ENG EK381: Probability, Statistics, and Data Science for Engineers

Course Information

Lecture Time and Place: PHO 211, Tu/Th 1:30-3:15.

Recitations/Discussion Sessions/GTF/UTF Office hours:

1. Monday, 5-6, PHO 404/428.
2. Tuesday, 5:30-6:30, PHO 404/428.
3. Wednesday, 4:40-5:30, EPC 204 (**Recitation A**).
4. Thursday, 12:30-1:30, PSY B33 (**Recitation B**).
5. Friday, 3:00-4:30, PHO 404/428.

You have to attend only one recitation. The other times are optional, in case you have questions or need homework help.

Lecturer: Yannis Paschalidis, 8 St. Mary's St., Room PHO 429, tel: 617-353-0434, yannisp@bu.edu, `http://sites.bu.edu/paschalidis`.
Office hours: Th: 3:15-4:15, Fr: 2:00-3:00. I would recommend to try Piazza first (see below) for getting answers to well-formulated questions. The best way to reach me is via e-mail; we can arrange a meeting time outside my regular office hours.

Graduate Teaching Fellows: Alp Acar, alpacar@bu.edu, Yujia Xue, yujiaxue@bu.edu.

Undergraduate Teaching Fellows: TBD.

Motivation: It is becoming increasingly clear that the next generation of engineers will need to draw upon concepts and techniques from probability and statistics to tackle the challenges posed by uncertain, complex systems and large, high-dimensional data sets. This course is intended to give all engineering students a strong foundation in probability and statistics as well as an introduction to ideas from data analytics and machine learning. Any student that successfully completes this course will be well-prepared to take upper-electives in machine learning, data analytics, random processes, as well as any other course that draws heavily upon probabilistic reasoning.

BU Hub Outcomes: Students who successfully complete this course will have satisfied:

- *Quantitative Reasoning II:* Students will learn how to translate complex engineering problems into formal probabilistic questions as well as the techniques needed to evaluate these statements analytically and algorithmically. They will gain intuition for statistical thinking, its modern application to large data sets, as well as the limitations of this approach and associated risks.

- *Critical Thinking:* Students will learn where and why "common-sense" intuition does not match up with formal reasoning through probability and statistics. They will develop the toolkit needed to assess modern engineering problems from the probabilistic lens and make and defend the validity of arguments, including their own.

Course Goals (ENG): Provide students with:

- A background in the foundations of probability theory.
- Intuition for probabilistic concepts and reasoning.
- Experience with standard probability distributions and their application to modeling engineering systems and data sets.
- An understanding of statistics and the application of statistical techniques to data sets.
- An appreciation for the applications of probability in the design and analysis of modern engineering systems and processes.
- An understanding of hypothesis testing and optimal decision rules.
- Experience with numerical software for generating and analyzing pseudorandom quantities.
- A basic understanding of modern methods in data science and machine learning.

Course Outcomes (ENG): As an outcome of completing this course, students will:

- Understand the basic foundations of probability theory.
- Develop intuition for probabilistic concepts and gain experience with principled probabilistic thinking and computation.
- Become familiar with standard probability distributions for discrete and continuous random variables.
- Acquire intuition for multiple random quantities and their relationships.
- Understand the role of probability for modeling and analyzing complex engineering systems.
- Understand the basic principles and techniques of statistics, and gain experience with their application to data sets.
- Gain experience using numerical software to simulate random events.
- Understand the principles of optimal hypothesis testing and detection.
- Be exposed to basic methods in data analytics and machine learning and their applications.

- Acquire the preparation needed to succeed in upper-level courses that rely on probability as a pre-requisite.

Course Websites:

Website:	`http://learn.bu.edu/`
Discussion board:	`http://piazza.com/` Signup Link: `http://piazza.com/bu/fall2018/ek381a1` Class Link: `http://piazza.com/bu/fall2018/ek381a1/home`

Piazza: We will be using Piazza as a discussion board. You have all been registered and you should have received an invitation to join. The system is highly catered to getting you help quickly and efficiently from both the course staff and your fellow classmates. You have the option of asking (and answering) questions anonymously, meaning that your name will only be displayed as "Anonymous" to everyone else. Rather than emailing questions, I encourage you to post your questions on Piazza.

Prerequisites: ENG EK103 Computational Linear Algebra.

Corequisites: CAS MA225 Multivariate Calculus.

Textbooks: There is a required textbook for the course:
[YG] Roy D. Yates and David J. Goodman, *Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers*, Wiley, 3rd Edition, 2014.
It is available at the BU bookstore and from other vendors. Compare prices as the BU bookstore usually sells it at a discount. We will refer to this book as [YG] in the course handouts.

The textbook has a companion website that includes solutions to odd-numbered problems, MATLAB files and reference material, and other resources: `https://bit.ly/2MNWclR`.

Grading: There will be several on-line quizzes, two in-class midterm exams, and a final exam. Your grade will be formed as follows:

1. 15% Homework and on-line quizzes. We will count the best $n-1$ scores out of n assignments.
2. (22.5+22.5=)45% Midterm exams.
3. 35% Final.
4. 5% Attendance and class participation **only if** your overall homework+on-line quizz score exceeds 85%.

Attendance: You will find that active class attendance and compilation of class notes are essential in this course. The concepts we will cover are fundamental, have applications in almost every area of engineering and science, and will be novel to most of you. Because the topics we will cover build upon each other, if you fall behind you may find that you are lost and not able to follow the lectures. The material introduced in class is at times conceptually subtle and will have to mature in your mind before you become comfortable with it. The probability that simply intensive studying immediately before exams will suffice is very close to zero!

Homework: Homeworks will be assigned weekly. They will be due one week after the date issued. We will only accept hard copies of your homework during class, either handwritten or typed (except for some computational assignments for which special instructions will be given). Homework submission by email will not be accepted. Deadlines will be strictly enforced. Although homeworks represent only 10% of the grade you will find that they are *essential* to the learning process. I strongly encourage you to work on them independently. Especially in probability, it is easy to follow another person's solution but much harder to come up with your own. Past experience has shown that the performance in the exam is highly correlated with your ability to solve problem sets on your own! We will offer homework help at the discussion meeting times and office hours.

Rules of Conduct: You *may* collaborate in study groups on the solution of homeworks. An *acceptable* form of collaboration is to discuss with others possible approaches for solving the problems and then fill the details and write your solutions independently. Copying the solution that someone else has written is *unacceptable* and at times transparent. If you do collaborate you *should* acknowledge your collaborators in the write-up for each problem. I view this as essential!

Needless to say that I expect students to adhere to basic, common sense concepts of academic honesty; presenting another's work as your own or cheating on exams will not be tolerated. Knowingly allowing others to represent your work as their own is as serious an offense as submitting another's work as your own. BU takes academic integrity very seriously. More information on BU's Academic Conduct Code, with examples, may be found at `http://www.bu.edu/academics/policies/academic-conduct-code`.

Make-up Exams: There will be no make-up exams. If there is a legitimate reason for missing an exam, then the scores of other exams will be used appropriately to compensate for the missed exam. If there is no legitimate reason provided for missing an exam, a grade of zero will be assigned for the missed exam.

Final Exam: The exam has been scheduled for **Tuesday, December 18, 3:00pm-5:00pm**. Students must be present for the final exam, except under exceptional circumstances that must be discussed with the instructor at the beginning of the semester. Going home early for the holidays *is not* an legitimate excuse.

Incomplete grades: Incomplete grades will not be given to students who wish to improve their grade by taking the course in a subsequent semester. An incomplete grade may be given for medical reasons if a doctor's note is provided. The purpose of an incomplete grade is to allow a student *who has essentially completed the course* and who has a legitimate interruption in the course, to complete the remaining material in another semester. Students will not be given an opportunity to improve their grades by doing extra work.

Drop dates: Students are responsible for being aware of the drop dates for the current semester. Drop forms will not be back-dated.

Course Material: Most of the course material will be posted on the course website.

Working with Data. Part of the homework assignments will ask students to apply the concepts they have learned to real data sets, using a numerical programming language such as MATLAB.

Other books on the topics covered in the course are:

- D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to Probability*, Athena Scientific, 2002.
  Similar coverage as [YG].

- W. Feller, *An Introduction to Probability Theory and its Applications*, Volume I, 3rd Edition, Wiley, 1968.
  The classic text for probability courses. Excellent and extensive coverage.

- W.W. Hines, D.C. Montgomery, D.M. Goldman, and C.M. Borror, *Probability and Statistics in Engineering*, 4th Edition, Wiley, 2003.
  Good reference book for quality control applications of probability and statistics.

- A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd Edition, McGraw Hill, 1991.
  Typically used for more advanced courses on Estimation and Stochastic Processes. Has a good coverage of introductory probability and statistics. Excellent reference book.

- S. Ross, *A First Course in Probability*, 9th Edition, Pearson, 2014.
  Similar coverage with [YG].

- S. Ross, *Introduction to Probability and Statistics for Engineers and Scientists*, 5th Edition, Academic Press, 2014.
  Discusses a number of applications in engineering.

Other resources. These are resources you may want to consult as you are studying, to help you see things from a different angle, provide sample problems you can work on, etc. In general, the more problems you work on, the better you become at solving a new problem and understanding the material better.

1. The [YG] student solutions manual (`https://bit.ly/2MJuPJV`) with solutions to odd-numbered problems.

2. MIT OpenCourseWare Material (MIT 6.041, `https://bit.ly/2LrP4a3`).

3. Course notes by Prof. Bruce Hajek at UIUC (`https://bit.ly/2weiJPh`).

Core Topics. Each department's section of EK381 is expected to cover the following topics. Examples and motivating applications will be discipline-specific, depending on the section. Note that each topic is broken down into *concepts*. This is how we will organize our thinking about probability throughout the semester (as opposed to formulas, etc.).

1. **Foundations of Probability**
    - Set Theory
    - Sample Space, Outcomes, Events
    - Probability Law
    - Conditional Probability
    - Total Probability Theorem
    - Bayes' Theorem
    - Independence
    - Conditional Independence
    - Counting Methods
    - Independent Trials

2. **Discrete Random Variables**
    - Probability Mass Function (PMF)
    - Cumulative Distribution Function (CDF)
    - Average and Expectation
    - Functions of Discrete Random Variables (RVs) and their Expectations
    - Variance and Standard Deviation
    - Important Families of Discrete RVs
    - Conditioning a Discrete RV by an Event

3. **Continuous Random Variables**
    - Cumulative Distribution Function (CDF)
    - Probability Density Function (PDF)

- Expectation of Continuous RVs
- Functions of Continuous RVs and their Expectations
- Variance and Standard Deviation
- Important Families of Continuous RVs
- Conditioning a Continuous RV by an Event

4. **Multiple Random Variables**
   - Joint CDFs, PMFs, and PDFs
   - Marginal PMFs and PDFs
   - Conditional PMFs and PDFs
   - Independent RVs
   - Functions of Multiple RVs
   - Covariance and Correlation
   - Jointly Gaussian RVs
   - Orthogonal Random Variables
   - Conditional Expectation
   - Iterated Expectation

5. **Detection**
   - Binary Hypothesis Testing
   - Maximum Likelihood (ML) Detection
   - Maximum a Priori (MAP) Detection
   - Minimum Mean-Squared Error (MMSE) Estimation
   - Linear Least-Squares Error (LLSE) Estimation

6. **Statistics**
   - Sample Mean and Variance
   - Law of Large Numbers
   - Central Limit Theorem
   - Confidence Intervals
   - Parametric Statistical Testing

7. **Intro to Data Science and Machine Learning**
   - Random Vectors
   - Training and Test Error

- Basic Classifiers (Nearest Neighbor, Linear)
- Principal Component Analysis (PCA)

8. **Markov Chains**

    - Finite State Automata
    - Markov Property
    - Transition Probabilities
    - Steady State Probabilities
    - State Classification
    - Irreducible Markov Chains
    - Multiple Communicating Classes