“2: Essential Statistics and Data Science Seminar”

Professor John W. Emerson (“Jay”)
Director of Graduate Studies
Department of Statistics & Data Science
Yale University

Course Description

This course provides intermediate-level coverage of topics in Statistic and Data Science, solidifying the foundation for further study and research. While some problems are pulled from my research, we may study newer “fresh” problems for the first time (including the most up-to-date COVID-19 data). Most data examples have important lessons that are critical for developing skillsets required to engage real data in any area of application: Finance, Consulting, Startups, Marketing, Science, and so on. Students will both review and learn additional topics in probability and statistics and will learn “hands-on” intermediate-level programming skills working in the R language and environment for statistical computing and graphics using R Studio.

The course has two main phases. The first phase consists of four weekends of mini-lectures together with in-class exercises and discussion. We will cover topics in probability and statistics, but will focus primarily on programming with the R language at a higher level that will be essential for the research projects. Most assigned work will not be collected or graded, and students are expected to work independently and collaborate with each other as instructed. On-line lectures will not simply duplicate assigned reading or data exploration, however, and student must be self-motivated to work outside of class even when a formal homework assignment is not collected. The course second phase is a short research seminar. Each of three research groups will meet with Professor Emerson for an hour each week for guidance on their projects and students must work independently and with the group to finish a good first draft of the project report during the final four weeks. When not meeting with Professor Emerson, the groups should work together to make progress on their own.

This course is the **second** of a series of three research seminars offered by Professor Emerson; it assumes some basic familiarity with R and R Studio and probability and statistics. The first, “Introductory Statistics and Data Science,” assumes no prior experience programming or in probability and statistics; students completing that introductory course should be prepared for this second course. The final course is Professor Emerson’s “Statistics and Data Science with Real-World Case Studies” research seminar, with the highest level of independent research projects; it should be accessible to students completing one (but preferably both) of the first courses, or to students who are comfortable programming and working with data in the R language and are familiar with standard core topics in statistics.

Textbook and Other Materials

We will use the free textbook **OpenIntro Statistics** to guide our explorations of topics in probability and statistics. Other freely-available resources relating to programming with R will be provided.

Potential Research Projects (intermediate-level exploration and analysis)

#1: COVID-19. The data continue to evolve. What can we learn about patterns of spread and rates of infection and fatalities? Can these patterns be tied to observed government policies or other socio-economic factors? The project may include the development of a web application for deploying interactive exploration of the data.

#2: This project will study market efficiencies in gambling point spreads in sports (a common choice is basketball, although other sports may be possible if data are available). This project requires mastery of skills including data parsing from web sources and linear regression and analysis of variance.

#3: This project considers a data set on real estate valuations in New Haven, CT, and seeks to model property values from housing characteristics and location.

#4: This project considers the question of nationalistic biases in the judging of Olympic diving.

Other research topics may be possible, if proposed by a group of students and accepted by Professor Emerson, but must be appropriate in scope given the limited length of this course.