THE GRADUATE STUDENT SECTION

WHAT IS...

Benford's Law?

Arno Berger and Theodore P. Hill

Communicated by Cesar E. Silva

Benford's law quantifies the surprising fact that in many datasets, such as populations of counties, numbers on the World Wide Web, or incomes and expenses on tax returns, the numbers are much more likely to start with small digits like 1 or 2 than with large digits like 8 or 9. The law actually provides a specific probability distribution on the (significant) digits, telling exactly how likely each sequence of digits is. The law depends on the raw data x only via the significand $S(x)$, which is obtained by moving the decimal point immediately to the right of the first nonzero digit and by ignoring signs; e.g., $S(2017) = S(-0.02017) = 2.017$, with $D_1(2017) = 2$ and $D_2(2017) = 0$ being the first and second significant (decimal) digit, respectively.

Numbers in data sets are much likelier to start with small digits than large digits

A real-valued random variable X is said to be *Benford* (base 10) if

$$(1) \qquad \mathbb{P}(S(X) \le s) = \log_{10} s \quad \text{for all } 1 \le s < 10 .$$

Note that every Benford random variable also satisfies the *first-digit law*,

$$\mathbb{P}(D_1(X) = d) = \log_{10} \frac{d+1}{d} \quad \text{for all } d = 1, \dots, 9 ,$$

so in particular the probability of $D_1(X) = 1$ is more than six times the probability of $D_1(X) = 9$; see Figure 1.

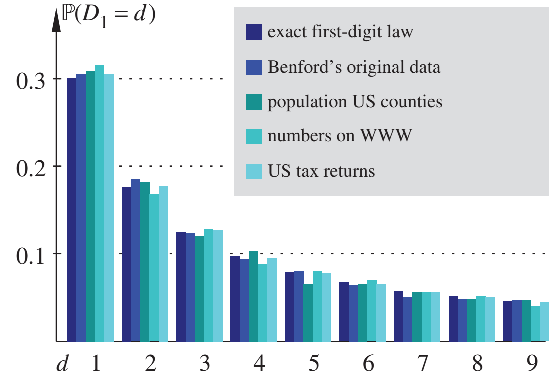

Figure 1. Comparisons to the first-digit law for four datasets.

The first known reference to the logarithmic distribution (1) is a two-page note in the *American Journal of Mathematics* in 1881 by renowned Canadian-American mathematician and astronomer Simon Newcomb, future president of the American Mathematical Society. Newcomb's discovery went largely unnoticed, and it came to be known as *Benford's law* after its rediscovery and popularization by physicist Frank Benford in 1938.

Stochastic Processes

One key to the analysis of (1) is the *significand σ-algebra* $\mathbb{S}$, the sub-σ-algebra of the Borel sets on $\mathbb{R}^+$ generated by the significand function S (or, equivalently, generated by the significant digit functions $D_1, D_2, D_3, \ldots$). This σ-algebra has several interesting properties: Every nonempty set $A \in \mathbb{S}$ is infinite with accumulation points at 0 and $+\infty$; $\mathbb{S}$ is self-similar with respect to multiplication by integral powers of 10, i.e., $10^k A = A$ for all $k \in \mathbb{Z}$ and $A \in \mathbb{S}$; and $\mathbb{S}$ is closed under (positive) scalar multiplication and integral roots, i.e.,

$$(2) \qquad aA^{1/n} \in \mathbb{S} \quad \text{for all } a > 0, A \in \mathbb{S}, n \in \mathbb{N},$$

but is *not* closed under integral powers.

The closure property (2), together with tools from Fourier analysis and ergodic theory, yields that Benford's

Arno Berger is professor of mathematics at the University of Alberta. His e-mail address is berger@ualberta.ca.

Ted Hill is professor emeritus of mathematics at Georgia Tech, currently research scholar in residence at Cal Poly San Luis Obispo. His e-mail address is hill@math.gatech.edu.

DOI: http://dx.doi.org/10.1090/noti1477

law (1) is the unique probability distribution on the significant digits or, equivalently, on $(\mathbb{R}^+, \mathbb{S})$, that is *scale-invariant* (i.e., the distributions of $S(X)$ and $S(aX)$ are identical for all $a > 0$), is the unique continuous distribution that is *base-invariant*, and is the unique distribution that is *sum-invariant*.

None of the classical probability distributions—uniform, normal, exponential, gamma, Cauchy, Poisson, etc.—are exactly Benford. Some Pareto and lognormal distributions are close to being Benford and some are not, say, with respect to the Kolmogorov (sup norm) distance between the cumulative distribution functions, whereas other families such as uniform distributions are bounded strictly away from being Benford.

Sums of random variables are generally not Benford, not even if the summands are independent and Benford. The product XY of two independent positive random variables, on the other hand, is Benford if *either* X *or* Y is Benford. The sequence $(X, X^2, X^3, \dots)$ of powers of a random variable X is a Benford sequence (see below) with probability one if and only if

$$\mathbb{P}(\log_{10}|X| \text{ is rational}) = 0. \tag{3}$$

The sequence $(X_1, X_1X_2, X_1X_2X_3, \dots)$ of products of independent and identically distributed copies X_j of X is Benford with probability one if and only if

$$\mathbb{P}\left(\log_{10}|X| \in m^{-1}\mathbb{Z}\right) < 1 \quad \text{for every } m \in \mathbb{N}.$$

Note that the latter follows from (3). Both conditions are satisfied whenever X is absolutely continuous, in which case $S(X^n)$ and $S(\prod_{j=1}^n X_j)$ also converge in distribution to a Benford random variable; see Figure 2.

In a central-limit-like context, taking random samples from random distributions in an "unbiased" manner will entail convergence to Benford's law with probability one in the sense that the empirical distributions of the significands of the mixed random samples will converge almost surely to the logarithmic distribution (1). This perhaps helps explain why numbers appearing in newspapers or on the Web have been found to be good fits to (1). However, rates of convergence in these settings, or even good rules of thumb analogous to those for the Central Limit Theorem, have yet to be discovered.

Deterministic Processes

In direct analogy to (1), a sequence $(x_n) = (x_1, x_2, x_3, \dots)$ of real numbers is said to be *Benford* if, for every $1 \le s < 10$, the natural density of the set $\{n \in \mathbb{N} : S(x_n) \le s\}$ exists and equals $\log_{10} s$. Many familiar sequences of real numbers, including the sequences of positive integers and prime numbers, are *not* Benford, but many others are, including the sequences of Fibonacci and Lucas numbers, factorials, and powers of 2.

To see, for instance, that (2^n) is Benford, check first that a sequence (x_n) of real numbers is Benford if and only if the sequence $(\log_{10}|x_n|)$ is uniformly distributed mod 1, note that $\log_{10} 2$ is irrational, and apply a theorem of Weyl about the uniform distribution under irrational rotations on the circle. Since (2^n) is Benford, the scale invariance of Benford's law implies that the orbit under the map $f(x) = 2x$, i.e., the sequence $(f^n(x_0)) = (2^n x_0)$, is Benford for all $x_0 \neq 0$.

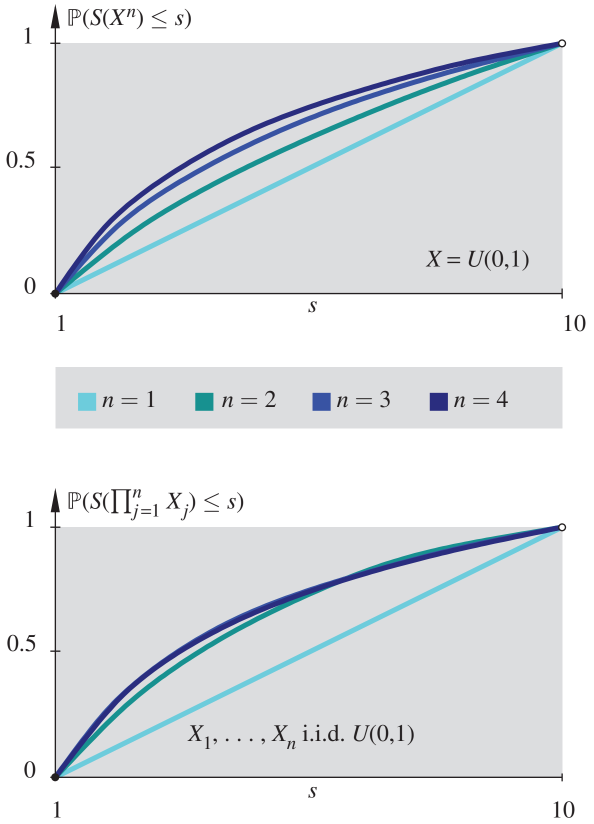

Figure 2. For absolutely continuous X, the distribution for the significands of powers (top) and i.i.d. products of X approach the logarithmic distribution (1).

A more subtle result, which can be established utilizing finer uniform distribution tools and the dynamical systems technique of *shadowing*, is that the orbit under any map $f(x) = ax + g(x)$ with $g = o(x/\log x)$ as $x \to +\infty$ is Benford for all sufficiently large x_0 if and only if $a > 1$ is not a rational power of 10. The orbit under a smooth map $f(x) = ax^2 + g(x)$ with $g' = o(x/\log x)$, on the other hand, is Benford without any restriction on $a > 0$, but only for Lebesgue-almost all sufficiently large x_0; the set of exceptional x_0 is nonempty (as illustrated by Figure 3) and in fact uncountable.

Even when a map produces Benford orbits for almost all starting points, determining exactly for which x_0 it does so can be a challenge. For example, the orbit under $f(x) = x^2 + 1$ is Benford for Lebesgue-almost all x_0, but it is not known whether the orbit of $x_0 = 0$, i.e., the integer sequence $(x_n) = (1, 2, 5, 26, 677, \dots)$, is Benford. (Incidentally, x_n for each $n \in \mathbb{N}$ equals the number of binary trees of height less than n.)

Analogous to the notion of a Benford sequence, a real-valued function f on $\mathbb{R}^+$ is *Benford* if

$$T^{-1}\text{length}\{0 < t \le T : S(f(t)) \le s\}$$

converges to $\log_{10} s$ as $T \to +\infty$ for every $1 \leq s < 10$ or, equivalently, if $\log_{10}|f|$ is (continuously) uniformly distributed. For instance, all functions $f(t) = e^{at}p(t)$, with $a \neq 0$ and any polynomial $p \neq 0$, are Benford. Note that these functions are solutions of autonomous linear differential equations. Perhaps not too surprisingly, the solutions of many other differential equations also turn out to be Benford functions.

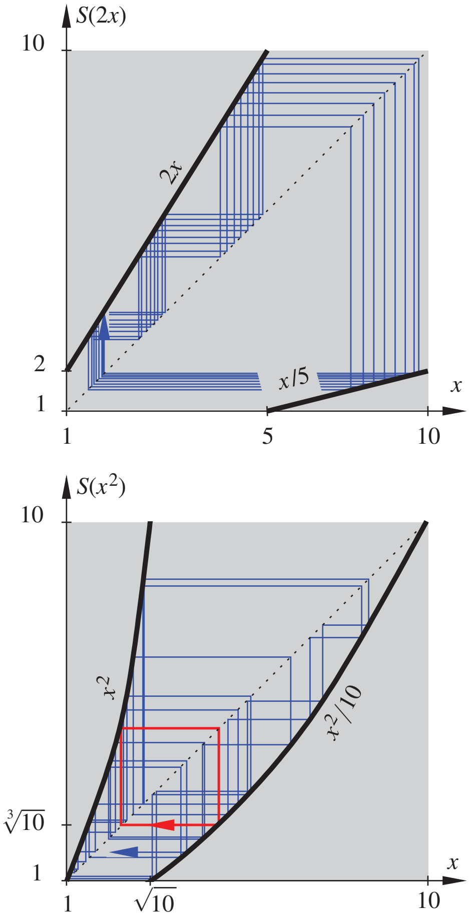

Figure 3. Significands for orbits under $f(x) = 2x$ (top) and $f(x) = x^2$. The second, quadratic, case exhibits periodic orbits like the one shown in red for $\sqrt[3]{10}$, whereas the first, linear case does not.

Applications

Knuth's classic *The Art of Computer Programming* noted the ubiquity of Benford's law in scientific calculations and its consequent use in analyzing the average behavior of floating point arithmetic algorithms. In the mid-1990s, accountant Mark Nigrini found that Benford goodness-of-fit tests can be very effective red-flag tests, since true tax data often is a close fit to Benford's law, whereas fabricated data is not. The statistical phenomenon of Benford's law has since seen widespread interest across a broad range of fields, from accounting and physics to digital-image forensics, biology, and medicine. Of the nearly thousand entries in the database `www.benfordonline.net`, more than two thirds have appeared since 2000.

Current Status. (Warning: Benford's law may be addictive!)

Many questions regarding Benford's law are natural variants of famous open problems. For instance, the question of which starting points generate Benford orbits is closely related to the problem of deciding which real numbers are normal (i.e., have all finite strings of digits equally represented in their decimal expansion). Even though a theorem of Borel guarantees that almost all numbers are normal, this remains a hard problem in general. The question of whether Benford random variables are the only continuous random variables for which $S(X), S(X^2), S(X^3)$ all have the same distribution is a variant of Furstenberg's question regarding invariant measures for $2x \bmod 1$ and $3x \bmod 1$. And the answer to the question of whether any solution (x_n) of the linear difference equation $x_n + 2x_{n-1} + 3x_{n-2} = 0$ is Benford hinges on Schanuel's conjecture on transcendence degrees.

There is no known back-of-the-envelope argument, not even a heuristic one, that explains the appearance of Benford's law across the board—in data that is pure or mixed, deterministic or stochastic, discrete or continuous-time, real-valued or multidimensional. The main mathematical challenge is to establish theories that help explain and predict when data will follow Benford's law and when it will not.

If useful conclusions can be drawn from huge datasets by looking at only a few significant digits, as has been reported to be the case for detecting financial fraud and alterations of digital images, as well as for detecting natural phenomena such as earthquakes and phase changes in quantum processes, then these same techniques may well provide useful tools for many other applications in the future.

References

[1] A. Berger and T. P. Hill, *An Introduction to Benford's Law*, Princeton University Press, Princeton, NJ, 2015. MR 3242822

[2] R. A. Raimi, The first digit problem, *Amer. Math. Monthly* **83** (1976), no. 7, 521–538. MR 0410850

Photo Credits

All illustrations are courtesy of Arno Berger.

ABOUT THE AUTHORS

Arno Berger is interested in dynamical systems, analysis, and probability theory. When not immersed in more mundane matters, he enjoys marvelling at the big Alberta sky from his office window.

Ted Hill spends much of his spare time in a cabin overlooking the Pacific Central Coast, and in addition to fair division, optimal stopping, and general probability, he loves hiking, diving, and mountain biking.