Bio 409S—Data Analysis for Biologists
Semester, Year

Curricular Codes: NS, S
Class Time: Date, Times
Location: TBD
Instructor: Jesse Granger, (she/her)
Preferred name: Professor Granger
Email: jesse.granger@duke.edu
Office Hours: TBD and by appointment
Office Hour Location: Biological Sciences 301 and *Zoom-Link* by appointment

Course Overview:
So, you collected your data—now what? Being a biologist requires careful experimental design, effective and compelling data presentation, and statistical testing to extract meaningful conclusions. All of these skills are a part of data analysis. This course focuses on helping you develop those essential data analysis tools required to be a biologist and learn to be a critical consumer of data in your daily life.

Students will gain quantitative reasoning skills through the use of data visualization, statistical analyses, and mathematical modeling, as well as become proficient in the programming language R. The goal of this course is to provide students with the necessary skills and tools to analyze their own data, as well as critically evaluate data analyses done by others.

This course is open to any STEM student but is designed for biologists and will focus on biological questions and case studies. There are no prerequisite classes for this course. It is not assumed that students will have taken a statistics course in the past, nor are students required to have any previous coding experience. Students will be expected to have access to laptop with the free statistical programming software R installed on it, and to bring it with them to class.

Course Objectives:
By the end of this course students will be able to—

- Articulate the limitations of statistical tests and discuss the ways in which differing data presentations affect how data are interpreted.
- Evaluate a dataset for quality and internal biases and identify flaws in experimental designs.
- Apply appropriate data visualization and statistical tests to existing datasets using R to communicate key findings to a general audience.

Course Resources:

- **Contact:** The course has an accompanying class Slack channel where you can discuss problem sets or questions with fellow classmates. It is encouraged that you seek help from your classmates with problem sets first; however, do not hesitate to also contact the instructor via email with any questions or concerns. Expect a response within two business days.
- **Coding Resources:** The Duke Center for Data and Visualization Sciences hosts weekly walk-in hours, as well as by-appointment meetings where you can get advice on data projects and other computing related problems. They also host a series of additional resources for coding in R: https://library.duke.edu/data/tutorials.

Assignments:

- **Attendance/In-Class Assignments:** Students are expected to attend and participate in every class meeting. Course attendance will be assessed through in-class assignments. There will be a series of 20 in-class assignments given randomly throughout the semester, to be submitted at the end of class, graded for completion. Students will be allowed to skip one with no penalty, and 1 point (out of 20) will be deducted for every additional missed assignment.
- **Problem Sets:** You can download the problem sets from Sakai at 5am every Saturday and the answers must be uploaded to Sakai every Monday at 10am. These problem sets will be graded for correctness. *Although collaboration with other classmates is encouraged, submitted work must be written up individually.* Late work will not be accepted.
- **Assessments:** There will be two assessments. These will be take-home, open note, and can be completed over the course of one week. *Collaboration with other classmates is not allowed.* Space will be provided by request for students in need of a distraction-free location to complete their assessments. Rubrics for both assessments will be available on Sakai and can be found in the appendix of this syllabus.
- **Final Project:** At the end of the semester, students will present and submit a final project. The project will involve using existing data to answer a hypothesis, testing the hypothesis using the proper statistical tests, and presenting conclusions using appropriate graphical figures. Students will give a short in-class oral presentation and submit a written 2-page project-summary. Rubrics for both the oral and written portion of this project will be available on Sakai and can be found in the appendix of this syllabus. Students are welcome to use their own data for this if they wish, and if not, a repository of existing data sets will be provided to choose from.
- **Large Language Models:** Students are allowed to use large language models (such as chatGPT) on all assignments; however, they must CITE their usage in the following manner (chatGPT assisted) and be prepared to potentially be asked to share their workflow with the instructor or the rest of the class.

Grading:

Final grades will be computed as follows:

In Class Assignments:	20%
Problem sets:	30%
Assessment 1:	15%
Assessment 2:	15%
Final Project:	20%

Diversity, Equity, and Inclusion:

We learn best when we feel safe and welcome. I strive to foster a classroom environment where students with diverse backgrounds, identities, and experiences can learn and thrive. If there is anything that you feel I need to know to facilitate your academic success, please do not hesitate to contact me.

Diverse perspectives promote scientific discovery and intellectual stimulation in the classroom, and nowhere is this truer than the realm of data analysis. Throughout the course, I will highlight the ways in which the fields of data analysis, statistics, and biology has been advanced by diverse scientists, as well as highlight the important and complicated racial history of the field of statistics.

Campus Resources:

- **Accessibility Services:** Students with learning differences who believe that they may need accommodations in the class are encouraged to contact the Student Disabilities Access Office at 919.668.1267 or disabilities@aas.duke.edu as soon as possible to better ensure that such accommodations are implemented in a timely fashion.
- **Wellness Services:** Students in need are encouraged to reach out to the Counseling and Psychological Services (CAPS) department or Duke Reach. In addition, DuWell hosts activities to promote holistic wellness, including guided meditations, yoga, and meditative art.
- **Learning Resources:** The Duke Academic Resource Center provides services like learning consultations, peer tutoring, learning differences support, and study groups
- **Resources for International Students:** International students can contact the Visa Services Office at visahelp@mc.duke.edu.
- **Resources for Student-Athletes:** Student-Athletes can find academic support resources from Student-Athlete Academic Support Services.

Syllabus:

If you have read this syllabus carefully and completely, you can receive 10-pts back on any problem set by emailing the instructor a picture of a cute animal with the subject line, “Syllabus.”

Academic Integrity

You are expected to uphold the Duke Community Standard in this course. The community standard is reproduced below:

> *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 non-academic 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.*

Bio 409S—Data Analysis for Biologists
Semester, Year

Tentative Calendar, (subject to change):

Date	Topics	In class projects	Assignment Due Dates
Week 1	-Introductions -Intro to R	Day 1: Introductions / Syllabus Day 2: Group practice with R	Day 1: Read the syllabus Day 2: Read Getting Started with R and R Markdown and R Notebooks
Week 2	-Data Visualization using ggplot2 -Biases in Data Visualization	Day 1: Group practice plotting with ggplot2 Day 2: Case Studies with data visualization (@GraphCrimes on twitter)	Homework 1—Due Mon. at 10am Day 1: Watch "visualization in R using ggplot" video (link to data and code files) OR read Introduction to ggplot2 Day 2: Read Chart Design Principles, Data Vis Rules, and **skim** Chapter 14 "Detect Lies and Reduce Bias"
Week 3	-Biases in Data Visualization (cont') -R basics	Day 1: Case Studies with data visualization (cont') Day 2: Group practice with For loops, ifelse and functions	Homework 2—Due Mon. at 10am Day 1: Make your own "graph crime" and bring to class Day 2: Read For loops, ifelse, and functions (optional additional reading with for loops)
Week 4	-Data Wrangling using Tidyverse and dplyr	Day 1: Group practice with dplyr Day 2: Group practice with data wrangling	Homework 3—Due Mon. at 10am Day 1: Read Intro to dplyr Day 2: Read Data wrangling
Week 5	-Review -Hypothesis testing, p-values, and confidence intervals	Day 1: Review and practice problems Day 2: Case studies on p-value manipulation, group exercises with hypothesis testing and calculating confidence intervals	Homework 4—Due Mon. at 10am Day 1: **Assessment One Opens** Day 2: Read "statistics" and "hypothesis testing", "The central limit theorem" "Confidence Intervals" and "Standard Errors and Confidence Intervals". Read **one** of the following: Option 1, Option 2
Week 6	-Hypothesis testing using t-tests -Hypothesis testing	Day 1: 2-sample t-tests by hand, group practice using different kinds of t-tests. Design your own experiments for each kind of t-test	Homework 5 (SHORT)—Due Mon. at 10am Day 1: Read "t-tests", SKIM "calculating t-tests" then read and follow the examples from the following articles (make your own data where you see fit): 1) one sample t-tests, 2) two sample t-tests, 3) paired t-tests

	(ANOVA, and flow charts)	Day 2: Practice with Anovas, and picking statistical tests	**Assessment One—Due Wednesday at 10pm** Day 2: Read ANOVAs, Statistical tests, and skim Choosing the correct statistical test
Week 7	-Biases in hypothesis testing in scientific papers -Intro to regression analysis	Day 1: Group practice in identifying bias (find an example of biased statistics in the news) Day 2: Group practice with Linear Regressions	Homework 6—due Mon. at 10am Day 1: Read p-hacking and post-hoc analyses, and The Misuse of Statistics and Data in the Digital Age. Read **two** of the following **carefully** and **skim** the others: One, Two, Three, Four Day 2: Read Intro to Linear Regressions, Linear Regression in R (skim the sections on multiple regression) and R Linear Regression Tutorial, Skim Chapter 3 A and B notes (on Sakai). Additional Resources: One, Two, Three
Week 8	-Multiple Regressions -Model Selection	Day 1: Group practice with multiple regressions Day 2: Group practice with model selection	Homework 7: Due Mon. at 10am Day 1: Read "Intro to Regression in R" and "Multiple Regression made simple", Skim Chapter 4A (on Sakai), and follow along with the examples in **one** of the following tutorials: One, Two, Three Day2: Read **one** of these two articles on nested models: One, Two. Read Linear Model Selection and Stepwise Selection, and Skim chapter 4B (on Sakai)
Week 9	-Logistic Regression analysis -The history of statistics	Day 1: Group practice with logistic regressions Day 2: Discussion from the readings,	Homework 8: Due Mon. at 10am Day 1: Read 'Intro to Logistic Regression' One and Two, and "Logistic Regressions in R". **Submit proposed dataset for final project** Day 2: Read "How Eugenics Shaped Statistics," read the first chapter of "Normality: A Critical Genealogy" and listen to this episode of Statistically Insignificant **Homework 9 (SHORT—Unusual due date to get comments back before assessment two is due): Friday at 10am**
Week 10	- Review - Final Project Workshop	Day 1: Review and practice problems	Day 1: **Assessment Two Opens**

		Day 2: Group edits on final project rough draft	Day 2: Submit project rough draft and bring a printed copy to class
Week 11	- Student's choice	Day 1: TBD Day 2: TBD	**Assessment Two—Due Monday at 10pm** Day 1: TBD Day 2: TBD. Submit written copy of final project
Week 12	-Project Presentations	Presentations	Day 1: Upload a copy of your PowerPoint slides to Sakai Day 2: Submit course evaluation

Student's Choice: Students will vote to cover two of any of the following topics: MANOVA, PCA's, Random Forest, time series, circular data, mixed-effect models, nearest-neighbors, intro to Bayesian statistics, or alternative topics of their choosing

Appendix 1: Rubric for Assessment 1—Figure Design and Data Wrangling

Grade:	4/4	3/4	2/4	1/4
Diagrams	Diagrams and/or sketches are clear and greatly add to the reader's understanding. They follow standard chart design principles/data visualization rules.	Diagrams and/or sketches are clear and easy to understand. They follow standard chart design principles/data visualization rules.	Diagrams and/or sketches are somewhat difficult to understand. They violate standard chart design principles/data visualization rules in one instance.	Diagrams and/or sketches are difficult to understand or are not used. They violate standard chart design principles/data visualization rules in multiple instances.
Completion	All problems are completed and match the given solution.	All problems are completed but deviate from the given solution in less than five places without being incorrect.	All problems are completed but deviate from the given solution in more than five places without being incorrect.	Any problem is not completed or is incorrect.
Strategy / Procedures	Typically, uses an efficient and effective strategy to solve the problem(s). All code is replicable and clearly commented.	Typically, uses an effective strategy to solve the problem(s). All code is replicable and clearly commented.	Sometimes uses an effective strategy to solve problems but not consistently. One or two parts of the code are not replicable or are not commented.	Rarely uses an effective strategy to solve problems. Multiple parts of the code are not replicable or are uncommented.
Terminology and Notation	Correct terminology and notation are always used, making it easy to understand what was done.	Correct terminology and notation are usually used, making it fairly easy to understand what was done.	Correct terminology and notation are used, but it is sometimes not easy to understand what was done.	There is little use, or a lot of inappropriate use, of terminology and notation.
Explanation	Explanation is detailed and clear.	Explanation is clear.	Explanation is a little difficult to understand but includes critical components.	Explanation is difficult to understand and is missing several components OR was not included.
Neatness and Organization	The work is presented in a neat, clear, organized fashion that is easy to read.	The work is presented in a neat and organized fashion that is usually easy to read.	The work is presented in an organized fashion but may be hard to read at times.	The work appears sloppy and unorganized. It is hard to know what information goes together.

Bio 409S—Data Analysis for Biologists
Semester, Year

Appendix 2: Rubric for Assessment 2—Statistical tests and regression analysis

CATEGORY	4	3	2	1
Statistical Concepts	Explanation shows complete understanding of the statistical concepts used to solve the problem(s).	Explanation shows substantial understanding of the statistical concepts used to solve the problem(s).	Explanation shows some understanding of the statistical concepts needed to solve the problem(s).	Explanation shows very limited understanding of the underlying concepts needed to solve the problem(s) OR is not written.
Statistical Errors	90-100% of the steps and solutions have no statistical errors.	Almost all (85-89%) of the steps and solutions have no statistical errors.	Most (75-84%) of the steps and solutions have no statistical errors.	More than 75% of the steps and solutions have statistical errors.
Explanation	Explanation is detailed and clear.	Explanation is clear.	Explanation is a little difficult to understand, but includes critical components.	Explanation is difficult to understand and is missing several components OR was not included.
Neatness and Organization	The work is presented in a neat, clear, organized fashion that is easy to read.	The work is presented in a neat and organized fashion that is usually easy to read.	The work is presented in an organized fashion but may be hard to read at times.	The work appears sloppy and unorganized. It is hard to know what information goes together.
Diagrams and Sketches	Diagrams and/or sketches are clear and greatly add to the reader's understanding of the procedure(s).	Diagrams and/or sketches are clear and easy to understand.	Diagrams and/or sketches are somewhat difficult to understand.	Diagrams and/or sketches are difficult to understand or are not used.
Completion	All problems are completed and match the given solution.	All problems are completed but deviate from the given solution in less than five places without being incorrect.	All problems are completed but deviate from the given solution in more than five places without being incorrect.	Any problem is not completed or is incorrect.
Statistical Terminology and Notation	Correct terminology and notation are always used, making it easy to understand what was done.	Correct terminology and notation are usually used, making it fairly easy to understand what was done.	Correct terminology and notation are used, but it is sometimes not easy to understand what was done.	There is little use, or a lot of inappropriate use, of terminology and notation.
Strategy / Procedures	Typically, uses an efficient and effective strategy to solve the problem(s). All code is replicable and clearly commented.	Typically, uses an effective strategy to solve the problem(s). All code is replicable and clearly commented.	Sometimes uses an effective strategy to solve problems but does not do it consistently. One or two parts of the code are not replicable or are not commented.	Rarely uses an effective strategy to solve problems. Multiple parts of the code are not replicable or uncommented.

Appendix 3: Rubric for Oral Presentation of Final Project

Grade:	4/4	3/4	2/4	1/4
Comprehension	Student is able to accurately answer almost all questions posed by classmates about the topic.	Student is able to accurately answer most questions posed by classmates about the topic.	Student is able to accurately answer a few questions posed by classmates about the topic.	Student is unable to accurately answer questions posed by classmates about the topic.
Preparedness	Student is completely prepared and has obviously rehearsed.	Student seems pretty prepared but might have needed a couple more rehearsals.	The student is somewhat prepared, but it is clear that rehearsal was lacking.	Student does not seem at all prepared to present.
Time-Limit	Presentation is 5-6 minutes long.	Presentation is 4 minutes long.	Presentation is 3 minutes long.	Presentation is less than 3 minutes OR more than 6 minutes.
Content	Shows a full understanding of the topic. Contains each of the required sections and applies methodologies correctly.	Shows a good understanding of the topic. Contains each of the required sections and applies methodologies correctly.	Shows a good understanding of parts of the topic. Is missing one of the required sections but methodologies are applied correctly.	Does not seem to understand the topic very well. Is either missing multiple of the required sections or applies methodologies incorrectly.
Volume	Volume is loud enough to be heard by all audience members throughout the presentation.	Volume is loud enough to be heard by all audience members at least 90% of the time.	Volume is loud enough to be heard by all audience members at least 80% of the time.	Volume often too soft to be heard by all audience members.
Stays on Topic	Stays on topic all (100%) of the time.	Stays on topic most (99-90%) of the time.	Stays on topic some (89%-75%) of the time.	It was hard to tell what the topic was.

Appendix 4: Rubric for Written Submission of Final Project

Grade:	4/4	3/4	2/4	1/4
Organization	Information is very organized with well-constructed paragraphs and subheadings.	Information is organized with well-constructed paragraphs.	Information is organized, but paragraphs are not well-constructed.	The information appears to be disorganized. 8)
Amount of Information	All topics are addressed and all questions answered with at least 2 sentences about each.	All topics are addressed and most questions answered with at least 2 sentences about each.	All topics are addressed, and most questions answered with 1 sentence about each.	One or more topics were not addressed.
Quality of Information	Information clearly relates to the main topic. It includes several supporting details and/or examples.	Information clearly relates to the main topic. It provides 1-2 supporting details and/or examples.	Information clearly relates to the main topic. No details and/or examples are given.	Information has little or nothing to do with the main topic.
Sources	All sources (information and graphics) are accurately documented in the desired format.	All sources (information and graphics) are accurately documented, but a few are not in the desired format.	All sources (information and graphics) are accurately documented, but many are not in the desired format.	Some sources are not accurately documented.
Mechanics	No grammatical, spelling or punctuation errors.	Almost no grammatical, spelling or punctuation errors	A few grammatical spelling, or punctuation errors.	Many grammatical, spelling, or punctuation errors.
First Draft (in class, week 10)	Detailed draft is neatly presented and includes all required information.	Draft includes all required information and is legible.	Draft includes most required information and is legible.	Draft is missing required information and is difficult to read.
Diagrams & Illustrations	Diagrams and illustrations are neat, accurate and add to the reader's understanding of the topic.	Diagrams and illustrations are accurate and add to the reader's understanding of the topic.	Diagrams and illustrations are neat and accurate and sometimes add to the reader's understanding of the topic.	Diagrams and illustrations are not accurate OR do not add to the reader's understanding of the topic.