Statistics for Data Science
Syllabus

Last Updated for Spring 2019

Course Developers:	Coye Cheshire coye@ischool.berkeley.edu	Paul Laskowski paul@ischool.berkeley.edu
Session Instructors:	Eric Penner eric.penner@ischool.berkeley.edu Gunnar Kleeman gunnarklee@ischool.berkeley.edu	Casper Joergensen cjoergensen@ischool.berkeley.edu
Teaching Assistants:	Gurdit Chahal g.s.chahal@berkeley.edu Todd Young todd.young@ischool.berkeley.edu	Chi long Ansjory ansjory@berkeley.edu

Office Hours: To be posted by the instructors on ISVC

Course 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:

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

Weekly Workflow:

A typical week of the course proceeds as follows:

- **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. Students should also complete any assigned pre-class exercises.
- **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 24 hours before the following live session. See individual labs for their due dates.

Communication:
Instructors will use Piazza for general course communication. Please post any questions regarding course content and logistics to Piazza so that other students can see them.

Required Textbooks:

1. Devore, J. L. (2015). *Probability and statistics for engineering and the sciences.*¡/em¿ Boston, MA: Cengage Learning.

   This will be our primary textbook for the first part of the course, including probability theory, estimation, and hypothesis testing. Devore includes enough mathematical detail to support our curriculum, but explains the intuition behind it slowly with a large number of examples.

2. Wooldridge, J. (2015). *Introductory econometrics: A modern approach* 6th ed. Boston, MA: Cengage Learning.

   For our study of classical linear regression, we will switch to this classic econometrics textbook. Wooldridge covers the classical linear model in more detail than Devore, explaining how to check assumptions and what to do if they don't appear to hold.

Recommended Textbooks:

1. Fox, J., & Weisberg, S. (2011). *An R companion to applied regression.* Thousand Oaks, CA: Sage Publications.

   We will read selections from the first few chapters of this book as we introduce R. It is not necessary to buy this book because these chapters will be available from study.net. On the other hand, this is a useful book to have on your shelf as you learn more about regression and need to translate your knowledge into R.

Grading:

1. Probability Theory Lab - 20%
2. Comparing Means Lab - 20%
3. Linear Regression Lab - 25%
4. 2 Quizzes - 10% (5% each)
5. Weekly Homework - 15%
6. Class Participation - 10%

Labs:
The majority of the final grade is based on three graded labs. Each of these focuses on a different topic:

1. Probability Theory (individual lab)
2. Comparing Means (group lab)
3. Classical Linear Regression (group lab)

In a typical lab, students will download a real-world dataset to analyze using the techniques learned in class. Each student must submit (1) a PDF report detailing the solutions and (2) an R-script, Jupyter notebook, or Rmd file that is used to generate the solutions. Failing to submitting one of these files will result in an automatic 20% grade reduction.

The Probability Theory lab is unique in that it requires a large number of pencil-and-paper calculations. Students may scan in their work for submission or use LaTex to type their solutions. This is an individual lab.

Lab 2 and Lab 3 are designed to be group labs. Students will work in teams of two or three to complete these.

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.

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.

Weekly Homework:
Most weeks of the course include a homework set that is designed to reinforce and extend the concepts covered in class. Each homework is due 24 hours before the following live session so that instructors have time to assess student progress. Homework will only be given a grade of 0, 1, 2, or 3. In general, students will not receive individual feedback on homework. Instead, it is their responsibility to bring any questions they have to office hours.

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%. Unfortunately, we are not able to accept any work after the live session in which we discuss the solutions.

Course Outline:

1. Introduction and Descriptive Statistics (2 lectures) The course begins with an introduction to quantitative research and tools for describing a sample of data.
    - Measurement
    - Types of variables
    - Operationalization of constructs
    - Descriptive statistics
    - Measures of location
    - Measures of dispersion
2. Probability Theory and Mathematical Statistics (5 lectures) We build up from mathematical foundations to understand how statistical models behave.
    - Axioms of probability
    - Random variables
    - Probability density and cumulative probability functions
    - Joint distributions
    - Unconditional and conditional expectation
    - Variance and covariance
    - Sampling
    - The Central Limit Theorem
3. Estimation and Hypothesis Testing (3 lectures) We introduce statistical inference - the process by which we use a sample to learn things about a population model.
    - Desirable properties of estimators
    - Maximum likelihood estimators
    - Method of moments estimators
    - Confidence intervals
    - The Frequentist approach to statistical inference
    - z-tests and t-tests for one sample
    - Parametric tests for comparing means
    - The reproducibility crisis
    - p-hacking
    - p-value corrections
    - Publication bias
    - Strategies for improving reproducibility
4. Classical Linear Regression (5 Lectures) We study linear regression with an emphasis on correctly checking assumptions, and on the flexibility inherent in the linear model.
    - Bivariate estimation
    - Multivariate estimation
    - Rubin's Causal Model
    - Omitted variable bias

- Factors that influence standard errors
- The classical linear model assumptions
- Key assumptions for large sample sizes
- The use of variable transformations, polynomials, indicator variables, and interaction terms
- Regression Diagnostics and formal statistical assumption testing
- True experiments