DS201: Statistical Programming

Dr. Anil Kumar Sao

Assignment 1

16.01.2025

Instructions for Submission: You can submit your solution as a Jupyter Notebook/Matlab file with comments and discussions on the results obtained in each step.

1. Follow Standard Report Format: Include sections like Introduction, Data, Methodology, Results, Discussion, and Conclusion.
2. File Naming Convention: Adhere to the specified naming convention for each file you submit (e.g., RollNumber FirstName Asg1).
3. Refrain from using zip files. If necessary, submit multiple files.
4. Include comments in the code explaining the logic and any assumptions made.
5. Include References: Cite any external sources or references used in your assignment.
6. Code Quality: Ensure your code follows best practices, is well-organized, and avoid plagiarism as a plagiarism check will be conducted.
7. Be aware that late submissions are not permitted; ensure timely submission.
8. Coding can be done in any language.

1. A publishing firm wants to develop special printing machines for English. For this, they need to determine the probability of occurrence of specific letters and words. You are given two large text files (fileA and fileB).

NOTE: characters like whitespace, special characters and punctuation are to be omitted.

(a) Determine the probability of each alphabet in the English language. Upper-case and lower-case alphabets are considered the same. List the top ten alphabets that occur in fileA.

(b) The measure of uncertainty is determined by its entropy. Entropy should be computed as

$$H = \sum(-p_i \log p_i)$$

where p_i is the probability of event i. If we consider the occurrence of alphabets in English as events of interest, determine the entropy. In other words, determine the uncertainty of alphabets in the English language using fileB.

(c) Repeat (a) and (b) if word is considered as an event and use fileC and fileD for analysis.

2. Imagine that you are playing with a random number generator that produces values between 0 and 1, perfectly spread (uniformly distributed). Now, what if you could generate n of these numbers, calculate their mean and variance, and then watch how these statistics change as n grows larger and larger? Does randomness settle into predictable patterns as you generate more numbers? Try experimenting with different values of n. What do you observe and why do you think it happens?

3. Repeat the same experiment as in Question 2 but for a Gaussian distribution having mean and standard deviation as 4 and 3 respectively.

NOTE: Random number, having a distribution (Gaussian and uniform), can be generated using inbuilt function.