Harvard Stat 111: Introduction to Statistical Inference
Spring 2024

Professors: Joe Blitzstein (blitz@g.harvard.edu) and Neil Shephard (shephard@fas.harvard.edu).

Lectures: TuTh 1:30 pm – 2:45 pm in Science Center Hall C.

Sections and Office Hours: There will be weekly sections and office hours, with many to choose from. You are expected to attend section once a week. Attending the same section regularly is recommended, but you have the flexibility to attend any section each week.

Course Webpage: https://canvas.harvard.edu/courses/126187

Description: Statistics centers around three main goals: *exploring* and *describing* data and a phenomenon of interest, *predicting* one variable using another variable, and drawing *causal* conclusions about the effect of one variable on another. We will delve into principles and methods for all three of these goals, focusing on *statistical inference*, which is a framework for *modeling* phenomena, *learning* from data, and making principled *decisions* under uncertainty. Due to the complexity of these problems, a three-pronged approach is often needed, combining *theory*, *simulation*, and *data*.

Prerequisites: Stat 110 concepts and techniques will be very important throughout the course. Single variable calculus will also be needed extensively.

References: Handouts will be provided in PDF format (draft chapters of a book that we are writing). There is no need to purchase any books for the course.

Computing: Some homework problems will require coding to run simulations and/or do statistical computations. We recommend using the R language via the RStudio interface, using R Markdown or Quarto to weave together text, images, math, and code. R is freely available, in contrast to some proprietary statistical software that can cost thousands of dollars.

No prior knowledge of R or coding is required. There will be R bootcamps at the start of the semester to help you get up and running with R if you haven't used it before. You can use another free, open source environment such as Python or Julia instead if you prefer (but not something like Excel!). However, far more resources and support will be available for R than for other languages.

Discussion Forum: We will use Ed for online discussions. Ed can be used for questions and discussions about the homework, and material from class, sections, and the notes. Ed can also be used for meeting people in the class, finding study groups, seeing section and office hour announcements, and various logistical and social conversations. See this link for a quick overview.

When discussing homework questions on Ed, please avoid posting large portions of your solution. General discussions of the relevant concepts and strategies for approaching a problem are fine, as are questions meant to help you get unstuck with a particular step in a solution. If you would like to post a major part of your work as part of a question, you can make the post private (in which case it will only be visible to the teaching staff), but please do that sparingly so that the whole class can benefit as much as possible from the discussions. Also, please *search* before posting a question (someone may have already asked a similar question, in which case it is usually better to continue that thread than to start a new thread), and when you start a new thread use the correct category (there is a designated category for each homework problem, e.g., HW 2.3 is for Problem 3 of Homework 2) and include a short description of your question in the title.

Please set your Canvas and Ed *notifications* carefully according to your preferences. See here for information about Canvas notifications. In Ed, click the Account button (on the upper right) and go to Settings; also, click the Watch button on threads that you want to stay notified about.

Grading: Grades will be based on a weighted average of scores from homework, an in-class 75-minute midterm exam on **Tuesday, March 5**, and an in-person final exam on **Thursday, May 2 from 2:00 pm to 5:00 pm**.

Let h, m, and f be your homework average, midterm score, and final exam score, respectively, each scaled out of 100. Then your overall score for the course is given by

$$s = \max(0.45 \cdot h + 0.20 \cdot m + 0.35 \cdot f,\ \ 0.45 \cdot h + 0.10 \cdot m + 0.45 \cdot f).$$

Attendance and Participation: Attending and participating in at least one lecture and/or section per week is expected. For students whose overall score is near but below a grade cutoff, we will take attendance and participation into account. This will only be done in a positive manner, i.e., some grades slightly *below* a cutoff may be bumped up due to strong attendance and participation, but no grades that are *above* a cutoff will be bumped down due to poor attendance and participation.

Homework: Since actively solving problems is crucial in learning statistics, there will be weekly problem sets, normally due on **Fridays at 5:00 pm**. There will be no homework due the week of the midterm or the week of Spring Break. Homework must be submitted via Gradescope; no submissions on paper or by email will be accepted.

Unless otherwise specified, please show your work, simplify fully, and give clear, careful justifications for your answers (using *words and sentences* to explain your logic, in addition to the relevant mathematical expressions and/or code). When code is used, both the *outputs* (such as plots and summary statistics) and the code itself must be included in your submission. Using R Markdown or Quarto is a convenient, useful way to keep things together. It's also fine to handwrite and scan some of your work, but you should later merge this work with your code and outputs of your code so that it is together for each problem. If you handwrite some of your work, you should start each problem on a new page, to make it easier to combine everything in the correct order (so work for a particular problem appears on consecutive pages).

To help with various circumstances (expected or unexpected), your lowest homework score will be dropped. Additionally, you can have **four** extensions until Monday at 5:00 pm (the Monday of the week after the Friday deadline). No further extensions will be granted.

You do *not* need to explicitly request to drop a homework or get a Monday extension; both are automatic. If you do not submit a homework, the score on that homework will be 0 and a Monday extension will *not* be incurred. Homework is normally due Fridays at 5:00 pm (Eastern). There

is a 15 minute grace period, after which a Monday extension will be applied to a submission (automatically and with no penalty), if you have not already used four Monday extensions.

Note that which homework is dropped is based only on your homework *scores* (so this is determined at the end of the semester, by taking the minimum of all your homework scores for the semester). In contrast, Monday extensions are based only on *times* of submission (so it's determined on the fly for a particular homework whether it has incurred a Monday extension). If you do not submit a particular assignment, then it will get a score of 0 and will *not* incur a Monday assignment.

Homework Academic Integrity Policy: You are welcome to discuss homework problems with others, but *you must write up your solutions yourself and in your own words*. For problems where coding is required, you must write and run your own code. Copying portions of someone else's work, or just making trivial changes for the sake of not copying verbatim, is not acceptable. We highly recommend starting problem sets early enough so that you have time to work hard on the problems on your own first, before discussing them with friends/collaborators.

Referring to previous years' homework solutions or cheating websites such as Chegg is not allowed.

Using generative AI tools such as ChatGPT to help with homework *is* allowed, but not recommended except possibly as a helpful but unreliable coding assistant. In our own testing of ChatGPT, it could do some Stat 111 problems correctly but also made many mistakes. Furthermore, working hard on the homework problems is crucial for learning the material and preparing for the exams, so even if ChatGPT did excellent work on the Stat 111 homeworks, relying on it too much would likely be harmful for overall understanding and for performance on the exams.

In any case, your solutions must reflect your own understanding of the material, explained in your own way, rather than being copied from any other source.

Additional Policies:

1. We allow students to take the course pass/fail, though if you are interested in that option you should discuss it with your advisor, and check whether or not it this option is feasible for your plan of study. The Statistics concentration and secondary field require Stat 111 to be taken for a letter grade.

2. We expect that the course will be video recorded, but make no guarantees about the video quality or the timing of the posting of lecture videos since that is beyond our control.

3. It *is* possible to register for Stat 111 and another course that meets at the same time or at an overlapping time. See oue.fas.harvard.edu/simultaneous-enrollment for information about Harvard's policies on simultaneous enrollment.