**DS201**

**Statistical Programming**

**Assignment 2**

**Ashutosh Kumar Jha**

**12340390**

**DSAI**

**2nd Year**

**Semester 4**

**---------------------------------------------------------------------------**

**1. Question 1: Analyzing the Behavior of Random Variables Under Exponential and Uniform Distributions**

Introduction:

-----------------------

The behavior of random variables under different distributions provides fundamental insights into probability theory and statistical analysis. This study investigates the behavior of two processes:

1. **Exponential Distribution**: Models events that occur continuously and independently at a constant average rate, with a parameter lambda.
2. **Uniform Distribution**: Represents events equally likely to occur over a defined range.

The objective is to compute and analyze the random variable Y, derived as the Cumulative Distribution Function (CDF) of the respective distributions for a varying number of samples (n).

Data:

------------------------

**Exponential Distribution**:

* Parameter (lambda): 2
* Generated random numbers: n∈{100,1000,5000}

**Uniform Distribution**:

* Range: [0, 1]
* Generated random numbers: n∈{100,1000,5000}

Methodology:

—-----------------------

**Random Sample Generation**:

* Exponential random samples are generated using the inverse transform method with scale= 1/lambda
* Uniform random samples are generated within the range [0, 1].

**Computing Y**:

* For exponential samples, Y is derived using the CDF formula: y\_exp = 1 - (e^(-lambda\*x))
* For uniform samples, Y\_uniform = x, as the CDF for uniform distribution equals the identity function.

**Visualization**:

* Histograms are plotted for Y\_exp and Y\_uniform​ using different sample sizes (n).
* Kernel Density Estimation (KDE) overlays are added to observe the distribution pattern.

Results:

-----------------------

The histograms and KDE plots for n=100,1000,and 5000 reveal:

1. **Exponential Distribution**:
   * Y\_exp​ follows a cumulative trend with values concentrated closer to 1 as sample size increases.
   * The density distribution becomes smoother and more consistent with increasing n.
2. **Uniform Distribution**:
   * Y\_uniform remains uniformly distributed across [0, 1], with an almost linear density pattern.
   * Increasing n makes the histogram closer to a uniform density function.

Discussions:

----------------------

**Exponential Distribution Analysis**:

* The concentration of Y\_exp​ towards higher values is due to the nature of the exponential CDF.
* Larger n reduces random noise, ensuring better approximation of the theoretical distribution.

**Uniform Distribution Analysis**:

* The Y\_uniform​ values demonstrate the uniform nature of the underlying distribution.
* With increasing n, the distribution approaches perfect uniformity, as expected.

**Effect of Sample Size (n)**:

* Larger sample sizes improve the approximation to the theoretical probability distributions for both processes.
* The histograms confirm the law of large numbers, where empirical distributions converge to the theoretical ones.

Conclusion:

-----------------------

This analysis demonstrates how Y, derived from two distinct distributions, behaves under varying sample sizes. Key takeaways include:

1. **Exponential Distribution** results in a cumulative density trend, with values leaning toward 1 as the sample size increases.
2. **Uniform Distribution** retains its uniformity, with Y\_uniform​ evenly spread over the range.
3. Larger n smoothens the distributions, validating theoretical expectations.

This study underscores the importance of sample size in statistical modeling and highlights the contrasting behaviors of exponential and uniform distributions in modeling random variables.

**2. Question 2: Analyzing Word Frequencies and Cumulative Distribution Function (CDF) of Text Data**

1. Introduction:

----------------------

Text analysis plays a crucial role in understanding patterns, trends, and distributions within textual data. This report focuses on determining the most frequently used words in a given text file and analyzing their cumulative frequency distribution. The insights can be used for applications such as natural language processing (NLP), linguistic analysis, and text summarization.

2. Data

--------------------

The data used for this analysis is a textual file containing unstructured content. Key attributes of the data:

1. **Content**: Text in various forms (articles, sentences, or words).
2. **Format**: A plain-text file (.txt) with lines of sentences.
3. **Structure**: Words are delimited by spaces and punctuations are ignored during preprocessing.

3. Methodology

---------------------

**Preprocessing**:

* The file content is read and converted to lowercase to maintain uniformity.
* Words are extracted by splitting the text on spaces.
* Punctuation and special characters are removed.

**Frequency Analysis**:

* The frequency of each word is calculated using Counter from Python's collections module.
* The top 30 most frequent words are identified.

**Cumulative Distribution Function (CDF)**:

* The cumulative frequency of the top 30 words is calculated as a proportion of their total frequency.
* The CDF helps understand how the word frequencies accumulate over the top-ranked words.

**Visualization**:

* A bar plot is used to show the frequencies of the top 30 words.
* A line plot is used to depict the CDF of these words.

4. Results

--------------------

**Top 30 Words**:

* The top 30 words and their respective frequencies are extracted from the data.
* These words reflect the most common terms in the text, often including stop words (e.g., "the", "is", "and").

**CDF**:

* The CDF reveals how the cumulative frequency of words builds up across the top 30 words.
* The plot demonstrates that a small subset of words accounts for a significant portion of the overall frequency

5. Discussion

-------------------

**Key Observations**:

* The top-ranked words often include high-frequency stop words.
* The cumulative frequency plot shows that the first few words contribute to a significant portion of the total frequency, indicating a power-law distribution common in language.

**Interpretation of CDF**:

* The steep rise at the start of the CDF plot suggests a high concentration of word frequency among the top-ranked words.
* Words ranked lower in frequency contribute less significantly to the cumulative total.

**Limitations**:

* The analysis does not exclude stop words, which dominate the frequency counts.
* The results may vary significantly with different types of text (e.g., technical vs. casual writing).

6. Conclusion

------------------

This analysis effectively identifies the most frequent words in the text and highlights their cumulative contribution through the CDF. The results underscore the uneven distribution of word frequencies, where a small subset of words dominates the text. Future enhancements, such as stop-word removal and stemming, could refine the insights and broaden the scope of analysis.

**3. Question 3: Analyzing Transformed Random Variables Through Inverse CDF Sampling**

1. Introduction

--------------------

Inverse CDF (Cumulative Distribution Function) sampling is a widely used method to generate random variables from a specified probability distribution. This report investigates the transformation of uniformly distributed random variables into random variables following two distinct distributions:

1. **Exponential Distribution**: A continuous probability distribution with a rate parameter lambda = 2
2. **Custom Distribution**: Defined by a Probability Density Function (PDF) f(x)=x and its CDF inverse is given by F−1(u)= sqrt(2\*u)

The goal of this analysis is to visualize and compare the transformed random variables for these distributions.

2. Data

----------

**Source**: Random numbers are generated programmatically using NumPy.

**Uniform Samples**: n=1000 random numbers uniformly distributed between 0 and 1.

**Transformed Samples**:

* Exponential: Computed using F−1(u)=−ln⁡(1−u)/lambda with lambda = 2.
* Custom: Computed using F−1(u)=sqrt(2\*u)

3. Methodology

---------------------

**Generation of Uniform Samples**:

* Random numbers are drawn from a uniform distribution U(0,1).

**Transformation Using Inverse CDF**:

* For the exponential distribution: Y=−ln⁡(1−u)/lambda
* For the custom distribution: Y=sqrt(2\*u)

**Visualization**:

* Histograms of the transformed variables are plotted.
* Kernel Density Estimation (KDE) overlays are added to analyze the probability density of the transformed variables.

4. Results

---------------

**Exponential Distribution**:

* The histogram and KDE show the characteristic shape of an exponential distribution.
* Higher frequencies are observed near lower values of Y, indicating the steep decline in probability density as Y increases.

**Custom Distribution**:

* The histogram and KDE display a distinct shape influenced by the linear nature of the custom PDF f(x)=x
* The probability density increases as Y grows, consistent with the increasing slope of the custom PDF.

5. Discussion:

------------------

**Key Observations**:

* The exponential distribution's histogram matches its theoretical shape, with a higher density of smaller values.
* The custom distribution demonstrates increasing density, as expected from f(x)=x
* Both transformations validate the effectiveness of the inverse CDF sampling method.

**Implications**:

* The method accurately generates random variables for both standard (exponential) and custom-defined distributions.
* The distinct shapes of the histograms highlight the flexibility of the inverse CDF method for sampling from various distributions.

**Limitations**:

* The uniform samples are limited to n=1000. Larger sample sizes may provide smoother KDE estimates.
* Real-world applications may involve more complex PDFs, requiring additional validation.

6. Conclusion

------------------

The analysis demonstrates the successful application of inverse CDF sampling to transform uniformly distributed random variables into variables following exponential and custom distributions. The histograms and KDE overlays validate the theoretical properties of the distributions, highlighting the utility and versatility of this sampling method in statistical modeling and simulation.

Code: [12340390 Ashutosh Asg2.ipynb](https://colab.research.google.com/drive/1aZvWOr_7n9XXY8zhQzoPufF2nrQ_P4Yu?usp=sharing)