Syllabus* for Foundations of Biostatistics (BIOL 350)

Fall 2021

Gregg Hartvigsen

Version date: August 29, 2021

* Assumes that we're mask-to-mask (as opposed to face-to-face). In the event of a change the syllabus will be updated to reflect the new learning environment.

"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 2:30 - 3:45 or 4:00 - 5:15, ISC 115
Contact me	ISC 360, 245.5448, hartvig@geneseo.edu
Office hrs†	Tues. & Th 8:00 - 9:00am (**Zoom** with your Geneseo account; enter waiting room) Tues. 11:30 - 1:30 (ISC 343) W 2:00 - 4:00 (ISC 343)
Required textbook	Hartvigsen (2021, 2nd edition!). I recommend you bring this to class. If you have the old one it will have different questions that won't earn you points. I don't have the new questions + answers available.
Required laptop (Mac, Windows, or Linux)	Bring to all mtgs. Note that all seats have a plug.
Required software (free)	`Excel` (or similar), `R` (version 4.1.0 or later), `RStudio` (version 1.4.1 or later), and LaTeX(MacTex for Macs and TexLive for Windows and Linux).
Mask, worn properly	Please comply - we all need to feel safe in the classroom.
Refreshments	Unfortunately, no refreshments can be consumed in class since masks are mandatory
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 Backup and Sync" (recommended), Dropbox, OneDrive, or iCloud. This protects you from have to say "I lost my work because my computer died."
Calculator	Absolutely forbidden.

1 Course Resources

1. **Me!** 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!**
3. **Book**: The required book (Hartvigsen, 2021) was written based on my teaching of this course. It is a bit different from the 1st edition, most notably with the chapter-ending problems. The book was written using `RStudio` and LaTeX, skills you will have by the end of the course!
4. **Post-lecture notes**. After lectures I'll post the notes on Canvas.
5. **ICEs**. These are in-class exercises that will have background information for a specific topic, instructions, sample code, and problems to complete, generally in class. These are integral parts of the course and your knowledge of these will be rewarded on `R` Challenges.

6. **Old assessments**. Some old challenges are available in my Geneseo "Outbox" in the `BDA` folder. For help on boxes see `https://wiki.geneseo.edu/display/cit/Inboxes+and+Outboxes`.
7. **Additional resources for help using R**.
   (a) Duckduckgo (or Google) search questions. This is great for solutions and for solving errors you might make.
   (b) A quick entry to the breadth of the `R` software packages can be found through the Cran Task Views.

2 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 *some* types 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 (or keep) is the attitude that you can fearlessly figure out how to solve new 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`. The most common suggest on my SOFIs is that I should lecture more. I like lecturing but I think you'll learn more by doing.

3 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. work with data. This will involve finding, entering, saving, organizing, and manipulating data. To accomplish this you will learn how to use both `Excel` and `R`;

2. identify correctly basic experimental design principles that are used to answer a variety of biological questions;
3. correctly develop and test hypotheses statistically using data;
4. understand 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 either la-tech or lay-tech);
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. conduct an original project that requires gathering, analyzing, and graphing data using `R` that assesses a comprehensive biological problem and culminates in an oral, poster-based presentation;
10. write and use your own functions;
11. 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 an exam. 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 Pfizer, Merck, Google, Facebook, Mozilla, Microsoft, NY Times, Shell, 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.**

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 Grading

Below I provide information on assessment tools used in this class. These include homework assignments, `R` Challenges during part of a class, a mid-term and final exam, and a final project. I may not post grades on Canvas - it is your responsibility to keep track of your grades and weights to know what your grade is during the class (how to calculate weighted means is discussed in your book!). Your final grade proportion will be converted 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...." They provide an example for a student who gets their 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 abuse.

Assessments

Item	Description	Number	Points
R Challenges	Solving problems using `R` script or Sweave files	5	seq(10,50,by=10) pts ea (150 pts total)
Homeworks	Problems with strict due dates/times.	5	10 ea
Project proposal	One page, paper copy	1	10
Speed Presentation	2 minutes, poster slide with at least one graph	1	10
Project presentation	Presentation and poster	1	80
Total			300

R challenges. These are assessments of your ability to solve problems like the ones you've seen up to this point. There are five. Each one is weighted more than the previous because I assume you're getting better at this and so you get to get more points toward your final grade this way. They need to be completed during class time and submitted by the end of class. 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 time is up.

Homeworks. These are completed outside of class but note strict deadlines (due by the beginning of class). These will be either a an `R` script or Sweave file, uploaded to Canvas. Check to see if you're building an `R` script file or Sweave'd doc. The plan is to have five but there may be fewer, depending on how crazy time gets.

Project proposal, speed presentation, final project. See below!

5 Semester project

This is a solo project that will give you, if done well, an awesome talking point during a job, grad school, or med school interview. You will present a projected poster (single slide) of a deep analysis of an original hypothesis, or hypotheses, using real data (e.g., from journals, your own original research (no part can have been previously presented), or using data found on the internet; e.g., CDC at `https://www.cdc.gov/`). **All the files for this project need to be archived into a single Zip file and emailed to me by 11:59pm on the last day of classes**. You must provide the data *and* the `R` script file(s) used to complete *all* the analyses from your presentation and produces your graphs so that I can simply run the script file. Each 24-hour period after the due date/time that the project is handed in late will accumulate a 10% grade deduction for your project. If the `R` script file doesn't work you lose 20% of the points for the project. You have 5-minutes only for the presentation. You will be describing the important parts of your poster and likely answering questions. You will not want to talk slowly and use filler/fluff to buy time and you don't want to get cut off at 5 minutes, unable to provide a conclusion, for instance. I will particularly seek to see you speak extemporaneously about what you did and found.

There are three assessments to this semester-long project:

1. **The Proposal**. This is a one page proposal that needs to convince me that you've done your homework to find a project you can complete. You need to convince me that you have a clear question, or set of questions that are biologically oriented and that you have found data that you can use. This should include citations and/or links to data. This proposal requires approval. If your project is not approved you get to keep trying. The

longer it takes to get approval the harder the project will be. It must be approved no more than one week from the time I hand it back to redo it and submit it for approval. Your grade will not change on this assignment but getting steered onto a new, better path will pay dividends on the presentation. If it not approved you don't get the points.

2. **Speed Presentation**. This is a 2 minute presentation that shows everyone the state of your Poster. It should have a title, your name, at least one decent graph that includes a related statistical analysis, a result in the "Results" section, and at least one citation for where data came from.
3. **Poster Presentation**. The presentations will be completed during a randomly determined time slot during the final exam period. You are welcome attend any other presentations but this is not required. You can arrive well before your time slot but the plan is you begin at your specified time. Coming late and still being able to present if there's an opening will result in a 25% loss in points. Missing your presentation entirely will result in a 50% loss in points.

For this project you will need data. Be warned: good science is hypothesis driven. You always should

1. develop a question;
2. phrase the question as a testable hypothesis;
3. get the appropriate data to answer the question;
4. answer the question.
5. share the result with clean data analysis and beautiful, revealing visualizations.

You should not look for data and then build a project around it. No data can come from an `R` package! This paper raised an important issue about whether people who analyze other people's data are just scientific parasites. An interesting rebuttal to that paper can be found here: **"The one true route to good science is..."**. Do check these out before spending a lot of time on your project. Be sure to chat with me to avoid going down a poorly chosen rabbit hole.

Here are some rules to keep in mind when thinking you're done with this project. Ignore these at your own risk.

1. **Submit one zip file before 5pm on the last day of classes**. The single zip file should be named Lastname-Firstname.zip. The poster file, inside this Zip archive, should be named "Lastname-Firstname.pdf". That's the file I'll put up for you to discuss.
2. **Data analyses must work**. You can't get a passing grade on this project if your code does not complete the analyses and make the visualizations provided in your poster.
3. **Avoid relying on an `R` package not discussed in class.** If you do this you should include installing and loading the package like this in your `R` script file:

```
if (!require(package.name)) {
  install.packages("package.name")
  library(package.name)
}
```

4. **Do not set your working directory in script files.** This might work great on your computer but on mine your code will not work!
5. **Poster structure**. There are many good examples of how to construct an effective scientific poster. A good project presentation will flow much like a good scientific paper:
   (a) **Descriptive title**. This shouldn't be a tease - it should be $\leq$ 10 words. Include your name underneath. I am not a co-author.

(b) **Abstract**. A 100 – 150 word summary. About one sentence for each section: Introduction, Methods, Results, Discussion.
(c) **Introduction**. Introduce the problem and why your audience should care. Citations expected.
(d) **Methods**. Describe how the data were originally collected and what you did to analyze the data. Results from normality tests, if necessary, would go here because it justifies why you did a particular test. If showing how the data are distributed is a result then it belongs in the results section.
(e) **Results**. Share what you found. There should be graphs and results from statistical tests (e.g., the ANOVA table).
(f) **Discussion**. What do your results mean? Pull together the significance of your results.
(g) **Acknowledgments**.
(h) **Citations**.

Good posters generally have few words and many self-explanatory visualizations. Below is the approximate rubric for your presentation evaluation. The actual rubric may differ slightly from this.

The poster:

(a) was professionally structured. Not too much text. Majority of the space was occupied by visualizations. The locations and lengths of the title, abstract, intro, methods, results, discussion, citations, and acknowledgements were professional;
(b) included publication-quality visualizations with appropriate captions;
(c) included results that were concise and referenced all visualizations;
(d) included appropriate and correctly executed statistical tests and results to tell a logical story. Appropriate statistical information was included;
(e) asked and answered appropriate questions for this class;
(f) relied on extensive/appropriate data to answer the questions asked;
(g) included correct/appropriate citations and acknowledgements.

The presentation of the poster:

(a) was well organized to lead listeners through a well structured story told by the data, visualizations, and analyses;
(b) was described verbally at an appropriate depth in the time permitted (5 minutes). You didn't waste an opportunity (it was too short) and you didn't get cut off and not finish because you planned more than 5 minutes;
(c) included, if appropriate, precise and concise answers to questions.

Note: You do NOT print the poster out. It is projected using my computer.

Project Ideas

Wow. So many possibilities! Projects have ranged from creating a simple stochastic model of logistic growth and analyzing the data to very large-scale, comprehensive analyses of different disease incidence rates across different countries.

One thing to keep in mind is that, at this point in time, `R` has become ubiquitous in scientific data analysis, particularly in biology. It is possible to find all sorts of data sets with associated `R` script files that will crunch the data in a variety of ways and produce amazing visualizations. You, however, will be evaluated on your creativity, difficulty, and the work you do. This is like an Olympic diving competition - it's the product of difficulty and execution. If you lift a project (difficulty = 0) then your learning and grade will suffer.

Keep in mind that other faculty members in the department can be great resources for understanding the system you're studying. My research is on understanding the dynamics of species (e.g., cooperation, competition) and their interactions (e.g., diseases/hosts, food webs). If you choose a project that is way over your head (try to use this project to finally learn about CRISPR) your grade might suffer. Given all this, here are few ideas:

1. You could work to analyze data from a professor, assuming the data have not yet been analyzed (no redoing projects).
2. If you're in a lab doing independent research you can "double-dip" to advance your project. You may NOT use someone else's analysis for this project. But if this is something you're working on and hope to present at GREAT Day, for instance, this would be great.
3. Check out data online, of course, such as Dryad Data Outlet, CDC data, State Cancer Profiles, Ecological Data Wiki, Nature Serve, Ecological Society of America's data archive site, and Blast genomic data.
4. Where not to get data: most `R` packages come with data sets. **These are not appropriate for this project** since code is provided to analyze these data.

6 Schedule

Note: bring your computer unless otherwise instructed.

Date	Day	Activity.Topic	Prep.for.Class
8/31/2021	Tue	Introduction to class + data analysis	Read syllabus. Read Primer Introduction (pg xiii - xviii). Have installed latest versions of R and Rstudio. Make sure they work. BRING COMPUTER!
9/2/2021	Thu	Introduction to Excel and R. Getting data into R	Read Chapt. 1 + 2. Complete "Getting Started Using R" ICE. Don't hand in but check yourself against solution set.
9/7/2021	Tue	Labor Day - No Class	
9/9/2021	Thu	Go over homework. Extending getting data into R - Webplot Digitizer	HW: R script file that has code that generates answers to all problems from Chapt 1 + 2. (name it: Lastname-Firstname1.r)
9/14/2021	Tue	R challenge 1 using R script file (10 pts) Crash course in statistics for the project	Be proficient with basic R tasks, using R script file, emailing an attachment to someone (me!).
9/16/2021	Thu	Working with data	Read Chapt. 3
9/21/2021	Tue	Introduction to Sweave/LaTeX	HW: Chapt. 3 problems (R script file named Lastname-Firstname3.r). Have LaTeX installed (takes forever!). Make sure that Sweave works.

9/23/2021	Thu	Tell me about my data	Read Chapt. 4. HW: Chapt. 4 problems. Hand in paper copy of pdf made using Sweave. Begin with template!
9/28/2021	Tue	Visualizing data	Project Proposal. Read Chapt. 5.
9/30/2021	Thu	R challenge 2 using Sweave (20 pts)	Organize code and practice Sweave for challenge
10/5/2021	Tue	Probability	
10/7/2021	Thu	Hypotheses + Experimental design	Read Chapt. 6, sections 6.1-6.5
10/12/2021	Tue	October Break - No Class!	
10/14/2021	Thu	Intro to statistical inference, p-values, and errors	Read Chapt. 6, sections 6.6-6.9
10/19/2021	Tue	One-sample tests	Read Chapt. 7, sections 7.1-7.3
10/21/2021	Thu	Two-sample tests	HW: chapter 5 + 6 problems, Sweaved. Read chapt. 7, sections 7.4-7.6
10/26/2021	Tue	Catch up and review of questions for R challenge #3	
10/28/2021	Thu	R challenge 3 using Sweave (30 pts)	Organize code and practice Sweave for challenge
11/2/2021	Tue	Speed Presentation: Project update - 2 min presentation using my computer. Explain one key result.	Send PPT or PDF by Sun, 10-31 @ 5:00pm. See syllabus. -5 pts per 24hrs late
11/4/2021	Thu	One-way ANOVA	Read Chapt. 8, sections 8.1-8.4
11/9/2021	Tue	Visualizing ANOVA data (error bars and post-hoc tests)	Read Chapt. 8, section 8.5
11/11/2021	Thu	Two-factor analysis and visualization	Read Chapt. 8, section 8.5-8.7
11/16/2021	Tue	Correlation	HW: Chapt 8 problems. Read Chapt. 9, section 9.1
11/18/2021	Thu	R Challenge 4 using Sweave (40 pts)	Organize code and practice Sweave for challenge
11/23/2021	Tue	Linear regression	Read Chapt. 9, section 9.2-9.3
11/25/2021	Thu	Categorical data + writing your own functions	Read Chapt. 10 + Chapt. 11, section 11.1
11/30/2021	Tue	Non-linear regression	Chapt. 11, section 11.5
12/2/2021	Thu	Thanksgiving - No Class!	Thank someone for something
12/7/2021	Tue	A little programming and the central limit theorem	Read Chapt. 12
12/9/2021	Thu	R challenge 5 using Sweave (50 pts)	Organize code and practice Sweave for challenge
12/12/2021	Sun	Submit all project files by 5pm	See syllabus for delivery instructions

12/16/2021	Thu	Final Presentations 2:30 section at 12:00-2:30, ISC 115. 4:00 section at 3:30-6:00, ISC 115.	Order will be random and provided a day before. You may attend any you like (but don't miss yours!).

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 classes we both agree not to text, chat, "do" Facebook or Instagram, recreationally watch YouTube videos, message, 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. For instance, you might want to ask a question during class so I should be paying attention to you so I can respond.

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 will do our best to make reasonable accommodations for students with documented physical, emotional, or cognitive disabilities. In addition, we will do our 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 exams in the Test Center (`https://www.geneseo.edu/is/testcenter/main`) during the regularly scheduled exam times.

8 Dishonesty

ALL WORK MUST BE YOUR OWN. Do not plagiarize from others, including classmates, previous classmates, and external sources. All code you provide that I didn't write must be your own. That means, if asked, you could tell me what it all does. It's okay to use external literature sources but cite them completely. Do not try to find code online that you think solves your problem but you have no idea what it does. The assignments and assessments are designed so that you can complete them yourself! Please see the College's policy on academic dishonesty.

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 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`. Columbia University Press.

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