w203: Statistics for Data Science

Course Overview and Policies

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

Course Components

- **Asynchronous Content** - Each week of the course begins with a unit of lecture videos, located on [ISVC]. As you work through each unit, you will encounter special "Reading" pages, which direct you to read specific pages in the textbooks. We recommend that you perform the reading when instructed to ensure you have mastered each concept before moving on. Many units are further subdivided into two or three "sprints." You might find it helpful to tackle each sprint on a separate day. **It is important that you complete all videos and readings before your live session.** Slides for each unit are available in this repository.

- **Live Session** - Live sessions take place over Zoom, and can be accessed from [ISVC]. During 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.

- **Homework** - Most weeks of the course include a homework assignment. The homework for each week will be available Tuesday at 2pm on [Github]. It will be due the following Tuesday at 2pm, and submitted via [Gradescope]. **It is important that you start the homework early, including before your live session, to give yourself a full week to complete it.**

- **Solution Files** - Solutions to each homework and each live session will be available on [Github]. You learning is not complete when you solve a problem; it is important that you view the solution to solidify your knowledge and identify ways to improve your arguments. Homework solutions are located in the semester's "central" repository, and will be updated Thursdays at 2pm. Each live session repository will be located in a separate repository, with a name like unit_XX_ls_sol. This allows us to give each student access to the solution immediately after live session.

- **Tests** - The course includes two tests, located on [Gradescope]. The tests are made available during week 4 and week 6 of the course, and must be completed by Tuesday 2pm of the following week. There is a 3-hour time limit for each test. It is up to you to decide when to begin each test. Once you start the test on [Gradescope], a 3-hour timer will begin to count down. Practice tests will be available to give you a sense of what the tests are like.

- **Labs** - There are two labs that are multi-week, group-projects. Labs are available on [Github] and will be submitted via [Gradescope]. Groups are typically between three and four students. Each lab includes a complete analysis, from data wrangling and statistical procedures, to writing a persuasive report.
- **Slack Channel** - We use a slack channel, datasci-203-2022-spring, for official course announcements and general discussion. This is a public channel within the ucbischool.slack.com workspace. Please also join any section-specific slack channels that your live session instructor will announce.
- **Practice Problems** - The course includes a set of curated [practice problems]. Use your slack credentials to log in. Each problem comes with a comment functionality that you can use to share and discuss solutions. While optional, we highly recommend solving lots of practice problems each week. This is the best way to get really comfortable with probability theory and great preparation for the tests.
- **Participation** - 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.

Assignment Submission Guidelines

Mathematical exercises may be typeset using Latex, or hand-written and scanned, but you must ensure that the final file is easy to read. Instructors will not grade submissions with unusual file types or formatting that makes access difficult.

For any assignment that includes R work, you must submit two files (if R was not used, submit a single "output" file):

1. An output file (.pdf, .html, or .md)
2. The source file used to generate your output (**.Rmd** or .ipynb)

Textbooks

Required Textbooks

- The main text for the class is *Foundations of Agnostic Statistics*, 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

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 instructional team.

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 some delays are unavoidable.

- **Homework**: 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 48 hours late.** This allows us to release the homework solutions each Thursday at 2pm.
- **Tests**: Tests cannot be turned in late. Please talk to your instructor in advance if you have a conflict that will interfere with a test.
- **Labs**: Since labs are group projects, they cannot be turned in late. Please talk to your instructor if your team is facing circumstances that may interfere with the lab.

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 accommodation.

Principles of Community

We affirm the UC Berkeley Principles of Community and we endeavor to create a positive, supportive, and inclusive learning environment for all students and all of students' identities. If there is a part of your identity that you would like your instructors or classmates to recognize, please let us know however you are most comfortable.

Calendar

Unit	Parts	Topic	Reading	Assigned	Due
1	Two	Probability Spaces	FOAS 1.0-1.1.4, Knox 2020	Unit 1 HW	
2	Two	Defining Random Variables	FOAS 1.2.0-1.3.4	Unit 2 HW	Unit 1 HW
3	Two	Summarizing Distributions	FOAS 2.0-2.2.2	Unit 3 HW	Unit 2 HW
4	Two	Conditional Expectation and The BLP	FOAS 2.2.3-2.3	Test 1	Unit 3 HW
5	Three	Learning from Random Samples	FOAS 3.0-3.3	Unit 5 HW	Test 1
6	Three	Hypothesis Testing	*Devore & Berk* 9.1, 9.2, 9.3, 9.4, (Optional 9.5)	Test 2	Unit 5 HW

Unit	Parts	Topic	Reading	Assigned	Due
7	Two	Comparing Two Groups	*Devore & Berk* 10.1, 10.2, 10.3, 10.4; 14.1, 14.2	Hypothesis Testing Lab	Test 2
8	Two	OLS Regression Estimates	FOAS 1	-	-
9	One	OLS Regression Inference	FOAS 1	Unit 9 HW	Hypothesis Testing Lab
10	Three	Descriptive Model Building	FOAS 1	Unit 10 HW	Unit 9 HW
11	Two	Explanatory Model Building	FOAS 1	Unit 11 HW & Regression Lab	Unit 10 HW
12	Three	The Classical Linear Model	FOAS 1	Unit 12 HW	Unit 11 HW
13	Two	Reproducible Research	FOAS 1		Unit 12 HW
14	One	Maximum Likelihood Estimation	FOAS 1		Regression Lab