Columbia University

Statistics W5701

Fall 2020

Probability and Statistics for Data Science

Course Overview

TABLE OF CONTENTS

OVERVIEW

This is a tentative course structure outline and is subject to change

Document updated May 29, 2020

THE COURSE

This course is an introduction to Probability and Statistics for Data Science.

Students will learn to apply various conceptual and computational techniques useful to tackle problems in statistics. Data Science deals with data but there is more than simply producing beautiful graphs. We will start with data and their simply presentation which naturally leads to the notion of statistics theory and practices. We will first study probability theory, different models and how to estimate parameters and measures. Then we will its role in hypothesis testing. After that we will cover different topics including analysis of variances, goodness of fits and more, depending on the pace of class and time allowed.

This course is designed to be both theoretical and practical. Students are challenged in the following aspects:

- Theoretical
  - Theorems, proofs (and how to write a coherent proof)
  - Geometric intuitions
  - Imagining algorithms, process and its outcome
- Practical
  - Concrete calculations, results interpretations
  - Applying algorithms to numerical examples
- Technology (if time and classroom setting permits)

- Experimenting using Excel possibly with a little bit VBA

This course is particularly interesting for those who want to acquire an understanding of the abstract theory as well its practical applications in different areas with a balance fine-tuned based on the class backgrounds compositions.

PRE-REQUISITES

Mathematics: Basic working knowledge with set theory (union, interaction, complement of sets); functions of one and several variables. Calculus especially differentiation and Integration (single and multiple), linear algebra (working knowledge with vectors and matrices, theory and computations)

Computing Skills: Implementation of computations and simple algorithms with a good working knowledge of Microsoft Excel.

WHO IS IT FOR?

This class is, as of writing, open to Data Science Institute students only.

If you intend to learn serious science or engineering subjects (e.g. Physics, Chemistry, Social Sciences, Mathematics) statistics is a language which you need to master. It also can serve as a good introductory course to abstract measure theory and at the same time sharpen your intuition when solving problems.

TEXTBOOKS

	Name	Author(s)	Details	Comments
1	Introduction to Probability and Statistics for Engineers and Scientists, 5th Edition	Sheldon Ross	eBook ISBN: 9780123948427 Hardcover ISBN: 9780123948113 Imprint: Academic Press/ Elsevier Published Date: 14th August 2014	Required

There will be homework problems assigned based on the textbook above. It would be a good idea to buy / borrow a copy so you have ready access. Statistics department has ordered copies for Columbia University Mathematical Library to place on reserve shelf.

COURSE CONTENTS

We intend to follow closely the textbook with some omissions of non-essential sections, and adding some topics if time permits.

Topics to cover and pace are very tentatively planned as below

Week 1: introduction; data and presentations; descriptive statistics

Week 2: sample spaces, set operations, axioms probability; conditional probability

Week 3: random variables (rv), pmfs/pdfs/cdfs; independence, expected value, joint rvs

Week 4: conditional rvs, variance/covariance/correlations; inequalities, weak law of large numbers (LLN)

Week 5: discrete named distributions (Bernoulli, binomial, hypergeometric, Poisson, etc)

Week 6: continuous named distributions (normal, exponential, uniform,Chi-square); Poisson process

Week 7: parameters, likelihood, maximum likelihood estimation (MLE), central limit theorem (CLT) continuity corrections, general confidence intervals

Week 8: z-scores, z-intervals, prediction intervals, sampling distribution of the mean estimator; t-intervals, binomial intervals

Week 9: hypothesis testing, z-tests under different situations

Week 10: one sample t-tests, two samples, paired, binomial tests

Week 11: simple regression overview, MLE estimates and their distributions; multiple regression optimization

Week 12: weighted least squares, multiple linear regression, one way/two way analysis of variance (ANOVA)

Week 13: Chi-square goodness of fit; review

REQUIRED WORK

Students are required to complete homework assignments. They concern both theoretical and practical aspects of the topics covered in class. For the theoretical section students are required to perform mathematical calculations and proofs. For the practical section (especially later part of the course where statistics formulas might be tedious to evaluate using just pencils and paper) students are required to perform tasks and experiments using Microsoft Excel and maybe some other technology. Having access to a laptop helps.

There will be midterm and final exams. Classroom participation and other factors will also contribute to the final grade. The exact proportions will be determined when class begins.

GRADING

We will determine the percentage contribution of homework, midterm, final exam, class participation towards the final grade when class begins

GRADING POLICY

INTEGRITY

All solutions to the homework, test and exams (take home or otherwise) should be your work. Academic common sense should provide a good guideline and if you are in doubt please consult the instructor. A substantiated violation of the code of integrity and/or academic dishonesty (homework copying for example) may result in serious academic disciplinary action (including but not limited to a failing Grade of this course)

LATE POLICY

Late assignment receives no points. If you still want to hand it in, it should be given directly to the TA.

Late or omitted assignments due to exceptional circumstances (e.g. serious illness with doctor's note or emergency) would be handled on a case-by-case basis.

DISABILITY-RELATED ACADEMIC ACCOMMODATIONS

In order to receive disability-related academic accommodations for this course, students must first be registered with their school Disability Services (DS) office. Detailed information is available online for both the Columbia and Barnard registration processes.

Refer to the appropriate website for information regarding deadlines, disability documentation requirements, and drop-in hours(Columbia)/intake session (Barnard).

For this course, students are not required to have testing forms or accommodation letters signed by faculty. However, students must do the following:

- The Instructor section of the form has already been completed and does not need to be signed by the professor.
- The student must complete the Student section of the form and submit the form to Disability Services.
- Master forms are available in the Disability Services office or online: https://health.columbia.edu/services/testing-accommodations

For further information concerning Disability Services, please contact

Disability Services, Columbia Health

Wien Hall, 1st Floor Suite 108A, 411 W. 116th Street, MC 3714, New York, NY 10027

Phone: 212.854.2388

www.health.columbia.edu/ods

ABOUT THE INSTRUCTOR

Tat Sang Fung holds a Ph.D. in Mathematics from Columbia University in the City of New York (1996). He has taught Differential Equations and Numerical Methods, Basic Mathematics, College Algebra and Analytic Geometry, Advanced Calculus, Linear Algebra. He coauthored the article "BGM numeraire alignment at will" published in Risk International, 2004. He has over 23 years of experience in mathematical finance specializing in financial engineering and quantitative techniques in Treasury and Capital Markets. He has been teaching a graduate level class “Numerical Methods in Finance” at Columbia University every spring semester since 2006.

Tat Sang Fung has been teaching graduate level classes at Columbia since Spring 2006. He can be reached at fts@math.columbia.edu

-end of document-