DSC 381: Probability and Simulation Based Inference for Data Science

Course overview

Students will learn the foundations of probability and inferential statistics, through both basic theoretical underpinnings and simulation examples. The concepts and skills that are taught in this course will provide students with the tools to understand many different models and approaches used in statistical learning and inference. Topics include basic probability calculus, random variables, probability functions and densities, useful inequalities, sampling distributions of statistics and confidence intervals for parameters, hypothesis testing, properties of estimators, maximum likelihood estimation, and exponential families. Students will learn the basic principles of statistical inference not through formulas and memorization, but by the basic theoretical underpinnings of the process through simulation and discussion of examples.

Course modality

All the lectures, assignments, and work in the course is served out in edX (edX.org). All assignments are to be submitted in edX, with Canvas as a backup. (See course FAQs.) https://canvas.utexas.edu/

Textbooks

For the probability part of the course we use the textbook
M. Mitzenmacher and E. Upfal, "Probability and Computing", Cambridge University Press (CUP), ISBN 9781107154889. The e-book sells for $56, and hardcopy for $70.
This is a required textbook. The 1st edition is also fine (of the material covered in this class, it is missing only §9).

For the statistics part, **you are expected to use some material from an applied statistics textbook**. The recommended textbook is Lock, Lock, et. al., *Statistics: Unlocking the Power of Data*, Wiley. (3rd ed. 2020, 2nd ed. 2016, 1st ed. 2012) The book is not required. Any edition is acceptable. No required homework is assigned from it. Online access from the publisher to the current edition, with additional helpful videos and complete solutions for all odd-numbered problems, can be purchased for about $70 (with a grace period of 14 days before paying.) See also the handout "Prerequisite Information", linked in edX (under "Week 0", "Handouts").

Instructional Staff:

Instructor: Mary Parker (she/her) (Call me any of these: Professor Parker, Dr. Parker, Mary)
Learning Facilitators and Teaching Assistant (TA): (See posting in edX)
Email: Contact the instructional staff through the course email address: onlineprobability@austin.utexas.edu

Communication and The Discussion Board:

This course is quite large, with several hundred students. We will be using the Discussion Board heavily. Please ask questions about course content and syllabus through the Discussion Board only.

Messages about content and syllabus questions will not be answered by email because those are useful to many students in the course.

The TAs and the instructor will monitor the Discussion Board and participate from time to time. We strongly urge you to do your best to answer each other's (and your own, if you find an answer)

questions. Trying to explain a concept you have learned is the best way to understand it more deeply. So, you will help yourselves, and also your classmates!

Office Hours:

Office hours: on Zoom as shown in Week 0 under "Course Communication."

Prerequisites

Calculus and knowledge of basic applied statistics methods. Students who have not had a fully applied statistics (or "elementary statistics) course will also be expected to do work from the recommended elementary statistics book during the course.

Computing

Some homework problems will need the use of a statistical programming language like R. Example code will be given in R. There are many excellent tutorials available on-line, for example the (free) edX course "Data Science: R Basics", by Rafael Irizary. If you do not yet know and use R, please reference this or a similar tutorial to acquire a basic working knowledge. However, R is not strictly required. You could use any other programming language with good statistics support. (Spreadsheets do not meet this criterion.)

In the statistics part of the course, examples and solution keys are prepared using StatKey, http://www.lock5stat.com/StatKey/. which provides basic statistical inference using randomization methods. It accompanies the recommended statistics textbook (Lock et al.) and is freely available at on the web. Instructions for its use are provided.

Grading Policy

The final grade for your UT transcript is a weighted average of your homework, quiz, and exam scores.
Weights: Homework: 29%, Quizzes: 30%, Exam 1: 20%, Exam 2: 20%, Other: 1%.
UT final grades are in letters: Grades: A: 93-100, A-: 90-92, B+: 87-89, B: 83-86, B-: 80-82, C+: 77-79, C: 73-76, C-: 70-72, D+: 67-69, D: 63-66, D-: 60-62, F: 0-59
Decimal averages are rounded to the nearest whole number percentage(with exactly 0.5 being rounded up.)
EdX certificates are awarded to students who meet the requirement stated in edX. The certificate does not display a grade.

The various assignments have different point values, and each assignment grade is converted to a percentage before any summaries are computed. In computing grades, edX converts each assignment grade to a percentage, keeping several decimal places in their system, but reporting rounded percentages. Your Progress Report shows a grade for each question on each assignment as well as summaries.

The "Other" category is for Orientation assignments and "reminder assignments" of various things from the orientation and readings which some students in the past skipped over. (I have to have those graded in order to track whether students look at them.) We do not promise that everything of which you might need to be reminded is included. Pay careful attention to all of the course materials.

Unfortunately, the grade report you will see in edX is set up in a way most students find inconvenient. It starts with a grade of 0 and you earn additions as you go through the course. Most students prefer to set up a spreadsheet for themselves and use their "Progress Report" in edX to update their own

spreadsheet each week. That is fairly easy to do and allows you to compute your grade as you go along.

Homework: Each weekly homework assignment counts equally in the final homework grade, regardless of the number of problems or points in it. The central goal of the homework assignments is to ensure that students are practicing using the main concepts presented in the class. Substantial effort in practicing these is needed in order to be able to use these concepts in later courses, so these are an important portion of the course grade. Homework problems will be posted approximately weekly (as we complete a major topic or chapter in the lecture.)

Quizzes: Each quiz assignment counts equally in the final quiz grade, regardless of the number of problems or points in it. Quizzes are short (usually 4 questions, each similar to shorter homework problems or a piece of a longer homework problem) so each question counts a high percentage of the grade. They are timed assignments (30 minutes.) Be sure to fully understand how to work all the homework problems and similar problems before taking the quiz. Quizzes are available in a window of seven days.

Some assignments have an overall problem which is broken down into parts. This allows students to earn some partial credit for correct work in the early parts of the problem even if they can't get the full problem done correctly. In this course, there is a very strong emphasis on starting the solution correctly as a crucial part of learning the material. It is also important to practice a certain type of carefulness in your work.

Exams: Exams have approximately 8 – 12 questions, which will be similar in scope and difficulty to medium-length and short homework problems. They are timed assignments (120 minutes.) Exams have a window of seven days in which to take them.

What are we testing / measuring in this course?

1. Skill in understanding and using the probability and statistics concepts covered.
2. In the homework we are also measuring how resourceful you are in finding other class members to work with in small groups. Discussion of your thoughts with others will be necessary to learn to use the material at the level expected in this graduate class. The course discussion board is NOT a small group. Very few of you will discuss enough there for it to be useful to you. It is important to form small groups of about 4 to 12 students.
3. In all assignments, be attentive to numerical details as you work through the problems. Part of your training as a data scientist is to develop skills in proofreading, etc. so that your answers can be trusted.
4. In all assignments, being careful in reading online material (including magnifying mathematical notation in your display as needed,) entering answers, and proofreading your final submission.
5. In quizzes and exams, being careful in managing your attention and your time. We strongly advise you to use a timer designate a specific amount of time at the end of each timed assignment to proofread the answers you have submitted.

Rules for Quizzes and Exams: All quizzes and exams are open book (any book, including e-book is fine), and open-notes (your course notes, lecture notes from edX). A hand-held non-graphing calculator or a spreadsheet is ok, as is R or any other statistics program. No other online resources are allowed. Also, you are not allowed to use help from anyone else (in person or online.)

Read the information near the end of the syllabus about "Sharing Materials" and "Academic Dishonesty" carefully to understand your obligations.

Discussion of how to handle difficulties in submitting your work.

The FAQ pages are an important part of this syllabus. Be sure to understand the use of Canvas in this course, as described there.

Rules for obtaining help on homework.

The FAQ pages are an important part of this syllabus.

Discussion of availability of solutions, grading, etc.

The FAQ pages are an important part of this syllabus.

Calendar and Topics

All availability and due date times are UTC 11:00 on the indicated date.

Note that the times, **for students in the United States** are a few hours after midnight thus the effective deadlines are the **end of the previous day of the week rather than the day listed here**. Use this as needed: https://www.timeanddate.com/worldclock/converter.html

Also note that the crucial times will shift when there is a local change to or from Daylight Savings Time because UTC times do not use Daylight Savings Time. Individual students are expected to keep up with any such changes, because you know when your local time shifts.

- Learning material for the week: Available Wednesday of the previous week
- Homework for the week: Available Sunday of the week. Due Wednesday of the following week
- Quizzes and Exams: Available Thursday of the week and due 7 days later
  (This is slightly modified for the assignments due in the last week of the course.)

Course Calendar and Outline of Topics

MU §x refers to chapters in the (required) textbook by Mitzenmacher and Upfal.

Before class begins: (about 3 days before the course begins)
Read through the Week 0 materials and do the Orientation Activities.

Week 1: Aug. 21 – 25, 2023
Events and probability (MU §1), lectures 1.1–1.11
Start working on some of HW 1

Week 2: Aug 28– Sep 1, 2023 (Jan 16 is a UT holiday)
Events and probability (continued), lectures 1.12-1.13;
HW 1 is assigned (due Wed of next week)
Discrete random variables (MU §2), lectures 2.1-2.7

Week 3: Sep 4 – 8, 2023
Discrete random variables (continued), lectures 2.8-2.18
Thur: Quiz 1 is assigned (covers Chapter 1 material in Weeks 1 and 2) (due Thur of next week)
HW 2 is assigned (due Wed of next week)

Week 4: Sep 11 – 15, 2023
Moments & Deviations (MU §3.1-3.3) , lectures 3.1-3.9
HW3 is assigned (due Wed of next week)
Thur: Quiz 1 is due.
Thur: Quiz 2 is assigned (covers Chapter 2 material in Weeks 2 and 3) (due Thur of next week)

Week 5: Sep 18 – 22, 2023
Introduction to Statistical Inference, lectures 4.1-4.5
HW 4 is assigned (due Wed of next week)
Thur: Quiz 2 is due

Week 6: Sep 25– 29, 2023
Cautions, lectures 5.1 and 5.2
Use of Simulation 6.1-6.2
HW 5 is assigned (due Wed of next week)
Thur: Exam 1 is assigned (covers all material through Week 5) (due Thur of next week)

Week 7: Oct 2 – 6, 2023
Moment generating functions (MU §4.1-4.2.1), lectures 7.1-7.3
Continuous random variables (MU §8.1 – 8.3.1), lectures 8.1-7
HW 6 is assigned (due Wed of next week)
Thur:Exam 1 is due (Covers all material through Week 5)
Thur: Quiz 3 is assigned (covers weeks 5 and 6) (due Thur of next week)

Week 8: Oct 9 – 13, 2023
Continuous random variables (continued), lectures 8.8-15
HW 7 is assigned (due Wed of next week)
Thur: Quiz 3 due

Week 9: Oct 16 – 20, 2023
Normal distribution (MU §9), lectures 9.1-7
HW 8 is assigned (due Wed of next week of classes)
Thur: Quiz 4 available (covers weeks 7 and 8) (due Thur of next week of classes)

Week 10: Oct 23 – 27, 2023
Inference with Simulation: Details, lectures 10.1-5
HW 9 is assigned (due Wed of next week)
Thur: Quiz 4 is due

Week 11: Oct 30 – Nov 3, 2023
Inference with Theoretical Distributions, lectures 11.1-6
HW 10 is assigned (due Wed of next week)
Thur: Quiz 5 is available (covers Weeks 5, 6, 10) (due Thur of next week)

Week 12: Nov 6 – 10, 2023
Chi-Squared and ANOVA Tests, lectures 11.7-10
HW 11 is assigned (due Wed of next week)
Thur: Quiz 5 is due

Week 13: Nov 13 – 17, 2023
Estimation, lectures 12.1-6 and Exponential Families 12.7-9
HW 12 covering Estimation 12.1-6 is assigned (due Wed of next week) (material is covered on a quiz and exam)
HW 13 covering Exponential Families 12.7-9 is assigned (due the day after Exam 2 is due) (material is **not** covered on a quiz or exam)
Thur: Quiz 6 is available (covers Weeks 5, 6, 10, 11, 12)

Week 14: Nov 27 – Dec 1, 2023 (Note the previous week is the UT Thanksgiving holiday)
No new material covered this week.
Thur: Quiz 6 is due
Fri Exam 2 is assigned and due **Friday of next week**

Week 15: Dec 4 – 8, 2023
Fri: Exam 2 is due
Sat: HW 13 is due

Ending the course: Dec 9 – 12, 2023
Sun: Deadline for email regarding grades for Exam 2 and HW 13 to be received at course email address.
Tue: Course ends.

- edX course is archived, with learning materials remaining available, but not assignments.
- Discussion Board is archived, so no new posts can be made

Program Requirements

For questions about program requirements and credit, please contact the MSDS Program Coordinator at msdsgradcoordinator@utexas.edu.

Sharing of Course Materials is Prohibited:

No materials used in this class, including, but not limited to, lecture hand-outs, videos, assessments (quizzes, exams, papers, projects, homework assignments), in-class materials, review sheets, and additional problem sets, may be shared online or with anyone outside of this class unless you have explicit, written permission. Unauthorized sharing of materials promotes cheating. It is a violation of the University's Student Honor Code and an act of academic dishonesty. We are well aware of the sites used for sharing materials, and any materials found online that are associated with you, or any suspected unauthorized sharing of materials, will be reported to Student Conduct and Academic Integrity in the Office of the Dean of Students. These reports can result in sanctions, including failure in the course. Allegations of Scholastic Dishonesty will be dealt with according to the procedures outlined in Appendix C, Chapter 11, of the General Information Bulletin, http://www.utexas.edu/student/registrar /catalogs/ .

Academic Dishonesty and Policies on Cheating

Faculty are committed to detecting and punishing all instances of academic dishonesty and will pursue cases of academic dishonesty in accordance with university policy. Academic dishonesty, in all its forms, is blight on our entire academic community. All parties in our community (professors, staff, and students) are responsible for creating an environment that educates outstanding professionals in our disciplines, and this goal entails excellence in technical skills, self-giving citizenry, and ethical integrity. Industry wants statisticians and data scientists who are competent and fully trustworthy, and both qualities must be developed day by day throughout an entire lifetime.

The general descriptions of assignments in this syllabus and the accompanying FAQs outline the acceptable methods of students working together in this course.
Summary: On quizzes and exams only work individually. On homework, it is acceptable for students to discuss problems on the graded homework among small groups of class members. Small groups means less than 12 students or so. The important part is that every student is participating and sharing their ideas and helping critique the ideas which arise.

Important details about UT policies and procedures involving academic integrity can be found in several places.

A description of how the UT Student Honor Code is to be interpreted in the MSDS online degree program is near the beginning of the Week 0 portion of this course. Please read that carefully. You will be asked multiple times in this course, and in other courses in the program, to acknowledge that you have read it and agree to abide by it.

More information is available at https://deanofstudents.utexas.edu/conduct/academicintegrity.php and the general statement about the requirements of the Student Honor Code are described here : https://catalog.utexas.edu/general-information/appendices/appendix-c/student-discipline-and-conduct/

All cheating will be reported directly to the college.

Documented Disability Statement

The University of Texas at Austin provides, upon request, appropriate academic adjustments for qualified students with disabilities. Any student with a documented disability who requires academic accommodations should contact Services for Students with Disabilities (SSD) using the procedures at from https://diversity.utexas.edu/disability/

Faculty are not required to provide accommodations without an official accommodation letter from SSD. Please notify us as quickly as possible if the material being presented in class is not accessible (e.g., instructional videos need captioning, course packets are not readable for proper alternative text conversion, etc.). Contact Services for Students with Disabilities through the above website or 512-471-6259 (voice) or 1-866-329-3986 (video phone.)