Benford's Law as an Extension of Zipf's Law

Analytically, samples taken from a log-uniform distribution comply with Benford's law (BL). A statistical derivation of Benford's law, originally given by Hill, relies on this fact. Zipf's law describes the occurrence of words in a given languages and follows a similar digit law to BL. When a language has an infinite number of words Zipf's law reduces to the Riemann Zeta function. By considering an extension of Zipf's law as a summation of an uncountable infinite number of languages, each with an infinite number of words, we show a connection between Zipf's law, the Zeta function and Benford's law. This immediately extends the BL beyond its classical definition and provides a rich mathematical structure to the theory which is related to the Zeta function.

Contents

1. Introduction

1.1. Benford's first digit law

Benford's law (BL)—also referred to as the Newcomb–Benford law—describes the phenomenon that the probability of the digits $1, 2, ..., 9$ to occur in the first index in a number for many real-world data sets is not uniformly distributed [2]. Newcomb gave the first statement of the first-digit law after realising that the first pages of logarithmic tables wear out faster than subsequent pages [10]. Benford extended this idea by manually analysing twenty naturally occurring datasets for conformity with Newcomb's observation.

Benford's law arises as many datasets, particularly naturally occurring datasets. This is counter-intuitive as, at the surface, there appears to be no reason why the occurrence of digits in the first index would not be uniformly distributed. However, Hill gave a statistical proof of Benford's law, which relied on the assumption that the underlying distribution was log-uniform

rather than uniformly distributed. For instance, the Fibonacci numbers, whose underling distribution is a log-uniform distribution, appear in the growth patterns of sunflowers, pinecones and other plants and flowers [11]. Fundamentally, naturally occurring datasets arise from the ratio of two distinct quantities, which in turn brings their distribution close to a log-uniform distribution; therefore, they are often compliant with Benford's law [**?**]. Benford's law has been applied to a variety of different scientific fields including, astrophysics [6], biophysics [5]; geophysics [7]; particle physics [13] and quantum critical phenomena [4]. This allows for data trends and anomaly detection in a computationally simple way. Further applications of Benford's law can be found at [1].

1.2. Benford's Law and Anomaly Detection

Benford's law allows researchers and analysts to summarise large amounts of data and test for conformity with a log-uniform distribution. As with all such analysis, there may be compelling reasons for a dataset conforming, or conversely not conforming, with Benford's law. Additionally, one of the challenges with analysing datasets for conformity with any digit based law is the availability of a reliable control group. Since Benford's is effective in analysing changes between two similar datasets [**?**], its usually appropriate to compare the results of a Benford analysis between datasets. However, as all log-uniform datasets are Benford, such distributions can be used as a basis to understand the conditions in which Benford's law applies to particular datasets. Considering small perturbations away from log-unfoirom distributions and distributions which span a relatively small number of orders of magnitude will give further insight into the applicability of Benford's law.

Benford's law has been used extensively in the field of anomaly and fraud detection ranging from music streaming [9], data quality assessments data [14], cryptocurrency transactions [**?** , chen2022antibenford]isrepresented socio-economic data [12] and psychological based pricing [3]. Intervention or manipulation of data often alters the conformity of a dataset with Benford's law. Such intervention may be intentional or a result of some external factor or systemic another change to the underlying dataset. For example, changes in the taxation of a financial asset my alter the buying and selling of that asset class in order to minimise the tax paid. A change in conformity with Benford's law does not necessarily mean that data has been fraudulently changes but often indicate that some fundamental change has taken place that should be investigated further.

2. Benford distributions

This section looks at the mathematical definition of Benford's law and log-uniform distributions. We also explore the connection between geometric series and Benford's law and give examples which are and are not Benford compliant.

2.1. Benford's Law

The term index is used to refer to a position in a number. At each index there will be an integer $0, 1, 2, ..., 9$. Indexes begin at the first significant digit in a number. For example, for the integer 6301, the first index D_1 references the first position with a digit of 6, the second index D_2 references the second position, with a digit of 3 and similarly for the third and fourth indexes. In this case we would stay that $D_1 = 6, D_2 = 3, ...$ or $(D_1, D_2, D_3, D_4) = (6, 3, 0, 1)$. By assumption D_1 is non-zero.

2.2. Zipf's Law

Zipf's Law is an empirical observation about the distribution of frequencies of items in a dataset. Named after linguist George Zipf, the law suggests that in many natural language corpora and other types of datasets, the frequency of the most common item is approximately inversely proportional to its rank. In simpler terms, the second most common item appears half as often as the most common one, the third most common item appears one-third as often, and so on. This power-law distribution indicates that a small number of items dominate, while the vast majority occur relatively infrequently.

Mathematically, Zipf's Law can be defined using the equation:

$$f(r) = \frac{c}{r^s}$$

where:

- $f(r)$ is the frequency of an item at rank r,
- c is a constant that depends on the dataset,
- s is a scaling parameter (usually close to 1) that determines the steepness of the distribution.

An illustrative example of Zipf's Law can be found in the word frequencies in a large text corpus. In the English language, a small set of high-frequency words (such as 'the', 'of', 'and') occur very frequently, while the vast majority of words occur less often. The most common word, 'the', appears far more often than the second most common word, 'of,'. This distribution is often called 'heavy-tailed' because it has a long tail of infrequent items.

If consider Zipf's law for a language with an infinite number of words, then the distribution converges to the Riemann Zeta Function[??]:

$$f(s) = \zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

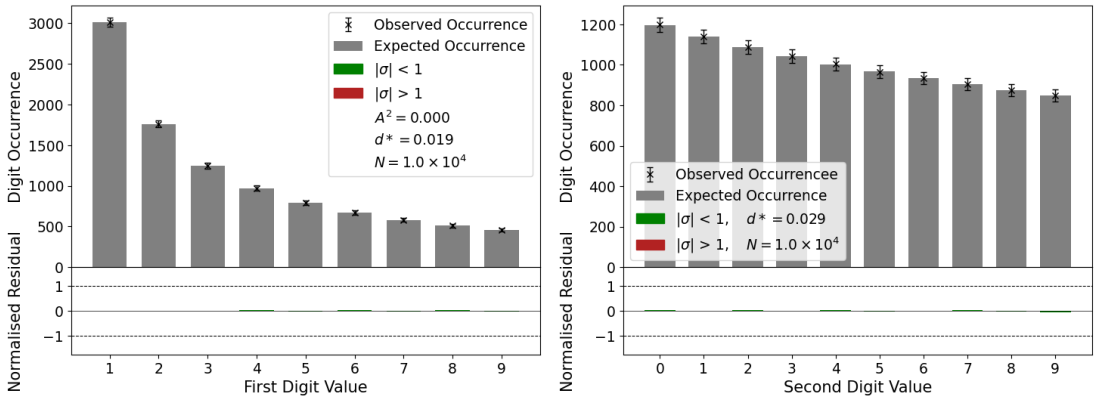

FIG. 1: The two plots show Benford's law applied to a log uniform dataset in the range $[1, 10^{10}]$ with a total of 10^4 datapoints. The lefthand plot shows the first digit law and the righthand plot the second digit test. The expected value of the occurrence, E_i , are in accordance with Benford's law, plotted as a cross, with error bars equal to the Poisson noise, $\sqrt{E_i}$. The actual observed occurrences, O_i , are shown for each digit as a bar. For conformity, we expect the residuals to be normally distributed with a variance less than one. In this case, all the residuals are within one standard deviation of the mean (zero) indicating conformity with Benford's law. d^* is a metric, which will later be defined, used to measure conformity of datasets with Benford's law, which will be discussed later in this paper.

2.3. Log-Uniform Distribution

A distribution is log-uniform if the logarithm of the analysed variable is uniformly distributed. In other words, if X is a random variable with a uniform distribution, then $\log_{10} X$ has a log-uniform distribution.

In terms of the probability distribution function (PDF) a log-uniform distribution can be classified in the following way. Let X be log-uniformly distributed over the range $[a, b]$ with $a < x < b, \forall x \in X$. Then the PDF is given by:

$$f(x; a, b) := \frac{1}{x[\log_e b - \log_e a]} = \frac{1}{x \log_e \frac{b}{a}}. \tag{1}$$

By taking the limiting case where $a := 0$ and $b \to \infty$ we can derive Benford's first digit law (or indeed any other digit law we require). This is done by considering the ranges for which the index D_1 admits a first digit d_1 and normalising over the entire range of data. This can be done across different base representations and over finite ranges, which gives rise to the finite range Benford Law's.

Figure 1. shows Benford's law applied to log-uniform data generated in the range $[0, 10^{10}$ which spans 10 orders of magnitude. The plot shows the expected and observed first digit counts and the normalised residuals for these values. As all residuals are within the Poisson error of the observed data we can conclude that the distribution complies with Benford's law. More rigorous methods to measure conformity with Benford's law will be discussed later in this paper.

As geometric series are log-uniform [3], generating log-uniform data is equivalent to generating a geometric series starting as a and ending at b with the desired number of terms in the

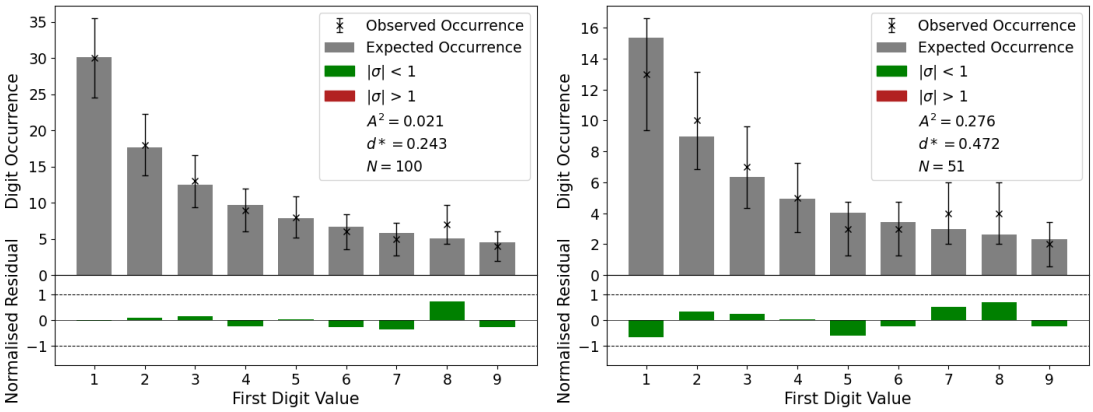

FIG. 2: The two plots show Benford's law applied to two naturally occurring datasets. The lefthand plot shows the first digit law applied to the first 100 Fibonacci numbers and the righthand plot shows the same law applied to all known generators of Mersenne primes (currently 51). The expected value of the occurrence according to Benford's law is plotted alongside the observed occurrences. The residuals are shown in the lower subplot in each figure. As all the residuals are within one standard deviation of the mean (zero), this indicates conformity with Benford's law. $d*$ is a metric, which will later be defined, used to measure conformity of datasets with Benford's law, which will be discussed later in this paper. The Fibonacci numbers are log-uniform at higher terms, as the series converges to a geometric series with the geometric ratio being the Golden ration. The Mersenne primes generators are roughly log-uniform as the distribution of the ratios between subsequent terms is normally distributed with a small standard deviation.

sequence. This process has been previously investigated by [3] in the context of Benford's law.

Several naturally occurring series, such as the Fibonacci series and the Mersenne prime generators are Benford; as shown in Figure.2. Indeed these distributions appear to be roughly log-uniform. Whilst conformity is less pronounced during an initial inspection of the plots, this is to be expected, as each plot has fewer associated data points leading to larger Possion noise. All residuals are within one standard deviation of the expected observation due to Benford's law, indicating conformity.

3. Benford's Law as an Edge Case of Zipf's Law

Consider Zipf's law applied to a language with an infinite number of words. The Zipf's law is equivalent to the Riemann Zeta function. Consider the following Benford ratio defined as:

$$B(s; D_1 = d_1) := \frac{\zeta(s)_{|D_1=d_1}}{\zeta(s)}, \tag{2}$$

which is the ratio of the Zeta function summed over the first digit D_1 to the Zeta function itself. Now Zipf's law is only defined for $s > 1$. Consider the edge case where $s = 1$ and consider a continuation of Zipf's where sums are replaced by integrals. Then:

$$B(s = 1; D_1 = d_1) = \frac{\zeta(1)_{|D_1=d_1}}{\zeta(1)} \to \frac{\int_{D_1=d_1} \frac{1}{x}\, dx}{\int_0^\infty \frac{1}{x}\, dx} = \log_e\left(1 + \frac{1}{d_1}\right), \tag{3}$$

which is mathematical statement of Benford's law. Zipf's law in the continuous case can be interpreted as a superposition of an uncountable infinite number of infinite word languages'[??]. In this case we arrive at Benford's law. Therefore, Benford's law can be loosely interpreted as the superposition of the Riemann zeta function[??] over different modes of vibration given by the state parameter s. As such, Benford's law and Benford ratios provide a rich mathematical structure which is related to the Zeta function and further generalisations of Benford's law.

3.1. Perturbations Around Zipf's Law

We have already discussed Zipf's law and its relation to the Riemann Zeta function as the case where the language considered has an infinite number of words. However, it is unclear how this relates to Benford's law and its generalisations. Firstly, we can consider small perturbations around $\zeta(1)$ by introducing the parameter ϵ. Consider the following ratio:

$$B(s = 1 + \epsilon; D_1 = d_1) = C\frac{\zeta(1 + \epsilon)_{|D_1=d_1}}{\zeta(1 + \epsilon)} \tag{4}$$

This is similar to 4 except we have introduced a prefactor term C to ensure that normalisation holds; that is, the sum over the possible values of D_1 is one. By considering different values of ϵ we can introduced perturbations around Zipf's law. We can then analyse conformity with Zipf's law at these varying perturbations.

3.1.1. Measuring Conformity

To measure conformity with Zipf's law we can use the widely adopted χ^2 and reduced χ^2 [8] test statistics. These statistics have their own benefits and drawbacks, however the motivation for their use is their simplicity to calculate and their clear and well understood interpretation.

The χ^2 statistic for discrete data takes the form [8]

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}, \tag{5}$$

where E_i is the expected occurrence as calculated from the probability distribution, and O_i is the observed occurrence. When calculating the likelihood of a value of χ^2 to occur, the reduced χ^2 statistic, χ^2_ν, is useful:

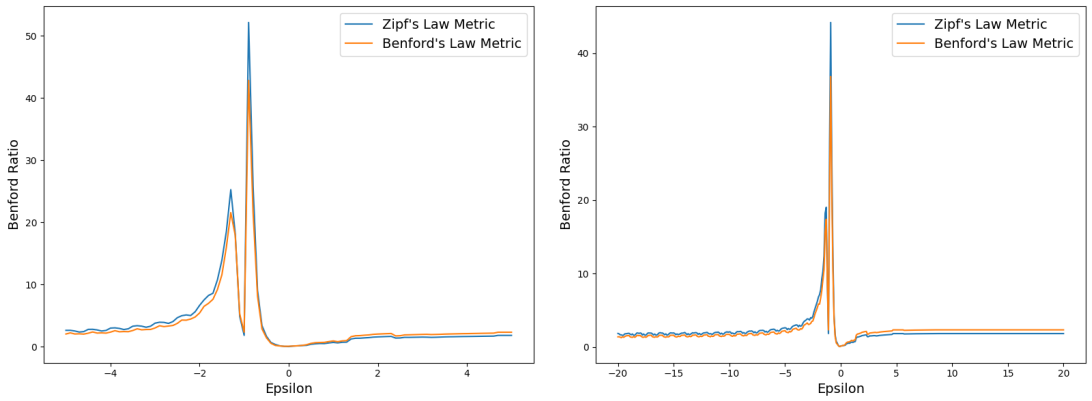

FIG. 3: The two plots show the change in the χ^2 statistics under perturbations away from Zipf's law. We have taken the limit where the language considered has an infinite number of words and have, as such, considered variations around $\zeta(1+\epsilon)$. Each calculation considers 10000 terms in the The lefthand graph shows variations away from both Zipf's law and the probability distribution given by Benford's law according to χ^2. In both cases a similar trend is followed with relatively good conformity around ϵ 0. Indeed χ^2 diverges at $\epsilon = -1.0$ with a local maxima and minima at $\epsilon \approx -1.0$ and $-1.34, 0.0$, respectively. The plot also appears to have an infinite discontinuity at $\epsilon = -1$; that is, at $\zeta(0)$. The similarity between the Zipf and Benford probability distributions suggests a connection between the two laws. The righthand plot shows the same result expanded over a larger set of values of ϵ.

$$\chi^2_\nu = \frac{\chi^2}{\nu}. \tag{6}$$

For a dataset with ν degrees of freedom a value of the χ^2_ν roughly equal to one indicates a good match with the parent distribution. In our case, if the perturbed dataset has a χ^2_ν value roughly equal to (or less than) one then the observed distribution complies well with Zipf's law.

3.2. Conformity With Zipf's Law

By calculating the χ^2 statistic for various values of ϵ we can observe any variability. 3 shows the χ^2 and χ^2_ν metrics for deviations around Zipf's law. Each language in question encompasses an infinite number of words. We've considered deviations around $\zeta(1+\epsilon)$ with each series being evaluated with 10^4 terms. The Benford ratio is then taken – the ratio between the sum of terms with first significant d_1 and the entire series.

In both plots we observe deviations from both Zipf's law and the probability distribution derived from Benford's law at certain values, evaluated via the the χ^2 statistic. Similar patterns emerge in both cases with good conformity when ϵ is proximate to zero. However, the ratios diverges at at $\epsilon = -1.0$ with an infinite discontinuity, suggesting the both Zipf's and Benford's law are not defined for $s = 0$. We therefore do not expect Zipf's law to have a continuous extension to arbitrary values of s in the Zeta function. Furthermore, intriguing local maxima and minima can be observed at $\epsilon \approx -1.0, -1.34$ and $\epsilon \approx -1.0, 0.0$, respectively. We therefore expect poor conformity with Zipf's law around the former maxima and better conformity around the latter minima.

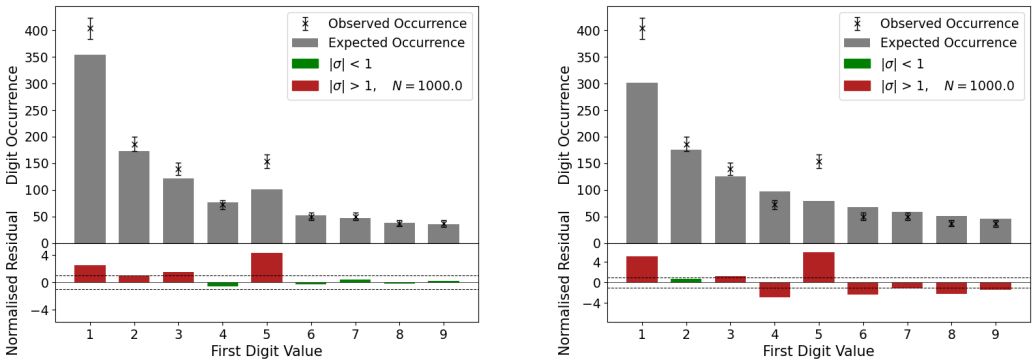

FIG. 4: The lefthand and righthand plots show Zipfs and Benford's law applied to the first 1000 terms in the Benford ratio respectively. The sum has been taken from 0 to the current term (less that 1000). The digit 5 is significantly more likely to occur when applying Zipf's law (when considering an infinite number of words in the language) compared with Benford's law. In general, the form of each distribution is similar with higher significants being less likely to occur than lower significants in the first index. However, this characteristic does not hold for the significant 5, showing that whilst Zipf's and Benford's law have some similarities, the underlying distributions do differ in some respects. For Zipf's law, the expected occurences have been determined by calculating the Benford ratio for each significant with 10^6 terms in each sequence.

Additionally, we see similar levels of conformity for both Zipf's and Benford's law suggesting that the two laws may be interlinked – that is, Benford's law is a continuous extension of Benford's law. However, it should be noted that the two observations yield different probability distributions. However, in the case that $s = 1$ they do yeild relatively similar properties. The main difference when compared with Benford's law is that for Zipf's law $P(D_1 = 4) < P(D_1 = 5)$, showing a greater proportion of the fist significant of 5. In the case of Benford's law $P(D_1 = x < D_1 = y$ for all $y > x$. However, this property does not appear to be evident when considering Zipf's law.

This phenomena can be clearly seen when viewing 4. The underlying dataset consists of the first 1000 terms of the Benford ratio. Zipf's and Benford's law has then been applied to the resultant distribution. For Benford's law, the occurrence of subsequent significants are predicted to be less likely than the current significant. This property does not hold for Zipf's law, with the significant 5 appears more frequently that the significant 4. This shows that whilst both distributions have a similar mathematical structure, the underlying distributions differ. Just because a distribution complies with one of these laws the same does not necessarily hold for the other. Indeed it will be unlikely that the Benford ratio of a sample distribution is compliant with both Benford's and Zipf's law.

Another point to note is that introducing deviations from Zipf's law will alter the normalisation required to ensure all probabilities sum to one. Therefore, when calculating generalisations of Zipf's law away from $s = 1$ it is important to ensure that all probabilities are suitably normalised. This cannot be done when $s = -1$ due to the infinite discontinuity in the Zeta function.

4. Conclusion

This research argues that there is a fundamental connection between Zipf' law, the Riemann Zeta function and Benford's law. Benford's law can be loosely interpreted as the superposition of the Riemann zeta function which in turn is the superposition of a unaccountably infinite number of infinite word languages. Therefore study of Benford's law could unlock a rich mathematical structure and allow for further extensions of Benford's law beyond log uniform distributions. However, significant differences do occur between Zipf's and Benford's law. Indeed, in the traditional sense, Zipf's law can only be applied to language based data and Benford's law to numerical datasets. However, when considering the extension via Benford ratios, it is possible to compare the two techniques.

References

[1] A. Berger, T. P. Hill, and E. Rogers. Benford Online Bibliography. Available at: https://www.benfordonline.net/. Accessed: 01.07.2021.

[2] Frank Benford. The law of anomalous numbers. Proceedings of the American Philosophical Society, 78(4):553, 1938.

[3] Alexander Long Benjamin Hull and Ifan G Hughes. Using residual heat maps to visualise Benford's multi-digit law. European Journal of Physics, 2021. Benjamin Hull et al 2022 Eur. J. Phys. 43 015803 10.1088/1361-6404/ac3671.

[4] Anindita Bera, Utkarsh Mishra, Sudipto Singha Roy, Anindya Biswas, Aditi Sen(De), and Ujjwal Sen. Benford analysis of quantum critical phenomena: First digit provides high finite-size scaling exponent while first two and further are not much better. Physics Letters A, 382(25):1639–1644, 2018.

[5] A.J. da Silva, S. Floquet, D.O.C. Santos, and R.F. Lima. On the validation of the newcomb-benford law and the weibull distribution in neuromuscular transmission. Physica A: Statistical Mechanics and its Applications, 553:124606, 2020.

[6] de Jong, Jurjen, de Bruijne, Jos, and De Ridder, Joris. Benford´s law in the gaia universe. A&A, 642:A205, 2020.

[7] Isadora A. S. de Macedo and Jose Jadsom S. de Figueiredo. Using benford's law on the seismic reflectivity analysis. Interpretation, 6(3):T689–T697, 2018.

[8] Ifan G Hughes and Thomas P A Hase. Measurements and Their Uncertainties : A Practical Guide to Modern Error Analysis. Oxford University Press, 2010.

[9] P Filzmoser N Mumic. A multivariate test for detecting fraud based on Benford's law, with application to music streaming data. Statistical Methods Applications, 2021. https://link.springer.com/article/10.1007/s10260-021-00582-6.

[10] Simon Newcomb. Note on the frequency of use of the different digits in natural numbers. American Journal of mathematics, 4(1):39–40, 1881.

[11] TO Omotehinwa and SO Ramon. Fibonacci numbers and golden ratio in mathematics and science. International Journal of Computer and Information Technology, 2(4):630–638, 2013.

[12] Timjerdine Said and Kaicer Mohammed. Detection of anomaly in socio-economic databases, by benford probability law. In 2020 IEEE 6th International Conference on Optimization and

Applications (ICOA), pages 1–4. IEEE, 2020.

[13] LIJING Shao and BO-QIANG Ma. First digit distribution of hadron full width. Modern Physics Letters A, 24(40):3275–3282, 2009.

[14] T Shyamili TS Rao T Shalini, TLS Rao. Data Quality Assessment Using Benford's Law and Excel. Institute of Electrical and Electronics Engineers, 2023. https://ieeexplore.ieee.org/abstract/document/10118030.