MATH 189: EXPLORATORY DATA ANALYSIS AND INFERENCE

JELENA BRADIC
ASSISTANT PROFESSOR OF STATISTICS
DEPARTMENT OF MATHEMATICS, UCSD,
SPRING 2018

LECTURES: TU&TH 6:30PM-7:50PM, SOLIS #107

Course Summary: A study of fundamental theoretical background of statistical methods and their applications, with the emphasis on real data analysis. The course covers the basic concepts of probability, estimation, testing and regression, all explained on a particual case study developed for it. After this course, you will be able to understand and speak the basic language of statistics and appreciate the strengths and limitations of each method and formulate conclusions accordingly. More importantly you will be able to perform statistical data analysis and formulate your conclusions in a statistical report. Finally, and most importantly, you will learn how to use open source software R and Python for an effective data analysis.

Instructor

PROFESSOR JELENA BRADIC
Email: jbradic@ucsd.edu.
Office: 5151 AP&M
Telephone: 858-534-3992
Office Hours: Tuesday/Thursday 10:00am-11:00am or by appointment

Teaching Assistants

HANBO LI	JIAGI GUO	ZIAN WANG
Email: lihanbo90@gmail.com	jig026@ucsd.edu	ziw105@ucsd.edu

Course webpage: All the lecture notes and class info (homework due dates, homework solutions) will be available through the class blog math189.edublogs.org. We will be using Piazza for forum-type discussions throughout the quarter. All posts will be anonymous to the student users but not the TA's and the professor. You will find basic class info on that webpage too.

Course Outline:

1. Introduction to data
2. Case Study 1: mean, variance, kurtosis, quantile-quantile plots, normal approximations
3. Case study 2: survey methodology, simple random sampling, confidence intervals, bootstrap
4. Case study 3: stratified sampling, parametric bootstrap, allocation
5. Case study 4: estimation and testing, goodness of fit tests, information, asymptotic variance
6. Case study 5: contingency tables, experimental design
7. Case study 6: regression, prediction

Time Permissible:

- Case study 7: random forests, classification, svm
- Case study 8: genome wide analysis (GWAS) analysis via simulations

Computation: Software package for this class is flexible. It can be either R, S-plus or excel or any other packages. When needed, I will demonstrate the examples through R. TA's will teach R for all those who are unfamiliar with it and they will cover details of examples used during lectures. TA's will teach Python to all interested as well.

Homework: Homework assignments will be handed out bi-weekly. You will typically have a week in which to work on a homework. Late homework will not be accepted since solutions might be posted online or covered in recitation. Each homework will contribute towards approximately 5% of the final grade.

You are allowed and are encouraged to work with other students on the homework problems, however, verbatim copying of homework is absolutely forbidden. Therefore each student must ultimately contribute his or her own part of the "group" homework to be handed in and graded.

Each homework assignment/project will deal with a real data set.
Projects are evaluated based on all of the following five criterions:

(1) **Composition and Presentation**
(literature review, scientific writing and presentation of logic arguments)
(2) **Basic analysis**
(mentioned during lectures and discussions)
(3) **Graphs and tables**
(must be of the best scientific quality)
(4) **Advanced analysis**
(additional materials prepared outside of topics mentioned during lectures)
(5) **Appendix for technical material**
(all statistics **must** be supported with mathematical/statistical arguments for its usage - consistency, convergence and expectations)

Grading:
The final grade will be made up as follows:

⋆ Homework:	45%	bi-Weekly
⋆ Class & Discussion Attendance	5%	random checks during lectures
⋆ Class Participation	5%	
⋆ Final Exam:	45%	Take home project

Attendance: Attendance of the class lectures and discussions is required. The class covers many conceptual issues and statistical reasoning that cannot be learned otherwise.

Academic integrity: Collaboration and discussions are allowed and are encouraged in this class, but copying or letting others copy your work amounts to plagiarism. Although I expect high academic awareness in this class, if such plagiarism occurs I will take the following action: a grade 0 will be assigned to all involved in the incident where cheating occurred and I will report the incident to the Academic Senate which will then decide appropriate course of action.