DA 503
APPLIED STATISTICS

Instructor : H. Sait Ölmez
Office : FENS 1088, Tuzla Campus
Contact : Email: olmez@sabanciuniv.edu / Tel: (216) 483 9578

Teaching assistant : Yalçın Can Kılıç
Contact : yalcinkilic@sabanciuniv.edu

Course schedule

Term : Fall Term, 2021-2022 academic year
Start/End : November 20th, 2021 – January 3rd, 2022
When : Monday 19:00 – 22:00 / Saturday: 13:00 – 16:00

Course description and outline

This course covers the major topics of descriptive and inferential statistics. Course content includes the basic procedures of choosing and conducting appropriate statistical tests for a given research or business problem. This is primarily a lecture course, with a fair amount of in-class and out of class coding work using Python. Students should gain a strong foundation in inferential statistics, including z-tests, t-tests, ANOVA, Chi-Square, Linear Regression and other Non-Parametric tests used in assessing the presence of an effect under investigation. Students will understand the theory behind these methods, and be able to apply them correctly and appropriately to analyze and infer from data, and be able to interpret the results.

Outline

- Statistics: Why do we need it and how do use it?
- Experimental design and setup for data collection
- Pre-processing data
- Exploratory Data Analysis and visualization
- Descriptive statistics using numerical and graphical representations
    - Univariate and Bivariate analysis of data
- A primer on basics of probability
    - Bayes' Theorem
    - Random variables
    - Discrete and continuous probability distributions
- Sampling theory and Point estimates
- Interval estimates (Confidence Intervals)
- Understanding and using Hypothesis Testing (classical vs. computational methods)
    - Concept of significance and p-values
    - Commonly used hypothesis tests
    - Types of errors in hypothesis tests
    - Power of a hypothesis test and the sample size
    - Effect size from a practical standpoint
- Correlation tests

    - Pearson's correlation
    - Spearman's-rank correlation
    - Point-Biserial correlation
    - Phi coefficient/Cramer's V correlation
- Chi-square tests
    - Goodness of fit test
    - Test of independence
- Analysis of Variance (ANOVA)
    - One-way ANOVA
- Simple and Multiple Regression techniques
    - Ordinary least squares and Gradient Descent solution
    - Interpretation of regression coefficients
    - Assessing the overall quality of a regression
    - Assumptions used in building a solution
    - Model selection process
    - Regularization in Regression
        - Ridge (L2) and Lasso (L1) Regression techniques
    - Robust Regression algorithms (if time permits)
        - Regression with outliers
        - Huber Regression, RANSAC Regression, Theil Sen Regression
        - Quantile regression

Course Format

Lectures will be supported with computational methods and programming work using Python. You are expected to work with Jupyter Notebooks during the course. All submissions (assignments, take-home etc.) must be in the .ipynb format uploaded to SUCourse+ as Jupyter Notebooks. File name convention for uploads is `lastname_firstname_hw#.ipynb`.

Course Notes

Course lectures will be uploaded to SUCourse+ in the pdf format. Python notebooks we use in class will also be available on SUCourse+, so you can download and practice on them.

Textbook

There is not a formally assigned textbook for this class. There is a vast amount of text and video material on the Internet. I will, however, suggest the following references:

- OpenIntro Statistics (3rd edition)
    - David M. Diez, Christopher D. Barr and Mine Çetinkaya-Rundel
    - https://www.openintro.org/download.php?file=os3&referrer=/stat/textbook.php
- An Introduction to Statistical Learning
    - G. James, D. Witten, T. Hastie and R. Tibshirani
    - http://www-bcf.usc.edu/~gareth/ISL/
- Think Stats (Probability and Statistics for Programmers)
    - Allen B. Downey
    - http://greenteapress.com/thinkstats/thinkstats.pdf
    - Note: Pythonic part of the book is somewhat useless as it's built on top of custom modules written by the author.

- Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python
    - Peter Bruce, Andrew Bruce, Peter Gedeck, 2020
- Head First Statistics: A Brain-Friendly Guide
    - Dawn Griffiths
- An Introduction to Statistics with Python: With Applications in the Life Sciences
    - Thomas Haslwanter

Assessment

Course grading will be based on the total number of points that the participant receives on quizzes, exams and homework assignments. Breakdown of the (**tentative**) overall course grading is:

- 1 Midterm exam 35%
- 1 Final (or a Take-home) exam 40%
- 3-to-4 Homework assignments 25%

Software:

We will use Python and its relevant libraries (Numpy, Scipy, stasmodels, scikit-learn) throughout the course (exercises, homework assignments and exams).

Course Communication

Course-related announcements will be communicated through SUCourse+. Lectures, handouts, solution sets and Python Notebooks will be posted on SUCourse+. E-mail is my preferred way of communication if you need to reach me as I happen to check my mail on a regular basis.

Collaboration

You are more than welcome, and even encouraged to discuss the problems in your homework assignments with your fellow classmates. You are, however, expected to submit your own work prepared in your own words. You are also expected to comply with the ethical standards of the university and stay away from any sort of dishonesty (cheating, copying somebody else's work, etc.).

Disclaimer

This is a tentative syllabus which is subject to change. Any changes in the syllabus will be announced via email or SUCourse+, or you'll be notified in class.