Research Article

n-Digit Benford Converges to Benford

Azar Khosravani and Constantin Rasinariu

Department of Science and Mathematics, Columbia College Chicago, Chicago, IL 60605, USA

Correspondence should be addressed to Constantin Rasinariu; crasinariu@colum.edu

Using the sum invariance property of Benford random variables, we prove that an n-digit Benford variable converges to a Benford variable as n approaches infinity.

1. Introduction

Given a positive real number y, and a positive integer i, we define $D_i(y)$ as the ith significant digit of y, where $D_1 : \mathbb{R}^+ \to \{1, \ldots, 9\}$ and $D_i : \mathbb{R}^+ \to \{0, 1, \ldots, 9\}$ for $i > 1$. Thus, $D_1(2.718) = 2$ and $D_3(2.718) = 1$. We assume base 10 throughout this paper.

Let $\mathscr{A}$ be the smallest sigma algebra generated by D_i. Then $D_i^{-1}(d) \in \mathscr{A}$ for all i and d. Within this framework, a random variable Y is Benford [1–3] if, for all $m \in \mathbb{N}$, $d_1 \in \{1, \ldots, 9\}$, and $d_i \in \{0, 1, \ldots, 9\}$ for $i > 1$ the probability that the first m digits of a real number are $d_1 d_2 \cdots d_m$ is given by

$$
\begin{aligned}
&P\left(D_1(Y) = d_1, \ldots, D_m(Y) = d_m\right) \\
&= \log\left(1 + \left(\sum_{j=1}^{m} 10^{m-j} d_j\right)^{-1}\right).
\end{aligned}
\tag{1}
$$

While Benford variables have logarithmic distributions in all of their digits, often times, in Benford literature the focus has only been on the distribution of the first digit. Such limitation may obscure the true nature of the quantity investigated. There are datasets which exhibit a perfect "Benford" distribution in the first digit but fail to do so in the second. Nigrini [4] provided such an example and consequently recommended the use of the first two-digit test in order to improve the recognition of the Benford datasets

and thus to identify financial fraud. He also recommended this approach for other accounting related analyses.

Such cases were generalized in [5], where a new class of random variables, called *n-digit Benford variables*, was introduced. These variables exhibit a logarithmic digit distribution only in their first n digits but are not guaranteed to be logarithmically distributed beyond the nth digit. Unlike Benford variables whose decimal logarithm is uniformly distributed mod 1, the decimal logarithm of n-digit Benford random variables has less stringent constraints; it must only satisfy prescribed areas over a given partition of the unit interval. This provides us with a collection of random variables that contains the Benford variables as a subset.

It is intuitive to assume that when n goes to infinity, an n-digit Benford variable converges to Benford. The purpose of this paper is to prove that this is indeed the case.

This paper is structured as follows: in the next section we introduce n-digit Benford variables together with some of their properties. In Section 3 we briefly discuss sum invariance, which is fundamental for our main result. Finally, using sum invariance, in Section 4 we show that n-digit Benford variable converges to Benford, as $n \to \infty$.

2. n-Digit Benford

An n-digit Benford random variable behaves as a Benford variable only in the first n-digits but may not have a logarithmic digit distribution beyond nth digit [5].

TABLE 1: Sum invariance illustration for the first 50000 Fibonacci numbers.

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
21714.0	21712.2	21717.8	21707.4	21713.2	21725.0	21702.7	21717.4	21715.5

Definition 1. Let $n \in \mathbb{N}$. A random variable Y is n-digit Benford if for all $d_1 \in \{1, \ldots, 9\}$ and all $d_i \in \{0, 1, \ldots, 9\}$, for $2 \le i \le n$,

$$P\left(D_1\left(Y\right) = d_1, \ldots, D_n\left(Y\right) = d_n\right)$$
$$= \log\left(1 + \left(\sum_{j=1}^{n} 10^{n-j} d_j\right)^{-1}\right). \tag{2}$$

Note that a Benford variable is n-digit Benford variable, for any n.

Lemma 2. *If Y is n-digit Benford, then it is k-digit Benford, for all $1 \le k < n$.*

Proof. Let $k = n - 1$. Then, by (2)

$$P\left(D_1\left(Y\right) = d_1, \ldots, D_{n-1}\left(Y\right) = d_{n-1}\right) = \sum_{d_n=0}^{9} P\left(D_1\left(Y\right)\right.$$

$$= d_1, \ldots, D_{n-1}\left(Y\right) = d_{n-1}, D_n\left(Y\right) = d_n)$$

$$= \sum_{d_n=0}^{9} \log\left(1 + \left(\sum_{j=1}^{n} 10^{n-j} d_j\right)^{-1}\right)$$

$$= \log\left(\frac{10^{n-1}d_1 + \cdots + 10d_{n-1} + 1}{10^{n-1}d_1 + \cdots + 10d_{n-1}} \times \cdots\right.$$

$$\left. \times \frac{10^{n-1}d_1 + \cdots + 10d_{n-1} + 10}{10^{n-1}d_1 + \cdots + 10d_{n-1} + 9}\right) \tag{3}$$

$$= \log\left(\frac{10^{n-1}d_1 + \cdots + 10d_{n-1} + 10}{10^{n-1}d_1 + \cdots + 10d_{n-1}}\right) = \log\left(1\right.$$

$$\left. + \left(\sum_{j=1}^{n-1} 10^{n-j} d_j\right)^{-1}\right). \qquad \square$$

As an example, let us consider the 2-digit Benford variable Y with the probability density function given by

$$f\left(y\right)$$
$$= \begin{cases} \dfrac{\pi}{2y \ln 10} \sin\left(\pi \beta_{d_1 d_2}\left(y\right)\right), & d_1 + \dfrac{d_2}{10} \le y < d_1 + \dfrac{d_2 + 1}{10} \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $\beta_{d_1 d_2}(y) = (\log(10y/(10d_1 + d_2)))/(\log((10d_1 + d_2 + 1)/(10d_1 + d_2)))$. Its graph is illustrated in Figure 1. We can check that $P(D_1(Y) = d_1, D_2(Y) = d_2) = \log(1 + (10d_1 + d_2)^{-1})$. From Lemma 2 this is a 1-digit Benford variable as well. However, Y is not a 3-digit Benford variable, since, for
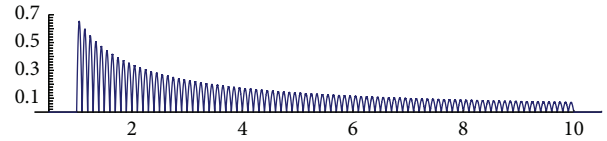

FIGURE 1: The pdf of a 2-digit Benford variable.

example, $P(D_1(Y) = 1, D_2(Y) = 1, D_3(Y) = 1) = 2.86 \times 10^{-3}$ instead of 3.89×10^{-3} as required by (2).

3. Sum Invariance

To define sum invariance, we first define the significand function, also known as the mantissa function.

Definition 3. The significand function $S : \mathbb{R}^+ \rightarrow [1, 10)$ is defined as

$$S\left(x\right) = 10^{\log x - \lfloor \log x \rfloor}, \tag{5}$$

where $\lfloor x \rfloor$ denotes the floor of x.

Let us consider a finite collection of positive real numbers K and define $S_{d_1 \cdots d_n}$ to be the sum of the significands of the numbers starting with the sequence of digits $d_1 \cdots d_n$. Sum invariance means that $S_{d_1 \cdots d_n}$ is digit independent. For instance, consider the Fibonacci sequence which is known to be Benford [6]. Then for the first 50000 Fibonacci numbers we obtain Table 1, where S_1 denotes the sum of all significands starting with 1, and so forth.

Nigrini was the first to notice sum invariance in some large collections of data [7]. Allaart [8] refined this concept, by defining it in connection with continuous random variables. Specifically, a distribution is sum invariant if the expected value of the significands of all entries starting with a fixed n-tuple of leading significant digits is the same as for any other n-tuple: $\mathbb{E}[S_{d_1 \cdots d_n} Y] = \mathbb{E}[S_{d_1' \cdots d_n'} Y]$. Allaart showed that a random variable is sum invariant if and only if it is Benford. Berger [3] proved that for sum invariant random variables

$$\mathbb{E}\left[S_{d_1 \cdots d_n} Y\right] = \frac{10^{1-n}}{\ln 10}. \tag{6}$$

For example, for a Benford sequence with 50000 elements, formula (6) yields $S_1 = \cdots = S_9 = 21714.7$ rounded to the tenths, which is very close to the actual values for the Fibonacci numbers illustrated in Table 1. Naturally, the more the numbers are taken from the sequence, the closer the one gets to the theoretical sum.

4. Main Result

A random variable is sum invariant if and only if it is Benford [3, 8]. Using this result, we will prove that an n-digit Benford variable converges to Benford as n approaches infinity by calculating the bounds for the expected value of its significand.

Given a function $g : \mathbb{R} \to \mathbb{R}$, we define $g^{\dagger} : \mathbb{R} \to [0, 1)$ as

$$
g^{\dagger}(x) = \begin{cases} \sum_{k \in \mathbb{Z}} g(x+k), & \forall x \in [0, 1), \\ 0, & \text{otherwise.} \end{cases} \tag{7}
$$

Lemma 4. *Let Y and $X = \log Y$ be two random variables with the probability density functions f and g, respectively. Then*

$$
\mathbb{E}\left[S_{d_1 \cdots d_m} Y\right] = \int_{\log(d_1 + \cdots + d_m/10^{m-1})}^{\log(d_1 + \cdots + (d_m+1)/10^{m-1})} 10^x g^{\dagger}(x)\, dx. \tag{8}
$$

Proof. Using $f(y) = g(\log y)/(y \ln 10)$, we get

$$
\begin{aligned}
\mathbb{E}\left[S_{d_1 \cdots d_m} Y\right] &= \int_{-\infty}^{\infty} S_{d_1 \cdots d_m}(y) f(y)\, dy \\
&= \sum_{k \in \mathbb{Z}} \int_{10^k(d_1 + \cdots + d_m/10^{m-1})}^{10^k(d_1 + \cdots + (d_m+1)/10^{m-1})} y 10^{-k} \frac{g(\log y)}{y \ln 10}\, dy \\
&= \int_{\log(d_1 + \cdots + d_m/10^{m-1})}^{\log(d_1 + \cdots + (d_m+1)/10^{m-1})} 10^x \sum_{k \in \mathbb{Z}} g(x+k)\, dx.
\end{aligned} \tag{9}
$$

$\square$

It is known that a necessary and sufficient condition for a random variable to be Benford is that $g^{\dagger} = 1$ [3, 9]. Consequently, (6) follows immediately from Lemma 4.

There are arbitrary many ways in which we can build an n-digit Benford variable. Let $\mathcal{B}_n$ be the infinite collection of all n-digit Benford variables. We use $\mathbb{E}[S_{d_1 \cdots d_n} \mathcal{B}_n]$ to denote the collection of the expected values of the significands of the elements of $\mathcal{B}_n$. The next theorem leads to the main result of our paper. It provides the bounds for the expected value $\mathbb{E}[S_{d_1 \cdots d_n} Y]$ for $Y \in \mathcal{B}_n$.

Theorem 5. *Let $Y \in \mathcal{B}_n$. Then*

$$
10^{1-n} \log\left(1 + \frac{1}{x_n}\right)^{x_n} \le \mathbb{E}\left[S_{d_1 \cdots d_n} Y\right]
$$
$$
\le 10^{1-n} \log\left(1 + \frac{1}{x_n}\right)^{x_n+1}, \tag{10}
$$

where $x_n = 10^{n-1} d_1 + \cdots + d_n$.

Proof. We will calculate the lower and upper bounds of $\mathbb{E}[S_{d_1 \cdots d_n} Y]$ using the fact that $\int_0^s g^{\dagger}(x) dx$ is monotonically increasing with s, where g is the probability density function of $\log Y$. From Lemma 4, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[S_{d_1 \cdots d_n} Y\right] \\
&= \left(d_1 + \cdots + \frac{d_n+1}{10^{n-1}}\right) \log\left(d_1 + \cdots + \frac{d_n+1}{10^{n-1}}\right) \\
&\quad - \left(d_1 + \cdots + \frac{d_n}{10^{n-1}}\right) \log\left(d_1 + \cdots + \frac{d_n}{10^{n-1}}\right) \\
&\quad - \int_{\log(d_1 + \cdots + d_n/10^{n-1})}^{\log(d_1 + \cdots + (d_n+1)/10^{n-1})} 10^s \ln 10 \int_0^s g^{\dagger}(x)\, dx\, ds.
\end{aligned} \tag{11}
$$

Since $Y \in \mathcal{B}_n$, we get

$$
\int_0^s g^{\dagger}(x)\, dx = \log\left(d_1 + \cdots + \frac{d_n}{10^{n-1}}\right) + \int_{\log(d_1 + \cdots + d_n/10^{n-1})}^{s} g^{\dagger}(x)\, dx. \tag{12}
$$

The second term in (12) can take any value between 0 and $\log(1 + 1/(10^{n-1}d_1 + \cdots + d_n))$, since $g^{\dagger}(x)$ is only constrained by its total area over the interval

$$
\left[\log\left(d_1 + \cdots + \frac{d_n}{10^{n-1}}\right), \log\left(d_1 + \cdots + \frac{d_n+1}{10^{n-1}}\right)\right]. \tag{13}
$$

It follows that

$$
10^{1-n} \log\left(1 + \frac{1}{x_n}\right)^{x_n} \le \mathbb{E}\left[S_{d_1 \cdots d_n} Y\right], \quad \forall Y \in \mathcal{B}_n, \tag{14}
$$

where $x_n = 10^{n-1} d_1 + \cdots + d_n$. Similarly we obtain

$$
\begin{aligned}
\mathbb{E}\left[S_{d_1 \cdots d_n} Y\right] &\le 10^{1-n} \log\left(1 + \frac{1}{x_n}\right)^{x_n} \\
&\quad + 10^{1-n} \log\left(1 + \frac{1}{x_n}\right), \quad \forall Y \in \mathcal{B}_n,
\end{aligned} \tag{15}
$$

which completes the proof. $\square$

As $n \to \infty$, both lower and upper bounds of $\mathbb{E}[S_{d_1 \cdots d_n} \mathcal{B}_n]$ approach $10^{1-n}/\ln 10$, proving the sum invariance [3]. Consequently, n-digit Benford variable converges to Benford.
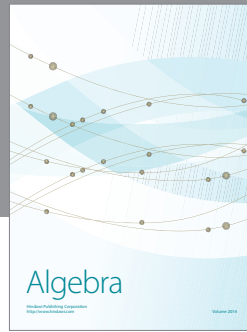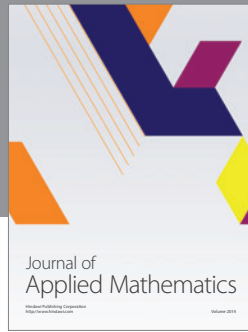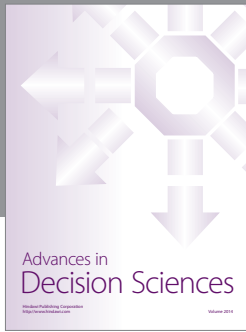
Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] F. Benford, "The law of anomalous numbers," *Proceedings of the American Philosophical Society*, vol. 78, no. 4, pp. 551–572, 1938.

[2] T. P. Hill, "Base-invariance implies Benford's law," *Proceedings of the American Mathematical Society*, vol. 123, no. 3, pp. 887–895, 1995.

[3] A. Berger and T. P. Hill, "A basic theory of Benford's law," *Probability Surveys*, vol. 8, pp. 1–126, 2011.

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

[5] A. Khosravani and C. Rasinariu, "n-digit benford distributed random variables," *Advances and Applications in Statistics*, vol. 36, no. 2, pp. 119–130, 2013.

[6] R. L. Duncan, "An application of uniform distribution to the Fibonacci numbers," *The Fibonacci Quarterly*, vol. 5, pp. 137–140, 1967.

[7] M. J. Nigrini, *The detection of income tax evasion through an analysis of digital frequencies [Ph.D. thesis]*, University of Cincinnati, Cincinnati, Ohio, USA, 1992.

[8] P. C. Allaart, "An invariant-sum characterization of Benford's law," *Journal of Applied Probability*, vol. 34, no. 1, pp. 288–291, 1997.

[9] P. Diaconis, "The distribution of leading digits and uniform distribution mod 1," *The Annals of Probability*, vol. 5, no. 1, pp. 72–81, 1977.