w203: Statistics for Data Science

Syllabus

The goal of this course is to provide students with a foundational understanding of classical statistics and how it fits within the broader context of data science. Students will learn to apply the most common statistical procedures correctly, checking assumptions and responding appropriately when they appear violated. Emphasis is placed on different practices that constitute an effective analysis, including formulating research questions, operationalizing variables, exploring data, selecting hypothesis tests, and communicating results.

The course begins with an introduction to probability theory, with pencil-and-paper problem sets to build intuition for the mathematical objects that comprise statistical models. Next, we describe the use of estimators to learn about model parameters, emphasizing guarantees that hold as sample sizes increase to infinity. We then turn to the logic of hypothesis testing and survey a variety of tests used to compare two groups. Finally, we devote several weeks to a discussion of classical linear regression, stressing its flexibility as a tool. We devote special units to the building of regression models in the context of description and of causal inference. Throughout the course, students will practice analyzing real-world data using the open-source language `R`.

Course Prerequisites

- Proficiency with calculus (including an ability to take simple derivatives and integrals)
- Familiarity with basic matrix operations
- Ability to write proofs

Weekly Workflow

- **Before Live Session** - The first task for each unit is to complete the asynchronous videos and corresponding readings. We recommend that you begin by watching the first videos. As you work forward, you will encounter special "Reading" pages, which direct you to read specific pages in the textbooks. Where possible, we have further organized each week into two or three "sprints." You may want to take a break after each sprint, or tackle each sprint on a separate day. **It is important that you complete all videos and readings before live session.**
- **During Live Session** - In live session, students engage in activities that build upon the asynchronous videos and readings. These include discussions, collaborative problem solving, and programming exercises. We expect that you dial in from a quiet location, that you will arrive on time, and that you connect over both audio and video. Most importantly, we ask that you treat all classmates with respect and help us create a supportive learning environment.
- **After Live Session** - After live session, students complete the homework for the corresponding unit, as well as any tests or lab assignments.

Textbooks

Required Textbooks

- The main text for the class is *Foundations of Agnostic Statistics*, written by Peter M. Aronow and Benjamin T. Miller. A physical copy of the book is available for approximately $30 (Amazon Link, Cambridge University Press Link). As well, for individuals who would like to have a digital copy of the book, it is available through the UC Library (link).
- There is a required course packet, which you may purchase at study.net. This includes select chapters from other textbooks, namely *Probability and Statistics, 8th Edition* by Devore.

Optional Textbooks

The following textbooks are not required but are available for those who would like extra practice problems and/or additional exploration of course concepts. Both are available through the UC library

- *Modern Mathematical Statistics with Applications* by Devore and Berk (library link) is very readable and has lots of worked examples. It's a good complement to Aronow and Miller's concise mathematical language.
- *Elementary Probability and Applications* by Durrett (library link) has a lot of examples for you to work through. It's mostly examples. Only one person can have access to this ebook at a time, so to avoid blockages consider buying a physical textbook for about $30 on amazon.

Required Compute Resources

Much of the work for this course can be completed on the UC Berkeley datahub. Students may also run the course materials in docker with the stable `rocker/verse` image, or through a local install of R and RStudio.

Grading

Grading for the course will follow this rubric:

Component	Percentage
All Homework Combined	25%
Test 1: Probability Theory	10%
Test 2: CE, BLP, & Sampling	10%
Hypothesis Testing Lab	20%
Linear Regression Lab	25%
Participation	10%

On homework, each question sub-part will be graded out of three points, for mastery. This means that:

- (3 points): A sub-part that was fully correct and demonstrates that a student has mastery of that concept.
- (2 points): A sub-part that has substantially correct work, but that has any error
- (1 point): A sub-part that has been attempted, but has considerable errors
- (0 points): A sub-part that has either not been attempted, or has been attempted in a way that communicates no understanding of the concept.

Students who are interested in knowing how to map percentage grades onto letter grades can use this table as a guideline:

Letter Grade	Percentage
A	x >= 93
A-	90 <= x < 93
B+	87 <= x < 90
B	83 <= x < 87
B-	80 <= x < 83
C+	77 <= x < 80
C	73 <= x < 77
C-	70 <= x < 73
Lower Grades	x < 70

Participation Policy

Participation is more than just showing up to class. It includes being prepared, engaging actively, and treating others with respect to sustain a thriving learning environment.

Academic Misconduct Policy

This section covers acceptable and unacceptable conduct for the course. If you have any questions reach out to a member of the w203 intsructional staff.

We treat academic misconduct seriously and will refer cases of academic misconduct to the Center for Student Conduct at UC Berkeley per the Student Code of Conduct. This is a long drawn out process that involves going to hearings and can result in consequences up to and including suspension from the program and revocation of your degree. Don't do it. It's not worth it.

Acceptable Resources

- Wolfram Alpha is an online tool that can be used to solve all kinds of difficult numerical manipulations like differentiation and integration. You can use this on the tests and homework as you please (*this isn't a calculus class*) but when you use it, provide a reference so that your graders know how you got from one stage to the other.
- Stack Overflow and other stack exchange sites. (*as long as you do not ask exact questions from the course*)
- **Your fellow students** are invariably excellent individuals who will understand different aspects of statistics than you. You are allowed to discuss homework with other students, but not tests. For homework problems, strategize with other students as much as you like, but write down the people you work with on your submission, and write down your final solution yourself (do not copy from another student).
- Office hours! Between the section instructors and the TA's there are about a dozen office hours a week. Usually we'll end up covering the homework. A lot of great learning happens during these hours and we strongly recommend them.

Unacceptable Resources

- Chegg, Slader, CourseHero and similar resources where students post exact questions from the course which are answered by professional "tutors" or students at other universities are **unacceptable in all**

circumstances. The instructional team regularly looks at solutions posted to these websites and if we notice solutions that are substantially similar to these sites we will refer the offending student to the Center for Student Conduct.
- Copying from other students homework, websites, old solutions, etc. is cheating and offenders will be reffered to the Center for Student Conduct.

Late Policy

We set deadlines for homework, tests, and labs to be completed. At the same time, we understand that MIDS is a single facet of the life that you live, and some delays are unavoidable.

- **Homework**: Homework assignments are chances for you to apply the content that you have learned in the async and the live sessions. These occur at a regular (nearly weekly) pace. **Homework is due *at the time that your live session meets* on the day that it meets.**
  - Students have **five "late days" that they can use without penalty** on homework through the semester. After those "late days" have been used, each day late will be assessed a 10% penalty from the final grade of that homework assignment. As an example, after using all five late days, a homework assignment turned in one day late could earn a maximum score of a 90% *on that homework*.
  - Students **cannot turn in any single homework more than two days late.** This balances the need for student flexibility, with the need to release feedback to other students in a section. For example, a student who has class on Monday at 4:00p cannot turn in an assignment later than Wednesday at 4:00p.
  - *Note on Gradescope due dates:* Gradescope is the online platform that we use for homework. The due dates shown on gradescope are there to accomodate the last section of the week. You still must turn in your assignments before your live session meets. Late days will be counted at the end of the semester. You must keep your own tally of how many late days you have used.
- **Tests**: Tests count for a larger proportion of the final course grade, and happen less frequently. As a result, we would like tests to be completed on time.
- **Labs**: The labs in the course are group projects. We will expect that labs are turned in for instructor review, in whatever state they are in, at the time they are due.

We appreciate that these are challenging times. Please, be encouraged to talk with your instructor if you have a challenge that arises. We want to support your learning, not cause anxiety or stress. We will work with you to find an accomodation.