STA 371H: Statistics and Modeling (Honors)

Course Syllabus

Spring 2015

Course overview

The promise of data—whether big, small, or somewhere in between—is the knowledge it can bring us. Used correctly, data can help answer hard questions in the face of uncertainty and randomness. 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? We usually cannot study questions like these using controlled experiments. Rather, we must carefully sift through a body of evidence, hoping to tease out relationships in complex, noisy systems that don't behave like an experiment drawn up in a laboratory.

In this course, you will learn to use the language of probability to study problems such as these in a formal quantitative way. My goal is to help you cultivate two important skills:

1. Building simplified models of real-world systems to aid cause-and-effect reasoning and guide intelligent behavior.

2. Using visual and quantitative evidence to evaluate hypotheses in loosely structured problems with no verifiably correct answer.

Throughout the course, I will emphasize the analysis of real data, and will provide examples from finance, politics, sports, marketing, economics, and science to help illustrate the concepts you'll be learning. By semester's end, you will have learned some lessons that will serve you well throughout the rest of your life, both as a producer and as a consumer of data.

The methodological focus of this course is on building models that aid us in analyzing data and making smart decisions. 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 mathematical properties, will be the focus.

Course Details

Section: 04380
Time: M W 9:30 AM –11:00 AM
Place: CBA 5.330

Instructor: James Scott
(james.scott@mccombs.utexas.edu)
Office hours: M W 3:30–4:30 PM
Office: CBA 6.478

Teaching Assistant: David Puelz
(david.puelz@phd.mccombs.utexas.edu)
Office hours: TBA
Office: TBA

`jgscott.github.io/STA371H_Spring2015`

Materials

Readings

You do not need to purchase a textbook: I hate the idea of assigning an expensive shelf decoration that is, at best, a halfway match with the goals of the course. Instead, we will have three main references, all free:

1. An online course packet, roughly 200 pages in length. I wrote the packet myself, so it is carefully integrated with the rest of the course material. This is in a state of continual editing, so I will post chapters to the course website as they become relevant over the semester.

2. The first five chapters of *Statistical Modeling: A Fresh Approach*, by Daniel Kaplan. These five chapters are available for free; I have provided links on the course website.

3. *Data Analysis for Politics and Policy*, by Edward Tufte. Out of copyright, with a PDF copy freely available on the course website.

In addition to these resources, there will also be shorter readings, all of which will be linked from the course website when the time comes.

Software

As for software, we will use an open-source statistical programming environment called R (www.r-project.org) for almost of what we need to do. R is freely available for Mac, Windows, and Linux. It's the real deal—no silly Excel limitations for us. I want you to come away from the course with a legitimate, industrial-strength platform—Google and Facebook, for example, both use R to do a large amount of their data analysis. Specifically, we will use a graphical front-end to R called RStudio: www.rstudio.org. You will almost surely like this better than the standard Mac or Windows interface.

Strictly speaking, R is recommended, not required: if you're an Excel whiz, you might be able to make it work for you for the entire course. If you are tempted to take this route, though, you should consider some things: 1) R is easy to learn, and the course will hold your hand a lot in the beginning; 2) R is drastically more powerful, efficient, and versatile than Excel for data analysis; 3) the plug-in you need to do real statistical modeling in Excel isn't available for Macs; 4) if you Google "R for business" you'll discover that knowing R is a competitive advantage for a lot of jobs these days; and 5) unless specifically stated otherwise for a given topic, neither I nor the TAs will answer questions about Excel. But you are 100% free to use any software package you like, including Matlab, STATA, SAS, Minitab, and so forth. Heck, you can even write your own stuff in Python, if that's your thing.

Course outline

The following day-by-day outline 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. There are 24 topics that must fit in 28 class days, not counting holidays and exam days. This gives us some leeway for extra days devoted to examples, review sessions, or simply slowing down. You'll also see that the topics are split into two overall sections: statistical modeling (roughly the first half of the course) and probability/decision-making (roughly the second half). Another important "major idea" of the course—Monte Carlo simulation—comes up in both halves.

Important dates:
March 11: In-class midterm exam
March 16–20: Spring Break (no class)
April 3: Projects due
May 6: Last class meeting

Statistical modeling

1. *Explanation and evidence.*
Topics: Confounding; dependence graphs; blinding and placebos; longitudinal/cross-sectional studies; natural experiments.
Read: Chapter 1 of the course packet; Kaplan, Chapter 1.

2. *Introduction to multivariate data.*
Topics: Continuous and grouping variables. Contingency tables. Simple summaries and graphics: histogram, boxplot/dotplot, scatter plot. Variation of a typical case; variation between and within groups. Simple group-wise models. Model/fitted values and residuals.
Read: Course packet, Chapter 2 (pages 25–36); Kaplan, Chapters 2–3.

3. *Two quantities varying together.*
Topics: Fitting straight lines via ordinary least squares. Interpreting the model parameters. Basic stories one can tell with a statistical model: prediction, summary, adjustment, reducing uncertainty.
Read: Course packet, Chapter 2 (pages 37–49); Tufte, Chapter 1.

4. *Simple nonlinear models.*
Topics: Residual diagnostics. Transforming and combining variables. Fitting nonlinear models by least squares using transformations.
Read: Course packet, Chapter 2 (pages 60–70); Tufte, Ex. 6 on page 108.

5. Predictable and unpredictable variation.
Topics: Quantiles and coverage intervals. Naïve prediction intervals. The decomposition of variance and R^2. Sample correlation.
Read: Course packet, Chapter 3; Kaplan, Chapter 4.

6. Parameter uncertainty: a first look.
Topics: The fundamental frequentist thought experiment. Sampling distributions. Bootstrapping, standard errors, confidence intervals.
Read: Course packet, Chapter 4 (pages 79–91); Kaplan, Ch. 5.

7. The Gaussian regression model.
Topics: The normal distribution; the signal-to-noise ratio and the t-statistic; inference and prediction.
Read: Course packet, Chapter 4 (pages 92–115). Tufte, Ch. 3.

8. Continuous and grouping variables together.
Topics: Aggregation paradoxes. Dummy variables and interactions.
Read: Course packet, Chapter 5.

9. Introduction to multiple regression.
Topics: Linear models with more than one predictor. Interpreting partial slopes: three tries. Statistical adjustment.
Read: Course packet, Chapter 6.

10. Multiple regression: further details.
Topics: Collinearity. Quantifying the relative importance of predictors.
Read: Tufte, Ch. 4.

11. Testing hypotheses: the basics.
Topics: Permutation tests in simple contingency tables. Neyman–Pearson versus Fisherian notions; α-level versus p-value.
Read: Course packet, Chapter 7 (first 12 pages).

12. Testing hypotheses: more complex cases.
Topics: Hypothesis testing in multivariate statistical models.
Read: Course packet, Chapter 7 (remainder).

13. Model checking and model selection.
Topics: Fine-tuning a statistical model. Choosing predictors to include. Occam's Razor and the fit/simplicity tradeoff.

14. Regression for discrete outcomes.
Topics: Link functions. Logistic regression (definitely); Poisson regression (maybe).
Read: Course packet, Chapter 8.

15. Data collected in time.
Topics: Forecasting; trends; seasonality; using lagged predictors.
Read: Course packet, Chapter 9.

Probability and decision-making

16. How not to make a decision.
Topics: Behavioral quirks, biases, and heuristics.
Read: Section 3 of Barberis and Thaler.

17. Introduction to probability.
Topics: Probability and odds; coherence; probability distributions.
Read: Lecture notes on probability (first 10 pages).

18. Conditional probability.
Topics: Conditioning and Bayes' rule.
Read: Lecture notes on probability (remainder).

19. Variability and the $\sqrt{n}$ law.
Topics: Mean, variance, and the law of large numbers. The fundamental relationship between sample size and statistical variability.
Read: Harold Wainer, "The Most Dangerous Equation."

20. Expected value.
Topics: Expected value; scenario analysis; decision trees.
Read: TBA

21. Expected utility.
Topics: St. Petersburg paradox. Expected utility; the von Neumann–Morgenstern axioms.
Read: Lecture notes on utility theory.

22. Simulation.
Topics: Using Monte Carlo simulation to estimate probability distributions for complex systems.
Read: Phil Laak on Kelly's criterion (links to be posted).

23. Correlated random variables.
Topics: Joint distributions; covariance and correlation.
Read: Lecture notes on random variables.

24. Investments, insurance, and gambling.
Topics: Applying modeling principles to investment problems.
Read: TBA

How the course is structured

On a day to day basis, this course revolves around independent inquiry, in addition to traditional lectures. The focus of class time is on building your capacity to think about open-ended problems. On any given day, there will be a mix of lecture, discussion, and hands-on modeling.

The upshot of all this? You will end up learning many important skills outside of class. This learning will take (at least) three forms:

Reading. Mostly this will be out of the online course packet, but there will also be supplemental readings from other sources.

Watching videos. Some of these are about important concepts, and some are about software.

Practice. There will be weekly problem sets, consisting mostly of open-ended modeling problems. You are encouraged, but not required, to work on these in groups of four people or fewer.

You'll find links to all the relevant material through the class website.

As you might imagine from this description, succeeding in this course will require substantial time devoted to out-of-class preparation. Will it be worth it? Absolutely! You will learn a lot about statistics if you make the effort. As a rule of thumb, you should expect to spend 3 hours per week in class; 1–3 hours per week reading and watching videos; and anywhere from 3-6 hours per week completing the exercises. As with all college classes, in some weeks there will be less than this, and in some weeks there will be more.

Prerequisites

The formal university prerequisites for this course are: Business Administration 324 or 324H; Management Information Systems 301 or 310; Mathematics 408D, 408L, or 408M; and Statistics 309 or 309H. Note: the calculus prerequisite is not there for show. In particular, I expect you to remember derivatives and basic material on logarithms.

Exams and grading

Grades will be determined by a midterm (25%), final exam (30%), project (20%), homework assignments (20% total), and scribing for one class day (5%).

Grading

Midterm:	25%
Final exam:	30%
Homework:	20%
Project:	20%
Scribing:	5%

Homework and project

Homework is assigned weekly, and will count for 20% of your final grade. All homework must be turned in at the beginning of class on the day it is due. You are allowed (but not required) to work on homework in groups of 4 people or fewer. If you work in a group, turn in a single write-up, with all your names on it.

Homework is graded on a 5-point scale. The default score for a good job is 4, with 5 reserved for exceptional efforts; grades will be curved accordingly at the end. No late homework will be accepted, but everybody gets one drop. Homework will be accepted only in hard-copy form. No electronic copies will be accepted without prior permission received before the homework is due. I will typically grant such permission in extenuating circumstances.

Homework: assigned weekly and due in hard-copy form at the beginning of class on the due date.

There is one course project, which you may complete in groups of four people or less if you wish. (Groups are optional.) This will involve getting your own data set on a question that interests you, running an appropriate statistical analysis, and writing up your conclusions. The details will be discussed in class, and posted on the course website. This project is worth 20% of your grade, and is due by 5 PM on Friday, April 3, 2015. As with the homework: no electronic copies will be accepted without prior permission.

Project 1: due Friday, April 3, 2015.

Exams

The in-class midterm is worth 25% of your final grade, and will take place on the last day of class before Spring Break. You will need a blue book and an ink pen, and nothing else.

Mid-term: in class on the last class day before spring break

The final exam is worth 30% of your grade and will take place during the usual University exam period in early May. As with the midterm, you will need a blue book, an ink pen, and nothing else. I will announce the time slot as soon as it is available from the University registrar.

Final
When: TBA, during offical exam period
Where: TBA

Scribing

Finally, 5% of your grade comes from scribing. This is easy: all you have to do is sign up for one day where you will provide the definitive record of what we did in class. That means taking comprehensive notes, prettying them up afterwards, and then posting them on the class website. This will give your classmates a completely independent (from me) answer to the question, “What did we do in class last week?” I will provide the sign-up sheet at the beginning of the semester.

Missed exams

You will not be allowed a make-up for a missed exam without a documented and verifiable medical excuse, or documentation that a family emergency prevented you from attending. The only documentation I will accept for this purpose is an electronic or written letter from Student Emergency Services in the Office of the Dean of Students notifying me of your absence. The Dean of Students will, in turn, require supporting documentation from you (e.g. a doctor's note or letter from primary care provider) in order to verify your illness, injury, or emergency. While this policy may seem strict, it is the only way we can be fair to everyone.

If you will be out of town representing the University on an academic, athletic, or student-organization trip, you must speak with me and provide me with appropriate documentation at least 2 weeks in advance. I will be glad to make arrangements for you to take the exam before you leave for your event.

Finally, if you must miss an exam for the observance of a religious holy day, inform me at least 2 weeks before the exam, so that alternative arrangements can be made in conjunction with the Dean and the relevant university offices.

If you miss an exam for any other reason—including personal travel or family holiday—you will get a zero.

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

There is no explicit attendance component to your course grade. Having said that, it is hard to do well on the rest of the course assignments without coming to class. I encourage you to form good habits.

Curving grades

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

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

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.

Other course details

Quantitative reasoning flag

This course carries the Quantitative Reasoning flag. Quantitative Reasoning courses are designed to equip you with skills that are necessary for understanding the types of quantitative arguments you will regularly encounter in your adult and professional life. For more details, see `www.utexas.edu/ugs/core/flags/quantitative-reasoning`.

Classroom etiquette

You are expected to participate in class; close and put away your laptops, unless it's a "hands-on" day where I ask you to look at data and run models in class; and to turn off your phones, iPods, and other cool 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 no excuses will be accepted. 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 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."

The bottom line when it comes to cheating is: just don't do it. You might fool me, if you're very lucky and very unscrupulous. But you are highly unlikely to fool the McKinsey interviewer you were hoping to impress with your knowledge of statistics. And you may find that the job market is far more ruthless than university judicial boards.

Now for the usual boilerplate.

The McCombs School of Business has no tolerance for acts of scholastic dishonesty. The responsibilities of both students and faculty with regard to scholastic dishonesty are described in detail in the BBA Program's Statement on Scholastic Dishonesty.[1] By teaching this course, I have agreed to observe all faculty responsibilities described in that document. By enrolling in this class, you have agreed to observe all student responsibilities described in that document. If the application of the Statement on Scholastic Dishonesty to this class or its assignments is unclear in any way, it is your responsibility to ask me for clarification. Students who violate University rules on scholastic dishonesty are subject to disciplinary penalties, including the possibility of failure in the course and/or dismissal from the University. Since dishonesty harms the individual, all students, the integrity of the University, and the value of our academic brand, policies on scholastic dishonesty will be strictly enforced. You should refer to the Student Judicial Services website[2] to access the official University policies and procedures on scholastic dishonesty as well as further elaboration on what constitutes scholastic dishonesty.

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

First of all, you should know that I am legally barred from discussing your course performance with anyone other than you and anyone that you explicitly designate. That includes your parents.

[1] http://www.mccombs.utexas.edu/BBA/Code-of-Ethics.aspx

[2] http://deanofstudents.utexas.edu/sjs/

Second, a note on Canvas. Canvas is a password-protected web site, and is created automatically for all accredited courses taught at The University. Site activities could include exchanging e-mail, engaging in class discussions and chats, and exchanging files. In addition, 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`.

Campus safety

Please note the following recommendations regarding emergency evacuation from the Office of Campus Safety and Security, 512-471-5767, http://www.utexas.edu/safety.

- Occupants of buildings on The University of Texas at Austin campus are required to evacuate buildings when a fire alarm is activated. Alarm activation or announcement requires exiting and assembling outside.
- Familiarize yourself with all exit doors of each classroom and building you may occupy. Remember that the nearest exit door may not be the one you used when entering the building.
- Students requiring assistance in evacuation should inform the instructor in writing during the first week of class.
- In the event of an evacuation, follow the instruction of faculty or class instructors.
- Do not re-enter a building unless given instructions by the following: Austin Fire Department, The University of Texas at Austin Police Department, or Fire Prevention Services office.
- Behavior Concerns Advice Line (BCAL): 512-232-5050
- Further information regarding emergency evacuation routes and emergency procedures can be found at: http://www.utexas.edu/emergency.