CSE 357: Statistical Methods for Data Science Fall 2020

News:

08/07: Piazza course sign-up link

07/15: Our first lecture will be on Aug 24th (Mon) at 2:40pm via zoom.

07/15: Course website up.

When: Mon Wed, 2:40pm - 4:00pm
Where: Online, via zoom (details below)

Instructor: Anshul Gandhi
Instructor Office Hours: TBD, via zoom

Course TAs: TBD
TA Office Hours: TBD, via zoom

Course Info

This undergraduate-level course covers probability and statistics topics required for data scientists to analyze and interpret data. The course will involve theoretical topics and some programming assignments. The course is targeted primarily for junior and senior undergraduate students who are comfortable with concepts relating to probability and are comfortable with basic programming. Undergraduates from Computer Science, Applied Mathematics and Statistics, and Electrical and Computer Engineering would be well suited for taking this class. Topics covered include Probability Theory, Random Variables, Stochastic Processes, Statistical Inference, Hypothesis Testing, and Regression. For more details, refer to the syllabus below.

Grading will be on a curve, and will tentatively be based on assignments, exams, and in-class quizzes; all components will be handled remotely. For more details, refer to the section on grading below.

Remote Instruction and Online Learning

The course will be entirely online, including lectures and office hours. Please read the below information carefully.

- Lectures will be synchronous (live), via zoom. The zoom link (likely a recurring meeting link) will be shared with all students who are enrolled or in the waiting list on SOLAR.
  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 recorded and posted on echo360 (accessible via blackboard) a few 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.
- Instructor and TA office hours will be via zoom; the links (likely a recurring meeting link) will be shared in a timely manner before the office hours.

- Exams will be held remotely in an online manner, most likely via blackboard.
- Assignments will be released via blackboard and will need to be submitted online (most likely via blackboard or google forms; details will be provided in a timely manner).
- Graded assignments and exams will be either emailed back to you or shared with you on request. Any regrading issues will be handled remotely as well.
- In-class quizzes (likely starting from the 2nd week) will be held during the lectures, and will be administered via blackboard or echo360; details will be confirmed in the first week of classes.

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
Aug 24 (Mon) [Lec 01]	Course introduction, class logistics		
Aug 26 (Wed) [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	
Aug 31 (Mon) [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**
Sep 02 (Wed) [Lec 04]	Random variables - 1 • Mean, Moments, Variance • pmf, pdf, cdf • Bernoulli(p) • Indicator RV • Binomial(n, p) • Geometric(p)	AoS 2.1 - 2.3, 3.1 - 3.4 MHB 3.7 - 3.9	Python scripts: draw_Bernoulli, draw_Binomial, draw_Geometric
Sep 07 (Mon)	Labor Day observed		No class
Sep 09 (Wed) [Lec 05]	Random variables - 2 • 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	**assignment 1 due** **assignment 2 out** Python scripts: draw_Uniform, draw_Exponential, draw_Normal
Sep 14 (Mon) [Lec 06]	Random variables - 3 • Joint probability distribution • Linearity and product of expectation • Linearity of variance	AoS 2.5 - 2.7 MHB 3.10, 3.13	
Sep 16 (Wed) [Lec 07]	Probability inequalities • Markov's Inequality • Chebyshev's inequality • Weak Law of Large Numbers • Central Limit Theorem	AoS 4.1 - 4.2, 5.3 - 5.4 MHB 3.14.2, 5.2	
Sep 21 (Mon) [Lec 08]	Non-parametric inference - 1 • Basics of inference • Empirical PMF • Sample mean • bias, se, MSE	AoS 6.1, 6.2, 6.3.1	**assignment 2 due** **assignment 3 out** Required collisions.csv dataset for A3. Python scripts: sample_Bernoulli, sample_Binomial, sample_Geometric

Sep 23 (Wed) [Lec 09]	Non-parametric inference - 2 • Empirical Distribution Function (or eCDF) • Statistical Functionals • Plug-in estimator	AoS 6.3.1, 7.1 - 7.2	Python scripts: sample_Uniform, sample_Exponential, sample_Normal, eCDF
Sep 28 (Mon) [Lec 10]	Confidence intervals • Percentiles, quantiles • Normal-based confidence intervals	AoS 6.3.2, 7.1	
Sep 30 (Wed) [Lec 11]	Parametric inference - 1 • Basics of parametric inference • Method of Moments Estimator (MME) • Properties of MME	AoS 6.3.1 - 6.3.2, 9.1 - 9.2	**assignment 3 due**
Oct 05 (Mon) [Lec 12]	Mid-term 1 review		
Oct 07 (Wed)	Mid-term 1		
Oct 12 (Mon) [Lec 13]	Parametric inference - 2 • Likelihood • Maximum Likelihood Estimator (MLE) • Properties of MLE	AoS 9.3 - 9.4, 9.6	**assignment 4 out** Required data: acceleration, model, mpg, q7_X.dat, q7_Y.dat.
Oct 14 (Wed) [Lec 14]	Hypothesis testing - 1 • Basics of hypothesis testing • The Wald test	AoS 10 - 10.1 DSD 5.3 - 5.3.1	
Oct 19 (Mon) [Lec 15]	Hypothesis testing - 2 • Type I and Type II errors • The Wald test	AoS 10 - 10.1 DSD 5.3.1	
Oct 21 (Wed) [Lec 16]	Statistics in Medicine		Guest lecture by Dr. Shrivastava
Oct 26 (Mon) [Lec 17]	Hypothesis testing - 3 • Z-test • t-test	AoS 10.10.2 DSD 5.3.2	**assignment 4 due** **assignment 5 out**
Oct 28 (Wed) [Lec 18]	Hypothesis testing - 4 • Kolmogorov-Smirnov test (KS test) • p-values	AoS 15.4, 10.2 DSD 5.3.3, 5.5	
Nov 02 (Mon) [Lec 19]	Hypothesis testing - 5 • p-values • Permutation test	AoS 10.2, 10.5 DSD 5.5	
Nov 04 (Wed) [Lec 20]	Hypothesis testing - 6 • Pearson correlation coefficient • Chi-square test for independence	AoS 3.3, 10.3 - 10.4 DSD 2.3	

Nov 09 (Mon)			No class **assignment 5 due**
Nov 11 (Wed) [Lec 21]	Bayesian inference - 1 - Bayesian reasoning - Bayesian inference	AoS 11.1 - 11.2 DSD 5.6	**assignment 6 out** Required datasets: q3_sigma3.dat, q3_sigma100.dat, q5.dat, q6.csv.
Nov 16 (Mon) [Lec 22]	Bayesian inference - 2 - Bayesian inference - Conjugate priors	AoS 11.1 - 11.2 DSD 5.6	
Nov 18 (Wed) [Lec 23]	Regression - 1 - Basics of Regression - Simple Linear Regression	AoS 13.1, 13.3 - 13.4 DSD 9.1	
Nov 23 (Mon)	Thanksgiving break		No class
Nov 25 (Wed)	Thanksgiving break		No class
Nov 30 (Mon) [Lec 24]	Regression - 2 - Multiple Linear Regression	AoS 13.5 DSD 9.1	
Dec 02 (Wed) [Lec 25]	Mid-term 2 review		**assignment 6 due**
Dec 07 (Mon)	Mid-term 2		

Resources

- Recommended 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: 48%
  - 6 assignments during the semester. Expect 5-7 questions per assignment, including some programming

    - questions (especially after mid-term 1).
    - 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.
    - **Assignments are due in class, at the beginning of the lecture. No late submissions allowed**.

- Exams: 45%
    - Two online exams, via blackboard.
    - Mid-term 1: 20%.
    - Mid-term 2: 25%.
    - Easier than the assignments and no long derivations or programming questions.

- In-class mini-quizzes: 7%
    - A simple, timed, in-class quiz, administered via blackboard or echo360, based on the material covered in that lecture.
    - Half the grade is for participation and half for getting the right answer.
    - Will count for attendance as well.

- **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. The instructor is obligated to uphold these policies.
      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.