36-743,36-744: (Some) Statistical methods for reproducibility

2 minis (Aug 26 to Oct 17, Oct 21 to Dec 5), 6 credits each
Fall 2019, Syllabus

August 22, 2019

1 Basic Course Information

Instructor Aaditya Ramdas, `aramdas@stat.cmu.edu` [Office hours: on demand, via email]

Time: 10:30-11:50am
Location: Baker Hall 150
Exceptions: No class on Sep 2 (Labor day), and Nov 25, 27 (Thanksgiving). See the academic calendar.

Website See http://www.stat.cmu.edu/∼aramdas/reproducibilty19 for basic course material.
Announcements All announcements will be made on the above course website.

Participants This course can be credited by PhD students with good mathematical background, but it can be audited by anyone who is curious about the topic. Students who want to sit through the course must officially audit.

Prerequisites The first mini is a prerequisite for the second mini. There are no formal prerequisites for the first mini, but non-PhD students must email the instructor for permission to enroll in the course. Enrolled students are expected to have completed at least one intermediate statistics course (like 10-705 at CMU). In particular, students must be familiar with hypothesis testing and confidence intervals.

Textbook There is no official textbook. References to papers will be mentioned in class or on the website.

2 Course Description

In the past decade or so, there have been serious concerns about an alleged "reproducibility crisis" in the sciences (typically the human sciences, like psychology, sociology, etc). While many factors have been identified, the *statistical* aspects are central to the debate, and recognition of this fact has led to the proposal of many new ways to model the problem and new solutions to address these problems. Indeed, when analyzing high-dimensional data in the modern day and age, one is often trying to find a needle in a haystack. The concept of "statistical significance" is especially hurt by the selection bias caused by the multiplicity issue: there are multiple leads that the data suggest and we often follow only the most promising ones. Identifying, quantifying and eliminating this selection bias is critical in our long-term quest for reproducibility.

In this two-mini course, we will identify and discuss several statistical methods that have been developed over several decades to improve reproducibility and replicability. Specifically, we will identify several sources of bias (usually selection bias, but also publication and confirmation biases) that could arise in subtle ways during experimentation and data analysis. These will be exemplified by caricatured examples, such as the *jellybeans-acne story*, the *garden of forking paths*, the *peeking problem*, *p-hacking and interval hacking*, the *Kaggle leaderboard paradox* and so on.

We will then examine a variety of state-of-the-art solutions that have been proposed to deal with these problems, ranging from classical multiple testing, to sequential confidence sequences, to closed testing, the reusable holdout, simultaneous inference, selective and post-selection inference and differential privacy. We will begin with some classical solutions that date back to the 1950s, but the majority of methods that we will touch upon are modern developments within the last decade.

3 Graded Components

(dates are tentative)

Mini 1 The grade will be based on one long homework (40%) whose questions will be progressively released, and a course project (10% for proposal, 50% for final report). Homework 1 will be due on Oct 5. The project proposal (approximately one page) is due on Sep 13, and the final report due on Oct 16 (Tue, when class is cancelled due to university policy).

Mini 2 40% of the grade will be based on one long homework whose questions will be progressively released. For the rest, students will preferably develop a significant extension of the course project in Mini 1 (50% for final report, 10% for a short in-class presentation). Homework 2 will be due on Nov 15. The in-class presentations will occur on Dec 4. The project final report (approximately ten pages) will be due on Dec 5.

Projects There are a wide variety of options available for course projects. Examples include:

- (Low risk) You can survey an area of the literature (covered in a textbook, or a set of advanced papers) that is related to the course, and is complementary to what is covered in class.
- (Medium-low risk) You can create a set of graphs, plots, or interactive figures, which allow the user to visualize several of the methods covered in the course. For inspiration, check out distill.pub, and specifically, a paper on why momentum works.
- (Medium-high risk) You can apply the contents of the class to your own research problem, for example by improving the guarantees you had achieved by using older tools, or by extending the analysis of your problem to hold for new settings.
- (High risk) If you are mathematically very mature, and want to work on a new research problem in this area from scratch, talk to the instructor privately in person.

Other ideas for course projects are also welcome. Grades will ultimately be awarded based on the instructor's judgment of the amount of work completed in the project, with some amount of subjective discounting for the risk of the project taken up. Students will be evaluate on both writing (project reports) and speaking (project presentations).

4 Learning Objectives

Mini 1 Upon successful completion of the first mini, the student will be able to

- Explain the jelly-beans acne case study to non-statisticians.
- Assess the pros and cons between various error rates in multiple testing.
- Recognize interval hacking and p-hacking, and understand their duality.
- Quote the assumptions under which different multiple testing methods work.
- Connect global null testing with familywise control using closed testing.
- Describe the intuition behind the knockoff method for variable selection.

Mini 2 Upon successful completion of the second mini, the student will be able to

- Recognize the peeking problem, why it arises and how to correct for it.
- Explain why the garden of forking paths hurts reproducibility.
- Describe the Kaggle leaderboard problem and an outline of its solution.
- Translate the outputs of a closed testing procedure into post-hoc bounds.
- Identify situations where doubly-sequential inference applies.

5 Approximate Schedule

Mini 1 (16 lectures)

- L1 (Aug 26): Motivation, syllabus, policies, discussion of projects
- L2 (Aug 28): The jellybeans-acne cartoon, multiple testing (coherence, consonance, error rates)
- L3 (Sep 2): no class (Labor day)
- L4 (Sep 4): Global null testing (Fisher, Stouffer, Simes, higher criticism)
- L5 (Sep 9): FWER control (Bonferroni, Sidak, Holm, closed testing)
- L6 (Sep 11): The principles of closure, partitioning and sequential rejection
- L7 (Sep 16): FDR (Benjamini-Hochberg and variants)
- L8 (Sep 18): Three different proofs of the BH procedure
- L9 (Sep 23): The Empirical Bayes perspective and local FDR methods
- L10 (Sep 25): From p-hacking to interval-hacking (false coverage rate)
- L11 (Sep 30): (spillover class or TBD)
- L12 (Oct 2): Variable selection in regression (knockoffs and model-X knockoffs)
- L13 (Oct 7): Interactive multiple testing with a human-in-the-loop (data carving)
- L14 (Oct 9): Online multiple testing
- L15 (Oct 14): Reading day: historical perspectives, recap and summary
- L16 (Oct 16): In-class exam: project midway presentations

Mini 2 (15 lectures)

- L17 (Oct 21): The peeking problem: power-posing, and sequential type-1 error
- L18 (Oct 23): The LIL (law of the iterated logarithm) and nonasymptotic variants, confidence sequences
- L19 (Oct 28): The doubly-sequential setting: online control of error rates
- L20 (Oct 30): The lasso post-selection inference problem: conditional inference using the polyhedral lemma
- L21 (Nov 4): Simultaneous inference using the PoSI framework
- L22 (Nov 6): Post-hoc, exploratory multiple testing: closure principle to the rescue
- L23 (Nov 11): Overfitting by adaptively querying a database (differential privacy)
- L24 (Nov 13): Overfitting the test set and the Kaggle leaderboard paradox (reusable holdout)
- L25 (Nov 18): The garden of forking paths
- L26 (Nov 20): A generalized-Kaggle solution by revealing only single bits
- L27 (Nov 25): no class (Thanksgiving)
- L28 (Nov 27): no class (Thanksgiving)
- L29 (Dec 2): Selection bias in multi-armed bandits
- L30 (Dec 4): (Project presentation?) Fresh data: the universal solution, recap and summary

6 Course policies

6.1 Attendance

On-time attendance is expected and highly recommended. Every research study on this topic that I have read concludes that academic performance is negatively affected by not showing up to class.

6.2 Collaboration

Discussion of class material is heavily encouraged. Additionally,

- After submission of a homework, discussion of answers is encouraged.
- Before submission of a homework, reasonable verbal discussion of homeworks is allowed. An example of unreasonable verbal discussion: one person reciting formulae orally while another one writes them down. Written discussion (in any form) is permitted in groups smaller than 3 (or in rare exceptions 4) students.
- No matter what discussions have taken place, every homework and cheat sheet and mini-project and self-test (in its entirety) must be written up or coded up alone.

6.3 Academic Integrity

I have a zero tolerance policy for violation of class policies. If you are in any doubt whether a form of collaboration or obtaining solutions is permitted, please clarify it with me before proceeding.

- For each question on each homework, collaborators for that question must be acknowledged. Copying solutions from the internet is explicitly disallowed. You may search for material to help you understand a concept better, but be sure to create your own final solution. If you happen to use results from Wikipedia or textbooks, you must cite the source and are expected to completely understand the result you are citing. However, it is disallowed to copy solutions to exercises from elsewhere on the internet, like other courses or papers. When quoting text from a textbook, paper or website, use the `\begin{quote}` option in Latex.
- Any deviation from the rules will be dealt with according to the severity of the case. For example: evidence of written discussion in a larger group than 3-4 will result in points earned for that question becoming zero for all those relevant students; blindly copying one solution from someone else or online will result in the maximum points that can be earned for that homework becoming zero (maximum eligible grade becomes B); repeat occurrences will result in a failing grade for the course.
- In line with university policy, all instances of cheating/plagiarism will be reported to your academic advisor and the dean of student affairs. See the university policy on academic integrity.

6.4 Use of Mobile Devices and Laptops in Class

The use of mobiles and laptops in class is heavily discouraged. Learning research shows that unexpected noises or movement automatically divert and capture people's attention, meaning that you are affecting everyone's learning experience. For this reason, I ask you turn off your mobile devices and close your laptops during class. If you must use your laptop or mobile, make sure you are sitting at the back of the class.

6.5 Late Assignments

Every student is allowed a total of 2 late days per mini. Beyond that, the maximum earnable points for that assignment will drop by 20% per day.

7 Additional information

7.1 Global Communication Center

For assistance with the written or oral communication assignments in this class, visit the Global Communication Center (GCC). GCC tutors can provide instruction on a range of communication topics and can help you improve your papers, presentations, and job application documents. The GCC is a free service, open to all students, and located in Hunt Library. You can make tutoring appointments directly on the GCC website: http://www.cmu.edu/gcc. You may also browse the GCC website to find out about communication workshops offered throughout the academic year.

7.2 Accommodations for Students with Disabilities

If you have a disability and are registered with the Office of Disability Resources, I encourage you to use their online system to notify me of your accommodations and discuss your needs with me as early in the semester as possible. I will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, I encourage you to contact them at access@andrew.cmu.edu.

7.3 Statement of Support for Students' Health & Well-being

Take care of yourself. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress.

If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit http://www.cmu.edu/counseling/. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.

If you or someone you know is feeling suicidal or in danger of self-harm, call someone immediately, day or night (CaPS: 412-268-2922, Resolve Crisis Network: 888-796-8226). If the situation is life threatening, call the police (On-campus CMU Police: 412-268-2323, Off-campus Police: 911).