Syllabus for Foundations of Biostatistics (BIOL 350)

Fall 2023

Gregg Hartvigsen

Version date: August 28, 2023

"Knowledge is power." Francis Bacon (1561-1626)
"To measure is to know." Lord Kelvin (1824-1907)
"Knowledge is good." Emil Faber. Founder of Faber College (see documentary: Animal House, 1978)
"Knowledge comes from collecting, analyzing, visualizing, and interpreting information correctly." Gregg Hartvigsen

Item	Details
Meetings	Tues/Thurs 2:00 - 3:15, ISC 115
Contact me	hartvig@geneseo.edu and ISC 360 (office).
Office hrs	Mon. 10:00am – 11:00am (Zoom only); Tues. 8:00 – 9:00am (ISC 360); Wed. 9:00am – 10:30am (Zoom only). Join me at `https://geneseo.zoom.us/j/4333683209`, passcode = 497118. For details see Section 10 on page 10.
Required textbook	Primer in Biological Data Analysis, 2nd ed. (Hartvigsen, 2021). I recommend you bring this to class and, particularly, statistical challenges.
Required laptop (Mac, Windows, or Linux)	Bring to all mtgs. Note that all seats have a plug. A Chromebook is not be sufficient to complete exercises in this class (see `https://wiki.geneseo.edu/display/cit/Student+Laptop+Requirement`).
Required software (free)	`Excel` (or similar), `R` (version 4.3.1 or later), `RStudio` (version 2023.06 or later), a PDF viewer, and LaTeX (MacTex for Macs and TexLive for Windows and/or Linux).
Cloud storage (free)	You are expected to keep all files from this class in a folder that automatically syncs in the cloud. This can be through "Google Drive" (recommended), Dropbox, OneDrive, or iCloud. This protects you from having to utter the words "I lost my work because my computer died."
Calculator	Absolutely forbidden.

Note: ALL COURSE MATERIALS ARE KEPT IN MY OUTBOX, NOT BLIGHTSPACE! Go to the following to learn how:
`https://geneseo.atlassian.net/wiki/spaces/HELP/pages/76778990/Inboxes+and+Outboxes`

1 Overview

This is a course on empowerment and discovery in biology. It's designed to be an introduction to the management, visualization, analysis, and interpretation of biological data. The process of understanding what data are telling us requires proficiency in a range of fundamental procedures, from the design of experiments through the generation of results to the interpretation of those results in a biological context to the presentation of data and ideas. Your ability to challenge ideas with data should be helpful and inspiring to you. If you seek to understand and solve problems in science or critically assess the meaning of biological information (e.g., as a doctor, PA, ecologist,

or reader of the news and interpreter of statements made by politicians) then you are in the right place.

In this course you will survey a variety of statistical approaches which should help you be able to think more creatively. Biological data analysis is, however, an ever-expanding area of inquiry. New methods are being developed daily. New data also are creating the need for new types of analytical tools. You will be learning to use the program `R` to perform your analyses and create professional-quality visualizations.

My goal is for you to learn how to solve problems. This will require that you push yourself to solve your own problems. I am glad to help but part of your learning will involve figuring out how to find answers to your questions independently.

You, no doubt, will forget how the data are supposed to be entered for a particular test or how to do something cool when making a graph. But you don't have to remember everything because you'll develop *many* script files that you can simply reuse. Be sure to name your files appropriately (do not name your work on a t-test "`lab7.r`"). What I hope you get (and keep) is the attitude that you can fearlessly figure out how to solve problems.

A peek at what happens during a typical class

We often begin discussing a few basic principles for the day regarding statistics, visualizations, experimental design, and/or the interpretation of information, most of which the book will have already helped get you started. We then apply this information to solve an in-class exercise (ICE) using `R`.

2 Expected Learning Outcomes

If you successfully complete this class your ability to critically assess commonly encountered biological information will improve. Specifically, you will be able to:

1. **manage** a wide range of data sets. This will involve finding, entering, saving, organizing, and manipulating data. To accomplish this you will learn how to use both `Excel` and `R`;
2. **implement** basic experimental design principles that are used to answer a variety of biological questions;
3. **correctly develop and test** hypotheses statistically using data;
4. **explain** the differences between parametric and non-parametric statistical tests;
5. **create** appropriate publication-quality visualizations of a wide range of results;
6. **correctly identify and report**, in appropriate scientific format, the results of a statistical test;
7. **write a professional-looking report** using `Sweave` and LaTeX (pronounced lay'-tek, not lay-teks)s;
8. **explain** the meaning of quantitative results (and accompanying visualizations), whether done by you or others (e.g., in books, papers, presentations, and news media).
9. **write and use** your own functions;
10. **write computer programs** in an object-oriented programming language, including "for" and "while" loops, and "if" conditional statements.

These outcomes are aligned with the classic report "Scientific Foundations for Future Physicians" produced by the Association of American Medical Colleges and the Howard Hughes Medical

Institute, 2009. This report states that

> ...biology students must be able to "apply quantitative reasoning and appropriate mathematics to describe or explain phenomena in the natural world." Additionally, "it is essential not only to read the medical and scientific literature of one's discipline, but to examine it critically to achieve lifelong learning. These activities require knowledge and skills in critical analysis, statistical inference, and experimental design."

(Schwartzstein et al., 2013) state that students taking the MCAT are required to demonstrate

1. "scientific reasoning and problem solving by reasoning with scientific principles, theories, and models and by analyzing and evaluating scientific explanations and predictions;
2. reasoning about the design and execution of research by demonstrating their understanding of important components of scientific research and by reasoning about ethical issues in research; and
3. data-based and statistical reasoning by interpreting patterns in data presented in tables, figures, and graphs and by reasoning about data and drawing conclusions from them."

We can achieve these outcomes by working together. You are more likely to achieve these if you actively ***play*** with data analysis and modeling techniques. You should work to solve all problems posed in the ICEs and your book. Those questions may appear on a "statistical challenge." Use `R` in your other classes to solve problems, too. And write up lab reports using `Sweave` from within `RStudio`. These skills will come back to reward you. The evidence of others using this platform are found in leading scientific journals, including *Science*, *Nature*, and *PNAS*. It also is an important analytics program used extensively by a variety of organizations, including the CDC, Pfizer, Merck, Google, Meta (Facebook), Mozilla, Microsoft, NY Times, and others. Also, the data skills you learn in this class will help you get into med school (see the MCAT test) or graduate school (see the GRE test) and help you succeed when there.

Finally, by the time you successfully complete this course you should be able to counter the following misguided thinking:

> **Why do we need to use statistics and mathematics in biology? If the hypothesis is clear, the experiment is designed correctly, and the data carefully collected, anyone should be able to just look at the data and clearly see whether or not the hypothesis is supported. Statistical procedures and mathematical tricks are simply safety nets and smoke screens to cover up sloppy science.**

3 Expected Learning Shortcomings (things you won't likely learn)

Biology is an extremely broad and rapidly changing, data rich discipline. Techniques used to understand and visualize biological information are being developed and implemented continuously. If you open a recent issue of the journal *Science* or *Nature* you'll also find analyses and visualizations that we won't cover in this class. However, as you develop your scientific skills you will better understand these mysteries you encounter. So, as you grow in ability, you will need to explore strange new statistics, to seek out new data and new visualizations, to boldly go where at least you have not gone before to be a part of modern biology.

4 Course Resources

1. **Dr. H.** Please consider me a member of your academic success team. It's not you against me. It's you and me against me. I can be helpful with `R`. Attend office hours.
2. **Your classmates!** Look left. Look right. These are your teammates.
3. **Olivia Lopez**. Olivia's responsibilities include helping you succeed. She'll also be looking after the homework assignments which will help you practice data analysis problem solving.
4. **Book.** The required book (Hartvigsen, 2021) was written based on my teaching of this course. It is designed to help you succeed. I've spent many hundreds of hours writing this and get, in return, hundreds of dollars (so, possibly as much as a dollar per hour). In exchange for your purchasing the book I will host a pizza party. Note that if you pirate the book, you're the pirate and the publisher, who I convinced to support me, is the primary victim.
5. **Provided Information**. I provide lecture notes, old exams, and ICEs through my OutBox. See https://geneseo.atlassian.net/wiki/spaces/HELP/pages/76778990/Inboxes+and+Outboxes. Note that on Macs the "Go menu" is in Finder. Be sure to go to my "Outbox," not my "Inbox."

What's my grade in this class?

In my OutBox you'll find an `R` script file called `What is my grade.r`. Use this to maintain your grade in the class. If you're not sure or want to verify your grade please come by any one of my office hours.

The challenges are usually emailed back to you before the next class meeting. **It is your responsibility to ensure the grades are accurate**. You have just one week after I return a challenge to appeal the grade.

Your final grade will be converted from a proportion of points earned into a letter grade using the following ranges.

Score		Letter Grade		Score
0.933	$\leq$	A		
0.900	$\leq$	A-	$<$	0.933
0.867	$\leq$	B+	$<$	0.900
0.833	$\leq$	B	$<$	0.867
0.800	$\leq$	B-	$<$	0.833
etc.				

Note that I will round your grade UP to three decimal places using `Excel`'s `ceiling()` function. The College Bulletin shares that "Grade point averages are truncated to two decimal places...." which means it's worse than rounding down. They provide an example for a student who gets their semester GPA of 2.728571 rounded down to 2.72! I bet they round down your cumulative GPA, as well! To help counter this I will take a grade of, say, 0.83210522 (a proportion), which would be a B- if *correctly* rounded to three decimal places, and round it up to 83.3, making it a B. Note, too, that the college gets you again by making a B+ 3.3 "quality points" instead of $3.\overline{3}$. You do get a GPA boost for getting a grade with a minus – they convert it to 3.7 instead of $3.\overline{6}$. So, maybe it's a wash, but you might go look at your grades. If you have more grades with a "+" then you're getting stiffed quantitatively. You also might check to see if your grades are normally distributed. Most students' grade distributions are not normally distributed, being skewed left. We know that

when the data are "skewed" the mean is an incorrect measure of central tendency. So, **statistics are being used to abuse you**. In this class you will learn how to identify such numerical abuses.

5 Assessments

Below are the items that will be used to assess your understanding of the material.

Item	Description	Number	Points
Homeworks	Homework problems occur at the end of each chapter.	5	10 each (50 pts total)
Statistical challenges	Solving problems in class using `R` script or Sweave files (these are cumulative)	4	20, 40, 50, and 60 pts each (170 pts total)
Statistical application	An optional assessment during the final time period that may, or may not, help your grade. Assess whether a data set supports a result	1	During final exam week, 10 pts
Total			230

Homeworks. These assignments are emailed to me. The file must include your full name (e.g., `Gregg-Hartvigsen-HW3.rnw`). Homeworks will be charged a value of 20% per 24-hour period the work is late. All book chapters have questions at the end - you should complete all these for all chapters.

Statistical challenges. These are assessments of your ability to solve problems like the ones you've seen up to the time of each challenge. They increase in value because you should be getting better at these and, therefore, get more rewards (points). These need to be completed during class time and submitted by the end of class (you email me your completed file). Note that 5% of the grade is deducted for each minute it is submitted late (I can see time stamps). You'll want to work efficiently and be sure everything works well before your time is up.

Missing a statistical challenge. Missing a challenge is a big deal. I am happy to work with you if you keep me informed with an email notification before the challenge and I receive from you evidence of your reason for missing class (e.g., notification from the health center).

But what if you can't make it because you're barfing that morning? Email me and tell me you're going to the health center instead. Afterwards, email me evidence that you saw them. It is your responsibility to communicate with me about this. Appropriate excuses include illness with evidence of your visit to an appropriate professional or notification through the Dean of Students Office (Phone: 585-245-5706; Email: deanstu@geneseo.edu).

Statistical application (the "final"). This will involve using data from a scientific paper and testing the result found by the authors. You likely will need to acquire the data using WebPlot Digitizer (`https://apps.automeris.io/wpd/`).

5.1 An approach to crushing assessments

Be your strongest, most consistent and honest ally. So, **study like you're taking an exam**. Each study session should be no longer than the length of the exam for which you are studying. You probably know you wouldn't do your best on an exam by constantly checking your phone, texting friends, and listening to music. Treat your study time with the same intensity/respect as when you are taking the test. Studying, therefore, should be tiring. It's just as important to take an awesome break between study sessions. Treat studying like a professional athlete who trains effectively and peaks at an event (the test). They do NOT prepare for an event by simply pulling an all-nighter. Also, get that sleep you have heard so many times as being important. And eating well is really important. Eat like you're training for a race - don't fill up on comfort junk food (donuts?) before a big exam. All I'm really suggesting is that you use what biologists have taught us. I know professional athletes are listening to the biologists - you should, too!

6 Schedule

Note: always bring your computer and your power cord.

Date	Day	Activity.Topic	Prep.for.Class	What.is.due
8/29/2023	Tue	Introduction to FoB; complete online survey.	Read syllabus. Read Primer Introduction (pg xiii - xviii). Have installed latest versions of R, Rstudio, and LaTeX. Make sure they work. BRING COMPUTER!	
8/31/2023	Thu	Introduction to Excel and R. Getting data into R	Read Chapt. 1 + 2. Complete Tuesday's "Getting Started Using R" ICE. Don't hand in but check yourself against solution set.	
9/5/2023	Tue	Answer questions about "Getting Started Using R." Getting data using Webplot Digitizer		HW: Chapter 1 problems - email me your R script file
9/7/2023	Thu	Working with data	Read Chapt. 3	Be able to do Chapt. 2 problems but don't hand in.
9/12/2023	Tue	Practice statistical challenge		HW: Chapt. 3 problems - email me your R script file
9/14/2023	Thu	Statistical challenge 1 using R script file (20 pts)	Be proficient with basic R tasks, using R script files, and emailing an attachment to me.	
9/19/2023	Tue	Introduction to Sweave/LaTeX	Have LaTeX installed (takes forever!). Make sure that Sweave works.	
9/21/2023	Thu	Tell me about my data	Read Chapt. 4	
9/26/2023	Tue	Visualizing data	Read Chapt. 5.	HW: Chapt. 4 problems - email me your .rnw file
9/28/2023	Thu	Probability		
10/3/2023	Tue	Hypotheses + Experimental design	Read Chapt. 6, sections 6.1 - 6.5	
10/5/2023	Thu	Intro to statistical inference, p-values, and errors	Read Chapt. 6, sections 6.6-6.9	

10/10/2023	Tue	October Break - No Class!		
10/12/2023	Thu	One-sample tests	Read Chapt. 7, sections 7.1-7.5	
10/17/2023	Tue	Two-sample tests	Read Chapt. 7, section 7.6	
10/19/2023	Thu	Statistical challenge 2 using Sweave (40 pts)	Organize code and practice Sweave for challenge	
10/24/2023	Tue	One-way ANOVA	Chapt. 8, sections 8.1-8.4	
10/26/2023	Thu	Visualizing ANOVA data (error bars and post-hoc tests)	Read Chapt. 8, section 8.5	
10/31/2023	Tue	Catch up and review of questions for statistical challenge 3	Review	HW: Chapt 8 problems - email me your .rnw file
11/2/2023	Thu	Two-factor analysis and visualization	Read Chapt. 8, section 8.5-8.7	
11/7/2023	Tue	Correlation	Read Chapt. 9, section 9.1	
11/9/2023	Thu	Statistical challenge 3 using Sweave (60 pts)	Organize code and practice Sweave for challenge	
11/14/2023	Tue	Linear regression	Read Chapt. 9, section 9.2-9.3	
11/16/2023	Thu	Categorical data	Read Chapt. 10	HW: Chapt 9 problems - email me your .rnw file
11/21/2023	Tue	Writing your own functions	Chapt. 11, section 11.1	
11/23/2023	Thu	Thanksgiving - No Class!	Thank someone for something	
11/28/2023	Tue	Non-linear regression	Chapt. 11, section 11.5	
11/30/2023	Thu	A little programming and the central limit theorem	Read Chapt. 12	
12/5/2023	Tue	Catch up and review of questions for statistical challenge 4		
12/7/2023	Thu	Statistical challenge 4 using Sweave (80 pts)	Organize code and practice Sweave for challenge	
12/18/2023	Mon	Statistical application (20 pts) 12:00 noon!	A challenge dealing with real data	

7 Electronic distraction devices, drugs, and other disabilities

https://www.quora.com/What-is-the-worst-thing-about-the-city-of-Tokyo-Japan

In my class we both agree not to text, chat, "do" Facebook (with grandma?), Instagram, or TikTok, recreationally watch YouTube videos, message people (or other animals), or do similar electronic gaming or distracting activities during class (laptops can be used for taking notes but please don't violate the expectations above). Why? These activities are distracting to those around you and me! I think everyone deserves respect in the classroom.

We also agree not to consume alcohol or other recreational drugs during class or come to class impaired by such activities. If either of us finds scheduling these activities (e.g., texting or doing drugs) around class time difficult then we should seek professional help (e.g., through the Lauderdale Center for Student Health & Counseling).

Additionally, those of us who teach at SUNY Geneseo always will do our best to make reasonable accommodations for documented physical, emotional, or cognitive challenges. In addition, I will do my best to accommodate challenges brought about through pregnancy, parenting, or care giving. Students should contact the Office of Accessibility Services (585-245-5112) and me to discuss needed accommodations as early as possible in the semester. Note that I happily will help you to take a challenge in the Test Center (`https://www.geneseo.edu/is/testcenter/main`) during the regularly scheduled exam times.

Can I listen to music on headphones during a challenge? Yes, but it's not a good idea. Never will you concentrate better with distractions like music playing.

8 Honesty

ALL WORK MUST BE YOUR OWN. Do not plagiarize from others, including classmates, previous classmates, and external sources. All of the code you provide in the challenges must be either your own or code provided by me. That means, if asked, you must be able to tell me what all your code does. Do not try to find code online (e.g., ChatGPT) that you think solves your problem – this will be considered plagiarism. The assessments are designed so that you can complete them yourself! Please see the College's policy on academic honesty.

9 Minimum Competence in Biology

To graduate with a Biology major, students must attain a grade of C- or better in all required Biology courses (excluding Biology electives). A grade of C- must be achieved in any course before it may be used as a prerequisite for another course. A student may only repeat a required Biology course or related requirement once for major credit and the course must be taken at the next offering of the class (provided there are seats available). If a student does not earn at least a C- on the second taking of the class, she/he will not be able to complete the Biology major.

10 Zoom Office Hours

Most of my office hrs are over Zoom (`https://geneseo.zoom.us/j/4333683209`, passcode = 497118). I only have 40 minutes of zoom time so it will end and then I will restart Zoom. If you get cut off, just come back. We may switch to Teams but I find that platform unpleasant. You might not get in right away because I'm with someone who wants a private conversation and so I'll let you in as soon as I can. If you can't wait send me an email with your question. In preparation for a zoom meeting you will be able to share your screen so you should get your questions ready before entering the zoom call.

11 Religious observances

It is my responsibility, as outlined in the College's Undergraduate Bulletin, to accommodate religious observances. No exams have been scheduled to occur on notable observance days. However, as stated in the 2016-2017 Bulletin, I am "to comply in good faith with the provisions of... " section 224-a of the Education Law of New York State. I am happy to meet your needs if you inform me of any such absence at least one week prior to the conflict. Without you providing me this information I may not be able to meet your learning expectations for the class.

References

Hartvigsen, G. 2021. A primer in biological data analysis and visualization using R, 2e. Columbia University Press.

Schwartzstein, R. M., G. C. Rosenfeld, R. Hilborn, S. H. Oyewole, and K. Mitchell. 2013. Redesigning the MCAT exam: balancing multiple perspectives. Academic Medicine, **88**.