W203 Statistics for Data Science

Summer 2019

Instructors

Eric Penner
eric.penner@ischool.berkeley.edu

Gunnar Kleeman
gunnarklee@ischool.berkeley.edu

Paul Laskowski
paul@ischool.berkeley.edu

Mark Labovitz
mlabovitz@ischool.berkeley.edu

Teaching Assistants

Gurdit Chahal
g.s.chahal@berkeley.edu

Todd Young
todd.young@ischool.berkeley.edu

Description

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 develop intuition for the key concepts that underlie statistical models. Next, we use the simple example of the mean to demonstrate the use of estimators and hypothesis tests. We then turn to classical linear regression, taking several weeks to build a strong understanding of this central topic. Our treatment stresses causal inference and includes a discussion of omitted variables. At the end, we describe some of the concerns that arise in the process of specifying linear models. Throughout the course, students will practice analyzing real-world data using the open-source language, R.(3 units)

Prerequisites

- Working knowledge of calculus. A good understanding of linear algebra is strongly recommended, as the course will make occasional use of matrix notation.
- At least one prior college-level statistics course is recommended.

Sites

Github This is where all course documents will be distributed to you. Throughout the course of the semester we will release (grant you access to specific private repositories) live session documents, homeworks, labs, datasets, and really anything else. It is your responsibility to clone, manage, and keep up to date the repos that you are given access to. We recommend that as soon as you are given access to a repo you clone that repo onto your machine and then immediately start a working branch of that repo so that any updates that we may later push will not cause any merge conflicts.

Piazza Forum This is where all communications between students and instructors will take place for this course. You can post questions, messages and the like to the whole class, other student or the instructors as yourself or anonymously. We encourage you to ask and answer question posed to the entire class those who do will be rewarded for it!

I School Virtual Campus This is where you will watch async video, turn in assignments, take quizzes, and get your grades. There are number of important requirement regarding the manner in which you will need to turn in assignments, some with penalties for non compliance it is your responsibility to know

and act in accordance with these any question should be posted to piazza asap.

Weekly Workflow

Before Live Session Students watch the asynchronous videos and study the assigned readings for a given unit. Note that the readings are mandatory and often include more examples than provided in the videos.

During Live Session Students engage in activities to reinforce and extend the materials they studied.

After Live Session Students complete the homework, lab, or other assignments corresponding to the given unit. Homeworks will be due at the start of the following live session. See individual labs for their due dates.

Required Textbooks

- Devore, J. L. (2015). Probability and statistics for engineering and the sciences Boston, MA: Cengage Learning. 9th Edition. In the past, students have downloaded the 8th edition, for less money or for free, and found this to be sufficient for the course. The largest differences we have found between the 8th and 9th editions are in Chapter 8. We have included this chapter in the study.net course packet, below.
- Wooldridge, J. (2015). Introductory econometrics: A modern approach 6th ed. Boston, MA: Cengage Learning.

Other Required Readings

- **Study.net** Further readings are provided in a course packet, which should be purchased through study.net. The link will be posted on ISVC

Grading

- Probability Theory Lab (individual lab) - 20 percent
- Comparing Means Lab (group lab) - 20 percent
- Linear Regression Lab (group lab) - 20 percent
- 2 Quizzes - 10 percent (5 percent each)
- Weekly Homework - 20 percent
- Class Participation - 10 percent

Weekly Homework

Most weeks of the course include a homework set that is designed to reinforce the concepts covered in class. Each homework is due at the start of the following live session.

Labs

The course includes three labs, which are larger assignments designed to extend your knowlege with a focus on connecting statistical theory to practical application.

The Probability Theory lab is a culmination of the first major section of the course. It includes pencil-and-paper exercises to build intuition for the mathematical building blocks that underly statistics. This is an individual lab.

The Comparing Means lab is an introduction to writing statistical analysis. Focus is placed on the proper selection and interpretation of statistical tests, as well as effective argument and writing style. Students will work in teams of two or three to complete this lab.

The Linear Regression Lab gives students a chance to synthesize knowledge gained throughout the semester and combine technical, inferential, and strategic thinking to produce a professional-level analysis. In the course of this assignment, student teams will have a chance to provide peer feedback to each other. This will be a chance to practice critical reading of statistical analysis, and enable the strongest possible final products. Students will work in teams of two or three to complete this lab.

Quizzes

The purpose of the quizzes is to test your ability to reason about the concepts covered in the course. Quizzes will be conducted under a time limit and may include multiple-choice questions, short-answer questions, and other question types.

File Types

For any assignment that includes R work, you must submit both 1. an output file (.pdf or .html) and 2. a source file used to generate your output (.ipynb or .Rmd or .R). Mathematical exercises may be typeset using Latex, or hand-written and scanned, but you must ensure that the final file is easy to read. Forgetting to submit any of the required files will result in a non-refundable 3-point deduction. Instructors will not grade submissions with unusual file types or formatting that makes access difficult.

Office Hours

Office hours are a central component of this course, giving instructors the chance to tailor explanations to individual students. Students may attend the office hours of any instructor, and they are encouraged to attend as many office hours as possible.

Participation

Students are expected to be active participants in class activities and to come to the live sessions prepared to discuss the videos and readings. Students should also come to class with questions that they would like to discuss with classmates and the instructor. Most importantly, we expect all students to behave professionally and help create a supportive learning environment.

Late Policy

Homework and labs submitted after the deadline will be docked an automatic 20 percent. Unfortunately, we are not able to accept any work after the live session in which we discuss the solutions.