M541b: Introduction to Mathematical Statistics
Fall 2021

Instructor: Dmitrii M. Ostrovskii, email: `dostrovs@usc.edu`, office: KAP 406H
TA/Grader: Jihoon Sohn, email: `jihoon.sohn@usc.edu`, available via zoom
Schedule/Classroom: M/W/F 11-11:50 am, KAP 140
Office hours: M/W 12-2 pm for the instructor, TBA for the grader

Outline of the course. This course is supposed to complement M541a as a graduate-level introduction to Mathematical Statistics. The goal is to understand the key principles behind statistical estimation and testing, and learn how to apply these principles in concrete scenarios.

We shall first briefly recap the main results of M541a: the maximum likelihood principle and the method of moments; sufficiency; the basics of asymptotic theory (consistency, optimality, and asymptotic efficiency). Then we shall cover hypothesis testing, including the concept of p-values and connection with interval estimation. This will take the first half of the course.

In the second half of the course, we shall consider "computationally intensive" methods, such as bootstrap and expectation-maximation (EM). Next we shall consider estimation in linear models, including the case of high dimensionality. Finally, if time permits, we shall cover some advanced topics, such as the Bayesian methods and posterior sampling (including Markov chain Monte Carlo and Metropolis-Hastings), generalized linear models, and robust estimation.

Contact. The best way to contact me is via email (put M541b into the subject field), or in-person during the office hours. We can also fix an appointment, in person or via zoom.

Prerequisites. Working knowledge of point estimation (M541a), Probability at the level of M407/505a, Analysis (M425a,b), and Linear Algebra (M471). Measure-theoretic Real Analysis (M525a) and Probability as per M507a are not required but will be useful.

Covered topics. (Some topics are optional due to limited time; they are marked with "?".)

1. <u>Background:</u>
   - Brief review of the probability "vocabulary": probability spaces, random variables, expectations, moment-generating function, basic inequalities (Jensen, Markov, Chebyshev, and Chernoff); conditional expectations and independence.
   - Brief recap of the key results in point estimation: consistency, laws of large numbers, central limit theorem, maximum likelihood principle and sufficiency.
2. <u>Hypothesis testing:</u>
   - Problem statement and terminology.
   - Likelihood ratio test (LRT) for two simple hypotheses; the Neyman-Pearson lemma.
   - Testing composite hypotheses: uniformly most powerful (UMP) tests and the Karlin-Rubin theorem; the Neyman structure and conditioning method.

- Permutation and rank tests; invariance.
- Asymptotics of the power function and Pittman's efficiency.
- LRT and Wilks' theorem.
- Connection between interval estimation and hypotheses testing.

3. "Data-intensive" inference:
   - Bootstrap and Jackknife.
   - Expectation-maximization (EM) algorithm.
   - ? "Computation-friendly" Bayesian inference: Markov chain Monte Carlo and Metropolis-Hastings algorithms.
4. Topics in modern statistics:
   - Estimation in linear models; finite-sample results.
   - ? Introduction to generalized linear models and logistic regression.
   - ? Robust estimation; median-of-means methodology.

Homework assignments. There will be 5 to 6 homework assignments, designed in a way to help you understanding the material covered in the course. Here is some advice:

- Always start to work on homework assignments as early as possible. (Well, at least *look* in there, simply in order to estimate the difficulty of what you'll have to deal with!)
- It's okay to work in a small group. However, always try to solve the problem yourself first—especially those of you who do research or planning to pursue it. If you are blocked, use the office hours to ask a question. Taking notes during a discussion is a useful practice.
- Always write the solution on your own; suspected plagiatrism will receive 0 credit, and further sanctions in case of a relapse. I recommend typesetting solutions in LaTeX; writing by hand is possible, but please make sure that it is readable—show respect to your grader!

Exams. The midterm will take place during a lecture somewhere after the Fall break (most likely, in the last week of October), in our regular classroom. The (comprehensive, two-hour) final exam will take place on the week of December 13 in our regular classroom.

Grading. 45%/25%/30% for homeworks/midterm/final exam; may be slightly adjusted later.

Attendance. I shall not keep track of attendance: I rely on your motivation. That said, note that the class is currently in-person, and the room is large enough to maintain social distancing.

Literature. We shall not follow a single textbook; that said, *Statistical Inference* by Casella and Berger (2nd ed.) is the main reference for the first half of the course; other useful references are: *A Course in Large Sample Theory* by Ferguson, *Asymptotic Statistics* by Van der Vaart, *Testing Statistical Hypotheses* by Lehman and Romano. For the second part of the course, I recommend section 2 of `https://web.stanford.edu/class/cs229t/notes.pdf` to get on grips with linear models. A comprehensive introduction to modern statistical theory from a parametric viewpoint can be found in *Basics of Modern Mathematical Statistics* by Spokoiny and Dickhaus; however, it is way beyond the scope of this course.