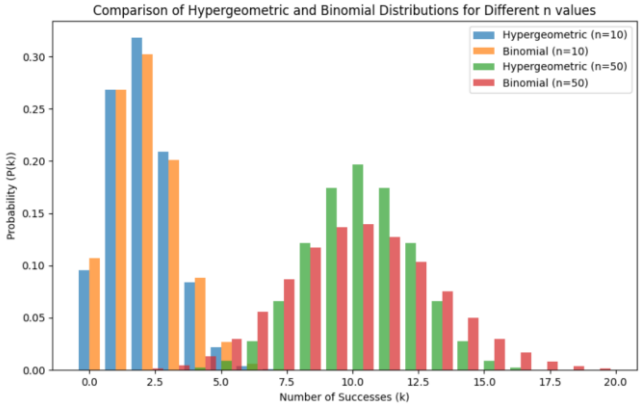
CA-1 Statistical Engineering

Rashin Rahnamoun
Faculty of Computer Science and Engineering
Shahid Beheshti University

***Abstract*— This study investigates the applicability of Benford's Law to distributions beyond the first digit, exploring its suitability for various digit positions and assessing the impact of different parameters. Utilizing a generalized formula for Benford's Law, incorporating a summation term, the research examines the distribution probabilities for digits in positions ranging from the first to the ninth. Additionally, the study introduces an approach to determine the expected value and variance of digits' probabilities in a given position. The methodology is extended to diverse probability distributions, including the Pareto, Weibull, and Log-Normal distributions, through the utilization of a genetic algorithm to optimize parameters. The investigation reveals that certain distributions, such as the Weibull with specific shape parameters, can effectively approximate Benford's Law. Moreover, the study introduces a J-divergence fitness function to enhance the genetic algorithm's optimization process. This research contributes to the understanding of generalized Benford's Law and its potential applications in assessing the conformity of various distributions to this mathematical phenomenon.**

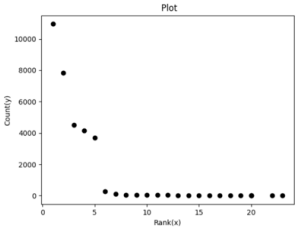
I. BINOMIAL AND HYPERGEOMETRIC

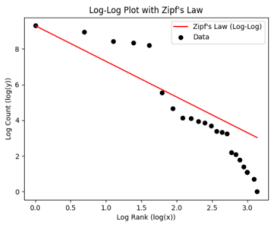
Increasing n in both the Binomial and Hypergeometric distributions tends to make the distributions more normal-like and brings the Hypergeometric distribution closer to the Binomial distribution due to the larger sample size.



II. DARWIN AND ZIPF'S LAW

In this investigation, we conducted a comparative analysis of word counts and their frequencies in Darwin's book. After obtaining the data, we visualized the relationship between the ranks of words and their respective counts in a log-log format. This log-log plot served as the basis for comparing the observed distribution with Zipf's Law, a statistical principle often used to describe the frequency distribution of words in natural language. The results of our analysis revealed intriguing patterns and deviations from Zipf's Law, providing insights into the underlying structures of word frequencies in Darwin's text. This investigation contributes to our understanding of the statistical properties of language and highlights the nuances in word distribution that may exist beyond the scope of traditional linguistic models.





The results of the word extraction process and their corresponding frequencies in the analyzed text are presented in the table below:

Rank	Final Results	
	Word	***Count***
1	the	10953
2	of	7832
3	and	4514
4	in	4161
5	to	3702
6	nature	260
7	within	106
8	size	62
9	distribution	60
10	body	52
11	extistence	48
12	head	40
13	eye	30
14	conclusion	28
15	second	26
16	words	9
17	hair	8
18	necessity	6
19	muscles	4
20	darwin	3
20	touch	3
20	revolution	3
23	publication	2
24	month	1

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

III. Benford's Law Analysis

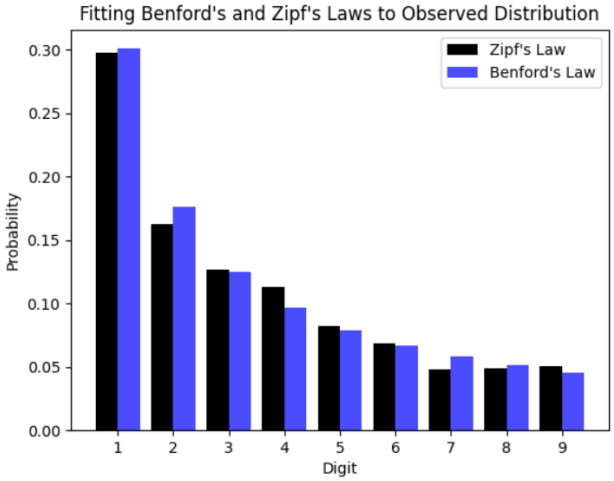
A. Analyzing Random Number Convergence

This study investigates the manifestation of Benford's Law in a controlled environment using a Python-based simulation. By generating random numbers in accordance with the probabilities dictated by Benford's distribution over 100 iterations, each comprising 1000 random numbers, the study aims to discern the pattern's consistency. The resulting histograms, computed for each iteration, are aggregated to present an average histogram, allowing an observation of the convergence toward the expected distribution. The study's outcomes provide insights into the stability and conformity of Benford's Law in the context of randomly generated numbers. The average histogram, showcasing the distribution of leading digits, demonstrates the law's applicability even in synthetic datasets, reinforcing its robustness and potential utility in various analytical and investigative domains.

$$P_{D_1}(d_1) = log_{10}^{\left(1+\frac{1}{d_1}\right)} \quad (1)$$

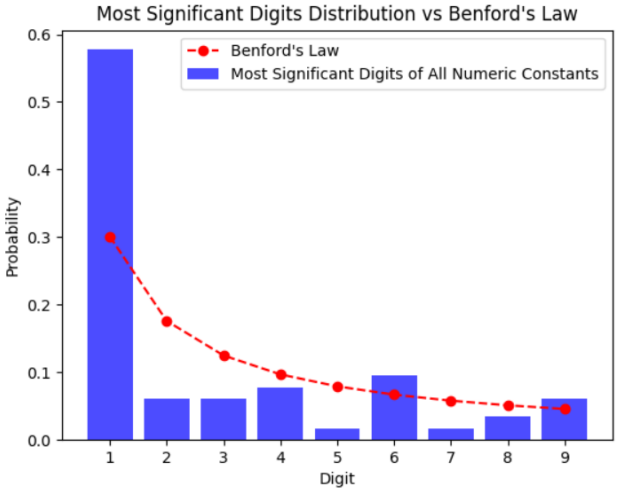
Benford's Law can be conceptualized through the application of the Riemann sum to Zipf's Law [1]. This connection becomes apparent when considering the scenario where the variable "n" is treated as a continuous parameter. In this context, the integration aspect of the Riemann sum offers insights into the emergence and characteristics of Benford's Law from the underlying principles of Zipf's Law. This relationship underscores the mathematical continuity between these two laws, shedding light on the intricacies of their interplay and providing a nuanced perspective on the statistical regularities observed in diverse datasets. The connection between these laws lies in their underlying principles of scale invariance and self-similarity, revealing a common thread in the patterns observed in both numerical and non-numerical datasets. This interplay offers valuable insights into the underlying order and regularities found in diverse datasets, transcending disciplinary boundaries.

$$\frac{1}{n} \approx \int_{n_{1=n}}^{n+1} \frac{dn'}{n'} = \ln\left(1+\frac{1}{n}\right) \quad (2)$$



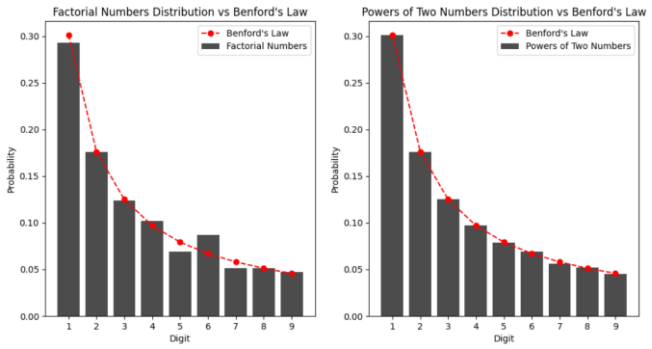
B. Fitting Benford's Law to Other Distributions

The provided simulation result employs statistical analysis to investigate the distribution of the most significant digits within a dataset comprising various numerical constants from the scipy.constants module. The primary objective is to compare the observed distribution of significant digits with Benford's Law, a mathematical phenomenon often observed in naturally occurring datasets. Benford's Law predicts that the occurrence of leading digits in many datasets follows a logarithmic distribution, with the digit '1' appearing more frequently than other digits. The code generates a larger dataset by repeating the selected constants, extracts the most significant digits from these constants, and then plots a histogram to visualize the distribution. The resulting graph is compared to the expected distribution according to Benford's Law. The printed digit counts further provide a quantitative analysis of the observed digit frequencies. This analysis serves as a practical illustration of how Benford's Law can be applied to assess the conformity of numerical datasets, providing insights into the inherent patterns and characteristics of the constants used.



The simulated results of the analysis reveal insightful observations about the distribution of most significant digits within datasets generated from factorial numbers and powers of two. By adhering to Benford's Law, which predicts a logarithmic distribution for leading digits, the simulations

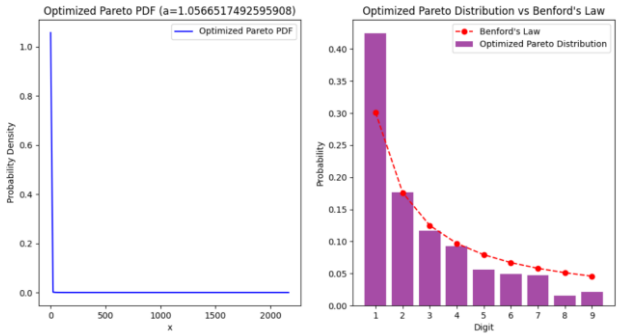
demonstrate that the digit '1' tends to occur more frequently than other digits. The investigation involves the generation of substantial datasets for factorial numbers and powers of two, followed by an analysis of the most significant digits within each dataset. The resulting bar plots visually depict the observed digit distributions alongside the anticipated distributions according to Benford's Law. The comparison offers valuable insights into how these mathematical sequences align or diverge from the expected patterns. This simulation-based approach provides a practical illustration of the application of Benford's Law, contributing to a nuanced understanding of the inherent characteristics of mathematical sequences.



The code presented amalgamates statistical principles, probability distributions, and optimization methodologies by employing a genetic algorithm. The overarching goal is to refine a parameter associated with a Pareto distribution, aiming to minimize the chi-squared statistic when contrasting the distribution of leading digits within a dataset with the characteristics outlined by Benford's Law.Initiating with the utilization of mathematical operations facilitated by NumPy, statistical functions from SciPy, and data visualization tools inherent in Matplotlib, the code establishes a framework. Within this framework, the Pareto distribution is characterized through its probability density function, and a mechanism is devised to isolate the most significant digits within the dataset. Benford's Law, a statistical phenomenon governing the distribution of leading digits, is also introduced into the analytical framework.A genetic algorithm, an optimization technique mimicking the principles of natural selection, is then employed. This algorithm iteratively refines the parameter of the Pareto distribution, progressively minimizing the chi-squared statistic. where O_i is the observed frequency for each category (leading digit), E_i is the expected frequency for each category, and the sum is taken over all categories.

$$\chi 2 = \sum \frac{(O_i - E_i)\text{^}2}{E_i} \qquad (3)$$

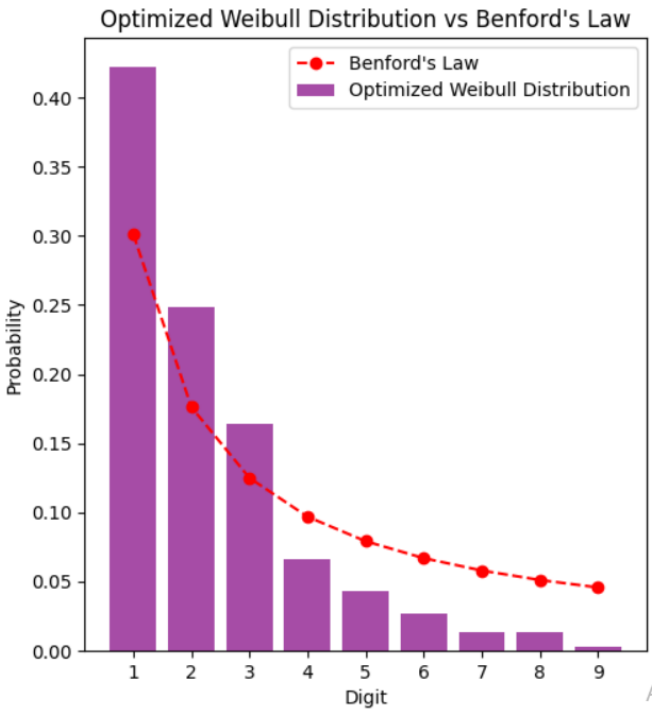
The objective is to enhance the fit of the distribution to Benford's Law, offering an optimized parameter that aligns more closely with the inherent characteristics of the dataset.The ensuing visualizations, including the optimized probability density function of the Pareto distribution, a histogram detailing the distribution of optimized leading digits, and a comparative analysis with Benford's Law, serve to elucidate the interplay between the dataset and the principles espoused by Benford's Law. This exemplifies the application of genetic algorithms in statistical analysis, accentuating their role in optimizing parameters and illuminating underlying patterns within datasets. The best *a* value is approximately 1.0566.



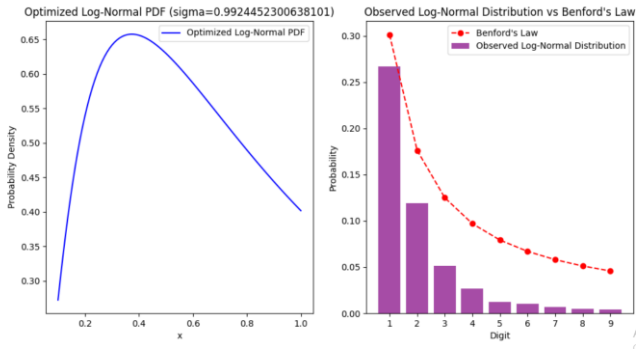
It can be demonstrated mathematically that when N is set to 10, it serves as an approximation for Benford's Law [2].

$$\frac{\bar{n}}{N+1} : \frac{N+1-\bar{n}}{N+1} \qquad (3)$$

It is possible to demonstrate that a Weibull distribution with a specific shape parameter (c) ≈ 1.06912 can serve as a meaningful approximation. Referring to [3], it is noted that a Weibull distribution with a shape parameter less than or equal to 0.5 is considered to be a favorable match for Benford's Law.



According [3] in Log-normal distribution if σ increases, it is better to fir Benford's law but μ has not effect.



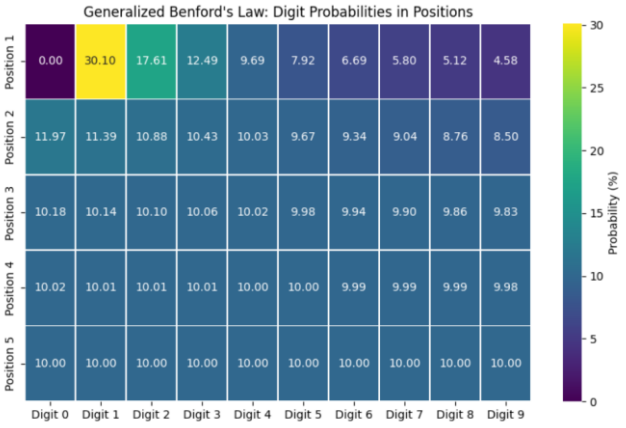
The Inverse Gamma distribution [3], which is intricately connected to the beta function, and ultimately linked to the Fisher distribution, demonstrates a notable compatibility with Benford's Law. Additionally, the Exponential distribution is frequently employed as a suitable approximation for conforming to the patterns observed in Benford's Law.

C. *Genreralized Bendord's Law*

This generalized formula extends Benford's Law to consider digits beyond the initial position, enabling a nuanced examination of digit distribution within numerical datasets. The resulting heatmap visually represents the calculated probabilities as percentages, offering insights into the distribution patterns of digits across various positions within multi-digit numbers.

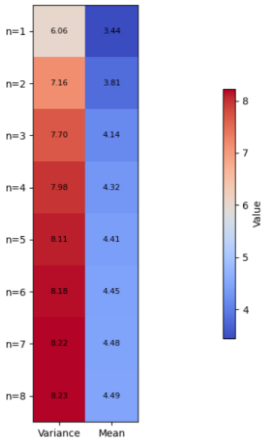
$$\sum_{k=10^{n-2}}^{10^{n-1}-1} \log_{10}\left(1+\frac{1}{10k+d}\right) \tag{4}$$

As the position of the digit within a numerical dataset increases (denoted as the n-th digit), the distribution of digits tends to converge towards a uniform distribution, with each of the ten digits having an approximately equal probability of 10%. This phenomenon becomes more evident with the increase in the digit position. For instance, when considering four-digit numbers, the distribution becomes notably uniform, with each digit expected to occur approximately 10% of the time. This is exemplified by the occurrence of "0" at 10.0176% and "9" at 9.9824% in the fourth digit, indicating a close approximation to the anticipated uniform distribution



D.*Analyzing the Mean and Variance*

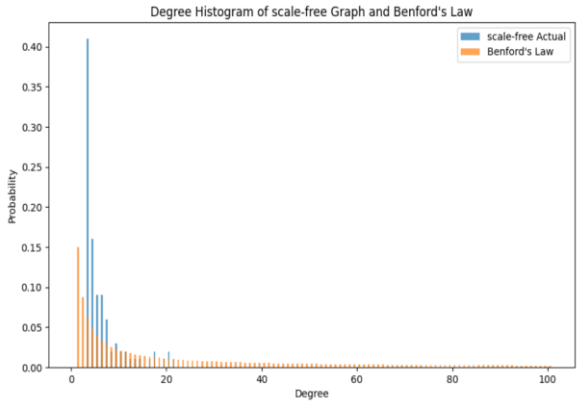
The provided algorithm explores the principles of Generalized Benford's Law, a statistical phenomenon applicable to diverse datasets. The core concept involves predicting the distribution of leading digits in numerical data. The law posits that smaller digits (1 through 9) appear more frequently than larger ones as the leading digit in naturally occurring datasets. In this context, the algorithm defines a probability function that calculates the likelihood of a specific digit appearing at a designated position within the dataset. Additionally, the algorithm computes the mean and variance for the digit distribution at various positions. The resulting insights are visualized in a table, offering a concise summary of the expected statistical characteristics. Further, a heatmap-style visualization provides a graphical representation of the mean and variance trends across different positions, enhancing the understanding of the generalized Benford's Law and its implications for digit distribution



IV. NETWORKS AND BENFORD'S LAW

Scale-free networks, random networks, and small-world networks are three distinct types of complex networks, each characterized by unique degree distributions. In scale-free networks, the degree distribution follows a power-law, featuring a small number of highly connected hubs alongside numerous nodes with lower degrees, reflecting a "rich-get-richer" phenomenon. Random networks, on the other hand, exhibit a more uniform degree distribution following a Poisson distribution, lacking the prominent hubs observed in scale-free networks. Small-world networks strike a balance, incorporating local clustering akin to regular networks while maintaining short paths between distant nodes, offering characteristics between those of random and regular networks. These degree distribution patterns are instrumental in understanding the structural properties and behaviors of these networks, influencing their resilience, connectivity, and overall dynamics.

Graph Name	Measurements		
	nodes	***edges***	***d***
Scale free	100	196	0.20
Random	100	136	0.42
Small world	100	100	0.50



The Jensen-Shannon Divergence (JSD) and Shannon Entropy are measures used to quantify the dissimilarity between probability distributions. In the context of complex networks, these measures can be applied to degree distributions, which represent the distribution of node degrees within a network.

The Jensen-Shannon Divergence (JSD) and Shannon Entropy are measures used to quantify the dissimilarity between probability distributions. In the context of complex networks, these measures can be applied to degree distributions, which represent the distribution of node degrees within a network.

Jensen-Shannon Divergence:

The Jensen-Shannon Divergence is a symmetrized and smoothed version of the Kullback-Leibler Divergence, which measures the difference between two probability distributions. For a given probability distribution P and Q, the JSD is calculated as the average of the Kullback-Leibler divergences between P and the average distribution M, where M is the midpoint distribution. In the context of complex networks, JSD can be used to compare the degree distributions of different types of networks.

Shannon Entropy:

Shannon Entropy is a measure of the uncertainty or disorder in a probability distribution. For a discrete probability distribution P, Shannon Entropy H(P) is calculated as the negative sum of P(i) times the logarithm base 2 of P(i) for each possible outcome i. In the context of complex networks, Shannon Entropy provides insights into the diversity of node degrees.

Graph Name	**Measurements**			
	nodes	***edges***	***J_S***	***J***
Scale free	100	196	0.16	3.33
Random	100	151	0.13	3.39
Small world	100	100	0.17	1.17

REFERENCES

[1] O. Kafri, “A novel approach to probability,” *Advances in Pure Mathematics*, vol. 06, no. 04, pp. 201–211, Jan. 2016, doi: 10.4236/apm.2016.64017.

[2] O. Kafri, “Zipf’s law, Benford’s law, and Pareto rule,” *Advances in Pure Mathematics*, vol. 13, no. 03, pp. 174–180, Jan. 2023, doi: 10.4236/apm.2023.133010.

[3] G. Fang and Q. Chen, “Several common probability distributions obey Benford’s law,” *Physica A: Statistical Mechanics and Its Applications*, vol. 540, p. 123129, Feb. 2020, doi: 10.1016/j.physa.2019.123129.