USC Viterbi School of Engineering

INF564: Probability and Statistics for Data Science (Fall 2018)

Units: 4

Instructor: Mohammad Reza Rajati, PhD
PHE 412
rajati@usc.edu – Include INF 564 in subject

Office Hours: Wednesday 3:30 –5:00 PM

TA(s): TBD
@usc.edu – Include INF 564 in subject
Office Hours: TBD
Office Location: TBD

TBD
@usc.edu –Include INF 564 in subject
Office Hours: TBD
Office Location: TBD

Lecture(s): Tuesday, Thursday, 4:00 - 5:50 pm in KAP 156

Webpages: Piazza Class Page for everything except grades
and Blackboard for grades and homework submission
– All HWs, handouts, solutions will be posted in PDF format
– *Student has the responsibility to stay current with webpage material*

Prerequisites: Prior courses in multivariate calculus, linear algebra, and linear system theory.

Other Requirements: Basic computer skills (e.g., plotting, Matlab, Excel, Python, etc.). Use of R is ma

Tentative Grading: Assignments 20%
Three Midterm Exams 40%
Final Exam 40%
Participation on Piazza* 5%
(Optional) Project TBD

Letter Grade Distribution:

$\geq$ 93.00	A	73.00 - 76.99	C
90.00 - 92.99	A-	70.00 - 72.99	C-
87.00 - 89.99	B+	67.00 - 69.99	D+
83.00 - 86.99	B	63.00 - 66.99	D
80.00 - 82.99	B-	60.00 - 62.99	D-
77.00 - 79.99	C+	$\leq$ 59.99	F

Disclaimer: Although the instructor does not expect this syllabus to drastically change, he reserves every right to change this syllabus any time in the semester.

Note on e-mail vs. Piazza: If you have a question about the material or logistics of the class and wish to ask it electronically, please post it on the piazza page (not e-mail). You may post it anonymously if you wish. Often times, if one student has a question/comment, other also have a similar question/comment. Use e-mail with the professor, TA, graders only for issues that are specific to your individually (e.g., a scheduling issue or grade issue).

Catalogue Description: This course introduces fundamental concepts in probability and statistics from a data science perspective. It aims at synergistically presenting rigorous probabilistic reasoning and problem solving as well as computerage statistical methods that are widely used in data science.

Course Objectives: Upon successful completion of this course a student will

- Understand axiomatic probability and know how to model real-world problems using it
- Understand Discrete and Continuous Random Variables, their distributions, their properties, moments, and correlations.
- Understand the limiting behavior of large amounts of data by limit theorems
- Understand sampling and sampling distributions
- Cunstruct models of distributions using histograms and density estimation techniques
- Estimate parameters of distributions using maximum Likelihood, maximum a-posteriori, and other estimation techniques
- Assess the properties of estimators
- Construct confidence intervals for point estimates
- Test hypotheses about different parameters of distributions of populations using samples of data
- Construct linear regression models for data, assess those models, and select variables using various techniques including statistical tests and regularization
- Use non-parametric and robust statistical tools to assess classification and regression models
- Apply resampling and Monte-Carlo methods to computation and statistical inference
- Use statistical software (R, SPSS, STATA, Python, etc.) to effectively use all of the techniques learned in the course
- Be ready for understanding and implementing machine learning and data mining methods that rely on statistical analysis.

Exam Dates:

- **Midterm Exam 1:** Friday, September 27, 8:00 - 9:50 AM
- **Midterm Exam 2:** Friday, October 25, 8:00 - 9:50 AM
- **Midterm Exam 3:** Friday, November 15, 8:00 - 9:50 AM
- **Final Exam:** Thursday, December 12, 4:30 - 6:30 PM as set by the university
- **Note**: Midterm Exams may be moved to 10-11:50 AM if rooms are available. Please make sure you can take the exam at 10-11:50 AM as well.

Textbooks:

- **Required Textbooks:**

  1. *Probability and Random Processes for Electrical and Computer Engineers*, 1st Edition
     Author: John A. Gubner; Cambridge University Press, 2006. **ISBN-13:** 978-0521864701
  2. *Probability with Applications and R*, 1st Edition
     Author: Robert P. Dobrow; Wiley, 2014. **ISBN-13:** 978-1-118-24125-7
  3. *Mathematical Statistics with Resampling and R*, 2nd Edition
     Authors: Laura M. Chihara and Tim C. Hesterberg; Wiley, 2019. **ISBN-13:** 978-1119416548

- **Recommended Textbooks:**

  1. *Probability and Random Processes*, 3rd Edition
     Authors: Geoffery R. Grimmet and David R. Stirzaker; Oxford University Press; 2001. **ISBN-13:** 978-0198572220
  2. *Introduction to Probability*, 2nd Edition
     Authors: Dimitri P. Bertsekas and John N. Tsitsiklis; Athena Scientific, 2008. **ISBN-13:** 978-1886529236
  3. *Introduction to Probability Models*, 11th Edition
     Authors: Sheldon M. Ross, Academic Press, 2010. **ISBN-13:** 978-0124079489
  4. *One Thousand Exercises in Probability*, 1st Edition
     Authors: Geoffery R. Grimmet and David R. Stirzaker; Oxford University Press; 2001. **ISBN-13:** 978-0198572213
  5. *Schaum's Outline of Probability, Random Variables, and Random Processes*, 3rd Edition
     Author: Hwei P. Hsu; McGraw-Hill Education; 2014. **ISBN-13:** 978-0071368100
  6. *Schaum's Outline of Probability and Statistics*, 4th Edition
     Authors: John J. Schiller Jr., R. Alu Srinivasan, Murray R Spiegel; McGraw-Hill Education; 2012. **ISBN-13:** 978-0071795579
  7. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, 1st Edition
     Authors: Bradly Efron and Trevor Hastie; Cambridge University Press, 2016. **ISBN-13:** 978-1107149892

8. *Probability and Statistics for Data Science*, 1st Edition **Author:** Norman Matloff; Chapman and Hall, 2019. **ISBN-13:** 978-1138393295
9. *Statistical Inference*, 2nd Edition
   Authors: George Casella and Roger L. Berger; Duxbury, 2001. **ISBN-13:** 978-0534243128
10. *An Introduction to Statistical Inference and Its Applications with R*, 1st Edition
    Author: Michael W. Trosset; CRC Press, 2009. **ISBN-13:** 978-1584889472
11. *Introduction to Mathematical Statistics*, 8th Edition
    Authors: Robert V. Hogg, Joseph W. McKean, and Allen T. Craig; Pearson, 2018. **ISBN-13:** 978-0-13-468699-8

Grading Policies:

- The letter grade distribution table guarantees the *minimum* grade each student will receive based on their final score. When appropriate, relative performance measures will be used to assign the final grade, at the discretion of the instructor.
  - Final grades are non-negotiable and are assigned at the discretion of the instructor. If you cannot accept this condition, you should not enroll in this course.
  - Three of your lowest homework grades will be dropped from the final grade.
  - 80% of your lowest midterm grade will be dropped. For example, if you receive 80, 60, and 32 in three midterms, your total midterm grade will be $\frac{80+60+0.2\times32}{2.2} = 66.55$ instead of $\frac{80+60+32}{3} = 57.33$.
  - *Participation on Piazza has up to 5% extra credit, which is granted on a competetive basis *at the discretion of the instructor*.

- **Homework Policy**
  - Homework is assigned on a weekly basis. *Absolutely no late homework will be accepted. A late assignment results in a zero grade.*
  - Homework solutions should be typed or *scanned* using scanners or mobile scanner applications like CamScanner and uploaded on the course website (photos taken by cell-phone cameras and in formats other than pdf will NOT be accepted). Programs and simulation results have to be uploaded on the course website as well.
  - Students are encouraged to discuss homework problems with one another, but each student must do their own work and submit individual solutions written/ coded in their own hand. Copying the solutions or submitting identical homework sets is written evidence of cheating. The penalty ranges from F on the homework or exam, to an F in the course, to recommended expulsion.
  - Posting the homework assignments and their solutions to online forums or sharing them with other students is strictly prohibited and infringes the copyright of the instructor. Instances will be reported to USC officials as academic dishonesty for disciplinary action.

- **Exam Policy**

  - **Make-up Exams:** No make-up exams will be given. If you cannot make the above dates due to a class schedule conflict or personal matter, you must drop the class. In the case of a required business trip or a medical emergency, a signed letter from your manager or physician has to be submitted. This letter must include the contact of your physician or manager.
  - Midterms and final exams will be closed book and notes. No calculators are allowed nor are computers and cell-phones or any devices that have internet capability. One letter size cheat sheet (back and front) is allowed for the midterms. Two letter size cheat sheets (back and front) are allowed for the final.
  - All exams are cumulative, with considerable emphasis on material presented since the last exam.

- **Attendance:**

  - Students are required to attend all the lectures and discussion sessions and actively participate in class discussions. Use of cellphones and laptops is prohibited in the classroom. If you need your electronic devices to take notes, you should discuss with the instructor at the beginning of the semester.

Important Notes:

- Textbooks are secondary to the lecture notes and homework assignments.
- Handouts and course material will be distributed.
- Please use your USC email to register on Piazza and to contact the instructor and TAs.

Tentative Course Outline

Tuesday	Thursday
Aug 27th **1** **Introduction** **Logic**	29th **2** **Set Theory**
Sep 3rd **3** **Set Theory** , **Probability Models** • Sample Space, • σ-algebra of events • Probability as An Additive Measure • Continuity of Probability • Conditional Probability	5th **4** **Probability Models and Independence** • Total Probability • The Baye's Rule • The Multiplication Rule
10th **5** **Random Variables** • Definitions • CDFs • Borel Sets	12th **6** **Random Variables** • CDFs • Independence • Multiple Random Variables **Combinatorics (Reading)**
17th **7** **Discrete Random Variables** PMFs Famous Discrete Random Variables	19th **8** **Discrete Random Variables** • Famous Discrete Random Variables • Multiple Random Variables • Joint PMFs • Marginal PMFs • Conditional PMFs • Total Probability • Substitution Law • Independence • Derived Distributions

Tuesday	Thursday
24th **9** **Moments of Discrete Random Variables** • Expectation • The Law of The Unconscious Statistician • Properties of Expectation • Higher Order Moments • Variance and Standard Deviation	26th **10** **Moments of Discrete Random Variables** • Moments of Famous Discrete Random Variables • Existence of Expectations* • Covariance and Correlation and Their Properties • Expectation As Norm and Inner Product • The Cauchy-Schwartz-Bunyakovsky Lemma
Oct 1st **11** **Moments of Discrete Random Variables** • Expectation As Norm and Inner Product • The Cauchy-Schwartz-Bunyakovsky Lemma **Conditional Expectation** • The Law of The Unconscious Statistician • Substitution Law for Conditional Expectation • Total Expectation	3rd **12** **Conditional Expectation** • Conditional Expectation as A Random Variable • Properties of Conditional Expectation • Existence of Conditional Expectation • Conditional Probability as Conditional Expectation • Wald's Equality • Projections, Projection Theorem, Principle of Orthogonality • Conditional Expectation as an Estimator

Tuesday	Thursday
8th **13** **Continuous Random Variables** • PDFs • Important Continuous Random Variables	10th **14** **Continuous Random Variables** • Important Continuous Random Variables • Multiple Random Variables and Joint PDFs • Marginal PDFs • Independence • Conditional Probability and Conditional PDFs • Moments of Continuous Random Variables
15th **15** **Continuous Random Variables** • Existence and Properties of Moments • Moments of Famous Continuous Random Variables • The Law of The Unconscious Statistician (LOTUS)	17th Fall Recess

Tuesday	Thursday
22nd **16** **Continuous Random Variables** • The Law of Total Probability • The Substitution Law • Total Probability • Total Expectation • Total Probability and Expectation for Multiple Random Variables • Conditional Expectation **The Bivariate Normal Distribution**	24th **17** **Random Vectors** • Expectation of A Random Vector • Linearity of Expectation • Auto-correlation Matrix • Covariance Matrix • Positive Definiteness • Cross-correlation Matrix • Cross-covariance Matrix • The Multivariate Normal Distribution **Derived Distributions** • Monotonic Functions • Linear Functions
29th **18** **Derived Distributions** • Non-Monotonic Functions • Multivariable Functions • Linear Mappings • A Single Function of Multiple Random Variables • Order Statistics	31st **19** **Derived Distributions** • Order Statistics • Sum of Independent Random Variables • Noraml Random Variables in Polar Cooridnates • The Rayleigh Distribution • Simulation of Random Variables • The Box-Muller Method • Rejection Sampling Algorithm

Tuesday	Thursday
Nov 5th **20** **Generating Functions** • Moment Generating Functions • Region of Convergence • Inversion of MGFs • Properties of MGFs	7th **21** **Generating Functions** • Random Sums of Random Variables • Laplace and Z transforms • Characteristic Functions • Generating Functions for Random Vectors • Joint Characteristic Functions
12th **22** **Concentration Inequalities** • Markov and Chebychev Inequalities **Stochastic Convergence** • Modes of Convergence • Hierarchy of Modes of Convergence	14th **23** **Limit Theorems** • Weak Law of Large Numbers • Strong Law of Large Numbers • Monte-Carlo Methods • Bootsrtap* • The Central Limit Theorem • Berry-Esseen Theorem • Binomial Approximation • Chi-squared Approximation

Tuesday	Thursday
19th **24** **Statistics** • Histograms • Kernel Density Estimation • Point and Interval Estimation of The Mean • One-Sided and Two-Sided Confidence Intervals • Interpretation of Confidence Intervals • Estimation of Variance • Student's T-Statistic	21st **25** **Statistics** • Point and Interval Estimation of Proportion • Two Sample Confidence Intervals • Interval Estimation of Difference between Means (Independent and Dependent Samples) • Interval Estimation of Ratio of Variances • The Fisher-Snedecor Statistic • Boostrap Confidence Intervals* • Frequentist (Fisherian) Hypothesis Testing • p-values* • Type-I and Type-II Errors • Power of A Test • Neyman-Pearson Lemma* • Testing for The Mean, Proportion, Difference in The Means, and Difference in The Proportions • The Kolmogorov-Smirnov Test • The Chi-Squared Test

Tuesday	Thursday
26th **26** **Statistics** • Parameter Estimation • Properties of Estimators • Method of Moments • Minimum Variance Unbiased Estimator* • Maximum Likelihood Estimation • The Cramér-Rao Bound • Maximum A-Posteriori Estimate • Minimum Mean-Squared Error Estimate	28th Thanksgiving Recess
Dec 3rd **27** **Statistics: Linear Regression*** • Simple Linear Regression • Multiple Regression • Least Squares • Confidence Intervals and Hypothesis Testing for Coefficients • Multicollinearity • Heteroscedasticity • F-test for ANOVA and Overal Significance of Model	5th **28** **Markov Chains*** • The Markovian Property • Markov Chains • Random Walks • Homogeneous Chains • Transition Matrix • Transition Graph • The Chapman-Kolmogorov Equation • Steady State Behavior of Markov Chains • Categories of States in Markov Chains • Ergodic Markov Chains

Notes:

- Items marked by * will be covered only if time permits.
- Instead of Markov Chains, Robust Statistical Methods may be covered.

Statement on Academic Integrity: USC seeks to maintain an optimal learning environment. General principles of academic honesty include the concept of respect for the intellectual property of others, the expectation that individual work will be submitted unless otherwise allowed by an instructor, and the obligations both to protect one's own academic work from misuse by others as well as to avoid using another's work as one's own. All students are expected to understand and abide by these principles. SCampus, the Student Guidebook, contains the University Student Conduct Code (see University Governance, Section 11.00), while the recommended sanctions are located in Appendix A. See: http://scampus.usc.edu.

Emergency Preparedness/Course Continuity in a Crisis In case of a declared emergency if travel to campus is not feasible, USC executive leadership will announce an electronic way for instructors to teach students in their residence halls or homes using a combination of Blackboard, teleconferencing, and other technologies. See the university's site on Campus Safety and Emergency Preparedness: http://preparedness.usc.edu

Statement for Students with Disabilities: Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to me (or to TA) as early in the semester as possible. DSP is located in STU 301 and is open 8:30 a.m.5:00 p.m., Monday through Friday. Website: http://sait.usc.edu/academicsupport/centerprograms/dsp/home_index.html

(213) 740-0776 (Phone), (213) 740-6948 (TDD only), (213) 740-8216 (FAX) ability@usc.edu.