Biology 3000, 5000: Biostatistics, Fall 2021

Professor: Corey Devin Anderson, Ph.D. (Evolution, Ecology, and Population Biology)

Preferred salutation: "Dr. Anderson"
E-mail: coreanderson@valdosta.edu

Course Format:

Traditional Face-to-Face (F2F): Face-to-face classes generally have the following:

- a scheduled meeting place and
- a scheduled time and day(s) of the week.

Lecture location: BSC 1202
Days and time: Tues/Thurrs, 2:00 to 2:50 PM.

Lab location: BSC 3018 (after each lecture)
Days and time: Tues /Thurs, 3 to 4:50 PM.

Final exam: scheduled time
Wednesday 08 Dec: 2:45 to 4:45 PM.

Office Hours: Tuesday 9:30 AM to 11:30 AM

The lectures provide a survey of key concepts, topics, and tests in biostatistics; the labs are intended to reinforce the lecture material, as well as to provide basic training in statistical programming with R.

Standards

Education outcomes for BS Degree in Biology: **BO1** (Develop and test hypotheses, collect and analyze data, and present conclusions in both written and oral formats used in peer-reviewed journal and scientific meetings).

VSU General Education Outcomes: **GE3** (Students will use computer and information technology when appropriate); **GE5** (Students will demonstrate knowledge of scientific and mathematical principles and proficiency in laboratory practices); **GE7** (Students will demonstrate the ability to analyze, to evaluate, and to make inferences from oral, written, and visual material).

***Policy on appointments and "drop-ins"**
If you need extra help or clarification, the best method is Email (I try to be very responsive); after class or lab is usually also a good time for help. I don't mind scheduling appointments outside of office hours (I encourage students who are struggling to seek help), but I ask that you please try to take advantage of scheduled class times and office hours, if possible, as I cannot always accommodate meetings outside of the course schedule.

Course overview

This is an upper division course on the statistical analysis of biological data ("biostatistics").

The catalog lists BIOL 1107, BIOL 1108, MATH 1112 or 1113, and MATH 1401 (old 2620) as prerequisites. I interpret this to mean that the course is upper division, that you have some basic math skills (precalculus or otherwise), and some sort of background in statistics. That said, I recognize that most students (even the graduate students) taking this course likely have a very rough understanding of statistics (and probably a fear of mathematics in general), even if they have had Math 1113 and Math 2620. By virtue of the course topic (biostatistics) you should expect this course to be quantitative (and computational), but I do not consider the math in this course to be advanced (so don't psyche yourself out!).

Statistics are an essential tool in the biological sciences (including biomedical research), yet most biology students are poorly trained in statistics, partly because biostatistics is not usually a required course for a degree in biology. That said, even graduate school entrance exams (such of the MCAT) have put renewed emphasis on data analysis; note that the passage questions in the new MCAT are adapted from scientific journal articles and reflect the increasing importance of research in medicine. This is because medical practitioners need to be able to interpret the results of medical tests, and *good* practitioners should also be able to vet primary literature and/or publish their own cases or studies.

More generally, it is important to consider that scientific facts are ultimately derived, not from text books, but from the empirical scientific literature, published primarily in scientific journals, where inferences are supported by data and inferential testing. Not all studies require fancy quantitative methods, but whenever somebody is trying to test something based on a sample from a larger population, inferential statistics are likely to be involved. It is important to be able to understand and vet methods, as well as be able to apply such methods to your own data sets.

Because the large number of tests available, and the complex mathematical and computational theory underlying some tests, statistics can be a humbling domain for a biologist; the learning curve is seemingly exponential and constantly expanding; it is impossible to know everything. However, a better understanding of commonly encountered lingo, concepts, and tests is an important starting point. If you want to be a good analyst, you probably need to understand some of the theory underlying the different statistical models (so that you apply them correctly), but the most basic goal should be to know what test to apply based on the nature of your question and the type of data that you have (and how it is distributed)…and to be able to interpret test results.

A good biostatistics course should probably be two semesters; in the first semester you would cover basic concepts and tests, and in the second semester you would examine special topics in biostatistics that you are likely to encounter (e.g., multivariate statistics, morphometrics, circular statistics, spatial statistics, meta-analysis, etc.). Since Biology 3000 is limited to one semester, we will focus our attention on the primary topics.

Statistical programming (with R)

In the modern era, computers have facilitated the application of statistics by scientists and some methods (such as permutation testing) would be virtually impossible to implement without the aid of a computer. There are many statistical software packages available and, for many years, there was a trend towards programs with simple-to-use graphical user interfaces ("GUIs": Macintosh/Windows-style menus that you can point and click with a mouse…or, now, with your finger). Around the year 2000, many academics started to abandon proprietary statistical software with GUIs in favor of free, open-source statistical programming platforms. The most popular of these platforms is called "R".

R is a descendant of the statistical programming language "S" and has many advantages over proprietary statistical software packages that use graphical user interfaces. When you download R (for free), there are many functions that are built into the base distribution, but additional functionality can be contributed as packages of functions (developed by other R users) that can be downloaded and applied. Since R is open-source, anybody can see (or modify) the source code. Moreover, R is also a programming language, which means that you can write your own scripts (within which you can call functions from the base package or other contributed packages). This makes R extremely powerful because you are not limited to the choices on a menu; almost anything is possible.

The downside of R is that there is a steeper learning curve and, while anything is possible, it is not always obvious or easy to do certain things (such as custom graphics). On the flip side, once you learn the basics of R, some things are far easier (and faster) to do in R than with GUI-based software and, once the code is written, outputs can be reproduced much faster than what is possible with a GUI (especially if you have a lot of repetitive/batch procedures).

R is a statistical programming language and is intended for doing statistical analysis. Some functions that can be called in R may be written in other languages, such as "C". R is similar to "Python" (both are "interpreted" languages) and both R and Python are the most popular platforms for "Data Science." In the research sector and industry, there is a huge demand for people with R and Python programming skills.

In this course we will focus on statistical programming, which includes an introduction to some basic concepts in computer programming, such as how to write a "function" and a "loop". I recognize that most biology students do not have a strong background in computer science, so part of the goal (beyond teaching you the state-of-the-art in statistics) is to improve your computational skillset. This is a very important skill in the day and age of "big data." Many data sets (such as genomic data sets and large medical research studies) are simply too large for point and click interfaces; computer programming is now a required skill in most realms of biology, including medical science.

While some people have a better natural aptitude for computer programming, nobody is born knowing how to use R or how to write a loop. You get better at doing this sort of thing by practicing, a lot (I would suggest *every day*). Programming is a one-step forward, two-steps backwards process: solving one problem often leads to another…but if you are determined and keep battling, eventually you will succeed…and that's how you get better.

I have had students cry and moan about R, but then get into graduate school or get a job because they had training with R (and inevitably they thank me for it). I am not teaching you R because it is free, I am teaching you R because it is the most powerful platform for doing statistical analysis, period.

Grading

I use a rank-based (or "stack rank") grading system; this means that you will be evaluated based on how well you perform (in terms of your point total) relative to other students in the class.

When possible, I like to use natural breaks in the point distribution to determine letter grades. For example, if there is a substantial point differential separating the top five students in the class from the remaining students, these top students would typically receive an "A". Conversely, natural breaks at the bottom of the distribution determine those students that do not pass (i.e., D/F). In the case that discrete natural breaks in the distribution do not exist, I will use quartiles of the distribution to assist in parsing the distribution.

There are approximately 930 points that can be earned in this course:

- 300 points from unit exams (lecture portion).
- 300 points from problem sets.
- 250 points for R exams.
- ~80 points for pop coding quizzes (~ 8 pop quizzes worth 10 points each).

There will be three unit exams (mainly multiple choice format, with some written answers), each worth 100 points. The first exam will be in the middle of September. Exams have both a multiple choice component as well as an R component, worth an additional 50 points. The third unit exam will be cumulative. The final R practicum will be given after Unit Exam 3.

Lab exercises are mainly intended to reinforce course concepts through the application of statistical programming (with R). Proficiency with lab exercises will be gauged via five problem sets (worth 60 points each), pop coding quizzes, and the R exams (following the lecture exam). At a certain point in the semester, you will be practicing R by doing chosen problems from the textbook, in R.

Note that problems sets make up a substantial portion of your final grade. This means that a strong performance on problem sets can raise your class rank considerably; conversely, blowing off problem sets will likely result in a poor grade (no matter how well you do on the tests). I will not drop any problem set or test grades, so make sure you give all tasks your full effort.

Problem sets, in particular, are a very important part of a statistics course. Sometimes an important part of the process is struggling with a solution, but, through sustained effort and learning, eventually solving the problem.

Grading for graduate students

Graduate students enrolled in the class will be graded on similar criteria. Letter grades for graduate students will be evaluated based on where they fall in the undergraduate student distribution and where they fall relative to other graduate students. Graduate students will be removed from the undergraduate distribution before determining grades for undergraduates.

Books

Required text:

1) The analysis of Biological Data by Whitlock and Schluter (3rd edition); the publisher is W.H. Freeman (Macmillan Learning).

https://www.macmillanlearning.com/college/us/product/Analysis-of-Biological-Data/p/131922623X#:~:text=The%20Analysis%20of%20Biological%20Data,of%20statistics%20for%20biology%20students.&text=These%20include%20new%20calculation%20practice,medical%20and%20human%20health%20research.

This is an excellent introductory textbook, and most of the lecture material will be follow the topics in the book. I have chosen this book because it is easy to read (relative to most statistics texts), it has lots of practice problems, and it does an excellent job at explaining some of the more challenging concepts.

Field trips

The 1 hour and 50 minute time slot for labs means that field based data collection is not feasible. Some semesters, we will do a class fishing trip (to compare types of bait and examine random effects of different fishermen). This will probably not happen in fall 2020 because of COVID-19 precautions.

Cheating policy

Do NOT cheat on exams. You will receive a zero on the exam and will be reported to the Dean of Undergraduate Academic Affairs. I consider copying of problem sets/computer code to be cheating! Use must sign the agreement on code plagiarism to be formally matriculated into this course.

Calculator policy

Some unit exam questions may require a calculator….so remember to bring one to the unit exams.

Cell phone and computer policy

Unless you have special permission, **cell phones and computers are strongly discouraged during lecture**. Students who have cell phones out during exams will receive a zero on that exam. Any student caught photographing an exam will get an automatic "F", and will also be banned from retaking the course with Dr. Anderson.

Policy on audio recordings

I prefer that my lectures and labs not be recorded (especially without my consent), but if you feel as if you need to record a lecture, please place your recording device in the front of the classroom, so that I am aware that I am being recorded.

Students with disabilities

Students requiring classroom or testing accommodations because of documented disabilities should discuss their needs with the instructor at the beginning of the semester. Students not registered must contact the Access Office, Farber Hall, Phone; 245-2498. Website: http://www.valdosta.edu/access/
For some students, the presence of a medical condition places them at high risk for COVID-19. These students can use the online form to submit documentation of the condition to the Access Office to ensure confidentiality.

https://www.valdosta.edu/student/disability/forms/request-for-covid19-course-modification.php

The Access Office will then contact the advisor and department to indicate the receipt of documentation that supports the request for course substitutions or appropriate alternative assignments and virtual access to lectures.

Fall 2021 (addendum): VSU COVID-19 policies:

VSU cares about student success both on and offline, and a variety of resources are available to help studentsboth academically and personally during the Fall 2021 semester. One of the best resources is VSU's Coronavirus FAQ page located at https://www.valdosta.edu/health-advisory/coronavirus.php. Information is available there about a variety of topics in VSU's return-to-campus plan.

I will only be broadcasting lectures when there is a student out with an official Email from VSU student affairs.

Many of the problem sets in this course will be based on writing computer code in R to solve various statistical problems.

Blatant copying of computer code from the internet* or from other students** is unacceptable and is grounds for failure in BIOL 3000/5000.

*When writing computer code, we often get stuck at certain steps (that we are unsure how to solve) and it is perfectly normal (if not requisite) that you use the internet for help. However, there is a big difference between using the internet as a resource to learn new functions or tricks when writing code and blatant copying of large tracts of code from the internet without any effort to do it yourself. The latter is obvious to detect and is not allowed in this course.

**Copying code from other students is also forbidden. Writing code is like writing a paper: you would never be allowed to turn in the same paper in an English course…or a slightly reworded version of the same text (in a sad attempt to disguise the fact that you copied code). I grade all of the assignments (with a close eye for detail) and it is obvious when code is copied.

By signing this form, you acknowledge that you have been warned that violation of the course policy on code plagiarism will result in a zero on that assignment and potential failure in the course.

_____ (signature)

_____ (date)