

# Introduction to Biostatistics (BIOL 288)

(a.k.a. Biological Data Analysis, BDA, and BIOL 250)  
Fall 2014

“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
Contact Info	Dr. Gregg Hartvigsen, ISC 360, hartvig@geneseo.edu, 245.5448
GH Office Hrs	M 5:00 - 6:00, W 4:00 - 5:30, F 4:00 - 4:45, and by appointment!
TA Marina Massaro	mlm26@geneseo.edu, Office hours: , ISC 343
TA Savannah Russ	smr27@geneseo.edu, Office hours: , ISC 343
Required textbook	Hampton and Havel (2013)
Your laptop	Bring this to all meetings, unless otherwise instructed. Also, bring power cord.
Required software	R, RStudio, and Excel or Excel-like program

This course fulfills the computer tools requirement for biology BS and biochemistry majors.

## Course Resources

1. **Post-lecture notes.** After lectures I'll post the notes on MyCourses.
2. **ICE.** These are in-class exercises, also 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.
3. **Presentations.** Here's an article on the [Ten Simple Rules for Making Good Oral Presentations](#).
4. **Help on R.**
  - (a) A quick entry to the breadth of the R software packages can be found through the [Cran Task Views](#).
  - (b) Use a simple Google search for how to do stuff in R. You just do a search and begin with the letter "r."
  - (c) From within R you can use the ? mark to search keywords at the command prompt. Use ?? to begin a search which will pull up help that contains that word.
  - (d) **Optional book:** I wrote a biostats book (Hartvigsen, 2014) that was developed primarily through my teaching of this course. I didn't make it required for you because it basically shows you how to do most everything and I'd rather we work out the solutions dynamically in class. It should help if you want to pick it up. Consider it, or one of its many competitors, if you find R intimidating.
5. **Let's do lunch!** I'd like to lunch with you and chat about anything - the course, life, jobs, 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 will change your view of the universe. If you seek to understand and/or solve a problem based in science 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. Learning even standard procedures in R will allow you to adapt and expand your ability to understand modern data.

In this class you also will be working to make graphs. Admittedly, many of your graphs (e.g., for your project) could be made with Excel, or even drawn carefully by hand on graph paper but, at this level, you are required to make professional-quality visualizations. This is not done using hand-drawn graphs or Excel.

Here's some bad news: **at some point you will wish the TAs and I were more helpful.** In this course you are encouraged and rewarded for independently developing solutions to problems. So, we believe that when we're not being helpful we're actually, in our devious, parental sort of way, actually helping! Your most important skill from your college education, and the one that people will admire you for and pay you for, is your ability to solve problems correctly and relatively quickly on your own and in groups. I believe, for instance, that you will quickly forget how the data are supposed to be entered for a chi-square test. I hope, however, you will have the confidence to say "I can figure that out."

### A peek at what happens during a typical class

We will begin discussing a few basic principles for the day regarding statistics, visualizations, experimental design, and the interpretation of information. We then will spend the remaining time working to solve an in-class exercise (ICE), provided as an electronic document on MyCourses.

## Expected Learning Outcomes

In this class you will be rewarded for achieving the following outcomes. If you successfully complete this class you will be able to:

1. develop and test hypotheses that are mutually exclusive and all inclusive using data;
2. correctly enter and manage data for different analytical procedures and visualizations 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. correctly determine the appropriate analysis given a set of data and the hypothesis/hypotheses to be tested. This includes parametric and non-parametric procedures;
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  $\text{\LaTeX}$ ;
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 a 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, and write and use (call) your own functions.
11. discuss how “big data” differs from “ordinary data.”

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. You should work to solve all problems posed in the ICEs. 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, Bing, NY Times, Shell, and others.

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

**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 mid-term and final 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. Your grade as a percentage will be determined by dividing your points earned by the number of points that were offered, multiplied by 100.

Item	Description	Points
<b>Pop quizzes</b>	Unannounced handwritten challenges for concepts and R coding. Number unknown	5 pts ea
<b>ICE</b>	In class exercises <i>may</i> be graded	10 pts ea
<b>Homeworks</b>	Number unknown	10 pts ea.
<b>Mid-term exams (2)</b>	Written and computer parts	100 pts ea.
<b>Final exam, written</b>		100 pts
<b>Final exam, computer</b>		100 (default) up to 150 pts (your choice)
<b>Semester project</b>	Includes assignments and group self-evaluation weightings	100 pts

Your final grade will be converted from a numerical value to a letter grade using the following rules. Note that I'll round your percentage to one decimal place (e.g., `round(your.grade, 1)`).

Score		Letter Grade		Score
93.3	$\leq$	A	$\leq$	$\infty$
90.0	$\leq$	A-	$<$	93.3
86.7	$\leq$	B+	$<$	90.0
83.3	$\leq$	B	$<$	86.7
80.0	$\leq$	B-	$<$	83.3
etc.				

**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 the final computer exam points range from 100 - 150. You choose the weight of these before submitted it.

## 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). This, along with the supporting data and R files, need to be dropped into the class folder in my InBox. You can learn about these at <https://wiki.geneseo.edu/display/cit/In+and+Out+Boxes>, long before the due date! 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 and all analyses and graphs are made. Each day any of these files is late will result in 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.

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. Most likely, however, you'd just throw this in with your result in the next section.
  - (e) Results. Share your results. The should be graphs and the statistics from statistical tests (e.g., the ANOVA table).
  - (f) Discussion. What do your results mean to the greater picture.
  - (g) Acknowledgments and citations.

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

Date	What's Due
<b>10-10</b>	One paragraph proposal - requires approval. Paper copy only. Everyone's printed name and signature must be on it.
<b>11-21</b>	Group project mini-presentation. This is a shortened version of your final presentation. Be sure to demonstrate that you have data, have made thorough analyses of the data, and have graphs showing clear results. The slides you present must be dropped in my InBox before class starts.
<b>12-8, 11:59 pm</b>	Drop in my InBox your presentation files: pptx, ppt, or pdf only, data files opened with your R script file(s), and your R script file(s). The filenames must include all your last names (see above). I'm totally flexible about the due date! You can hand it in <b>any time before</b> this deadline. You can even deliver it after this time but will be charged a late fee = 10% for each 24 hr period following the due date/time.

## 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 diseases 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 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. 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.

Finally, it can be dangerous to proceed without talking to me. 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.

## 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	Topic	Prep.for.Class	ICE
8/25/2014	Mon	Introduction to quantitative analyses	Syllabus, Chapt 1, install R and RStudio	Intro to R, "simple calculations," and Day 1 Survey
8/29/2014	Fri	Data management in Excel and Thinking about what Science is	Last day to withdraw. Chapt 2	Working with data using Excel and R
9/1/2014	Mon	Labor Day		
9/5/2014	Fri	Overview of visualizing data	Google search "R graphs" and try stuff	Visualizations
9/8/2014	Mon	Central tendency and variability	Chapt 4	Stats - summary and BMI
9/12/2014	Fri	Frequency Distributions; Build Groups for final projects	Chapt 3	The Prudential Ad
9/15/2014	Mon	Probability	Chapt 5	Probability
9/19/2014	Fri	Group presentations of data	Bring your idea and sample data	
9/22/2014	Mon	Introduction to Sweave	Intro to Sweave; install Tex programs (see Intro to Sweave ICE doc)	Intro to Sweave
9/26/2014	Fri	More on visualizations		Influenza - Google data
9/29/2014	Mon	Intro to statistical inference	Chapt 6	Inference
10/3/2014	Fri	Exam I	review previous work + readings	
10/6/2014	Mon	Go over Exam I		
10/10/2014	Fri	A little programming and the central limit theorem and genetic drift	One paragraph group project proposal	Intro to Programming
10/13/2014	Mon	Fall Break		
10/17/2014	Fri	Optimization problem		Can optimization (use .rnw provided for your write-up)
10/20/2014	Mon	Graph challenge presentations	Bring a printout of your original and recreated graph (Sweave it!)	

10/24/2014	Fri	“The only normal people are those you don’t know very well”	Chapt 8	Normality
10/27/2014	Mon	One- and two tailed one-sample + paired tests	Chapt 9	One-sample tests
10/31/2014	Fri	Two-sample tests	Chapt 10	Two-sample tests
11/3/2014	Mon	Non-parametric paired and independent tests		Non-parametric tests
11/7/2014	Fri	Testing hypotheses about frequencies	Chapt 7	Chi-square tests
11/10/2014	Mon	One-way ANOVA	Chapt 11	One-way ANOVA
11/14/2014	Fri	ANOVA design + two-way ANOVA	Chapt 12	Two-way ANOVA
11/17/2014	Mon	Exam II	review previous work + readings	Graphing climate data
11/21/2014	Fri	Group Progress Reports	Prepare and give PPT presentation with 5 slides ( $\leq 5$ minutes)	
11/24/2014	Mon	Correlation	Chapt 13	Correlation
11/28/2014	Fri	Linear Regression	Chapt 14	Regression
12/1/2014	Mon	Non-linear regression		Non-linear Regression
12/5/2014	Fri	Final - Written		
12/8/2014	Mon	Final - Computer		
12/8/2014	Mon	Submit by 11:59 pm all presentation files.		
12/12/2014	Fri	Final Presentations (3:30 - 6:30)		

## 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

- Hampton, R. and J. Havel, 2013. *Introductory Biological Statistics*, 3/e. Waveland Press.
- Hartvigsen, G., 2014. *A primer in biological data analysis and visualization using R*. Columbia University Press.