

Biological Data Analysis (BIOL 250)

Fall 2015

Created: August 30, 2015, 1:45pm

“Knowledge is power.” Francis Bacon (1561-1626)

“To measure is to know.” Lord Kelvin (1824-1907)

“Heavier-than-air flying machines are impossible.” Lord Kelvin (1824-1907)

“Knowledge is good.” Emil Faber. Founder of Faber College (see documentary: [Animal House](#))

“Knowledge comes from collecting, analyzing, visualizing, and interpreting data correctly.” Gregg Hartvigsen

Item	Course details
Meetings	MF 2:30 - 3:45, ISC 115
Gregg’s contact info	Office = ISC 360 Office phone = 245.5448 Email = hartvig@geneseo.edu
Gregg’s office hrs	Mon. 3:50 - 5:30, Wed. 3:45 - 5:00, Fri. 3:50 - 4:50
TA Tom Hartvigsen	th11@geneseo.edu Office hours: W 2:30 - 4:00, Th 10-12, ISC 343
TA Savannah Russ	smr27@geneseo.edu Office hours: TBA, ISC 343
Required textbook	Hartvigsen (2014)
Your laptop	Bring this to all meetings, unless otherwise instructed. Also, bring power cord.
Required software	R, RStudio, and L ^A T _E X(all free)
Recommended software	Excel or Excel-like program

Note that this course fulfills the computer tools requirement for biology BS and biochemistry majors.

Course Resources

1. **Book:** The required book (Hartvigsen, 2014) was written based on my teaching of this course. It is a brief introduction to biological data analysis and visualization.
2. **Post-lecture notes.** After lectures I’ll post the notes on MyCourses.
3. **ICEs.** These are in-class exercises, provided on MyCourses, that will have background information for a specific topic, instructions, sample code, and problems to complete. These are integral parts of the course and your knowledge of these will be rewarded on exams.
4. **Help on R.**
 - (a) Your book (e.g., see section 1.3, “Getting help with R”, in your book.
 - (b) Go to the website <http://rseek.org/> and enter your search terms.
 - (c) A quick entry to the breadth of the R software packages can be found through the [Cran Task Views](#).
5. **SOFI comments** from last time I taught the course. I hope these can be helpful to you, too!
6. **Let’s do lunch!** I’d like to lunch with you and chat about anything - the course, life, jobs, grad school, juggling, hitchhiking, motorcycling, music, tennis, racquetball, squash, nature, cycling, the next exam, whatever. Please gather a group of two or more. The venue is your choice. We each pay for ourselves. Note: I’m a vegetarian.

Overview

This is a course on empowerment and discovery in biology. It's designed to be an introduction to the management, visualization, and analysis 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 biological information 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 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 because all new analytical procedures are being developed in this environment (not Excel, not Minitab, not SPSS, not SAS).

Here's some bad news: **at some point you will wish the TAs and I were more helpful.** We will encourage you to figure R out (but will always answer your questions). So, if you can figure it out yourself you'll:

1. get the answer;
2. have figured out *how* to get the answer;
3. hopefully have shared your knowledge with others.

You, no doubt, forget how the data are supposed to be entered for a particular test or how to do something cool when making a graph. Saving your work in script files will be quite helpful, as well as naming those files with meaningful names (do not name your work on a t-test "BDAlab17.r"). What I hope for you is that you'll soon have the attitude that you can figure anything out.

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. We then apply this information to solve an in-class exercise (ICE) using R.

Expected Learning Outcomes

If you successfully complete this class you will be able to:

1. correctly develop and test hypotheses statistically using data;
2. correctly enter and manage data using Excel;
3. get data into the R statistical and computing environment through directly entering the data, reading them in from a spreadsheet-like format, and by downloading them from a website;
4. understand the differences between parametric and non-parametric statistical tests;
5. create appropriate publication-quality visualizations of data;
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 ;
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. work in a team on an original, semester-long project that requires gathering, analyzing, and graphing data using R that assesses a comprehensive biological problem and culminates in an oral presentation;
10. write computer programs in an object-oriented programming language, including “for” and “while” loops, “if” conditional statements;
11. write and use your own functions;
12. find, save, organize, manipulate, and read in data files on your computer.

These outcomes are aligned with the 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.”

and:

“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.”

We can achieve these outcomes if we work 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 R. 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, if you’re interested in going to med school you will find that analytical skills ([click here](#)) are an important part of the MCAT test. It’s integral to the environmental sciences because of the amount of data in those fields.

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.

Grading

Below are the basic assessment tools for this class. Most of the points will be earned through two exams and the semester-long research project. There is some flexibility in how many points will

be available due to the unknown number of pop quizzes and homework assignments. The number of these is inversely related to a vague quantification of in-class participation and engagement. Feel free to encourage me to provide these opportunities. Your grade will be determined by dividing your points earned by the number of points that were offered.

Item	Description	Points
Pop quizzes	Unannounced handwritten challenges for concepts and R coding. Number unknown	~ 5 pts ea
ICE	In class exercises <i>may</i> be graded	~ 10 pts ea
Homework	Number unknown	~ 10 pts ea.
Mid-term exam	Written and computer parts	100 pts (total)
Final exam (written + computer)	Cumulative	150 pts
Semester project	Includes progress reports	100 pts

Your final grade will be converted from a numerical value to a letter grade using the following rules. I will round your score up[♥] to three decimal places using Excel's function `CEILING(number, 0.001)`. The values below are proportions of possible points earned.

Score	Letter Grade	Score
0.933	A	∞^*
0.900	A-	0.933
0.867	B+	0.900
0.833	B	0.867
0.800	B-	0.833
etc.		

* Extra credit points may lead to scores > 1.0 .

♥ The college rounds down (truncates) the average “quality points” you earn for each semester before calculating the GPA for each semester. In the 2014-2015 Bulletin's example (pages 38-39) they show a student earning a GPA of 2.728571, which then becomes 2.72! They probably round down your GPA for your cumulative GPA, as well! To help counter this I will take a grade of, say, 0.832105, which would be a B-, and round it up to 0.833, making it a B.

Pop Quizzes. These are unplanned opportunities to reward you for preparing for class. You are expected to be able to complete basic, rudimentary tasks performed in recent classes and understand the reading for the current week. These may include function calls from R and concepts from lecture and lab. They are to be answered using pen and paper. A sample question might be “Provide all the necessary R code that would read in a data file and create a professional-quality scatterplot of the data.”

Exams. All three exams are cumulative. They will have two parts. The first will be a written part that takes about a third of the time. Once you hand that in you can proceed to open your computer and do the computer part. The written exams are closed book. The computer exams are open (book, notes, computer, and internet). For all exams you are not allowed to communicate with anyone (chatting, talking, texting, or emailing, unless otherwise instructed). Note that you can adjust the weight on the written and computer parts. This is done by making the second part (computer part) worth between 40 - 60%. You choose this weight on the computer part of the exam *before* submitting it (you email the computer part to me when done).

Semester project

This is a project you should feel comfortable showing someone when you want to get a job, do research with someone, or show someone who might consider funding you as a grad student or accept you into medical school. Working in groups of four you will present a deep analysis of an original hypothesis, or hypotheses using real data (e.g., from journals, your own research, or using data found the internet; e.g., CDC). All the files for this project need to be emailed to me before the closing date/time. You must provide the data *and* the R script files 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 accumulates a 10% grade deduction in the project. If the R script file doesn't work expect to lose 50% of the points for the project. The presentation should be 12-minutes long and then allow three minutes for questions.

Group presentations will be completed during the final exam time slot (see schedule below). Attendance throughout all presentations is required because your grade will be contingent on your evaluation of the presentations.

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 build a project around whatever data you can find. [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.

Here are some rules to keep in mind for this project.

1. **Name the project files.** Name the presentation file with your last names (e.g., "Dewey, Cheatem, and Howe.pptx"). Do not name the file "presentation1.pptx." Data files should have the same names but with different "extensions." (I will take the liberty of subtracting 5% for confusing naming of the files.) Name your R and Excel files similarly.
2. **Talk structure.** A good project presentation will flow much like a good scientific paper:
 - (a) Title slide. The first slide should show your project's title and your names.
 - (b) Possibly include an outline slide, which reappears at the beginning of each new section.
 - (c) An Introduction. Introduce the problem and why your audience should care.
 - (d) Methods. Describe how the data were collected and what you did to analyze the data. A normality test, 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 the statistics from statistical tests (e.g., the ANOVA table).
 - (f) Discussion. What do your results mean? Pull together the significance of your results.
 - (g) Acknowledgments and citations.

Check out this article on the [Ten Simple Rules for Making Good Oral Presentations](#).

3. **Project dates** for presentation components, due at the beginning of class on these days:

Date	What's Due
9-11	Topic idea (a few sentences)
9-21	One paragraph proposal. Paper copy only. Everyone's printed name and signature must be on it. Include a title.
10-30	Group project mini-presentation (~ 5 minutes). This is a shortened version of your final presentation. Be sure to demonstrate that you have data, have made at least one thorough analysis of your data, and have graphs showing clear results. Provide title slide, intro, three visualization slides showing data, at least one formal hypothesis test, and a brief discussion of what these results mean. The slides you present must be dropped in my InBox before class starts. Name your file with your last names in alphabetical order.
12-14, 11:59 pm	Email me your presentation files: pptx, ppt, or pdf only, data files opened with your single, complete R script file. The filenames must include all your last names (see above). I'm flexible about the due date! You can hand it in any time before this deadline. You can deliver it after this time but will be charged a late fee = 10% for delivery within each 24 hr period following the due date/time. The presentation file you hand in will be on my computer waiting for you on presentation day. It is your job to know the fonts and formatting are fine on a Windows machine.
12-22, 3:30 - 6:30pm	Final presentations. Miss this and flunk the class.

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 a combination of difficulty and execution. If you lift a project (e.g., see [R Climate Graphs](#)) then your learning and grade will suffer.

Keep in mind that the TAs and I can *usually* help you. Sometime, however, I can't be of much help. I'm an ecologist but even that is far too broad for any one person to be an expert. If you work with a system that I understand I can help you more. But if you find ecology revolting that's OK! We have other resources for you to tap to get into the project that really interests you, and that's what I think will work best for you. Keep in mind that if you choose a project that is way over your head, and it remains that way until the end, your grade will suffer! Given all this, here are few ideas:

1. Talk to a professor and, possibly, develop a research project with them that involves helping them analyze some of their data.
2. You or your friends are working on a project and want help or double-dip so you can better understand the data. Please don't have a research buddy just hand you the finished

solution that you present. But if this is something you're working on and hope to present at GREAT Day, for instance, you're in business! Be sure to check with me....

3. Check out the the [Ecological Data Wiki](#) or [Nature Serve](#), the [Ecological Society of America's data archive site](#), and [Blast genomic data](#).
4. Most add-on packages in R come with data sets (e.g., see [R data sets](#)). Type `> data()` at the console and a file will be opened. These are often pretty small and are not suitable for a semester project.

Note that proceeding without talking to me is not recommended. There always seems to be a group that goes off in a direction with the best of intentions but without realizing they have no oars, no rudder, no life vests, no nothing. Don't be that group! Use the resources available to you. Consider me and the TA(s) members of your group.

Building Groups

This is *always* difficult and, in this class, important. We're going to spend a day working on this with an effort to do it right. I recognize, however, that group work always is challenging.

Schedule

Bring your computer unless otherwise instructed. Note that the in-class exercises (ICE) should be available on MyCourses the night before class. Some students like to take a look at them (and/or complete them!) before they come to class. This is great!

Date	Day	Activity.Topic	Prep.for.Class
8/31/2015	Mon	Introduction to class and R. Complete survey	Chapt. 0 - the book's introduction
9/4/2015	Fri	Introduction to Excel and R. Getting data into R	Chapt. 1 + 2. Have installed R and Rstudio.
9/7/2015	Mon	Labor Day - No Class!	
9/11/2015	Fri	Form groups, hand in proposed topic	HW: Even probs, Chapt 1 + 2. Consider what makes a good group
9/14/2015	Mon	Working with your data	Chapt. 3.
9/18/2015	Fri	Tell me about the data.	Chapt. 4.
9/21/2015	Mon	Visualizing your data. Google influenza data.	Chapt. 5 - Group proposal.
9/25/2015	Fri	Data challenge 1 (previous work). Begin probability	Organize code for data challenge
9/28/2015	Mon	Introduction to Sweave	Install Latex on your system before class and test. Give yourself an hour. See Seave-startup.pdf.
10/2/2015	Fri	Calculations and visualizations in Sweave	Modify sweave lab report
10/5/2015	Mon	The Prudential ad	HW (sweave ICE write-up)
10/9/2015	Fri	Quiz (done using sweave) + the meaning of statistics	Make sure you can Sweave! Chapt. 6, sections 6.1-6.3
10/12/2015	Mon	October Break - No Class!	
10/16/2015	Fri	Experimental design	Chapt. 6, sections 6.4-6.5

10/19/2015	Mon	Intro to statistical inference, p-values, and errors	Chapt. 6, sections 6.6-6.8
10/23/2015	Fri	Mid-term exam	Written (15 minutes) and computer (1 hr).
10/26/2015	Mon	One-sample tests	Chapt. 7, sections 7.1-7.2
10/30/2015	Fri	Group project update - presentation (~ 5 minutes using my computer)	Send by Thurs. @ 11:59pm: powerpoint file with title slide (names!), intro slide, three visualization slides showing data, and a brief discussion of meaning.
11/2/2015	Mon	Two-sample tests	Chapt. 7, section 7.3
11/6/2015	Fri	More than two samples	Chapt. 8, sections 8.1-8.2
11/9/2015	Mon	Visualizing sample data, error bars, and post-hoc tests	Chapt. 11, sections 11.2-11.3
11/13/2015	Fri	Two-factor analysis	Chapt. 8, section 8.3
11/16/2015	Mon	Sweave-based quiz, statistical analyses, visualizations, and interpretation	
11/20/2015	Fri	Correlation	Chapt. 9, section 9.1
11/23/2015	Mon	Linear regression	Chapt. 9, section 9.2
11/27/2015	Fri	Thanks Giving - No Class!	
11/30/2015	Mon	Categorical data + writing your own functions	Chapt. 10 + Chapt. 11, section 11.1
12/4/2015	Fri	Non-linear regression	Chapt. 11, section 11.5
12/7/2015	Mon	A little programming and the central limit theorem	Chapt. 12
12/11/2015	Fri	Final - Written	
12/14/2015	Mon	Final - Computer	
12/22/2015	Tue	Final Presentations - 3:30 - 6:30pm	

Electronic distraction devices, drugs, and other disabilities

Taking this course means you agree not to text, chat, “do Facebook,” or do similar electronic gaming or distracting activities during class. Also, you agree not to consume alcohol or other recreational drugs during class or come to class impaired by such activities. If you find scheduling any of these activities around class time difficult then you should seek professional help (e.g., through the [Lauderdale Center for Student Health & Counseling](#)).

Additionally, SUNY Geneseo and I will work to make reasonable accommodations for persons with documented physical, emotional, or cognitive disabilities. I also will work to accommodate medical conditions related to pregnancy or parenting. Students should contact Assistant Dean Tabitha Buggie-Hunt in the [Office of Disability Services](#) (tbuggieh@geneseo.edu or 585-245-5112) and me to discuss needed accommodations as early as possible in the semester and no less than one week prior to any and all assessment experiences.

References

Hartvigsen, G., 2014. A primer in biological data analysis and visualization using R. Columbia University Press.