Multiple origins of the Newcomb-Benford law: rational numbers, exponential growth and random fragmentation

Peter Ryder

Institut für Festkörperphysik, Universität Bremen, D-28334 Bremen, Germany
ryder@ifp.uni-bremen.de

The Newcomb-Benford law states that, in data drawn randomly from many different sources, the probability that the first significant digit is n is given by $\log(1 + 1/n)$. In a previous paper [1], it was shown that there are at least two basic mechanisms for this phenomenon, depending on the origin of the data. In the case of physical quantities measured with arbitrarily defined units, it was shown to be a consequence of the properties of the rational numbers, whereas for data sets consisting of natural numbers, such as population data, it follows from the assumption of exponential growth. It was also shown that, contrary to what has been maintained in the literature, the requirement of scale invariance alone is not sufficient to account for the law. The present paper expands on [1], and it is shown that the finite set of rational numbers to which all measurements belong automatically satisfies the requirement of scale invariance. Further, a third mechanism, termed "random fragmentation", is proposed for natural number data which are not subject to exponential growth. In this case, however, the Newcomb-Benford is only approximately reproduced, and only under a certain range of initial conditions.

1 Introduction

1.1 History: Newcomb and Benford

Simon Newcomb was born on March 12th, 1835 in Nova Scotia, Canada, but migrated at a early age to the United States, where he became a distinguished astronomer. In a paper published in 1881 [2], when he was director of the American Nautical Almanac Office in Washington, he reported his observation that in large sets of empirical data taken from a wide variety of sources — in his case astronomical data — the first significant figures are not distributed evenly among the digits 1 to 9, as most people would probably intuitively expect, but the digit 1 occurs most frequently as the first significant digit, and the frequency decreases continuously from 1 to 9. His attention was drawn to this phenomenon by the fact that booklets with tables of logarithms showed more signs of wear on the first pages than on later ones.

On the basis of a reasoning which is discussed in more detail below, Newcomb concluded that the numbers are distributed uniformly on a logarithmic scale. In that case, refering to figure 1, the probability that the first significant figure is 1 is given by $AB/AJ = (\log(2) - \log(1))/(\log(10) - \log(1)) = \log(2)$. The probabiity for 2 is BC/AJ, for 3 CD/AJ, and so forth.

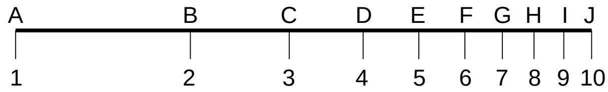

Figure 1: If the numbers are distributed evenly on a logarithmic scale, the probability that the first significant figure is 1 is given by AB/AJ = (log(2) − log(1))/(log(10) − log(1)) = log(2).

.

Thust the probability P_n that the first significant digit is n is given by

$$P_n = \frac{\log(n+1) - \log(n)}{\log(10) - \log(1)} = \log(1 + 1/n). \tag{1}$$

The probabilities, or frequencies, expressed in percent, given by equation (1) are shown in table 1. It is seen that 1 appears as the first significant digit in more than 30% of the numbers, and 9 in less than 5%.

Table 1: Distribution of the first significant digits according to equation (1).

Digit:	1	2	3	4	5	6	7	8	9
%:	30.10	17.61	12.49	9.69	7.92	6.69	5.80	5.12	4.58

Newcomb did not give any empirical data in support of his law, and no such data was apparently published until 1938, when Benford [3] published a long paper on the subject, in which he demonstrates that (1) is more or less closely followed by data from many different sources: geographical data, populations, death rates, street addresses, molecular weights, specific heats, physical constants, etc. He also shows that some mathematical series, such as $1/n$ or $\sqrt{n}$ ($N \in \mathbb{N}$) have a tendency to follow the same distribution of the first significant digits. In some cases, however, the precise nature of the data is not clear ("Pressure Lost, Air Flow", for example), and the sources are not cited[1].

1.2 Two examples

Before looking into the possible reasons for the existence of the uneven distribution of the first digits, we discuss briefly two simple examples. The first is taken from data provided by the US National Institute for Science and Technology (NIST). On NIST's web site you can find, amongst a plethora of other data, the latest values of the fundamental physical constants (such as Planck's constant, the gravitation constant, the electron mass, the elementary charge, etc.). The complete list[2] has 318 entries, with values stretching over 115 orders of magnitude and involving all kinds of units. A comparison of the distribution of the first significant digits of these constants with the data from table 1 is shown in figure 2. The agreement is very good, especially considering the fact that the sample is relatively small.

[1] In fact, neither Benford nor Newcomb give any references!

[2] http://physics.nist.gov/constants

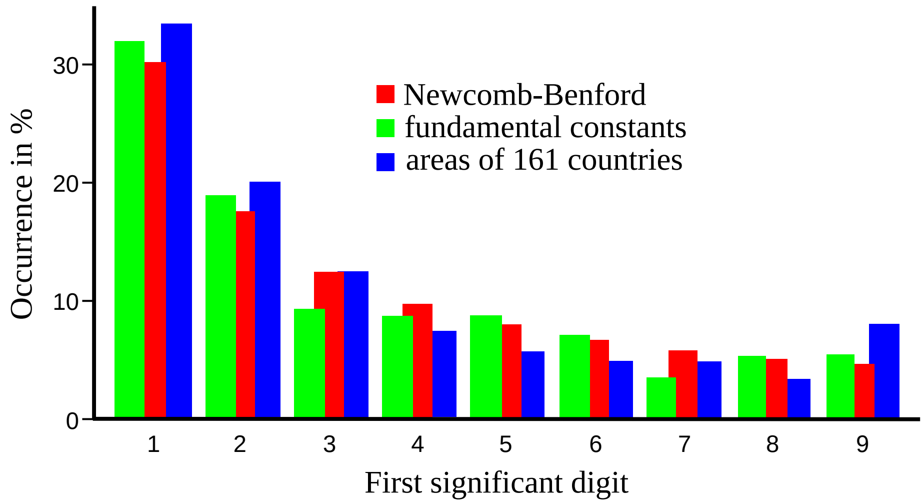

Figure 2: The distribution of the first significant digits of 318 fundamental physical constants (green) and the surface areas of 161 countries (blue) compared to the Newcomb-Benford Law (red).

The second example concerns the surface areas (in km^2) of 161 countries, also plotted in figure 2. These figures show the same tendency, though with slightly more statistical scatter than the data from the list of physical constants.

1.3 Previous explanations

Newcomb's explanation [2] starts with the observation: "As natural numbers occur in nature, they are to be considered as ratios of quantities." It is obvious that he does not use the term "natural numbers" in the strict mathematical sense (the positive integers). The sentence must be taken to mean, in modern terminology, that the results of physical measurements are *rational* numbers. He assumes further that very many repeated divisions are required: "To solve the problem we may form an indefinite number of such ratios, taken independently, [...] and continue the process so as to find the limit towards which the probability approaches." The reason why the divisions must be repeated indefinitely is not quite clear. He also points out that, since multiplication with a power of ten does not affect the first significant digit, we can treat all numbers as being in one decade, e.g. between 1 and 10. This decade, which he visualizes as a circle, will fill up as the divisions are repeated, and then "It is evident that [...] the fractions will approach to an equal distribution around the circle". This is also not clearly argued: the density will increase, but he has not proved that it is constant over the circle. He tacitly assumes a logarithmic scale along the circumference of the circle, but if one postulates a linear scale, it is hard to see why the same argument should not still apply, and the density cannot be constant both on a linear and on a logarithmic scale. However, as we will see in section 3.2, the formation of ratios in the representation of physical measurements does in fact appear to be the fundamental cause of the Newcomb-Benford law.

Pinkham [4] claims that the Newcomb-Benford law is a direct consequence of the requirement of **scale invariance**, but, as we will show in the next section, this requirement alone is not sufficient. Hill [5] discusses the consequences of scale and base invariance and uses formal probability theory to show that "if distributions are selected at random [...] and random samples are then taken from these distributions, the significant digits of the combined sample will converge to the logarithmic (Benford) distribution". Benford [3] concludes that "[the logarithmic law] can be interpreted as meaning that [natural phenomena] proceed on a logarithmic or geometric scale".

2 The mathematics of the Newcomb-Benford Law

We first determine the density function $f(x)$ on a linear scale which leads to a constant density on a log scale. Obviously, to give a constant density on a log scale, $f(x)$ must satisfy

$$\int_A^B f(x)\mathrm{d}x = (F(B) - F(A)) = c(\log(B) - \log(A)) = c'(\ln(B) - \ln(A)),$$

where c and c' are constants, and $f(x) = \mathrm{d}F(x)/\mathrm{d}x$. Thus $f(x)$ is proportional to the derivative of $\ln(x)$, i.e. $1/x$:

$$f(x) = k/x, \qquad k = \text{constant}.$$

This means that the three properties (a) even distribution on a log scale, (b) $1/x$ distribution on a linear scale and (c) the distribution of the first digits given by (1) are mathematically equivalent.

The Newcomb-Benford Law has been empirically verified for a large number of data sources (for a concise review and bibliography see e.g. Weinstein [6]). However, many special distributions, such as the size distributions of crushed rocks [7], follow other laws, even if they appear to be "random". The best agreement is found with large data sets drawn from many different sources.

The Newcomb-Benford law should be scale invariant, since any universal law must be independent of the units of measurement used, otherwise it would not be universal. Pinkham [4] claims that (1) is the *only* function for the digit probability which satisfies this criterion. This is equivalent to the assertion that the $1/x$ density distribution follows from the requirement of scale invariance. If the function $f(x)$ remains completely unchanged when the units are changed, this means that, for any constant α, it must satisfy the condition

$$f(x)\mathrm{d}x = f(\alpha x)\mathrm{d}(\alpha x) \quad \Rightarrow \quad f(\alpha x) = f(x)/\alpha. \tag{2}$$

Let us assume a polynomial form $f(x) = \sum_i a_i x^i$ for the density function. Inserting this in the above equation gives

$$\sum_i a_i \left(\alpha^i - 1/\alpha\right) x^i = 0 \qquad \text{for all } x.$$

This holds only when all $a_i = 0$ for all i except $i = -1$.

This confirms Pinkham's result in a strict sense. However, one may question the validity of equation (2) as the requirement of scale invariance. In practice, the absolute value of the density function cannot be determined from a finite sample. Suppose we have a set of M numbers ranging from n_1 to n_2. The density function must satisfy the condition

$$\int_{n_1}^{n_2} f(x)\mathrm{d}x = M.$$

This means that the function $f(x)$ must carry a constant factor, which is determined by the size and range of the sample and is consequently not universal. We must therefore modify equation (2) as follows:

$$f(x)\mathrm{d}x = Af(\alpha x)\mathrm{d}(\alpha x) \quad \Rightarrow \quad f(\alpha x) = f(x)/(A\alpha), \quad A = \text{const.}$$

and the condition for the polynomial solution is now

$$\sum_i a_i \left(\alpha^i - 1/(A\alpha)\right) x^i = 0 \qquad \text{for all } x.$$

This is satisfied for *any* digital value $i = j$ with $a_{i \neq j} = 0$ and $A = \alpha^{-(i+1)}$, so that the requirement of scale invariance is satisfied, for all practical purposes, by any density function of the form

$$f(x) = kx^j, \quad k = \text{const.}, \quad i \in \mathbb{Z}. \tag{3}$$

Thus any simple power law x^j, not just $1/x$ satisfies the requirement of scale invariance. Note that one possibility is, with $j = 0$, $f(x) = k$, i.e. an even distribution on a linear scale. Of course only $j = -1$ leads to the Newcomb-Benford law.

Interestingly, all other possible values of j lead to a definite distribution of the first significant digits, as we can show by integrating within any decade. For $j \neq -1$, the probability that the first significant digit is n is given by:

$$P_{jn} = \frac{\int_{n\cdot 10^m}^{(n+1)\cdot 10^m} x^j}{\int_{10^m}^{10^{m+1}} x^j} = \frac{(n+1)^{j+1} - n^{j+1}}{10^{j+1} - 1},$$

which is the same for all decades.

Starting from the $1/x$ distribution function it is also possible to deduce the relative frequencies of the 2nd, 3rd, etc. digits, which of course include 0, and also the conditional probabilities (e.g. the probability that the second digit will be m when the first is n). Further, the corresponding probabilities can also be given for other number systems, e.g. octal or hexadecimal.

To take first a simple example, the probability P_{31} that the first digit is 3 and the second 1 is equal to the probability that the number is in the interval $(31, 32[$ or any power of 10 of this interval, i.e.

$$P_{31} = \frac{\int_{31}^{32} \mathrm{d}x/x}{\int_{10}^{100} \mathrm{d}x/x} = \log\left(1 + \frac{1}{31}\right).$$

This may be generalized to any number of digits and any number system as follows: let $\{D\}$ be the sequence of digits of the number system with the base B which represents the

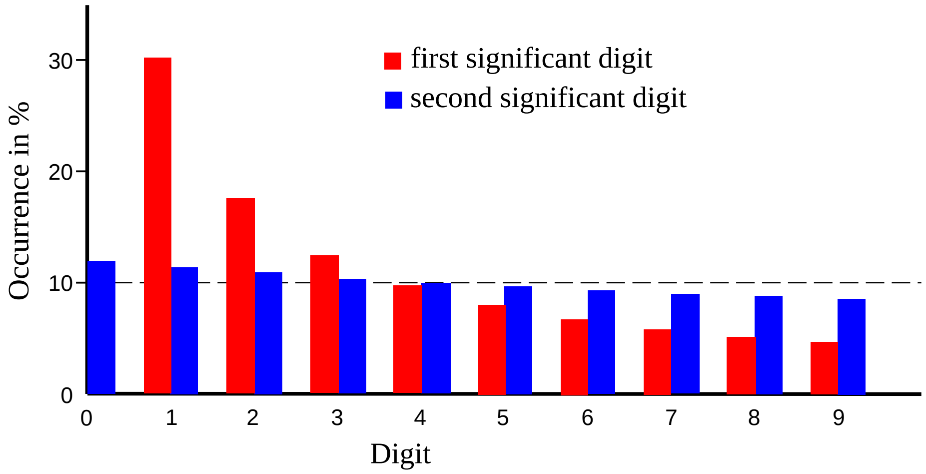

Figure 3: The distribution of the first and second significant digits according to the Newcomb-Benford Law. The distribution of the second digits differs only slightly from a uniform distribution (dashed line).

natural number D. Then the probability $P_{\{D\}}$ that the number starts with the sequence $\{D\}$ is

$$P_{\{D\}} = \log_B\left(1 + \frac{1}{D}\right).$$

The probability that the second significant digit of a number in the decimal system is k is given by

$$P_k^{(2)} = \sum_{i=1}^{9} \log\left(1 + \frac{1}{10i + k}\right) \quad (k \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}). \tag{4}$$

Figure 3 shows a comparison of the relative frequencies of the first significant digits from equation (1) with those of the second digits calculated from equation (4). It is seen that the deviation from a uniform distribution (indicated by the dashed line in the figure) is much less marked for the second digits. Further calculations show that the distribution approaches uniformity more and more as we go to third, fourth etc. digits.

When we wish to test the validity of the Newcomb-Benford law for any given set of data, it is therefore best to use the distribution of the first significant digits, because this gives the best statistics. Looking at the distribution of the remaining digits does not bring any further useful information. On the other hand, when we are looking for the theoretical basis of the Newcomb-Benford law, it is best to show how the $1/x$ distribution comes about, because this is universal and independent of the number base used .

3 The mechanisms behind the Newcomb-Benford law

3.1 The nature of physical measurements

Let us look first at a simple example of a physical measurement: the determination of the mass of an object with a laboratory balance. The result of a weighing experiment is a reading of some physical quantity x, e.g. the extension of a spring, the deflection of a pointer, etc., which is proportional to the weight and hence to the mass of the body. Then we must compare x to the reading $x_\circ$ obtained for a unit mass (1 kg). The mass of the body in the chosen units is then given simply by $m = x/x_\circ$, whereby x and $x_\circ$ are rational numbers. The number of significant digits of each of these numbers is, however, limited to some finite value due to the fact that every physical measurement is subject to an unavoidable experimental error. Both numbers may therefore be converted to *natural* numbers by multiplying with a suitable integral power of ten:

$$m = \frac{x}{x_\circ} = \frac{p}{q}.10^n \quad p, q \in \mathbb{N}, n \in \mathbb{Z}. \tag{5}$$

The factor 10^n does not affect the first significant digit, so we only need to consider the properties of the fraction p/q. This is of course a rational number. However, due to the above-mentioned limits to the experimental accuracy, neither p nor q can be indefinitely large. Thus all possible numbers p/q which may appear as the result of a physical measurement belong to a *finite subset* of the set of rational numbers $\mathbb{Q}$.

It might be argued that, in day-to-day practice, weight measurements are not carried out in the way described above: to find my own body mass, for example, I just step onto the bathroom scales and read off the value on an analogue scale or a digital indicator. However, this not alter the basic principle: it is just that the calibration – the weighing of the standard mass – has already been carried out by the manufacturer and is implicitly contained in the reading.

Further, it is clear that the vast majority of physical measurements are not obtained by such a simple process as the weighing example, but by indirect methods involving several measurements. For example, we cannot "weigh" the earth or an electron, but we can determine the mass of the earth from satellite data or the mass of the electron from the deflection of electron beams in electric or magnetic fields. To take the satellite example, the mass of the earth can be calculated from the formula

$$M = \frac{4\pi^2 a^3}{GT^2},$$

where a is the semi-major axis of the orbit, T the orbital period and G the gravitational constant. Putting the values into this equation, we again have a division of rational numbers which may be reduced to the form of equation (5).

3.2 The Newcomb-Benford law as consequence of the properties of rational numbers

We now investigate the properties of the finite subset S of the rational numbers. A convenient way to define this set is to use the matrix shown in figure 4, which may be employed to show that the set of rational numbers is countable.

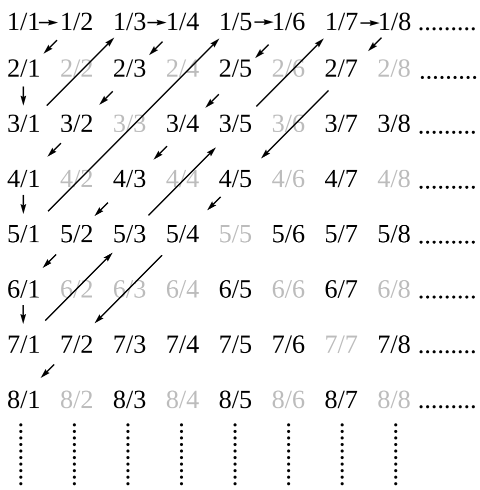

Figure 4: The rational numbers arranged in a matrix with p/q in row p and column q. Starting in the top left corner and following the path along the diagonals as indicated by the arrows, each rational number may be assigned to a natural number. The fractions shown in gray are reducible and must therefore be omitted in the process.

The numbers are arranged so that p/q is in the pth row and the qth column. If we start in the top left corner (1/1) and follow the path indicated in the figure along the diagonals, we can "count" the rational numbers, i.e. associate each rational number with a natural number. Each diagonal is defined by $p+q=N, (N \in \mathbb{N})$. Thus, if we stop the process at the end of a diagonal, the resulting finite subset S of $\mathbb{Q}$ is defined as

$$\mathsf{S} = \{x | x = p/q, p \in N, q \in \mathbb{N}, p+q \leq N\}. \tag{6}$$

It is clear from this definition that, if p/q is an element of S, then so is its reciprocal value q/p, i.e. S has the symmetry property

$$x \in \mathsf{S} \leftrightarrow 1/x \in \mathsf{S}. \tag{7}$$

It is then obvious that the numbers cannot be distributed evenly on a linear scale: the first number in the list is 1, and from the rest half are between 0 and 1, and the other half between 1 and some large number $N-1$.

Let x_1 and x_2 be two elements of the set S. Then it follows from the above that $1/x_1$ and $1/x_2$ are also elements of S. Further, for each x between x_1 and x_2 there will a reciprocal number $1/x$ between $1/x_1$ and $1/x_2$. Hence the number of elements of S in the interval $[x_1, x_2)$ is equal to the number in the interval $(1/x_2, 1/x_1]$. In the limit of large numbers, this means that the density function $f(x)$ for the set S must have the property

$$f(x)\mathrm{d}x = -f(1/x)\mathrm{d}(1/x) = f(1/x)\mathrm{d}x/x^2$$

and hence

$$f(1/x) = x^2 f(x) \tag{8}$$

This equation is satisfied by any functions of the form

$$f(x) = \frac{1}{x} \sum_p A_p (x^p + x^{-p}) \qquad p \in \mathbb{N}.$$

However, the only function which satisfies both the requirement (3) for scale invariance and equation (8) is obtained by putting all $A_p = 0$ except for $p = 0$, i.e.

$$f(x) = k/x, \qquad k = \text{const.} \tag{9}$$

Thus the symmetry property (8) and the scale invariance requirement (2) together lead directly to the Newcomb-Benford law.

Table 2: The distribution of first digits in the finite set S, defined by equation (6) with $N = 1000$ and $N = 10000$ compared with the Newcomb-Benford law (occurrences in %).

First digit	1	2	3	4	5	6	7	8	9
S $(N = 1,000)$	30.16	17.54	12.44	9.71	7.87	6.65	5.87	5.21	4.56
S $(N = 10,000)$	30.19	17.52	12.43	9.66	7.92	6.70	5.84	5.15	4.59
Equation (1)	30.10	17.61	12.49	9.69	7.92	6.69	5.80	5.12	4.58

In order to test this result on special sets of the form S, a program was written to determine the relative frequencies of the first significant digits for all rational numbers p/q with $p + q \leq N$ for any given N. The results for $N = 1,000$ and $10,000$ are shown in table 2. The sets contain 164,903 and 12,225,985 numbers, respectively. The agreement with the Newcomb-Benford law is very good — in any case better than for any set of real data. The root mean square deviation from the Newcomb-Benford values is approximately 0.0005 in both cases.

It would thus appear that the set S defined by (6) already in itself fulfils the requirement of scale invariance, so that we do not need to introduce this as an extra condition. This is in fact the case, as can be shown by considering what happens when the units are changed. As explained above, the result of any measurement can be reduced to the form $(p/q)10^n$, where the rational number p/q is an element of the set S. A change of units will mean that the result of the measurement is multiplied by some rational number α, giving a new number, which can again be reduced to a quotient of two natural numbers and a power of ten:

$$\alpha(p/q)10^n = (p'/q')10^{n'}.$$

However, due to the limited experimental accuracy, the number p'/q' is subject to the same restrictions as p/q and hence is also an element of the set S. Applying this to *all* the elements of the set S, each is replaced by another number which is also an element of the set, so that the set itself remains unchanged.

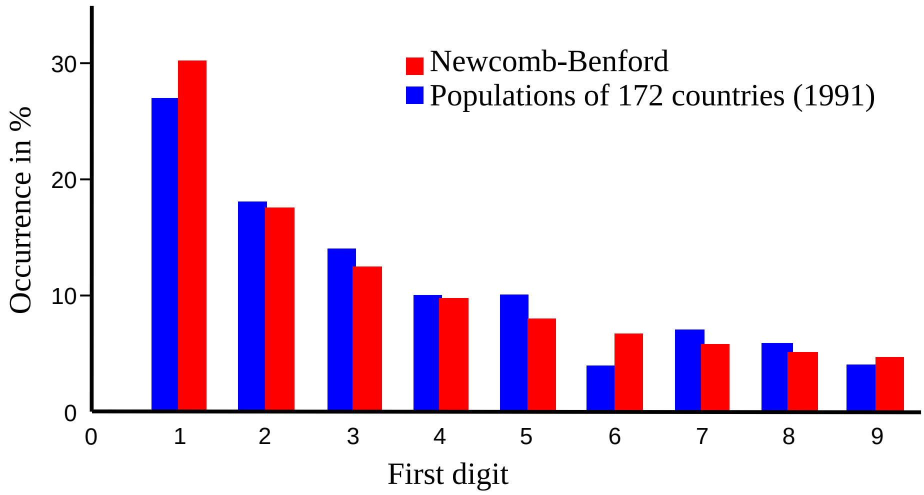

Figure 5: The relative frequency of the first significant digit in the populations of 172 countries (1991) compared to the Newcomb-Benford law.

3.3 Population data and exponential growth

Figure 5 shows the frequencies of the first significant digits in the population figures for 172 countries (taken from the statistics for the year 1991). There is good agreement with the Newcomb-Benford law, although the explanations given in the previous sections cannot be applied to this case: population figures are natural numbers, not ratios, and the argument of scale invariance is not relevant, because there are no arbitrarily defined units. There must therefore be some other explanation for the validity of the Newcomb-Benford law in this case.

A possible explanation is that populations tend to grow exponentially, at least over a certain period of time, i.e. the population X at time t is given by $X(t) = X_\circ \exp(t/\tau)$, where $X_\circ$ and τ are constants for any given population. We may invert this equation to find the time at which the populations reaches a certain value:

$$t = \tau \ln(X/X_\circ).$$

The time required for the population to grow from X to $X + \mathrm{d}X$ is

$$\mathrm{d}t = \tau \mathrm{d}X/X.$$

If we now consider a large set of independent populations with different time constants τ_i, $i = 1, 2, 3, \ldots$, the probability that any one population will be in the interval $(X, X + \mathrm{d}X)$ at any randomly chosen time is proportional to $\tau_i \mathrm{d}X/X$, and the mean number of population figures in this interval is

$$f(X)\mathrm{d}X \sim \sum_i \tau_i \mathrm{d}X/X \qquad \Rightarrow \qquad f(X) = k/X \quad (k = \text{const.}).$$

Thus we have again the $1/x$ density function, which, as we have seen above, is the basis of the Newcomb-Benford law.

It might be argued that the populations of many countries, for example those in Western Europe, are not growing exponentially, so that the above argument does not apply. However, as we shall see in the next section, natural-number data sometimes obey the Newcomb-Benford law even if they are not subject to exponential growth.

3.4 The Newcomb-Benford Law and Election Votes

On September 27, 2009, elections were held in Germany for the new parliament (Bundestag). Detailed results can be found, for example, on the web pages of the Federal Statistical Office. The following discussion concerns data from tables showing the number of valid votes in each of the 16 federal states and for each state the distribution of these votes amongst the different political parties. For the present analysis, only the five parties which qualified for seats in the new parliament were taken into account. The total number of valid votes was 81,691,389 (each person has 2 votes).

When the frequencies of the first significant digits of the numbers appearing in these tables are calculated, the result shown in figure 6 are obtained. It is seen that there is fairly good agreement with the Newcomb-Benford law, but we cannot explain this with the models presented so far. If the population of Germany were growing exponentially — and with it the number of voters — the model of the previous section would be valid, but this is not the case. The population is at present stagnating, or even declining. We must therefore look for another explanation in this case.

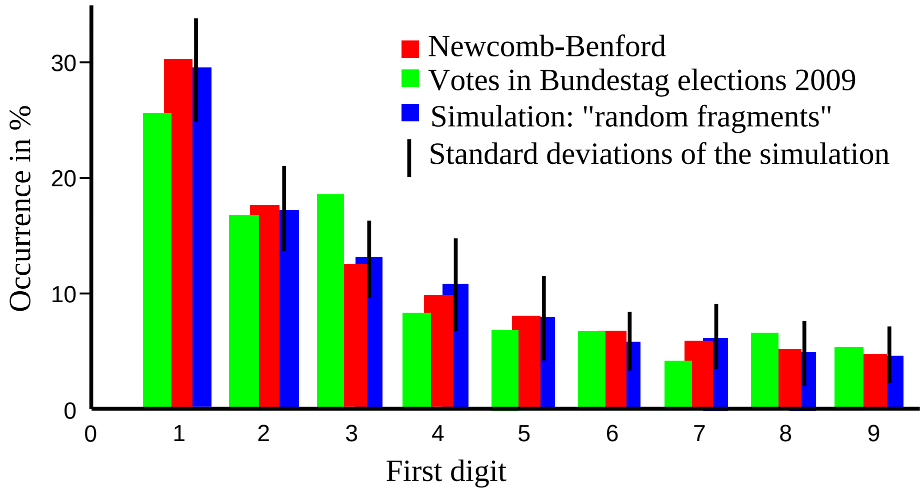

Figure 6: The relative frequency of the first significant digit in the voting results of the 2009 German elections, compared with the Newcomb-Benford law and the "random fragmentation" simulation (see text).

A possible explanation may be found in the well-known fact that, when objects, or numbers, are broken down into fragments by some random process, the small objects are more numerous than the large ones. For example, the number of grains of sand on a beach is much greater than the number of pebbles. It is not easy to calculate theoretically the expected size distribution for such a process. A simple numerical simulation was therefore performed to see if such a stochastic fragmentation could, at least in principle, lead to something like the Newcomb-Benford distribution. The following model, which we will call the “random fragmentation” model for convenience, was used.

Starting with some large number N, representing for example the total number of votes in a country, the first step is to split this number randomly into two parts, whereby all possible results $N, N-m$, with $0 < m < N$, are considered to have equal probability. In the computer program this was achieved by creating a random number in the interval $(1, N-1)$ and subtracting it from N. This process is repeated with the set of two numbers $N, N-m$, producing a set of four numbers, and so on. If the number 1 occurs, it is of course left unchanged. To simulate approximately the situation of the German elections results, N was set to 82,000,000, and the fragmentation procedure was repeated 6 times, giving finally a set of 64 numbers. The whole process was repeated 20 times, order to get an idea of the statistical scatter. The results of the simulation — the mean values and standard deviations — are plotted in figure 6. There is good agreement with both the election results and the Newcomb-Benford distribution.

In order to see if better agreement with the Newcomb-Benford law could be obtained with better statistics, the means of 100,000 repetitions, with the same initial data, were calculated. The results (see table 3) do indeed show very good agreement. However, these results cannot be generalized, as will be explained in the following section.

Table 3: The distribution of first digits in the mean of 100,000 repetitions of the random fragmentation model, compared with the Newcomb-Benford law (occurrences in %).

First digit	1	2	3	4	5	6	7	8	9
Random fragments	30.06	17.59	12.50	9.68	7.94	6.71	5.81	5.13	4.57
Equation (1)	30.10	17.61	12.49	9.69	7.92	6.69	5.80	5.12	4.58

4 Discussion and conclusions

The results presented in this paper show clearly that there is no single universal explanation for the Newcomb-Benford law, but several mechanisms, depending on the nature and origin of the data. It is obvious, for example, that physical constants have nothing to do with exponential growth, and population figures are not ratios of rational numbers.

In the case of physical measurements expressed as rational numbers with limited accuracy, the properties of the finite set S defined by (6) account fully for the observed distribution of the first significant digits and its scale invariance; it is not necessary to introduce scale invariance as an additional requirement. The maximum natural number N in the definition of S was not quantified. However, table 2 shows that the distribution of the first significant digits does not depend sensitively on the size of the set. There is some finite value of N for which the set includes all possible results of a physical measurements, apart from power-of-ten factor which is not relevant for this discussion.

The results of the random fragmentation model depend strongly on the choice of the initial conditions, i.e. the number N and the number of fragmentation steps. This is clear from the following considerations: After the first step, repeated many times for averaging, we have a uniform distribution, because all possible divisions are equally probable. On the other hand, if the process is continued indefinitely, we will eventually end up with nothing but 1s, because the number 1 cannot be split up further. The distribution of the first significant figures will then give 100% for 1 and 0% for all other digits. Thus, in the process, we will pass through intermediate distributions, some of which may resemble the Newcomb-Benford distribution. This is illustrated in figure 7 by means of a concrete numerical example. Starting from the number 1,000, the fragmentation process was carried to 1, 2, 3, 4 and 5 steps, and in each case the mean values were calculated from 1000 repetitions of the whole process. The results are shown in figure 7.

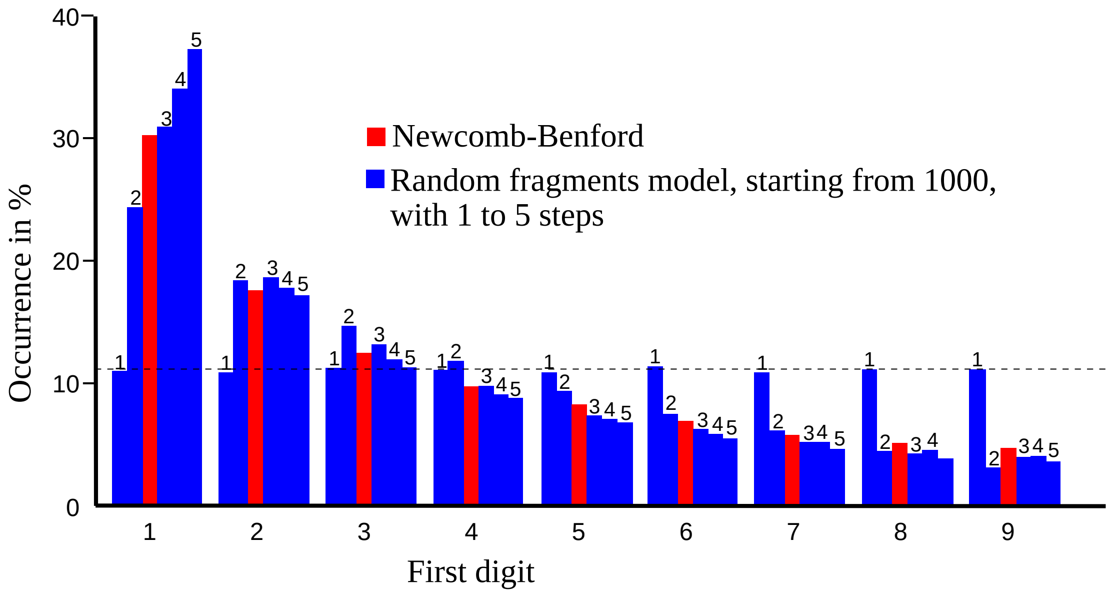

Figure 7: Comparison of the results of the random fragmentation model, starting from 1000 and using different numbers of fragmentation steps, with the Newcomb-Benford distributions

The distribution obtained after the 3rd step agrees best with the Newcomb-Benford distribution; further fragmentation steps deviate more an more from this distribution, as can be seen best in the distributions of the number 1. It is therefore questionable whether the voting example is in fact an instance of the Newcomb-Benford law. The fragmentation process leads to a monotonically decreasing size function, but this is not necessarily the $1/x$, so that the resulting distribution of the first significant digits does not follow (1) exactly, but is only similar to it, in that the frequency decreases monotonically from 1 to 9. This view is supported by experiment evidence published by Kreiner [7], who investigated the size distribution of rock particles obtained by mechanical crushing and found significant deviations from the Newcomb-Benford law.

References

[1] Peter Ryder, Z. Naturforsch. **64a,** 615 (2009)

[2] Simon Newcomb, Am. J. Math. **4**, 39 (1881)

[3] Frank Benford, Proc. Am. Phil. Soc. **78**, 551 (1938)

[4] Roger S. Pinkham, Ann. Math. Stat. **32**, 1223 (1961)

[5] Theodore P. Hill, Statistical Science **10**, 354 (1995)

[6] Eric W. Weinstein, http://mathworld.wolfram.com/BenfordsLaw.html

[7] W. A. Kreiner, Z. Naturforsch. **58a**, 618 (2003)