California Institute of Technology
Department of Computing + Mathematical Sciences

IDS/ACM/CS 157 Statistical Inference

Spring 2024

Lectures:	Tue & Thu 9:00am-10:25am in Baxter Lecture Hall
Instructor:	Konstantin (Kostia) Zuev
Office:	114 Annenberg
Email:	kostia@caltech.edu (please include "157" in the subject line)
Office Hour:	Thu 1pm-2pm, or by appointment (please, send an email to schedule)
Head TA:	Harsh Gandhi (hgandhi@caltech.edu)
TAs and OHs:	https://piazza.com/caltech/spring2024/idsacmcs157/staff You are welcome to attend as many office hours as you like.

Course Goals

Statistical Inference is a branch of Mathematical Engineering that studies ways of extracting reliable information from limited data for learning, prediction, and decision making in the presence of uncertainty. The main goals of this course are:

- Develop statistical thinking and intuitive feel for the subject,
- Introduce the most fundamental ideas, concepts, and methods of Statistical Inference, and
- Explain how and why they work, and when they don't.

If you do well in the class, you should be able to read (and understand) most contemporary papers that use statistical inference and perform statistical analysis yourself.

Prerequisites

This is an introductory course on statistical inference. No prior knowledge of statistics is assumed. However, a solid understanding of Probability is required. Ma 3 or ACM/EE/IDS 116 (or equivalent) is a "hard" prerequisite. A key part of the course is problem sets, where you will get experience in using the learned methods and models in applications via simulations in MATLAB. So, some familiarity with MATLAB (and programming in general) is desired, but this is a "soft" prerequisite: MATLAB is easy to pick up on the fly, especially for the purposes of this course.

Textbooks

There is not a single book the course is based on (I am writing the one!). Good news: I will provide comprehensive lecture notes. After each lecture, I will be uploading the corresponding notes to the course Piazza page together with supplementary materials for further reading. The course was developed using the following books, which can be used as supplementary (but not required) textbooks:

- G. Casella & R.L. Berger, *Statistical Inference,* 2002.
- A.C. Davison, *Statistical Models*, 2003.
- L.A. Wasserman, *A Concise Course in Statistical Inference*, 2005.
- M. Lavine, *Introduction to Statistical Thought*, 2013.
- S.L. Lohr, *Sampling: Design and Analysis*, 2010.
- D.C. Montgomery, E.A. Peck, & G.G. Vining, *Introduction to Linear Regression Analysis,* 2006.
- D. Nolan & T. Speed, *Stat Labs: Mathematical Statistics Through Applications*, 2000.
- S. Weisberg, *Applied Linear Regression*, 2005.

Course Plan

The following is a tentative outline of the topics that I plan to cover this term.

Week 1	Introduction, Summarizing Data
Week 2	Classical Statistics: Fundamentals of Survey Sampling
Week 3	Modeling and Inference: A Big Picture, Statistical Functionals
Week 4	Jackknife, Bootstrap, Method of Moments
Week 5	Maximum Likelihood Estimation
Week 6	Hypothesis Testing: General Framework, p-Values
Week 7	The Wald, t-, Permutation, and Likelihood Ratio tests
Week 8	Regression Function, Scatterplots, Simple Linear Regression, Ordinary Least Squares
Week 9	Properties of OLS Estimates, Interval Estimation, Prediction, Graphical Residual Analysis

Grading

Your final grade will be based on your total score. Your total score is a weighted average of Problem Sets (60%), Midterm Exam (20%), and Final Exam (20%). You can increase your total score by up to 5% if you participate actively in Piazza discussions in the Q&A section[1]. Every answer submitted before TAs or instructor answer, which is later endorsed as "good answer" by TAs or instructor, gets 1% of the total score. There are no fixed thresholds for grades, but if your total score is 90% (80%, 70%, 60%), you are guaranteed at least "A" ("B", "C", "D").

Problem Sets

There will be six Problem Sets. Problems (and solutions) will be posted on Piazza. For assignment and due dates see "Important Dates" below. Late submissions will not be accepted, but the Problem Set with the lowest score will be dropped and not counted toward your total score. Submitting wrong files or files in a wrong format is considered as a late submission. Extensions may be granted for academic, personal, or medical reasons. For extensions, please email the Head TA.

Exams

There will be two exams: Midterm (based on Lectures 1-9) and Final (based on Lectures 10-16). The Head TA will provide a review session before each exam. Both exams are take-home, self-timed, and "open-book": you can use notes and books, but not your classmates and the Internet. You can use your computer only as a typing device and for basic arithmetic operations. No other electronic devices are permitted.

Collaboration Policy

A detailed collaboration policy is given on the course website at:
http://www.its.caltech.edu/~zuev/teaching/2024Spring/CollaborationIDS157.pdf
In general, collaboration is encouraged everywhere except for the exams. Let's help each other and learn together! If you get stuck with a homework problem, I encourage you to discuss it with other students (offline or online on Piazza). But remember that you will have to prepare and submit your solution by yourself. No collaboration is allowed on the exams.

Important Dates (All times are Pacific Times)

	Available	Due
Problem Set 1	1pm Thu, Apr 11	9pm Thu, Apr 18
Problem Set 2	1pm Thu, Apr 18	9pm Thu, Apr 25
Problem Set 3	1pm Thu, Apr 25	9pm Thu, May 02
Head TA Review	9am Thu, May 02	
Midterm Exam	1pm Thu, May 02	9pm Tue, May 07
Problem Set 4	1pm Tue, May 07	9pm Tue, May 14
Problem Set 5	1pm Tue, May 14	9pm Tue, May 21
Problem Set 6	1pm Tue, May 21	9pm Tue, May 28
Head TA Review	9am Thu, May 30	
Final Exam	1pm Thu, May 30	9pm Thu, June 06

Websites

- Course website:
  http://www.its.caltech.edu/~zuev/teaching/2024Spring/IDS157.htm
- Lecture notes, further reading materials, problem sets, exams, data sets, solutions, announcements, and class discussions will be managed via Piazza, which is designed such that you can get a quick help from your classmates, TA(s), and instructor. Instead of emailing questions to the teaching staff, I encourage you to post your questions on Piazza because
  - You will get the answers faster
  - Your classmates may also benefit from seeing the answers to your questions.

  Here is the Piazza page:
  http://www.piazza.com/caltech/spring2024/idsacmcs157/home

[1] If you are interested in being a TA next year, try to be active on Piazza and help other students by answering their questions.

- Problem sets and exams will be graded via Gradescope.
  - If you are a **registered student**, you will be enrolled on Gradescope by the end of the 1st week of classes, and you will receive a notification from Gradescope about your enrollment.
    - Please make sure that the email that you use on Gradescope is your official Caltech email.
  - If you are a **registered student**, but have not been enrolled on Gradescope by the end of the 1st week of classes, please email the Head TA as soon as possible and ask to enroll you to Gradescope. Your absence on Gradescope means that, according to my records, you are not registered for the course.
  - If you want just to **audit the course**, it is fine, you will have access to Piazza and all course materials there (please email me and I will enroll you on Piazza), but you will not have access to Gradescope and your submissions will not be graded. If you audit the course this term, you should not register for the course in the future.

  To submit your solution via Gradescope, your need to create a single PDF (not images) that contains the whole solution (for example, by scanning your solution), and then upload it to Gradescope. Here are some useful links:
  - Scanning on a mobile device: https://help.gradescope.com/article/0chl25eed3
  - Submitting an assignment: https://help.gradescope.com/article/ccbpppziu9

  Should you have any questions regarding Gradescope, please ask on Piazza: we will have many experts there.

Suggested Study Process

To get the most out of IDS 157, here is my suggestion on the study process:

- Attend Lectures, focus on understanding the big picture of what is going on.
- Review Lecture Notes (ideally on the same day they are released), make sure that everything is clear.
- If something is not clear, ask on Piazza, and help your classmates by answering their questions.
- After each Lecture, very briefly summarize my notes in Your Own Notes, extract the essence.
- Start working on each Problem Set on the same day it is released (or as soon as possible after that).
- Aim at finishing each Problem Set and Exam at least 1 day before they are due.
- If you get stuck with a problem, ask for hints on Piazza (unless it is an exam problem, and then you are screwed ;-))

Keep in Mind

My goal is to help you understand and learn the material. Understanding is a creative and time- and effort-consuming process. If you don't understand something, please ask to me. If you are struggling with balancing the workload, please talk to me. If you have any concerns, please let me know. Keep in mind that I am here to help.

Honor Code

"No member of the Caltech community shall take unfair advantage of any other member of the Caltech community."