IE 551 Adv Topics: Statistics for Machine Learning	Instructor: Dr. Hoang Pham
Spring 2024	hopham@soe.rutgers.edu
Office hours: Wednesday 2:00 - 4:00pm	Wednesday, 6-9pm, ARC-206

Project #1

(due 2pm, Wednesday, February 28, 2024)

Following is a brief problem description of group project. Work in groups of 3 members and select one as the team leader. A project proposal (typed) report, no more than tone page, is due by 6pm, Wednesday, January 31, 2024, which includes names of your group members, your group sample data set, and brief outline & directions of the project.

PROJECT DESCRIPTION

Project Theme: Population versus Samples

Population refers to the entire set of groups or individuals that you want to draw conclusions about, while a sample is a subset of the population that you will use to draw conclusions. Ideally, a sample should be randomly selected and representative of the population.

A parameter is a measure that describes the entire population, whereas a statistic is a measure that describes the sample. The key difference between a population parameter and a sample statistic is that the former describes the entire population, while the latter describes only a sample from that population. Population parameters are, obviously, more precise and accurate, as they are calculated using the entire population data.

You can use statistical inferences or hypothesis testing to estimate the likelihood that a sample statistic differs from the population parameter.

A **sampling error** is the difference between a population parameter and a sample statistic. In your project, for example, the sampling error represents the difference between the mean median household income based on your sample dataset and the true mean median household income of the entire population in the United States.

The goal of your project is to develop methods, models, algorithms, etc., to generalize findings from the sample your group obtained in class to the population data. This aims to minimize or keep the results of the sampling error very low.

Specifically, the data (in the box today in class, 1/24/24) contains the median household income for the entire United States and each state and county for the recent year with a total of approximately 4000 observations. Your group is tasked with drawing a sample of size 50 from the entire population dataset.

In this project, your group is tasked with developing methods, models, algorithms, etc., to generalize findings from the sample obtained in class to the entire population data. The objective is to minimize sampling errors based on certain criteria. Please provide a detailed demonstration of how you have applied various techniques, methods, and approaches learned in this class to the project.

PROJECT REPORT: Due by 2pm, Wednesday, February 28, 2024

A typed REPORT (no more than 8 pages) is due by 2pm, Wednesday, February 28, 2024

Your Report must include the following:

Introduction, Objectives, & Problem Statements
Data Collection & Analysis
Methodologies, Modeling Analysis, and Results
Conclusions and Findings.

PROJECT PRESENTATIONS: Wednesday, February 28, 2024

Project Grading:

- Each group will receive a single total group score consisting of the final Report (90%) and oral presentation and peer comments (10%). Individual student grades will be assigned based on self and teammate confidential peer review comments (20%). For example, if your group total score is 90 and your peer review score is 95% of total score. Then your final score is: 90 (0.8) + (0.95) (90) (0.2) = 89.1
- Every one of the team members should understand and should be capable of explaining every problem, data collection and analysis, and results and findings in your group report.
- The presentation is 12 minutes long including Q&A on Wednesday, 2/28/24, so it is important to plan and practice the presentation in advance. The 12-minute time limit will be strictly enforced because we need to allow sufficient time to accommodate all of the presentations. If your presentation is too long and you are not able to give your ending summary, your grade will suffer.