CSE 544:

News:
02/02: No class on 2/2 owing to snow day. Our first lecture will be on 2/4 via Zoom. Meeting link on Blackboard, under Zoom Meeting tab on the left.
01/12: Piazza course sign-up link

Probability & Statistics for Data Science Spring 2021

When: Tu Th, 1:15pm - 2:35pm
Where: Online, via zoom (details below)

Instructor: Anshul Gandhi
Instructor Office Hours: Tu Fr, 11am - 12pm

Course TA and Graders: Supreeth Narasimhaswamy, Michael Yao, Srikar Pothumahanti
TA Office Hours: TBD

Course Info

This grad-level course covers probability and statistics topics required for data scientists to analyze and interpret data. The course is also part of the Data Science and Engineering Specialization. The course is targeted primarily at PhD and Masters students in the Computer Science Department. Topics covered include Probability Theory, Random Variables, Stochastic Processes, Statistical Inference, Hypothesis Testing, Regression, and Time Series Analysis. For more details, refer to the syllabus below.

The class is expected to be interactive and students are encouraged to participate in class discussions.

Grading will be on a curve, and will be based on assignments, exams, and a semester-end mini data analysis project. For more details, see the section on grading below.

Hybrid Instruction and Online Learning

The course will be primarily online, including lectures, as mentioned below. The hybrid component requirements below apply on to Section 544-02 (in-person students).

- Lectures will be synchronous (live), via zoom. The zoom link can be found on Blackboard under the Zoom Meeting tab on the left.
  Please do not distribute the zoom link and please do not cause any disruption on the zoom lectures. Any such incidents will be reported to the Office of Judicial Affairs.
- All lectures will be automatically recorded by the Blackboard zoom account and will be available on the Zoom Meeting tab on Blackboard (usually a couple hours after the lecture).
- All lecture slides and code used in the class will be made available on this website (under the syllabus section) after class.

- For the hybrid component (those enrolled in Section 544-02), students will be required to either attend some lectures in-person in NCS 120 (during the regularly scheduled class times) or attend in-person office hours (during the regularly scheduled office hours). These in-person options will be available likely around March. The instructor will announce the availability of these options and further details in class once they are finalized.

- All offline communication will be via piazza; please sign up. You are responsible for monitoring piazza to ensure you do not miss important class announcements.

- Exams will be held remotely in an online manner via Blackboard.
- Assignments will be released via Blackboard (under the Assignments tab) and will need to be submitted online via the link that Blackboard provides.
- Assignment, exam, and project grades will be uploaded on Blackboard by the TAs along with summary comments on the grading scheme. Any regrading issues must be directed to the TAs.

Please email the instructor if you have any problems with remote instruction, such as a poor network connection, unaccommodating environment, or time zone issues.

Syllabus & Schedule

Date	Topic	Readings	Notes
Feb 02 (Tu)	Snow day		No class
Feb 04 (Th)	Course introduction, class logistics		

[Lec 01]			
Feb 09 (Tu) [Lec 02]	Probability review - 1 • Basics: sample space, outcomes, probability • Events: mutually exclusive, independent • Calculating probability: sets, counting, tree diagram	AoS 1.1 - 1.5 MHB 3.1 - 3.4	
Feb 11 (Th) [Lec 03]	Probability review - 2 • Conditional probability • Law of total probability • Bayes' theorem	AoS 1.6, 1.7 MHB 3.3 - 3.6	assignment 1 out
Feb 16 (Tu) [Lec 04]	Random variables - 1: Overview and Discrete RVs • Discrete and Continuous RVs • Mean, Moments, Variance • pmf, pdf, cdf • Discrete RVs: Bernoulli, Binomial, Geometric, Indicator	AoS 2.1 - 2.3, 3.1 - 3.4 MHB 3.7 - 3.9	Python scripts: draw_Bernoulli, draw_Binomial, draw_Geometric
Feb 18 (Th) [Lec 05]	Random variables - 2: Continuous RVs • Uniform(a, b) • Exponential(λ) • Normal(μ, σ^2), and its several properties	AoS 2.4, 3.1 - 3.4 MHB 3.7 - 3.9, 3.14.1	Python scripts: draw_Uniform, draw_Exponential, draw_Normal
Feb 23 (Tu) [Lec 06]	Random variables - 3: Joint distributions & conditioning • Joint probability distribution • Linearity and product of expectation • Conditional expectation	AoS 2.5 - 2.8 MHB 3.10 - 3.13, 3.15	assignment 2 out **assignment 1 due**
Feb	Probability Inequalities	AoS 5.3,	

25 (Th) [Lec 07]	• Weak Law of Large Numbers • Central Limit Theorem	5.4 MHB 3.14.2, 5.2	
Mar 02 (Tu) [Lec 08]	Markov chains • Stochastic processes • Setting up Markov chains • Balance equations	AoS 23.1 - 23.3 MHB 8.1 - 8.7	
Mar 04 (Th) [Lec 09]	Non-parametric inference - 1 • Basics of inference • Simple examples • Empirical PMF • Sample mean • bias, se, MSE	AoS 6.1 - 6.2, 6.3.1	
Mar 09 (Tu) [Lec 10]	Non-parametric inference - 2 • Empirical Distribution Function (or eCDF) • Kernel Density Estimation (KDE) • Statistical Functionals • Plug-in estimator	AoS 7.1 - 7.2	assignment 3 out. Required data: q2.dat, weather.dat **assignment 2 due** Python scripts: sample_Bernoulli, sample_Binomial, sample_Geometric, sample_Uniform, sample_Exponential, sample_Normal, draw_eCDF
Mar 11 (Th) [Lec 11]	Confidence intervals • Percentiles, quantiles • Normal-based confidence intervals • DKW inequality	AoS 6.3.2, 7.1	
Mar 16 (Tu) [Lec 12]	Parametric inference - 1 • Consistency, Asymptotic Normality • Basics of parametric inference • Method of Moments Estimator (MME)	AoS 6.3.1 - 6.3.2, 9.1 - 9.2	

Mar 18 (Th) [Lec 13]	Parametric inference - 2 • Properties of MME • Basics of MLE • Maximum Likelihood Estimator (MLE) • Properties of MLE	AoS 9.3, 9.4, 9.6	**assignment 3 due**
Mar 23 (Tu)	Mid-term 1		Via Blackboard

Resources

- **Required text: (AoS) "All of Statistics : A Concise Course in Statistical Inference" by Larry Wasserman (Springer publication)**.
  - Students are strongly suggested to purchase a copy of this book.
- Recommended text: (MHB) "Performance Modeling and Design of Computer Systems: Queueing Theory in Action" by Mor Harchol-Balter (Cambridge University Press)
  - Suggested for probability review and stochastic processes.
  - There is copy placed on reserve in the library. The instructor also has a few personal copies that you can borrow.
- Recommended text: (DSD) "The Data Science Design Manual" by (our very own) Steven Skiena (Springer publication).
  - Suggested for data science topics in the second half of the course.

- Others:
  - S.M. Ross, Introduction to Probability Models, Academic Press
  - S.M. Ross, Stochastic Processes, Wiley

Grading (tentative)

- Assignments: 45%
  - 6 assignments during the semester. Expect 5-8 questions per assignment, including some programming questions.
  - Collaboration is allowed (max group size 4). You are free to form your own groups, and group membership can change between assignments.
  - Submit one softcopy solution per group, typed or handwritten, but should be legible. Should include all required plots and work. Include group member names and ID numbers on the first page.

    - **Assignments are due via Blackboard, at the beginning of the lecture (1:15pm). No late submissions allowed**.

- Exams: 45%
    - Two online (via Blackboard), timed exams.
    - Mid-term 1: 20%.
    - Mid-term 2: 25%.
    - Easier than the assignments but the questions will be on the same lines.

- Mini group project: 10%
    - One semester-end project to be done in groups. The project work is expected to begin in the last month of the course.
    - Further details on the project will be provided in class around mid-March.

- **Important:**

    - **Academic dishonesty will immediately result in an F** and the student will be referred to the Academic Judiciary. See below section on Academic Integrity.
    - Grading will be on a curve.
    - **Assignment of grades by the instructor will be final; no regrading requests will be entertained**.
    - There is a University policy on grading, as well as a set of grading guidelines agreed upon by the CS faculty.
      No exceptions will be made for any student and no special circumstances will be entertained.

Academic Integrity

Each student must pursue his or her academic goals honestly and be personally accountable for all submitted work. Representing another person's work as your own is always wrong. Faculty are required to report any suspected instances of academic dishonesty to the Academic Judiciary. For more comprehensive information on academic integrity, including categories of academic dishonesty, please refer to the academic judiciary website at http://www.stonybrook.edu/commcms/academic_integrity. Please note that any incident of academic dishonesty will *immediately result in an F grade* for the student.

Critical Incident Management

Stony Brook University expects students to respect the rights, privileges, and property of other people. Faculty are required to report to the Office of Judicial Affairs any disruptive behavior that interrupts their ability to teach, compromises the safety of the learning environment, or inhibits students' ability to learn.

Student Accessibility Support Services

If you have a physical, psychological, medical, or learning disability that may impact your course work, please contact the Student Accessibility Support Center, 128 ECC Building, (631) 632-6748, or at sasc@stonybrook.edu. They will determine with you what accommodations are necessary and appropriate. All information and documentation is confidential. https://www.stonybrook.edu/sasc.

Please report any errors to the Instructor.