SDS 325H: Honors Statistics

Course syllabus

Spring 2015

Course overview

Course Details

Time: M W 11:00AM–12:30PM
Place: FAC 101B

Instructor: James Scott
email: james.scott@mccombs.utexas.edu
Office hours: W 3:30–5:30 PM
Office: CBA 6.478

`http://jgscott.github.com/SDS325H_Spring2015/`

No undergraduate education in science is complete without thorough training in statistics. This is true for at least three reasons.

(1) Throughout your career as a student and scientist, you will be called upon to make sense of data—both to understand it yourself and to communicate its message to others.

(2) Your colleagues keep score with statistics. Pick up any recent edition of *Science* or *Nature*. You will encounter few articles whose intellectual content can be fully appreciated without understanding the statistical methods used by the authors.

(3) For the rest of your adult life you will face many difficult questions—"What should I think about this policy? What should I do about this major career decision?"—where uncertainty and randomness play a major role. What policies contribute the most to creating sustained economic growth? Do charter schools work miracles for their pupils, or benefit from self selection? Should your company become an early adopter of an expensive new technology whose ultimate payoff is in doubt? To answer questions such as these, we must carefully sift through a body of evidence, hoping to tease out relationships in complex, noisy systems that don't always behave like they would if we could draw things up in a laboratory.

Together, these sum up the two different capacities—as a participant in the scientific method, and as a citizen in a democracy—in which learning statistics will enrich your life.

In this course, we will build up towards an understanding of the kind of multivariate statistical models that you are likely to encounter in reading a modern scientific journal article. We will be concerned mainly with understanding and interpretation, rather than theoretical details. This does not mean that we won't encounter some math along the way—just that the models themselves, rather than their formal properties, will be the focus.

How the course is structured

This course revolves around independent inquiry, in addition to traditional lectures. On any given day, you should expect to be called upon to present your results from a data analysis, or to enter into a discussion with me or with your classmates about important issues that we've encountered. You will cultivate two important skills: (1) building simplified models of real-world systems to make predictions, aid cause-and-effect reasoning, and guide intelligent behavior; and (2) using visual and quantitative evidence to evaluate hypotheses in loosely structured problems with no verifiably correct answer. These are skills for which there is no blueprint or flowchart, and which can be learned only by practice.

Succeeding in this course will require substantial time devoted to out-of-class preparation:

(1) *Reading.* No single textbook quite does the job for this course. Thus in lieu of an official text, you will receive readings over the course of the semester, all of which I will post online. Many of these readings will be a course packet that I have prepared myself; treat it as a free book-like reference. The rest will come from a free online book called OpenIntro: Stats (there is a link on Blackboard). This is a perfectly good book that has been used with success in other introductory statistics courses. No other book or printed course packet is required, although I am happy to suggest supplemental books if you find the course notes lacking. I assure you this won't hurt my feelings; my style just might not work for you.

(2) *Watching video tutorials.* I have taken some of the more nuts-and-bolts stuff that you'd expect to see conveyed in lectures and broken it up into shorter, more easily digestible chunks of video. These videos are available online, and can be revisited whenever they become relevant for . . .

(3) *Practice.* Weekly exercises. Expect these to stretch your capabilities. You are allowed, but not required, to work on these in groups.

Will it be worth it? Absolutely! You will learn a lot about statistics if you make the effort. But to put it as forthrightly as possible: there will be a lot of work. As a rule of thumb, you should expect to spend 3 hours per week in class; 1–3 hours per week reading or watching videos; and anywhere from 2–8 hours per week completing the homework. If this sounds incompatible with your other pursuits this semester, then this course is probably not for you.

What you will learn

The course will be organized around specific data sets, and the lessons contained therein. This doesn't necessarily break down cleanly into a "this concept this week, that concept next" outline. But here's a rough breakdown of what you should expect to learn.

Exploratory data analysis, or basic descriptive statistics. This involves two overlapping questions. First, what quantities should you compute, and what algorithms should you run, in order to get a preliminary sense of the underlying structure in your data? We'll learn some common exploratory techniques, along with their shortcomings.

Second, how should you visually display the information that arises in your scientific work? Good statistical plotting is a lot like good writing. It is a skill borne of practice and experience, it requires many do-overs, and it gives the (erroneous) impression that its construction was effortless.

Key terms and concepts: boxplot, histogram, quantiles, mean, median, mode, standard deviation, robustness, scatter plots, association and causation, summarizing quantitative data, summarizing ordinal or categorical data

Basic probability: Some important topics here include:

- Bayes' rule and conditional probability
- Basic material on random variables: discrete versus continuous; densities and CDFs; independence; joint, conditional, and marginal distributions; covariance, correlation
- Distributions: binomial, Poisson, normal, chi-squared, t, F
- Sampling distributions, the law of large numbers and the central limit theorem
- Monte Carlo simulation

Classical inference: We will discuss the four inferential frameworks that are widely used in practice: maximum likelihood, the method of moments, classical decision theory, and (briefly) Bayesian inference. We will also focus on two important, and widely misunderstood, ideas: confidence intervals and classical significance tests. You will, of course, learn the "standard" battery of tests involving means, proportions, and contingency tables. But I care less that you remember the specific mechanical details of, for example, a Kruskal–Wallis test, and far more that you emerge from the course able to articulate the correct interpretation of *all* significance tests, which really do have a unifying (if somewhat convoluted) logic. You will also come to appreciate difference between the Fisherian and Neyman–Pearson views of inference, which are historically opposed but systematically conflated in most modern textbooks. If you come away from the course with a well-earned sense of skepticism regarding how p-values are reported in journal articles, I'll be thrilled.

Key terms and concepts: hypothesis testing, z test, t test, χ^2-test, difference of means, difference of proportions, tests of independence, confidence intervals, permutation tests and other nonparametric tests, likelihood, method of moments

Modeling: The ultimate goal of the course is to learn how to use statistical models to extract insights from data. We will start with a few simple, one-variable parametric models for discrete and continuous outcomes. But we will cover as many multivariate models as time allows. This will certainly including linear regression and logistic regression. I hope we can also cover some from the following list: survival models; models for multiple testing; additive models; ordinal and multinomial models; longitudinal models; and hierarchical models.

Key terms and concepts: simple linear regression, multiple linear regression, least squares, F test, ANOVA, logistic regression, multivariate models

The outline on the following page is subject to review if we need to slow down or speed up, but it should give you a more detailed idea of how the course will proceed. In particular, the advanced topics in Week 15 can easily be cut if we need to spend more time on earlier material.

Prerequisites and software

I assume that everyone in the course will have had calculus. But you won't need it at a particularly advanced level. Here's a simple diagnostic: can you compute the instantanteous rate of change of the function $f(x) = \ln x$, as a function of x? If you can, then you know enough calculus to succeed in this course. If you needed to look up a rule or two in a textbook or online resource, but you get the gist of it, then you're probably OK. If you don't understand what the question is asking, then you need a math refresher before you take this course.

The course will use a free, open-source statistical computing language called R. More specifically, we will use a third-party front-end

Download R from `www.r-project.org/`

Week	Topic	Day by day
1–2	Explanation and evidence	An introduction to data-based reasoning
	Exploring multivariate data	Variation within and between categories
		Fitting simple parametric curves by linear least squares
3	Predictable and unpredictable variation	Sums of squares in group-wise models and regression
		Dummy variables and interactions
4–5	Quantifying uncertainty	Some probability basics
		Sampling distributions
		Bootstrapping
		Permutation tests
6–7	Parametric inference	Distributions and likelihood
		Estimation and prediction in parametric models
		Testing in parametric models
		Power
8	Midterm week	Catch-up day
		Midterm exam
9–10	Multiple regression	Fun with charismatic megafauna: a first look at multiple regression
		Statistical adjustment
		The analysis of variance, revisited
		Estimation, prediction, and testing in the multiple regression model
11	Modeling discrete outcomes	Binary data
		Count data
12	Causality	Basic theory of causal inference
		Instruments, mechanisms, and *d*-separation
13–15	Further topics	Some of: hierarchical models; curve fitting; Bayesian methods; survival models; multiple testing

interface to R called RStudio. It is freely available at `http://http://www.rstudio.com`. You will need to download both R and RStudio; the latter requires the former to work. If you've ever used a stats package such as Stata or SAS, or a scripting language such as Matlab or Python, you'll find this easy. If not, that's OK, too—I will teach R as if you've never seen it, or a similar software package, before.

I personally use R for about 90% of the statistical computing needs that arise in my research. It can do, or can be made to do, most data analyses you'll ever need to conduct. Google and Facebook use it to benchmark changes to their algorithms; the New York Times uses it to produce its graphics; Barack Obama's election campaign used it for voter analytics. It is the *lingua franca* of modern statistics. Learning it will serve you well.

Exams and grading

Grades will be determined by one in-class midterm exam; one open-book, take-home final exam; and regular homework assignments.

Grading
Homework: 50%
Mid-term: 20%
Final: 30%

Homework will count for 50% of your final grade. To receive full credit on an assignment, you must show your work and/or explain your reasoning. The assignments will typically be posted on Monday and due on the Monday of the following week. All homework must be turned in at the beginning of class on the day it is due. No late homework will be accepted. But your lowest homework grade of the semester will be dropped, thereby allowing for the occassional illness or other difficulty with finishing the assignments.

Mid-term Week before Spring Break

The mid-term is worth 20% of your final grade, and will take place during the last week of class before Spring Break. You will be allowed to bring a calculator, but it is not necessary to have one. The exam will be graded such that, if you set up all calculations in the appropriate way, you will get full credit even if it is not possible to get the final answer without a calculator.

If you must miss the exam for the observance of a religious holy day, inform me as far in advance of the day as possible, so that alternative arrangements can be made in conjunction with the Dean and the relevant university offices. If you miss the mid-term for any other reason—including illness or travel—then you must inform me in advance, and I will allow you to count your final exam grade as your midterm grade. This option cannot be exercised retroactively, and it is not available unless you inform me in advance. (If you are sick, then an e-mail on the morning of the mid-term will be fine.)

The final is a take-home exam, and will count for 30% of your grade. The final will be available online at 5:00 P.M. on the last class day of the semester, and is due one week later at 5:00 P.M. There will be extra office hours during the final week of class to answer questions in advance of the final.

Final
Distributed: Last day of class
Due: One week later

Re-grade requests

On occasion you may notice a simple clerical error in the recording of a grade, which I am happy to correct without hassle. Other regrading requests must be submitted in writing within 7 days of the marked paper being returned. Keep in mind that the entire paper will then be subject to re-grading, and that your grade may go up or down as a result.

Attendance

I have gone to a lot of effort to provide you with course notes and video tutorials that can help you learn. These notes and videos are intended to supplement, not replace, your attendance in class. If you stay home and try to learn from the supplemental material alone, you are unlikely to do well on the exams, and are even more unlikely to retain much of what you have learned. Passivity is the enemy of learning.

Beyond the obvious correlation between coming to class and overall course performance, attendance does not play a role in course grading.

Curving grades

The raw percentage scores to the right will guarantee you *at least* the corresponding grade.

I reserve the right to curve grades up. But I will never curve them down. That means these grades are a floor, not a ceiling, on the final grade that someone with the corresponding raw score would receive. The precise details of any curve are at my sole discretion, and if I should choose to use a curve, I will detail the cutoffs used when course grades are submitted.

Percentage	Grade
93–100	A
90–92	A-
87–89	B+
83–86	B
80–82	B-
70–79	C
60–69	D

Other course policies

Classroom etiquette

You are expected to participate in class; keep your browser windows free of distractions on those "hands-on" day where I ask you to look at data and run models in class; and to turn off your phones and gizmos. I also ask that you arrive on time to class, since late arrivals disrupt things for all other students. In turn, I will make sure we finish on time so that students may reach their next lectures/hot dates.

Cheating, plagiarism, and such

Acts of academic dishonesty are ethically wrong; they harm the reputation of the school and demean the honest efforts of the majority of students. You know it; I know it; and if I catch you, I'll bring the thunder and lightning. Additionally, you should consider three things:

1. Cheaters are a tiny minority. The vast majority of students who preceded you did it the honest way. Follow their lead.
2. You play like you practice. The habits you form now will predict the headlines that people write about you, or your lab or company, later in life. Try Googling "Jeff Skilling" or "Fabulous Fab" if you don't believe me.
3. If you cheat, you're playing with fire. The minimum penalty will be a zero for that assignment or exam. You also risk failing the course and being dismissed from the University.

My first hit for "Fabulous Fab" is the *New York Daily News* from 27 April 2010, which wrote: "Fabrice Tourre, who calls himself Fabulous Fab, is not so much. Actually, the 31-year-old Frenchman of the racy e-mails came across like a weenie when he appeared before a Senate subcommittee to be grilled about Goldman Sachs' role in a deal the SEC says wasn't kosher." Cheat at your own risk, weenie.

Now for the usual boilerplate. The responsibilities of both students and faculty with regard to scholastic dishonesty are described in detail in the Policy Statement on Academic Dishonesty for the College of Natural Sciences. By enrolling in this class, you have agreed to observe all of the student responsibilities described in that document. By teaching this course, I have agreed to observe all of the faculty responsibilities described in that document.

Students with disabilities

The University of Texas at Austin provides upon request appropriate academic accommodations for qualified students with disabilities. Services for Students with Disabilities (SSD) is housed in the Office of the Dean of Students, located on the fourth floor of the Student Services Building. Information on how to register, downloadable forms, including guidelines for documentation, accommodation request letters, and releases of information are available online at `deanofstudents.utexas.edu/ssd/index.php`. For more information, contact the Office of the Dean of Students at 471-6259, or 471-4641 TTY.

Student privacy

I am legally barred from discussing your course performance with anyone other than you and anyone that you explicitly designate. That includes your parents.

Second, a note on Canvas. Canvas is a password-protected web site, and is created automatically for all accredited courses taught at The University. I won't use Canvas except to e-mail the class. But you're free to use it, and I'm therefore required to state the following. Canvas include a class e-mail roster. Students who do not want their names included in such an electronic class rosters must restrict their directory information in the Office of the Registrar, Main Building, Room 1. For information on restricting directory information, see `www.utexas.edu/student/registrar/catalogs/gi02-03/app/appc09.html`.