CS 7200: Statistical Methods for Computer Science

Spring 2020

January 23, 2020

Location: Tue 11:45am-1:25pm and Thu 2:50pm - 4:30pm, Behrakis Health Sciences Center 325

Instructor: Olga Vitek, WVH 310, o.vitek@neu.edu
Office hours Tue 1:30-2:30pm and Fri 4:30-5:30pm, or by appointment, WVH 310F.

Tecahing assistant: Mr. Sicheng Hao, WVH 310 hao.sic@husky.neu.edu
Office hours Tue 2-3pm, or by appointment.

Goals of the course: The course introduces concepts in applied statistics. It overviews Bayesian and frequentist characterization of uncertainty for continuous and categorical data, principles of experimental design, and methods of causal inference. The course discusses the methodological foundations, as well as issues of practical implementation and use.

The methods discussed in the course are useful in any area of science and industry that collects and analyzes data. First, many areas rely on empirical research. Students working in these areas will benefit from a formal exposure to the scientific method, principles of experimental design, and analysis of data from designed and observational studies. Second, the success of most new scientific methods is determined by the quality of their evaluation. The concepts presented in this course will help students design and analyze data from method evaluation experiments.

The course discusses the following topics:

- Basics frequentist statistical inference for continuous data: measures of association, confidence and prediction intervals, hypothesis testing, benefits and limitations of p-values.
- Experimental design: ways to select data for the experiment and maximize its statistical efficiency, factorial, fractional factorial and block designs; response surface exploration.
- Introduction to causal inference: graphical models, adjustments for confounders, interventions and counterfactual inference.
- Statistical inference for categorical data: measures of associations, generalized linear models and log-linear models.

At the end of the course the students will be able to (1) recognize the problems of inferential nature and understand the underlying principles, (2) use statistical inference to design experiments and analyze data, and appropriately document the process, and (3) draw valid conclusions supported by the experimental design and data analysis, and clearly present the results.

Pre-requisite: Proficiency in linear algebra, probability and programming languages such as Python, R, or Matlab.

Software: Students can work on homework assignments and projects in Python, R, or Matlab. Examples of implementations of statistical methods will be provided in R.

Course web page: https://ovitek.github.io/CS7200/S19/index.html
Daily updates on the schedule, handouts and homework assignments will be posted on the course page.

Attendance: Attendance is optional, but you are responsible for all the material covered in class.

Communication: The course will be using the discussion board Piazza
piazza.com/northeastern/spring2020/cs7200 You are encouraged to ask and answer questions on the discussion board. All important announcements will be made through Piazza. Once the course begins, course-related email inquiries will be left unanswered.

Textbook: The main textbook is
Kutner, Nachtsheim, Neter & Li (2005). *Applied Linear Statistical Models*, 5th Ed, McGraw-Hill.

Additional texts will be posted dynamically on the course website and on Piazza.

Homework: Expect roughly 4 biweekly homeworks during the semester. Extensions to homework deadlines can be obtained if requested **at least 48 hours** before the deadline, and duly justified. Homeworks turned in after the deadline will not receive credit.

Answers to the homework problems should be submitted in a format that is easy to both read and reproduce. Computational problems can be done in any programming language (R preferred), and solutions should include the data file (.rdata preferred), the source (.rmd preferred), and the output file (.pdf or .html).

Although some aspects of the homeworks can be discussed with your colleagues and on Piazza, and asking/answering questions on Piazza is encouraged, each homework should be done independently. A homework having any degree of similarity with that of another student (current or past, at Northeastern or outside) is considered plagiarism, and will not be accepted. The homework will be assigned a grade of 0. Additional consequences are described at
http://www.northeastern.edu/osccr/pdfs/Resources/Faculty_Guide_to_Academic_Integrity.pdf

Exams: One in-class midterm, and one in-class final exam.

Grades: All grades will be distributed via Blackboard.

Re-grades of homeworks and exams: All re-grading requests should be made in writing, within **one week** after receiving the grade. The request should state the specific question that needs to be re-graded, as well as a short (1-2 sentences) explanation of why re-grading is necessary. The new grade can potentially be lower than the original grade.

Project: At the end of the semester the students will perform a group project working with a real-world problem.

The project grade consists of project proposal (20%), project report (oral 30% and written, 30%), and project review (20%).

Projects having any degree of similarity with work by any other group, or with any other document (e.g., found online) is considered plagiarism, and will not be accepted. The minimal consequence is that all the group members will receive the project score of 0, and the best possible overall course grade will be C. Additional consequences are described at
http://www.northeastern.edu/osccr/pdfs/Resources/Faculty_Guide_to_Academic_Integrity.pdf

Breakdown of the final grade: The final grade is based on a total of 400 points broken down into homeworks (100 pts), midterm (100 pts), project (100 pts), final exam (100 pts).

The final letter grades will follow the usual scale:

90-100% = A-range (i..e, A+, A or A-)
80-89% = B-range (i.e., B+, B or B-)
70-79% = C-range (i.e., C+, C or C-)
60-69% = D
0-59% = F
The cutoffs for '+' and '-' grades will be determined at the end of the semester, at the discretion of the intructor. This scale is subject to change at any time, at the discretion of the instructor.

Changes to final course grade: Changes to the final course grade should be requested in writing, within **one week** after receiving the final course grade. The request should contain a technical explanation of why re-grading is necessary. If the request is justified, the instructor will regrade **all the submissions**, including all the homeworks, the exams and the project, to determine the new grade. The new grade can potentially be lower than the original grade.