ELSEVIER

Physica A 293 (2001) 297–304

PHYSICA A

www.elsevier.com/locate/physa

Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf

L. Pietronero[a,*], E. Tosatti[b,c], V. Tosatti[b], A. Vespignani[c]

[a] *Dipartimento di Fisica and Unitá INFM dell'Universitá di Roma "La Sapienza", P.le A. Moro 2, I-00185 Roma, Italy*

[b] *SISSA/ISAS and Unita' INFM Trieste SISSA, Via Beirut 2-4, 34014 Trieste, Italy*

[c] *The Abdus Salam International Centre for Theoretical Physics (ICTP), P.O. Box 586, 34100 Trieste, Italy*

Received 28 November 2000

Abstract

The distribution of first digits in numbers series obtained from very different origins shows a marked asymmetry in favor of small digits that goes under the name of Benford's law. We analyze in detail this property for different data sets and give a general explanation for the origin of the Benford's law in terms of multiplicative processes. We show that this law can be also generalized to series of numbers generated from more complex systems like the catalogs of seismic activity. Finally, we derive a relation between the generalized Benford's law and the popular Zipf's law which characterize the rank order statistics and has been extensively applied to many problems ranging from city population to linguistics. © 2001 Published by Elsevier Science B.V.

Suppose you look at today's stock prices and bet on the value of the first digit. One could guess that a fair bet should correspond to the frequency of $\frac{1}{9} = 11.11\%$ for each digit from 1 to 9. This is by no means the case, and one can easily observe a strong prevalence of the small digits over the large ones. The first three integers 1–3 alone have globally a frequency of 60% while the other six values 4–9 appear only in 40% of the cases. This situation is actually much more general than the stock market and it occurs in a variety of number catalogs related to natural phenomena. The first observation of this property traces back to Newcomb in 1881 [1] but a more precise account was given by Benford in 1938 [2,3]. He investigated some 20

* Corresponding author. Fax: +39-06-446-3158.

E-mail address: luciano@pil.phys.uniromal.it (L. Pietronero).

0378-4371/01/$ - see front matter
PII: S0378-4371(00)00633-6

tables of numbers ranging from the area of lakes and the length of rivers to the molecular weights of molecular compounds. In all cases he found the same behavior for which he guessed the first digit probability distribution $P(n) = \log[(n+1)/n]$, where $n = 1, 2, \ldots, 9$ is the first digit. Since then this observation has been repeatedly reported as a mathematical curiosity [4,5] with some computer science applications [6] and even tax-fraud detection [7].[1] In this paper we illustrate these observations with the specific example of the stock market [8,9], and we propose a more general framework that leads to the definition of a generalized Benford's law which can be linked to the Zipf's law [10]. We test the occurrence of the proposed generalization by analyzing for the first time the distribution of catalogs of seismic activity. We further identify a general mechanism for the origin of this uneven distribution in the multiplicative nature of fluctuations in economics and in many natural phenomena. This provides a natural explanation for the ubiquitous presence of the Benford's law in many different phenomena with the common element that their fluctuations consists of a fraction of their values. This brings us close to the problem of the spontaneous origin of scale invariant properties in various phenomena which is a debated question at the frontier of different fields.

Consider the values of the Athens, Madrid, Vienna and Zurich stock markets of January 23, 1998. The stock prices N are expressed in the local currencies. If the values of N were randomly distributed we would expect a uniform distribution for the value of the first digit $n = 1, 2, \ldots, 9$. This would lead to $P(n) = \frac{1}{9} = 11.11\%$. In Fig. 1 we can see instead that $P(1)$=30% and then the frequency decreases continuously with n until the minimum value $P(9) = 4.5\%$. This is the asymmetric distribution of first digits pointed out by Benford [2,3] who investigated mostly tables of natural numbers. Here we can see that such a behavior applies to economic data as well. The prevalence of the number 1 in the first digit may suggest ad hoc explanations like the fact that there may be a tendency, in each country, to assign price values of the order of unity in local currency. This hypothesis can be easily checked by expressing the stock prices in a different currency. To this purpose, we repeat the analysis re-expressing the Madrid stocks in Swiss francs and the Zurich ones in Pesetas. Remarkably, we observe that Benford's distribution is independent of the currencies adopted. This fact may appear strange at first sight, but is in fact just another example of scale invariance [11]. This property should be found in the probability distribution of the original stock prices $P(N)$ and then it should reflect also in the distribution of their first digit $P(n)$. Changing the unit of measure of a price corresponds to multiplying it by a factor b, in our case the exchange rate between Swiss Francs and Pesetas. If we change our units by a factor b, the values N will become $N' = bN$ and the corresponding distribution should be identical to the original one, apart from a constant rescaling factor $A(b)$, that may depend upon b but not on N. In mathematical terms scale invariance corresponds

[1] Recently, the 10 July 1999 issue of *New Scientist* featured an article on Benford's law that started a very interesting debate that can be found on the journal web site (www.newscientist.com/forums/benford/benford.html).

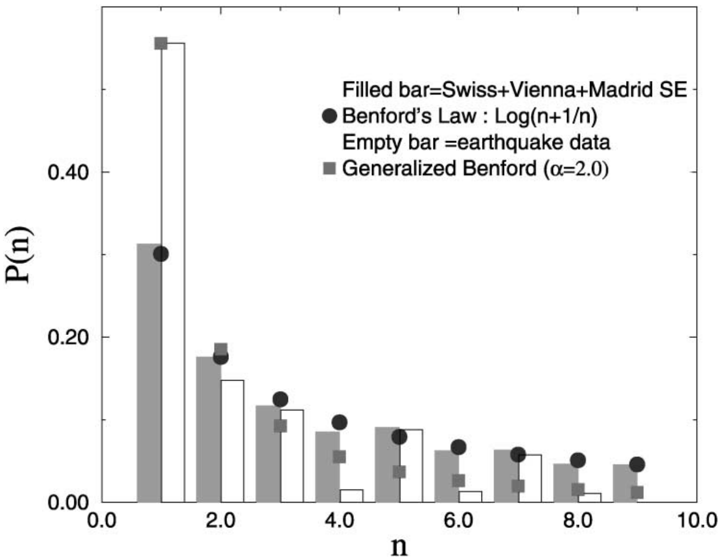

Fig. 1. Distribution $P(n)$ of the first digits of the stock prices N, averaging over three data sets (Madrid, Vienna and Zurich stock exchanges). The distribution is strongly asymmetric and it is well reproduced by the Benford's law shown for comparison with circles. Distribution of first digits from the earthquake magnitude catalog (see text) is reported in comparison with the generalized Benford's law with $\alpha = 2$ (squares). The agreement is fairly good for $n = 1$, 2 and 3 that represent more than 80% of the cases.

to the following functional relation:

$$P(N') = P(bN) = A(b)P(N)\,. \tag{1}$$

It is well known from statistical physics and critical phenomena that the general solution of this equation is any power law behavior with exponent α:

$$P(N') = N'^{-\alpha} = b^{-\alpha}N^{-\alpha}\,. \tag{2}$$

For this type of distributions we can compute the probability of the first digit by observing that for each decade we have the same relative probability for the various integers, so we can write

$$P(n) = \int_n^{n+1} N^{-\alpha}\,\mathrm{d}N = \frac{1}{1-\alpha}[(n+1)^{(1-\alpha)} - (n)^{(1-\alpha)}] \tag{3}$$

for $\alpha \neq 1$. For $\alpha = 1$, we have instead

$$P(n) = \int_n^{n+1} N^{-1}\,\mathrm{d}N = \int_n^{n+1} \mathrm{d}(\log N) = \log\left(\frac{(n+1)}{n}\right)\,, \tag{4}$$

that is precisely Benford's law as derived from the data analysis. From Eq. (4), we can see that the case $\alpha = 1$ corresponds to a uniform distribution in logarithmic space as from Eq. (4). This is the exponent which describes the stockmarket data. For $\alpha > 1$, Eq. (3) defines a generalized Benford's law in which, for increasing α, the first digit $n = 1$ becomes even more frequent, further increasing the unevenness of Benford's distribution. We can check if nature provides us with examples with $\alpha \neq 1$. We have analyzed in particular the southern California catalog for earthquake magnitudes. [2] It is well known that the earthquake magnitude distribution follows the famous Gutenberg–Richter law [12], that is a power law distribution with exponent ranging from 1.7 to 2.0. The first digit distribution of numbers on this catalog should then correspond to a generalized Benford's law with $\alpha \simeq 2.0$. In Fig. 1 we report the data analysis in comparison with the expected result. Despite the large statistical noise – attributable among other things to rounding errors in the seismic data – we obtain a good agreement for the first three digits which statistically are the most significant ones. This observation enlarges the occurrence and relevance of Benford-like asymmetries, by relating them to the physics of the vast class of systems showing power law behavior.

These observations raise two further questions: (i) The first is to understand why some data set naturally show scale invariant properties. (ii) The second is why, among the various scale invariant distributions corresponding to different α values, Benford's $\alpha = 1$ is that more frequently realized in nature. In the following we address these questions, and find the general mathematical origin of the distribution $P(N) = 1/N$ that can apply to a variety data set in several fields ranging from economics to physics and geology.

The most general mathematical property which applies to the statistics of a large variety of phenomena is the central limit theorem that governs the probability distribution corresponding to the sum of random numbers. Consider the value of a variable N which changes with time t by the addition of a random variable ξ, yielding the Brownian process

$$N(t+1) = \xi + N(t) . \tag{5}$$

If the random variable ξ is symmetrically distributed with finite variance, the probability distribution $P(N,t)$ to have a given value N after t additional steps will be Gaussian with variance $\sigma \sim t^{1/2}$. In the infinite time limit the variance is diverging and the probability distribution will approach the uniform one. This is indeed very far from the scale invariance properties we are looking for. Now, it is easy to realize that the Brownian dynamics does not realistically apply to many stochastic dynamical phenomena. Brownian dynamics is ruled by a noise term whose intensity is independent of the variable value N. Fluctuations are in this way independent and related to some external dynamical parameters. Clearly, many systems do not follow such a dynamical description. For instance, it is intuitive to consider that a stock price has fluctuations which are relative to the price itself. In practice, each stock suffers of percent oscillations.

[2] This is a public catalog available at the web site: http://scec.gps.caltech.edu. Magnitude reported range from 2 to 6 in a time window spanning from 1981 up to nowadays.

Hence

$$N(t+1) = \xi N(t)\,, \tag{6}$$

where ξ is again a stochastic variable, which in this case must be positive definite. The nature of this process is completely different from the usual Brownian motion. We can, however, relate the two processes by a simple transformation. If we take the variable in logarithmic space we get

$$\log N(t+1) = \log \xi + \log N(t)\,. \tag{7}$$

If we consider $\log \xi$ as the new stochastic variable, we recover a Brownian dynamics in a logarithmic space; i.e., a random multiplicative process corresponds to a random additive process in logarithmic space. This implies that for $t \to \infty$ the distribution $P(\log N)$ approaches a uniform distribution, and by transforming back to linear space we have

$$\int P(\log N)\,\mathrm{d}(\log N) = C \int \frac{1}{N}\,\mathrm{d}N\,, \tag{8}$$

where C represents the normalization factor. This immediately gives $P(N) \sim N^{-1}$ as the distribution of variable values N. Accordingly, the distribution of first digits n will follow an ideal Benford law with $\alpha = 1$.

In order to illustrate this result numerically, we have considered a uniform distribution of numbers in the interval $(0, 999)$. Clearly, such a case corresponds to a uniform distribution for both the actual numbers and their first digits. We apply to these numbers a multiplicative dynamics as described previously. At each iteration step, numbers are multiplied or divided by a factor Δ at random, following a bimodal distribution characterized by two delta function as $p(\xi) = \frac{1}{2}\delta(\xi - \Delta) + \frac{1}{2}\delta(\xi - \Delta^{-1})$. In this way we have that in the logarithmic space the stochastic variable $\log \xi$ has zero mean and is symmetrically distributed. We then repeat the updating process of variables, which in logarithmic space corresponds to the random addition of $\pm\log \Delta$. After enough multiplicative steps we can observe the resulting distribution of number N. Fig. 2 shows the outcome of 100 iteration steps with $\Delta = 1.5$. Both the first digit and the cumulative numbers distribution evolved according to Benford's law of Eq. (4). This simple exercise shows that the numbers N characterizing some physical quantities or objects, will naturally follow Benford's law if their time evolution is ruled by multiplicative fluctuations. It is worth remarking that many physical phenomena show scale invariant behavior and are characterized by power law distributions, due to cooperative effects, the onset of critical points and other nonlinear dynamical effects [11,13]. Here too, in the first digit distribution, scale invariance manifests itself with nontrivial power law exponents. Benford's law will apply to these phenomena in the generalized form of Eq. (3). In this sense, it is interesting to explore also connections with other well-known scale invariant features such as the Zipf's law [10].

The Zipf's law links a set of numbers like, for instance, the population of cities (W) in a given country to their rank order k. The empirical relation of Zipf yields

$$W(k) = \frac{A}{k}\,, \tag{9}$$

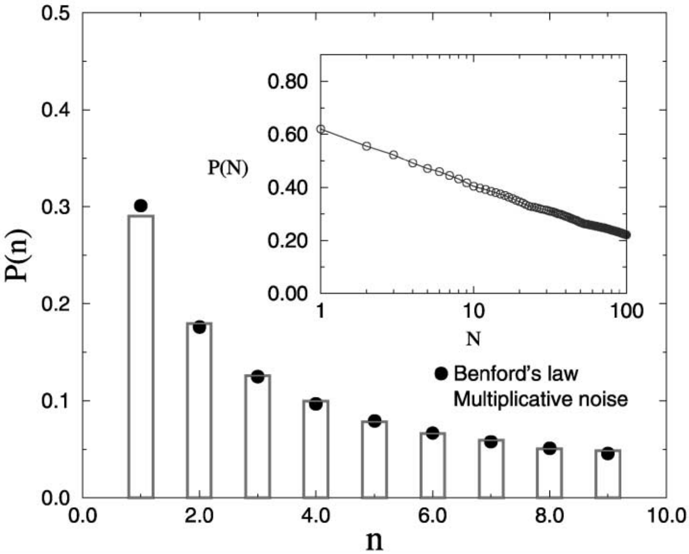

Fig. 2. Probability distribution of the first digit of a flat number distribution after 100 iterative applications of multiplicative noise (see text). The ideal Benford's law is strikingly satisfied. Small deviations disappear with more iterative noise applications. In the inset is shown the cumulative distribution of all the numbers $P_c(N)=\int_N^{K^*} P(N')\,\mathrm{d}N'$, where K^* is the upper cut-off; i.e., the largest number generated. For a Benford-like distribution we have $P_c(N) = C(-\log N + \log K^*)$ where C is the normalization factor. This behavior is nicely followed by the generated distribution as it appears from the linear plot on the semi-logarithmic scale.

where A is the population of the largest city with rank $k = 1$. The second largest city has rank $k = 2$ and population $W(2) = A/2$, and so on for higher ranking. This law has been tested and applied to various situations ranging from the distribution of population in cities and nations to the frequency of occurrence of a given word in linguistics, and it has been proposed by Gell-Mann as the prototype example of scaling [14].

It is interesting to enlight the relations between the Benford and Zipf laws by analyzing the rank properties of a set of $\mathcal{N}$ of numbers extracted from the general distribution $P(N) \sim N^{-\alpha}$. Considering that a finite value $N_{\max}$, corresponding to the rank $k = 1$, is the largest value of these $\mathcal{N}$ numbers, the rank k of the $N(k)$ number is given by all the numbers (out of $\mathcal{N}$) whose value is between $N(k)$ and $N_{\max}$:

$$k = \mathcal{N} \int_{N(k)}^{N_{\max}} P(N)\,\mathrm{d}N \sim N(k)^{1-\alpha}\,, \tag{10}$$

where we have considered that $N_{\max} \gg 1$ and gives only a small correction to the above expression. By inverting this result we obtain that

$$N(k) \sim k^{1/(1-\alpha)}\,, \tag{11}$$

which gives a connection between the generalized Benford's law and the Zipf's law with the following properties. (i) The original Benford's law ($\alpha = 1$) does not lead to a Zipf-type law because $N_{\max}$ diverges. (ii) Any power law distribution with $\alpha > 1$ leads to a generalized Zipf's law with exponent $1/(1 - \alpha)$. (iii) The standard Zipf's law corresponds only to distributions with $\alpha = 2$. The latter situation is not as common as the simple multiplicative process which lead to $\alpha = 1$. As we have seen previously, however, the earthquake magnitude distribution is close to $\alpha = 2$, and this value can be also found in several others situations. [3]

In summary, the fact that Benford's law is naturally explained in terms of a dynamics governed by multiplicative fluctuations can provide new insights on many scale invariant natural phenomena. For the stock market and several other phenomena, the simple multiplicative dynamics we have described appears to be quantitatively accurate and, in this respect, it explains the generality of Benford's law. On the other hand, phenomena such as earthquakes are also described by a scale invariant distribution, but not of the precise Benford type. This is not surprising in view of the complex dynamics of these phenomena that cannot apparently be reproduced by a simple multiplicative process. The understanding of the origin of scale invariance has been one of the fundamental tasks of modern statistical physics. How systems with many interacting degrees of freedom can spontaneously organize into critical or scale invariant states is a subject that is of upsurging interest to many researchers.

Acknowledgements

We thank G. Mussardo for useful discussions and comments. We are grateful to P. Shebalin for pointing out to us the earthquakes catalog. L.P. thanks the hospitality of ICTP where part of this work has been completed. L.P and A.V. acknowledge partial support from the TMR European Program ERBFMRXCT980183.

References

[1] S. Newcomb, Note on the frequency of use of the different digits in natural numbers, Am. J. Math. 4 (1881) 39–40.
[2] F. Benford, The law of anomalous numbers, Proc. Am. Philos. Soc. 78 (1938) 551–572.
[3] S.P. Richards, (pseudonym of Lines, M.E.), A Number for Your Thoughts (published by Richards, S.P. 1982).
[4] R.A. Raimi, The first digit problem, Am. Math. Monthly 83 (1976) 521–538.
[5] R.A. Raimi, The peculiar distribution of first digits, Sci. Am. (1969) 109–119.
[6] P. Schatte, On mantissa distributions in computing and Benford's law, J. Inform. Process. Cybern. EIK 24 (1988) 443–455.
[7] M. Nigrini, A taxpayer compliance application of Benford's law, J. Am. Taxation Assoc. (1996) 72–91.
[8] E. Ley, On the peculiar distribution of the US stock indices digits, Am. Stat. 50 (1996) 311–313.
[9] T.P. Hill, The first digit phenomenon, Am. Sci. 86 (1998) 358–363.

[3] See for instance the cluster–cluster aggregation phenomenon [15].

[10] G.K. Zipf, Human Behavior and the Principle of Least Effort, Addison-Wesley Press Inc., Cambridge, 1949.
[11] B.B. Mandelbrot, The Fractal Geometry of Nature, Freeman, New York, 1982.
[12] B. Gutenberg, C.F. Richter, Frequency of earthquakes in California, Bull. Seismol. Soc. Am. 34 (1944) 185–188.
[13] P. Bak, How Nature Works: The Science of Self-organized Criticality, Copernicus, New York, 1996.
[14] M. Gell-Mann, The Quark and the Jaguar: Adventures in the Simple and the Complex, Little, Brown and Co., London, 1994.
[15] T. Vicsek, M. Vicsek, Fractal Growth Phenomena, World Scientific, Singapore, 1992.