J. Korean Math. Soc. **47** (2010), No. 5, pp. 1055–1075
DOI 10.4134/JKMS.2010.47.5.1055

ON SOME PROPERTIES OF BENFORD'S LAW

Dominik Strzałka

Abstract. In presented paper there were studied some properties of Benford's law. The existence of this law in not necessary large sets of numbers is a very interesting example that can show how the complex phenomena can appear in the positional number systems. Such systems seem to be very simple and intuitive and help us proceed with numbers. However, their simplicity in the case of usage in our lifetime is not necessary connected with the simplicity in the case of laws that govern them. Even if this laws indicate the existence of self-similar properties.

1. Introduction

As it is well known the Arabic number system is based on a decimal system, which has got ten digits. It is (probably) one of the most natural systems, which is a matter of fact that the human has got ten fingers and this for example facilitate the learning of counting. With the aid of ten digits one can show in an understandable manner the whole known numbers. There are also other number systems such as: binary, octal, hexadecimal and even Roman but they aren't so popular.

Without any effort one can notice that the numbers are everywhere: shares prices, stock markets, physical and mathematical constants, weather forecast, sport results, timetables, bills, shop prices, etc. Even more, this huge set is everyday build up because we measure, count, calculate. When we see the series of numbers, almost no one wonders is it matters what is the first digit in these numbers. Almost everyone will hazard a guess that as well the digit 1 and the digit 9 initiate the same amount of numbers, i.e., they appear in $\approx 11.111\%$ of cases (digit 0 initiate the numbers from interval $(-1, 1)$ but all numbers can be also shown in scientific notation, which means that they can have a non-zero first digit). Is it true?

Received January 28, 2009; Revised March 3, 2009.
2000 *Mathematics Subject Classification.* 11B99, 11Z05, 28A80.
Key words and phrases. Benford's law, self-similarity, complex systems.

2. General character of Benford's law

In 1881 an astronomer Simon Newcomb noticed a very interesting property. In one of his published articles [8] he stated that in the observed series of numbers a digit 1 appears as a first significant digit more frequently than other digits. In 1938 this observation was rediscovered by doctor Frank Benford, who worked as a physicist in general electric laboratories and frequently used the tables of logarithms [1]. He noticed that the pages in books with mathematical tables of logarithms with values starting from 1 where more dirtier than the others. Benford concluded that the scientists follow special preferences choosing logarithms with first digit 1. He analyzed a large set of data [1, 12] and it turned out that the digit 1 is the first significant digit in $\approx 30\%$ of cases. He also formulated a thesis that the probability of occurrence digit d as a first significant digit in any number equals [1]

$$P\left(d\right)=\log_{10}\left(1+\frac{1}{d}\right). \tag{1}$$

Because Benford rediscovered this phenomenon and gave it more formal character (1) nowadays we use the term Benford's law. One can quickly convince about this law analyzing a not necessary big set of data. Probably the fact that in Benford's law the logarithm appears is a little bit surprising but it shouldn't be, because our senses e.g. vision and hearing work similarly.

3. Why is that?

It should be noticed that the big set of numbers, which was previously mentioned, consists of numbers, which have got different units, such as: length, volume, mass, velocity, price, etc. Each of these numbers can be written in scientific notation, i.e., $a \cdot 10^b$. If one determines for them their leading digit can assume that the distribution of these digits is governed by the function $f\left(x\right)$. But this function should be independent on units. If someone will multiply all analyzed numbers by any positive constant this distribution should stay unchanged. For example, lets assume that each number will be multiplied by 3. The amount of numbers with leading digit 1 (i.e., from interval $[1.0, 1.999\cdots)$ multiplied by 10 to the some power) should be exactly the same like the amount of numbers with leading digit 3, 4 or 5 (i.e., from interval $[3.0, 5.999\cdots)$ multiplied by 10 to the some power), because it isn't matter what unit is used [3].

Having the function $f\left(x\right)$, which is the distribution of leading digits in numbers, one can calculate the cumulative distribution function

$$\int_a^b f\left(x\right)dx. \tag{2}$$

The equation (2) can be also a proportion of numbers that have the leading digits from interval $[a, b]$ (see Fig.1).

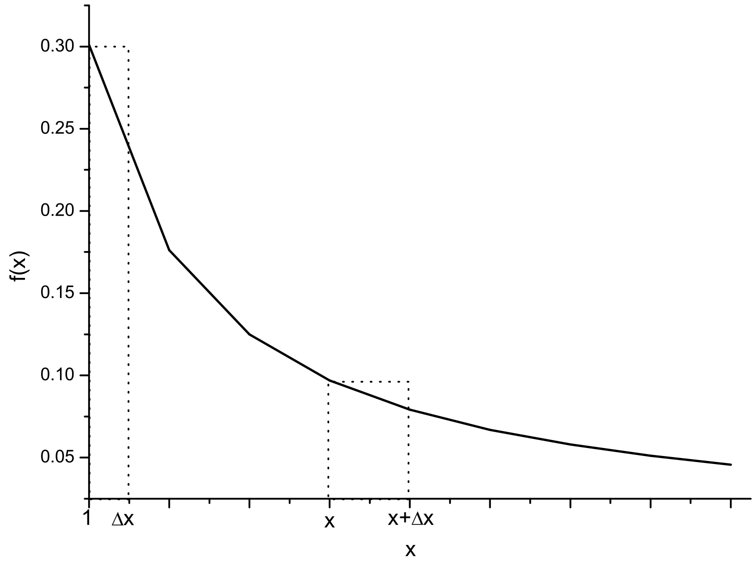

FIGURE 1. Theoretical shape of probability distribution of leading digits in system with base $B = 10$

For a very small increment Δx the value of f should fulfil

$$f(1) \cdot \Delta x = f(x) \cdot x\Delta x \tag{3}$$

because $f(1) \cdot \Delta x$ is a proportion of those numbers, which belong to the interval $[1, 1 + \Delta x)$, whereas $f(x) \cdot x\Delta x$ is a proportion of numbers that belong to the interval $[x, x + x\Delta x)$. The second set is the first set multiplied by x, thus the proportions should be equal (3), i.e.,

$$f(x) = \frac{f(1)}{x}. \tag{4}$$

The area under curve (1) should be equal 1 in limits $[1, 10)$, thus

$$\begin{aligned}
&\int_1^{10} f(x)\,dx = \int_1^{10} \frac{f(1)}{x} dx = f(1) \ln x|_1^{10} = 1 \\
&f(1) \ln 10 - f(1) \ln 1 = 1 \\
&f(1) = \frac{1}{\ln 10}
\end{aligned}$$

The proportion of numbers with leading digits from interval $[a, b]$ where $1 \leq a \leq b < 10$ from (2) and (4) equals

$$\int_a^b \frac{1}{x \ln 10} dx = \left. \frac{\ln x}{\ln 10} \right|_a^b = \frac{\ln \frac{b}{a}}{\ln 10}$$

because $\log_x y / \log_x z = \log_z y$, thus

$$\int_a^b \frac{1}{x \ln 10} dx = \log_{10} \frac{b}{a}. \tag{5}$$

From the equation (5) one can quickly compute the formula (1) assuming that for example if the probability of digit 1 occurrence is computed, the integral (5) is computed in limits $[1, 2)$, while for digit 2 in limits $[2, 3)$ thus $b = a + 1$ and the fraction b/a equals $1 + 1/a$ that is exactly like in (1).

4. What about the other number systems?

Benford's law can be used in any number system with the base B. In such a case the probability of occurrence as a first digit d (from interval $[1, B - 1]$) is given by [3]

$$P(d) = \frac{\log_{10} \left(1 + \frac{1}{d}\right)}{\log_{10} (B)}. \tag{6}$$

It should be emphasized that the main idea of Benford's law is a calculation of probability of digit d occurrence by the formula that uses the logarithm. No matter what kind of logarithm will be used – natural, decimal or with any other base, because

$$\frac{\log_a(b)}{\log_a(c)} = \frac{\frac{\log_d(b)}{\log_d(a)}}{\frac{\log_d(c)}{\log_d(a)}} = \frac{\log_d(b)}{\log_d(c)}.$$

Thus particularly the base of used logarithm can be equal B. Thanks this, the equation (6) can be rewritten to the following form

$$P(d) = \frac{\log_B \left(1 + \frac{1}{d}\right)}{\log_B (B)} = \log_B \left(1 + \frac{1}{d}\right). \tag{7}$$

Having the formula (7) it is possible to calculate the distributions of leading digits in any system with base B. The details for systems with $B \in [2, 10]$ are on Fig. 2.

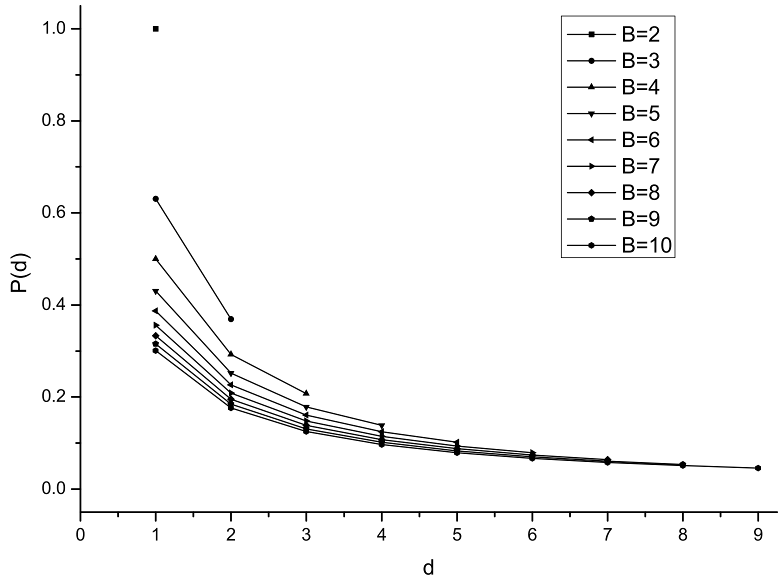

FIGURE 2. Benford's law in different systems

It can be also shown that the sum of all probabilities of first digit d occurrence in any number system with base B equals 1

$$\begin{aligned}\sum_{d=1}^{B-1} P(d) &= \frac{\ln\left(1+\frac{1}{1}\right)}{\ln(B)} + \frac{\ln\left(1+\frac{1}{2}\right)}{\ln(B)} + \cdots + \frac{\ln\left(1+\frac{1}{B-1}\right)}{\ln(B)} \\ &= \frac{\ln\left(\frac{2}{1}\cdot\frac{3}{2}\cdot\cdots\cdot\frac{B}{B-1}\right)}{\ln(B)} = \frac{\ln(B)}{\ln(B)} = 1.\end{aligned}$$

5. Second significant digit

The so far presented deliberations can lead to a simple question: is it possible to calculate such probabilities for second leading digits? To obtain this it should be noticed that for example the probability of occurrence as a second digit 7 is a sum of probabilities of occurrence of two first digits from the intervals: $[1.7, 1.8)$, $[2.7, 2.8)$, …, $[9.7, 9.8)$. In the first interval one obtains [4]

$$\frac{\log_{10}(1.8) - \log_{10}(1.7)}{\log_{10}(10) - \log_{10}(1)} = \frac{\log_{10}\left(\frac{18}{17}\right)}{\log_{10}(10)}$$

while in second interval it is

$$\frac{\log_{10}(2.8)-\log_{10}(2.7)}{\log_{10}(10)-\log_{10}(1)}=\frac{\log_{10}\left(\frac{28}{27}\right)}{\log_{10}(10)}.$$

Proceeding in a similar way it is possible to obtain the probabilities for successive intervals and the probability $P_1(d)$ of digit $d=7$ occurrence as a second significant digit after first non-zero digit should be

$$\begin{aligned} P_1(d=7) &= \frac{\log_{10}\left(\frac{18}{17}\right)}{\log_{10}(10)}+\frac{\log_{10}\left(\frac{28}{27}\right)}{\log_{10}(10)}+\cdots+\frac{\log_{10}\left(\frac{98}{97}\right)}{\log_{10}(10)} \\ &= \sum_{k=1}^{B-1}\frac{\log_{10}\left(\frac{kB+d+1}{kB+d}\right)}{\log_{10}(B)}. \end{aligned} \tag{8}$$

Remembering that: $\log_{10}(a)+\log_{10}(b)=\log_{10}(a\cdot b)$ it is possible to rewrite the equation (8) and obtain

$$P_1(d)=\sum_{k=1}^{B-1}\frac{\log_{10}\left(\frac{kB+d+1}{kB+d}\right)}{\log_{10}(B)}=\frac{1}{\log_{10}(B)}\log_{10}\left[\prod_{k=1}^{B-1}\left(1+\frac{1}{kB+d}\right)\right]. \tag{9}$$

The equation (8) allows computing the probability of occurrence of any digit at second position. In this case exists the possibility to calculate such a probability for digit 0, which in the case of formula (1) couldn't be taken into account because 0 as a first significant digit has no matter. It can be noticed that the probability for digit 0 at second position will be greater than the probability for any other digit, because $\frac{1}{kB+d}$ is the greatest when $d=0$. It suggests that the probability of digit 0 occurrence at other positions in numbers is also greater than the probability for rest of digits.

Let's consider for example the quadruple system. It comes from the formulas (6) and (8) that the probabilities of occurrence of the first and second leading digit d in numbers are following (Table 1):

TABLE 1. Benford's law probabilities in quadruple system

position	digit d			
n	0	1	2	3
0	0	0.5	0.292481	0.207519
1	0.303665	0.260976	0.229716	0.205643

It is clear that for second significant digits all probabilities except digit $d=0$ fall. Thus immediately a question arises: what is going on at the next positions?

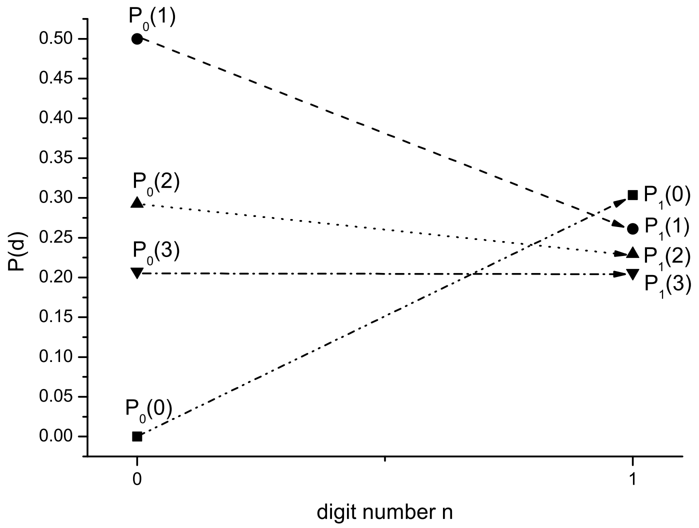

FIGURE 3. Changes in digits probabilities for first and second significant position in quadruple system according to Benford's law

5.1. Dependencies for other significant digits

It is possible to generalize the equation (8) to obtain a formula that guarantees the probabilities of appearance of n-th significant digit d after first non-zero digit. Such a generalization is given by [4]

$$P_n(d) = \frac{1}{\log_{10}(B)} \sum_{k=B^{n-1}}^{B^n-1} \log_{10}\left(1 + \frac{1}{kB+d}\right). \tag{10}$$

The equation (10) shows that the non-uniformity of digit distribution, if successive positions are considered, vanishes. For example in mentioned above quadruple system (i.e., $B = 4$) in the case of digit $d = 2$ one has got: $P_0(2) = 0.2924\cdots$, $P_1(2) = 0.2297\cdots$, $P_2(2) = 0.2454\cdots$, $P_3(2) = 0.2489\cdots$ and the probability of occurrence digit $d = 2$ quickly goes to 0.25 (i.e., $1/B$ as it should be expected). The dominance of digit $d = 0$, similarly like in the case of second significant digit, also exists (see Fig. 4).

6. Normalization of Benford's law

Benford's law has got many interesting properties. For example it can be normalized, i.e., all values of first digit probabilities in different number systems

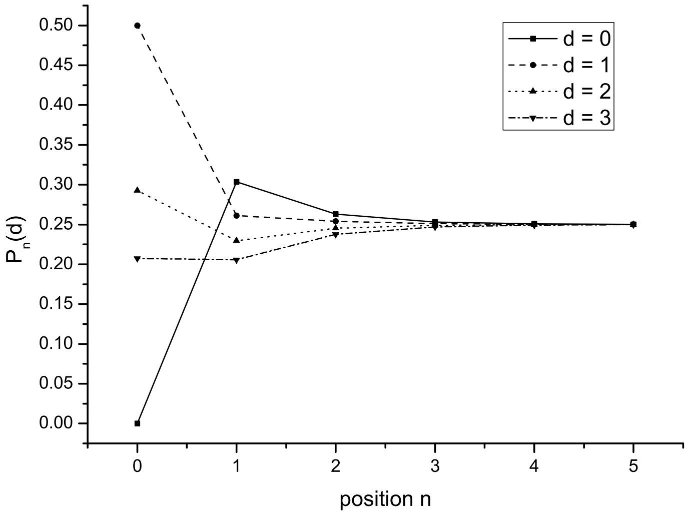

FIGURE 4. Changes in digits probabilities from first to sixth significant position according to Benford's law in quadruple system, $B = 4$

B will be divided by the probability of occurrence as a first digit 1. This will allow the calculation of participation of probabilities of other digits in comparison to the digit 1.

The calculations show that independently on assumed base B of number system the proportion of probabilities of first digits occurrence is always the same. Why is this? The answer is very simple and it comes from the observation that the probabilities for successive digits are always divided by the probability for digit $d = 1$, i.e.,

$$\frac{\log_{10}\left(1+\frac{1}{d}\right)}{\log_{10}\left(1+\frac{1}{1}\right)} = \frac{\log_{10}\left(1+\frac{1}{d}\right)}{\log_{10}(2)}$$

or equivalently taking into account (7)

$$\log_2\left(1+\frac{1}{d}\right)$$

and this relation is independent on base B thus it applies in all number systems. Such a phenomenon shows also that in the case of first significant digits Benford's law rules "in self".

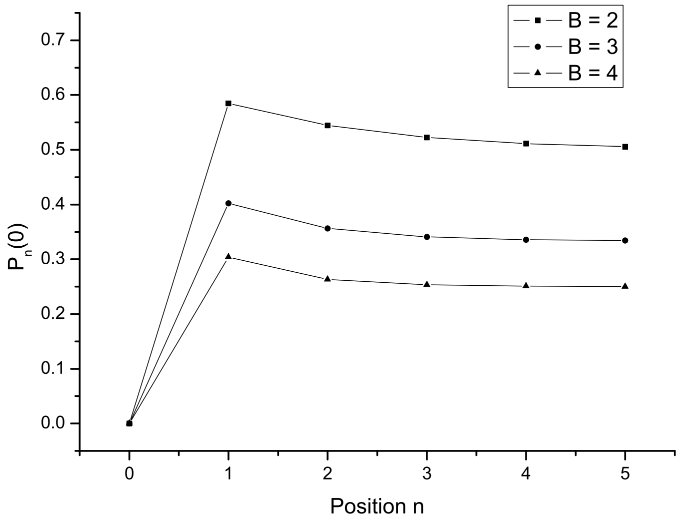

FIGURE 5. Changes in probabilities for digit $d = 0$ from first to sixth significant position according to Benford's law in three systems: binary, ternary and quadruple

7. Problem of digits dependence

Lets denote by D_i the function that shows the significant digit from position i in any number, i.e., $D_1(465) = 4$, $D_2(465) = 6$, $D_3(465) = 5$, etc. For any positive k, $d_1 \in \{1, 2, \ldots, 9\}$ and $d_j \in \{0, 1, \ldots, 9\}$, $j = 2, \ldots, k$ one has got (in decimal system) [4]

$$P\left(D_1 = d_1, \ldots, D_k = d_k\right) = \log_{10}\left[1 + \left(\sum_{i=1}^{k} d_i \cdot 10^{k-i}\right)^{-1}\right] \tag{11}$$

thus $P\left(D_1 = 4, D_2 = 6, D_3 = 5\right) = \log_{10}\left(1 + (465)^{-1}\right) \approx 9.32 \cdot 10^{-4}$.

From the equation (11) one can compute the probability of appearance of any mantissa with length k that consists of combination of digits according to system base B. The normalization condition for a given k is fulfilled [4]

$$\sum_{i=1}^{k} P\left(D_1 = d_1, \ldots, D_k = d_k\right) = 1.$$

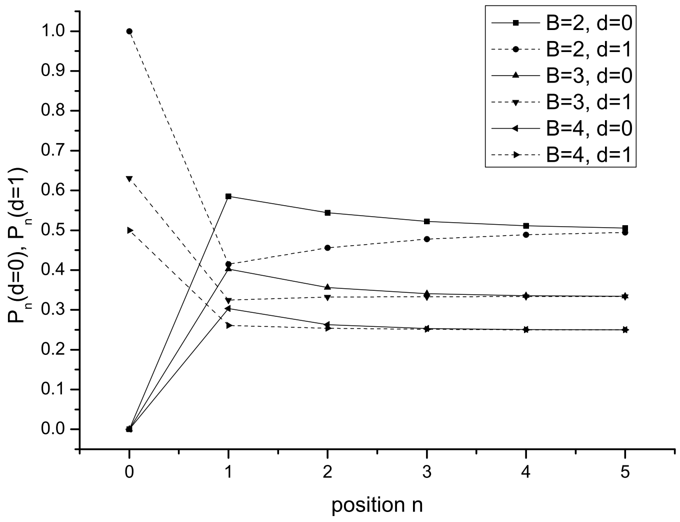

FIGURE 6. Changes in probabilities for digits $d = 0$ and $d = 1$ from first to sixth significant position for systems with base $B = 2, 3, 4$

This condition causes that the sum of probabilities of all possible combinations of digits for mantissas with length k is 1 (see Fig. 7 where the case of system with $B = 2$ is considered).

Having the equation (11) it is also possible to consider one more interesting feature in the case of Benford's law. The so far presented equations calculate the unconditional probabilities of any digit or mantissa appearance. For example if digit $d = 2$ after first non-zero digit will be considered in decimal system ($B = 10$) from (4) one will see that

$$P_1(2) = \sum_{k=1}^{10-1} \frac{\log_{10}\left(\frac{k\cdot 10+2+1}{k\cdot 10+2}\right)}{\log_{10}(10)} \approx 0.109.$$

But one can calculate the conditional probability of this digit appearance, i.e., the probability for this digit at second significant position having the knowledge about the probability for first digit. This problem can be solved taking into account the well known in probability calculus formula [10]

$$P(B|A) = \frac{P(AB)}{P(A)}, \tag{12}$$

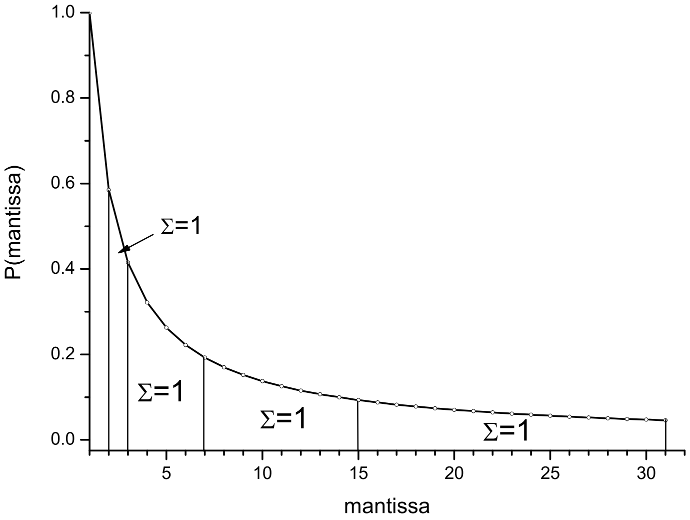

FIGURE 7. Sum of mantissa probabilities according to Benford's law for system with $B = 2$

where $P(A)$ denotes the probability of first significant digit appearance (from (1)), $P(AB)$ denotes the probability of mantissa occurrence that consists of first two digits (mantissa can have any length) in number (i.e., the first digit – an event A, the second digit – an event B), and $P(A|B)$ is the conditional probability of occurrence the event B after the event A. For example let's consider this problem when the first digit $d = 1$, i.e., the event A – the digit 1 at first significant position, the event B – the digit $d = 2$ at second significant position and the event AB – the mantissa 12. From (1) one will have that $P(A) \approx 0.301$, from (11) will see that $P(AB) \approx 0.0347$, thus from (12)

$$P(B|A) \approx \frac{0.0347}{0.301} \approx 0.115.$$

Two events are unconditional if $P(AB) = P(A) \cdot P(B)$, but if $P(B) = 0$, then another condition is considered, namely if $P(A|B) = P(B)$, then two events are independent. Because from (8) $P(B) = P_1(2) \approx 0.109$ and this isn't equal $P(A|B) \approx 0.115$ one has a very surprising conclusion: the digits that has appeared in numbers according to Benford's law aren't independent.

The equation (11) can be rewritten for any system with basis B. Then for any positive k, $d_1 \in \{1, 2, \ldots, B-1\}$ and $d_j \in \{0, 1, \ldots, B-1\}$, $j = 2, \ldots, k$

one has got

$$P\left(D_1 = d_1, \ldots, D_k = d_k\right) = \log_B\left(1 + \left(\sum_{i=1}^{k} d_i \cdot B^{k-i}\right)^{-1}\right). \tag{13}$$

These deliberations can be illustrated by a simple example – the case of the simplest binary system.

7.1. Probabilities for binary system

In each number system the occurrence of digit 0 as a first significant digit can be excluded due to the fact that each number can be represented by scientific notation, i.e., with first significant digit that is different than 0. Basing on Benford's law (1) it isn't also possible to calculate the probability for it. Thus the deliberations should be started form digit 1. In binary system its probability of occurrence basing on the equation (7) is equal 1 because this is the only one possible digit in this system that is different than 0. At second position the digits 0 and 1 can occur. The probability for $d = 0$ is $P_1(0) \approx 0.584962$, while for $d = 1$ it is $P_1(1) \approx 0.415037$. They are the unconditional probabilities. Taking into account the formulas (12) and (13) it should be assumed that: $P(A)$ is the probability of occurrence the digit 1 at first significant position (according to (7)), $P(AB)$ is the probability of occurrence mantissas 10 or 11 (according to (13)) and $P(A|B)$ is the conditional probability of occurrence 0 or 1 after first digit 1 according to (12).

It can be noted that in such a case the values of conditional probabilities will be exactly the same like the values for unconditional probabilities because the occurrence of digit 1 ($P(A)$ event) is a certain event. Because at second significant position there can appear digits 0 or 1 the deliberations can be carried two ways – the probability of occurrence mantissas 10 and 11 are: $P(10) = 0.584962\cdots$ and $P(11) = 0.415037\cdots$. When the third position will be considered the unconditional probabilities of occurrence digits 0 and 1 are respectively: $P_2(0) = 0.54432\cdots$ and $P_2(1) = 0.455679\cdots$. In the case of conditional probabilities after mantissa 10 can appear 0 or 1 (similarly in the case of mantissa 11). Thus there are possible four mantissas: 100, 101, 110, 111 that can appear with the following probabilities (from (13)): $P(100) = 0.3219\cdots$, $P(101) = 0.26303\cdots$, $P(110) = 0.22239\cdots$, $P(111) = 0.1926\cdots$. The conditional probabilities of occurrence 0 or 1 at third position should be considered in two cases: in the first one they will be preceded by 10 while in the second one by 11. In the case of predecessor 10 it will be (from (12) and (13)): for 0 – $P(0|10) = 0.55033\cdots$ and for 1 – $P(1|10) = 0.44966\cdots$; while for predecessor 11 it will be: for 0 – $P(0|11) = 0.53583\cdots$ and for 1 – $P(1|11) = 0.46414\cdots$. The sum of probabilities at both paths will be always equal 1.

If the following value: $P(10) \cdot P(0|10) + P(11) \cdot P(0|11)$ will be computed (from (13) and (12)) it will be equal $0.5432\cdots$ and it will be exactly the same

Table 2. Conditional and unconditional probabilities for binary system

P(mantissa)	mantissa		P_{cond}	
	DEC	BIN		
0	0	0	-	-
1	1	1	-	-
0.5849625	2	10	0.584962501	P(0|1)
0.4150375	3	11	0.415037499	P(1|1)
0.32192809	4	100	0.550339713	P(0|10)
0.26303441	5	101	0.449660287	P(1|10)
0.22239242	6	110	0.535836935	P(0|11)
0.19264508	7	111	0.464163065	P(1|11)
0.169925	8	1000	0.527835266	P(0|100)
0.15200309	9	1001	0.472164734	P(1|100)
0.13750352	10	1010	0.522758699	P(0|101)
0.12553088	11	1011	0.477241301	P(1|101)
0.11547722	12	1100	0.519249787	P(0|110)
0.1069152	13	1101	0.480750213	P(1|110)
0.09953567	14	1110	0.516679038	P(0|111)
0.0931094	15	1111	0.483320962	P(1|111)
0.08746284	16	10000	0.514714377	P(0|1000)
0.08246216	17	10001	0.485285623	P(1|1000)
0.07800251	18	10010	0.513163977	P(0|1001)
0.07400058	19	10011	0.486836023	P(1|1001)
0.07038933	20	10100	0.511909266	P(0|1010)
0.0671142	21	10101	0.488090734	P(1|1010)
0.06413034	22	10110	0.510872993	P(0|1011)
0.06140054	23	10111	0.489127007	P(1|1011)
0.05889369	24	11000	0.510002669	P(0|1100)
0.05658353	25	11001	0.489997331	P(1|1100)
0.05444778	26	11010	0.509261378	P(0|1101)
0.05246742	27	11011	0.490738622	P(1|1101)
0.05062607	28	11100	0.508622399	P(0|1110)
0.0489096	29	11101	0.491377601	P(1|1110)
0.04730571	30	11110	0.508065915	P(0|1111)
0.04580369	31	11111	0.491934085	P(1|1111)

like the probability of occurrence 0 at third significant position according to Benford's law from the equation (10). The same can be shown in the case of digit 1.

8. Self-similarity of Benford's law

Benoit Mandelbrot is usually considered as a father of fractal geometry. However it is worth to mention other people that had a connection with fractals: George Cantor, Giuseppe Peano, David Hilbert, Helge von Koch, Wacław Sierpiński, Gaston Julia or Felix Hausdorf. Their constructions (sets, curves, etc.) were for a long time considered as a mathematical "*monsters*" and "*oddities*" or were given as a counterexamples. In famous Mandelbrot's book "*Fractal geometry of nature*" [7] this interpretation was changed and now it is commonly known that these first fractals have got many connections with shapes that are in reality.

The self-similarity is a property that is connected with fractals. Probably this word doesn't require to give its interpretation, but the best explanation can be given by the example of cauliflower: it's small parts are similar to the whole. But the self-similarity property can be found not only in Nature but also in many things that are normally used. Such a property exist also in the case of decimal (but not only this) system. This feature is so natural that many people don't even know or notice this fact.

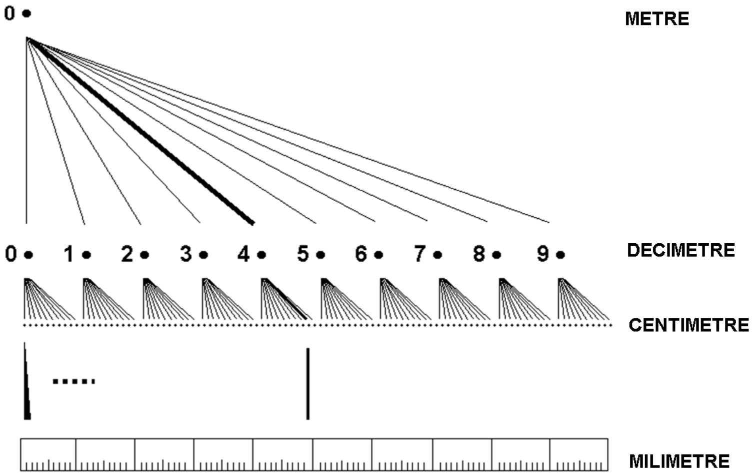

FIGURE 8. Self-similarity of decimal system [11]

To understand the self-similarity of decimal system it should be noted that the whole meter looks exactly the same like the decimeter, and the decimeter looks exactly the same like centimeter and so on [11]. The simple example can be given: 483 mm consists of 4 decimeters, 8 centimeters and 3 millimeters. The successive digits (Fig. 8) denote their position on a ruler and there is no need to count 483 millimeters; it is enough to find 4th decimeter then starting from it 8th centimeter and similarly 3rd millimeter. This is very simple and useful method and it can be compared with the problem of moving on decimal tree.

However, the simple explanation given above about self-similarity in positional systems doesn't give us any information about mechanisms that govern moving on this tree. One can ask for example about conditional and unconditional probabilities of choosing one or other ways in this tree or for example ask how is it related to Benford's law. In some sense this is possible when one notes that in previous section there were considered the conditional probabilities of occurrence the successive digits in mantissas for binary system. The conditional probabilities of occurrence digit 0 or 1 for successive positions go to 0.5 for $n \to \infty$ (see Fig. 9 and Tab. 2).

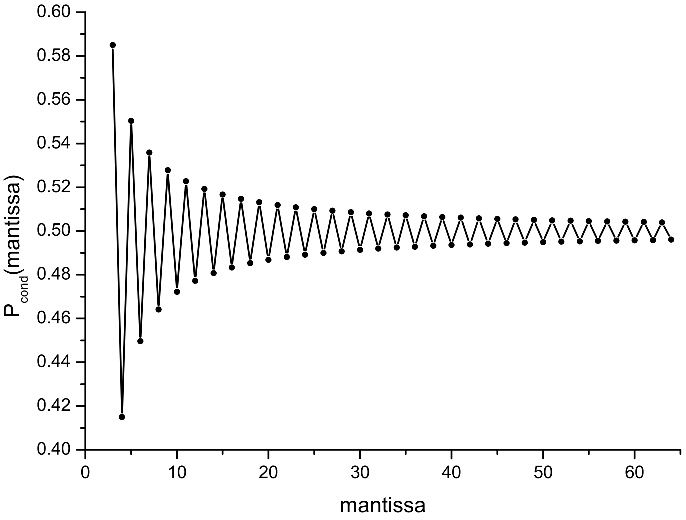

FIGURE 9. Conditional probabilities in system with base $B = 2$ according to Benford's law

But there can be also noted one more interesting thing. If for example the conditional probability of occurrence digit 0 after mantissa 101, i.e., $P(0|101)$ is analyzed it can be told that there is calculated the conditional probability of occurrence the number 10 after 5 (both values in decimal system). It comes from the fact that there is calculated the conditional probability of occurrence digit 0 after 101 (the binary representation of 5) in series 101**0** at fourth position and this series also denotes the number 10 in decimal system. But if the conditional probability of occurrence 1 after 101, i.e., $P(1|101)$ is calculated then according to what was written above the conditional probability of occurrence the number 11 after 5 is calculated. Thus in each case there is calculated the probability of occurrence the number $2x$ or $2x + 1$. As it can be seen the whole operation is in reality the shift of bits into the left and add 0 or 1. This allows making a connection between the equation (11) and the equation (12) to show the dependence between the unconditional and conditional probabilities in binary system (this can be done for other systems).

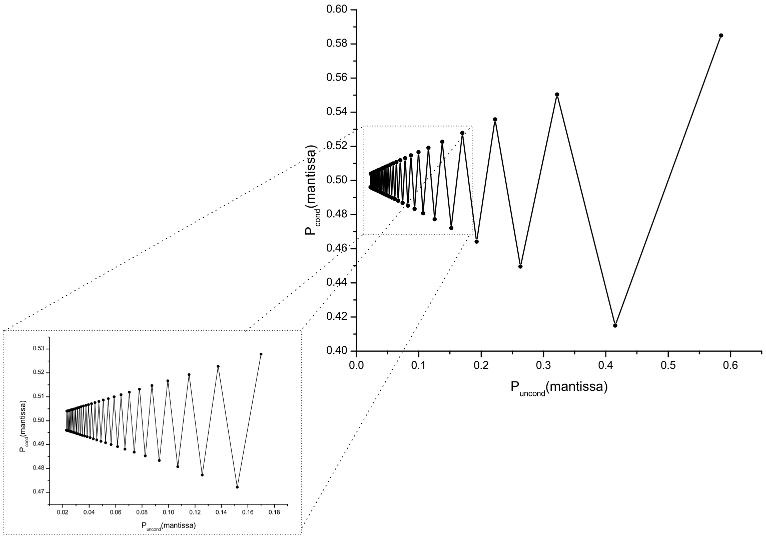

FIGURE 10. Conditional probabilities versus unconditional for binary system. The magnification shows the self-similarity property in Benford's law

From Fig. 10 it can be seen that the self-similarity property is visible. In this case one can reference this visualization to the examples of types of geometrical self-similarity given in Peitgen et al. book [11]. This is a some kind of similarity that can be related to the self-similarity at point.

9. Self-similar paths of unconditional probabilities – binary system

Taking into account the results that are presented in Fig. 10 it should be supposed that moving on "*tree*" (see Fig. 8) that is created by the self-similarity property of positional number systems in reality isn't *as simple as it seems to be.* If in set of numbers exists Benford's law this feature may influence the whole picture of system self-similarity because there is a question: is the moving on tree governed by any rules that take into account the conditional and unconditional probabilities of successive digits appearance? As it was written in Section 7 there is a problem of digits appearance dependence; generally according to Benford's law the probabilities of appearance of successive digits in numbers aren't independent. In Section 7.1 the case of the simplest system (binary one) was analyzed. This leads to the Fig. 10 but the whole problem can be analyzed in more complicated way. For example: if one takes the number 101001 in binary system can say that this number appeared in the following way. First we have mantissa 1 with the probability of occurrence according to Benford's law (eq. (7)) equal $P_0(d = 1) = 1$, then we have a mantissa 10 with the conditional probability of occurrence 0 after 1, given by (12), $P(0|1) = 0.584962\cdots$. Next we have a mantissa 101 with the conditional probability of occurrence 1 after 10, $P(1|10) = 0.44966\cdots$, etc. As a result for each mantissa one can have a set of conditional probabilities of successive digits appearance e.g. $P(0|1)$, $P(1|10)$, $P(0|101)$, $P(0|1010)$, $P(1|10100)$. This allows us to trace for each mantissa paths of unconditional probabilities versus conditional probabilities for each number (see Fig. 11 and its zoom – Fig. 12).

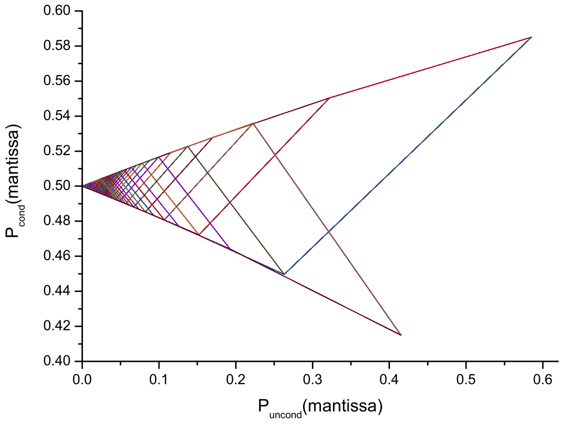

FIGURE 11. Moving on a self-similar tree for binary system with Benford's law according to conditional versus unconditional probabilities for successive digits

A binary system is the simplest one but even for quite short mantissas (Fig. 11 and 12 show probabilities for mantissas no longer than 11 bits, i.e., the numbers are smaller than $2048_{(10)}$) one can see that the self-similarity property of positional system with Benford's law can be visualized by complicated images of conditional and unconditional probabilities dependencies.

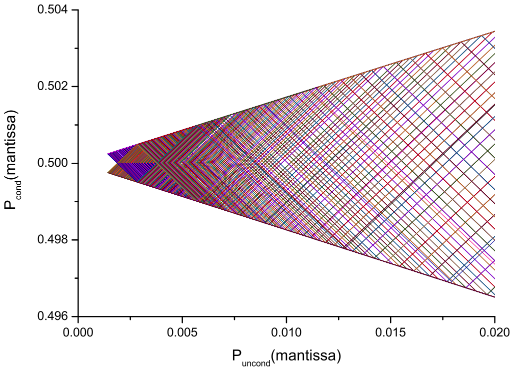

FIGURE 12. Magnification of Fig. 11 for $\mathrm{P}_{\mathrm{uncond}} < 0.02$

9.1. More complicated example – ternary system

A more complicated example can be given in the case of ternary system. It is still a quite simple positional number system but at each position there will be 9 possibilities (i.e., digit 0 after 0, 1 or 2, digit 1, after 0, 1 or 2 and digit 2 after 0, 1 or 2, i.e., there will be the following conditional probabilities $P(0|x)$, $P(1|x)$ and $P(2|x)$ where x stands for any mantissa with digits 0, 1 and 2) – let's note that previously it was only four such choices (generally it will be B^2 such choices for each position).

Visualizations for ternary system can be found in Fig. 13 and Fig. 14 where again the complicated nature of system self-similarity property is visible. Such visualizations can be made for other bases B; each of them will be more and more complicated but still will be governed by similar dependencies.

10. Connections with Zipf's law

It seems that the amount of words that are used by any human and the way that they are used is an individual matter of everyone. Many will also assume that there aren't any laws that govern this process, however it is known that there is a law, called Zipf's law [13] (for the first time it was noted by J. B. Estoup [2] in 1922), that for set of words X with the number of occurrences x_r ordered by the relation

$$x_1 \geq x_2 \geq \cdots \geq x_r \geq \cdots \geq x_n, \tag{14}$$

where r is a position of x_r in this order, makes a connection between r and x_r, which shows that

$$rx_r = \text{const.} \tag{15}$$

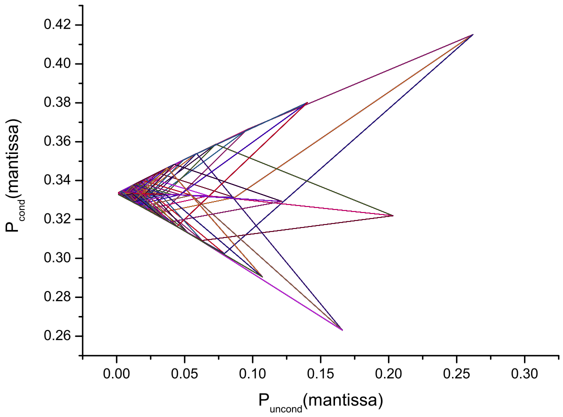

FIGURE 13. Moving on a self-similar tree for ternary system with Benford's law according to conditional versus unconditional probabilities for successive digits

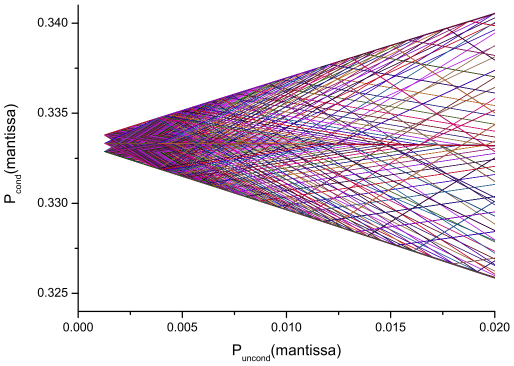

FIGURE 14. Magnification of Fig. 13 for $\mathrm{P}_{\mathrm{uncond}} < 0.02$

In other words, for long texts the frequency f of given word w occurrence multiplied by its position r on a sorted list is constant, i.e., $r(w) \cdot f(w) = \text{const.}$ Zipf analyzed James Joyce's Ulysses and discovered that the frequency of words occurrence is a power-law function $P_r(w) \approx 1/r^a$ with the exponent a close to unity. This is a very interesting phenomenon that can be also related not only to the texts but also to the cities sizes, the income or revenue of a company, etc.

as functions of the rank. Usually this power-law is also connected with fractals and self-similarity property [7] and can be also related to Pareto distributions.

The connections between Zipf's and Benford's laws are known in literature (see for example [5]), however in this paper it will be shown that such a connection can be obtained taking into account a little bit different approach than so far. Firstly, let's note that some kind of analogy can be made between the texts considered as sets of paragraphs, sentences, words where "*word*" is the smallest element of whole and sets of data, which are built by different numbers, because numbers can be considered as "*words*" in the case of numerical data. In the case of long texts one has Zipf's law with specific function of words probability occurrence, while for sets of numbers Benford's law seems to be the same case.

The equation (7) and the Fig. 2 show that if system base B grows, the non-homogeneity of first digits d probability occurrence falls (especially when $d \to B$). For example when $B = 3$ then $P(d = 1) = 0.630929\cdots$ and $P(d = 2) = 0.36907\cdots$ while for $B = 9$ it is $P(d = 1) = 0.315464\cdots$ and $P(d = 2) = 0.184535123\cdots$ and between $P(d = 7) = 0.060772\cdots$ and $P(d = 8) = 0.053605\cdots$ the differences are quite small (≈ 0.007167). Even for small values of B one can plot probabilities obtained for first significant digits on a log-log scale (i.e., Fig. 2 will be plotted log(digit) vs. log(probability)). As it will turn out in such a scale Benford's law for different bases B will be almost straight line with some exponent a, which values are given in Table 3 (the results were obtained by least mean square method).

Table 3. Values of exponent for different bases $B \leq 10$

Base	Exponent
3	-0.7736
4	-0.7976
5	-0.8152
6	-0.8289
7	-0.8399
8	-0.8491
9	-0.8569
10	-0.8637

Obviously one can use any system base B, not necessary equal 10 but even equal 100 or 1000. For such bases the number of first leading digits will be equal 99 or 999, etc. Such systems aren't specially useful but they will show very interesting connections between Benford's and Zipf's laws. Basing on results from Table 3 it might be expected that the values of exponent a can tend -1 and for large bases B it is confirmed – see Table 4.

Thus if $B \to \infty$ then the value of exponent $a \to -1$ showing that Benford's law for large bases B becomes similar to Zipf's law.

TABLE 4. Values of exponent for large bases B

Base	Exponent
100	-0.95958
1000	-0.990777
2000	-0.994332
5000	-0.997095

11. Conclusions

Benford's law itself is very surprising and as it was shown above can have many interesting properties. Some of them were indicated in this paper. Despite the fact that it seems to be a some kind of "oddity" it has got some interesting applications. The most known example is the problem of tax fraud detection that was given by Mark Nigrini [9]. It is also possible to use this law in the case of pictures analysis because their existence was there indicated [6].

References

[1] F. Benford, *The law of anomalous numbers*, Proc. Amer. Phil. Soc. **78** (1938), 551–572. (1938)

[2] J. B. Estoup, *Les Gammes Stenographiques*, Institut Stenographique de France, Paris, 1916.

[3] T. P. Hill, *Base-invariance implies Benford's law*, Proc. Amer. Math. Soc. **123** (1995), no. 3, 887–895.

[4] ———, *The first digit phenomenon*, Amer. Sci. **86** (1998), 358–363.

[5] S. Irmay, *The relationship between Zipf's law and the distribution of first digits*, J. Appl. Statist. **24** (1997), no. 4, 383–393.

[6] J.-M. Jolion, *Images and Benford's law*, J. Math. Imaging Vision **14** (2001), no. 1, 73–81.

[7] B. B. Mandelbrot, *The Fractal Geometry of Nature*, W. H. Freeman and Co., San Francisco, Calif., 1982.

[8] S. Newcomb, *Note on the frequency of use of the different digits in natural numbers*, Amer. J. Math. **4** (1881), no. 1-4, 39–40.

[9] M. J. Nigrini and L. J. Mittermaier, *The use of Benford's law as an aid in analytical procedures*, Auditing: A Journal of Practice and Theory **16** (1997), 52–67.

[10] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill Companies, 3rd edition, 1991.

[11] H.-O. Peitgen, H. Jurgens, and D. Saupe, *Chaos and Fractals: New Frontiers of Science*, New York, Springer-Verlag, 1992.

[12] L. Pietronero, E. Tosatti, V. Tosatti, and A. Vespignani, *Explaining the uneven distribution of numbers in nature*, Physica A **293** (2001), no. 1, 297–304.

[13] G. K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge, 1949.

Department of Distributed Systems
Rzeszów University of Technology
W. Pola 2
35-959 Rzeszów, Poland
E-mail address: strzalka@prz.edu.pl