Statistics 109 Syllabus
Spring 2020
Tentative

Instructor: Michael Parzen
Email: mparzen@fas.harvard.edu or michaelparzen@gmail.com
Office: SC-300B. Office Hours: Monday & Wednesday, 2-3pm and Online.

Caution: *This course has been substantially modified from previous offerings. An emphasis on interactive class sessions and the addition of take-home components to exams are the more obvious changes. Be sure to read this syllabus carefully and watch the first lecture to see if it is a good fit. We are not sure how this new format will work in a distance learning capacity but have a strong support team in place.*

Course Lectures: The campus lectures are Monday and Wednesday from 12pm-1:15pm in Science Center, Hall A (feel free to attend if you are in town). My lectures will be videotaped and placed online, usually a few hours later the same day (officially the extension school has 24 hours from when class ends to put the videos up).

Office Hours: I will have video office hours each week-details forthcoming.

Weekly Section: There is an optional weekly section where a teaching assistant reviews topics from lecture, goes over additional example problems and answers questions about the homework. This will be held online, videotaped, and can also be attended in person. This section will be scheduling is currently to be determined.

Teaching Assistant office hours: There will be at least two hours of additional video conferencing enabled office hours each week-details forthcoming.

Teaching Assistants: To be announced by the first day of class.

Availability: My teaching staff and I want to be as accessible as possible, given the constraints of a distance class. We welcome you to be in contact with us as much as possible and to let us know of any issues or problems as soon as possible.

Course website: https://canvas.harvard.edu/courses/67602

In Canvas you will be able to:

- check the homepage for weekly updates and reminders
- find copies of power point slides for lecture
- find any updates to the Syllabus
- check the Course Schedule for upcoming quiz, exam or homework dates
- watch the lectures as streaming video and a few short videos on difficult topics
- take the online quizzes
- check your grades

Lecture Slides: Slides for each lecture will be available for download at least 24 hours before each class.

Textbook: We will use several references all which are available on the course web site. A primary book will be *Data Analysis and Graphics Using R: An Example-Based Approach* (3rd edition) by John Maindonald and W. John Braun. This book is available electronically free of charge from the Harvard library system.

Course Objective:

Stat 109 is a second course in statistical inference and is a further examination of statistics and data analysis beyond the introductory course. Topics include t-tools and permutation-based alternatives including bootstrapping, analysis of variance, linear regression, model checking and refinement. Statistical computing and simulation based emphasis will also be covered as well as basic programming in the R statistical package. Emphasis is made on thinking statistically, evaluating assumptions, and developing tools for real-life applications. Note that Stat 109 cannot be taken for credit if Stat 139 has already been taken.

By the end of the course, students should be able to evaluate the strengths and weaknesses of a variety of statistical techniques appearing in the media, scientific literature, or students' own work. Given a data set, students should be able to

- state hypotheses,
- explore the data using statistical software,
- determine which statistical model may be appropriate,
- apply corresponding hypotheses tests,
- check the assumptions behind these tests and models,
- interpret the results of the analysis to draw conclusions about the hypotheses.

Prerequisites:
Knowledge found in a typical introduction to statistics course is assumed. A list of topics assumed known is given at the end of this syllabus.

Sections:
Optional (but **strongly** suggested) TA-led sections will be held throughout the course. Sections will mostly meet on Wednesday and Thursdays. Sections will go over practice problems and review difficult material.

Computing:
We will be heavily using the statistical software package, *R* and RStudo. Students can choose to work locally or via a free account at rstudio.cloud. See the course website for details on using rstudio.cloud.

No previous knowledge of computer programming or the R software is required.

Accommodations for students with disabilities:
Students needing academic adjustments or accommodations because of a documented disability must present their Faculty Letter from the Accessible Education Office (AEO) and speak with Mike by the end of the third week of the term. Failure to do so may result in us being unable to respond in a timely manner. All discussions will remain confidential.

Collaboration:
You are encouraged to discuss homework with other students (and with the instructor and TAs, of course), but you must write your final answers yourself, in your own words. Solutions prepared "in committee" or by copying or paraphrasing someone else's work are not acceptable; your handed-in assignment must represent your own thoughts. All computer output you submit must come from work that you have done yourself. All exams (midterm and final) are individual work.

Grading Guidelines:
Your final score for the course will be computed using the following weights; 20% for homework, 30% for project, 20% for midterm and 30% for final exam.

Homework:
There will be approximately 6 homework assignments. No homework scores will be "dropped." Late homework submissions will not be accepted without a note from a health professional. Homework will be submitted via Canvas in pdf format.

Project:
A group project will be due the first day of reading period. It will be based on a data analysis of your choice, and will result in a 6-10 page paper.

Exams:
There will be one midterm (on Thursday, March 14), and a final exam (date yet to be determined).

Graduate Credit
Students taking the course for graduate credit must work in a project group with other graduate students.

Policies on Accessibility and Academic Conduct

Official Harvard Extension School Policies

The Extension School is committed to providing an accessible academic community. The Accessibility Office offers a variety of accommodations and services to students with documented disabilities. Please visit https://www.extension.harvard.edu/resources-policies/resources/disability-services-accessibility for more information.

You are responsible for understanding Harvard Extension School policies on academic integrity (https://www.extension.harvard.edu/resources-policies/student-conduct/academic-integrity) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting the wrong draft, or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity. To support your learning about academic citation rules, please visit the Harvard Extension School Tips to Avoid Plagiarism (https://www.extension.harvard.edu/resources-policies/resources/tips-avoid-plagiarism), where you'll find links to the Harvard Guide to Using Sources and two free online 15-minute tutorials to test your knowledge of academic citation policy. The tutorials are anonymous open-learning tools.

Some general course points are as follows:

- You are responsible for ensuring that you have the necessary computer hardware and software, including any course-specific software needed to complete course assignments. Check the course website to determine if any special software is needed. Harvard University does not provide equipment or software. You must have Internet access available throughout the term.
- Distance courses are not self-paced. You are expected to watch the video lectures on a weekly basis and be an active participant in the course throughout the term.
- You must adhere to deadlines and due dates provided in the course syllabus and may not join distance courses after the end of registration.
- Recorded lectures for the distance education courses are available only to registered students; lectures are password protected after the first week of class.
- Many distance courses have in-class hour, midterm, and final online exams. You may take exams on campus or arrange to take them with a web-based proctor.

Projected Course Outline

- Introduction and R
- Random Variables, Normality, Central Limit Theorem
- Central Limit Theorem and Confidence Intervals
- Hypothesis Testing
- In depth, Testing a Population Proportion
- In depth, the Two Sample t Test
- The Bootstrap and Permutation Tests
- Power and Sample Size Calculations
- Multiple Comparisons and ANOVA
- Simple Linear Regression
- Assumptions for Simple Linear Regression
- Diagnostics and Transformations for SLR, I
- Diagnostics and Transformations for SLR, II
- Multiple Regression
- Diagnostics and Transformations for MLR, I
- Diagnostics and Transformations for MLR, II
- Model Checking and Refinement
- Strategies for Variable Selection
- Extensions to Regression (Quantile, Robust)
- Regression with Time Series Data
- Introduction to Logistic Regression

Although you are responsible for all material taught in an introductory statistics class such as the high school Advanced Placement statistics class or Statistics 100,101,102 or 104, here is a checklist of material that we will assume you already know before Stat 109.

Introduction

a) Population vs. Sample
b) Parameter vs. Statistic
c) Descriptive vs. Inferential statistics

Sample Data

a) Categorical vs. Numerical data
b) Discrete vs. Continuous numerical data
c) Observational studies vs. Randomized experiments
d) Confounding variables
e) Sampling techniques (simple random sample)
f) Types of sampling bias (response, selection, non-response) and their effects.

Graphical Methods

a) Frequency, Relative frequency (distribution)
b) Histogram, Boxplot, Scatterplot
c) Describing these plots (shape, skew, center, spread, outliers, etc.)

Numerical Measures

a) Mean, Median
b) Proportion
c) Standard deviation, Variance
d) Quartiles, Interquartile range, outliers
e) Z-scores
f) Percentiles
g) The Empirical Rule (or 68-95-99.7 Rule)

Probability and Sampling Distributions

a) What is a probability?
b) Population Distribution
c) Discrete vs. Continuous variables
d) Density curve
e) Normal distribution
f) Standardized variable, z-score
g) Standard normal distribution
h) Sampling distributions
i) Sampling variability
j) Central Limit Theorem
k) Sampling distribution of the sample mean
l) Sampling distribution of the sample proportion

One Sample Inference for the Population Mean and Population Proportion

a) The t-distribution
b) Confidence Intervals for the Population Mean and Proportion
c) Point estimate
d) Confidence level and critical values
e) Standard error
f) Sample size calculation
g) Hypothesis Testing for the Population Mean and Proportion
    i. Null and alternative hypothesis
    ii. 1-sided and 2-sided tests
    iii. Type I error, Type II error
    iv. Level of significance
    v. Test statistics
    vi. P-value

Two Sample Inference for the Difference in Population Means and Proportions

a) Confidence Intervals for the difference in population means: paired samples, independent samples.
b) Hypothesis testing for the difference in population means: paired samples, independent samples.
c) Confidence intervals and hypothesis test for the different in two proportions.

Simple Linear Regression

a) Scatterplots
b) Pearson's correlation
c) Fitting a line: Least squares method
d) Coefficient of determination.
e) Assumptions of simple linear regression (known to some extent)
f) The F test