The Relationship Between the Newcomb-Benford Law and the Distribution of Rational Numbers

Peter Ryder

Institut für Festkörperphysik, Universität Bremen, 28334 Bremen, Germany

Reprint requests to Prof. P. R.; E-mail: ryder@ifp.uni-bremen.de

Z. Naturforsch. **64a,** 615 – 617 (2009); received November 20, 2008 / revised January 26, 2009

The Newcomb-Benford law, also known as Benford's law or the first-digit law, applies to many tabulated sets of real-world data. It states that the probability that the first significant digit is n, ($n \in \{1,2,3,4,5,6,7,8,9\}$) is given by $\log(1+1/n)$. The law has been verified empirically with widely differing data sets. In the present paper it is shown that it does not necessarily follow from the requirement of scale invariance alone, as has been claimed. This condition is necessary, but not sufficient. In addition, it is necessary to consider the properties of certain finite subsets of the set of rational numbers.

Key words: Newcomb; Benford; First Digit Law; Rational Numbers.

In 1881 the astronomer Newcomb [1] published a short paper in which he reported the empirical fact that, in data drawn from astronomical measurements, the first significant figures are not distributed equally among the digits 1 to 9. He showed that his data agreed well with the assumption that the figures are distributed equally on a logarithmic scale, so that the probability, the first significant digit of a number in the interval $[10^m, 10^{m+1})$ is n, is given by

$$P_n = \frac{\log((n+1)\cdot 10^m) - \log(n\cdot 10^m)}{\log(10^{m+1}) - \log(10^m)} = \log(1+1/n). \quad (1)$$

The probability is thus the same for all decades and hence applies to the whole set.

Newcomb's attention was drawn to this apparent anomaly by the fact that booklets with tables of logarithms, which he and his colleagues used for their calculations, showed more signs of wear on the first pages than on later ones[1]. He attempts to justify the logarithmic distribution by assuming that "numbers occurring in nature are to be considered as ratios of quantities" and by considering what happens when, starting from a set of (rational) numbers with an arbitrary distribution, one forms ratios of these numbers, then ratios of the ratios, and so on indefinitely. He claims that this process will eventually result in an even distribution of the numbers thus formed on a logarithmic scale, whatever the distribution of the starting set. However, his proof is not rigorous (the key sentence begins with "It is evident that…"), and it is not clear why the numbers occurring in nature should be the result of indefinitely repeated divisions.

A paper published in 1938 by Benford [2] begins, like Newcomb's, with a remark on the state of well-used logarithmic tables[2]. It is remarkable that both this 21-page article and Newcomb's short paper are devoid of references, so it is not clear whether Benford was aware of Newcomb's paper or not. Benford gives many examples of data collected from different sources (geographical data, populations, newspapers, physical properties, and constants, etc.) which follow the first-digit law closely. He gives a lengthy discussion of the phenomenon and concludes that "[the logarithmic law] can be interpreted as meaning that [natural phenomena] proceed on a logarithmic or geometric scale".

Equation (1) is equivalent to the statement that the density of numbers on a linear scale follows a function of the form $f(x) = k/x$, where k is a constant, since

$$P_n = \frac{\int_{n\cdot 10^m}^{(n+1)\cdot 10^m} k\,\mathrm{d}x/x}{\int_{10^m}^{10^{m+1}} k\,\mathrm{d}x/x} = \frac{\ln(1+1/n)}{\ln(10)} = \log(1+1/n).$$

[1] "That the digits do not occur with equal frequency must be evident to anyone making much use of logarithmic tables and noticing how much faster the first pages wear out than the last ones."

[2] "It has been observed that the pages of a well used table of common logarithms show evidence of a selective use of the natural numbers".

Table 1. Relative frequencies of the first digits in the finite set S compared with the Newcomb-Benford law.

First digit	1	2	3	4	5	6	7	8	9
$p+q \leq 1000$	0.3016	0.1754	0.1244	0.0971	0.0787	0.0665	0.0587	0.0521	0.0456
Equation (1)	0.3010	0.1761	0.1249	0.0960	0.0792	0.0669	0.0580	0.0512	0.0458

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 [3]). However, many special distributions, such as the size distributions of crushed rocks [4], 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 [5] has shown that (1) is the only function for the digit probability which satisfies the strict requirement of scale invariance. However, this condition should be relaxed, because we cannot determine the absolute density, so if the change of units has only the effect of multiplying the density function with a constant, this cannot be detected by the analysis of real data. This means that the density function $f(x)$ must satisfy the condition

$$f(x)\mathrm{d}x = \beta f(\alpha x)\mathrm{d}(\beta\alpha x) \;\Rightarrow\; f(\alpha x) = f(x)/\alpha, \quad (2)$$

where α and β are constants. It is easily shown that this condition is satisfied by any simple power of the form $f(x) = kx^i$, where k is a constant, and i is any integer, not only -1, including 0 (even distribution).

Hill [6] 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".

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. The following discussion is limited to the linear density function, since everything else follows from that.

We first note that the representation of any physical measurement is an element of the set of rational numbers. Since the sign of the number is not important in the present analysis, the discussion may be limited to the positive (non-zero) rational numbers defined by $\mathbb{Q}^+ = \{p/q | p,q \in \mathbb{N}\}$. Clearly, if we disregard any power-of-ten factor, which will not affect the distribution of the first significant digits, there is a limit to the size of p and q for natural numbers which can be represented in print; thus all rational numbers which appear in tables belong to a finite subset S of $\mathbb{Q}^+$, defined e. g. by the condition that both p and q are less than or equal to some (large) natural number N, or $p+q \leq N$. Independently of the precise limits imposed on p and q, S will have the property that

$$x \in \mathsf{S} \leftrightarrow 1/x \in \mathsf{S}, \quad (3)$$

since any pair of natural numbers (p,q) may be used to represent the rational number $x = p/q$ and its reciprocal $1/x = q/p$. We now show that (3), together with (2) leads to a $1/x$ distribution of the elements of S on a linear scale of the rational numbers.

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

$$f(1/x) = x^2 f(x). \quad (4)$$

This does not uniquely define $f(x)$, but if, in addition, we take into account the requirement of scale invariance (2), the only value of i which satisfies (4) is in fact -1. In other words, the Newcomb-Benford law applies to the elements of S.

Of course the agreement is only approximate for any finite set S. Table 1 shows the relative frequencies of

the first digits of all rational numbers in the set S for $p+q \leq 1000$ compared with the theory. The set contains 164,903 numbers (reducible fractions such as 2/4 or 6/3 etc. were not counted). Even for such a relatively small set, the agreement is better than for any empirical data set.

All measurements which have ever been published in print are necessarily elements of a large but finite subset of the rational numbers with the property (3). Thus, when such numbers are selected randomly from many different sources, their distribution is expected to approach the $1/x$ function, which follows from (3). It is therefore concluded the underlying reason for the Newcomb-Benford law is this special property (3) of the rational numbers. The above discussion does not, however, apply to population data, which are also found to obey the Newcomb-Benford law. The reason in this case is most probably the fact that populations tend to grow exponentially.

[1] S. Newcomb, Am. J. Math. **4**, 39 (1881).
[2] F. Benford, Proc. Am. Phil. Soc. **78**, 551 (1938).
[3] E. W. Weinstein, http://mathwrld.wolfram.com/BenfordsLaw.html
[4] W. A. Kreiner, Z. Naturforsch. **58a**, 618 (2003).
[5] R. S. Pinkham, Ann. Math. Stat. **32**, 1223 (1961).
[6] T. P. Hill, Statistical Science **10**, 354 (1995).