Statistics for Data Science
Syllabus

2016 Fall

Course Leads:	Coye Cheshire coye@ischool.berkeley.edu	Paul Laskowski paul@ischool.berkeley.edu	
Instructional Team:	Jennifer Shin jennshin@berkeley.edu	Ryan Kappedal rkappedal@gmail.com	Ali Sanaei sanaei@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. They will also learn to evaluate the design of a study and how the variables being measured relate to research questions.

The course begins with a focus on exploratory analysis and descriptive statistics. From there, we learn how statistical models are built using the structure of probability theory. 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. The course concludes with a brief look at causal inference and reproducibility issues in research. 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.
- **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 a Slack channel for general course communication. Please post any questions regarding course content and logistics to the Slack channel so that other students can see them.

Required Textbooks:

1. Devore, J. L. (2015). *Probability and statistics for engineering and the sciences.* 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. Overall, this is a useful book to have on your shelf as you learn more about regression and need to translate your knowledge into R.

   It may be possible to rent our two required textbooks (e.g. from Amazon.com), which may cost substantially less than purchasing them. Used copies are also available from a variety of sellers. For students that will go on to take W271: Statistical Methods for Discrete Response, Time Series, and Panel Data, note that the textbook by Wooldridge is also required for that course.

Grading:

1. 4 Labs - 60% (15% each)
2. 2 Quizzes - 20% (10% each)
3. Weekly Homework - 10%
4. Class Participation - 10%

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

1. Exploratory Data Analysis
2. Probability Theory
3. Comparing Means
4. Classical Linear Regression

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.

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 graded with a check, a check-minus, or a zero. At the start of each live session, however, one or more students will present their solutions to the class.

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. Descriptive Statistics and Exploratory Data Analysis (2 lectures) The course begins with an introduction to quantitative research and techniques for exploring a dataset.
    - Measurement
    - Types of variables
    - Operationalization of constructs
    - Descriptive statistics
    - Measures of location
    - Measures of dispersion
    - Tools for visualizing Data
    - Guidelines for exploratory analysis
2. Probability Theory and Mathematical Statistics (4 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
    - non-parametric tests for comparing means
4. Classical Linear Regression (4 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
    - 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

5. Special Topics (2 lectures) The course concludes with a brief look at causal inference and reproducibility issues in research.

   - Rubin's Causal Model
   - Omitted variable bias
   - Exogeneity
   - True experiments
   - Error rate inflation
   - p-value corrections
   - Multiple comparisons
   - Stopping rules
   - Planned versus post-hoc comparisons
   - Publication bias
   - Strategies for improving reproducibility