DS201/DSL253: Statistical Programming

Assignment 02

23/01/2025

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.

Question 1

Imagine you are studying two distinct process that generate random numbers between 0 and 1, modeled as continuous random variables X with different distributions:

a. The first process generates numbers, following an exponential distribution with $\lambda = 2$.

b. The second process generates numbers uniformly distributed between 0 and 1.

You collect n random numbers from each process and define a new random variable, Y, given by $Y = F_X(x)$ where $F_X(x)$ represents the Cumulative Distribution Function or CDF of X. Derive the Probability Distribution Function (PDF) of Y for both processes and Observe the histogram of generated n numbers for both processes for different values of n. Write down your observations.

Question 2

Imagine you are an archivist analyzing a dusty old text file from a forgotten library. Your task is to uncover the hidden patterns in the text by doing the following:

a. Count the Words: Create a histogram of how often each word appears.

b. Focus on the Key Players: Identify the top 30 most frequent words from your list.

Once you have this list of important words, calculate the Cumulative Distribution Function (CDF) for their frequencies. Use this CDF as a transformation function to remap the word frequencies. What new insights can you uncover about the text after applying this transformation?

Question 3

Imagine you are working with a random number U, which is drawn from a uniform distribution between 0 and 1. You have a tool that allows you to transform this number by using the inverse CDF of a distribution X. By applying this transformation, you create a new number Y.
Now, your task is to determine what kind of random variable Y becomes after the transformation. Consider the following two cases:

a. When $X \sim$ Exponential

b. When the PDF of distribution is as follows: