Math 540: Intro to Probability Theory
Individual Project Standards

Instructor: Hadi Safari Katesari **Email:** hsafarik@stevens.edu

Individual project work in this course is intended to stimulate creative thinking and to develop the skills of independent analysis of real data problems.

- For the final project, you address some questions that interest you with the probabilistic and statistical methodology we learn in this course. You choose the question; you decide how to collect data; you do the analyses. The questions can address almost any topic, including topics in finance, Business, psychology, sociology, natural science, medicine, public policy, sports, law, etc.
- The final project requires you to synthesize all the material from the course. Hence, it's one of the best ways to solidify your understanding of statistical methods. Plus, you get answers to issues that pique your intellectual curiosity.
- Choosing appropriate probabilistic/statistical techniques and providing evidence that the suggested methods are suitable for the chosen data.
- Providing conclusions on your analysis.

The cover page of your project report should contain the year, name of the course, topic of your project and name of the student. The project should explain and justify all steps of your analysis. Include necessary references, program code and outputs to facilitate the reproducing of your results. **You should limit your report to 45 pages**.

You should get started on the project as early as possible, particularly in thinking about procuring data and collecting background information. Keep in mind that by the end of lectures, you will have learned many probabilistic/statistical techniques, not limited to: combinatorial analysis, axims of probability, conditional probability and independence, continuous and discrete random variables, jointly distributed random variables, properties of expectation, limit theorems, simulations, additional topics in probability, dimension reduction methods such as factor analysis. These techniques will help you address your question of interest. The format of individual research project submissions should be only ONE pdf file.

Project Formation

Here is the main (not all) formation of projects that is usual in internship and full-time positions in industry. The details of the material which should be presented in your projects are as follows. You need to have an appropriate justification if you want to leave out one or more items below. This formation is recommended to be used in the course project:

- Title: A proper project title describes the whole assignment in one sentence. It helps the team to refer the Project with the assigned Name. Project titles makes you to understand the main goal of the Project work and deliverables.

- Abstract: An abstract is a brief summary of your project: for Endeavor abstracts, fewer than 350 words. It is used typically for industrial/academic presentations to give the reader a synopsis of the research project, and it can also be used to summarize a creativity project.

- Chapter 1: Introduction (maximum 1 page): A project introduction is a paragraph or paragraphs explaining what a project is about. It should include key details about the project that give the reader enough information to understand the purpose and scope of the project. Start with the problem. Describe the general aspects of your project, and why it matters to do what you have done in your project.

- Chapter 2: Data Description (maximum 3 pages): You need to explain about your data here. If you are using a dataset from online resources, then you need to provide the link of dataset in this part. If you collected your dataset, then you need to explain how you collect it. Introduce your data, and where you have gotten it from, or how you have collected it (in the case that you have collected your own data). Perform some descriptive analysis on your data including, but not limited to, the plot of the data. You can provide tables, or graphs to familiarize the reader with your data. This is the place to discuss statistical tools, outliers, missing values and/or any problems existing in your data (e.g. missing values). Provide your solutions to these problems at this time. Please just provide reference (the online source link of your dataset(s)) here. Please, do NOT share csv file (or other formats) of your dataset(s) with Google Drive (or other clouds). Your dataset(s) should be unique and should not be analyzed anywhere based on what we are covering in this course. Moreover, there should not be any overlap with the datasets of other students in this sheet. Please check the VIEW-ONLY spreadsheet here. Priority is with the one who added the link sooner. If you are collecting data by your own, then just type "collect". You have access to the form through Stevens Email not from your personal Email such as Gmail. Please do not request access through your personal Email. Failure to satisfy these conditions result to a penalty for your project. Your descriptive statistics should be addressed here.

- Chapter 3: Methodology (maximum 3 pages): Essentially, a methodology is a collection of methods, practices, processes, techniques, procedures, and rules. For this project you need briefly explain about mean, variance, standard deviation, mode, first quantile, third quantile, joint distributions, conditional expectations, Bayes' rule, discrete and continuous random variables, order statistics, correlation, Markov Chains, simulation techniques, factor analysis, and other techniques that you are applying here.

- Chapter 4: Analysis and Results (minimum 20 pages): This is the main part of your project. The analysis of your project should be explained in this part. The results of your analysis should be provided. Do not forget to interpret the results in this part which has a high weight in your project.

  ### 4.1 Simulation Data Analysis

  Start by understanding the basics of simulation, its purpose, and how it's used in various fields. You can then choose a specific application area (e.g., finance, healthcare, transportation) and explore how simulation can be applied to solve a problem in that domain. This could involve modeling and simulating a simplified version of the problem. Keep in mind that simulations are applied to solve real-world problems.

  4.1.1 **Simulating Continuous Random Variables:** Simulating continuous random variables, such as normal, exponential, uniform, gamma, lognormal, pareto, beta and other distributions. You should implement various techniques discussed in the book, like inverse transform sampling and acceptance-rejection sampling. This could involve generating data that follows these distributions and analyzing the results.

  - **Statistical Analysis:** Calculate measures such as mean, variance, standard deviation, first quantile, third quantile, mode, order statistics skewness, and kurtosis. This will help you understand the properties of the continuous distribution.

- **Visualization:** create appropriate visualizations of the simulated data. You can generate histograms, density plots, or box plots to visualize the distribution's shape and characteristics. Visualization can provide insights into the data's behavior.
- **Central Limit Theorem Verification:** Verify the Central Limit Theorem by taking random samples from the simulated data and calculating sample means. You can observe how the sample means approximate a normal distribution and understand the importance of the CLT in statistics.
- **Outlier Detection:** Identifying potential outliers in the simulated data. You can use various outlier detection methods and assess whether the outliers conform to expectations for a continuous distribution.
- **Probability Calculations:** Calculate probabilities related to the continuous (such as normal) distribution. For example, you can find the probability that a randomly selected value falls within a specified range or above/below a certain threshold.

4.1.2 **Simulating from Discrete Distributions:** Simulating data from discrete probability distributions like the binomial, Poisson, geometric, hypergeometric, discrete uniform, negative binomial, Zeta (Zipof) and other distributions. You should explore techniques such as the inverse transform method or convolution. The goal is to generate data that follows these discrete distributions and analyze the simulated outcomes.

- **Statistical Analysis:** Calculate measures such as mean, variance, standard deviation, first quantile, third quantile, mode, skewness, and kurtosis. This will help you understand the properties of the discrete distribution.
- **Visualization:** create appropriate visualizations of the simulated data. You can generate histograms, density plots, or box plots to visualize the distribution's shape and characteristics.
- **Central Limit Theorem Verification:** Verify the Central Limit Theorem by taking random samples from the simulated data and calculating sample means. You can observe how the sample means approximate a normal distribution and understand the importance of the CLT in statistics.
- **Outlier Detection:** Identifying potential outliers in the simulated data. You can use various outlier detection methods and assess whether the outliers conform to expectations for a discrete distribution.
- **Probability Calculations:** Calculate probabilities related to the discrete (such as binomial) distribution. For example, you can find the probability that a randomly selected value falls within a specified range or above/below a certain threshold.

4.1.3 **Markov Chains:**

- **Transition Matrix Simulation:** Create a simulation of a simple Markov chain with a specified transition matrix. You can model transitions between states and calculate the state probabilities after a certain number of steps.
- **Recurrent Events:** Model recurrent events using a Markov chain. This can involve simulating events like customer arrivals, service times, and departures in a queueing system.
- **Ergodicity:** Simulate a Markov chain and comparing the time-averaged behavior to the state probabilities. This helps you understand the connection between long-term behavior and steady-state probabilities.
- **Sensitivity Analysis:** Perform sensitivity analysis by simulating Markov chains with varying transition probabilities or initial conditions. You should assess how small changes in parameters affect the system's behavior.
- **Visualization:** Visualize the behavior of the Markov chain by creating state transition diagrams, probability heatmaps, or time series plots to illustrate the evolution of the system over time.

4.1.4 **Variance Reduction Techniques:** Investigate variance reduction techniques in simulation, such as importance sampling, control variates, and antithetic variates. You can apply these techniques to a specific problem or simulation scenario. For example, you can model a complex system and use variance reduction techniques to improve the efficiency and accuracy of your simulations.

4.1.5 **Comparison of Different Simulation Methods:** Compare and contrast different simulation methods (e.g., Markov Chain, Variance Reduction) for solving a specific problem. This can help you understand the advantages and disadvantages of various simulation techniques.

4.1.6 **Simulation for Combinatorial Analysis:** Conduct a simulation for a combinatorial problem, such as simulating various card games or counting the number of different paths in a graph. This can help you understand the principles of probability through practical applications.

4.2 Real Data Analysis

4.2.1 **Bayes' Theorem:** Analyze a dataset involving Bayes' theorem, such as Bayesian inference in medical testing or spam email detection.

4.2.2 **Joint Distribution Analysis:** Find a real dataset where conditional probability and independence concepts can be applied. For example, you can explore conditional probability in medical diagnoses, A/B testing in marketing, or independence in customer preferences. Visualize the correlation between random variables. Conduct normality tests: Apply normality tests to check whether the simulated data follows a normal distribution. You can use statistical tests like the Kolmogorov-Smirnov test or the Shapiro-Wilk test and interpret the results.

4.2.3 **Factor Analysis**

The above-mentioned topics are mandatory to cover the materials of the course. Therefore, put all your efforts to cover these materials.

- Conclusion (maximum 1 page): Provide your findings in the context of the project. Fore example, highlight the implications/interpretations that you have come to during your modelling/forecasting process in the context of the project. You should avoid technical (mathematical/statistical) language as much as possible in this part of the project. This is the section in reports which is usually read carefully by managers, who may not have any statistical background. Compare your models, and pick the best candidate with respect to some criteria. This is where you provide your statistical conclusion on the models.

- References (maximum 1 page): Any sources including the dataset which is used in the project should be referred here.

Project Grades will be based on the following:

a) Data selection, writing in Markdown and following the structure of project (10%)

b) Presentation (10%) c) Chapter 4: (60%) d) Other chapters (20%)

If chapter 4 of your project is satisfactory (which is assessed by your instructor), then the grades of other chapters will be given to you. Otherwise, you will not be given any grades of the other chapters. Please implement the project report in markdown (jupyter notebook using markdown or R Markdown) including code, output and interpretation. Word file (and screenshots from code) is not accepted. **Interpretation is one of the most important part.** The following options are not necessary in this project but you may use it for your future jobs: Cover Page & Title Page, Bonafide Certificate from the project supervisor(s), counter signed by the HoD/Division or Group Head Declaration by author(s), Table of Contents, List of Symbols, Abbreviations and Nomenclature, Appendices.

Some On-line Data Sources

You can, but not limited to, check the following links for data.

- https://guides.emich.edu/data/free-data
- https://infoguides.gmu.edu/find-data/practice
- https://www.mat.univie.ac.at/ neum/statdat.html
- https://www.itl.nist.gov/div898/handbook/datasets.htm
- https://methods.sagepub.com/Datasets
- https://www.sheffield.ac.uk/mash/statistics/datasets
- link

GO DUCKS!