# Entropy and Randomness

*Submitted to Dr. Anil Kumar Sao*
*DSL253-Assignment-1*
*Prepared by Amay Dixit - 12340220*

### Notebook Link:

https://colab.research.google.com/drive/1o6nFTad4ZSJoeHF6j1IKj0g4iJUQ00et?usp=sharing

### Github Link:

https://github.com/amaydixit11/Academics/tree/main/DSL253/assignment_1

### Docs Link:

https://docs.google.com/document/d/143dyjNots-8kMUZzZPAzHFNwkEHArXHWc22G-gjbrmY/edit?usp=sharing

## INTRODUCTION

This report presents a comprehensive analysis of letter and word frequencies in English text, alongside an investigation of random number distributions. The analysis focuses on three main areas:

1. Letter frequency analysis in English text
2. Entropy calculation for both letters and words
3. Statistical behavior of random number generators

## DATA

The analysis utilizes four text files (fileA, fileB, fileC, and fileD). The data processing involved:

1. Fetching the data from github repositories
2. Removing special characters, punctuation, and whitespace
3. Converting all text to lowercase for consistency
4. Treating words as distinct entities for word-level analysis

# METHODOLOGY

## Letter and Word Frequency Analysis

1. **Text Preprocessing**:
    a. Converted all text to lowercase
    b. Filtered out non-alphabetic characters
    c. Separated text into individual words for word-level analysis
2. **Probability Calculation**:
    a. Computed frequency counts for each letter/word
    b. Calculated probability as: $P(x) = count(x) / total\_count$
3. **Entropy Calculation**:
    a. Applied Shannon's entropy formula: $H = -\sum(pi * log2(pi))$
    b. Calculated separately for both letter and word distributions

## Random Number Analysis

1. **Uniform Distribution:**
    a. Generated numbers between 0 and 1
    b. Calculated mean and variance for increasing sample sizes up to n=10000
    c. Plotted statistical measures against sample size
2. **Gaussian Distribution:**
    a. Generated numbers with mean=4 and standard deviation=3
    b. Tracked mean and variance evolution with increasing sample size
    c. Visualized convergence through plots

# RESULTS

## Letter Frequency Analysis (FileA)

The top ten most frequent letters and their probabilities are:

1. 's': 4.24%
2. 'z': 4.13%
3. 'y': 4.13%
4. 'f': 4.08%

5. 'w': 4.03%
6. 't': 4.01%
7. 'u': 4.01%
8. 'x': 4.00%
9. 'j': 3.98%
10. 'k': 3.96%'

## Entropy Analysis

The entropy calculated from FileB was 4.1760 bits, indicating moderate uncertainty in letter distribution.

## Word Frequency Analysis

1. FileC Results:
   a. Most frequent word: "the" (7.83%)
   b. Total entropy: 9.0691 bits
2. FileD Results:
   a. Most frequent word: "the" (6.66%)
   b. Total entropy: 9.2599 bits

## Random Number Generation Analysis

### Uniform Distribution

As the sample size (n) increases:

1. The sample mean converges to 0.5
2. The sample variance converges to $1/12 \approx 0.0833$
3. Convergence becomes more stable after $n > 1000$

### Gaussian Distribution ($\mu=4$, $\sigma=3$)

As the sample size (n) increases:

1. The sample mean converges to 4
2. The sample variance converges to 9
3. Convergence rate is similar to uniform distribution
4. Demonstrated expected statistical properties of the Central Limit Theorem

# DISCUSSION

## Letter and Word Frequencies

The analysis reveals patterns consistent with known English language characteristics:

1. The distribution shows relatively even frequencies among top letters
2. The letter 'e' is predominantly the most frequent
3. Vowels generally appear more frequently than consonants
4. Common articles and prepositions dominate frequency rankings
5. The entropy value indicates significant variability in letter usage
6. Word-based entropy is higher than letter-based entropy, indicating greater uncertainty at the word level

## Random Number Behavior

The observed convergence of sample statistics demonstrates the Law of Large Numbers:

1. Larger sample sizes lead to more stable and predictable statistics
2. The speed of convergence differs between uniform and Gaussian distributions
3. Variance estimates require larger samples for stable convergence compared to means
4. Demonstrated the Law of Large Numbers, i.e., sample mean converges to true mean given a sample of independent and identically distributed values

# CONCLUSION

1. The letter frequency analysis provides valuable insights for optimizing printing machinery for English text
2. The entropy calculations quantify the inherent uncertainty in English language at both letter and word levels
3. The random number generation experiments demonstrate fundamental statistical principles of convergence and distribution properties
4. The results validate both linguistic patterns and statistical theories
5. As the token size increases, the entropy, meaning the randomness and uncertainty increases

## DATA SOURCES

1. fileA
   https://raw.githubusercontent.com/amaydixit11/Academics/refs/heads/main/DSL253/assignment_1/fileA.txt
2. fileB
   https://raw.githubusercontent.com/amaydixit11/Academics/refs/heads/main/DSL253/assignment_1/fileB.txt
3. fileC
   https://raw.githubusercontent.com/amaydixit11/Academics/refs/heads/main/DSL253/assignment_1/fileC.txt
4. fileD
   https://raw.githubusercontent.com/amaydixit11/Academics/refs/heads/main/DSL253/assignment_1/fileD.txt