STAT 135: Intro to Statistics via Modeling

About the Course

Instructor Nicholas Horton (nhorton@amherst.edu, Seeley Mudd 303, 413-542-5655). My office hours will be held on Mondays 10:30-11:50am, Wednesdays 4:00-5:00pm, Fridays 3:00-4:00pm, and by appointment.

Description Introduction to Statistics via Modeling is an introductory statistics course that uses modeling as a unifying framework for much of statistics. The course provides a basic foundation in statistics with a major emphasis on constructing models from data. Students learn important concepts of statistics by mastering powerful and relatively advanced statistical techniques using computational tools. Topics include descriptive and inferential statistics, probability (including conditional probabilities and Bayes' rule), multiple regression and an introduction to causal inference. This is a more mathematically rigorous version of STAT 111 (students may not receive credit for both STAT 111 [formerly MATH 130] and STAT 135). Prerequisite: MATH 111.

Textbook We will be using "Introductory Statistics with Randomization and Simulation" (First Edition) http://www.openintro.org/stat/textbook.php, ISBN: 978-1500576691) as the text for the course. While the book is available as a free download, I strongly recommend that you purchase a copy of this book, read it closely and bring it to class on your laptop. A copy will be on reserve. The book is available for purchase from Amazon, for an affordable $6.48 (plus shipping) as of September 1st, 2014.

In a college course, the textbook is not just a reference to use after the instructor has presented new material but a sourcebook to use at every stage of learning. When all students read the text before class, the nature of the class meeting changes to the benefit of everyone. You have thought about the material, and you arrive with your own questions. You're ready to discuss what you understand, to clarify what you don't understand, and to hear more on the topic. You need to read ahead in the book prior to class as well as review the material after we've discussed it in class. There will also be reading questions due before we discuss material in class, as we will be focusing class time on activities to help understand concepts. Self-study in advance is needed to be successful.

Classes Class meets 2:00-2:50pm on Mondays and Fridays and Wednesdays from 2:00-3:50pm in Merrill 131. I expect you to attend class. Your participation is an important part of the learning process. If you cannot attend a particular class I would appreciate the courtesy of advanced notice and an explanation for your absence. Class participation and attendance plus presentations contribute 10% to your final grade.

After most classes, you will submit a short response answering the following three questions. There is a link to enter the essay on the Moodle front page. These are due by midnight of the day of the lecture. Don't feel you need to spend more than a few minutes, but if there is an issue that's important to you, take whatever time and space you need. If you were not in class, just say so.

1. What was the main topic of today's class?
2. What was the most important thing that you learned during today's class?
3. What important question do you still have?

While the class is in session, you should not use your computer or cell phone for personal email, web browsing, Facebook, or any activity that's not related to the class. I know that it's tempting to do this, but doing so creates an atmosphere of inattention in the classroom that interferes with learning.

Please plan to bring your laptop to class (we'll need to have sufficient numbers to allow you to work in pairs and take advantage of the technology available to us). Please contact me if you do not have access to a working laptop.

Policies

Attendance Your attendance in class is crucial, as is your punctuality. We are all going to learn this material together, so I need to have everyone present and working. I will make my best effort to provide accommodation for an unavoidable absence if you notify me well in advance. One necessary absence during the semester is not unusual; having more than two is uncommon.

Collaboration Much of this course will operate on a collaborative basis, and you are expected and encouraged to work together with a partner or in small groups to study, complete homework assignments, and prepare for exams. However, every word that you write must be your own. Copying and pasting sentences, paragraphs, or large blocks of R code from another student is not acceptable and will receive no credit or a penalty. No interaction with anyone but the instructor is allowed on any exams or quizzes. All students, staff and faculty are bound by the Amherst College Honor Code.

Amherst College Honor Code Statement (excerpt)

> Every person's education is the product of his or her intellectual effort and participation in a process of critical exchange. Amherst cannot educate those who are unwilling to submit their own work and ideas to critical assessment. Nor can it tolerate those who interfere with the participation of others in the critical process. Therefore, the College considers it a violation of the requirements of intellectual responsibility to submit work that is not one's own or otherwise to subvert the conditions under which academic work is performed by oneself or by others.

Cases of dishonesty, plagiarism, etc., will be reported, per the full statement of intellectual responsibility.

Resources

Moodle The Moodle site will be regularly updated with lecture handouts, project information, assignments, reading quizzes and other course resources. I have created discussion forums for questions that arise in and out of class.

Computing The use of the R statistical environment with the RStudio interface (downloadable from `http://rstudio.org`) is thoroughly integrated into the course. You'll be using R primarily through a version of RStudio accessible on the web at the URL (`http://r.amherst.edu`). This software is also installed on most computer labs on campus.

RStudio is free software that can be installed using the version of R on your own machine (download information for R can be found at `http://r-project.org`). Running it as a client on your own machine may be helpful during the projects.

More resources and short videos can be found at `http://www.amherst.edu/~nhorton/rstudio`. This includes pointers to the *Five Colleges Guide to R and RStudio* manual plus the one-page list of commands. Add-on routines (from the `mosaic` package) will be used extensively. There will be a quiz on R early in the semester (worry not: I'll provide a detailed study guide).

As mentioned earlier, it will be important to *bring your laptop to class*, to be able to follow along with some of the examples that I run and work through in class. Much of this work will be done in pairs (but we need to ensure that there are sufficient machines). There are several desktop computers available in the classroom. Please let me know if this presents any issues.

You will need to bring a scientific calculator to each midterm exam. This may not be shared a calculator with another student, or use your phone (which must be stored in your bag). You do not need a graphing calculator: any simple scientific calculator will suffice.

Writing Your ability to communicate results, which may be technical in nature, to your audience, which is likely to be non-technical, is critical to your success as a data analyst. The assignments in this class will place an emphasis on the clarity of your writing.

Extra Help In addition to my office hours (which I encourage you to attend), or emails to me, other resources are available. There is drop-in student help from Sunday–Thursday evenings, 7-9pm in room 300B in the Science Library. the Moss Quantitative Center in Merrill has drop in hours for statistics help. Your fellow students are also an excellent source for explanations, tips, etc.

Assignments

Homework Homework is the most effective way to reinforce concepts learned in class. There will be regular homework assignments. Often, questions will relate to material in the reading that will be covered in class. Some assignments will be submitted online, and some in class. Homework is due at the *start* of class, and unless indicated otherwise, will be accepted with a 25% penalty if turned in within 48 hours (and no credit otherwise).

Final Project Each group of three students will complete a research project during the term, and will present your results in a final report and oral presentation. We'll talk a lot more about the project as the semester proceeds. The project constitutes 20% of your final grade.

A major deliverable in this class will be to undertake (with a group) a statistical investigation on a question of interest to you. Rather than collect primary data, you will use data available on the Internet or from faculty research. You will prepare a project proposal describing your study and obtain approval from me before you begin the investigation. During the last week of class, you (and your group) will give a 10 minute oral presentation of your study. The project will give you experience planning a statistical study, acquiring data, creating and testing a linear model, and

writing a technical report. We will spend time in class looking at what data is available on the web and about writing a project proposal.

Exams There will be three exams and occasional quizzes to be taken during class, which will be worth 50% of your grade (on Wednesday, October 1st, Wednesday, October 29th, and Wednesday, December 3rd). All exams are closed book, while some may include an open-notes take home component. You may bring a calculator and a specified number of pages of paper with notes on both sides to the exams (these may be turned in with each exam).

Extra Credit Extra credit is available in several ways: attending an out-of-class lecture (as will be announced) and writing a short review of it; pointing out a substantial mistake in the book, a homework exercise or exam solution; drawing my attention to an interesting data set or news article; etc. The extra credit is applied when a student is near the boundary of a letter grade.

Grading When grading your written work, I am looking for solutions that are technically correct and reasoning that is clearly explained. Numerically correct answers, alone, are not sufficient on homework, tests or quizzes. Neatness and organization are valued, with brief, clear answers that explain your thinking. If I cannot read or follow your work, I cannot give you full credit for it. Your grade for this course will be a weighted average of several components:

Project & Presentation	20%
Homework	20%
Exams and Quizzes	50%
Participation & Presentations	10%

In order to extract maximum information from assessment, I do not work with a system where 90% means A, 80% B, and so on. I use the following approach to grading:

- A: all assignments completed, active participation in class, extraordinary and brilliant work
- B: all assignments completed, active participation in class, superior work
- C: almost all assignments completed, average work
- D: missing assignments, quality of work is less than satisfactory, attendance and participation unsatisfactory
- E: missing assignments, poor work, attendance and participation unsatisfactory

Schedule

Tentative Schedule The following outline lists each class date and gives the topic that will be discussed in that class. The reading assignment from the textbook is also given for each class date (unless indicated otherwise, these are in the Intro Stats text, the FC refers to the *Five College Guide to R*). Please complete the reading assignment *before* coming to class so that you can participate fully in the discussion. I reserve the right to revise this schedule – updates will be posted on Moodle. Last Modified: September 4, 2014

Date	Day	Topic	Reading/Due
Sept 3	Wed	Administrivia/Motivation	Syllabus
Sept 5	Fri	Introduction to data & R	1.1-1.4, FC1, and FC2
Sept 8	Mon	Introduction to data + probability	1.5-1.7, A.1 (first half), FC3
Sept 10	Wed	Introduction to probability	A.1 (remainder), FC4
Sept 12	Fri	Introduction to regression	5.1, FC5
Sept 15	Mon	Introduction to regression	5.2
Sept 17	Wed	Introduction to regression	5.3
Sept 19	Fri	Introduction to regression	
Sept 22	Mon	Multiple regression	6.1, FC9
Sept 24	Wed	Multiple regression	
Sept 26	Fri	Multiple regression	6.3
Sept 29	Mon	Conditional probability	A.2
Oct 1	Wed	In-class exam 1	
Oct 3	Fri	Foundations for inference	2.1, 2.2, FC6
Oct 6	Mon	Foundations for inference	2.3, 2.4
Oct 8	Wed	Random variables	A.3, FC10
Oct 10	Fri	Foundations for inference	2.5, 2.6
Oct 13	Mon	Fall Break	
Oct 15	Wed	Foundations for inference	2.7, 2.8
Oct 17	Fri	No class	
Oct 20	Mon	Inference for categorical data	3.1
Oct 22	Wed	Inference for categorical data	3.2, 3.3, 3.4
Oct 24	Fri	Data management	FC12
Oct 27	Mon	Inference for numerical data	4.1, 4.2
Oct 29	Wed	In-class exam 2	
Oct 31	Fri	Inference for numerical data	4.3
Nov 3	Mon	Inference for numerical data	4.4
Nov 5	Wed	Inference for numerical data	4.5
Nov 7	Fri	Inference for regression	5.4, FC7
Nov 10	Mon	Inference for regression	
Nov 12	Wed	Inference for multiple regression	6.2
Nov 14	Fri	No class	
Nov 17	Mon	Inference for multiple regression	
Nov 19	Wed	Prediction	6.4, FC8
Nov 21	Fri	Prediction	
Nov 24-28		Thanksgiving	
Dec 1	Mon	Review and closing thoughts	
Dec 3	Wed	In-class exam 3	
Dec 5	Fri	Project time	
Dec 8	Mon	Presentations	
Dec 10	Wed	Presentations	
Dec 11	Thu	Project deliverables	