On the Newcomb-Benford Law

W. A. Kreiner

Abteilung Chemische Physik, Arbeitsgruppe Laseranwendungen, Universität Ulm, D-89069 Ulm

Reprint requests to Prof.W. A. K.; E-mail: welf.kreiner@chemie.uni-ulm.de

Z. Naturforsch. **58a,** 618 – 622 (2003); received October 6, 2003

It is well known that the leading digit in tables of statistical and physical data is not evenly distributed among the digits 1 to 9. Benford (who collected a large number of data) assumed that they follow a logarithmic law, commonly known as Benford's Law, although it was proposed earlier by Newcomb. We suggest, however, that the probability of the first digit being a 1, 2, ... depends on the particular distribution function of the data. For example, the size distribution of objects which grow exponentially is found to follow the Newcomb-Benford law. On the other hand, as the experimental data discussed in this paper show, the function governing the probability of the first digit of the weight of fragments obtained from crushing a stone deviates substantially from the Newcomb-Benford Law.

Key words: Benford; First Digit Law; Newcomb; Minerals.

In most numbers one has to deal with in everyday life, the first significant digit is much more often a 1 than a 9 [1]*, [2]. In case of street addresses, atomic weights, page numbers and paragraphs of law books, the reason is simply the fact that practically all numbering starts with one. There are eleven numbers between one and 20 with a leading 1. If the count is extended further on, one will find that each new decade starts with numbers contributing ten times more first 1's than the previous decade. Wherever the counting is stopped, the starting numbers of the highest decade with its many first 1's will always contribute to statistics. That's why averaging over many numberings produces many more first 1's than any other digit.

There are other reasons for the higher probability of having a 1 for a first digit than a 9. Newcomb [1] assumed that, for natural numbers, the mantissae of their logarithms are equally distributed. From this he derived his first digit law, giving the probability for the first digit x to be $1 \dots 9$ as

$$W_x = \log \frac{x+1}{x} \tag{1}$$

It can be shown that the size distribution of an object or entity which grows exponentially follows this function, the Newcomb-Benford law, exactly. Examples of such objects are cultures of bacteria, or to a good approximation the mass of trees during their initial growth period, and also, incomes and taxes. These all remain at a small size for a long period. As they grow, they spend less time within a particular size interval. If samples are taken from an ensemble of objects which have started to grow at different times, the majority is in the small size stage and hence values obtained correspond to small numbers.

For simplicity, let us first consider tree like objects, with a mass $X = 1$ when they are planted. In addition, let us assume that they are planted at regular time intervals (with frequency f), the first one at $t = 0$, the second at $t = 1$ and so on and that they exhibit exponential growth. After some time, the complete size distribution will extend over exactly two decades (Fig. 1), *i.e.*, the oldest tree exhibits mass $X = 100$ and the youngest has just been planted,

* From his publication: As natural numbers occur in nature, they are to be considered as the ratios of quantities. We must select two numbers and inquire what is the probability that the first significant digit of their ratio is n. ... Let us suppose the numbers to be arranged according to the characteristics of their logarithms of the form $c+s$, $c'+s' \dots$ (c and c' being integers). The significant figures of the ratio will be independent of the integers c, $c' \dots$, since changing the integers will only change the decimal point. ... We have a series of numbers represented by a distribution of exponents s, $s' \dots$ which may be arranged around a circle. Finally, the fractions will approach to an equal distribution around the circle. We thus reach the conclusion:
The law of probability of the occurrence of numbers is such that all mantissae of their logarithms are equally probable.

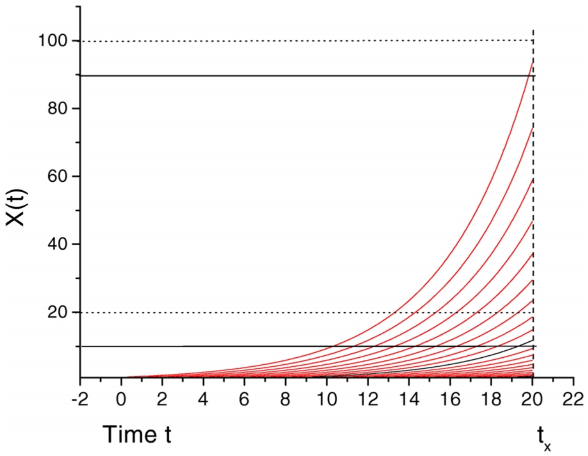

Fig. 1. Mass X of trees as a function of t when planted at different times t_n. The function shown is $X(t) = e^{0.2303t}$. All starting values are $X(t_n) = 1$. The goal is to find the size distribution at $t_x = 20$.

$X = 1$.

With the growth function $X(t) = e^{at}$ and the number of trees $n = f \cdot t$, as a function of time, the fraction of trees found within the mass interval dX at time t is

$$\frac{\mathrm{d}n}{\mathrm{d}X} = \frac{f \cdot \mathrm{d}t}{a \cdot e^{at}\mathrm{d}t} = \frac{f}{a}\frac{1}{e^{at}} = \frac{f}{a}\frac{1}{X} = D(X). \qquad (2)$$

$D(X)$ is the probability density function which is $\sim X^{-1}$ in this particular case and gives the probability of X being found within the mass interval $[X, X + \mathrm{d}X)$. Integration yields the number of masses within the interval $[X1, X2)$

$$n[X1, X2) = \int_{X1}^{X2} \frac{f}{a}\frac{1}{X}\mathrm{d}X = \frac{f}{a}[\ln X2 - \ln X1]. \qquad (3)$$

This result is identical to the Newcomb-Benford law. In this particular case, the probability for the first digit to be a "1" is $(\log 2 - \log 1) = \log 2$, corresponding to the length of time the objects remain within the size interval $[1, 2)$ or $[10, 20)$, respectively.

The probability distribution of the first digits is not altered at all when changing the units (*e.g.*, lb to kg), as long as the measuring system is linear, nor when the objects start to grow at a size $X \neq 1$. For each complete decade, the statistical weight of the first digits will follow this particular distribution, (1).

In general, the same number of objects will be found within all size intervals that are evenly spaced on a logarithmic scale. This applies to any decade, too. Therefore, considering many consecutive decades separately, the statistical weight of the first digit exhibits a saw tooth function. Newcomb has investigated the probability distribution of the second digit as well. Zero occurs 30% more frequently than 9. (For the third digit, there is still a difference of about 4%.)

Things are similar when entities are divided. Division of any finite number by two corresponds to subtraction of $\log 2$ on a logarithmic scale. On this scale, the range betweeen the numbers 1 and 2 comprises more than 30% of the range of a complete decade. As a consequence, consecutive division by two (or any number $\neq 0$) leads to a series of numbers where the probability of the first digit can be approximated by (1). This can be verified by starting with a randomly chosen number.

The probability of the first digit being a 1 increases beyond the Newcomb-Benford law, when the size distribution of the objects follows the function

$$D(X) = \frac{c}{X^P} \qquad (P > 1, c \text{ being a constant}), \quad (4)$$

as it is often the case with fractals. In the following, X means the mass of a sample.

As an example, let (4) describe the distribution of fragments obatained from crushing a stone. To start with, only the decade $[1, 10)$ is considered. Further, the samples are arranged in groups $z = 1, \ldots, 9$, where the weights of all samples belonging to one group exhibit the corresponding leading digit, $x = 1, \ldots, 9$. The statistical weight $W(z)$ of the first digit being $x = 1 \ldots 9$ is given by the integral of the size distribution function $D(X)$ within the limits of the interval $[x, x + 1)$:

$$W(z) = P1 \cdot \int_{x}^{x+1} X^{-P}\mathrm{d}X = \frac{P1}{1-P}\left[(x+1)^{1-P} - x^{1-P}\right]$$

$$(\text{for } x = 1, 2, \ldots, 9). \qquad (5)$$

When the sum over all samples is taken as 100%, $P1$ can be eleminated, leading to the expression

$$W(z) = \frac{100}{10^{(1-P)} - 1}\left[(x+1)^{1-P} - x^{1-P}\right]. \qquad (6)$$

For other decades, say $[0.01, 0.1)$ or $[100, 1000)$, the distribution of the first digits is the same, although the number of objects within each of the decades is different for $P > 1$. (In case of the Newcomb distribution, $P = 1$).

First digit / Sample	1	2	3	4	5	6	7	8	9	Number of fragments	Power law exponent P
Weißer Jura 3115 g	1339	591	354	233	157	118	106	72	59	3029	1.555(13)
4230 g	1868	797	461	308	195	141	121	98	86	4075	1.618(13)
1523 g	1867	810	457	328	226	145	114	97	77	4121	1.597(16)
363.4 g	238	105	73	39	36	22	14	19	12	558	1.493(31)
Σ	5312	2303	1345	908	614	426	355	286	234	11783	Pav 1.587(13)

Table 1. Distribution of First Digits.

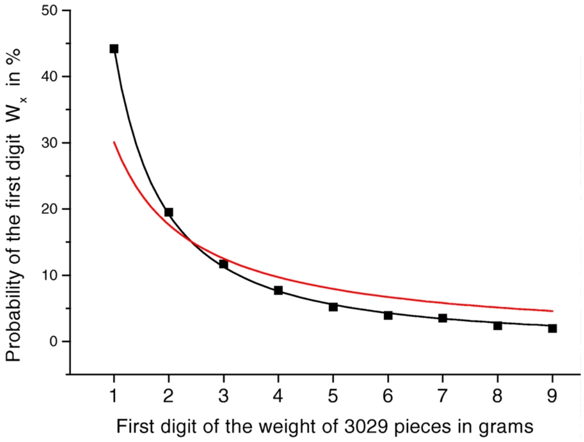

Fig. 2. Probability of the first digits 1...9 from the weight of 3029 pieces obtained by crushing a sample of Weißer Jura. The flater curve corresponds to the distribution proposed by Newcomb. The large probability of the first 1's and 2's is a consequence of $P > 1$ in the mass distribution function (4). $P = 1.555(13)$ was obtained from fitting (6) to the data points indicated by squares. The Newcomb distribution corresponds to $P = 1$ (2). Chi^2 is 0.136 when the power law is fitted and 30.3 for the Newcomb function.

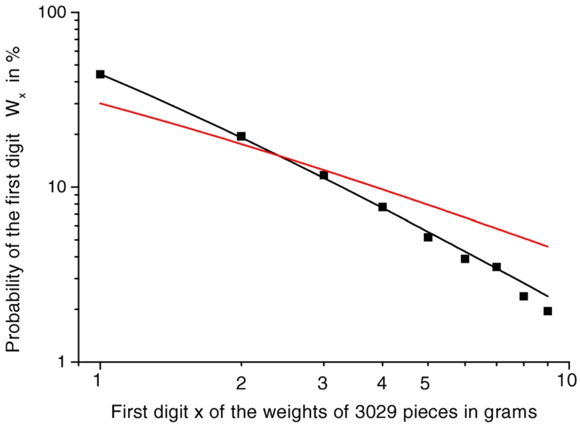

Fig. 3. Probability of first digits 1...9 from pieces of Weißer Jura. Plot of $\log W(z)$ vs. log x shows the deviation from Newcomb's law. While, due to Newcomb, the 1 should occur 6.6 times as frequent as the 9, the corresponding probability factor is approximately 20 in the actual case.

To apply (6) to a realistic distribution, a piece of mineral (Weißer Jura, Upper Jurassic period, from the Schwaebische Alb; weight 3115 g) was crushed repeatedly with a press. All pieces between 0.01 and 636.8 g were weighed. The mass of the weighed pieces amounted to 97.43% of the total mass. From 3029 measurements covering nearly five orders of magnitude, the distribution function $W(z)$ of the first digits was fitted. This is shown in Figs. 2 and 3 (linear and logarithmic plot, respectively), along with a fit to (6). The result $[P = 1.555(13)]$ shows that the statistical weight of the first 1's is considerably higher than expected from the law proposed by Newcomb (and verified by Benford (2) on over 20 000 data points). The significance level of the fit is above 0.95. Chi^2 amounts to 0.136. For comparison, the Newcomb distribution is plotted in grey. According to Newcomb, the 1 should occur ≈ 6.6 times more frequently than the 9, whereas, in the actual case, the corresponding probability factor is 19.03 (from the fitted function) and 22.69 (experiment).

From a total of four samples of Weißer Jura from the Schwaebische Alb (ranging from 363 g to 4230 g in weight) an average power law exponent $P = 1.587(13)$ was derived (Table 1). One sample of Schwarzer Jura from the Holzmaden area achieved $P = 1.579(24)$, while the result for a sample of green schist is $P = 1.772(9)$. However, no obvious dependence on the size of the sample was found so far.

To give an impression of the average distribution (or number density) function of Weißer Jura, the average P values have been determined separately for each of the five decades. For this purpose the numbers of all fragments falling into the interval $z = [10\text{ mg}, 20\text{ mg})$ were added, then for $z = [20\text{ mg}, 30\text{ mg})$ and so on, up to $z = [600\text{ g}, 700\text{ g})$. This is shown in Fig. 4 (kind of saw tooth function). See also Table 2. Then, using (6), five P values have been fitted for each of the decades and from the nine values each. From these five parameters distribution functions have been plotted and composed to an over- all function. The semi logarithimic plot is given in Figure 5.

First digit / Decade	1	2	3	4	5	6	7	8	9	Number of fragments	Power law exponent P
[100 g, 1000 g)	6	4	2	1	0	3	0	0	0	16	1.342(226)
[10 g, 100 g)	39	9	9	6	3	2	1	5	0	74	1.860(126)
[1 g, 10 g)	169	83	29	29	18	13	15	11	8	375	1.602(59)
[100 mg, 1000 mg)	923	368	216	134	115	63	59	44	31	1953	1.668(16)
[10 mg, 100 mg)	4181	1841	1089	739	478	345	280	226	195	9374	1.571(15)

Table 2. Total Number of Fragments per z Interval and per Decade. Four Samples of Weißer Jura.

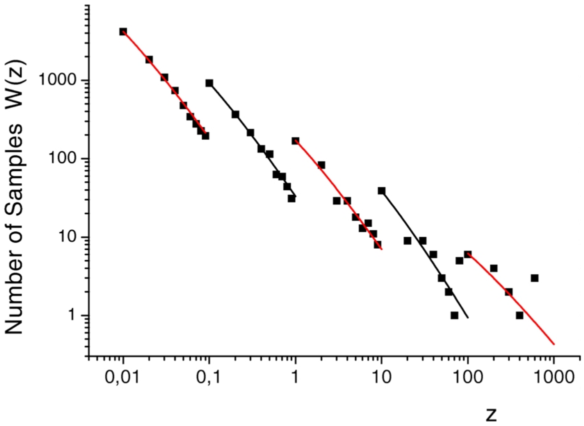

Fig. 4. Plot of the average probability function $W(z)$ of four samples of Weißer Jura. The squares indicate the number of fragments per z interval. Due to the logarithmic scale zero values of $W(z)$ in the two highest decades cannot be shown.

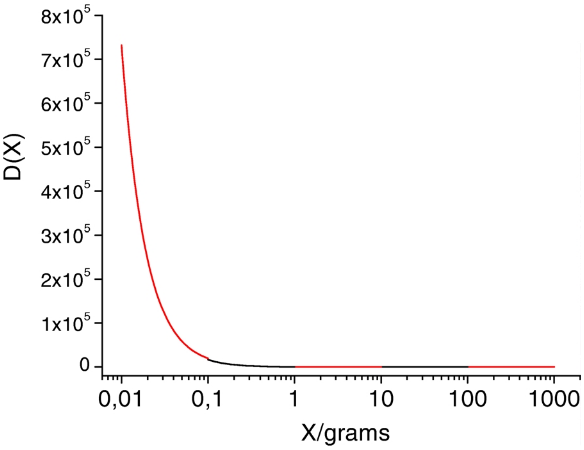

Fig. 5. Plot of the average distribution function of four samples of Weißer Jura. The sections correspond to the P values which have been fitted separately. This provides some indication of a single slope distribution over nearly five decades.

In conclusion, one can say that every distribution of measured values has a unique distribution of the first digits. In many cases, the 1 occurs more often as a first decimal digit than the 9. In general, a non-uniform distribution among the digits 1 to 9 at the first significant place is observed, when the values obtained from measurement do not follow a uniform distribution on a linear scale.

As already discussed here, the size distribution of objects which grow exponentially is identical to the one which is derived from the Newcomb hypothesis. An example where with this is not the case, is the distribution of the weight of fragments obtained from crushing a mineral. There, the probability of the occurrence of first digits is considerably different from the Newcomb law in the sense that there is an abundance of small fragments.

The question arises, why is the Newcomb law not obeyed in this case? Or perhaps the question should be put in the form of why should it?

In general, the exponent of the distribution function depends on the variable. When, in (4), the mass X of the fragments is replaced by an expression proportional r^3 (r being their linear dimensions), P will be multiplied by 3. With the density function and the P values determined experimentally it is possible to estimate the sum of all diameters of all fragments falling into a certain interval, say, one decade of mass. It turns out that a chain formed in this way gets longer and longer when the fragments get smaller. The sum of surfaces, however, stays nearly constant, therefore being the entity which comes closest to Newcomb's law.

In a crushing process, even when force is applied continously in a press, fragmentation occurs in a stepwise manner. The fragments surface is proportional to the number of atoms which have to be separated on cleavage. The number and the size distribution of fragments is most probably a function of the dynamic processes involved as soon as the stored energy is released.

With regard to power law distributions in nature, there is literature on grain size distributions in soils and sediments [3, 4]. Attempts have been made to fit the size distribution by two or three slope power law functions. Similar results have been obtained for the size distribution of asteroids [5] and from fragmentation by impact experiments, *e.g.*, by shooting a projectile on a basalt target [6]. Examples of terrestrial, lunar and interplanetary power law rock fragmentation have been collected by Hartmann [7]. A power law exponent (or differential power law index) around 4 is found for the frequency of meteor impact craters versus diameter on the Jupiter moon Ganymed (8). In addition, it should be mentioned that the average value of $P[= 1.587(13)]$ in the mass distribution function can be compared with the exponent of a function resulting from computer simulations of earthquakes [9]. There it was found that the occurrence frequency of clusters of masses moved during an earthquake scales with cluster size s as $s^{-\tau}$, with a power law exponent $\tau = 1.6$.

Among all distributions occurring naturally, an equal distribution among the digits 1 to 9 seems to be rare. This would be the case if, in the first example, the objects would not grow exponentially, but linearly. One can even think of size distributions where the first 9's are most frequent (growth function $\sim t^k, 0 < k < 1$). Of course, there is no distribution of first digits at all as soon as the distribution of values narrows down to a very small range.

[1] Simon Newcomb, Note on the Frequency of Use of the Different Digits in Natural Numbers, Amer. Journ. of Math. **4**, 39 (1881).
[2] Frank Benford, The Law of Anomalous Numbers, Proceed. American Phil. Society **78**, 551 (1938).
[3] M. Bittelli, G. S. Campbell, and M. Flury, Characterization of Particle-Size Distribution in Soils with a Fragmentation Model, Soil Sci. Soc. Amer. J. **63**, 782 (1999).
[4] Gabor Korvin, Fractal Models in the Earth Sciences, Elsevier Amsterdam, 1992.
[5] E. V. Ryan, Asteroid Fragmentation and Evolution of Asteroids, Annual Review of Earth and Planetary Sciences **28**, 367 (2000).
[6] A. Fujiwara, G. Kamimoto, and A. Tsukamoto, Destruction of Basaltic Bodies by High-Velocity Impact, Icarus **31**, 277 (1977).
[7] William K. Hartmann, Terrestral, Lunar, and Interplanetary Rock Fragmentation, Icarus **10**, 201 (1969).
[8] C. R. Chapman, W. J. Merline, P. Tamblyn, B. Bierhaus, and S. Brooks, Galileo Research, http://www.boulder.swri.edu/clark/galileo.html.
[9] Charles D. Ferguson, W. Klein, and John B. Rundle, Long-Range Earthquake Fault Models, Computers in Physics **12**, 34 (1998).