DATA 606 Statistics & Probability for Data Analytics

CUNY SPS Master of Science in Data Science

Spring 2021

Instructor: Jason Bryer, Ph.D.
Class Meetup: Wednesday 8:30pm to 9:30pm
Office Hours: Friday 12pm to 1pm & by appointment
Email: jason.bryer@cuny.edu
Website:: spring2021.data606.net

Course Description

This course covers basic techniques in probability and statistics that are important in the field of data analytics. Discrete probability models, sampling from infinite and finite populations, statistical distributions, basic Bayesian statistics, and non-parametric statistical techniques for categorical data are covered in this course. Each of these statistical concepts will be applied in a variety of real-world scenarios through the use of case studies and customized data sets.

Course Learning Outcomes:

By then end of the course, students should be able to:

- Understand the foundations of probability theory and perform basic probability calculations.
- Build basic stochastic models for commonly encountered business problems.
- Model situations involving uncertainty using appropriate probability distributions and conditional techniques.
- Explore and summarize data using descriptive statistics.
- Test hypotheses using classical and modern computational techniques.
- Construct estimators and calculate intervals using classical and modern computational techniques.
- Perform basic Bayesian statistical techniques for estimation and testing hypotheses.

Program Learning Outcomes addressed by the course:

- Business Understanding. Learn when probabilistic techniques apply to certain categories of business problems, discuss the sorts of solutions that are possible, and understand the limitations of these techniques.
- Foundational Math Skills. Explore and analyze data, build probabilistic and statistical models, construct estimators, and test hypotheses.
- Predictive Modeling. Learn foundational techniques that underlie predictive modeling algorithms, such as Naïve Bayes.
- Presentation. Complete and submit collaborative assignments using techniques from the course.

How is this course relevant for data analytics professionals?

Probabilistic techniques are the foundation of many data science applications from data exploration and visualization to outlier analysis, stochastic modeling, and data mining algorithms. This course will ensure that students have a strong understanding of these foundations.

Grading

- DAACS (6%)
- Participation (10%)
- Homework (18%)
- Labs (36%)
- Data Project (20%)
- Final exam (10%)

Grade Distribution

Quality of Performance	Letter Grade	Range %	GPA
Excellent - work is of exceptional quality	A	93 - 100	4
Excellent	A-	90 - 92.9	3.7
Good - work is above average	B+	87 - 89.9	3.3
Satisfactory	B	83 - 86.9	3
Below Average	B-	80 - 82.9	2.7
Poor	C+	77 - 79.9	2.3
Poor	C	70 - 76.9	2
Failure	F	< 70	0

How This Course Works

This course is conducted entirely online. Each week, you will have various resources made available, including weekly readings from the textbooks and occasionally additional readings provided by the instructor. Most weeks will have homework assignments and labs to be submitted (although some chapters will take more than one week, see the schedule for details). There will also be a presentation required and a forum post introduction required. You are expected to complete all assignments by their due dates.

You are expected to attend or watch every Meetup. I highly recommend attending the Meetups live if possible but I understand that may not be possible for everyone. Recordings will be made available by the next morning on the Meetups page. In addition to highlighting key concepts from each learning module, some topics will be discussed that are not in the textbook. Moreover, I regularly make announcements in the Meetups that will be important to being successful in this course. At the end of each Meetup there will be a short reflective exercise. These will contribute to your participation grade.

Meetup presentations will comprise the solution and presentation to the class of one of the suggested problems for study from the weekly materials (not the graded homework problems). Each student must present one problem during the semester. Problems are chosen by entering your name and problem in the Google Spreadsheet. Note there is a maximum of three presentations per meetup and presentations should be no more than five minutes. Prepare your presentation so that the slides or document (I suggest using R Markdown) will be shared on the course website. Problems are assigned first come, first served, so any problem not already chosen by another student is available.

The culmination of the course will be the presentation of the analysis of a dataset of your choosing. There will be a number of time slots available to present. You will be **required to attend one presentation session**, present yoru analysis and provide peer feedback for other students in that timeslot. See the project for more information.

Schedule

Note: Schedule is subject to change.

Dates	Topic
Jan-29 to Feb-07	Chapter 1 - Intro to Data, R, and Rstudio
Feb-08 to Feb-14	Chatper 2 - Summarizing Data
Feb-15 to Feb-21	Chapter 3 - Probability
Feb-22 to Mar-07	Chapter 4 - Distributions
Mar-08 to Mar-14	Chatper 5 - Foundation for Inference
Mar-15 to Mar-21	Chapter 6 - Inference for Categorical Data
Mar-22 to Apr-04	Chapter 7 - Inference for Numerical Data
Apr-05 to Apr-18	Chapter 8 - Linear Regression
Apr-19 to May-02	Chapter 9 - Multiple & Logistic Regression
May-03 to May-16	Intro to Bayesian Analysis

Accessibility and Accommodations

The CUNY School of Professional Studies is firmly committed to making higher education accessible to students with disabilities by removing architectural barriers and providing programs and support services necessary for them to benefit from the instruction and resources of the University. Early planning is essential for many of the resources and accommodations provided. Please see: http://sps.cuny.edu/student_services/disabilityservices.html

Online Etiquette and Anti-Harassment Policy

The University strictly prohibits the use of University online resources or facilities, including Blackboard, for the purpose of harassment of any individual or for the posting of any material that is scandalous, libelous, offensive or otherwise against the University's policies. Please see: http://media.sps.cuny.edu/filestore/8/4/9_d018dae29d76f89/849_3c7d075b32c268e.pdf

Academic Integrity

Academic dishonesty is unacceptable and will not be tolerated. Cheating, forgery, plagiarism and collusion in dishonest acts undermine the educational mission of the City University of New York and the students' personal and intellectual growth. Please see: http://media.sps.cuny.edu/filestore/8/3/9_dea303d5822ab91/839_1753cee9c9d90e9.pdf

Student Support Services

If you need any additional help, please visit Student Support Services: http://sps.cuny.edu/student_resources/