CS7280: Topics in Statistics and Data Analysis

WF 11:45-1:25

Olga Vitek
WVH310
o.vitek@neu.edu
http://olga-vitek-lab.org/

This is a basic course in applied statistics. The course introduces concepts of statistical modeling, inference, and experimental design that are particularly relevant to computer scientists, and which complement other courses such as machine learning. The topics include the concepts of random sampling, point and interval estimation, hypothesis testing and prediction, and optimal allocation of resources for data collection. The course discusses both frequentist and Bayesian approaches to inference. It emphasizes the challenges and opportunities for statistical inference in situations that involve large and/or complex datasets.

The course is driven by practical examples, hands-on homeworks, and a project. The course will use the programming language R. At the end of the course the students will be able to (1) recognize the problems of statistical nature when working with data, (2) use statistical inference to reason with data, and (3) draw valid conclusions and clearly present the results.

Required topics include:

- Introduction to R
  - R syntax
  - Reproducible data analysis: markdown and Sweave
  - Data management and data visualization with dplyr and ggplot2
- Descriptive data analysis
  - Basic summaries of the data, boxplots and histograms
- Basics of statistical inference.
  - Populations and random samples, sampling distributions, bias and variance
  - Point estimates and confidence intervals
  - Hypothesis testing: two-sample t-test, A/B testing
  - Sample size calculations for hypothesis testing
- One-variable linear regression
  - Parameter estimation and testing
  - Confidence intervals and prediction intervals
  - Joint interval
- Multi-variable linear regression
  - Multicollinearity
  - Interpretation of categorical predictors
  - Selection of a subset of predictors
- Analysis of categorical data
  - Tests of association for count data
  - Statistical inference in logistic regression
  - Sample size calculation for classification
- Extra topics in regression
  - Weighted regression
  - Permutations and bootstrap

  - Multiple testing
- Basics of experimental design
  - Analysis of Variance
  - Allocation of resources in multi-factor studies

Prerequisits:

The course is designed for graduate students in computer science, but is also open to students from other majors. The course is self-contained and can be taken by students without an extensive prior training in statistics. However, the mathematical and computational literacy at the beginner graduate student level (Calculus 2, linear algebra) is expected. Prior exposure to at least one programming language is required. Prior exposure to R is desirable but not required.

Textbooks for the course include

- *Statistical Methods for the Social Sciences*. A. Agresti, 2008. 4th Edition.
- *Applied Linear Statistical Models*. M. H. Kutner, C. J. Nachtsheim, J. Neter and W. Li, McGraw-Hill/Irwin, 2004. 5th edition.

Additional texts for the course include

- *R for data science.* G. Grolemund and H. Wickham, O'Reilly Media, 2016. (Free online)
- *Hands-On Programming with R: Write Your Own Functions and Simulations.* G. Grolemund, O'Reilly Media, 2014.
- *Reproducible Research with R and R Studio.* C. Gandrud, Chapman and Hall/CRC, 2013.

After completing this course, the student will be able to demonstrate the following competencies:

- Recognize the problems of inferential nature and understand the underlying principles
- Use statistical inference in designing experiments and analyzing data
- Draw valid conclusions and clearly present the results.