6.S098 Introduction to Statistical Hypothesis Testing

IAP 2024 - 6 units (2-0-4)

1 Course Description

Statistical hypothesis testing is used in many fields to evaluate whether a result is statistically significant. This course will cover how to formulate a statistical hypothesis and draw conclusions from a test procedure. Starting with t-tests—the primary workhorse of hypothesis testing, students will then become well-versed in common approaches to test for differences in means, goodness-of-fit, independence, equality of distributions, and significance in linear regressions. Two essential topics students will also learn are correcting for multiple hypotheses and power to detect an alternative hypothesis. Each class session will include a lecture and a problem-solving session. Problem sets will include theoretical and computational exercises, and the project will be an open-ended data analysis. The objective of this course is for students to gain a clear understanding of when to use hypothesis tests, how to select a test and perform it correctly, and how to avoid common pitfalls when interpreting the results.

These are some reasons to take this course:

- To get a head start for an undergraduate-level statistics course, such as 6.3720, 18.650, 7.093, 9.07, 16.09, 15.075, or IDS.012
- To prepare for a data science job
- To learn how to analyze data from experiments in many fields, including clinical science, finance, and tech
- To have fun in a class focused on problem solving!

This course was inspired by the hypothesis testing unit in 6.3720/6.3722 and borrows some material from Professor Polyanskiy's lectures. As such, this course covers roughly the same concepts but at a more introductory level and with more in-class practice. Students are encouraged to take 6.3720/6.3722 after to learn more about statistical data analysis.

2 Staff

Instructor: Christina Ji, cji@mit.edu
Lab assistant: Mauricio Barba
Faculty Advisor: Yury Polyanskiy

3 Logistics

Class sessions will be held Monday and Wednesday 2-4pm in 24-121. The schedule on the next page lists the 7 topics. Each session will have a lecture followed by a problem solving component. Attendance is required to complete exercises at each session. Please see the course philosophy and grading sections for how additional time and absences may be handled.

Listeners will be required to attend class and participate in the in-class problem solving sessions but will not have assignments graded. Course material will be restricted to students who are in class.

2 problem sets and 1 project will be due on Friday at 4pm in weeks 2, 3, and 4. They must be submitted individually on Gradescope. Collaboration in the form of joint problem solving with a few classmates is encouraged. However, the write-up and code must be your own. Please list your collaborators at the top of your submission. AI assistance, such as Chat-GPT, may be used but cannot take the place of writing and

understanding your own code or solution.

Office hours: Thursday 2-4pm in 26-210
Canvas: `https://canvas.mit.edu/courses/24087`
Piazza: `https://piazza.com/mit/spring2024/6s098`
Gradescope: `https://www.gradescope.com/courses/693455`
Absence tracker form: `https://forms.gle/qqDpXdabvnxnYYQk7`
Extension request form: `https://forms.gle/GtDKU6z14QNKv7uC6`

4 Prior Knowledge

The following prior knowledge will be helpful. Students are encouraged to ask questions about these areas during problem solving sessions and office hours.

- Basics of probability: The concepts that need to be understood are covered in section 3 of lecture 1 notes. If you would like another refresher, please refer to the corresponding sections in chapters 1-3 of the All of Statistics textbook, which will be available on Canvas for the class.
- Familiarity with a statistical programming language, such as Python, R, MATLAB, SAS, or Julia: Students may use any programming language to complete the problems. Lectures will introduce packages in Python. Course staff may provide some help with R and MATLAB. Support for other languages will be limited.

5 Course philosophy

The goal of this course is for students to develop their statistical reasoning skills and understand hypothesis testing well enough that they will be able to understand how to carry out the correct testing procedure for their own work. The lectures will introduce common hypothesis tests and good statistical practices.

The in-class problem solving sessions are designed to be a learning environment where students can receive more guidance from classmates and course staff. As such, working on the exercises in class is an important component of the class. To reduce time pressure, if you are unable to complete the exercises in class, please give your name to the instructor at the end of the session. Students who have attempted to complete the exercises in class may email their submissions to the instructor by 4pm the following day with no late penalty.

The problem sets will include longer exercises that students can spend more time working through. For the project, students will be given a real-world dataset and open-ended questions to answer with the data. Students will present their results in a short written report.

6 Grading

This course is graded on pass/fail and will count for 6 units at the undergraduate level.

Category	Individual weight	Count	Total weight
Problem solving sessions	10	7	70
Problem sets	40	2	80
Project	40	1	40
Participation			10
Total			200

For at most 2 sessions, if you are unable to attend, please submit the absence tracker form before 2pm the day of class. If you are taking the class for credit, please email your completed exercises to the instructor before 4pm 1 day after the session. The recording, notes, and exercises will be posted on Canvas after each session. Additional absences or late submissions will result in a grade of 0 for that day's exercises. For listeners, additional absences will result in loss of access to course material. The intent of this policy is to help students keep up with the class.

To allow for some flexibility, a 3-day extension may be requested for one pset by submitting the extension request form prior to the original deadline. That is, the extension must be requested prior to Friday 4pm, and the pset may be submitted by 4pm the following Monday on Gradescope. If an extension is requested for pset 1, it cannot be moved to pset 2. Extensions are not allowed for the project since it is due on the last day of IAP. Late assignment submissions will result in a grade of 0.

For any assignment, extensions due to illness or other reasons may be requested with support from Student Support Services or Grad Support. The same policy applies to absences.

7 Schedule

Date	Session	Pset
Mon 1/8	S1: Hypotheses, probability review, p-values, one-sample t-test	Pset 1 released
Wed 1/10	S2: More tests for means: z-tests and two-sample t-tests, Type I & II errors	
Mon 1/15	MLK - no class	Pset 2 released
Wed 1/17	S3: Testing goodness-of-fit & independence: Kolmogorov-Smirnov (KS), generalized likelihood ratio test (GLRT)	
Fri 1/19		Pset 1 due at 4pm
Mon 1/22	S4: Testing equality of distributions: two-sample KS, Wilcoxon rank-sum, permutation	Project released
Wed 1/24	S5: Testing for regression: Wald, F-test	
Fri 1/26		Pset 2 due at 4pm
Mon 1/29	S6: Multiple hypothesis correction: Bonferroni, Benjamini-Hochberg	
Wed 1/31	S7: Power: Neyman-Pearson, likelihood ratio test, sample size calculations	
Fri 2/2		Project due at 4pm