Statistics 215A, Fall 2017

Instructor: Professor Bin Yu
Lectures: T/Th: 11:00 am -12:30 pm, 344 (?) Evans
Discussion: Friday: 9-11 am, 332 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: Rebecca Barter (rebeccabarter@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:

- 60% assignments (homework and labs)
- 5% class/discussion and participation
- 15% midterm and written exam
- 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 story 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: 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 ask questions and brainstorm. 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.

This course will demonstrate what is like to work with people of domain expertise in order to answer questions outside statistics using data. At the same time, you will gain an understanding of the many steps involved in the iterative process of information extraction for a variety of purposes including prediction, inference, and interpretation. The lectures (and labs) will be based on real-data problems, and students will learn useful statistical concepts and methods in the contexts of these problems. The goal is to illustrate how judgement and common-sense are 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.

Many of these ideas are captured in my piece titled "Data Wisdom" (`http://www.odbms.org/2015/04/data-wisdom-for-data-science/`).

The class format will be a combination of lecture and discussion groups. 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 a grade based on several criteria including clarity of writing, validity of analysis and informativeness of visualizations. The final grade of each lab will be decided by the GSI who will use the student grades as a guide.

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.**

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 confident 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 technical topics:

In addition to the technical topics listed below, there will be a focus on oral and written communication skills in both the labs and class discussion.

- **Overview of the class. Logistics.** (0.5 weeks) (Aug. 24)
- **Starting with a high-level question, discovery-driven Exploratory Data Analysis (EDA) with a stability consideration. Numerical summary and visual descriptions of data.** (2 weeks: Aug. 29, 31, Sept. 5, 7)
  - Problem and data source: American time use survey
  - Numerical summaries or descriptive statistics: mean, median, mode, standard deviation (variance), interquartile range
  - Visual summaries: histogram, kernel smoother, box-plot, scatter plot, lowess.
  - Dimension reduction through principal component analysis (PCA), and multi-dimensional scaling (MDS)
  - Clustering (K-means, hierarchical clustering). Spectral clustering. Superheat.
- Prediction and assessment. Least squares. Data perturbation (1 week) (Sept. 12 & 14)
- Sources of randomness in data (2 weeks, Sept 19, 21, 26, 28)
  - Problem and data source: Ames housing price data (sampling from a population: density estimation and bias-variance trade-off)
  - Problem and data source: Collected data from the first class? (Neyman-Rubin model)
  - Problem and data source: ? (cluster sampling (EM))
  - Problem and data source: Snow data (natural experiment)
- Problem and data source: Ames Housing. Linear regression models and their interpretations (1 week: Oct. 3, 5; Oct. 6 Lab time as lecture time)
- Bin is on travel in the week of Oct. 9. Oct. 10: Linear regression as natural experiment – paper discussion. Rebecca in charge; Oct. 12 Lecture time as Lab time (swapped from the previous week).
- Problem and data source: paper reading??? LS as adjustment in Neyman-Rubin model (Oct. 17)
- Problem and data source: UCI data set? (enhancer data?) Classification: SVM, Logistic regression, weighted LS for logistic regression computation, and inference in Logistic regression. (1 week: Oct. 17, 19)
- Midterm week (Oct. 24 Review; Oct. 26 Midterm)
- Exponential family and GLMs (1 week: Oct. 31, Nov. 2)
- Problem and data source: Ames housing continues. Multiple hypothesis testing and FDR (Nov. 7)
- Regularization in Regression and GLMs I: PCA-based, Ridge, PLS (Nov. 9)
- Regularization in Regression and GLMs II: model selection, forward selection, L2boosting, Lasso and Sparse PCA via penalized regression. Inference related to Lasso. (1 week: Nov. 14, 16)

- Final project assigned on Nov. 17 in discussion session.
- Dimensionality reduction via random projection (Nov. 21)
- Advanced topics (Nov. 28, 30)
- Rebecca runs lab/discussion session on Dec. 1. No in-class final exam, but there is a final project.

Final Project Due: Dec. 8 (Friday), 5 pm.