STAT 88 FALL 2017

Probability and Mathematical Statistics in Data Science

ABOUT THE COURSE:

In this connector course we will state precisely and prove results discovered in DS8 through working with data. Topics include: total variation distance between discrete distributions; the mean, standard deviation, and tail bounds; correlation, and the derivation of the regression equation; probabilities, random variables, and the Central Limit Theorem; probabilistic models; symmetries in random permutations; prior and posterior distributions, and Bayes' rule.

This course is a *technical depth connector*, and not an *applications connector*. In this course, we will take the concepts that are revealed and explored through working with data in the main course, and develop the statistical or probabilistic foundation for them. For example, students in Data 8 will discover that "for almost all lists of numbers, an overwhelming majority of the entries are within a few standard deviations of the mean." We will state this result, called Chebychev's Inequality, formally in this course, and prove it.

At the end of the course, students will have the mathematical knowledge currently contained in STAT 20, 21, and 25. With the combination of this connector and the main course, the students' knowledge of theoretical statistics will exceed that in the current lower division Statistics courses. In addition, their knowledge will be grounded in working with real data using the computation-intensive techniques of modern data analysis in Data 8.

No more than one semester of calculus is required for the topics studied in Stat 88, and we will try to be as synchronous as possible with the sequence of the Foundations course. It will not be possible to be perfectly synchronous since Data 8 spends a lot of time on material related to computing, and we will cover material that is beyond what is expected mathematically of Data 8. Some weeks we will go deep into concepts already introduced in Data 8 and some weeks we will preview a topic theoretically that will be explored at a later time in the Foundations course.

TIME AND LOCATION:

The lectures will be M from 2-4PM in 10 Evans.

TEXT:

Theory meets Data, by Ani Adhikari. A copy is available on bcourses.

INSTRUCTOR:

Shobhana Murali Stoyanov (shobhana@berkeley.edu), 325 Evans

GSIS:

TBA

OFFICE HOURS:

Shobhana: Tuesdays and Wednesdays from 11:30-1 in Evans 325

DISCUSSION FORUM:

We will be using Piazza for discussions. If you have a question (that is not of a personal nature, but about the material) please post it to the class piazza site.

SECTIONS:

There is no mandatory section for this course. Rather, the GSI will hold office hours.

MIDTERM:

There will be an **in-class midterm** on Monday, October 23 that will cover the material of weeks 1-8.

QUIZZES:

Every class meeting, beginning September 18 - except the week before the midterm, and the day of the midterm - there will be an (approximately) 30 minute long quiz at the start of the session. The last quiz will be longer, and you may complete it at home, turning it in on December 4 at 5pm. The last quiz will count twice as much as the other 8, and cannot be dropped. I will drop your lowest quiz score (out of the first 8 quizzes) while computing your grade.

The topic for each quiz will be the material of the previous week's lecture and the related problems provided to the class for that week. Therefore your overall quiz score will be based on your seven best quizzes of the first eight, and the last quiz.

ACADEMIC INTEGRITY:

Please read the university's statement on academic integrity. You will be held to the UC Berkeley Honor Code.

ACCOMMODATIONS FOR STUDENTS WITH DISABILITIES:

Please see me as soon as possible if you need particular accommodations so that we can work out the necessary arrangements for the quizzes and exams. You are responsible for making sure that we know about your accommodations sufficiently in advance to schedule your proctoring.

GRADING:

The grading scheme will be:

- 7% each for the seven out of 8 quizzes that will be counted, for a total of 49%
- 14% for the last take-home quiz
- 37% for the midterm

No letter grades will be given for the quizzes or for the midterm. Course grades will be based on the distribution of overall scores.