Robustness of Extended Benford's Law Distribution and Its Properties

Shar Nizam Sharif and Saiful Hafizah Jaaman-Sharman(✉)

Department of Mathematics, Faculty of Science and Technology, Universiti Kebangsaan
Malaysia, UKM, 43600 Bangi, Selangor, Malaysia
shj@ukm.edu.my

Abstract. It was anticipated more than a century ago that the distribution of first digits in real-world observations would not be uniform, instead follow a trend in which measurements with lower first digits occur more frequently than measurements with higher first digits. Frank Benford coined the term "First Digit Phenomena" to describe this phenomenon, which is now known as Benford's Law distribution. Benford's Law distribution has long been recognized but was widely dismissed as a mathematical oddity in the natural sciences. There is a theoretical requirement to analyze such disparities as departures from Benford's Law have been observed. The use of parametric extensions to existing Benford's Law is justified, as evidenced by the inclusion of k-tuples as a new parameter in the study. A k-tuples can be interpreted as a set of order and cardinality of first significant leading digit in datasets. Therefore, a convenience and concise method for deriving parametric analytical expansions of Benford's Law for first significant leading digits is proposed by embedding k-tuples. A new probabilistic explanation for the appearance of extended Benford's Law distribution has been discovered. As a result, a one-parameter analytical extension of Benford's Law for first significant leading digits is proposed. The new distribution generated by embedding k-tuples is scale invariant and robust to existing Benford's Law properties which a sum of first digit proportion is equal to 1, unimodality, logarithmic distribution and positive skewness. Then, mathematical features are investigated and a new generic class of moments generating functions is created. Based on natural phenomenon number, extended Benford's Law shows lesser values than existing one. This study found that the extended Benford's Law distribution to be better than the existing Benford's Law with measurements of lower digit occur more frequently.

Keywords: Benford's Law · mean · variance · skewness · kurtosis

1 Introduction

Benford's Law determines the predicted frequency of occurrence of digits in tabulated data. The anticipated digit frequencies collection is named for Frank Benford, a physicist who wrote the fundamental study on the subject (Benford 1938). Contrary to popular belief, Benford (1938) discovered that the digits in tabulated data are not equally likely and have a biased skewness in favor of the lower digits. Benford (1938) notes at the

N. Wahi et al. (Eds.): ICMSS 2022, ACSR 98, pp. 195–204, 2023.
https://doi.org/10.2991/978-94-6463-014-5_19

outset of investigation that the first few pages of a book of common logarithms exhibit more wear than the last few pages. Benford (1938) therefore finds that the initial few pages are more frequently used than the final few pages. The first few pages of logarithm books illustrate the logarithms of numbers with low initial digits, such as 1, 2, and 3. Benford (1938) hypothesized that the worn initial pages were caused by the fact that the majority of the world's "used" numbers had a low first digit. The first digit of a number is the leftmost digit. Zero is not permitted as the initial digit, thus there are nine potential initial digits (1, 2, …, 9). Additionally, the indications of negative numbers are disregarded (Nigrini 2012).

One of the important advances in Benford's Law literature growth was by Pinkham (1961). Pinkham (1961) presented the question that if there were truly some laws governing digital distributions then this law should be scale invariant. That is, if the digits of the areas of the world's islands, or the length of the world's rivers followed a law of some type, then it should be immaterial if these measurements were stated in miles or kilometers. Pikham (1961) then established that Benford's Law is scale invariant under multiplication. So, if all the numbers in a data table that followed Benford's Law were multiplied by a (nonzero) constant, then the new data table would also follow Benford's Law. A list of numbers that correspond to Benford's Law is known as a Benford Set. Hill (1995), Pietrenero et al. (2001), Nigrini (2015) and Berger and Hill (2011) have corroborated the discovery of Pinkham (1961).

Since Hill (1995), considerable attention has been paid to the theory of random variables in probability spaces with the goal of improving the understanding of probability distributions that follow or nearly approximate Benford's law. However, deviations from Benford's Law have been identified along this path in a variety of data sets from the natural sciences (Hürlimann 2015). As a result, various research, including those by Cho and Gains (2007), Carrera (2015), and Ausloos et al. (2017), have questioned the veracity of Benford's Law. Numerous research indicates that there is a theoretical requirement to account for such disparities. In accordance with Hill (1995) and Leemis et al. (2000), Hürlimann (2015) proposed refocusing on probability distributions that follow or closely approximate Benford's law.

The purpose of this study is to extend the existing Benford's Law mathematical model with k-tuples. Therefore, the new Extended Benford's Law is obtained. Based on Hürlimann (2015), an existing Benford's Law with $\alpha \in (-\infty, \infty)$ can be defined by Eq. (1).

$$GB(d;\alpha) = \left\{ \begin{array}{l} \dfrac{d^{-\alpha} - (1+d)^{-\alpha}}{1 - 10^{-\alpha}}, if\, \alpha \neq 0 \\ \log\left(1 + d^{-1}\right), if\, \alpha = 0 \end{array} \right\} \qquad (1)$$

$$d = 1, \ldots, 9$$

$$\alpha \in (-\infty, \infty)$$

Simply put, let $X = X(\alpha)$, and let D denote the integer-valued random variable satisfying the inequality $10^D \leq X < 10^{D+1}$. Then, the first significant digit Y of X can be written as $Y = \lfloor X.10^{-D} \rfloor$, where $\lfloor . \rfloor$ is the floor function and one has Eq. (2).

$$\begin{aligned} P(Y = d) &= \sum_{k=0}^{N-1} P\Big(d.10^k \leq X < (1+d).10^k\Big) \\ &= \sum_{k=0}^{N-1} \int\limits_{d.10^k}^{(1+d).10^k} f_X(x)dx \end{aligned} \tag{2}$$

Integrating Eq. (2) one obtains the first significant leading digit in Eq. (1) (Hürlimann 2015).

2 Methodology

First, The Extended Benford's Law mathematical equation must satisfy and robust to the Benford's Law properties. As such, the properties of Benford's Law are (Berger and Hill, 2011; Berger et al. 2017; Nigrini 2011; Pomykacz et al. 2017):

1. unimodal
2. decreasing logarithmic distribution‘
3. positively skewness
4. scale invariance

To excel the robustness, the following steps are taken:

Step 1: Fathom the Classical Benford's Law which the basis for Hürlimann existing Benford's Law.

Step 2: Understand the existence of k-tuples definition in General Benford's Law.

Step 3: Extend existing Benford's Law based on probabilistic process.

Step 4: Generate first four moments of Extended Benford's Law.

Step 1

Classical Benford's Law states that each proportion of the first significant leading digit $D_1 = \{d_1\} = \{1, 2, ...9\}$ will be decreasing distributed and follow a very particular logarithmic distribution. Classical Benford's Law can be represented through Eq. (3) for all $d_1 = 1, 2, ...9$ with base-10 (Benford, 1938).

$$P(D_1 = \{d_1\}) = log_{10}\Big(1 + d_1^{-1}\Big) \tag{3}$$

As depicted in Eq. (3), the distribution of the first significant leading digits in Classical Benford's Law is unimodal and heavily positively skewed toward the smaller significant leading digits.

Step 2

Hill (1995) has derived a General Benford's Law with adapting k -tuples representing cardinality of first significant leading digit. Encyclopedia of Mathematics edited by Hazewinkel (2000) has defined a k-tuples as a list and an ordered set of k elements. Specifically, k-tuple can be interpreted as a set of order and mathematically can be written as k-vector. Furthermore, Hazewinkel (2000) states tuple is a finite sequence which accept repetition of elements from some set X. Hazewinkel (2000) has expressed a tuple by $\langle x_1, ...x_k\rangle$, (x_i), $(x_i)_{i=1}^{k}$, $(x_i)_{i\in\{1,...k\}}$, $(x_1...x_k)$, or $x_1...x_k$. Generalized Benford's Law can be represented through Eq. (4) for all positive integers k and all $d_1 \in \{1, 2, ..., 9\}$ (Hill 1995).

$$P\left(D_1 = d_{1,...}D_k = d_k\right) = log\left[1 + \left(\frac{1}{\sum_{i=1}^{k} d_i \times 10^{k-i}}\right)\right] \tag{4}$$

After embedding k-tuples, Hill (1995) claims that Eq. (4) gives unimodal distribution, positive skew distribution, decreasing logarithmic distribution and scale invariant to the Classical Benford's Law.

Step 3

Research aims to adapt k-tuples into existing Benford's Law which represented by Eq. (1). Implicitly, Extended Benford's Law is obtained. To achieve the goal, the following stages are taken:

Stage 1: Create a sample space (Ω) for Extended Benford's Law

Stage 2: Identify the constrains for Extended Benford's Law

Stage 3: Determine probability to sample space elements for Extended Benford's Law

Stage 1: Research symbolize the set of outcomes by (Ω). Then, events are subsets $A, A \subseteq \Omega$ which A are assigned probabilities. The probability is a mapping $A \mapsto P(A) \in [0, 1], A \subseteq \Omega$ which satisfy the following axioms (Hassler and Hosseinkouchack 2019):

- $P(A) \geq 0$
- $P(\Omega) = 1$
- $P(\cup_i A_i) = \sum_i P(A_i)$ for $A_i \cap A_j = \varnothing$ with $i \neq j$

The sample space respect to Extended Benford's Law consist of all potentially for first significant leading digit outcomes: $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

Stage 2: For Extended Benford's Law must satisfy and robust to existing Benford's Law. Therefore, the constraints are remaining the same with Benford's Law properties as being mention by Berger (2017).

Stage 3: With consideration of all constraints, the probability of first significant leading digit proportion of Extended Benford's Law with k-tuples can be represented by

Eq. (5).

$$B(d;\alpha) = \left\{ \begin{array}{l} \dfrac{\left[\sum\limits_{i=1}^{k} d.10^k\right]^{-\alpha} - \left[1 + \left(d.10^k\right)\right]^{-\alpha}}{1 - 10^{-\alpha}}, \mathit{if}\, \alpha \neq 0 \\ \log\left(1 + d^{-1}\right), \mathit{if}\, \alpha = 0 \end{array} \right\} \quad (5)$$
$$d = \{1, ..., 9\}$$

Step 4

Generally, for a discreet distribution of X (Mohd Alias Lazim 2011) states that the raw moments of the distribution are determined as $E(X^m), m \in \{1, 2, 3...\}$ with $m =$ moment. By referring to Scott and Fasli (2001) whom adjusted the first four moment to Benford's Law, thus study is able to generate the first moment which can be denoted as mean value of the Extended Benford's Law distribution as shown in Eq. (6):

$$E\left(D_1^m\right) = \sum_{\forall d_1} \frac{d_1^m \times P_{D_1}(d_1)}{N} = \sum_{\forall d_1} \frac{d_1^m \times \frac{\left[\sum_{i=1}^{k} d.10^k\right]^{-\alpha} - \left[1 + \left(d.10^k\right)\right]^{-\alpha}}{1 - 10^{-\alpha}}}{N} \quad (6)$$

Next, the Extended Benford's Law second moment is the variance, $V(D_1)$ and can be written as in Eq. (7):

$$E\left[\left(D_1 - E(D_1)^2\right)\right] =: V(D_1)$$
$$V(D_1) = \sum_{\forall d_1} \frac{(d_1 - E(D_1))^2 \times \frac{\left[\sum\limits_{i=1}^{k} d.10^k\right]^{-\alpha} - \left[1 + \left(d.10^k\right)\right]^{-\alpha}}{1 - 10^{-\alpha}}}{N} \quad (7)$$

Then, the third moment of Extended Benford's Law is known as skewness, γ_1. Skewness of Extended Benford's Law measures the asymmetry of the distribution and can be represented as in Eq. (8). Based on Benford's Law properties, Eq. (8) gives positive skew indicating a heavy tail on the right.

$$\gamma_1 = \frac{E\left[(D_1 - E(D_1))^3\right]}{V(d_1)^{3/2}}$$
$$= \sum_{d_1=1}^{9} \frac{(d_1 - E(D_1))^3 \times \frac{\left[\sum\limits_{i=1}^{k} d.10^k\right]^{-\alpha} - \left[1 + \left(d.10^k\right)\right]^{-\alpha}}{1 - 10^{-\alpha}}}{V(d_1)^{3/2}} \quad (8)$$

Lastly, the fourth moment is kurtosis and can be computed by Eq. (9).

$$\gamma_2 = \frac{E\left[(D_1 - E(D_1))^4\right]}{V(d_1)^2}$$

$$= \sum_{d_1=1}^{9} \frac{d_1 - E(D_1)^4 \times \frac{\left[\sum_{i=1}^{k} d.10^k\right]^{-\alpha} - \left[1+\left(d.10^k\right)\right]^{-\alpha}}{1-10^{-\alpha}}}{V(d_1)^2} \tag{9}$$

Equation (9) take into consideration of all first significant leading digit as sample size. From Eq. (9), one can notice that the exponent is 4, thus the term always gives positive value in the summation. By using Eq. (9), normal distribution is often assumed if the kurtosis is close to 0 and the term for such property is being called mesokurtic distribution (Westfall 2014). Meanwhile, the term is being called platykurtic distribution or if the kurtosis is less than zero represent light tails distribution and leptokurtic distribution if the kurtosis is bigger than zero represent heavy tails distribution. Kurtosis is a measurement to interpret on the order of tail extremity (Westfall 2014). In summation, Westfall (2014) and Scott and Fasli (2001) summarizes that the kurtosis is decreasing as the tails of distribution becomes lighter and vice versa.

3 Result

In summary, study has conveyed several findings. First and foremost, study has provided the theoretical Extended Benford's Law frequency probability distribution by embedding *k*-tuple through Table 1.

Next, study has visualized theoretical Extended Benford's Law frequency distribution generated through Fig. 1. The generated Extended Benford's Law distribution by embedding *k*-tuple is robust to its classical Benford's Law properties. The generated Exyended benford's Law are sum of first digit frequencies equal to 1, unimodality, monotonous logarithmic distribution and positive skewness as can be seen through Table 1 and Fig. 1. In comprehensive manner, study has generated moment functions which are mean, variance, skewness, and kurtosis respect to Extended Benford's Law distribution. Finally, study has showed the comparison between existing and Extended Benford's Law empirically by using natural number respect to mean, variance, skewness, and kurtosis value.

As being mentioned, the mathematical development of Extended Benford's Law by embedding *k*-tuple is scale invariant to existing Benford's Law. Therefore, the probability distribution for first significant leading digit frequencies to scatter naturally in natural phenomenon setting is expected to be identical for both existing and Extended Benford's Law. Therefore, by using Eq. 1 (existing Benford's Law) and Eq. 5 (Extended Benford's Law), the probability of first significant leading digit is being presented in Table 1 and the distribution for first significant leading digit of Extended Benford's Law in Fig. 1.

The distribution for the first significant leading digit of Extended Benford's Law satisfies the theoretical features of both classical and existing Benford's Law. Therefore, the Extended Benford's Law is unaffected by its theoretical features and robust to its properties. As illustrated in Fig. 1, the distribution is unimodal, logarithmic, and has a monotonous declining function.

As aforementioned, Benford's Law establishes the projected frequency of occurrences for the first significant leading digits in natural datasets derived from natural

Table 1. Benford's Law Frequency Distribution

digit	probability
1	0.303
2	0.177
3	0.125
4	0.097
5	0.079
6	0.066
7	0.057
8	0.051
9	0.045
Total	1

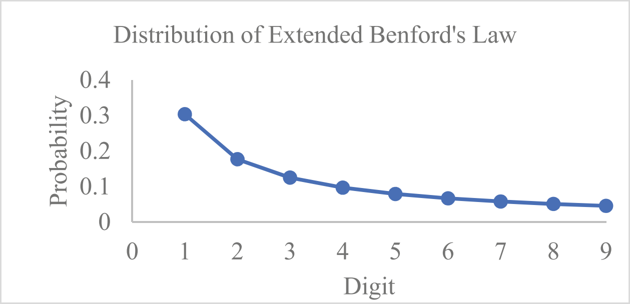

Fig. 1. Distribution for first significant leading digit of Benford's Law

processes. From Benford's law theoretical point of view, first significant leading digit derived from natural phenomena is bounded with minimum lower bound is 1 and maximum upper bound is 9 ([1, 9]). Digit 0 as first significant leading digit is not permissible and negative sign is being ignored. From previous studies, it is observed that many empirical data in the real world follow Benford's law such as Fibonacci numbers. The Benford's Law literature of the early 1970s included a of articles by Fibonacci theorists who showed that the familiar Fibonacci sequence (1, 1, 2, 3, 5, 8,…) follows Benford's Law perfectly. The Fibonacci Quarterly was the first journal to publish six Benford's Law papers in the same decade. Parenthetically, numbers from the naturally occurring Fibonacci sequences obey Benford's Law. Therefore, study has utilized the first 1000 Fibonacci numbers starting with the number 1. To see a reasonably good fit to Benford's Law, the distribution of first significant leading digit derived from Fibonacci sequence is being represented in Fig. 2.

Moving forward, mean, variance, skewness, and kurtosis of Fibonacci sequence are constructed for Extended and existing Benford's Law. Table 2 summarized the empirical

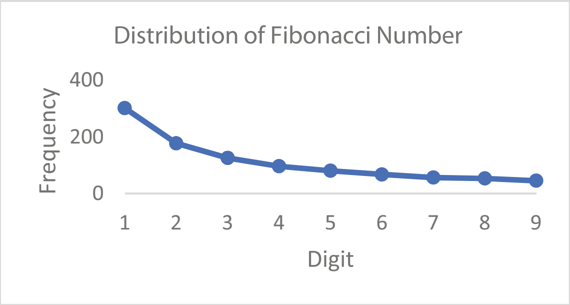

Fig. 2. Distribution for first significant leading digit of Fibonacci number

Table 2. Generated Moment Function

Moment	Extended		Existing
Mean	0.417	<	18.462
Variance	45.915	<	3476.328
Skewness	219.498	>	37.96
Kurtosis	7203.572	>	171.006

finding of mean, variance, skewness, and kurtosis of Fibonacci sequence for Extended and existing Benford's Law.

Based on Table 2, the mean and variance value of Extended Benford's law is lesser than existing Benford's Law towards smaller digit. Moreover, the value of skewness for Extended Benford's Law is higher than existing which indicates that the Extended Benford's Law is positively skewed to the lower digit. Finally, Extended Benford's Law has fat-tailed distribution compared to existing Benford's Law. In natural phenomena data, it has been agreed that the distribution of Extended Benford's Law in natural sciences data is more concentrated at the lower digit compared to higher digit.

4 Conclusion

In conclusion, study has demonstrated several vital and practical findings. First and foremost, study has presented the theoretical explanation for the development of Extended Benford's Law distribution. Robust to its properties, the Extended Benford's Law distribution is scale invariant, unimodal, positively skewed and the summation of all first significant leading digit probability frequency is equal to 1. Comprehensively, study is able to generate mean, variance, skewness and kurtosis for Extended Benford's Law distribution. Last but not least, study has showed the comparison between existing and Extended Benford's Law respect to mean, variance, skewness and kurtosis value in natural phenomenon number setting. By comparison, study has found that the Extended Benford's Law give lesser value in mean and variance. Then, the Extended Benford's

Law is more skewed to the right and demonstrates higher kurtosis value compared to existing Benford's Law.

References

Ausloos, M., Cerqueti, R., Lupi, C.: Long-range properties and data validity for hydrogeological time series: the case of the Paglia river. Physica A **470**, 39–50 (2017)

Benford, F.: The law of anomalous numbers. Am. Phil. Soc. **78**(4), 551–572 (1938)

Berger, A., Hill, T.P.: A basic theory of Benford's law. Probab. Surv. **8**(1), 1–126 (2011)

Berger, A., Hill, T.P., Silva, C.E.: WHAT IS … Benford ' s Law ? Notices Am. Math. Soc. **64**(2), 132–134 (2017)

Carrera, C.: Tracking exchange rate management in Latin America. Rev. Finan. Econ. **25**, 35–41 (2015)

Cho, W.K.T., Gaines, B.J.: Breaking the (Benford) law: statistical fraud detection in campaign finance. Am. Stat. **61**(3), 218–223 (2007)

Hassler, U., Hosseinkouchack, M.: Testing the Newcomb-Benford Law: experimental evidence. Appl. Econ. Lett. **26**(21), 1762–1769 (2019)

Hazewinkel, M.C.: In: Hazewinkel, M. (eds) Encyclopaedia of Mathematics. Encyclopaedia of Mathematics. Springer, Dordrecht (2000). https://doi.org/10.1007/978-94-015-1279-4_3

Hill, T.P.: A statistical derivation of the sidnificant-digit law. Stat. Sci. **10**(4), 354–368 (1995)

Hürlimann, W.: Benford's Law in scientific research. Int. J. Sci. Eng. Res. **6**(7), 143–148 (2015)

Leemis, L.M., Schmeiser, B.W., Evans, D.L.: Survival distributions satisfying Benford's law. Am. Stat. **54**(4), 236–241 (2000)

Lazim, M.A.: Introductory Business Forecasting a Practical Approach. D. of U. Press (Pnyt.) hlm, 3 edn. UITM Press, UITM, Shah Alam (2011)

Nigrini, M.J.: Benford's law application for forensic accounting, auditing, and fraud dectection. In: Wells, J.T. (Pnyt.) (ed.) John Wley & Sons, Inc, hlm. John Wley & Sons, Inc., Hoboken (2012)

Nigrini, M.J.: Persistent patterns in stock returns, stock volumes, and accounting data in the U.S. capital markets. J. Account. Audit. Finan. **30**, 1–17 (2015)

Nigrini, M.J.: Forensic Analytics Methods and Techniques for Forensic Accounting Investigations. John Wiley & Sons, Inc., hlm, Hoboken (2011)

Pietronero, L., Tosatti, E., Tosatti, V., Vespignani, A.: Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf. Physica A: Stat. Mech. Appl. **293**(1–2), 297–304 (2001)

Pomykacz, M., Olmsted, C., Tantinan, K.: Benford's Law in Appraisal (2017)

Pinkham, R.S: On the distribution of first significant digits. Ann. Math. Statist. **32**(4), 1223–1230 (1961)

Scott, P., Fasli, M.: Benford's Law: An Empirical Investigation and a Novel Explanation, pp. 1–21. Não publicado (2001)

Westfall, P.H.: Kurtosis as peakedness, 1905–2014. R.I.P. Am. Stat. **68**(3), 191–195 (2014)