PHYSA: 17247 Model 3Gsc pp. 1–7 (col. fig: NIL)

ARTICLE IN PRESS

Physica A xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Q1 Revisiting the Benford law: When the Benford-like distribution of leading digits in sets of numerical data is expectable?

Q2 G. Whyman [a], N. Ohtori [b], E. Shulzinger [a], Ed. Bormashenko [a,*]

[a] Ariel University, Faculty of Natural Sciences, Department of Physics, Ariel, P.O.B.3, 407000, Israel

[b] Department of Chemistry, Niigata University, 8050 Ikarashi 2-no-cho, Nishi-ku, Niigata 950-2181, Japan

HIGHLIGHTS

- The Benford distribution is a particular case of a more general mathematical statement.
- The exact Benford distribution is valid for the exponential function only.
- Explicit expressions for frequencies are obtained for the power, logarithmic, and tangent functions.
- The kinematic experiment illustrating proposed distributions is reported.
- Analysis of leading digits frequencies in thermal conductivities of liquids data is presented.

ARTICLE INFO

Article history:
Received 12 January 2016
Received in revised form 29 March 2016
Available online xxxx

Keywords:
Benford's law
Generalization
Exemplifications
Functional dependence between quantities

ABSTRACT

The Benford law states that the frequencies of decimal digits at the first place of numbers corresponding to various kinds of statistical or experimental data are not equal changing from 0.3 for 1 to 0.04 for 9. The corresponding frequencies' distribution is described by the logarithmic function. As is shown in the present article, the Benford distribution is a particular case of a more general mathematical statement. Namely, if a function describing the dependence between two measurable quantities has a positive second derivative, then the mentioned above frequencies decrease for digits from 1 to 9. The exact Benford distribution is valid for the exponential function only. Explicit expressions for frequencies of leading digits are obtained and specified for the power, logarithmic, and tangent functions as examples. The kinematic experiment was performed to illustrate the above results. Also the tabulated data on thermal conductivities of liquids confirm the proposed formula for frequencies' distribution.

1. Introduction

Our research is devoted to phenomenological, contra-intuitive law, observed in a diversity of statistical data, also called the first digit law, first digit phenomenon, or leading digit phenomenon. Benford's law states that in listings, statistical, scientific and engineering data the leading digit 1 tends to occur with probability of 30%, much greater than the expected 11.1% (i.e., one digit out of 9). Benford's distribution was reported first in 1881, when the American astronomer Simon Newcomb noticed that in logarithm tables the earlier pages (which contained numbers that started with 1) were much

* Corresponding author.
E-mail address: edward@ariel.ac.il (Ed. Bormashenko).

http://dx.doi.org/10.1016/j.physa.2016.06.054
0378-4371/© 2016 Published by Elsevier B.V.

more worn and smudged than the later pages. Newcomb noted "that the ten digits do not occur with equal frequency must be evident to any making use of logarithmic tables, and noticing how much faster first pages wear out than the last ones" [1].

The phenomenon was re-discovered in 1938 by the physicist Frank Benford, who tested it on data extracted from 20 different domains, as different as the surface areas of rivers, physical constants, molecular weights, etc. [2]. Since that, the law is credited to Benford. The Benford law is expressed by the following statement: the occurrence of first significant digits in data sets follows a logarithmic distribution:

$$P(n) = \log_{10}\left(\frac{n+1}{n}\right) \quad (1)$$

where $P(n)$ is the probability of a number having the first nonzero digit n.

Since that, Benford's law was applied for the analysis of the statistical data related to a broad variety of statistical data, including atomic spectra [3], population dynamic [4], magnitude and depth of earthquakes [5], genomic data [6,7], mantissa distributions of pulsars [8], infrared spectra of polymers [9], statistical analysis of the turbulent eddy motion [10] and economic data [11,12]. Alternative to Benford distribution formulae have been proposed [13]. Distinguishing noise from chaos by the analysis of the first digits distribution was discussed in Ref. [14].

While Benford's law definitely applies to many situations in the real world, a satisfactory explanation has been given only recently through the works of Hill, who called the Benford distribution "the law of statistical folklore" and "the mathematical gem" [15–17]. Various sources of the Benford distribution were discussed in the literature: from the physical arguments, relating the origin of the Benford law [18] to the scaling invariance of physical systems up to the properties of the positional (place-value) notation accepted for representing various sets of data [19]. The possible thermodynamic structure of the Benford distribution was treated in Ref. [20].

Engel et al. demonstrated that the Benford law takes place approximatively for exponentially distributed numbers [21]. The breakdown of the Benford law was reported for certain sets of statistical data [9,22].

It should be mentioned that the grounding and applicability of the Benford law remain highly debatable [17]. In spite of this, the Benford law was effectively exploited for detecting fraud in accounting data [23,24]. Our paper is devoted to the revealing mathematical roots of the Benford law.

2. Mathematical roots of the Benford distribution

Consider two measurable quantities, x and $y = f(x)$, where $f(x)$ is a positive monotonic function of x with a positive second derivative $f''(x) > 0$ defined within some sufficiently large interval $x_{\min} < x < x_{\max}$. Let the values of x are measured with the equal probabilities. Then, the frequency $P(n)$ of decimal digits $n = 1, 2, 3, \ldots, 9$ appearing at the first decimal place of $f(x)$ values is a decreasing function of n. For the exponential function $f(x) = a^x$ the frequency $P(n)$ fulfill the Benford law (1) but for other functions satisfying the above conditions, these decreasing dependences are different.

Proof. The interval of y values $[1 \cdot 10^m, 1 \cdot 10^{m+1})$ where m is a positive or negative integer contains all of 9 digits at the first decimal place. The corresponding values of x belong to the interval $\left[f^{-1}\left(10^m\right), f^{-1}\left(10^{m+1}\right)\right)$ where $f^{-1}(y)$ is the inverse function of $f(x)$. The only restriction on m is $\left[f^{-1}\left(10^m\right), f^{-1}\left(10^{m+1}\right)\right) \subset [x_{\min}, x_{\max}]$. The probability distribution of the measured quantity x is a constant on the above assumption, thus the probability to measure the value x between values x_1 and x_2 may be written as $(x_1 < x_2)$

$$P(x_1, x_2) = \frac{x_2 - x_1}{\left|f^{-1}\left(10^{m+1}\right) - f^{-1}\left(10^m\right)\right|}. \quad (2)$$

Thus, the probability $P(n)$ to measure the value of the dependent quantity beginning with the digit n is

$$P(n) = \frac{f^{-1}(10^m \cdot (n+1)) - f^{-1}(10^m \cdot n)}{f^{-1}\left(10^{m+1}\right) - f^{-1}\left(10^m\right)}. \quad (3)$$

Obviously the probability $P(n)$ is normalized on the interval $\left[f^{-1}\left(10^m\right), f^{-1}\left(10^{m+1}\right)\right)$:

$$\sum_{n=1}^{9} P(n) = 1. \quad (4)$$

The relation (3) allows one to obtain the appropriate probabilities for particular functions satisfying the above assumptions (see below).

Now consider the continuous analog $P(y)$ of $P(n)$ in Eq. (3)

$$P(y) = \frac{f^{-1}(y+a) - f^{-1}(y)}{f^{-1}\left(10^{m+1}\right) - f^{-1}\left(10^m\right)}, \quad a > 0. \quad (5)$$

Please cite this article in press as: G. Whyman, et al., Revisiting the Benford law: When the Benford-like distribution of leading digits in sets of numerical data is expectable?, Physica A (2016), http://dx.doi.org/10.1016/j.physa.2016.06.054