BSTA001: Population Health Data Science - I

About the Course

Instructor tom mcandrew
Email: mcandrew@lehigh.edu
Office Coordinates: Virtual Office this semester
Office Hours: TBD by students | by Appt.

Class times Monday, Wednesday 12:10 - 13:25, Wednesday 15:00 - 15:50 in Rauch Business Center, Room 070

Class materials Please plan to bring your laptop to class. Contact me if you do not have access to a working laptop.

Course Website I will update course website at http://thomasmcandrew.com/classes/2020F_PHDSI/public/ regularly with lecture notes and materials used in class. Lab and homework assignments will be distributed on GitHub.

Description In Population Health Data Science I (PHDS-I) students will spend the semester learning the fundamentals of probability theory, univariate statistics, statistical computing, and machine learning. A mix of traditional and experiential learning will focus on how to build an analysis pipeline to answer pressing questions in population health. In-class examples and projects will use real data sets. Examples include: comparing cardiovascular interventions in clinical trials, evaluating the incidence of influenza in the United States, and visualizing international health expenditures and burdens. Students will propose a small data-driven project focused in population health, and use their newly-acquired data science skills to collect, analyze, and present their work. We will plan to cover the following topics:

- Fundamentals of probability theory
  - Basic set theory and counting principles
  - Kolmogorov's Axioms
  - Baye's Theorem
- Univariate statistical distributions
  - Bernoulli and Binomial
  - Geometric and Poisson
  - Normal and Chi-Square
- Statistical computing and data munging
  - Python

    - Types of data
    - Flow control (loops etc)
    - Methods for visualizing data (scatter plot, histogram, etc)
  - Building an analysis pipeline
  - Basics of version control for software development
    - GitHub
  - Learning a univariate distribution from clinical data

- Machine learning
  - Big data, algorithms, and ethics
  - Differences between supervised and un-supervised learning
  - The perceptron algorithm
    - Objective functions
    - Exploring a simple learning algorithm to classify data

Textbook We will use the following open source (free) materials for class

- Introductory Statistics for the Life and Biomedical Sciences

A pdf of the book can be downloaded for free from the author's website at `https://www.openintro.org/go/?id=biostat0&referrer=/book/biostat/index.php`.

- Computational and Inferential Thinking

I will occasionally assign reading from Computational and Inferential Thinking. This is a free textbook available at `https://www.inferentialthinking.com/chapters/intro.html`

Time commitment I recommend budgeting approximately three out-of-class hours for every in-class hour to complete the reading, assignments, and homework. Spending twelve hours per week should be enough time to complete class requirements. If you are spending more than 12 hours per week on a regular basis, I would encourage you to check in with me.

Policies

Attendance Your attendance in class is crucial. If you are sick or otherwise cannot attend class, please let me know and stay home and rest.

Collaboration Much of this course will operate on a collaborative basis, and you are expected and encouraged to work together with a partner or in small groups to study, complete homework assignments, and prepare for exams. However, every word that you write must be your own. Copying and pasting sentences, paragraphs, or large blocks of python code from another student is not acceptable and will receive no credit or a penalty. No interaction with anyone but the instructor is allowed on any exams or quizzes. All students, staff, and faculty are bound by the Lehigh University Honor Code.

To sum up: On homeworks and labs, I want you to work together, but you must write up your answers yourself.. Dishonesty, plagiarism, etc., will be reported.

Technology

Computing with Python 3 Modern statistics can't be done without computation. We will use Python 3 in this course. Python is one of the most commonly used programming languages and is often used in industry level statistics and data science positions. Knowing Python is a marketable skill. In this class, we will use Python as much as we can, and for many homework problems.
We will use Python3 via Jupyter Notebooks. Lehigh University has their own server for running Jupyter Notebooks that can be accessed at `https://hpcportal.cc.lehigh.edu/`. If you are not on campus, you will need to be logged onto LU's VPN network. Instructions for the VPN are here. You are also welcome to work locally on your own computer if you have Python3 and Jupyter Notebooks. When you install Python, please make sure you install Python version 3.5 or higher.

It will be important to bring your laptop to class so that you can use Python for in class exercises. Much of this work will be done in pairs, but we need to ensure that there is a sufficient number of computers. Please let me know if this presents any issues.

Version Control with Git and GitHub Git is a version control system that facilitates working on coding and writing projects collaboratively, and allows you to revert your code to a previous version if you realize that you made a mistake. Version control systems such as git are used in most modern data science and statistics positions in industry. Part of my goal is to ensure you are prepared to enter the work force, and for that reason, the basic use of git is a learning objective for this course. This means all labs and the computational portion of homework assignments will be distributed to you in git repositories and submitted by committing and pushing the completed assignment to GitHub. I will provide further details and walk through this process, as well as basic interaction with git, in class.

Assignments

Your grade for this course will be a weighted average of scores from several components:

Item	Weight
Participation, Labs, and Homework	35%
Quizzes and Midterms	45%
Final Project	20%

Participation, Labs, and Homework The best way to learn statistics is to do it. This class will be built around a series of labs that we will do in class. Although most labs will not be graded for correctness, unless indicated, I expect you to complete them and push your work to GitHub. I will occasionally look at submitted labs to see how everyone is doing and whether there are any points I need to address in class. I am always happy to answer any questions you have about these labs. Additionally, we will have regular homework assignments to be completed outside of class.

Exams There will be two “midterm” exams and occasional in-class quizzes. Midterms will be in-class written exams. Midterms and quizzes will always be announced at least one class session in advance. We will not have a cumulative final, but instead a final project (see Project below). No communication with anyone besides the instructor is allowed on these assessments.

Project A large component of the course will be a project which will be presented to your classmates. Briefly, this project will entail application of statistical techniques and data visualization to a population health data set of your choosing. The goal will be to pose a hypothesis and attempt to answer it using skills learned in class. A separate handout will provide additional details.

Extra Credit Extra credit is available in several ways: attending an out-of-class lecture (as will be announced) and writing a short review of it; pointing out a substantial mistake in the book, a homework exercise or exam solution; if you read closely and to this point in the syllabus you can receive five percent extra credit on your first quiz by emailing me the name of you favorite prehistoric dinosaur, and if you don’t have a favorite, make one up; drawing my attention to an interesting data set or news article; etc. The extra credit is typically applied when a student is near the boundary of a letter grade.

Grading When grading your written work, I am looking for solutions that are technically correct and reasoning that is clearly explained. *Numerically correct answers, alone, are not sufficient* on homework, tests or quizzes. Neatness and organization are valued, with brief, clear answers that explain your thinking. If I cannot read or follow your work, I cannot give you full credit for it.

Accommodations for Students with Disabilities Lehigh University is committed to maintaining an equitable and inclusive community and welcomes students with disabilities into all of the University’s educational programs. In order to receive consideration for reasonable accommodations, a student with a disability must contact Disability Support Services (DSS), provide documentation, and participate in an interactive review process. If the documentation supports a request for reasonable accommodations, DSS will provide students with a Letter of Accommodations. Students who are approved for accommodations at Lehigh should share this letter and discuss their accommodations and learning needs with instructors as early in the semester as possible. For more information or to request services, please contact Disability Support Services in person in Williams Hall, Suite 301, via phone at 610-758-4152, via email at indss@lehigh.edu, or online at `https://studentaffairs.lehigh.edu/disabilities`.

The Principles of Our Equitable Community: Lehigh University endorses The Principles of Our Equitable Community. We expect each member of this class to acknowledge and practice these Principles. Respect for each other and for differing viewpoints is a vital component of the learning environment inside and outside the classroom.