BONTA AALAYA
12340530
DSL253 - statistical programming

1

| Introduction

In this experiment, we study 2 distinct processes which generate a random number between 0 and 1 (continuous random variable X) with different distributions, and try to conclude the behaviour of the probability density function for different sample sizes.

| Data

The dataset is obtained from inbuilt functions in Matlab

| Methodology

Matlab extension ‘Descriptive Statistics and Visualization’ consists inbuilt functions to generate exponential and uniform distributions and also has an in-built ‘cdf’ function to calculate cumulative distribution functions of bot distributions across varying sample sizes. Without this function, we can compute

For exponential distribution $X \sim Exp(\lambda)$

CDF $F_X(x) = 1 - e^{-\lambda x} for\, x \geq 0$

Hence PDF $Y \sim U(0, 1)$ [1]

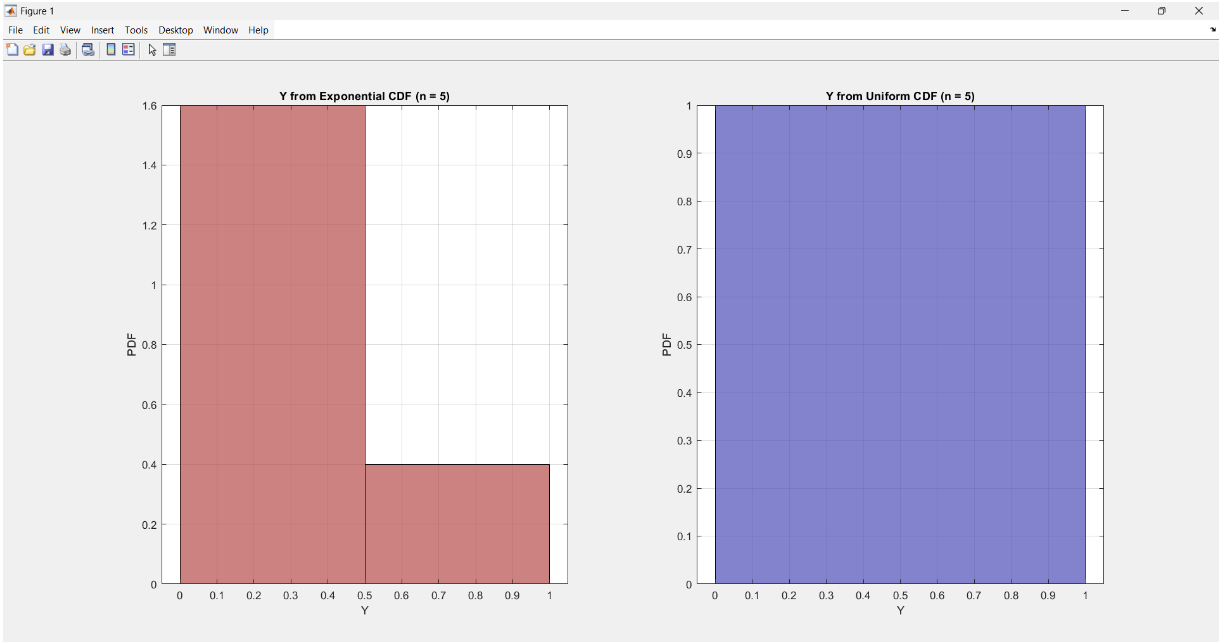
For uniform distribution $X \sim U(0, 1)$

CDF $F_X(x) = x \;\; for \;\; x \in [0, 1]$

Hence PDF $Y \sim U(0, 1)$ [2]

| Results

The following are the histograms obtained (Exponential left, Uniform right) in ascending order of sample size



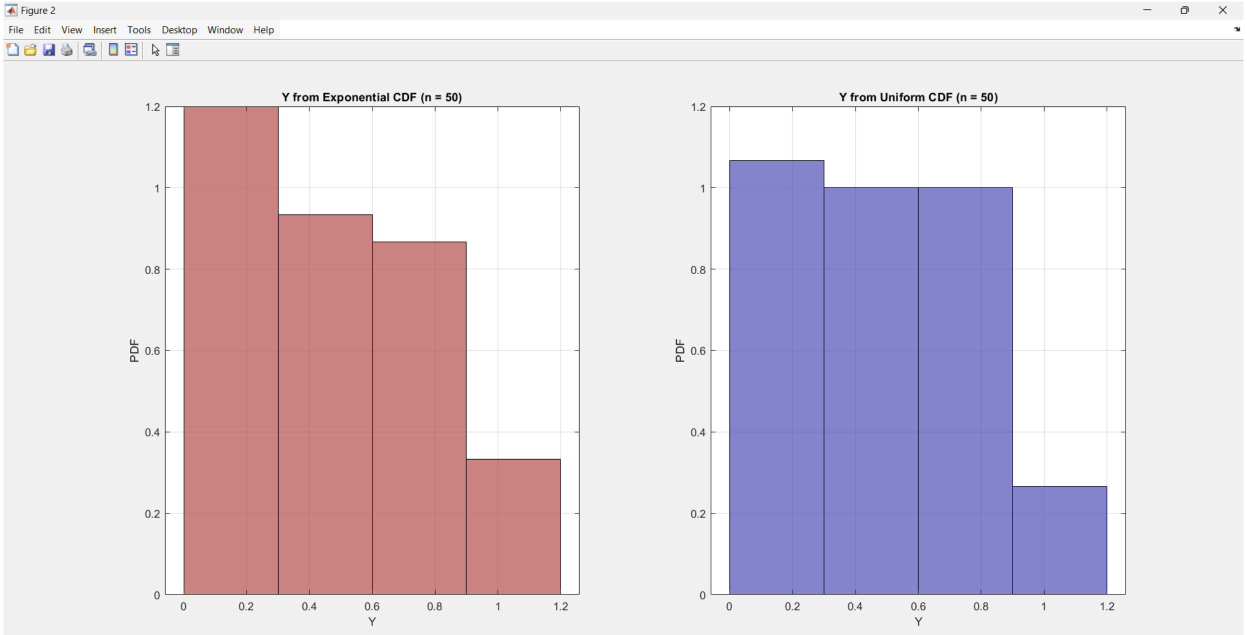


Figure 2
File Edit View Insert Tools Desktop Window Help
Y from Exponential CDF (n = 50)
PDF
0 0.2 0.4 0.6 0.8 1 1.2
Y
0 0.2 0.4 0.6 0.8 1 1.2
Y from Uniform CDF (n = 50)
PDF
0 0.2 0.4 0.6 0.8 1 1.2
Y
0 0.2 0.4 0.6 0.8 1 1.2

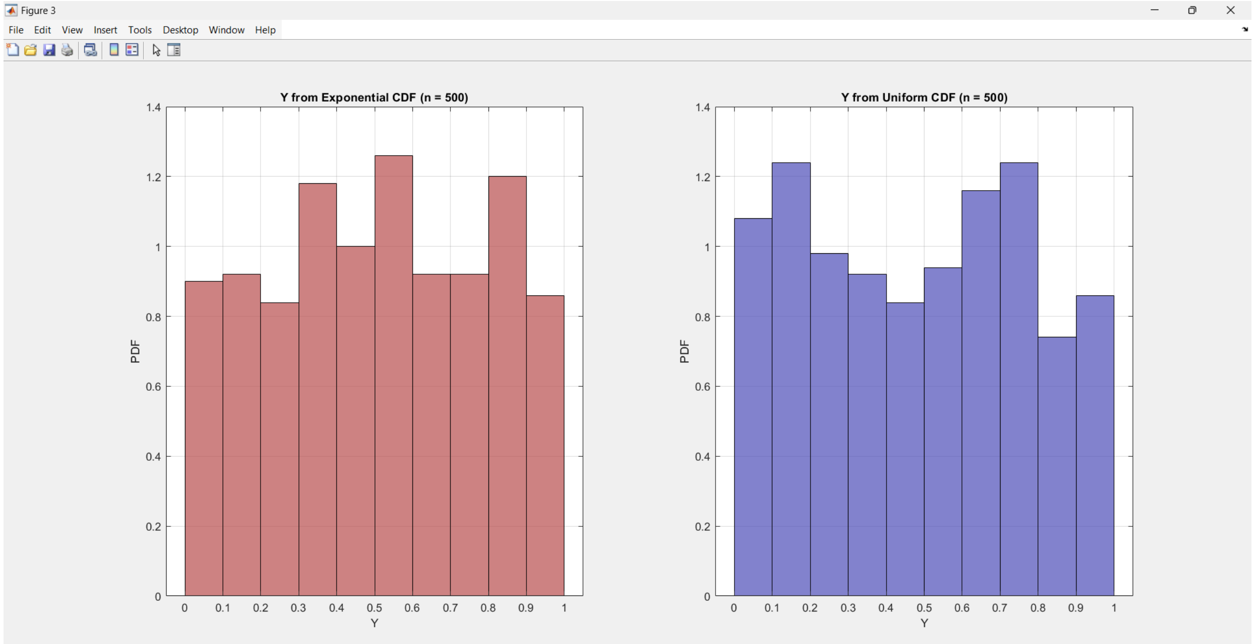


Figure 3
File Edit View Insert Tools Desktop Window Help
Y from Exponential CDF (n = 500)
PDF
0 0.2 0.4 0.6 0.8 1 1.2 1.4
Y
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Y from Uniform CDF (n = 500)
PDF
0 0.2 0.4 0.6 0.8 1 1.2 1.4
Y
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

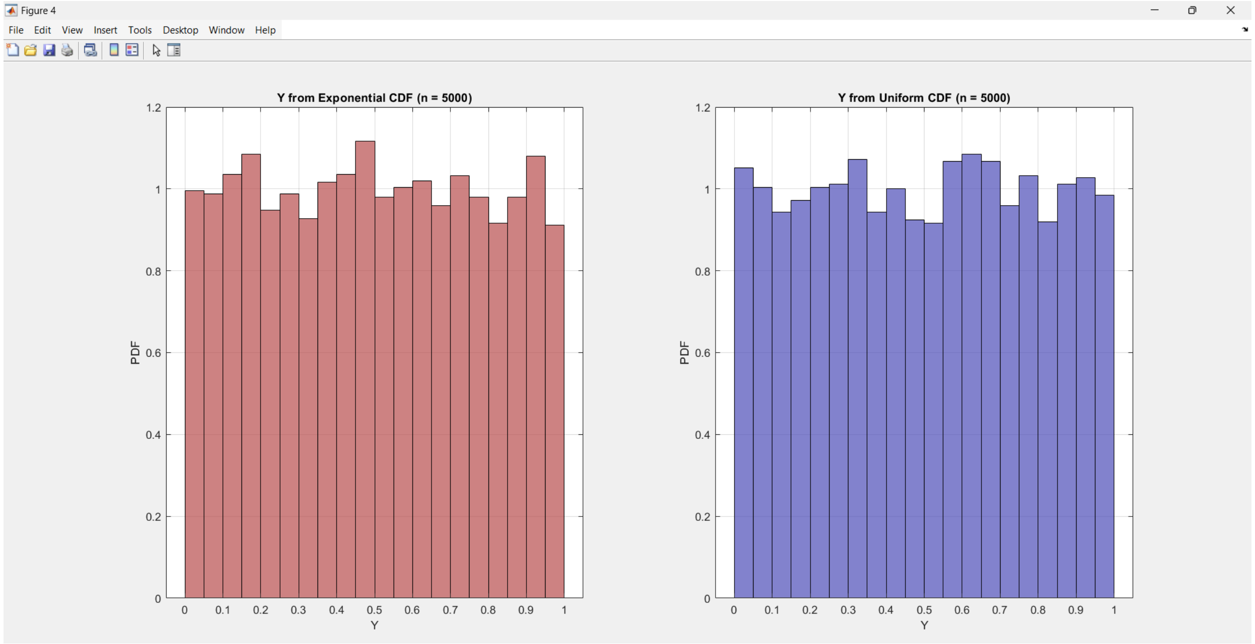


Figure 4
File Edit View Insert Tools Desktop Window Help
Y from Exponential CDF (n = 5000)
Y from Uniform CDF (n = 5000)
PDF
0
0.2
0.4
0.6
0.8
1
1.2
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Y

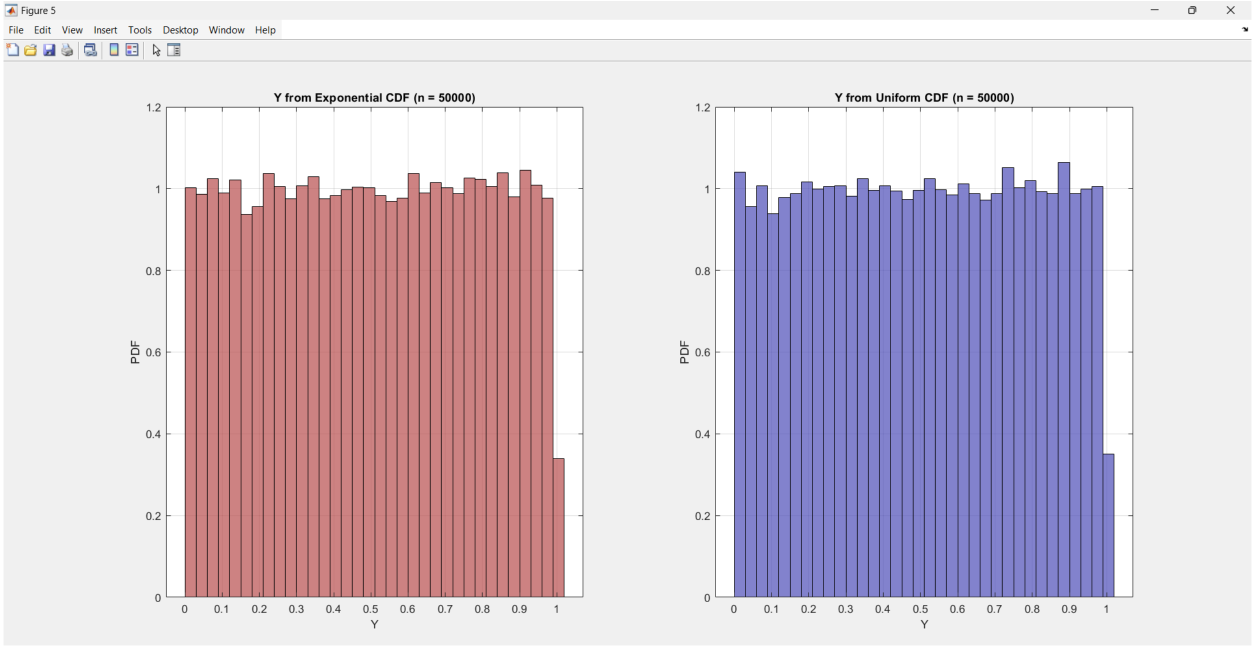


Figure 5
File Edit View Insert Tools Desktop Window Help
Y from Exponential CDF (n = 50000)
Y from Uniform CDF (n = 50000)
PDF
0
0.2
0.4
0.6
0.8
1
1.2
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Y

| Discussion

We can infer from the histrograms that for Exponential distribution, for smaller sample sizes, the original X values are heavily skewed towards values lesser than 0.5. This evens out when greater sample sizes are taken into consideration. Whereas uniform distribution remains consistent across given range [0,1]

| Conclusion

I hereby conclude that CDFs are extremely useful to transform random variables, showing how any random variable can be standardized. While the exponential distribution was initially skewed, its CDF becomes identical to that of uniform distribution when a larger sample size is taken

| References

[1]https://www.math.umd.edu/~millson/teaching/STAT400fall18/slides/article13.pdf
[2]https://courses.grainger.illinois.edu/bioe582/fa2017/Lecture_10_Continuous_Probability_Distributions_Unifiorm.pdf

2

| Introduction

In this experiment, we observe the change in word-frequency histogram after undergoing CDF transformation, and try to understand how Cummulative Density Function affects word frequency directly, and the effect it has on the histogram.

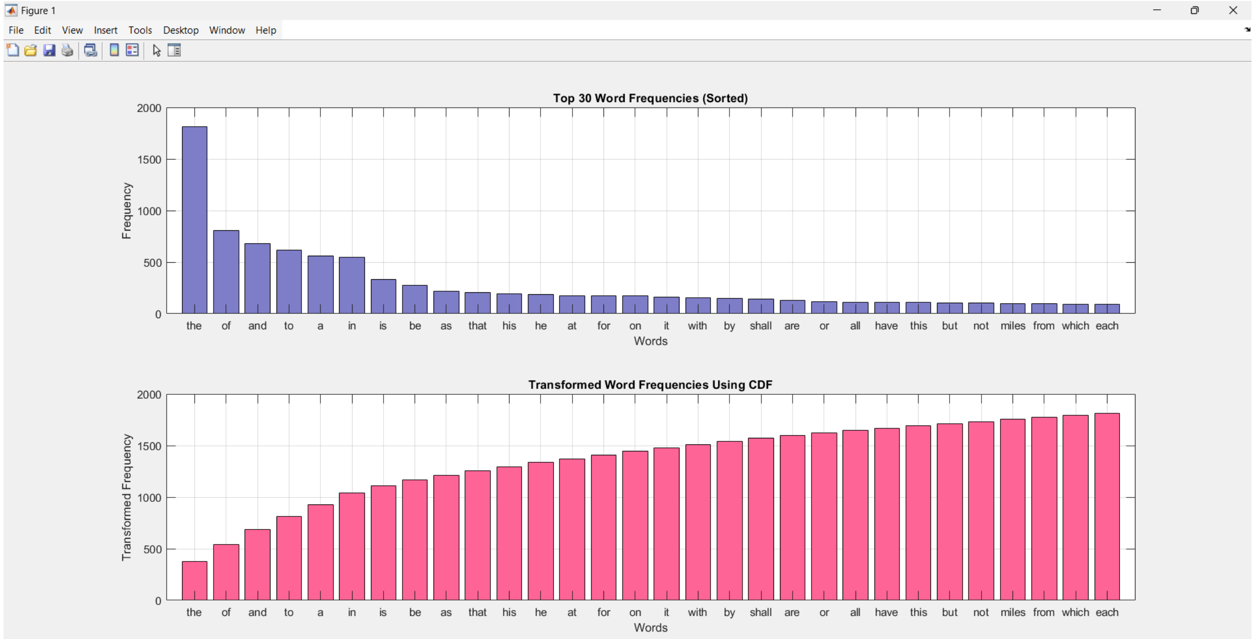
| Data

The dataset is available as text_file.txt in the assignment zip folder

| Methodology

Similar to the previous assignment, we remove all special characters, convert all letters into lowercase, and create a dictionary with all the words. Then, we select the unique words and count the frequency of each word's occurrence and index it with the corresponding word. We then sort the array in descending order and plot histograms of the top 30 words. CDF is calculated by using the cdf function again and plotted against the words.

| Results



| Discussion

The pattern (topmost words being ‘the’, ‘and’, and ‘of’) suggests that the prose is likely formal or narrative, using plenty of function words. The transformation distributes the word frequencies more evenly, emphasising the cumulative importance of less frequent words than dominant ones like "the." This process reduces the skewness caused by a few common words, drawing attention to the diversity of the vocabulary in the text.

| Conclusion

The original/raw frequencies confirm that the text heavily relies on common stop words. (The dataset of the text can be further adjusted if the focus is on the more meaningful words and not said stop words). The CDF transformation moves the focus away from the dominance of individual words, highlighting instead how less frequent words shape the text structure collectively.

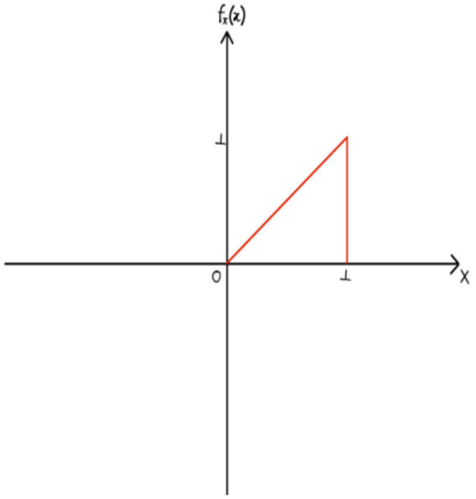
3

| Introduction

In this assignment, we aim to compute and visualise a random variable Y, which can be obtained by transforming a uniform distribution $U \sim (0, 1)$

| Data

The dataset is obtained from inbuilt function ‘rand’ in Matlab. The triangular distribution used is given in the question as follows:



| Methodology

Uniform distribution $U \sim (0, 1)$ is generated using rand, and inbuilt function ‘icdf’ is used to map U to both Exponential and the given triangular distributions. The attributes ‘l, m, and h’ are ‘low, mode, and high’, and the comparison line in final plot is generated using the first portion of [1]

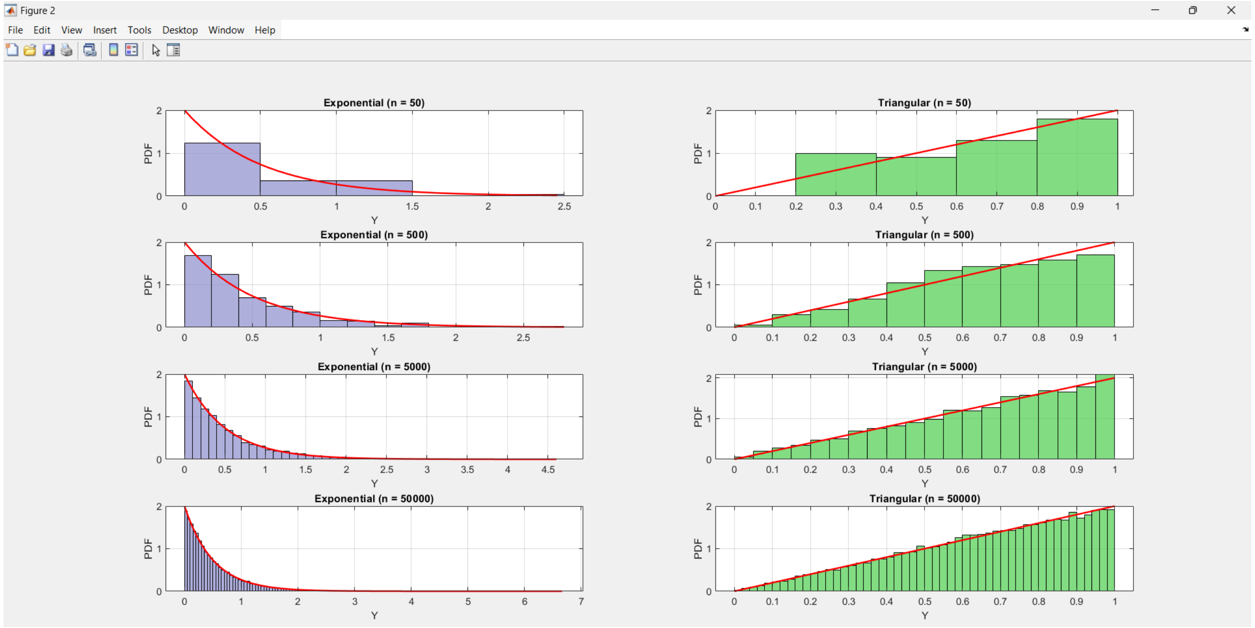
$$\boxed{f(x) = \frac{2(x-l)}{(m-l)(h-l)}} \quad l <= x < m$$

$$\boxed{f(x) = \frac{2(h-x)}{(h-m)(h-l)}} \quad m <= x <= h$$

The inverse cdf of an exponential distribution $f_X(x, \lambda) = \lambda e^{-\lambda x}$ is $F_X^{-1}(x) =- \frac{ln(1-x)}{\lambda}$ [2]

| Results

The plots obtained by considering the pdfs are given below



| Discussion

As sample size increases, both transformations converge towards the theoretical line (red line). This process effectively demonstrates the utility and flexibility of inverse CDF in generating samples from arbitrary probability distributions.

| Conclusion

This exercise demonstrates how the inverse CDF method is a universal tool for converting uniform random samples into samples from a target distribution, as long as the CDF of the target is known. By understanding and applying these transformations, we see the broader utility of the inverse CDF method in probability, statistics, and simulation, enabling us to effectively mimic real-world behaviours and patterns in random data generation

| References

[1] https://www.mhnederlof.nl/triangular.html
[2] https://www.mathworks.com/help/stats/exponential-distribution.html