THE LOGARITHMIC DISTRIBUTION OF LEADING DIGITS AND FINITELY ADDITIVE MEASURES

THOMAS JECH

In memoriam Zdeněk Frolík

1. Introduction.

When the "1" key on my old computer gave out I was not surprised. That this particular key was first to break was just another manifestation of the long observed phenomenon, namely that more numbers begin with digit 1 than with any other digit. The empirical *logarithmic distribution law* states that for a "randomly chosen" number, the leading digit will be 1 with probability $\log_{10} 2$. In general, the leading digit d occurs with probability $\log_{10}(1 + \frac{1}{d})$, and in fact for any positive integer k, the probability that the decimal expansion of a number begins with k is $\log_{10}(1 + \frac{1}{k})$.

This empirical logarithmic law was first observed and formulated by Simon Newcomb in 1881 [8]: "That the ten digits do not occur with equal frequency must be evident to any one making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones." *"The law of probability of the occurrence of numbers is such that all mantissæ of their logarithms are equally probable."*

A number of papers have been written to deal with this "first digit phenomenon", giving explanations using various summation methods or definitions of probability. On close inspection, all these methods attempt to introduce a (finitely additive) probability measure for which the distribution of leading digits satisfies the logarithmic law.

In this note we state necessary and sufficient conditions for a probability measure to satisfy the first digit law, and discuss various results and proposed explanations in the light of these conditions.

Supported in part by NSF grant DMS–8918299

For a detailed survey of the vast literature on the problem we refer the reader to the 1976 American Mathematical Monthly article [11].

2. Finitely additive probability measures that satisfy the first digit law.

By a *probability measure* on a set E we mean a finitely additive function μ defined on all subsets of E, with the property that $\mu(E) = 1$. If k is a positive integer, we denote D_k the set of all real numbers $x > 0$ that begin with the string of digits k. (The number π begins with 314, .025 begins with 2500, etc.) By $\log\ x$ we mean the common logarithm (base 10), and $\{x\}$ denotes the fractional part of a real number $x \quad (0 \leq \{x\} < 1)$.

Definition. A probability measure μ on a set $E \leq (0, \infty)$ satisfies the *leading digit law* if for every positive integer k,

$$\mu(D_k \cap E) = \log\,(1 + \frac{1}{k}).$$

Note that the condition $x \in D_k$ is equivalent to

$$\{\log\ k\} \leq \{\log\ x\} < \{\log\,(k+1)\}$$

and that $\log\,(1 + \frac{1}{k}) = \log\,(k+1) - \log\ k$.

Theorem. *The following are equivalent, for any probability measure μ on a set E of positive reals:*

(1) *μ satisfies the leading digit law.*

(2) *For any a and b such that $0 \leq a < b \leq 1$,*

$$\mu(\{x \epsilon E : a \leq \{\log\ x\} < b\}) = b - a.$$

(3) *For any Riemann-integrable function f on $[0, 1)$,*

$$\int f(\{\log\ x\}) d\mu = \int_0^1 f(t)dt.$$

(4) *For any integer $m \neq 0$,*

$$\int\ e^{2\pi i\, m\ \log\ x} d\mu = 0.$$

Proof. If μ satisfies the leading digit law and if $a = \{\log\ k\} < b = \{\log(k+1)\}$ then $\mu(D_k \cap E) = b - a$. Since every interval $[a, b)$ with $0 \le a < b \le 1$ contains a subinterval of the form $[\{\log\ k\}, \{\log(k+1)\})$, condition (2) follows.

(3) follows from (2) by standard methods.

To see that (3) implies (4), notice that when $m \neq 0$, we have

$$\int e^{2\pi i m \log\ x} d\mu = \int e^{2\pi i m \{\log x\}} d\mu = \int_0^1 e^{2\pi i m t} dt = 0.$$

To prove that (4) implies that μ satisfies the leading digit law, it suffices to show that if $0 \le a < b \le 1$ then (3) holds for the characteristic function $\chi_{[a,b)}$ of the interval $[a, b)$. As $\chi_{[a,b]}$ can be approximated by continuous functions (i.e. $f < \chi_{[a,b)} < g$ such that $\int_0^1 (g - f)dt < \varepsilon$), it suffices to prove (3) for every continuous function, periodic mod 1.

Thus let f be such a function. For every $\varepsilon > 0$ there exist trigonometric polynomials $\varphi(t) = a_0 + (a_1 \cos 2\pi t + b_1 \sin 2\pi t) + \cdots + (a_m \cos 2\pi m\, t + b_m \sin 2\pi\, m\, t)$ and $\psi(t) = c_0 + \ldots$ such that $\varphi < f < \psi$ and $\psi - \varphi < \varepsilon$. Clearly, $\int_0^1 \varphi(t)dt = a_0$ and $\int_0^1 \psi(t)dt = c_0$, while by (4), $\int \varphi(\{\log x\})d\mu = a_0$ and $\int \psi(\{\log\ x\})d\mu = c_0$. As $c_0 - a_0 < \varepsilon$, it follows that $\int f(\{\log\ x\})d\mu$ is equal to $\int_0^1 f(t)dt$.

3. Examples.

We start with the case when x is a continuous variable, the simplest instance being when E is the interval $E = [1, 10)$. Let ν be any finitely additive extension of the Lebesgue measure on $[0, 1)$, and let

$$\mu(X) = \nu(\log X) = \nu(\{\log\ x : x \in X\}) \qquad \text{for any } X \subseteq E.$$

Clearly, μ satisfies condition (2) and thus the leading digit law. This case admits a probabilistic interpretation, as μ can be regarded as probability (on $[1, 10)$) whose distribution is the function $F(x) = 10^x$; for every Lebesgue measurable $X \subseteq E$ we have $\mu(X) = \int_X dF$.

If large cardinals exist in the set theoretic universe (namely the real-valued measurable kind) then μ can even be found which is countably additive, rather than just finitely additive.

It should be noted that this case corresponds most closely to Newcomb's formulation that "all mantissæ of their logarithms are equally probable"; we can also interpret the distribution F as distance on the slide rule [10].

Staying with the continuous case, consider now the space $E = (0, \infty)$ of all positive reals. Instead of the measure μ on $(0, \infty)$, consider a measure ν on $\mathbf{R}$, and let

$$\mu(X) = \nu(\log\ X) \qquad\qquad (X \subseteq E);$$

ν is a finitely additive probability measure on $\mathbf{R}$ and we use the notation $\mu = \log^{-1}(\nu)$.

Condition (2) imposed on μ becomes

$$\nu(\bigcup_{m \in \mathbf{Z}} [m + a, m + b)) = b - a \qquad\qquad (0 \leq a < b \leq 1). \tag{5}$$

Measures that satisfy (5) exist. A sufficient condition for (5) is for instance the condition that ν is translation invariant (which is equivalent to the condition of "scale invariance" of μ stated in [9]). If this is the case then $\nu[m, m + 1) = 0$ for every m and so μ cannot be countably additive (and neither can it result from a distribution function).

We shall now address the discrete case, namely when E is enumerated by an increasing sequence $\{a_1, a_2, \ldots, a_n, \ldots\}$, the simplest case being $E = \mathbf{N}$.

Much of the appeal of the first digit problem owes to the fact that the set $D_k \cap \mathbf{N}$ do not have (assymptotic) density . The *density* of a set X is the arithmetic, or Cesàro, mean (if it exists) of the characteristic function χ_X of X:

$$\delta(X) = \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{k=n} \chi_X(k).$$

The upper density of D_1 is $\frac{5}{9}$ and that lower density is $\frac{1}{9}$ (to see this, consider the first 2,000 or the first 10,000 numbers).

That the sets D_k do not have density is equivalent to the fact that $\log n$ is not uniformly distributed mod 1. More generally [4], the limits

$$lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{j=n} \chi_{D_k}(a_j) \tag{6}$$

exist if and only if the sequence $(\log a_n : n = 1, 2, \dots)$ is uniformly distributed mod 1, and condition (4) becomes the well known Weyl criterion

$$(7) \qquad lim_{n\to\infty}\ \frac{1}{n}\sum_{j=1}^{j=n} e^{2\pi\, i\, m\ \log a_j} = 0.$$

If (a_n) is a geometric sequence $(a\,q^n)$ with $q > 1$ and $\log q$ irrational, then (7) is easily seen to hold for every integer $m \neq 0$, so the limits (6) exist. (This is of course related to Weyl's theorem [13] that the sequences $(n\alpha)$ with irrational α are uniformly distributed mod 1.) The literature abounds in examples of sequences for which $(\log a_n)$ is uniformly distributed mod 1.

The sequence $(\log n)$ is not uniformly distributed mod 1, and in fact

$$(8) \qquad \frac{1}{n}\sum_{j=1}^{j=n} e^{2\pi\, i\, m\ \log\ j} \simeq \frac{1}{n}\int_0^n e^{2\pi\, i\, m\ \log\ t}dt = \frac{e^{2\pi\, i\, m\, log\, n}}{1 + 2\pi\, i\, m \log\ e},$$

so the partial Cesàro sums in (7) wind up around the circles of radii $\dfrac{1}{|1 + 2\pi\, i\, m \log\ e|}$ (where $m \in \mathbf{Z} - \{0\}$).

One possible way of explaining the first digit law is to replace the sums (6) by a more general summation method, such as in [2]. In [6], (6) is replaced by an infinite iteration of Cesàro sums (see also [7]). It is clear from (8) that this method of averaging yields 0 (for each $m \neq 0$) when applied to the sequence $(e^{2\pi\, i\, m\ \log\ n})$, and so $(\chi_{D_k}(n))$ averages out to $\log(1 + \dfrac{1}{k})$.

Another argument [5] replaces density by a more general "logarithmic" or "harmonic" density

$$lim_{n\to\infty}\ \frac{1}{\ln\ n}\ \sum_{k=1}^{k=n}\ \frac{1}{\chi_X(k)}$$

and since $(\log n)$ is uniformly distributed mod 1 with respect to this method of summation [12], the first digit law follows.

Finally, we mention some conditions under which a measure on $\mathbf{N}$ satisfies the leading digit law. A natural condition has been suggested in [1]:

$$(9) \qquad \mu(X + 1) = \mu(X) \quad \text{and} \quad \mu(2X) = \frac{1}{2}\,\mu(X) \qquad (X \subseteq N)$$

and an even weaker condition was formulated in [3]:

$$\mu(2X \cup (2X+1)) = \mu(X) \qquad (X \subseteq N).$$

Either of these two conditons implies the leading digit law ([1], [3]). Measures that satisfy condition (9) are constructed in [1], with the additional property that $\mu(X) = \delta(X)$ whenever X has density.

References

1. R. Bumby and E. Ellentuck, *Finitely additive measures and the first digit problem*, Fundamenta Math. **65** (1969), 33–42.
2. J. Cigler, *Methods of summability and uniform distribution mod 1*, Compositio Math. **16** (1964), 44–51.
3. D. I. A. Cohen, *An explanation of the first digit phenomenon*, J. Combinatorial Theory (A) **20** (1976), 367–370.
4. P. Diaconis, *The distribution of leading digits and uniform distribution mod 1*, Annals of Probability **5** (1977), 72–81.
5. R. L. Duncan, *Note on the initial digit problem*, Fibonacci Quart. **7** (1969), 474–475.
6. B. J. Flehinger, *On the probability that a random integer has initial digit A*, American Math. Monthly **73** (1966), 1056–1061.
7. D. E. Knuth, *The art of computer programming*, 2nd ed., vol. 2, Addison-Wesley, Reading, MA, 1981, pp. 238–249.
8. S. Newcomb, *Note on the frequency of use of the different digits in natural numbers*, Amer. J. Math. **4** (1881), 39–40.
9. R. A. Raimi, *On the distribution of first significant figures*, American Math. Monthly **76** (1969), 342–348.
10. _____, *The peculiar distribution of first digits*, Scientific American (Dec. 1969), 109–120.
11. _____, *The first digit problem*, American Math. Monthly **83** (1976), 521–538.
12. M. Tsuji, *On the uniform distribution of numbers mod 1*, J. Math. Soc. Japan **4** (1952), 313–322.
13. H. Weyl, *Über die Gleichverteilung von Zahlen mod. Eins*, Math. Ann. **77** (1916), 313–352.

Department of Mathematics, The Pennsylvania State University, University Park, PA 16802, U.S.A.

E-mail address: jech@math.psu.edu