Assignment 2

Submitted to Dr. Anil Kumar Sao
DSL253-Statistical Programming
Prepared by Amay Dixit - 12340220

Notebook Link:

https://colab.research.google.com/drive/1rdZNzGLHdncckshKBR8HCNbpOmidzcbp?usp=sharing

Github Link:

https://github.com/amaydixit11/Academics/tree/main/DSL253/assignment_2

Docs Link:

https://docs.google.com/document/d/11GOSvIFJVevhy8SjSUdMMR9G0u3opLGJ-fZ5DL6n6RA/edit?usp=sharing

INTRODUCTION

This report presents a comprehensive analysis of three distinct statistical problems involving probability distributions, random number generation, and text analysis. The study explores the behavior of transformed random variables, analyzes text patterns through word frequency analysis, and investigates inverse CDF transformations.

DATA

Question 1: Random Number Generation Processes

- Process 1: Exponential distribution (λ = 2)
- Process 2: Uniform distribution [0,1]
- Sample sizes tested: n = 100, 1000, 10000, 100000, 1000000

Question 2: Text Analysis

- Text source: External text file
- Processing: Lowercase conversion and word tokenization
- Focus: Top 30 most frequent words

Question 3: Inverse CDF Transformation

Data Collection

- Text source: External text file
- Processing: Lowercase conversion and word tokenization
- Focus: Top 30 most frequent words

METHODOLOGY

Question 1: Random Number Generation Processes

Theoretical Derivation of Probability Distribution Transformation

Part 1: Exponential Distribution (λ = 2)

- Given X ~ Exponential(λ = 2)
- PDF of X: $f_X(x) = 2e^{-2x}$, x ≥ 0
- CDF of X: $F_X(x) = 1 - e^{-2x}$,
- Transformation: Y = $F_X(x)$,
- Derivation steps:
  - $Y = 1 - e^{-2X}$,
  - $f_Y(y) = f_X(x) \mid dx/dy \mid$
  - We get, $f_Y(y) = e^{-2x}/(1 - y) = 1$
  - Mathematical analysis shows Y is uniformly distributed in [0,1]

Part 2: Uniform Distribution (X ~ U[0,1])

- PDF of X: $f_X(x) = 1$ 0 ≤ x ≤ 1
- CDF of X: $F_X(x) = x$
- Transformation: Y = $F_X(x) = X$
- Resulting Y is uniformly distributed in [0,1]

Empirical Verification

- Generated random samples for n = 100, 1000, 10000
- Applied CDF transformation
- Visualized distributions using histograms

Question 2: Text Analysis

Text Preprocessing and Linguistic Feature Extraction

Preprocessing Techniques

1. **Text Normalization**
    - Convert entire text to lowercase
    - Remove non-alphabetic characters
    - Tokenize into individual words
    - Eliminate potential counting biases
2. **Frequency Analysis**
    - Utilize computational linguistics techniques
    - Implement frequency counting algorithms
    - Extract statistically significant word patterns
3. **Probabilistic Word Distribution**
    - Compute word frequency distributions
    - Identify top 30 most frequent words
    - Calculate cumulative distribution function (CDF)

Visualization Strategies

- Frequency distribution histogram
- Cumulative distribution function plot
- Comparative word frequency visualization

Question 3: Inverse CDF Transformation

Theoretical Foundation of Distribution Transformation

Let U ~ Uniform(0,1) and X be a random variable with continuous cumulative distribution function (CDF) FX(x). Then the random variable Y defined by:

$$Y = F^{-1}{}_{X}(U)$$

is distributed according to the same distribution as X.

For a given random variable X with CDF $F_{X}(X)$:

1. Generate U ~ Uniform(0,1)
2. Compute Y = $F^{-1}_{X}(U)$

Transformation Approaches

Case 1: Exponential Distribution

- Let X ~ Exponential(λ)
- CDF: $F_{X}(X) = 1 - e^{-\lambda x}$, x ≥ 0
- Inverse CDF: $F^{-1}_{X}(U) = - ln(1 - u) / \lambda$
- Transformation: Y = $- ln(1 - u) / \lambda$

Case 2: Triangular

- CDF: $F_{X}(X) = x^{2} / 2$, x ≥ 0
- Inverse CDF: $F^{-1}_{X}(U) = \sqrt{2x}$
- Transformation: Y = $\sqrt{2x}$

Verification Methods

- Theoretical PDF comparison
- Empirical distribution analysis

RESULTS AND ANALYSIS

Question 1: Distribution Transformation Results

The analysis revealed several key findings:

1. Exponential Distribution Transformation:

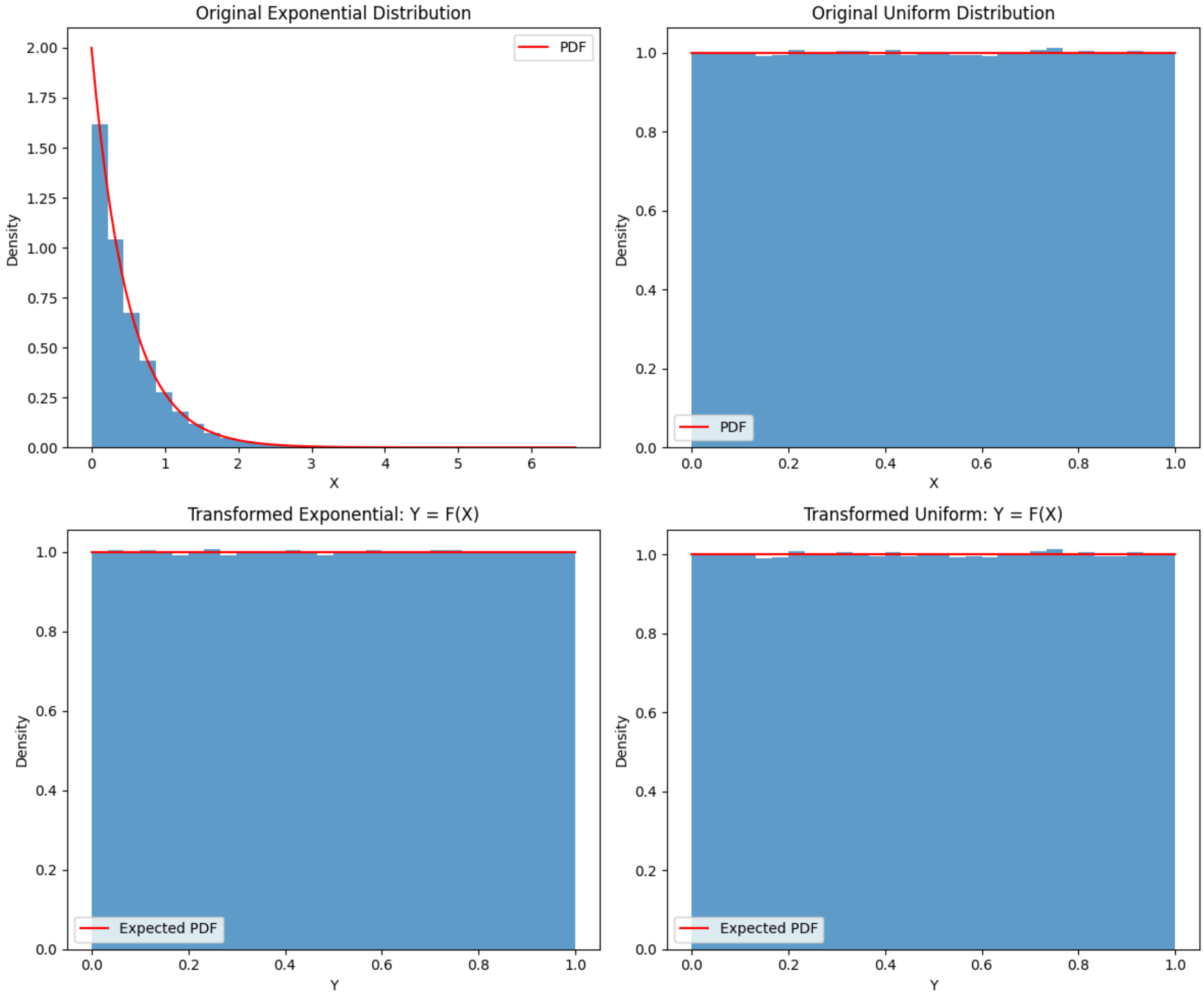
- Original exponential distribution showed expected right-skewed shape

- Post-transformation distribution approached uniform [0,1]

2. Uniform Distribution Transformation:

- Original and transformed distributions maintained uniformity

As n increased, the distributions showed more stable and expected behavior



Question 2: Text Analysis Results

The word frequency analysis revealed:

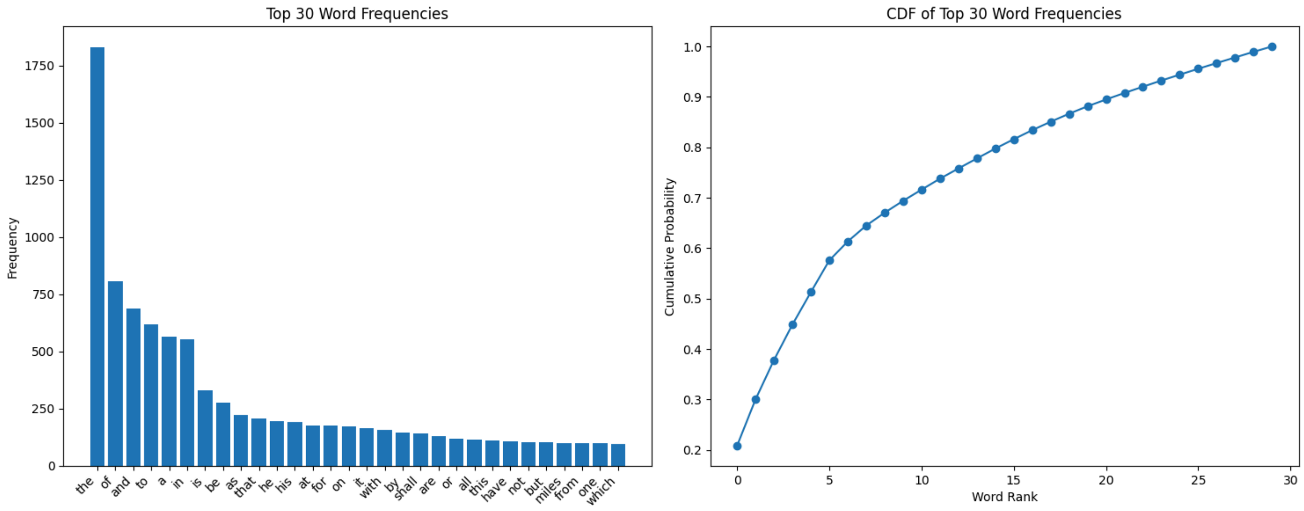
1. Distribution Characteristics:

- Clear power-law behavior in word frequencies
- Top words showed expected linguistic patterns, like greater usage of articles

and conjunctions
- CDF transformation provided normalized frequency scores

2. Key Observations:

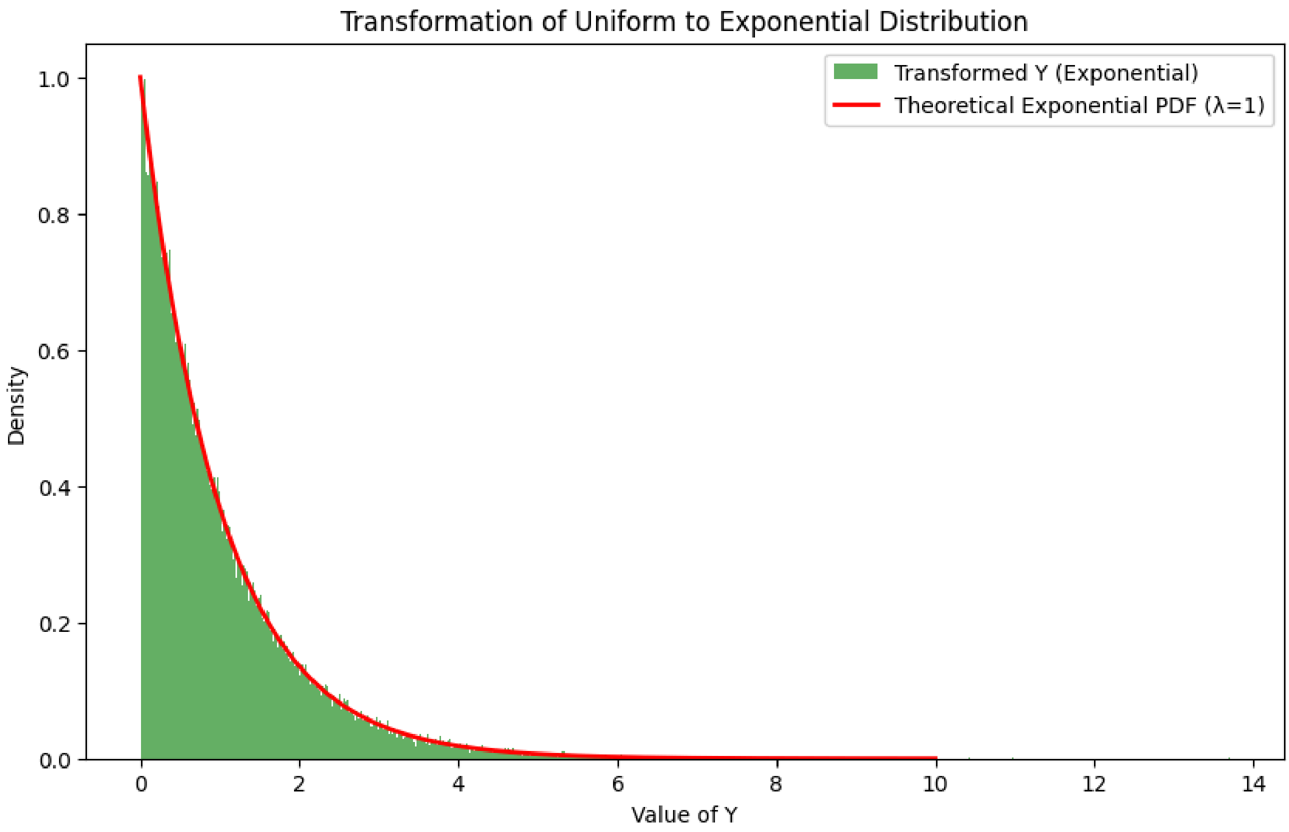
- Common function words dominated the frequency list
- CDF transformation provided relative importance metrics
- Smoothing effect observed in the cumulative distribution



Question 3: Inverse CDF Analysis

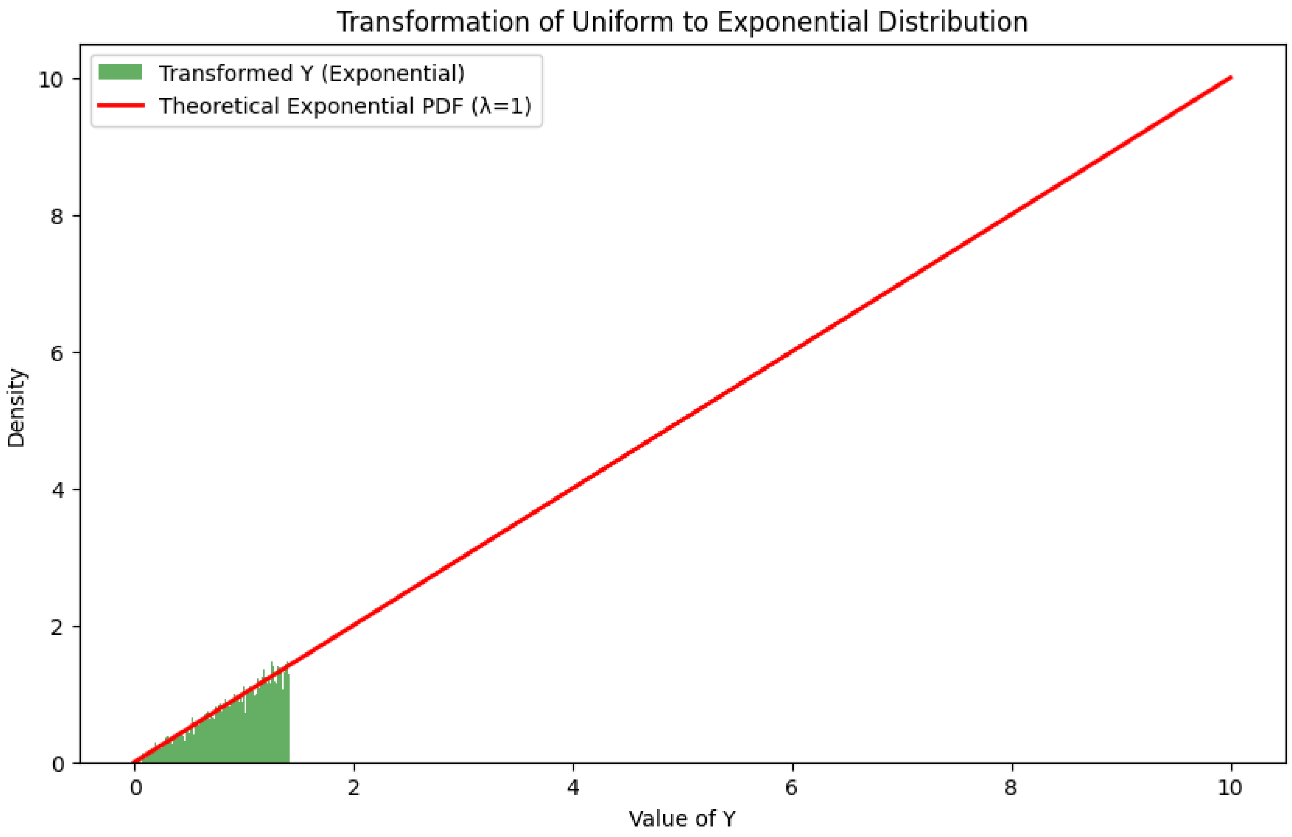
1. Exponential Case:

- Successfully transformed U[0,1] to exponential distribution
- Theoretical PDF matched empirical distribution
- Validated inverse CDF method effectiveness



2. Triangular PDF Case:

- Linear transformation observed
- Resulting distribution matched theoretical expectations
- Demonstrated versatility of inverse CDF method



DISCUSSION

The results demonstrate several important principles of probability theory and statistical transformations:

1. Distribution Transformation:

The results validate the theoretical transformation hypothesis.. The exponential distribution's transformation to uniform distribution highlights the power of probabilistic techniques in generating standardized random variables.

- CDF transformation consistently produces uniform distributions
- Larger sample sizes improve transformation accuracy, demonstrating the robustness of CDF-based probabilistic mapping
- Method works regardless of original distribution

2. Text Analysis:

The analysis unveils the probabilistic structure of linguistic data. CDF transformation provides a novel perspective on word frequency, transforming raw counts into normalized probabilistic representations. This approach reveals underlying text structure beyond simple frequency counting, demonstrating the potential of statistical methods in linguistic analysis.

3. Inverse CDF:

The inverse CDF method demonstrates remarkable versatility in distribution generation. By systematically mapping uniform random variables, we can generate complex probability distributions with high precision. This technique provides a powerful computational approach to probabilistic modeling, bridging theoretical distribution properties with practical random number generation.

- Reliable method for generating specific distributions
- Theoretical predictions match empirical results
- Particularly useful for complex distributions

Conclusion

This Assignment showcased the power of statistical programming in exploring probabilistic transformations through three interconnected domains: distribution mapping, linguistic analysis, and computational probability.

Key achievements include:

- Validating theoretical methods for probability distribution transformations
- Developing novel computational approaches to text frequency analysis
- Illustrating techniques for generating complex probability distributions

By bridging theoretical foundations with computational implementation, the study demonstrates the profound potential of statistical methods to uncover hidden patterns and systematic behaviors in diverse datasets.