Palestine Journal of Mathematics

Vol. 11(3)(2022) , 379–385

BENFORD'S LAW: A NUMBER-THEORETICAL PERSPECTIVE

Nicolò Cangiotti and Mattia Sensi

Communicated by H. M. Srivastava

MSC 2010 Classifications: Primary: 37A45, 62E10. Secondary: 11K06, 11K16.

Keywords and phrases: Benford's law, first-digit law, normal numbers.

Abstract The historical journey of the Benford's law and its most important definitions and properties are shortly reviewed. Firstly, we define a new class of numbers based on the first-digit law. Secondly, we investigate the relation between Benford's sequence and normal numbers.

1 Introduction

In 1881, on the pages of the *American Journal of Mathematics*, a curious paper written by Simon Newcomb [10] appeared. It began with the following sentence:

> That the ten digits do not occur with equal frequency must be evident to any one making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones.

This first attempt to understand the deep behaviour of the distributions of the digits of sets (actually, sequences) of real numbers led Frank Benford to analyze many real-life sets of numerical data from different fields. In 1938, he published a manuscript [1], in which he formalized the distribution of such digits defining a suitable probability function. In particular, in its first formulation, Benford proposed the following *first-digit law*:

$$F(n) = \log_{10}\left(\frac{n+1}{n}\right), \quad n \in \{1, 2, \dots, 9\},$$

where $F(n)$ denotes the frequency of the digit n in the first place of the base-ten representation of a sequence of real numbers. That pioneeristic paper allowed Benford to tie his surname to the first-digit laws, which is indeed also known as *Benford law*[1]. In the following years, many other authors carried out studies on the properties linked with the sequences, which respect the first-digit law, the so-called *Benford sequences*. Moreover, generalizations including the distribution of all digits of a number have been proposed, and nowadays it is possible to find an extensive literature on this topic. Yet another interesting point of view is given by the probabilistic version of the Benford's law, which consists in defining a random variable with a suitable probability function. However, in our context, it is highly non-trivial to formalize the main idea in the sense of probability theory, as we explain in the following.

For the interested reader, we suggest to consult [2, 3, 4, 9, 11] for a general framework on these topics. Moreover, for a very extensive collection of works concerning the Benford's law we refer to the online database *Benford Online Bibliography* [5].

In this work, we propose a new approach on the first-digit law that could be considered as the first brick of the investigation in such a direction. The foundational idea is to provide a number-theoretical characterization of the sequences for which first-digit law holds. To achieve that, we define a particular class of numbers, which are closely related with the Benford sequences. Furthermore, we propose a brief study on the relation between these sequences and the normal number, another bizarre mathematical object, defined in 1909 by Émile Borel [6]. All these results are collected in Section 3. In Section 2, we introduce some useful notations and definitions,

[1] Sometimes also the name of the first discoverer accompanies his, and the law is hence called the *Benford-Newcomb law*.

and summarize some interesting results, which shall be used in the following. Section 4 is devoted to take stock of what we have studied in this manuscript, and to provide some insights for future developments.

2 Definitions and basic properties

In the following, we recall some fundamental definitions and basic properties underlying the concept of Benford's law. The notations we shall adopt are the same used in [4]. In particular, we denote with $\mathbb{N}$ the set of positive integer numbers (i.e. natural numbers), with $\mathbb{Z}$ the set of integer numbers, and with $\mathbb{R}$ the set of real numbers. A sequence of real numbers shall be denoted by $(\cdot)$, for instance the sequence of natural numbers is given by (n), with $n \in \mathbb{N}$. The *integer part* and the *fractional part* of a real number ω are denoted by $\lfloor \omega \rfloor$ and $\langle \omega \rangle$, respectively. For the remainder of the paper, we use the base 10 (the same results can be recovered for the other bases by an easy adaptation). Hence, for ease of reading, we will not specify the base in the continuation of the work; for instance $\log_{10}(\cdot)$ will be simply denoted by $\log(\cdot)$.

Firstly, we need to to formalize the notion of *significand* of a number (also known as *mantissa*).

Definition 2.1. Let $\omega \in \mathbb{R}^+$, we call the *(decimal) significand of* ω the unique number $S(\omega) = s$, such that $s \in [1, 10)$ and $\omega = 10^k s$ for some $k \in \mathbb{Z}$. Moreover, for $\omega \in \mathbb{R}^-$, we set $S(-\omega) = S(\omega)$, and $S(0) = 0$.

Definition 2.2. The *first (decimal) significand of* $\omega \in \mathbb{R}$, denoted by $D_1(\omega)$, is the first (leftmost) digit of $S(\omega)$, where we consider by convention the terminating decimal representation if $S(\omega)$ has two decimal representations (*e.g.*, between $0.999\ldots$ and 1, we choose the latter). Analogously, $D_2(\omega)$ is the second digit of $S(\omega)$, and so on.

Example 2.3. We have, for instance, $S(2021) = S(0.2021) = S(202.1) = 2.021$. Moreover, the first significand is $D_1(2021) = D_1(0.2021) = D_1(202.1) = 2$ and also $D_4(2021) = D_4(0.2021) = D_4(202.1) = 1$. We notice that $D_i(2021) = 0$, for every $i \geq 5$.

We are now ready to give the formal definition of *Benford sequence*, which generalizes the formula introduced by Benford to describe the first-digit law.

Definition 2.4. A sequence of real numbers (ω_n) is said to be a *Benford sequence* (*Benford* for short), if the following holds:

$$\lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : S(\omega_n) \leq t\}}{N} = \log t,$$

for every $t \in [1, 10)$.

Example 2.5. The sequences (2^n) and (3^n), as well as the Fibonacci sequence, are Benford. Instead, the sequence of integer numbers (n) is not Benford. For these and many other examples, the interested reader can refer to [3].

A very useful characterization of Benford sequences is given by the following statement.

Proposition 2.6. *A sequence* (x_n) *of real numbers is Benford if and only if*

$$\lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : D_1(x_n) = d_1, D_2(x_n) = d_2, \ldots, D_m(x_n) = d_m\}}{N} = \\ = \log\left(1 + \frac{1}{10^{m-1}d_1 + 10^{m-2}d_2 + \cdots + d_m}\right),$$

for all $m \in \mathbb{N}$, *all* $d_1 \in \{1, \ldots, 9\}$, *and all* $d_j \in \{0, \ldots, 9\}$, *with* $2 \leq j \leq m$.

Remark 2.7. Sometimes (see, e.g., [3, Sect. 3]), Prop. 2.6 is given directly as an equivalent definition of Def. 2.4.

Remark 2.8. From Prop. 2.6 we obtain immediately the well-known *first-digit law*, which we are going to use in the following sections. Indeed, for every Benford sequence of real numbers (x_n), we have:

$$\lim_{N\to\infty} \frac{\#\{1 \le n \le N : D_1(x_n) = d\}}{N} = \log\left(1 + \frac{1}{d}\right), \quad d \in \{1, \dots, 9\}. \tag{2.1}$$

We notice that satisfying the first-digit law does not imply that a sequence is necessarily Benford (we refer to [3] for an elegant proof of this result).

Theorem 2.9. *A sequence of real numbers (x_n) is Benford if and only if the sequence $(\log|x_n|) = (\log|x_1|, \log|x_2|, \log|x_3|, \dots)$ is uniformly distributed modulo 1.*

Remark 2.10. Thanks to Thm. 2.9 we can prove easily that the sequence (2^n) is Benford. In fact, the classical equidistribution theorem of Weyl [13] states that a sequence $(na) = (a, 2a, 3a, \dots)$ is uniformly distributed mod 1 if and only if a is irrational. Thus, we notice that $\log(2^n) = n\log(2)$ and, since $\log(2)$ is irrational, we can conclude.

3 The number-theoretical approach

The following paragraphs are devoted to provide a new perspective on Benford sequences. Firstly, we construct a class of number strictly linked to the first-digit law. Then we study the relation between Benford's sequences and normal number.

3.1 The class of fd-numbers

Number theory represents a useful tool in the studies of Benford sequences. The first step in this direction is to understand how it is possible to formalize a number-theoretical concept that coincides with Benford's structure.

For ease of notation, we introduce a new object, based on Def. 2.1 in Sect. 2.

Definition 3.1. Let ω be a real number. The *(decimal) full fractional representation* of ω is given by

$$R(\omega) = 10^{-1}S(\omega).$$

Example 3.2. We have, for instance, $R(2021) = R(0.2021) = R(202.1) = 0.2021$.

We now introduce a new class of real numbers, deeply intertwined with the first-digit law.

Definition 3.3. A real number ω is a *first-digit number* (concisely, in the rest of the manuscript, a *fd-number*) if and only if the sequence

$$(\mathcal{R}_n(\omega)) := \left(\langle R(\omega) \cdot 10^{n-1}\rangle\right) \tag{3.1}$$

respects the first-digit law. The set of all fd-numbers shall be denoted by $\mathfrak{fd}$.

Example 3.4. We provide now a construction of a fd-number. Let us take the Benford sequence (2^n), and let us consider the number constructed by taking the first-digits of each number of the sequence and concatenated them, namely

$$\omega = 248136125124\dots.$$

Its full fractional representation is given by

$$R(\omega) = 0.248136125124\dots.$$

It is not difficult to verify that the sequence $(\mathcal{R}_n(\omega))$ satisfies the first-digit law.

Theorem 3.5. *Let (x_n) be a sequence for which the first-digit law holds. Then the real number ω, constructed by concatenating all the leading digits of (x_n), is a fd-number.*

Proof. By Rmk. 2.8, we know that a sequence (x_n) respect the first-digit law if

$$\lim_{N\to\infty} \frac{\#\{1 \le n \le N : D_1(x_n) = d\}}{N} = \log\left(1 + \frac{1}{d}\right).$$

Let ω be the number constructed by concatenating all the leading digits of (x_n). Thus, by considering $(\mathcal{R}_n(\omega))$, we obtain immediately

$$\begin{aligned}\lim_{N\to\infty} \frac{\#\{1 \le n \le N : D_1(\mathcal{R}_n(w)) = d\}}{N} &= \lim_{N\to\infty} \frac{\#\{1 \le n \le N : D_1(x_n) = d\}}{N} \\ &= \log\left(1 + \frac{1}{d}\right).\end{aligned}$$

□

Remark 3.6. We notice that, given a fd-number, it is easy to construct other fd-numbers using the first one as model. For example, one can consider the number ω of Ex. 3.4 and insert a *zero* between every two consecutive digits:

$$\nu = 2040801030601020501020 40\ldots,$$

obtaining again a fd-number.

Remark 3.7. An equivalent way to define fd-numbers is the following. A real number ω is a first-digit number if and only if the sequence:

$$\left(\widetilde{\mathcal{R}}_n(\omega)\right) := \left(R\left(\mathcal{R}_n(\omega)\right)\right) \tag{3.2}$$

respects the first-digit law. The two sequences (3.1) and (3.2) are usually the same, but not always. In fact, let us take, for instance, the number ν of Rmk. 3.6. We immediately obtain:

$$\begin{aligned}(\mathcal{R}_n(\omega)) &= (0.204080\ldots, 0.040801\ldots, 0.408010, \ldots, 0.080103\ldots, \cdots); \\ \left(\widetilde{\mathcal{R}}_n(\omega)\right) &= (0.204080\ldots, 0.408010\ldots, 0.408010, \ldots, 0.801030\ldots, \cdots).\end{aligned}$$

It is not difficult to verify the equivalence of these two definitions.

As we mentioned in the introduction, constructing a probabilistic space for fd-numbers is a non-trivial issue. In fact, the count of the *zeros* in the construction of a suitable probability function requires (and prompts) further reflection.
One may proceed as follows. We denote the set of base-10 digits by $\mathfrak{N} := \{0, \ldots, 9\}$. Given a real number $\omega \in \mathbb{R}$, we denote the string of digits in its decimal expansion by $[\omega]$, indexed by the inverse of correspondent exponent of 10 in its 10-representation, namely

$$\sum_{i=k}^{\infty} s_i 10^{-i} = [s_k, s_{k+1}, \ldots].$$

We recall now the definition of *Borel number*, as it was introduced in [7].

Definition 3.8. We say that $\omega \in \mathbb{R}$ is a *Borel number* if for every $n \in \mathbb{N} \setminus \{0\}$ and every string $[s_1, \ldots, s_n] \in \mathfrak{N}^n$, the following limit

$$\lim_{m\to\infty} \frac{\left|\{[b_k, \ldots, b_{k+n-1}] \subset [\omega] \ : \ [b_k, \ldots, b_{k+n-1}] = [s_1, \ldots, s_n]\}_{1\le k\le m-n+1}\right|}{m} \tag{3.3}$$

exists. When it exists, we denote it by $\mathbb{P}([s_1, \ldots, s_n])$.

In the same spirit of Def. 3.8, we simply denote by $[s]$ the event that a generic digit in the decimal representation of ω is equal to $[s]$. Unfortunately, Def. 3.8 should be amended to omit the *zeros* from the count. Thus, we denote by $[\omega]_0$ the string of digits in the decimal expansion

of $\omega \in \mathbb{R}$ in which we have removed the *zeros*. Thanks to this new construction, we can define a new probability by the following limit:

$$\lim_{m\to\infty} \frac{|\{[b_k,\dots,b_{k+n-1}] \subset [\omega]_0 \ : \ [b_k,\dots,b_{k+n-1}] = [s_1,\dots,s_n]\}_{1\le k\le m-n+1}|}{m}.$$

when it exists or every $n \in \mathbb{N}\setminus\{0\}$ and every string $[s_1,\dots,s_n] \in \mathfrak{N}_0^n$, with $\mathfrak{N}_0 := \mathfrak{N}\setminus\{0\}$. We shall denote it by $\hat{\mathbb{P}}([s_1,\dots,s_n])$.

Example 3.9. Let us consider the following real number

$$\omega = 0.10203040506070809.$$

Thus we have:

$$\begin{aligned}[\omega] &= [0,1,0,2,0,3,0,4,0,5,0,6,0,7,0,8,0,9];\\ [\omega]_0 &= [1,2,3,4,5,6,7,8,9].\end{aligned}$$

Remark 3.10. We notice that existence of limit (3.3) is not guaranteed as explained in [7, Rmk. 2.4].

In this new framework, it is possible to give the definition of fd-numbers as follows.

Definition 3.11. A Borel number ω is said to be a *fd-number* if the equality

$$\hat{\mathbb{P}}([s]) = \log\left(1+\frac{1}{s}\right)$$

holds for any $s \in \mathfrak{N}_0$.

This probabilistic construction, strongly related to Def. 3.8, could be used to build a class of numbers starting from the more general Benford law. However, one would consider a very small class of numbers with no *zeros* in their decimal expansion, to generalize Def. 3.11 in the most natural way. A more "inclusive" probabilistic interpretation represents a stimulating challenge that could be tackled in the near future.

Remark 3.12. Ergodic theory can be used to show that specific sequences are Benford.
In [8, Ex. 6.2.2], the author shows that the first digits sequence

$$\{1,2,4,8,1,3,\dots\},$$

obtained by considering only the first digit of each number in the sequence

$$\{2^n \mid n \ge 0\}, \tag{3.4}$$

satisfies condition (2.1). For each $k = 1,2,\dots,9$ we can define an interval $J_k := [\log k, \log(k+1)) \subset [0,1)$. The result is obtained by exploiting the unique ergodicity of a specific irrational rotation, T_θ with $\theta = \log 2$, on $[0,1)$.
This approach can clearly be generalized to show that any sequence obtained by taking the first digits of $\{k^n \mid n \ge 0\}$, for any $k \in \mathbb{N}\setminus\{10^m \mid m = 0,1,\dots\}$, is actually Benford; the correct irrational rotation is the one with $\theta = \log k$.
Moreover, the procedure can be generalized to obtain a result related to Prop. 2.6. We show a sketch of the procedure for the sequence (3.4). We want to quantify the frequency of numbers whose first digit is d_1, second digit is d_2, …, m-th digit is d_m. A power of 2^n satisfies this requirement if, for some $r \in \mathbb{N}$,

$$d_1 10^r + d_2 10^{r-1} + \cdots + d_m 10^{r-m+1} \le 2^n < d_1 10^r + d_2 10^{r-1} + \cdots + (d_m+1)10^{r-m+1}.$$

We take the log of the inequality, and we factor out 10^{r-m+1}.
Then, the size of the interval we are considering is

$$\log\left(\frac{10^{r-m+1}(d_1 10^{m-1} + d_2 10^{m-2} + \cdots + d_m + 1)}{10^{r-m+1}(d_1 10^{m-1} + d_2 10^{m-2} + \cdots + d_m)}\right) = \log\left(1 + \frac{1}{d_1 10^{m-1} + d_2 10^{m-2} + \cdots + d_m}\right),$$

which is precisely the desired formula. The same application of the Ergodic Theorem used in [8, Ex. 6.2.2] allows us to conclude.

3.2 Relation with normal numbers

A real number is called *normal* (in its base-ten representation) if every finite sequence of digits is uniformly distributed. In this brief subsection, we present an interesting link between this particular class of numbers and the Benford sequences, which represents a starting point for the an investigation in such a direction.

Firstly, it seems appropriate to point out the connection between normal number and sequences which are uniformly distributed modulo 1 (see, e.g, [12, Thm. 8.15]), as we do in the next theorem.

Theorem 3.13. *A real number ω is normal to base 10 if and only if the sequence $(10^{n-1}\omega) = (\omega, 10\omega, 10^2\omega, \dots)$ is uniformly distributed modulo 1.*

Lemma 3.14. *The sequence $(x_n) = \left(10^{10^{n-1}\omega}\right)$ is Benford if and only if ω is 10-normal.*

Proof. From Thm. 2.9, we have that (x_n) is Benford if and only if $(\log|x_n|)$ is uniformly distributed modulo 1. Thus, we immediately obtain that $(10^n\omega)$ has to be uniformly distributed modulo 1 and, by Thm. 3.13, this last requirement holds if and only if ω is 10-normal. □

Remark 3.15. Since almost all real number are normal, we have that almost all sequences of the form $(10^{10^{n-1}\omega})$ are Benford.

Remark 3.16. It is important to highlight that no fd-number is also a normal number (this is trivial to check). Thus, for the same reason as in Rmk. 3.15, the class of fd-numbers is smaller than the class of normal numbers.

4 Conclusions

In this paper, we have presented a number-theoretical approach to the study of Benford's law. In particular, we analyze three possible direction of investigation. The first is based on the definition of a new class of numbers, the fd-numbers, which are based on the classical first digit law. Then, we deepened the relation between Benford sequences and normal numbers. Finally, we give an interesting insight related to the ergodic theory, which may lead to fascinating development.

The work contained on this paper inspires many other issues, such as the properties of fd-numbers or additional links between Benford's sequences and particular classes of number (as in the case of normal numbers). We leave these, and other promising leads on Benford's law and number theory, as outlook for future research.

Acknowledgments. The authors would like to thank Marco Capolli and Daniele Taufer for the fruitful exchanges during the writing of this paper.

References

[1] F. Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572, 1938.

[2] A. Berger and T. P. Hill. Benford's law strikes back: No simple explanation in sight for mathematical gem. *The Mathematical Intelligencer*, 33:85–91, 2011.

[3] A. Berger and T. P. Hill. *An Introduction to Benford's Law*. Princeton University Press, 2015.

[4] A. Berger and T. P. Hill. The mathematics of Benford's law: a primer. *Stat. Methods Appl.*, 2020.

[5] A. Berger, T. P. Hill, and E. E. Rogers. Benford online bibliography.

[6] M. É. Borel. Les probabilités dénombrables et leurs applications arithméti- ques. *Rendiconti del Circolo Mat. di Palermo*, 27(1):247–271, 1909.

[7] N. Cangiotti and D. Taufer. Normal and pseudonormal numbers, 2021.

[8] K. Dajani. Introduction to ergodic theory and its applications to number theory, 2014.

[9] S. Miller. *Benford's Law: Theory and Applications*. Princeton University Press, 2015.

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

[11] M. J. Nigrini. *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*. John Wiley & Sons, 2012.

[12] I. Niven. *Irrational Numbers*. Carus Mathematical Monographs. Mathematical Association of America, 1956.

[13] H. Weyl. über die gibbs'sche erscheinung und verwandte konvergenzphäno- mene. *Rendiconti del Circolo Matematico di Palermo*, 330:377–407, 1910.

Author information

Nicolò Cangiotti, Polytechnic University of Milan, Department of Mathematics, via Bonardi 9, 20133, Milan, Italy.
E-mail: nicolo.cangiotti@polimi.it

Mattia Sensi, Delft University of Technology, Network Architectures and Services Group, Mekelweg 4, 2628CD, Delft, The Netherlands.
E-mail: m.sensi@tudelft.nl

Received: June 5, 2021
Accepted: June 22, 2021