An Invariant-Sum Characterization of Benford's Law

Pieter C. Allaart

Vrije Universiteit Amsterdam[1]

Abstract

The accountant Nigrini remarked that in tables of data distributed according to Benford's Law, the sum of all elements with first digit d $(d = 1, 2, .., 9)$ is approximately constant. In this note, a mathematical formulation of Nigrini's observation is given and it is shown that Benford's Law is the unique probability distribution such that the expected sum of all elements with first digits $d_1, .., d_k$ is constant for every fixed k.

Keywords and phrases: First significant digit, Benford's law, mantissa function, sum-invariance.

AMS 1990 subject classification: 60A10.

1 Introduction

The main goal of this article is to give a mathematical proof of an empirical observation of the accountant M. Nigrini. In his Ph.D. thesis (1992), Nigrini observed that tables of unmanipulated accounting data closely follow Benford's Law (see §2 below), and that in sufficiently long lists of data for which Benford's Law holds,

> *the sum of all entries with leading digit d is constant for various d.*

(cf. Nigrini, 1992, pp. 70/71).

This paper introduces a natural extension of the above observation to constancy of sums of all k-tuples of leading digits (called *sum-invariance* below), and the main result (Theorem 4.1 below) establishes both the corresponding generalization of Nigrini's observation and its converse:

> *A distribution is sum-invariant if and only if it is the Benford distribution ((2) below).*

[1] Author's address: Department of Mathematics and Computer Science, De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands.

2 Benford's Law

Benford's Law is an empirical law saying that in tables of physical constants and statistical data the first significant digit is distributed not uniformly, but logarithmically, i.e.

$$\text{Prob}(\text{first significant digit} = d) = \log_{10}(1 + d^{-1}), \qquad d = 1, 2, .., 9 \tag{1}$$

In its more general form, Benford's Law is a statement about mantissa distributions:

$$\text{Prob}(\text{mantissa} < x) = \log_{10} x, \qquad x \in [1, 10), \tag{2}$$

where the mantissa of a real number x is the number obtained from x by shifting the decimal point to the place immediately after the first significant (non-zero) digit. It is easily seen that (2) implies (1).

For an historical survey of Benford's Law, see for example Raimi (1976) or Schatte (1988). Hill (1994) has a brief discussion of attempts to explain the empirically induced law, and adds to them a new explanation assuming base-invariance. Interesting applications of Benford's Law can be found, among others, in Hamming (1960), Varian (1972) and Nigrini (1992).

3 Sum-invariance

While Nigrini states his findings in a number-theoretic setting, in this note a precise *probabilistic* formulation of his observation is given.

In order to arrive at a suitable formulation, the following three points are essential: First, observe that it is the mantissae of the numbers in the tables, not the numbers themselves, which are to be added. (Otherwise, for example, a single astronomically large number in a table would dominate all other sums; adding numbers of different orders of magnitude does not seem to lead to any meaningful conclusion).

Second, the word 'constant' in Nigrini's statement is translated to be 'constant in expectation'. One reason is that for any finite random sample from Benford's distribution, the sums are almost surely *not* constant. And demanding equality in distribution is far too much: it can be seen that, in case of the Benford distribution (2), assuming independent entries, the nine sums have different second moments. The first moment, however, suits the problem perfectly well, as will be made clear.

Finally, to establish uniqueness, it is necessary to consider also second and third significant digits, and so on. For example, the (expected) sum of all entries starting with 1.2 is equal to the sum of all entries starting with 7.4, the sum of entries starting with 2.7182 equals that of entries starting with 3.1415, etcetera.

With these points in mind, sum-invariance can be defined informally as

> *A distribution is* sum-invariant *if for any natural number k, the expected sum of the mantissae of all entries starting with a fixed k-tuple of leading significant digits is the same as that for any other k-tuple.*

To formalize this definition, the following preliminaries are needed. Let $\mathbb{R}^+$ denote the positive real numbers $(0,\infty)$, $\mathbb{Z}$ the integers and $\mathbb{N}$ the natural numbers; $\mathcal{B}$ the Borel σ-algebra on $\mathbb{R}^+$ and $\mathcal{B}(A)$ the Borel subsets of A. Let $\uplus$ signify union of disjoint sets. For $\mathrm{E} \subset \mathbb{R}$ and $a \in \mathbb{R}$, aE is the set $\{ae : e \in E\}$; and for a random variable X, $\mathbb{E}X$ is the expectation of X.

In what follows, only the familiar decimal case (base 10) will be considered. However, the base value is not essential and all results and definitions carry over easily to other bases.

Definition 3.1 The *mantissa function* M is the function $M : \mathbb{R}^+ \to [1,10)$ such that $M(x) = r$, where r is the unique number in $[1,10)$ with $x = r \cdot 10^n$ for some $n \in \mathbb{Z}$. For example, $M(9) = M(0.09) = M(90) = 9$.

Definition 3.2 For $k \in \mathbb{N}$, $d_1 \in \{1,...,9\}$ and $d_2,...,d_k \in \{0,1,...,9\}$, $A(d_1,...,d_k)$ is the set of all positive real numbers whose first k significant digits are $d_1,...,d_k$, respectively, and $\bar{A}(d_1,...,d_k)$ is the restriction of this set to $[1,10)$.

The next definition is convenient to reduce the problem to measures on $[1,10)$.

Definition 3.3 For a probability measure P on $(\mathbb{R}^+,\mathcal{B})$, its corresponding *mantissa distribution* is defined to be the measure P_M on $\mathcal{B}([1,10))$ given by

$$P_M(E) = P(\biguplus_{n\in\mathbb{Z}} 10^n E) \tag{3}$$

In other words, if P is the distribution of a random variable X, then P_M is the distribution of its mantissa $M(X)$.

Example 3.4 Suppose that P is the uniform distribution on $(0,1)$. Then by (3), for $x \in [1,10)$, $P_M([1,x)) = \sum_{n\in\mathbb{Z}} P([10^n,10^n x)) = \sum_{n=1}^{\infty} P([10^{-n},10^{-n}x)) = \sum_{n=1}^{\infty} 10^{-n}(x-1) = \frac{1}{9}(x-1)$. In this case P_M is the uniform distribution on $[1,10)$, which has its first significant digit uniformly distributed on the integers $1,2,..,9$, and therefore clearly does not satisfy Benford's Law.

Example 3.5 For $m \in \mathbb{N}$, let P be the distribution with probability density function g_m on $\mathbb{R}^+$ given by

$$g_m(x) = \begin{cases} (2m\ln(10)\cdot x)^{-1} & \text{if } x \in [10^{-m},10^m) \\ 0 & \text{otherwise} \end{cases} .$$

Then, using a calculation as above, it follows that $P_M([1,x)) = \log_{10} x, 1 \le x < 10$, so P satisfies Benford's Law for every $m \in \mathbb{N}$.

The following definition is the formal restatement of (2).

Definition 3.6 P_M is called *Benford's Law* if it satisfies

$$P_M([1,x)) = \log_{10} x, \qquad 1 \le x < 10. \tag{4}$$

After these preparations, a formal definition of sum-invariance can now be given.

Definition 3.7 A probability measure P on $(\mathbb{R}^+, \mathcal{B})$ is said to be *sum-invariant*, if for any random variable X with distribution P, the expectations

$$\mathbb{E}\left[\mathrm{M(X)1_{A(d_1,...,d_k)}(X)}\right], \qquad d_1 \in \{1, ..., 9\}, d_2, ..., d_k \in \{0, 1, ..., 9\} \tag{5}$$

are constant for every fixed $k \in \mathbb{N}$.

4 The main theorem

The following theorem is the main result of this article.

Theorem 4.1 *A probability measure P on $(\mathbb{R}^+, \mathcal{B})$ is sum-invariant if and only if its corresponding mantissa distribution P_M is Benford's Law* (4)*.*

Corollary 4.2 *Let $X_1, X_2, ..., X_n$ be random variables with a common distribution P. Then the expected sums*

$$\mathbb{E}\left[\sum\{\mathrm{M(X_i) : X_i \in A(d_1, ..., d_k)}\}\right], \ \mathrm{d_1 \in \{1, .., 9\}, d_2, .., d_k \in \{0, 1, .., 9\}} \tag{6}$$

are constant for every fixed $k \in \mathbb{N}$ if and only if P_M is Benford's Law (4)*.*

Proof: observe that the expression in (6) equals

$$\sum_{i=1}^{n} \mathbb{E}\left[\mathrm{M(X_i)1_{A(d_1,...,d_k)}(X_i)}\right]$$

and apply Theorem 4.1. □

Proof of Theorem 4.1:

It is easy to check that a Borel probability measure P on $\mathbb{R}^+$ is sum-invariant if and only if P_M satisfies

$$\int_A x dP_M(x) = \frac{1}{9}\lambda(A)\int_{[1,10)} x dP_M(x) \tag{7}$$

for all A of the form $\bar{A}(d_1, ..., d_k)$. Here λ denotes Lebesgue measure on $[1, 10)$.

That P_M in (4) satisfies (7) is an easy substitution. Conversely, suppose that (7) holds for all $A = \bar{A}(d_1, ..., d_k)$. Using countable additivity and Carathéodory's extension theorem (cf. Royden, 1988, p.295), it follows that (7) holds for every Borel measurable A. In other words, λ is absolutely continuous with respect to P_M, with a strictly positive density proportional to x. This implies that, conversely, P_M is absolutely continuous with respect to λ with density proportional to $1/x$. □

Remark 4.3 The essential feature of a probability distribution used here seems to be its density function (when continuous). Since for fixed $k \in \mathbb{N}$ the intervals $\bar{A}(d_1, ..., d_k)$ have constant length, the integrals in (7) are constant only if, after substitution, a constant function is integrated. This means that the density must cancel the multiplying factor x, and therefore can only be $(x \ln 10)^{-1}$.

Acknowledgement

The author wishes to thank Prof. T.P. Hill for his extensive support and several useful conversations; Prof. P.J. Holewijn for some useful suggestions and corrections; Professors M.J. Nigrini and P. Schatte for their interest and for sending some references; and the referee for his suggestions, which led to a considerably shorter proof of Theorem 4.1.

References

Hamming, R.W. (1970), On the distribution of numbers. *Bell System Tech. J.* **49**, 1609-1625.

Schatte, P. (1988), On mantissa distributions in computing and Benford's Law. *J. Inf. Process. Cybern.* EIK **24**, 443-455.

Raimi, R. (1976), The first digit problem. *Amer. Math. Monthly* **83**, 521-538.

Hill, T.P. (1994), Base-invariance implies Benford's Law. *Proc. Amer. Math. Soc.* **123**, 887-895.

Varian, H. (1972), Benford's Law. *Amer. Statistician* **26**, 65-66.

Nigrini, M. (1992) *The detection of income evasion through an analysis of digital distributions.* Ph.D. Thesis, Department of Accounting, University of Cincinatti.

Royden, H.L. (1988) *Real Analysis* (MacMillan, New York, 3rd ed.)