Konstantin P. Kobzar

ON A SENSE OF THE BENFORD`S LAW: DIGIT OR NUMBER?

In 1881, Simon Newcomb published article [1] based on logarithmic tables usage analysis which demonstrated that the first digits of numbers in data sets do not give uniform distribution as it could be expected. The first digits appear with different frequency, the logarithmic dependence showing. Newcomb managed to carry out in-depth analysis in his small paper. He revealed inhomogeneous distribution of the first digits, quantitatively described it and offered his explanation. Perhaps the manner of proving seemed too easy and "unscientific" or the study was ahead of time, but it was not noticed and appreciated by contemporaries. That is why when in 1938 Frank Benford rediscovered the described law on the basis of analysis of first digits occurrence in 20 different data sets [2] it was taken as novel, interesting and meaningful and was called Benford`s law. The carried researches showed that there is nothing special in this law about metric system or about decimal base, the law is invariant under changes of scale and under changes of the logarithm base and inhomogeneous distribution occurs not only for the first but also for the following significant digits.

The Newcomb-Benford law explanation was interesting for many analysts. S. Newcomb believed that all natural numbers are result of multiple divisions and it conditions of a logarithmical distribution of first digits. Subsequently many different explanations of the law were offered. References can be looked, for example, in [3]. Created by T. Hill the theory shows, that the Benford`s law has a statistical character [4-5].

The first significant digits $X = k$ ($k = 1, 2, \ldots, 9$) probability distribution, characterized by the Benford's law, is described by formula

$$P(k) = \log_{10}\left(1 + \frac{1}{k}\right) \qquad (1)$$

Expressions for the second and the following digits can be written in a similar way.

Since the law has probabilistic nature, discussed probability should be focused on. Probability is fundamentally relative value, it is always defined relative to the total number of possible events. From formal considerations, expressions (1) and others, characterizing Benford's law digits distribution, should have denominator that reproduce net digits totality. In standard law formulation denominator is always the same and thus can be ignored. However there are real data sets when some significant digits are missed in principle or happen very seldom. For example, it is height of people, weight of electrons, etc. From the position of probability Benford's law conception can be expanded and digits occurrence probability can be evaluated relative to part of digits, not all of them. It allows analysing the mentioned examples.

This law, called also the Significant-digit law or the First-digit law, is formulated for digits. But digits are determining a basis of the Benford`s law or they characterize an essence of this law indirectly only?

Analysis of bulk concrete data sets resulted in summary that some data sets are well described with Benford's law. However in other numerical data the deviations from theoretical meanings are observed, besides there are exceptions - certain data sets do not follow Benford's law [6]. Therefore law seems to act randomly. Is it really, maybe, as some authors speak, a mathematically strict sense of Benford's law has proven elusive [6] and its ubiquity in real data sets is mysterious [3]? There is necessity to emphasize here, that the statistical description of the Benford`s law answers a question "as all occurs" more, than on a question "why it occurs", i.e. actually it is necessary to speak not about a statistical derivation, but about the statistical form or statistical show of the law.

When studying actual data sets perfect coincidence between obtained distributions and theoretical values (e.g., following from the law) can not be always expected. The question is different. Two main kinds of distribution are found in data sets [7]. The first one is normal distribution. This is, for example, numbers of houses, book pages and other ordered sequences. Lognormal distribution that corresponds to Benford's law occurs for digits from random numerical sets characterizing natural or artificial objects [1, 2]. That shows that there are two types of mathematical objects, each of them with its own kind of distribution. Therefore, it is important to understand what the types of objects are.

One crucial aspect should be specified. F. Benford as well as S. Newcomb speak about the first significant digits of numbers, i.e., about discrete system. Also about discrete systems there is spoken at the extended interpretation of the Benford`s law, when the distributions of seconds, thirds or any other digits are described. At the same time, the law is described by logarithmic function, which continuous distribution of sizes characterizes in the essence. The given contradiction is logically solved if to recognize, that the law characterizes actually not digits, but numbers. Such representation reflects the law logic, because the law characterizes not digits per se, but numbers, which begin with specific digits and consist of specific digits. Benford's law focuses on digits. Thus, numbers consideration is shaded and digits seem to play the leading role. If to refuse stereotypes, it becomes clear that actually the law expresses the numbers distribution.

Expression (1) can be represented in form $P = \log_{10}(k+1) - \log_{10} k$. In this case the law is evident to characterize not separate digits but difference between them expressed in a certain way. Actually digits k and $k+1$ reflect limits of interval that include numbers starting with k digit. Thus, the law discovered by S. Newcomb and F. Benford characterizes distribution of intervals limited with numbers that begin with specific digits.

At transition to numbers the probability estimation is fulfilled for interval numbers $[a,b)$ on interval $[c,d)$, where $[a,b) \in [c,d)$. All numbers *a, b, c* and *d* – are real numbers since numerical series include not only natural but also fractional and irrational numbers. The decimal basis of the logarithm ceases to matter in sense of presentation even and reasonably to proceed to the natural logarithms. Probability P for interval numbers X

$$P(a < X < b) = \frac{\ln b - \ln a}{\ln d - \ln c} \tag{2}$$

For example, according to expression (2) the occurrence probability of numbers of interval $[e, \pi)$ on interval $[2{,}5)$ is about 0.158. Thus, expression (2) is a theoretical formula of the law.

On the other hand, totality of numbers that belongs to the interval can be described with corresponding integral. At this approach probability distribution expression is given as

$$P[a, b) = \frac{\int_a^b f(x)dx}{\int_c^d f(y)dy} \tag{3}$$

where x are numbers in interval $[a, b)$, and y are numbers in interval $[c, d)$ by $[a, b) \in [c, d)$. Since expressions (2) and (3) define the same object they can be equated each other.

$$\frac{\int_a^b f(x)dx}{\int_c^d f(y)dy} = \frac{\ln b - \ln a}{\ln d - \ln c} \tag{4}$$

In expression (4) numerators and denominators are fairly independent since they characterize different intervals. Thus, they can be equalled separately. For example, for numerators equation

$$\int f(x)dx = \ln b - \ln a \tag{5}$$

Integrand form can be defined

$$\int_a^b \frac{1}{x} dx = \ln b - \ln a \tag{6}$$

Denominator equation is similar, and in whole probability equation

$$P[a,b) = \frac{\int_a^b \frac{1}{x} dx}{\int_c^d \frac{1}{y} dy} = \frac{\ln b - \ln a}{\ln d - \ln c} \tag{7}$$

Thus, expression $1/x$ for number occurrence probability defines probability dependence for intervals, i.e., observed logarithmic distribution indirectly described with the Benford's law. The numbers occurrence frequency in form $1/x$ implies that the larger the number, the more infrequent it is, and vice versa. For example, number 2 should occur less frequent than 1, and 3 less frequent than 2 and so on. If continue till 10, it will occur less frequent than 9 though it begins with 1. And only continuing numerical sequence to the next order of magnitude founds that numbers that begin with unit occur more often than numbers beginning with 2, 3, 4…

In 1896, V. Pareto revealed hyperbolic form of wealth distribution [8], and in 1935 G.K. Zipf discovered that word occurrence frequency in a text changes by law $1/x$ from the most frequent words to other [9]. This relationship was called Zipf's law. In 2001, L. Pietronero et al. took notice to certain relation between Benford's and Zipf's laws [10], it was studied on specific examples by S. Irmay [11]. As follows from the formula (7), a relation between Benford's and Zipf's laws has deep mathematical meaning.

Evident summary follows from the revealed expression $1/x$. Since occurrence probability for each number is inversely to the number value, then product of number value and its occurrence probability is equal for all numbers, that is $x \, {}^{1}\!/\!_{x} = const$. Thus, quantitative contribution of each number to number totality is equal; therefore the product of number value and its occurrence probability is equal in any numerical set.

Consequently, discovered by S. Newcomb and F. Benford the law characterizes not digits, but numbers and logarithmic distribution relates to numbers. At the same time digits can be supposed to have normal distribution that is intuitively obvious. However that "digit" concept refinement is required. In fact the concept must be expanded to involve all values that have equal occurrence probability, regardless of number of symbols the “digit” contain. Perhaps, it is reasonable to save “digit” concept as a “symbol”. Then new object type that will cover a kind of digits, which possess certain number properties, must be separated. These objects can be called, for example, “pseudo numbers”. So ordered sequence will be pseudo numbers. Undoubtedly, winning “numbers” in lotteries are pseudo numbers that is what lotteries are based on. Of course, pseudo numbers can contain one digit only. Therewith it should be realized that data sets can contain both numbers and pseudo numbers.

Thus, the first and other digits have no specific inherent distribution. The distribution depends on object types (numbers, pseudo numbers or their combinations) mentioned digits belong to. Pseudo numbers occurrence in data sets is equally probable. Numbers occurrence probability in data sets is inversely to number values. Occurrence probability of number intervals in numerical sets conforms to logarithmic law. Consequently, in new realization Benford’s law can be construed as a distinction criterion between numbers and pseudo numbers.

References

1. Newcomb S. Note on the frequency of the use of digits in natural numbers// Amer. J. Math. – 1881. Vol. 4.№ 1. – P. 39-40.

2. Benford F. The law of anomalous numbers // Proc. Amer. Phil. Soc. – 1938. – № 78. – P. 551-572.

3. Berger A., Hill T. P. Fundamental flaws in Feller’s classical derivation of Benford’s law / University of Alberta, Edmonton, 2010. – 8 p.

4. Hill T. P. Base-invariance implies Benford's Law // Proc. Amer. Math. Soc. – 1995. Vol. 123.№ 3. – P. 887–895.

5. Hill T. P. The first digit phenomenon // Amer. Scientist. – 1998. Vol. 86. № 4. – P. 358–363.

6. Leemis L. M., Schmeiser B. W., Evans D. L. Survival distributions satisfying Benford`s Law // Amer. Stat. – 2000. Vol. 54.№ 4. – P. 236-241.

7. Aldous D., Phan T. When can one test an explanation? Compare and Contrast Benford's Law and the Fuzzy CLT // Amer. Stat. – 2010. Vol. 64.№ 3. – P. 221-227.

8. Pareto V. Cours d'economie politique: 2 vols / Geneva Switzerland – Lausanne et Paris, 1896-1897 – 430 p.

9. Zipf G. K. Psycho-biology of languages / Houghton-Mifflin, 1935 – 336 p.

10. Pietronero L., Tosatti E., Tosatti V., Vespignani A. Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf // Phys. Stat. Mech. Appl. – 2001. Vol. 293.№ 1-2. – P. 297-304.

11. Irmay S. The relationship between Zipf's law and the distribution of first digits // J. Appl. Stat. – 1997. Vol. 24.№ 4. – P. 383-393.