Botany 575 — Special Topic: *Intro to Modern Statistical Methods for Biologists*

Fall 2016

(Provisionally, meets with **Statistics 479**)

Course Structure and Motivation

Structure: *Introduction to Modern Statistical Methods for Biologists* (IMSMB) is an introductory statistics course for biologists that differs from Statistcs 371, *Introductory Applied Statistics for the Life Sciences*, in a few key ways, despite being aimed at the same audience of students and containing a large overlap in topical coverage. First, IMSMB will be linked with some of the discussion/lab sections of Biology/Botany/Zoology 151 much as freshman interest group (FIG) courses are with concurrent enrollment. Second, the order in which statistical topics are introduced will be modified to be more compatible with the needs to use certain statistical methods in the Biology 151 lab. Third, IMSMB will be blended in that one third of lectures will be held in a WisCEL classroom where students will use the statistical software R (and RStudio) to do data analysis on biological case studies that are selected to have connections with the biology they are learning concurrently in Biology 151 and the statistical methods they are learning in IMSMB with some other activities traditionally done during lecture shifted to the students' own time. Fourth, gaining the mastery to integrate reproducible statistical analysis into report writing is an explicit learning goal of the course. Fifth, the Bayesian approach to statistics will be integrated with the classical approach to statistics that is taught in almost all introductory statistics courses.

Motivation: Despite the facts that: (1) the majority of students that take introductory biology have already also satisfied the college algebra competency that is a prerequisite for introductory statistics, (2) using statistical methods is an important part of the lab component of Biology 151 and the independent research component of Biology 152, and (3) most biology students will take one or more courses in statistics to satisfy degree requirements before graduation; most biology students that take introductory statistics do so in their junior and senior years, long after they have completed the Biology 151/152 sequence. Part of the reason for this is that the demand for all introductory statistics courses (Statistics 224, 301, 302, 324, 371, and 571) is so high that very few seats remain at the time that (mostly sophomore) students enroll in Biology 151. Linking a new course with additional capacity will help to alleviate this situation and guarantee places in introductory statistics for some 151 students. (When IMSMB is taught as a pilot in Fall 2016, the capacity will only be 66 students, but if successful and if additional TA and WisCEL resources become available, the capacity could grow, perhaps two to three times as large with a single lecture, but additional 22 student discussion sections and 66 student WisCEL sections.)

Biology students, some of whom are mathphobic and may not understand the need to learn statistics, may perform better when the statistics course and material is explicity connected to the biology they are learning concurrently. The practice of biology includes quantitative reasoning and students will benefit when they learn data graphics and analysis specifically motivated by the biology in their coursework.

In my previous experience, teaching biology students to learn the statistics software R has been ineffective and unpopular (the source of most negative comments on student course evaluations). The primary reason is that all student R work has been homework and when students have struggled, they have not had access to immediate help, which can be very frustrating. When first learning R, most people do not have adequate knowledge to be able to access and understand online help. The

initial learning curve is very steep. By using the WisCEL classroom, all students will actively use R in an environment where help from the instructor, the TA, and peers, will be available immediately. Also, as students will apply their knowledge of R concurrently with their data analysis in Biology 151 labs and have their TA in that course as another resource, students should experience a much greater degree of success in developing mastery in a statistics package that is widely used in many areas of biology.

The free software RStudio provides an environment for reproducible data analysis and report writing using a single tool. RStudio makes some aspects of using the R language simpler, especially for novices. Students will be able to use the software and skills for creating reports on case studies in IMSMB that they can also use for their lab work in Biology 151.

Most introductory statistics courses for nonstatisticians aim for students to gain an understanding of several statistical concepts, but do not explicity prepare students to continue their education in statistics or to become able data analysts. The IMSMB course aims to augment conceptual understanding with a foundation that students will build on as they continue their education in statistics, either with additional formal coursework (as many students that progress to graduate school will do), or as part of training to become a medical professional, or in many jobs for biologists that do not require graduate training. There are many areas in biology where the Bayesian approach to statistics is widely used and the trend is for this branch of statistical application to become increasingly widespread within the biological sciences (and more generally). While fully teaching the Bayesian approach in addition to all of the standard introductory statistics topics is too much for a single course, failing to prepare students to learn anything about Bayesian statistics in what is very likely the only statistics course to be part of their undergraduate experience is doing a great disservice to all studentsof biology and especially to those whom will be expected to use and apply modern statistical methods after graduation. Many modern statistical methods are Bayesian and many require advanced computation; this course is designed to prepare biologists to learn what they will need to know long after the course is complete.

Course Description

The course will provide students with a practical and conceptual introduction to modern statistical methods for data analysis common to biological and medical application. Topics include: effective statistical graphics; probability and conditional probability as used in Bayesian and classical statistical models; random variables and common distributions; hypothesis testing and confidence intervals from simulation, resampling, and formulas; contingency table analysis; basic linear models (analysis of variance, regression analysis); model checking; statistical concepts of experimental design; biological and medical applications.

Course Materials

Required textbook: *The Analysis of Biological Data*, Second Edition, by Whitlock and Schluter (eventually replaced with a new textbook);

Course notes: *A first course in modern statistics for biologists* by Larget;

Free Web Primer: *An R companion to A first course in modern statistics for biologists* by Larget.

Course Objectives

IMSMB aims to teach biologists to *think statistically* when planning and conducting data analysis and when evaluating statistical results and interpretations from scientific and general publications. The course aims to motivate biologists to recognize and appreciate the value of statistical science, to introduce biologists to modern statistical methods in concept and in practice; to provide students with mastery of modern statistical software for basic biological applications, and to prepare students for continued education in statistics beyond the course.

Learning Objectives

By the end of the course, the successful student will:

1. be able to produce effective graphical displays of data;
2. have gained conceptual and practical mastery of statistical methods in many common settings;
3. have obtained proficiency in the statistical software package R for basic statistical models;
4. have reached a comfortable understanding of many statistical concepts, including estimation, hypothesis testing, statistical modeling, and how the process of data collection affects inference;
5. be able to undertand conceptually, distinguish, and apply a variety of statistical approaches to data analysis from classical and Bayesian points of view;
6. to understand how to use simulation for statistical inference;
7. to write reports that integrate reproducible data analysis.

Computing

The statistical software R is widely used, within the statistics community, but also across many disciplines including biology. The R language is extended with thousands of packages. Many of these packages are developed for specific biological application areas, including packages for data analysis in ecology, evolution, and bioinformatics. Students will learn R working collaboratively on data analysis projects for biology case studies in a WisCEL classroom environment. The R skills

students develop will be applicable in the lab assignments in Biology 151. Mastering R will require regular and extensive practice, but the skills gained will be very useful after the course ends.

Quick Questions

Prior to each regular lecture, students will be asked to read sections of the textbook and course notes and take an online quiz that tests their understanding of basic concepts. This outside of class work will enable us to use in-person lecture time more effectively for activities that promote deeper understanding of more challenging ideas and concepts in the course.

Homework

There will be weekly homework assignments from textbook problems weighted equally on a scale from 0 to 5. Your homework solutions should be organized and neat with solutions in order the order problems were assigned. Each problem solution should include a brief description of the problem (that may be paraphrased from the actual problem) as well as your work. *If your assignment is not neatly organized with problems in order and is not clearly legible, making it easy for the grader to follow the approach you take for each solution, your grade for the assignment will be lowered by 2 points.*

For well-organized and neat assignments, this is the grading rubric.

Points	**Characteristics**
5	Almost all problems are essentially correct with no major conceptual flaws. There may be some minor errors or calculation mistakes.
4	One problem is incomplete or contains a major conceptual flaw, but most problems are essentially correct. There may also be some minor errors or calculation mistakes.
3	At least two problems are incomplete or contain a major conceptual flaw, but most problems are essentially correct. There may also be some minor errors or calculation mistakes.
2	More than half the problems are incomplete or contain a major conceptual flaw, but there is evidence that the student made a serious attempt to solve most problems. The student gets some parts of some problems correct.
1	The assignment shows little progress toward a correct solution on any problem, but there is evidence that some serious effort was put forth on at least one problem.
0	The assignment is not turned in or contains no evidence that the student put forth serious effort on any problem.

Case Studies

Each week will include a data analysis case study and short report. Much of the work for each case study will be able to be completed during a lecture in the WisCEL classroom. Students will learn to use RStudio to write reports using the R markup language. Such reports will combine writing, graphics, and data analysis in a single document, promoting reproducible data analysis.

Exams

There will be two midterm examinations during the semester and a final examination during finals week. *At my sole discretion*, I may permit alternative examination times for students who give me *ample prior notice* of an acceptable reason, such as a university-related conflict (travel to an academic conference or participation in a sporting event). I will not change examination times during the semester for attending family functions, extending breaks, sleeping in, or missing a bus. I will not provide alternative examinations. If an examination is missed and I grant an excuse, the student may use the score on the final examination to replace the missing score.

Grading

The final course grade will be determined by a score made up from these weighted sources.

Quick Questions (lowest 3 scores dropped)	15%
Weekly Homework (lowest score dropped)	15%
Case Study Reports	15%
Midterm Examinations (15% each)	30%
Final Examination	25%

Discussion Sections

Assuming there is sufficient space, you may attend any discussion section without changing your registration. Time in discussion section will typically be used to solve problems similar to those on assignments, to ask questions, and to review past assignments.

Academic Honesty

You are permitted and, in fact, *encouraged* to talk to other students, your teaching assistant, or me about homework. Your TA or I may give you clues or discuss similar problems without doing your homework for you. You may look through books or Web pages for solutions to problems. However, you may not present other people's work as your own. Make sure to include with any submitted solutions to problems references to any sources of direct assistance. If you work with other students solving problems, make sure that you write up your own solution independently. It is not acceptable for one student to write a solution for another student to copy.

You must work independently during exams. You may not share calculators, pass notes, or use a laptop computer during the exams.