Introduction to Computational Statistics
M15645

Instructor Info

Steven Swee

M15645-teachers@esp.mit.edu

Course Info

Prereq: Statistics

Sundays

2-3p EST

Zoom

TA Info

Snow Petrel

Office Hrs: Sunday

(See HSSP Website)

Overview

Computational statistics is an exciting field in which we augment computational methods with traditional statistics (i.e. AP Statistics). During the course, we will survey various methods such as bootstrapping and Monte Carlo simulations. Unlike most other statistics courses, where methods are heavily emphasized, we will focus on developing a statistical intuition and how to develop domain knowledge. According to the statistician Nate Silver, if you want to be a good data scientist, you should spend 49% of your time developing your statistical intuition (i.e. how to ask good questions of the data), and 49% of your time on domain knowledge (improving overall understanding of your field). Only 2% on methods per se." Ultimately, the keys to become a more effective statistician or data scientist is to know how to ask the right questions and to know how to absorb new information quickly.

Material

Programming IDE
Google Colab

Alternative IDEs
VSCode, Pycharm, Spyder

Recommended text for those interested in studying statistics
An Introduction to Statistical Learning (https://www.statlearning.com/)

Mini-Projects

Throughout the course, there will be mini-projects for you to apply what you learned in the class with an open-source dataset. These mini-projects are completely optional and can serve as starting points for your own bigger projects if you wish to work on one.

Learning Objectives

- Develop an intuitive understanding of how some methods in computational statistics work
- Be able to compare and contrast classical statistics and computational statistics
- Gain some programming skills and understand basic algorithms
- Learn how to ask important and insightful research questions
- Learn how to write an effective report to showcase your findings

FAQs

What is computational statistics? Is it like data science?

I personally like to think of computational statistics as a subset of data science. Computational statistics focuses on using algorithms and computational tools to solve statistical problems.

Which programming language will we be using?

We will be using Python. For those interested in data science, other languages to learn are R (my personal preference) and SQL.

What are the prerequisites?

A background and/or interest in statistics and data science is a must. While we won't cover too much formal math (many of the explanations will be qualitative and designed to build intuition), I will be using quite a bit of jargon from statistics. A programming background is not required as any programming concepts we cover will be relatively easy to pick up.

What exactly can I take away from this class?

The goal is to introduce you to one part of the world of data science. There will be optional mini-projects and you are free to use what you learn to conduct a small research project on your own.

Google Colab

We will be using Google Colab throughout the course. There's no installation required and only a Google account is required (which you probably already have). Google Colab features Jupyter Notebooks, which are relatively easy to use.

GitHub

GitHub is a platform that allows users to share coding projects with each other. To access the in-class code and mini-projects, you can pull (fancy for download) the file from my GitHub repository directly into Google Colab.

Optionally, you can upload your code to GitHub to share with anyone you like. You are free to use the code provided by me.

Report Writing

Part of this class will focus on reading and writing an effective scientific paper. We will survey open-access publications and work on technical writing. Remember, fancy statistical analysis mean little when the discussion and results are poorly written. A good paper can make modest use of statistics and present interesting results.

Fun fact: For an upcoming publication, the statistical methods I used included a chi-square test (covered in AP Stats) and odds ratio (covered in undergraduate/graduate biostatistics). Even in medical publications, the statistics you learn in high school are still used!

Class Schedule

MODULE 1: The Big Picture

Week 1	Introduction and Setting Up	Learn the basics of computational statistics and programming in Google Colab
Week 2	Ask Questions, Gather Data, and Write Reports	Learn the soft skills necessary to be a successful statistician or data scientist

MODULE 2: The Methods

Week 3	Monte Carlo and Bootstrapping	Learn how to use computers to approximate solutions to problems
Week 4	Data Analysis	Learn how to interpret statistical results

MODULE 3: Putting It All Together

Week 5	Statistical Analysis Day 1	Combine the ideas covered in class to ask a research question and conduct an experiment
Week 6	Statistical Analysis Day 2	Combine the ideas covered in class to write about your results and present them

Mini-Projects

Week 1	Intro to Python Programming	Students will familiarize themselves with obtaining code from GitHub and programming in Google Colab.
Week 2	Interpreting Statistics Effectively	Students will review current research on common pitfalls in statistics education. Students will also critique written analyses and draft their own analysis.
Week 3	Running Simulations	Students will run a classical coding problem of estimating π using simulations.
Week 4	Kaggle Datasets	Students will learn how to download datasets from Kaggle, a platform for data scientists. Students will also learn how to perform basic data analysis using the methods they learned in class.
Week 5	Finding Your Own Dataset	In the first part of this two-part mini-project, students will learn how to develop an interesting research question, find relevant data on Kaggle, and conduct appropriate statistical analysis.
Week 6	Writing and Presenting	In the second part of this two-part mini-project, students will learn how to write about their findings and create their own presentations.