A Unified Approach to Fraudulent Detection

Anurag Dutta1 Manan Roy Choudhury2 Arnab Kumar De3

1, 2 Undergrad, Computer Science and Engineering, Government College of Engineering and Textile Technology, Serampore-712201, India.

3Assistant Professor, Government College of Engineering and Textile Technology, Serampore-712201, India. E-mail: arnab7339@yahoo.co.in

*Corresponding Author

Abstract
With the increase in demands and price of goods and services, fraudulency has caught a great height. But, it can’t be prohibited completely in the first stage. The detection of fraud has attracted continuous attention from academia, industry and regulatory agencies, and it is a challenging task for the researchers to develop a fraud detection framework. Starting from the late 1900s, ‘Benford’s law’ has served this purpose well. Abruptly, within a decade of its application lots and lots of fraudulency started getting seized. Later on, this law was used for detecting fairness of the elections, forensics, finances, etc. This article proposes a formula specifically derived from Zipf’s law that can detect fairness and fallacies in datasets involving forensics, finances, elections, and similar socio-economic issues. Unlike Benford’s law, our proposed formula is not dependent on any sort of observations, rather it is backboned by rigorous proof. Finally, we have done a comparison analysis between Benford’s law and our proposed formula graphically. All the data sets used by us have been rigorously studied, and many fitting tests have been applied to them.

Keywords: Fraud Detection, Benford’s law, Zipf’s law, Lagrange’s Interpolation, Mantissa Arc test.

1 INTRODUCTION
Fraud, the word itself means actions, methods or schemes without abiding by the laws that have been enacted upon children of mother earth. Starting from the first fraud in around 300 B.C, nothing much has changed to date. Time passed on, society grew up, colonization happened, the economy boomed and doomed but the inner impulse of mankind remained the same. To date, fraud is being enacted every day. Though the fraudsters are quite sly, still no one can get rid of the laws of nature. The difficulty though is with achieving a conclusive climatic assessment or result of an election based on direct and indirect observations that regimes can perform this, as Russia did in 2008, erect and intangible, formidable administrative barriers that presume and concludes any objective and viable oversight an impossibility of an imposter, or, as it occurred in Ukraine in 2004, both sides of a conflict can field their cadre of observers asserting or denying an indirect or direct fraud, whereas the rest of us are left in a state of dilemma to debate whom to believe. Furthermore, Benford’s law and Zipf’s law can be punched together to give us solid conclusive results on most kinds of frauds happening in our day-to-day life like income tax fraud, credit card fraud, GDP fraud, election fraud, etc.

Benford’s law (Frank Benford 1938), also known as the first digit law or law of anomalous digits, is a logarithmic probability distribution function for the first digits of a random, large and diverse dataset.

The first significant digit of a number is the first non-zero digit on its extreme left like 6 for 6897, 9 for 99 and 7 for 0.007895.

According to proposed Benford’s law, in a delineated dataset likelihood of prevalence of a definite digit as on its extreme left like 6 for 6897, 9 for 99 and 7 for 0.007895. Using Lagrange Interpolation, Mantissa Arc test.

<table>
<thead>
<tr>
<th>Digit</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.301029</td>
</tr>
<tr>
<td>2</td>
<td>0.176091</td>
</tr>
<tr>
<td>3</td>
<td>0.124938</td>
</tr>
<tr>
<td>4</td>
<td>0.096910</td>
</tr>
<tr>
<td>5</td>
<td>0.079181</td>
</tr>
<tr>
<td>6</td>
<td>0.066946</td>
</tr>
<tr>
<td>7</td>
<td>0.057991</td>
</tr>
<tr>
<td>8</td>
<td>0.051152</td>
</tr>
<tr>
<td>9</td>
<td>0.045757</td>
</tr>
</tbody>
</table>

Zipf’s law is a pragmatic law that has been developed using mathematical statistics and likelihood, which refers to the very fact that for several information studied within the physical and mathematical sciences, the rank-frequency distribution has an inverse relation between them. The Zipfian distribution falls under the class of discrete power-law probability distributions. It is related to the zeta distributions it deals with similar kinds of datasets but is not identical.

Zipf’s law was originally formulated for linguistics for the
study of the anomalous occurrence of words, stating that given some corpus or dataset of natural language occurrences, the frequency of occurrence of any word is inversely proportional to its rank in the accrodiun to the frequency table. Thus, the most frequent word will be occurring approximately twice as often as the second most frequently occurring word, three times as often as the third most frequently occurring word, and so on. For example, in the 1Brown Corpus of American English text, the word "the" is the most frequently occurring word, and by itself accounts for nearly 7.0 % of all occurrences of the words.

The law is coined after the American linguist George Kingsley Zipf (1902–1950). Though he wasn’t the one to give birth to that logic or better say the law. French stenographer Jean-Baptiste Estoup (1868–1950) had noticed this pattern too before Zipf stated it. Furthermore, it was also noticed in 1913 by German physicist Felix Auerbach (1856–1933).

[1] Mark J. Nigrini, was the person to use Benford’s practice within the field of fraud detection and forensic analysis. His analysis concerned various advanced theoretical works on Benford’s law and varied legal process that surrounds fraud convictions. Mark J. Nigrini is the author of forensics Analytics (Wiley Publications), which describes tests i.e., Benford’s law to discover frauds, errors, estimates, and biases in monetary and electoral information. He has been felicitated by the national media together with the “Wall Street Journal” and has printed many papers on Benford’s law.

[2] Arno Berger and Theodore P. Hill’s research paper stated about the randomness of Benford’s law, that this law should be applied to only some specific datasets to ensure a correct and conclusive result otherwise this law has more cons than pros.

[3] Hill, Theodore’s research paper tried to explain the diverse usability of Benford’s law in the fields like computer design, mathematical modelling and detection of fraud in accounting data.

[4] Innocent Mbona, Jan H. P. Eloff’s research was about finding a replacement for malicious social media bots. The study evidenced that the feature choice closely followed Benford’s law on a median human dataset, whereas an equivalent option violated Benford’s law on a malicious bot dataset. This study proves that options recognized by Benford’s law area are consistent and therefore the same with the info provided by PCA and follows the random forest technique on an equivalent dataset.

[5] Aleksandar Tošić, Jernej Vičič’s research focuses on the application of Benford’s law to scientific cooperative networks. The paper proposed a unique methodology following to assess the maturity of the research system. The paper digs in about the irregularities found between different and diverse research fields within Slovenia.

[6] Joanne Horton, Dhanya Krishna Kumar, and Anthony Wood’s research paper titled discussed the potential of Benford’s law for detecting the validity of the cardinal data used in different academic and research papers.

[7] Neumann, Peter G has shown a pavement about the usage of Zipf’s law in statistical metalinguistics laying a fine pellucid path in front of us to bid the usability of this law in various and diverse fields of statistics.

[8] Piantadosi, Steven’s research article states the diverse usability of Zipf’s law starting from fields like statistical linguistics to fraud detection. Piantadosi and Steven worked on the verification of Zipf’s law taking the dataset as a corpus of various words studying the occurrence of some particular frequent words and ranking them henceforth for applying Zipf’s law.

[9] Belevitch V’s research paper was amongst the first few papers written on Zipf’s law and its application. This paper discussed the various prospects of statistical methodology that can be implemented to Zipf’s law. The core of this paper is about the anomalous distribution of certain words that occur more frequently than the other words and hence studying its frequency distribution using different statistical methods.

[10] Jing Wei, Jianjun Zhang, Bofeng Cai, Ke Wang, Sen Liang, Yuhuan Geng’s research paper dealt with the datasets regarding CO2 emission from different cities in China and verifies it concerning the empirical data test. They proposed a revised model to explore the development of Chinese cities.

[11] Quiping A. Wang’s research paper proposed a new method to derive the Zipf-Pareto law using the law of least effort. An approach relating higher calculus and probability of non-additivity of efficient thermodynamic engine results in the derivation of Zipf-Pareto law.

[12] Michael E. Adel’s research paper proposed that ideas regarding business insights can be inferable from a patent topography by using Pareto-Zipf analysis. The main motto of this research paper is to calibrate scales and ascendancy of a patent landscape. Here, the law of nature concerning the fraudsters have been Statistical Laws with immense power – namely Benford’s law and Zipf’s law. These laws have been enacted upon many frauds and have successfully caught fraud in many cases. Some other aspects on which Benford’s law have been used quite often are checking Election Result, checking someone’s Income Tax details, Credit card transactions, etc. One interesting ² factual readout regarding the aforementioned topic is “The degree to which Benford’s law can be used as an indicator of electoral fraud has been debated by academicians, but the application of the rule to the leading digit of local vote tallies is problematic and apparent deviation from the law cannot be used alone to prove electoral fraud.”

Alongside, in this paper, work has been done on deriving a new formula, which in – fact has been erected using the famous Zipf’s law. The formulation has been successful in detecting
scrupulousness in social-economic datasets. One thing that we think we should brag about in our paper is that our formula is valid not only for discrete data sets with randomness, but we can also use it for data – sets dealing with linguistics as well and also a variety of more data sets, which perhaps Benford’s law would be incapable to compute (according to the definition of Benford’s law).

2 Benford’s Law

Benford’s law is an observation-based law whose, the discovery of which was dated back in the 1800s, when a Canadian – American astronomer [13]Simon Newcomb observed that in his log notebook, the earlier pages, especially those beginning with ‘1’ were in more worsened state than the latter ones. This observation created an impulse of thought in his mind which later took the shape of a formulation.

Newcomb proposed a law that stated, the probability of being the first digit of a number for a single number λ was equal to log (λ + 1) − log(λ).

Late on in the early 1900s, this phenomenon was again noted by a physicist, Frank Benford, who applied the then known formulation on a lot of datasets, and to his surprise, nearly all the datasets showed correlation with the formulation. The total number of observations used by Benford in his paper was nearly 20,000 which was a gigantic figure to deal with at that time.

Anyways, later Benford was accredited for this.

Briefly, Benford’s law states (or, rather, observes) that a variety of processes or measurements that gives rise to numbers (e.g., returns on investment, the population of huge cities, addresses of locations, sales of firms, heights of towers, and buildings) establish patterns in the digits that might otherwise seem counterintuitive wherein lower digits are more common than larger ones. The mathematical formulation of Benford’s law is

\[ \pi(\delta) = \log_{10}(\delta + 1) - \log_{10}(\delta) = \log_{10}(1 + \frac{1}{\delta}) \]

where,

\[ \pi(\delta) = \text{Probability of occurrence of the digit } \delta \text{ as the first digit } \forall 1 \leq \delta \leq 9 \]

The graph for Benford’s law (First digit) is

In the plot given above, P(δ) was confined on the y – axes while δ on the x – axes.

The aforementioned formulation works only for the occurrence of digit δ as the 1’st digit only, though another formulation has been generated that tells us about the probability of occurrence of digit δ as the ζ’th digit.

The formulation is as follows:

\[ \pi(\delta) = \text{Probability of Occurrence of the digit } \delta \text{ as the } \zeta \text{'th digit } \exists 0 \leq \delta \leq 9 \forall \zeta > 1 \]

On a general note, Benford’s law can be stated as

\[ \pi(\delta) = \begin{cases} \log_{10}(1 + \frac{1}{\delta}) & \exists \forall 1 \leq \delta \leq 9 \text{ and } \zeta = 1 \\ \sum_{k=10^k-1}^{10^{k+1}-2} \log_{10}(1 + \frac{1}{10k + \delta}) & \exists \forall 0 \leq \delta \leq 9 \text{ and } \zeta > 1 \end{cases} \]

In late 1900, an economist, Hal Varian suggested that, Benford’s law could be used to check for fraudulency in the socio-economic datasets and so was done. Benford’s law became very famous, and why shouldn’t it be? After all, it had such a vast domain of action. Benford’s law found it’s application in a lot of spheres, some are:

- [14]It’s used to detect fraudulency by checking irregularities in the election dataset.
- [15]It’s used to study price digits.
- It’s used to test genome data.
- It’s used to test for fallacies in scientific works.

Not even that, Benford’s law was even used as evidence in the US Criminal Courts.

In recent times, Benford’s law seems to be a plot fuel for many entertainment series. Benford’s law is used as a plot fuel to catch frauds in the 2016 movie – The Accountant. Apart from that, it’s used as a plot in the Netflix Series – Ozark and many more. As we know, everything has some merits as well as demerits, so does Benford’s law. Though Benford’s law has a wide range of applicability, still it’s not fool proof, the biggest example of it being the 1'total dataset of the 2020 US election, which didn’t follow Benford’s law and hence was reevaluated but no discrepancy was found.

3 Zipf’s Law

George Kingsley Zipf, an American linguist and philologist who studied statistical occurrences, the eponym of Zipf’s law suggested in 1935 that some words were repeated a few often while many or most are rarely used. This was the main impulse of Zipf’s law.

Briefly, [16]Zipf’s law states (or, rather, observes) for various
collection of data that is used to study the physical science and also, the social sciences, the rank-frequency distribution exists in an inverse relation. Zipf’s law was originally formulated in terms of perceptible linguistics, which stated, for any given corpus of natural language, some words occurred much frequently which the others occurred quite scarcely. Now, if any word appears in the frequency table, then its frequency of occurrence will be inversely proportional to the rank that the word has achieved in the accordion to the frequency table. Thus, the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc., but this law is not only restricted to linguistics but it can be applied to numeric as well. Let us have a set of data sorted in ascending order, in such a case, frequency times the rank will be a constant. Mathematically,

\[ f_\varsigma \times \varsigma = \psi \]

where,

- \( f_\varsigma \) = Frequency of data with rank \( \varsigma \).
- \( \varsigma \) = Rank.
- \( \psi \) = Constant.

We often use log – log graph for Zipf’s law where the log of the rank is the predefined scale for the x-axis and the log of its corresponding frequency is the scale for the y-axis. Log – log graph is used to show better details and clarity in the extremities of the graph. If we take log on both sides of the equation of Zipf’s law mentioned above we get the following equation -

\[ \log(f_\varsigma) + \log(\varsigma) = \log(\psi). \]

The above-mentioned equation is an equation of a straight line with intercept as \( \log(\psi) \). So, a dataset closely follows Zipf’s law if the log – log curve resembles a straight line with intercept \( \log(\psi) \).

The graph for Zipf’s law is given below :-

![Graph for Zipf's law](image)

In the plot given above, \( f_\varsigma \) was confined on the y – axes while \( \varsigma \) on the x – axes. Actually, in the plot given above, it’s a rectangular hyperbola of the form \( x \times y = c \) and the plot vary with varying values of ‘c’.

One interesting history regarding this law is that though this law was popularized and explained by Zipf, but he didn’t claim to have originated it.

The Zipf’s is similar to that of the Benford’s law conceptually though it’s different in distribution. In the later part of the paper, apart from Benford’s law, we have also used Zipf’s law to verify datasets. Alongside, Zipf’s law finds its importance in the:

- Information Theory
- Wide range of linguistics.
- To add on, it is also used as a part of extra-terrestrial intelligence.

4 Proposed Methodology

This section contains the collaborative work of all the contributors, we have proposed a new formulation capable enough to detect fraudulences and fallacies within the data sets. The added advantage and credibility of our formula is that its application is not limited to only determining the probability of the first digits, it is applicable in different fields such as Linguistics, Information Theory, Extra-terrestrial activities, Forensics, Finances, etc. This is because our is applicable in all those fields where Zipf’s law is applicable. The only problem is that our interpolated formula is only valid up to rank 9. The formula can definitely be tailor made for different fields such that the becomes valid for any rank given as interpolation can be done on rank starting from 1 to \( n \) where ‘n’ is the required rank in that field.

Let we are given a set of data, and we have ranked the occurrences of 1’st digit in ascending order accordingly.
These can be represented in a tabular manner as such:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Frequency of 1’s Digit</th>
<th>Zipf’s Column</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F₁</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>F₂</td>
<td>2 F₂</td>
</tr>
<tr>
<td>3</td>
<td>F₃</td>
<td>3 F₃</td>
</tr>
<tr>
<td>4</td>
<td>F₄</td>
<td>4 F₄</td>
</tr>
<tr>
<td>5</td>
<td>F₅</td>
<td>5 F₅</td>
</tr>
<tr>
<td>6</td>
<td>F₆</td>
<td>6 F₆</td>
</tr>
<tr>
<td>7</td>
<td>F₇</td>
<td>7 F₇</td>
</tr>
<tr>
<td>8</td>
<td>F₈</td>
<td>8 F₈</td>
</tr>
<tr>
<td>9</td>
<td>F₉</td>
<td>9 F₉</td>
</tr>
</tbody>
</table>

Here, F₁ > F₂ > F₃ ... > F₉ (According to Zipf’s law)

According to Zipf’s law

F₁ = 2 F₂ = 3 F₃ = ... = 9 F₉ (Considering Ideal Datasets)

F₂ = F₁/2

F₃ = F₁/3

...

F₉ = F₁/9

Now,

\[ \sum F_i = F_1 + F_2 + F_3 + \ldots + F_9 \]

\[ = F_1 (1 + 1/2 + 1/3 + \ldots + 1/9) \]

\[ \sum_{i=1}^{n} F_i = F_1 (1 + 1/2 + 1/3 + \ldots + 1/n) \]

Generally, \( n \leq 9 \) (as the first digit of a number range between 1 to 9)

\[ P(n = 1) = \frac{F_1}{\sum F_i} = \frac{1}{1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{9}} = \text{Probability of occurrence of digit with frequency } F_1 \text{ as the first digit.} \]

\[ P(n = 2) = \frac{F_2}{\sum F_i} = \frac{1}{1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{9}} = \text{Probability of occurrence of digit with frequency } F_2 \text{ as the first digit.} \]

\[ P(n = 3) = \frac{F_3}{\sum F_i} = \frac{1}{1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{9}} = \text{Probability of occurrence of digit with frequency } F_3 \text{ as the first digit.} \]

... (i)

Now, after further simplification we get:

\[ P(n = 1) = 0.353485762 \]

\[ P(n = 2) = 0.176742881 \]

\[ P(n = 3) = 0.117828587 \]

\[ P(n = 4) = 0.083714411 \]

\[ P(n = 5) = 0.070697152 \]

\[ P(n = 6) = 0.058914294 \]

\[ P(n = 7) = 0.050497966 \]

\[ P(n = 8) = 0.044185720 \]

\[ P(n = 9) = 0.039376196 \]

Now, by Lagrange’s Interpolation

\[ \rho_\tau(y) = \sum_{\omega=0}^{\tau} \beta_\omega \left( \prod_{\theta=0,\theta \neq \omega}^{\tau} \frac{y - \alpha_\theta}{\alpha_\omega - \alpha_\theta} \right) \] ... (ii)

with the Lagrange Polynomial, \( \rho_\tau(y) \) satisfying the points \((\alpha_0, \beta_0), (\alpha_1, \beta_1), \ldots, (\alpha_{\tau-1}, \beta_{\tau-1})\).

where,

\[ \rho_\tau(y) = \text{Lagrange’s Polynomial in } \gamma \text{ of degree } \tau - 1. \]

Here, \( \alpha_0 = 1, \alpha_1 = 2, \ldots, \alpha_9 = 9 \) and \( \beta_\mu = P(n = \mu) \forall \mu \in \{1,2,3,\ldots,9\} : \)

Substituting the values of \( P(n = 1), P(n = 2), P(n = 3), \ldots, P(n = 9) \) i.e., the values of \( \beta_\mu \forall \mu \in \{1,2,3,\ldots,9\} \) in Equation (ii),
we get the Lagrange polynomial:

\[ \rho_n(y) = \left(\left(\left(\left(9.741 \times 10^{-7}y - 4\right) - 5.845 \times 10^{-4}y + 0.2 \times 10^{-2}\right)(y - 2) - 0.1 \times 10^{-3}\right)(y - 8) + 0.3 \times 10^{-2}\right)(y - 3) - 0.2 \times 10^{-2}\right)(y - 5) + 0.8 \times 10^{-2}\right)(y - 1) - 0.4 \times 10^{-2}\right)(y - 9) + 0.4 \times 10^{-2} \]

\[ = \rho_n(y) = 9.741 \times 10^{-7}y^8 - 0.4 \times 10^{-4}y^7 + 0.8 \times 10^{-3}y^6 - 0.009y^5 + 0.062y^4 - 0.262y^3 + 0.704y^2 - 1.142y + 1 \]

\[ P(n = \mu) = 9.741 \times 10^{-7}\mu^8 - 0.4 \times 10^{-4}\mu^7 + 0.8 \times 10^{-3}\mu^6 - 0.009\mu^5 + 0.062\mu^4 - 0.262\mu^3 + 0.704\mu^2 - 1.142\mu + 1 \]

\[ \forall \mu \in \mathbb{Z}^+ \& \mu \in [1, 9] \quad \ldots \ (iii) \]

When plotted graphically the interpolated polynomial i.e., \( \rho_n(y) \), we get the following plot:

**Fig. 3.** Graph of our proposed formula where x-axis denotes the first digits and the y-axis denotes the probability of their occurrences.

The graph when compared that of the plot of Benford’s law, gives a slight deviation for the digit with highest frequency or with rank = 1.

**Fig. 4.** Comparison graph between Benford’s law (blue) and our proposed formula (red).

In the case of our interpolated formula the percentage of occurrence of the first digit is around 35% but in case of Benford’s law it’s around 30%. Apart from the first digit as 1 there is not much deviation of our interpolated formula from Benford’s law.

5 Application

In this section, we have tried to apply the aforementioned laws, Benford’s law, the Zipf’s law, and our proposed formula in some well fledged dataset. The datasets on which we have worked are:

- 6COVID – 19 Worldwide Cases
- 7Credit Card Transaction data

All these datasets are available online on Kaggle.

Before applying Benford’s law, we will perform some tests to confirm whether the dataset is fit for applying Benford’s law or not.

We will perform three tests namely:

- Mean Absolute Deviation (MAD)
- Chi-square test
- Mantissa Test

We have used R-Studio to perform these tests.

- Mean Absolute Deviation: The Mean Absolute Deviation (MAD) of a dataset is the mean of the absolute deviations from a central point i.e., ‘Mean’.

**MAD of a dataset \{ z_1, z_2, z_3, ..., z_n \} = \frac{1}{n} \sum_{i=1}^{n} |z_i - A(z)|**

**A(z) = Mean value of the dataset.**

**N = Total number of data values.**

**z_i = Data values in the dataset.**

Conformity range for Mean Absolute Deviation test for a dataset to be fit for applying Benford’s law.

**Table 3.** 8Conformity Range for the 1’st Digits.

<table>
<thead>
<tr>
<th>Conformity Range</th>
<th>First Digits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Close Conformity</td>
<td>0.000 - 0.006</td>
</tr>
<tr>
<td>Acceptable Conformity</td>
<td>0.006 - 0.012</td>
</tr>
<tr>
<td>Marginal Conformity</td>
<td>0.012 - 0.015</td>
</tr>
<tr>
<td>Non-conformity</td>
<td>Above 0.015</td>
</tr>
</tbody>
</table>

- Chi-Square test: Pearson’s Chi-Squared test is a fitting test that is used to determine whether there is a permissible gap between the theoretical statistics and the practical

---

1 In numerical analysis, for a given set of points \((x_i, y_j)\) with no two \(x_i\) values equal, the Lagrange polynomial is the polynomial of lowest degree that assumes at each value \(x_i\) the corresponding value \(y_j\).

2 Data for this dataset was collected from a lot of sources like Johns Hopkins University, MoBS Labs, World Health Organization, etc.

3 Abstract dataset for detecting Credit Card Fraud.

4 This Conformity Range have been taken from a verified paper by Mark J. Nigrini.

© Research India Publications. http://www.ripublication.com
statistics.

Given that null hypothesis are correct as n tends to infinity, the $\chi^2$ distribution is:

$$Z^2 = \sum_{i=1}^{k} \left( \frac{(z_i - m_i)^2}{m_i} \right) = \sum_{i=1}^{k} \left( \frac{z_i^2}{m_i} - n \right)$$

$m_i = np_i \forall i \in \mathbb{N}$, where $p_i$ are the probabilities given by null hypothesis and $\sum_{i=1}^{k} p_i = 1$

- Mantissa Arc test: This test helps us to determine the centre of mass, for a given of set of mantissas that have been well distributed on a unit circle. If the mantissa of numbers is uniformly distributed around the unit circle, then the centre of mass i.e., the mean vector is $(0,0)$.

The ‘Difference’ graph i.e., scatter plot which denotes the difference between the Synthetic probability and the probability by our proposed formula closely conforms the validity of the dataset if the points on the Scatter plot are approximately equal to zero. Based on this Scatter plot we can decide whether the dataset deviates from our proposed formula or not i.e., judge its fairness.

Note: Mean Absolute Deviation test is the most accurate test for testing whether the data is fit for Benford’s law or not. If the result of MAD test comes as “moderate conformity” or “non-conformity” then Chi-Square test and Mantissa Arc test is used for further conformity.

5.1 COVID – 19 Worldwide Cases

From World Health Organization - On 31 December 2019, WHO was alerted and noticed about several cases of pneumonia in Wuhan City, Hubei Province of China. The virus traced did not match any other existing known virus. This issue was the first to raise concern amongst all because when a virus is new, we do not know how it can affect people.

We have used the dataset of the Confirmed COVID Cases all over the world in this study. The dataset we have used can be assessed from 9(here).

We considered the first digits of the dataset i.e., number of confirmed COVID-19 cases worldwide per day and we obtained the following result:

<table>
<thead>
<tr>
<th>Digit</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>92499</td>
</tr>
<tr>
<td>2</td>
<td>52889</td>
</tr>
<tr>
<td>3</td>
<td>38844</td>
</tr>
<tr>
<td>4</td>
<td>29036</td>
</tr>
<tr>
<td>5</td>
<td>24122</td>
</tr>
<tr>
<td>6</td>
<td>19027</td>
</tr>
<tr>
<td>7</td>
<td>16517</td>
</tr>
<tr>
<td>8</td>
<td>14370</td>
</tr>
<tr>
<td>9</td>
<td>14383</td>
</tr>
</tbody>
</table>

Adding up all the frequencies, we get $\sum f_i = 301687$.

After that, we calculated the synthetic probability (or better say the practical probability) by using the formula, $P = \frac{f_i}{\sum f_i}$ and compared that to the theoretical probabilities obtained by our proposed formula.

Now, we will apply our proposed formula to the above-mentioned dataset and will try to draw conclusions about the fairness of the COVID-19 Cases Worldwide dataset.

### Table 4. Digit – wise frequency for 1’st Digit place for COVID – 19 datasets.

<table>
<thead>
<tr>
<th>Digit</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>92499</td>
</tr>
<tr>
<td>2</td>
<td>52889</td>
</tr>
<tr>
<td>3</td>
<td>38844</td>
</tr>
<tr>
<td>4</td>
<td>29036</td>
</tr>
<tr>
<td>5</td>
<td>24122</td>
</tr>
<tr>
<td>6</td>
<td>19027</td>
</tr>
<tr>
<td>7</td>
<td>16517</td>
</tr>
<tr>
<td>8</td>
<td>14370</td>
</tr>
<tr>
<td>9</td>
<td>14383</td>
</tr>
</tbody>
</table>

### Table 5. Digit-wise comparison between synthetic probability and probability by our proposed formula.

<table>
<thead>
<tr>
<th>Practical Probability</th>
<th>Theoretical Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.306605853</td>
<td>0.353485762</td>
</tr>
<tr>
<td>0.175310835</td>
<td>0.176742881</td>
</tr>
<tr>
<td>0.128755962</td>
<td>0.117828587</td>
</tr>
<tr>
<td>0.096245446</td>
<td>0.088371441</td>
</tr>
<tr>
<td>0.079957042</td>
<td>0.070697152</td>
</tr>
<tr>
<td>0.063068677</td>
<td>0.058914294</td>
</tr>
<tr>
<td>0.054748796</td>
<td>0.050497966</td>
</tr>
<tr>
<td>0.047632149</td>
<td>0.044185720</td>
</tr>
<tr>
<td>0.047675240</td>
<td>0.039276196</td>
</tr>
</tbody>
</table>

9 Due to huge dataset size, the data can’t be shown in proper tabular format, though it can be download in tabular format without any hustle.
Fig. 5. Comparison bar graph between synthetic probability and the probability by proposed formulation.

The aforementioned graph compares the probabilities according to the practical observation and by our proposed formulation.

Further, we have also plotted the difference between the practical probabilities and the probabilities according to our proposed formulation, and we got discrete points, which on merging gave nearly a straight-line curve.

Fig. 6. Scatter plot for difference between the Synthetic probability and the probability by our proposed formula.

Thus, the points in the above scatter plot are very close to zero confirming that the above dataset is fair.

Now, we will apply Zipf’s law in the dataset,

In the Zipf’s Column I, we multiplied Frequency with the rank.
Table 6. Digit, frequency, rank and Zipf’s Column for the COVID – 19 datasets.

<table>
<thead>
<tr>
<th>Digit</th>
<th>Frequency ((V_m))</th>
<th>Rank ((m))</th>
<th>Zipf’s Column I</th>
<th>Zipf’s Column II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>92499</td>
<td>1</td>
<td>92499</td>
<td>0.09</td>
</tr>
<tr>
<td>2</td>
<td>52889</td>
<td>2</td>
<td>105778</td>
<td>0.1</td>
</tr>
<tr>
<td>3</td>
<td>38844</td>
<td>3</td>
<td>116532</td>
<td>0.11</td>
</tr>
<tr>
<td>4</td>
<td>29036</td>
<td>4</td>
<td>116144</td>
<td>0.11</td>
</tr>
<tr>
<td>5</td>
<td>24122</td>
<td>5</td>
<td>120610</td>
<td>0.12</td>
</tr>
<tr>
<td>6</td>
<td>19027</td>
<td>6</td>
<td>114162</td>
<td>0.11</td>
</tr>
<tr>
<td>7</td>
<td>16517</td>
<td>7</td>
<td>115619</td>
<td>0.11</td>
</tr>
<tr>
<td>8</td>
<td>14370</td>
<td>9</td>
<td>129330</td>
<td>0.13</td>
</tr>
<tr>
<td>9</td>
<td>14383</td>
<td>8</td>
<td>115064</td>
<td>0.11</td>
</tr>
</tbody>
</table>

In the 2'nd Zipf’s Column, we rounded of the result of \(\frac{ZC_1}{\sum ZC_1}\) up to 2 decimal places.

where, \(ZC_1 = \text{Zipf’s First Column}\)

and we can say that the \(^10\)values of the data of Zipf’s Column II were nearly the same.

The log – log plot of the Zipf’s Column I has been shown below.

Fig. 7. log – log graph of Zipf’s law for COVID-19 Worldwide Cases.

The above log – log graph closely resembles a straight-line curve.

By all these studies, we can say that

\[P = \frac{f_i}{\sum f_i}\]

"The Dataset of the Worldwide COVID 19 cases seem not to be fraudulent."

Now we will perform the fitting tests on the COVID-19 Worldwide cases dataset to see whether the dataset is fit for Benford’s law to be applied on it. We have already mentioned the details of the fitting tests above.

Results of the Mean Absolute Deviation, Chi-square test and Mantissa Arc tests for COVID-19 Worldwide Cases dataset.

1. Mantissa:
   - Mean = 0.493
   - Var = 0.084
   - Ex. Kurtosis = -1.182
   - Skewness= 0.013

2. Mean Absolute Deviation (MAD): 0.0008645937
   - MAD Conformity = Close Conformity
   - Distortion Factor = -1.389082

3. Pearson Chi-Square test:
   - \(\chi^2 = 3480.4\)
   - df = 89
   - p-value < 2.2e-16

4. Mantissa Arc Test:
   - \(L_2 = 3.6422e-05\)
   - df = 2
   - p-value = 1.69e-05

Hence, this dataset is fit for Benford’s law to be applied.

After that, we calculated the synthetic probability (or better say the practical probability) by using the formula, \(P = \frac{f_i}{\sum f_i}\) and compared that to the theoretical probabilities obtained by Benford’s law for the case of \(\zeta = 1\).

Table 7. Tabular comparison between Practical and Theoretical Probabilities (Benford’s law) for 1’st Digit place for COVID – 19 datasets.

<table>
<thead>
<tr>
<th>Practical Probability</th>
<th>Theoretical Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.306605853</td>
<td>0.301029996</td>
</tr>
<tr>
<td>0.175310835</td>
<td>0.176091259</td>
</tr>
<tr>
<td>0.128755962</td>
<td>0.124938737</td>
</tr>
<tr>
<td>0.096245446</td>
<td>0.096910013</td>
</tr>
<tr>
<td>0.079957042</td>
<td>0.079181246</td>
</tr>
<tr>
<td>0.063068677</td>
<td>0.066946790</td>
</tr>
<tr>
<td>0.054748796</td>
<td>0.057991947</td>
</tr>
<tr>
<td>0.047632149</td>
<td>0.051152522</td>
</tr>
<tr>
<td>0.047675240</td>
<td>0.045757491</td>
</tr>
</tbody>
</table>

\(^{10}\)In correlation to Zipf’s law the Zipf’s Column I should have been constant but due to randomness of values, the Column 1 was unable to maintain the consistency that was expected by Zipf’s law, which was further shrunk to visualize the consistency well in Zipf’s Column II.
The practical probability was nearly similar to that of the Theoretical Probability.

![Graph for Digit distribution, Digit distribution (second order test), summation distribution by digits, Chi-Square difference and Absolute summation difference.](image)

**Fig. 8.** Graph for Digit distribution, Digit distribution (second order test), summation distribution by digits, Chi-Square difference and Absolute summation difference.

To create a more mesmerizing visualization, we have tried to plot the Probabilities (Synthetic) or better say, practical Probabilities and the Probabilities according to our proposed formulation along with those according to the pre-existing Benford’s law in the form of a Comparison Plot which is shown below.

![Comparison Chart plot between Synthetic probability, probability by Benford’s law and probability by our proposed formula.](image)

**Fig. 9.** Comparison Chart plot between Synthetic probability, probability by Benford’s law and probability by our proposed formula.

A more detailed and distinguishable 3 – D plot can be obtained from (here).

All these justifications, indicates that our formula can also detect fairness in datasets and this dataset seems to be fair considering all the test results and the above graphs.

Moreover, this dataset i.e., Worldwide COVID 19 cases seem to be fair without any underlying fallacies or fraud.

### 5.2 Credit Card Transaction Data

Card transaction data is data claimed on financial background, generally collected through the transfer of funds between a card holder’s account and a business’s account. It consists of the use of either a debit card or a credit card to generate data on the transfer for the purchase of goods or services. We can’t mention the source this dataset due to legal rules and guidelines.

In this subsection, we have used an abstract credit card transaction data. The dataset we have used can be assessed from (here).

We considered the first digits of the dataset i.e., Monthly Transaction amount and we obtained the following result.
Table 8. Digit-wise frequency for 1’st Digit place for Credit Card Transaction Dataset.

<table>
<thead>
<tr>
<th>Digit</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>341</td>
</tr>
<tr>
<td>2</td>
<td>334</td>
</tr>
<tr>
<td>3</td>
<td>371</td>
</tr>
<tr>
<td>4</td>
<td>332</td>
</tr>
<tr>
<td>5</td>
<td>333</td>
</tr>
<tr>
<td>6</td>
<td>324</td>
</tr>
<tr>
<td>7</td>
<td>356</td>
</tr>
<tr>
<td>8</td>
<td>338</td>
</tr>
<tr>
<td>9</td>
<td>346</td>
</tr>
</tbody>
</table>

Adding up all the frequencies, we get \( \sum f_i = 3075 \)

After that, we calculated the synthetic probability (or better say the practical probability) by using the formula, \( P = \frac{f_i}{\sum f_i} \) and compared that to the theoretical probabilities obtained by our proposed formula.

Now, we will apply our proposed formula to the above-mentioned dataset and will try to draw conclusions about the fairness of the given dataset.

Table 9. Digit-wise comparison between synthetic probability and probability by our proposed formula.

<table>
<thead>
<tr>
<th>Practical Probability</th>
<th>Theoretical Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.110894309</td>
<td>0.353485762</td>
</tr>
<tr>
<td>0.108617886</td>
<td>0.176742881</td>
</tr>
<tr>
<td>0.120650407</td>
<td>0.117828587</td>
</tr>
<tr>
<td>0.10796748</td>
<td>0.088371441</td>
</tr>
<tr>
<td>0.108292683</td>
<td>0.070697152</td>
</tr>
<tr>
<td>0.105365854</td>
<td>0.058914294</td>
</tr>
<tr>
<td>0.115772358</td>
<td>0.050497966</td>
</tr>
<tr>
<td>0.109918699</td>
<td>0.044185720</td>
</tr>
<tr>
<td>0.112520325</td>
<td>0.039276196</td>
</tr>
</tbody>
</table>

The aforementioned graph compares the probabilities according to the practical observation and by our proposed formulation.

Fig. 10. Comparison bar graph between synthetic probability and the probability by proposed formulation.

Further, we have also plotted the difference between the practical probabilities and the probabilities according to our proposed formulation, and we got discrete points, which on merging didn’t give a straight-line curve.
Thus, the points in the above scatter plot are not close to zero confirming that the above dataset is not fair. So, it’s better to cross-check this dataset again to verify it’s fairness. Also, this dataset doesn’t closely resemble the graphs of Benford’s law and our proposed formula.

Now, we will apply Zipf’s law in the dataset.

In the Zipf’s Column I, we multiplied Frequency with the rank.

In the 2’nd Zipf’s Column, we rounded of the result of \( \frac{Z_{C1i}}{\sum Z_{C1i}} \) up to 2 decimal places.

where, \( Z_{C1} \) = Zipf’s First Column

and we can say that the values of the data of Zipf’s Column II were strictly different.

The log – log plot of the Zipf’s Column I has been shown below.

### Table 10. Digit, frequency, rank and Zipf’s Column for the Credit Card dataset.

<table>
<thead>
<tr>
<th>Digit</th>
<th>Frequency (( V_m ))</th>
<th>Rank (( m ))</th>
<th>Zipf’s Column I</th>
<th>Zipf’s Column II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>341</td>
<td>4</td>
<td>1364</td>
<td>0.09</td>
</tr>
<tr>
<td>2</td>
<td>334</td>
<td>6</td>
<td>2004</td>
<td>0.13</td>
</tr>
<tr>
<td>3</td>
<td>371</td>
<td>1</td>
<td>371</td>
<td>0.02</td>
</tr>
<tr>
<td>4</td>
<td>332</td>
<td>8</td>
<td>2656</td>
<td>0.18</td>
</tr>
<tr>
<td>5</td>
<td>333</td>
<td>7</td>
<td>2331</td>
<td>0.15</td>
</tr>
<tr>
<td>6</td>
<td>324</td>
<td>9</td>
<td>2916</td>
<td>0.19</td>
</tr>
<tr>
<td>7</td>
<td>356</td>
<td>2</td>
<td>712</td>
<td>0.05</td>
</tr>
<tr>
<td>8</td>
<td>338</td>
<td>5</td>
<td>1690</td>
<td>0.11</td>
</tr>
<tr>
<td>9</td>
<td>346</td>
<td>3</td>
<td>1038</td>
<td>0.07</td>
</tr>
</tbody>
</table>
The above log–log graph doesn’t resemble a straight line curve; thus, no conclusion can be derived from this log–log graph of Zipf’s law.

By all these studies, we can say that

“The Dataset of the Credit Card Transaction seems to be fraudulent or there is some kind of underlying error or miscalculation in the dataset”

Now we will perform the fitting tests on the Credit Card Transaction dataset to see whether the dataset is fit for Benford’s law to be applied on it. We have already mentioned the details of the fitting tests above.

Results of the Mean Absolute Deviation, Chi-square test and Mantissa Arc tests for Credit Card Monthly Transaction dataset.

1. Mantissa:
   - Mean = 0.676
   - Var = 0.066
   - Ex. Kurtosis = -0.300
   - Skewness = -0.781

2. Mean Absolute Deviation (MAD): 0.0008645937
   - MAD Conformity = Close Conformity
   - Distortion Factor = 40.45302

3. Pearson Chi-Square test:
   - \( \chi^2 = 1414.3 \)
   - df = 89
   - p-value < 2.2e-16

4. Mantissa Arc Test:
   - L2 = 0.11627
   - df = 2
   - p-value = 2.2e-16

Hence, this dataset is fit for Benford’s law to be applied.
To create a more mesmerizing visualization, we have tried to plot the Probabilities (Synthetic) or better say, practical Probabilities and the Probabilities according to our proposed formulation along with those according to the pre-existing Benford’s law in the form of a Comparison Plot which is shown below.

![Comparison Plot](image)

**Fig. 14.** Comparison Chart plot between Synthetic probability, probability by Benford’s law and probability by our proposed formula.

A more detailed and distinguishable 3-D plot can be obtained from (here).

All these justifications, indicates that our formula can also detect fallacies in datasets and this dataset seems to not be fair considering all the test results and the above graphs.

Moreover, this dataset i.e., Credit Card Transaction seems to have some underlying error or miscalculation proving the dataset to fallacious or fraud.

### 6 Scope of further Research

- We have interpolated the polynomial for the range \([1,9]\). Though interpolating it in the range \([-9,9]\) – \(\{0\}\) can be considered for a scope of further research that would give a better range to work up on and would create a new field of finding probabilities with negatives and would be a great pillar to hunch upon.

- Unlike Benford’s law our proposed formula can be applied in various fields like linguistics, forensics, finances, elections and other socio-economic data. For different kind of datasets, we can have a tailor-made formula for variable ranks i.e., if a dataset requires rank up to say ‘n’ then we can apply Lagrange’s Interpolation for \(x = 1, 2, \ldots, n\) and the corresponding \(f(x)\) values. We have performed Lagrange’s Interpolation for rank up to 9 as first digit of a number can’t exceed 9 but in case of other datasets, we can derive a different interpolated formula with rank ‘n’.

### 7 Conclusion

By all counts, and with proven results, we would hereby like to conclude all of our work that we have put across our paper. Numerous datasets have been worked up on and have been operated on to verify their flawlessness and fairness, using Benford’s law and Zipf’s law. Not only that, we have used a lot of fitting tests that ensured us to apply the pre-existing laws on the datasets. ‘Prediction matches reality if it’s made in reality’ have been gladly proven by these two laws with predicted assimilation. Further, we would like to put across the fact that, the datasets of the COVID – 19 (uploaded by World Health Organization) is pristine. At the same time, some irksome fallacies have been desolated in the abstract dataset of the Credit Card Transaction details.

Asunder from that, the formula that have been pinned up by us in this study, is also a great scope to work upon. Our formula, in spite of being erected using observation, unlike other laws, have been justified with colossal clarifications. Finally, it is hoped that our proposed methodology may open up new vistas in the way of making decisions in the field of fraud detection.

### References


