Statistics 215A, Fall 2018

Instructor: Professor Bin Yu
Lectures: T/Th: 11:00 am -12:30 pm, 204 Wheeler
Discussion: Friday: 9-11 am, 344 Evans

Text books:

- Draft of book "Data Sciene in action" by Bin Yu and Rebecca Bater.
- Statistical models, D. Freedman (required).
- The Elements of statistical learning, Hastie et al (recommended).

Bin's Office Hours in 409 Evans: to be announced.

Phone: 642-2021 (Office), 642-2781 (dept, messages), **email:** binyu@stat.berkeley.edu

Comments, Suggestions, Gripes: in person, email, anonymous notes in my box or under the door. All feedback is welcome.

GSI and office hours: Zoe Vernon (zoe_vernon@berkeley.edu): Office hours to be announced. GSI will be in charge of the discussion sessions and the labs/homework.

Phone: 642-2781 (Stat. Dept. main no.)

Grading:

- 55% assignments (homework and labs)
- 5% class/discussion and participation
- 15% midterm (written exam in class)
- 25% final project

Assignments: There will be 4 or 5 assignments given out on Friday in the discussion session and usually due in two weeks (there will be an announcement if otherwise). **The assignments actually require two weeks of work to satisfactorily complete. So, it is a good idea to start very early.** The assignments contain homework problems and data analysis labs. For the data labs, each student will produce a 12-page (maximum) report presenting a narrative that connects the motivating questions, the analysis conducted and the conclusions drawn. The reports will be made using Knitr/Sweave and the final pdf output should not contain any code whatsoever. Each report will be hosted in a github repository containing both the code and the written report. **No late assignments** will be accepted, *for any reason.*

Course description:

- **Overview** Information technology advances have made it possible to collect huge amounts of data in every walk of our life and beyond. These vast amounts of data have enabled scientists, social scientists, government agencies, and companies to ask increasingly complex questions aimed at understanding the physical and human world, making pubic policies, and improve productivity. However having data alone is not enough; statistics is indispensable in the process of obtaining meaningful answers from collected data. Not only are the common statistical models incredibly powerful, but statistical experimental design itself provides principles and methods to collect data in order to effectively address the questions asked.

The most influential contributions can be made when domain experts (scientists, for example) and statisticians work together to brainstorm and ask questions. These domain experts not only are key to formalizing the ideas, but they also are integral in generating the data. Engaging with the individuals who collected the data in the first place allows the statistician to learn about all the context in which the data lives, and subsequently, to conduct an effective analysis capable of actually answering the question being asked.

- **Collaborative learning in context** This course will demonstrate what is like to be an applied statistician or data scientist in today's data-rich world. We emphasize working with people of domain expertise in order to answer questions outside statistics using data. We illustrate through lectures, class discussions, data labs, and homework assignments for students to gain an understanding of the many steps involved in the iterative process of information extraction or in a data science life cycle. This process includes problem formulation, data generation, data cleaning, data exploration, prediction, *identification of sources of randomness in data*, inference, and interpretation. Three principles of data science: predictability, computability and stability (PCS) are overaching principles throughout the statistical investigation process or the data science life cycle.

  The lectures (and labs) will be based on real-data problems, and students will learn and practice quantitative critical thinking skills, useful statistical concepts and methods in the contexts of these problems. The goal is to illustrate how judgement and common-sense are also crucial to the process of conducting data analysis and drawing conclusions. While the statistical techniques will be introduced through a first-principles approach, students will learn to develop custom techniques in less familiar situations.

  The essential elements of applied statistics are captured in Bin's piece titled "Data Wisdom" (`http://www.odbms.org/2015/04/data-wisdom-for-data-science/`) and students are asked to read the piece after the first lecture.

- **Data lab format and peer-grading** The data labs will be done individually, except for one group lab later in the semester. The goal of writing the lab report is not only to gain data analysis experience, but is also an exercise in communication. We ask that particular attention is given to the writing of the report, as your peers will be reading them. So that the students can learn from one another, the labs will be peer-reviewed. Each student will review 2-3 labs from their peers, and will provide comments based on several criteria including clarity of writing, validity of analysis and informativeness of visualizations. The grade of each lab will be decided by the GSI who will use the student comments as a guide.

  In some sense, the students will act as the "referees" of a paper (the lab report) and the GSI will act as the "associate editor" as if a paper is submitted to a journal. Hence students will obtain refereeing training as well through the course.

- **A full commitment to the class is necessary** Please be aware that this is a heavy-load class. If you are not sure that you can commit, please audit the class instead, since there are many students on the waitlist. Further, because class discussions are an integral part of the course, registered students are required to attend all classes unless permitted by the instructor under justifiable circumstances. *After the first 3 lectures, students are expected to read sections from the draft book BEFORE each lecture so we could devote more class time to group discussions.*

- **Pre-requisites** In this class, we require knowledge of upper division mathematical statistics and probability courses (Stat 134 and 135) at UC Berkeley. In terms of computing, at a minimum you should be comfortable manipulating files in Unix and writing your own functions, manipulating and cleaning data and creating and customizing graphics in R. Ideally students

will already have a basic fluency in the "tidyverse" in R as well as confidence using GitHub. While we will be providing a short introduction to these topics in the labs, students who are entirely unfamiliar with these tools will need to put in some work to ensure that they meet the standards expected of the course.

Tentative list of topics:

- Overview of the class and logistics. Where do applied statistics problems come from? Who are the people involved? First introduction of stability in problem formulation. (0.5 weeks) (Aug. 23)
- Problem formulation and many ways of data generation. Data cleaning (in discussion session). With a high-level question, discovery-driven Exploratory Data Analysis (EDA) with. Numerical summaries and visual descriptions of data. Stability considerations in every step. (2 weeks: Aug. 28, 30, Sept. 4, 6)
- Prediction and assessment. Least squares for prediction and class discussion on prediction. Evaluation or scruitiny of results. (1 week) (Sept. 11 & 13; Bin is on travel on Sept 13 and Zoe will lead a discussion on research papers)
- Stability consideration through appropriate data perturbations. Cross validation. Regularized LS. Interpretable linear models. Model selection, forward selection, L2boosting, Lasso and Sparse PCA via penalized regression. (1.5 week) (Sept. 18 & 20, 25)
- Why do we need the concept and notation of random variable (or the probabilistic language) in data analysis? Randomization through physical constructs: simple and cluster sampling. Neyman-Rubin randomized experiment, natural experiment. LS as adjustment in Neyman-Rubin model. (1.5 weeks, Sept. 28; Oct. 2, 4)
- Linear regression models. Bootstrap. Interpretations. (1.5 weeks: Oct. 10, 12, 13 (discussion session for lecture))
- Bin will be on travel in the week of Oct. 15. One lecture will be for a class discussion on a research paper using linear regression model for interpretable results. The other lecture will be used as Lab time.
- Classification: SVM, Logistic regression, weighted LS for logistic regression computation, and inference in Logistic regression and interpretations of data results. Exponential family. (2 weeks: Oct. 23, 25, 30, Nov. 6)
- Midterm week (Nov. 8 Review; Nov. 13 (Tuesday) Midterm in class)
- GLMs and regularized GLMs (0.5 weeks: Nov. 15)
- Final project assigned on Nov. 16 in the discussion session.
- Hypothesis testing and multiple hypothesis testing (0.5 weeks: Nov. 20)
- Advanced topics (e.g. random projection, Random forests, deep learning) (1 week: Nov. 27, 29)
- No in-class final exam, but there is a final project.

Final Project Due: Dec. 7 (Friday), 5 pm.