".

How can I prepare for the class sessions to be successful?

Applied statistics is best learned by doing. In both lecture and lab, some class time will be spent working in groups on data analysis. For this format to be successful, you must be prepared for class. Which means…?

Which means that you have completed all assigned videos or reading before class meetings. During class we will discuss analyses and run R code; therefore, be ready to work in-class problems on your computer.

If you didn't understand something on the videos, you probably aren't alone. Please submit questions to the online forum to be answered in class.

We will rely heavily on the Sakai course management system for information distribution, communication, online quizzes and homework.

The best way to learn is to plunge into data analysis with fearless abandon and lots of patience. This leads us to R. Learning to write code is analogous to learning a foreign language. Like writing a sentence or paragraph, there are grammatical rules that you must follow. There are also multiple ways to express the same idea. Some people will find it easy, others not so much. The best advice to becoming proficient at R is practice… practice… practice.

What required texts, materials, and technologies will I need?

Text (helpful, but not required): There are no required texts for this class. I am working on a text for the course, but for now the below background reading can be helpful.

Michael C. Whitlock and Dolph Schluter. 2015. The Analysis of Biological Data, 2nd Edition. New York, NY: W.H. Freeman & Co.

Software (required): We will be using R for manipulating data and conducting analysis. R and RStudio (see below) are free, open access tools. I provide information on how to download R in class, but it can also be found here:

R: an open source implementation of the S Language, available from CRAN at http://cran.r-project.org.

It isn't necessary to buy a text on R as there are literally hundreds of websites, blogs and other resources for learning the program. However, if you prefer to read a book, this suggested text focuses on using R to learn statistics.

Reference Text for R (not required): Eric Goh Ming Hui. 2019. Learn R for Applied Statistics: With Data Visualizations, Regressions, and Statistics. Apress. http://doi.org/10.1007/978-1-4842-4200-1

We will use RStudio as a shell or gui for R (https://rstudio.com/products/rstudio/download/). We provide information on how to use this in class as well so don't worry about downloading it ahead of time. Within RStudio, we will use R Markdown https://rmarkdown.rstudio.com/ for creating reports and lab briefs. R Markdown embeds your text, code, results and figures in the same document so that your work is elegantly formatted and fully reproducible.

We may use Zoom and Sakai Forums for online discussions. Please contact NSOE IT if you have problems with Zoom, Sakai, or internet connectivity.

How will my grade be determined?

Grades will be assessed based on the below activities.

Lab briefs: Labs are a combination of problem sets and data analysis using R. All labs should be completed individually, unless the assignment instructions specifically state otherwise. Lab write-ups are graded both on getting the correct answer and presenting the information in a clear, concise, grammatically correct and visually attractive manner. All labs must be written in R Markdown and knitted to a pdf document. Think of the lab write-ups as technical briefs that you are providing to your boss on a specific subject.

Participation: Participate actively in class (or Zoom meetings) – it's more fun and you learn more. Participation will be assessed through questions and engagement in class, attendance in labs, and the group project presentation (see Group Project below).

Quizzes: I hated quizzes when I was a student and assume you do too. Alas, after having taught this course a few times, I see their relevance. Online quizzes will test you on the reading and/or video assignments. Rather than treating them as a source of anxiety, try to view them as a self-assessment of your understanding. Quizzes are closed-book and timed, so you have 20 minutes to complete each quiz. Online quizzes will be posted on Tuesday and must be completed by Friday at midnight.

Exam 1: The first exam will test your understanding of probability and statistical principles. The exam will be an open book, take home exam with a 24-hour limit: you can use your notes and online resources, but you are prohibited from getting help from any other person (living or dead). There will be problems that require you to write out chunks of R-code.

Group Project: You will work in groups of 2-3 students to conduct a group project: the members of your Group Project will be assigned by the instructor. Together, you will develop a research question and obtain and analyze a dataset employing the statistical tools we cover in the course. The dataset must be multivariate and the statistical analysis must include, at a minimum, a linear or generalized linear model with multiple covariates.

- Mid-semester, groups will first submit an abstract that describes their research questions and dataset (the abstract will be graded as a lab brief). The dataset must be attached to the abstract as an appendix.
- At the end of the semester, groups will submit the full paper in which they present their findings.
- After submitting the full paper, each group member will individually submit a one-page project reflection.
- A description of each of these assignments is provided in the group-project-instructions.pdf in the Group Projects folder under Resources on Sakai.

Grading: Coursework will be weighted as follows:

Labs & Group Project Abstract	35%
In-class participation	5%
Online quizzes*	10%
Exam 1	20%
Group Project	30%

Final scores will be the weighted average of the five categories listed above. Course grades will be curved based on the final score: 100 (A+), 93 (A), 90 (A-), 87 (B+), 83 (B), 80 (B-), 77 (C+), 73 (C), 70 (C-), 67 (D+), 63 (D), 60 (D-). A passing grade is 60% or better.

Late assignments will be reduced by 5% of possible points for each day they are late.

*For online quizzes, we will disregard the fully completed quiz with the lowest score. Please note that any incompleted quizzes will be counted as a 0 in the grade assessment.

Teaching Assistants will grade labs based on clearly defined grading rubrics. If you have a question about the grading or your grade, please refer your question to the Instructor not the TAs.

What are the course policies?

Communications

We will use Sakai Conversations for communications. Conversations is a learning management system that is directly linked to Sakai and that allows students to ask questions in a forum-type format. Students can publicly and anonymously ask questions, answer questions, and post notes. Each question prompts a collective answer to which any user can contribute and the instructor answer. This will allow all students to learn from questions that are posted, whether they concern statistical ideas or logistical questions like 'should we assume that the data are normally distributed for problem #2'.

You should check the course Sakai site and Conversations frequently for announcements from me and the TAs and potential answers to questions that may arise.

Discussion Guidelines

Civility is an essential ingredient for academic discourse. All communications for this course should be conducted constructively, civily, and respectfully. Differences in beliefs, opinions, and approaches are to be expected. Please bring any communications you believe to be in violation of this class policy to the attention of your instructor. Active interaction with peers and your instructor is essential to success in this course, paying particular attention to the following:

- Be respectful of others and their opinions, valuing diversity in backgrounds, abilities, and experiences.
- Challenging the ideas held by others is an integral aspect of critical thinking and the academic process. Please word your responses carefully, and recognize that others are expected to challenge your ideas. A positive atmosphere of healthy debate is encouraged.
- Read your online discussion posts carefully before submitting them.

Diversity Statement

It is my intent that students from all diverse backgrounds, perspectives, and abilities be well served by this course, that students' learning needs be addressed both in and out of class, and that the diversity that students bring to this class be viewed as a resource, strength and benefit. It is my intent to present materials and activities that are respectful of diversity: gender, sexuality, disability, age, socioeconomic status, ethnicity, race, and culture. Your suggestions are encouraged and appreciated. Please let me know ways to improve the effectiveness of the course for you personally or for other students or student groups.

Data Acknowledgement

In this class we use a variety of different datasets selected because they have the appropriate numerical and design characteristics to illustrate data analysis and statistical concepts. In dealing with these data, we often overlook their context, and therefore, explicit and implicit biases as well as sensitive issues regarding gender, race, sexual orientation and diversity. The intent is not to ignore or minimize these issues of paramount importance, but rather to focus on the course objectives of understanding and using methods of data analysis.

Academic Integrity

As a student, you should abide by the academic honesty standard of the Duke University. Its Community Standard states: Duke University is a community dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Citizens of this community commit to reflect upon and uphold these principles in all academic and nonacademic endeavors, and to protect and promote a culture of integrity.

To uphold the Duke Community Standard:

- I will not lie, cheat, or steal in my academic endeavors;
- I will conduct myself honorably in all my endeavors; and
- I will act if the Standard is compromised.

Academic Policy & Procedures

You are responsible for knowing and adhering to academic policy and procedures as published in the Duke Community Standard Guide. Please note, an incident of behavioral infraction or academic dishonesty (cheating on a test, plagiarizing, etc.) will result in immediate action from me, in consultation with university administration (e.g., Dean of Undergraduate Studies, the Office of Student Conduct, Academic Advising).

Academic Accommodations

If you need to request accommodation for a disability, you need to contact the Disability Management System (DMS) office. I will work with that office to provide you with equal access to course materials and make accommodations for exams and other assessments.

What university resources can help me during this course?

Student Support

Please consult with me about appropriate course preparation and readiness strategies, as needed. The Academic Resource Center (ARC) offers free services to all students (e.g., peer tutoring, learning consultations, ADHD/LD coaching, and help with online learning.

If you are concerned about your physical or mental health DukeReach can connect you with departments across campus to get you help and you can contact CAPS directly for counseling services.

Academic Advising

Consult your academic advisors on course performance (i.e., poor grades) and academic decisions (e.g., course changes, incompletes, withdrawals) to ensure you stay on track with degree and graduation requirements. The Academic Advising Center can help you navigate who to contact. The university publishes a full list of academic policies for undergraduates to review.

A few last comments…

Understanding statistics is a valuable life skill, having the ability to analyze data, particularly in R, is an employable skill and valuable for many careers. This class is designed to provide students with the knowledge and tools to analyze data for their masters projects and dissertations. Hopefully it achieves that goal and inspires some students to go even further into more advanced statistical modeling. In addition to being a core class for the MEM degree, the class also satisfies MF requirements.

MF course learning outcomes

ENV 710 is a required course for Duke's Master of Forestry (MF) program. The Society of American Foresters (SAF), which accredits the MF program, requires course syllabi to list course learning outcomes (CLOs) related to forestry competency areas. CLOs in ENV 710 include Competencies B: Measurement of Forest Resources, 2. Increased ability to use appropriate, contemporary technology.

See the expected course schedule below.

What is the expected course schedule?

Week	Date	Topics	Pre-class work	In-class activities	Assignments
1	1/11	L0: Introduction to course	Class Introduction	Zoom meeting	**Online quiz** (due Fri. before 12 am)
2	1/18	L1: Descriptive statistics	Vid-1: Descriptive statistics	Group work	**Online quiz** (due Fri. before 12 am)
		Lab 1: Intro. to R			
3	1/23	L2: Discrete probability distributions	Vid-2: Bernoulli random variables Vid-3: Binomial distribution	Group work	**Online quiz** (due Fri. before 12 am)
	1/25	L3: Discrete probability distributions	Vid-4: Poisson distribution	Group work	
		Lab 2: Probability distributions and descriptive statistics			
4	1/30	L4: Continuous probability distributions	Vid-5: Uniform distribution Vid-6: Normal distribution	Group work	**Online quiz** (due Fri. before 12 am)
	2/1	L5: Sampling, Central Limit Theorem	Vid-7: Sampling Vid-8: Central Limit Theorem	Group work	
		Lab 3: Sampling			**Lab Report 2** (due day before Lab 3)
5	2/6	L6: Statistical inference	Vid-9: t-distribution Vid-10: Confidence intervals Krzywinski & Altman 2013	Group work	**Online quiz** (due Fri. before 12 am)
	2/8	L7: Hypothesis testing, p-values	Vid-11: Hypothesis testing Vid-12: p-value	Group work	
		Lab 4: Exploring confidence intervals			**Lab Report 3** (due day before Lab 4)
6	2/13	L8: Assumptions and transformations	Vid-13: Transformations	Group work	**Online quiz** (due Fri. before 12 am)
	2/15	L9: Type I and II errors, power	Vid-14: Types of error	Group work	
					Lab Report 4 (due 2/24)

7	2/20	L10: Experimental design	Vid-15: Experimental design	Group work	
	2/22	**Exam 1**	**Exam on L1-L10**		

8	2/27	L11: Linear models	Vid-16: Linear Models-1	Group work	**Online quiz** (due Fri. before 12 am)
	3/1	Group Project		Group work	
		Lab: Work on Group Projects	**Group Project Abstract**		**Group Project Abstract**

9	3/6	L12: Linear models	Vid-17: Linear Models-2	Group work	**Online quiz** (due Fri. before 12 am)
	3/8	L13: Linear models: Assumptions	Vid-18: Assumptions of linear models	Group work	
		Lab 5: Linear models I			

Spring Break

11	3/20	L14: Linear models: Multiple variables	Vid-19: Multivariate models	Group work	**Online quiz** (due Fri. before 12 am)
	3/22	L15: Hierarchical models: Fixed and random effects	Vid-20: Fixed and random effects	Group work	
		Lab 6: Linear models II			**Lab Report 5** (due day before Lab 6)

12	3/27	L16: Model assessment and comparison	Vid-21: Model comparison *Johnson & Omland *TREE* 2004	Group work	**Online quiz** (due Fri. before 12 am)
	3/29	L17: Generalized linear models – Poisson regression	Vid-22: Generalized linear models Vid-23: Poisson regression O'Hara and Kotze 2010	Group work	
		Lab 7: Poisson regression	**Introduction & Outline to Group Project**		**Lab Report 6** (due day before Lab 7) **Group Project Intro due 3/31**

13	4/3	L18: Generalized linear models – Poisson regression		Group work	
	4/5	L19: Generalized linear models – logistic regression	Vid-24: Logistic regression	Group work	**Online quiz** (due Fri. before 12 am)

		Lab 8: Logistic regression			**Lab Report 7** (due day before Lab 8)

14	4/10	L20: Generalized linear models – logistic regression	Vid-25: Binomial regression	Group work	
	4/12	L21: Generalized linear mixed models	Vid-26: Generalized linear mixed models - 1	Group work	
		Lab 9: Generalized linear mixed models			**Lab Report 8** (due day before Lab 9)

15	4/17	L22: Wrap-up			
	4/19	L22: Wrap-up			**Lab Report 9** (due 4/19)

		– No Final Exam – Group Papers Due 11:59 pm, April 21			