

Statistics/Forestry/Horticulture 571 Fall 2011
STATISTICAL METHODS FOR BIOSCIENCE I

| | | |
|-------------|-------------------------|-------------------------|
| Instructors | Bret Hanlon (lecture 1) | Bret Larget (lecture 2) |
| Office | MSC 1241 | MSC 1275A |
| email | hanlon@stat.wisc.edu | brlarget@wisc.edu |

| Lecture 1 | | Lecture 2 | |
|---------------------|-----------------------|---------------------|----------------------|
| TuTh 9:30 – 10:45am | SMI 331 | TuTh 11am – 12:15pm | Ingraham 22 |
| Discussion sections | | Discussion sections | |
| Tu 1:00 – 2:15pm | Engineering Hall 2535 | Tu 1:00 – 2:15pm | Gym.-Natatorium 1140 |
| Th 8:00 – 9:15am | Social Science 4314 | We 2:30 – 3:45pm | Sterling 1333 |
| We 1:00 – 2:15pm | Noland 579 | We 1:00 – 2:15pm | Grainger 2185 |

About the course: Statistics 571 is an introductory course in statistics aimed at graduate students in the biological sciences across multiple departments at UW. Let's say more about this:

Statistics 571 is an *introductory course*. We will not cover in a single semester a comprehensive selection of topics that will meet the needs for all of the statistical methods that each of you will require for your graduate research. Many of you should continue formal course work with Statistics 572 and a few of you would benefit by taking courses from among dozens of statistics courses offered by the Department.

Statistics 571 is a course for *graduate students*. The Department offers many different introductory statistics courses, mostly for undergraduates, with different expectations for mathematical preparation and area of application. As a graduate course, Statistics 571 aims to move more rapidly and more deeply through course topics than our undergraduate course Statistics 371. We will use a textbook extensively for the course, but will expect students to have the maturity to respond positively to our choices to cover some topics not in the textbook and to jump around in order to present topics in the order we choose. In addition, we expect graduate students to take more ownership of their own education than we would of undergraduates. We expect students in the course to seek and use outside resources as needed if the demands of the course are uncomfortable. Some students in the course have never taken statistics before, or if they have, may have lost whatever mastery they once possessed of the material. Others of you mastered a previous course in statistics and have deepened your understanding of the field through substantial application of methods in your research and/or reading the literature in your field. It is impossible to design a pace that is ideal for both extremes; we aim for the target audience who has seen some of this material previously, but requires review, and who is ready to obtain a deeper understanding of both statistical concepts and methods than they had previously. Others need to adjust expectations and their approach to the course.

Statistics 571 serves students *across the biological sciences*. Some of you are dairy scientists who will design experiments with dairy cattle to assess the effects of different factors on quantitative measures of milk production and cattle health. Some of you are ecologists who want to know how to model and assess the relationships among multiple variables you can measure to address questions about nature. Some of you need expertise in methods for analyzing data that arises from high throughput biological molecular measurements. Most of you have research interests and statistical needs distinct from these examples. In this course, we will use real data and examples from a variety of areas within biology, teach concepts that are generally important to understanding statistical inference, and teach methods that are widely applicable. However, we cannot be comprehensive; we may not teach the specific method you need and we may not use an example directly from your field of study. We will do the best we can to be useful to all of you.

Course logistics: Statistics 571 is taught in two different lectures by two different instructors. However, as much as possible, the two lectures will cover approximately the same material at the same speed, and, in fact, are based on identical lecture notes. In practice, we treat both lecture sections as part of a single course as much as is feasible. Students may attend the alternative lecture when desired, if there is sufficient space in the lecture hall. The exams, homeworks, policies, and grading will be identical. Students may attend the office hours of any professor or TA involved with the course. Students should select one discussion section to attend regularly as this makes it easier for us to return homework to you. However, *the discussion section you attend need not be the same discussion section in which you are enrolled, space permitting*. Discussion sections *will not meet* during the first week of classes, during the two weeks when midterm exams are given, nor during the week of Thanksgiving.

Course websites: See <http://www.stat.wisc.edu/courses/st571-larget/> for handouts, homework assignments and other course information. See <https://learnuw.wisc.edu/> for the gradebook. See <https://wischolar.wisc.edu/stat571/> for a course blog for discussion about the course.

Discussion board/Blog: We will pilot a new online tool, WiScholar, as a common discussion board for all students in the course. All people associated with the class should be able to login with their university netid and read and post messages to the blog. This blog is primarily a forum for peer-to-peer discussion about the course. The blog will supplement, but not replace, face-to-face communication with course instructors and TAs. Good usage includes: (1) posting issues with the course or areas of confusion; (2) posting questions about homework problems hoping for *hints* to solutions; (3) clearing up class logistics and policy. The instructors and TAs will respond to and participate in the blog when appropriate, but we will not be monitoring the blog 24/7. You should not expect a rapid response from a professor or TA for a specific question, but can hope for more timely help from the larger community in the course.

Communication with instructors: Questions regarding course content should be made via the blog on the WiScholar site. The reasons are that responses from instructors, TAs, and knowledgeable peer students to one student are shared and easily accessed by all, the communication is automatically archived, and the blog provides an opportunity for students in the course who understand the material well to provide answers from a student perspective. Direct email to instructors should be used only for private correspondence.

Course objectives: The primary goal is to provide graduate students in the biological sciences with a deep understanding of statistical concepts and methodology, so that they may apply it to their own research and answer questions like: How can I visualize my data? Which statistical method is most appropriate? How can I check that the assumptions of the method are met? How do I interpret the results in the context of the biology? A secondary goal is to instruct students in the use of the software package R so that they become competent users for basic analyses by the end of the semester. We will not require you to use R, but recommend that you take advantage of your time in this course to gain some mastery of this valuable tool. We will use case studies and genuine data examples whenever possible to illustrate statistical concepts. Mathematical and computational complexity will be minimized when deriving/presenting methods and their justifications; however, students should expect to use both mathematics and computation as necessary for practical data analysis for completing homework.

Textbook: We have selected the text *The Analysis of Biological Data* by Michael Whitlock and Dolph Schluter as a required text for the course. This text is readable for the audience, is reasonably complete for the topics we wish to cover, contains many interesting biological examples, contains many well-written *interleaf* essays on important statistical concepts, and does a good job of presenting up-to-date statistical methods. We believe the text to be a valuable resource during the course and beyond. We will not, however, limit ourselves to the topics in the textbook or to the order/fashion in which the ideas are presented. In particular, we expect you to be able to read many early chapters on your own to review/learn the material

therein and we will incorporate advanced methods from later chapters as the applications for which they are appropriate are introduced. In addition, we have opted for a *just-in-time* approach for teaching both statistical graphics and probability. Rather than presenting these ideas in a complete chunks (or as complete chapters in the textbook), we introduce methods and concepts from graphics and probability in the context of data analysis or model-based inference appropriate for the statistical topic at hand.

Furthermore, we recommend the textbook *Statistics for the Life Sciences*, 3rd edition, by Samuels and Witmer, as a supplemental textbook. We will post suggested additional problems from this textbook and their solutions for those of you who need to work more problems than what we assign for homework in order to achieve mastery of the material. Both textbooks are on reserve at the Wendt Library and the Samuels/Witmer text is also on reserve in the Steenbock Memorial library.

Computing: Students may use whichever statistical package they choose for the course. However, we strongly encourage the use of R. The instructors and TAs will demonstrate R in class and will answer questions on its use. We will not provide this assistance for alternative packages. We are convinced that R is the best statistical package available for the analysis of biological data and that the effort extended to learn R during the course will provide you with a valuable tool for your career. Even if R is not the dominant statistical package in your discipline today, it is likely to be so (or at least highly influential) when your career as a biologist is still young. In addition, you must learn R if you plan to take Statistics 572 (or other courses in the statistics department).

If, however, you wish to leverage previous expertise or to use the statistical package that is most common in your field today, we will not disallow this. Beware, however, that support for using any alternative statistical package must come from outside resources. We know that R can be used to do everything we ask you to do for statistical computing in this course and we will teach you to do so. If you choose a different package, the risk and consequences are yours.

R is available for free download online at <http://cran.r-project.org/>. R is available on all commonly used platforms (Windows, Mac and Linux). R is the *de facto* standard statistical computing package used in graduate programs in statistics and among a majority of professional statisticians. Its influence and importance is growing in many other academic fields, including biology, and in major corporations. There have been over 100 books published on using R for statistical computing, including over 45 since 2009, and several are aimed at biological audiences.¹ The basic software has been extended with more than 1000 packages contributed by users, many of which are specifically tailored for use in biological applications.

The course assumes no prior experience with R. Developing proficiency with R requires practice and learning to interact with the computer by typing commands rather than using a mouse and menus. R includes a high-level programming language. Advantages include an extensive code base that performs many tasks, flexibility to customize functions and graphs, and the ability to keep a commented history of steps of an analysis which allows one to reconstruct and redo analyses from scratch with a simple command.

Learning R does not come easily to all, and while existing help and documentation is quite plentiful, finding help at the appropriate level can be challenging. To help you, we will provide links to R tutorials on the course web page and each section of lecture notes will contain a supplement with R commands used for the graphs and statistical analyses from lecture. Some discussion section meetings will spend time teaching you to use R, so you are especially encouraged to bring your laptop to class these days, if you have one.

Exams: There will be two in-class midterm exams and a 2-hour final exam. All exams will be open book, open notes, and you will be allowed a calculator (but not a laptop). The first midterm exam will be on Thursday, October 20, and the second midterm exam will be on Thursday, November 17, the week before Thanksgiving. There will be no homework assigned the week before the midterms. Discussion section will not meet the week of exams, but TAs will be available in their offices during normal discussion times.

¹See <http://www.r-project.org/doc/bib/R-publications.html> for a complete list.

The final exam will be on Monday, December 19 , from 5:05pm to 7:05pm, in a location to be announced. Notice of any conflict with these dates must be given to the instructor within the first week.

Grading: Semester grades will be based on homework (20%), two midterms (20% each), and the final exam (40%). Letter grades will not be given for midterm exams, but information will be provided to let you know how you are doing in the class.

Homework: Assignments will be assigned and posted on the course webpage on Thursdays, to be handed in by the following Friday by 4pm to your TA's mailbox. Solutions will be posted the following Monday. (This schedule will be modified near Thanksgiving.) The mailboxes are in the hallway just inside the main University Avenue entrance to the Medical Sciences Center. You must show your work and organize it to get full credit. Homework should be neat, clearly legible (typed or written legibly), order the problems in the order assigned, and use complete sentences in proper English. Assignment that ask for the use of a statistical package should include only short samples of computer output. Including output from R or other packages does not preclude the necessity to write out answers: do not expect the grader to find the answers somewhere in an included excerpt of computer output.

Late homework will be accepted only under extenuating circumstances and: (1) only with prior notice to and permission from your instructor; and (2) only if it is turned in before solutions are posted. Unexcused late homework will receive no credit. Each student is allowed to drop two assignments (which might have been skipped due to illness or other factors). The lowest two scores will be dropped if a student completes more than the minimum number of assignments.

Academic honesty: You are encouraged to work together with classmates and talk to your teaching assistant or instructor about your homework. We are convinced it is very beneficial to share and discuss ideas. However, you may not present other people's work as your own. Even if you work with other students solving problems, you still must independently write up your own solutions, run your own statistical computations, and produce your own graphs. You must work independently during exams. You may not share calculators or pass notes during exams.

Laptop policy: You may enjoy the wireless capability of the classroom so long as you stay on task. Advantages to using a laptop include: taking notes, viewing lecture notes rather than printing them, experimenting with R or other computing software, and so on. There are also limitations; figures and sketches drawn on the board cannot easily be replicated on a notebook in the classroom, for instance. In addition, activities such as emailing, web surfing, and gaming are not allowed in class. These activities are a distraction to classmates: be respectful of others. Be sure the sound is off at the beginning of the class.

H1N1 influenza: We are all encouraged to stay home when sick (including instructors and TAs). Colleague coverage will be used as much as possible to continue instruction in the case that instructors become ill. Students who are sick will be responsible for getting class notes that they have missed and for making up assignments or exams within a reasonable period of time. Students do not need to communicate with instructors if they miss class or discussion for sickness. They do need to contact the instructor by email *in a timely manner* in case they need to miss an exam. Students will not need to provide medical excuses for absences from flu-like symptoms. The situation calls for trust and responsibility among all of us.

Tentative schedule

| Day | Date | Topics | Chapter/Section |
|------------|-------------|--|---------------------------|
| T | Sept. 6 | Introduction; Data | 1.1, 1.3, 1.5 |
| R | Sept. 8 | Samples and Populations | 1.2, Interleaf 2 |
| T | Sept. 13 | Proportions: Graphs and the Binomial Distribution, other Discrete Distributions | 3.4, 5.1–5.6, 7.1, 7.4 |
| R | Sept. 15 | Probability: Sampling distributions; Mean; Standard error | 4.1–4.2 |
| T | Sept. 20 | Proportions: Estimation; Likelihood; Confidence Intervals | 7.3 |
| R | Sept. 22 | Proportions: Hypothesis Testing | 6.1–6.7, 7.2, Interleaf 3 |
| T | Sept. 27 | Probability: Conditional probability; independence | 5.1–5.6 |
| R | Sept. 29 | Probability: Probability Trees; Bayes' Theorem | 5.7–5.10 |
| T | Oct. 4 | Contingency tables: | 9.1–9.7 |
| R | Oct. 6 | Proportions: Comparing two or more samples | |
| T | Oct. 11 | Probability: The normal and t distributions continuous distributions | 10.1–10.4, 11.1 |
| R | Oct. 13 | Probability: Sampling distribution of the sample mean The central limit theorem | 10.5–10.6 |
| T | Oct. 18 | One sample inference: graphing; | 2.1–2.1, 11.1–11.2 |
| R | Oct. 20 | MIDTERM 1 sampling distributions and estimation | |
| T | Oct. 25 | One sample inference: the bootstrap | 19.3 |
| R | Oct. 27 | One sample Inference: hypothesis testing | 11.3–11.4, 19.1 |
| T | Nov. 1 | Two sample inference: graphs; randomization/permutation tests | 19.2 |
| R | Nov. 3 | Two sample inference: confidence intervals and testing | 12.1–12.6 |
| T | Nov. 8 | Power and sample size determination | 14.7 |
| R | Nov. 10 | Assumptions and Transformations | 13.1–13.3, 13.8–13.9 |
| T | Nov. 15 | Elements of Experimental Design | 14.1–14.9 |
| R | Nov. 17 | MIDTERM 2 | |
| T | Nov. 22 | ANOVA: The F-test and estimation of variance; Randomization | 15.1–15.8, 19.1 |
| R | Nov. 24 | THANKSGIVING BREAK | |
| T | Nov. 29 | ANOVA: Inference for contrasts, Multiple Comparisons | |
| R | Dec. 1 | Probability: Correlation | Interleaf 4 |
| T | Dec. 6 | Regression and related topics | 17.1–17.10 |
| R | Dec. 8 | | |
| T | Dec. 13 | | |
| T | Dec. 15 | | |
| M | Dec. 19 | FINAL EXAMINATION, 5:05–7:05pm | |