Benford's Law Strikes Back: No Simple Explanation in Sight for Mathematical Gem

ARNO BERGER* AND THEODORE P. HILL

The widely known phenomenon called Benford's Law continues to defy attempts at an easy derivation. This article briefly reviews recurring flaws in "back-of-the-envelope" explanations of the law, and then analyzes in more detail some of the recently published attempts, many of which replicate an apparently unnoticed error in Feller's classic 1966 text **An Introduction to Probability Theory and Its Applications**. Specifically, the claim by Feller and subsequent authors that "regularity and large spread implies Benford's Law" is fallacious for any reasonable definitions of regularity and spread (measure of dispersion). The fallacy is brought to light by means of concrete examples and a new inequality. As for replacing the wrong assertions by an equally simple explanation which is valid, now—that is a task for the future.

It's All About Digits

The eminent logician, mathematician, and philosopher C.S. Peirce once observed [Ga, p.273] that "in no other branch of mathematics is it so easy for experts to blunder as in probability theory". As the reader as well will see, this is all too true for *Benford's Law*, also known as the *First-Digit Phenomenon*.

Benford's Law, abbreviated henceforth as BL, is one of the gems of statistical folklore. It is the observation that in many collections of numbers, be they mathematical tables, real-life data, or combinations thereof, the leading significant digits are not uniformly distributed, as might be expected, but are heavily skewed toward the smaller digits. More precisely, BL says that the significant digits in many datasets follow a very particular logarithmic distribution. In its most common formulation, the special case of first significant *decimal* (i.e., base 10) digits, BL reads

$$\text{Prob}(D_1 = d_1) = \log_{10}\left(1 + d_1^{-1}\right), \quad \text{for all } d_1 = 1, \ldots, 9; \qquad \text{(BL1)}$$

*This author was supported by an NSERC Discovery Grant.

here D_1 denotes the first significant decimal digit, e.g.,

$$D_1(\sqrt{2}) = D_1(1.414\ldots) = 1,$$
$$D_1(\pi^{-1}) = D_1(0.3183\ldots) = 3,$$
$$D_1(e^{\pi}) = D_1(23.14\ldots) = 2.$$

A crucial part of the content of (BL1), of course, is an appropriate formulation or interpretation of "Prob". For sequences of real numbers or real datasets, for example, Prob usually refers to the proportion (or relative frequency) of entries for which an event such as $D_1 = 1$ occurs, whereas for a random variable, Prob is simply the probability on the underlying probability space. Figure 1 illustrates several of these settings, including mathematical sequences such as the powers of 2, and real-life data from Benford's original paper as well as recent census statistics.

In a form more complete than (BL1), BL is a statement about the joint distribution of *all* decimal digits: For every natural number n, this version states that

$$\mathrm{Prob}((D_1, D_2, \ldots, D_n) = (d_1, d_2, \ldots, d_n)) = \log_{10}\left(1 + \left(\sum\nolimits_{j=1}^{n} 10^{n-j} d_j\right)^{-1}\right) \qquad \text{(BL2)}$$

holds for all n-tuples $(d_1, d_2, \ldots, d_n)$, where d_1 is an integer in $1, 2, \ldots, 9$ and where for $j > 1$, d_j is an integer in $0, 1, \ldots, 9$. Here D_2, D_3, D_4, etc. represent the second, third, fourth, etc. significant decimal digit, so that, for example,

$$D_2(\sqrt{2}) = 4, \quad D_3(\pi^{-1}) = 8, \quad D_4(e^{\pi}) = 4.$$

The "laws" (BL1) and (BL2) were apparently first discovered by polymath S. Newcomb in the 1880s [N]. They were rediscovered by physicist F. Benford [Ben]; Newcomb's article having been forgotten at the time, they came to be known as *Benford's* Law. Today, BL appears in a broad spectrum of mathematics, ranging from differential equations to number theory to statistics. Simultaneously, the applications of BL are mushrooming—from diagnostic tests for mathematical models in biology and finance to fraud detection. For instance, the U.S. Internal Revenue Service uses BL to ferret out suspicious tax returns, political scientists use it to identify voter fraud, and engineers to detect altered digital images. As Raimi already observed some 25 years ago [R1, p.512],

> This particular logarithmic distribution of the first digits, while not universal, is so common and yet so surprising at first glance that it has given rise to a varied literature, among the authors of which are mathematicians, statisticians, economists, engineers, physicists, and amateurs.

The online database [BH] now contains more than 600 articles on the subject. Many of these articles, including some of the very recent ones, purport to provide easy derivations or proofs of BL. The present article sets out to identify some of the prevalent fallacies in those arguments.

Simple Explanations? Are You Sure?

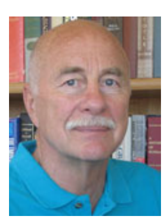
Let's start with the purely mathematical framework. Many familiar sequences, including the Fibonacci numbers, the powers of 2, and the factorial sequence ($n!$), all follow BL *exactly*; in particular, exactly a proportion of $\log_{10} 2 = 0.3010\ldots$, that is, approximately 30.1% of all entries of those sequences begin with the decimal digit 1. Similarly, start with any positive number and multiply by 3 repeatedly (i.e., iterate the function $x \mapsto 3x$), or multiply alternately by 3 and by 4, or iterate the function $x \mapsto 2x + 1$. Each of these iterations results in a sequence that follows BL exactly, no matter what positive number was chosen in the beginning. Thus, even though many common sequences such as the natural numbers and the primes do not follow BL, those that do are so ubiquitous that many authors have assumed that a simple explanation must exist.

Raimi [R1, R2] reviews many of the attempts at such explanations: Some authors simply labeled BL self-evident; thus Benford himself wrote that "the logarithmic law applies particularly to those outlaw numbers that are without known relationship", Goudsmit and Furry opined that it "is merely the result of our way of writing numbers", and likewise Weaver claimed that BL "is a built-in characteristic of our number system". For the more mathematical ones among the back-of-the-envelope derivations, Raimi

AUTHORS

ARNO BERGER is an Associate Professor at the University of Alberta. Before coming to Canada, he had held faculty positions in Austria and New Zealand. His main passion in mathematics is dynamical systems theory, so naturally he is also interested in analysis and probability.

Department of Mathematical and Statistical Sciences
University of Alberta
Edmonton, T6G 2G1
Canada
e-mail: berger@ualberta.ca



THEODORE P. HILL earned his doctorate at University of California, Berkeley. He is now Professor Emeritus at Georgia Tech (after many visiting stints in Costa Rica, The Netherlands, and elsewhere) and is a visiting scholar at California Polytechnic State University. His primary research areas are mathematical probability, especially optimal-stopping theory, fair-division problems – and Benford's Law.

School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332
USA
e-mail: hill@math.gatech.edu

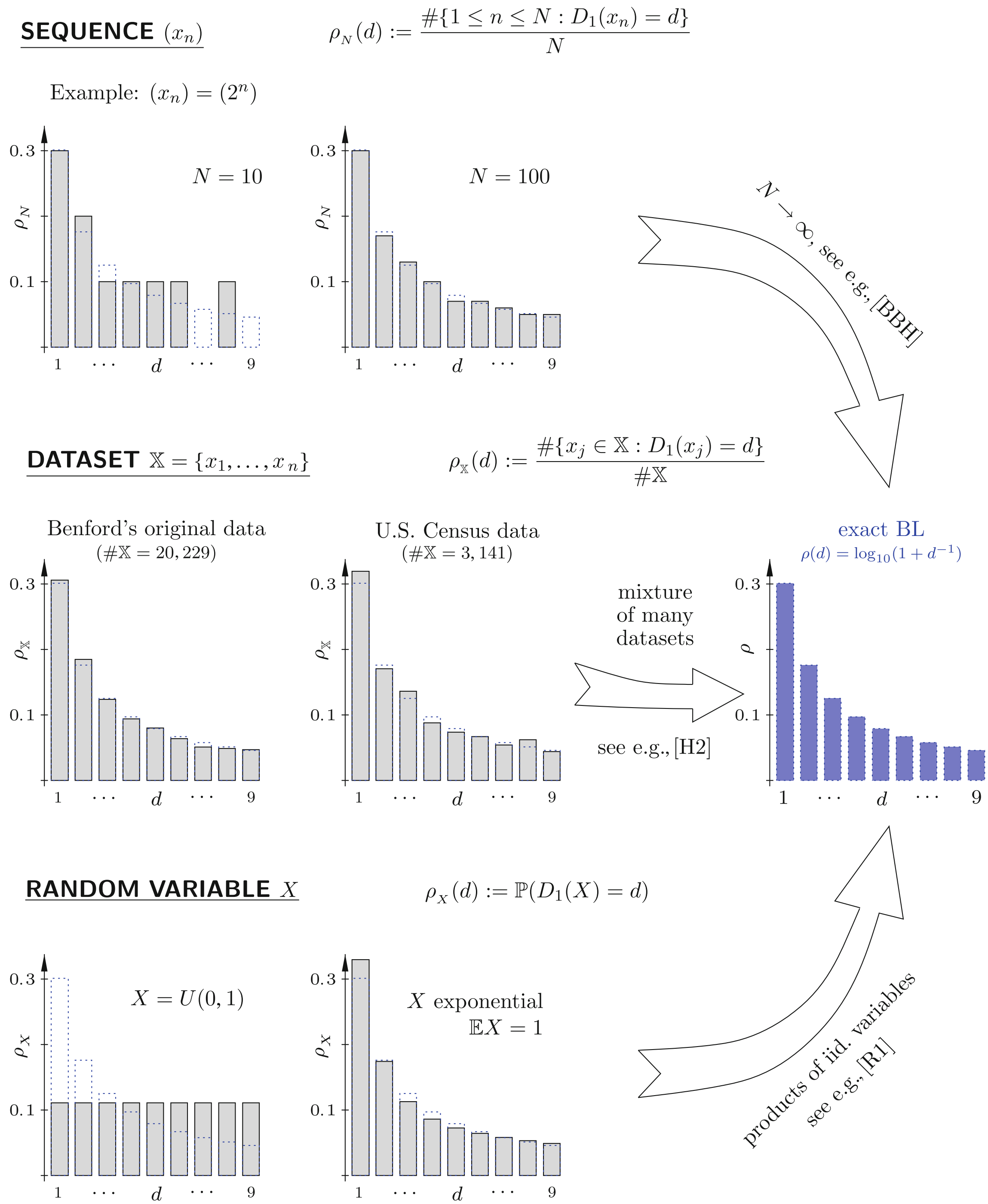


Figure 1. Different interpretations of (BL1) for sequences, datasets, and random variables, respectively, and scenarios that may lead to exact conformance with BL.

carefully points out their shortcomings, e.g., Flehinger's Cesàro-summation method, Herzel's urn model, and the two different fallacies in Logan and Goudsmit's urn model derivation.

In [R1, sec.7], Raimi also explains the basic flaw in Pinkham's widely cited scale-invariance argument. That is the argument that BL is the only distribution on significant digits that is invariant under changes of scale, meaning that (BL1) and (BL2) remain unchanged if a sequence, dataset, or random variable is multiplied by any positive constant. Raimi credits Knuth [K] for the discovery that the error is in Pinkham's implicit assumption that there exists a scale-invariant probability distribution on the positive real numbers, when clearly there is no such distribution. To see this, simply note that for instance multiplying any positive random variable X by 2 doubles the value of its median, and hence X cannot be scale-invariant. To the best of the authors' knowledge, the first correct proof that BL indeed is the unique scale-invariant probability distribution (and also the unique continuous base-invariant distribution) on the significant digits is in [H2].

A closer look at sequences of numbers reveals some surprises that may help explain why correct and quick derivations of BL in a purely mathematical context may be hard to come by. For instance, iterating the function $x \mapsto x^2 + 1$ results in a sequence following BL for (Lebesgue) *almost all* starting points, but not for *all* starting points. Thus, if the initial value x_1 is chosen randomly from any positive distribution with a density, such as, say, the uniform distribution on (0,1) or an exponential distribution, then the resulting sequence (x_n) with $x_{n+1} = x_n^2 + 1$ will follow BL exactly with probability 1. But there are exceptional points also. For example, choosing $x_1 = 9.9496230\ldots$ implies $D_1(x_n) = 9$ for all n, that is, *every* number x_n begins with a decimal digit 9. (See [BBH, exp.4.3] to find out what is special about this remarkable value for x_1.) Whether or not the sequence (x_n) follows BL when $x_1 = 0$ is still an open problem.

All in all, even though it would be highly desirable to have both a rigorous formal proof and a reasonably sound heuristic explanation, it seems unlikely that any quick derivation has much hope of explaining BL mathematically.

It is in the realm of real-life data that assertions about easy derivations of BL become especially treacherous. One type of erroneous shortcut in particular continues to propagate, and the remainder of this article is devoted to identifying and illuminating it.

A variety of formal mathematical proofs is available for sequences and random variables (see e.g., [BBH, H2]). But in teaching probability and statistics, a correct general explanation of a principle is often as valuable as a detailed formal argument. In his December 2009 column in the *IMS Bulletin*, UC Berkeley statistics professor T. Speed extols the virtues of derivations in statistics [S]:

> I think in statistics we need derivations, not proofs. That is, lines of reasoning from some assumptions to a formula, or a procedure, which may or may not have certain properties in a given context, but which, all going well, might provide some insight.

For illustration, Speed quotes two examples of the convolution property for the Gamma and Cauchy distributions from the classic 1966 text *An Introduction to Probability Theory and Its Applications* by W. Feller [Fel]. On page 63, Feller also gave a brief derivation, in Speed's sense, of BL.

For the purposes of this note, a simple and very useful characterization of BL in the stochastic setting can be given in terms of uniform distribution modulo one. Recall that a random variable X is *uniformly distributed modulo one*, or *u.d.* mod 1 for short, if the fractional part $X \bmod 1 := X - \lfloor X \rfloor$ of X has the same distribution as $U(0,1)$; here $\lfloor\, x \,\rfloor$ denotes, for any real x, the largest integer not larger than x, and $U(0,1)$ is a random variable uniformly distributed on (0,1). In these terms, the promised characterization of BL (see also Figure 2) is

(1) A positive random variable X follows BL if and only if $\log_{10} X$ is u.d. mod 1.

Since Feller has inspired so many who teach probability and statistics today, and since many undergraduate courses now include a brief introduction to BL, it is not surprising that Feller's derivation is still in frequent use to "provide some insight" about this phenomenon. For example, a class project report for a 2009 upper-division course in statistics at UC Berkeley [AP1, p.3] said,

> ...like the birthday paradox, an explanation [of BL] occurs quickly to those with appropriate mathematical background ... To a mathematical statistician, Feller's paragraph says all there is to say ... Feller's derivation has been common knowledge in the academic community throughout the last 40 years.

The online database [BH] lists about twenty published references since 2000 alone to Feller's argument (e.g., [AP1, Few]) the crux of which is Feller's claim (trivially edited) that

(2) If the spread of a random variable X is very large, then $\log_{10} X$ will be approximately u.d. mod 1.

The implication of (1) and (2) is that all random variables with large spread will approximately follow BL. That sounds quite plausible, but true to C.S. Peirce's observation, even Feller blundered on Benford's Law, and he took many other experts with him. Claim (2) is simply false under any reasonable definition of "spread" and any reasonable measure of dispersion, including *range*, *interquartile range*, *standard deviation* and *mean difference*, no matter how smooth or level a density the random variable X may have. To see this, one does not have to look far. Concretely, no positive uniformly distributed random variable even comes close to following BL, regardless of how large (or small) its spread is. This statement can be quantified explicitly via the following new inequality which is stated in terms of the so-called the Kolmogorov-Smirnov distance $d_{KS}(X, Y)$ between two random variables X and Y, defined as $d_{KS}(X, Y) = \sup_{x \in \mathbb{R}} \left| \mathbb{P}(X \le x) - \mathbb{P}(Y \le x) \right|$.

Proposition 1 ([Ber]) *For every positive uniformly distributed random variable X,*

$$d_{KS}(\log_{10} X \bmod 1, U(0,1)) \ge \frac{-9 + \ln 10 + 9 \ln 9 - 9 \ln \ln 10}{18 \ln 10} = 0.1334\ldots,$$

and this bound is sharp.

There is nothing special about the use of the Kolmogorov-Smirnov distance or of decimal base in this regard; similar universal bounds hold for the Wasserstein distance, for example, and other bases. Likewise, there is nothing special about the choice of the uniform distribution as a source of counterexamples here and below; its usage is solely motivated by the simplicity of the uniform distribution and its role in many applications. For example, if X_α is exponentially distributed with mean α then

$$\mathbb{P}(D_1(X_\alpha) = 1) = \sum_{k \in \mathbb{Z}} \left(e^{-10^k} - e^{-2 \cdot 10^k} \right) = 0.3296\ldots > \log_{10} 2,$$

whenever α is an integer power of 10. Thus in this case as well, it follows immediately from (1) that X_α is not

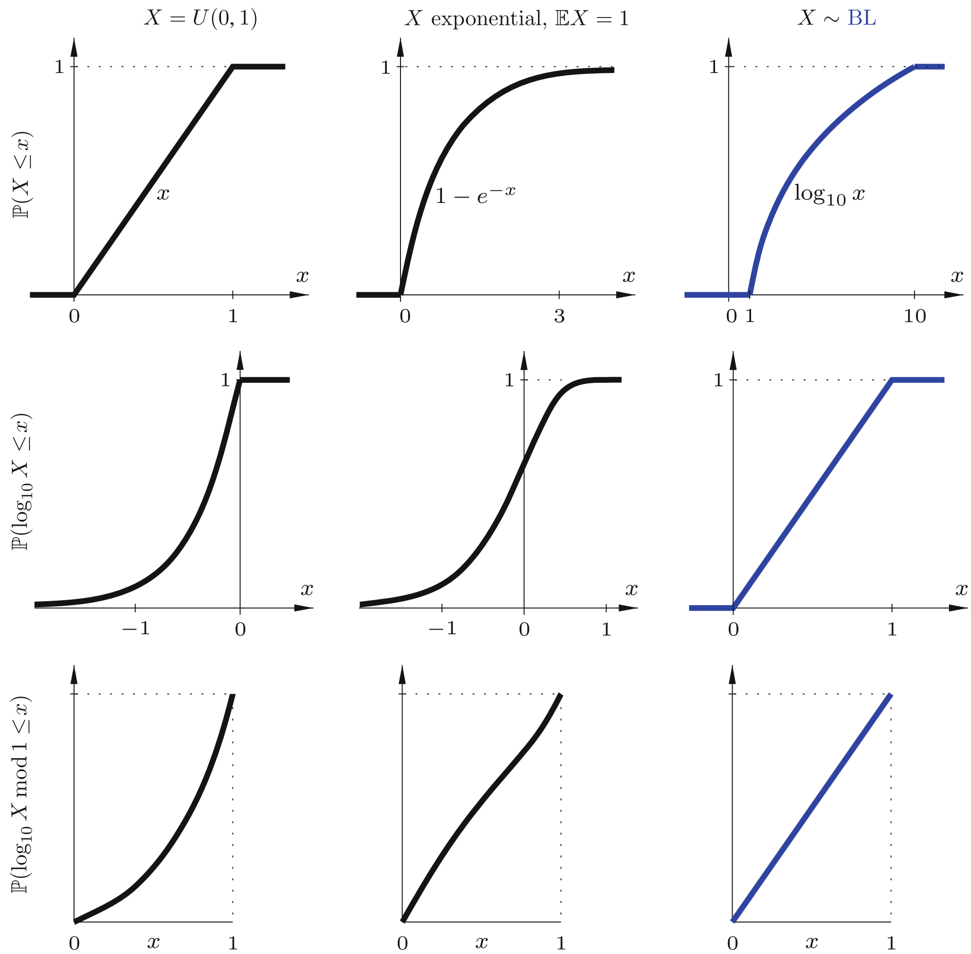


Figure 2. Uniform (left column) and exponential (center column) random variables do not follow BL as $\log_{10}X$ is not uniformly distributed modulo one, see bottom row. However, note that in the exponential case the deviation from BL is quite small.

approaching BL, even though the spread (range, interquartile range, standard deviation, mean difference, etc.) of X_α goes to infinity as $\alpha \to \infty$.

How could Feller's error have persisted in the academic community, among students and experts alike, for over 40 years? Part of the reason, as one colleague put it, is simply that "Feller, after all, is Feller", and Feller's word on probability has just been taken as gospel. Another reason for the long-lived propagation of the error has apparently been the confusion of (2) with the similar claim

(3) If the spread of a random variable X is very large, then X will be approximately u.d. mod 1.

For example, [AP1, p.3] cites Feller's claim (2), but on p. 8 the same article states Feller's claim as (3). A third possible explanation for the persistence of the error is the common assumption that (3) implies (2). For example, [GD, p.1] states:

> An elementary new explanation has recently been published, based on the fact that any X whose distribution is "smooth" and "scattered" enough is Benford. The scattering and smoothness of usual data ensures that log (X) is itself smooth and scattered, which in turn implies the Benford characteristic of X.

Now (3) is also intuitive and plausible, but unlike (2), it is often accurate if the distribution is fairly uniform. And if the distribution is not fairly uniform, then without further information, no interesting conclusions at all can be made about the significant digits: most of the values could for instance start with a "7". Now it seems obvious that X has very, very large spread if and only if log X has very large spread, so on the surface (2) and (3) appear to be equivalent. After all, what difference can one tiny extra "very" make? But the obvious again is simply false, as can easily be seen, for instance, when X has a Pareto distribution with parameter 2, that is, $\mathbb{P}(X > x) = x^{-2}$ for all $x \ge 1$. Then X has *infinite* variance, whereas the variance of $\log_{10}X$ equals $\frac{1}{4}(\log_{10} e)^2$ and hence is less than 0.05. Thus (2) and (3) are not at all equivalent, and (2) is false under practically any interpretation of "spread".

Although (3) is perhaps more accurate than (2), unfortunately it does nothing to explain BL, for the criterion in (1) says that X follows BL if and only if the *logarithm* of X—and not X itself—is uniformly distributed modulo one. Some authors partially explain the ubiquity of BL based on an assumption of a "large spread on a logarithmic scale" (e.g., [AP1, Few, W]), and some, when confronted with the evidence that (2) is false, claim that "what Feller obviously *meant*" [AP2, italics in original] by spread was log spread, i.e., that when Feller wrote (2) he really meant to say that

(4) If $\log_{10} X$ has very large spread, then $\log_{10} X$ will be approximately u.d. mod 1,

which is but an unnecessarily convoluted version of (3). They then apply (3) or (4) to conclude that if $\log_{10}X$ has large spread, then X approximately follows BL. This avoids Feller's error (2), but still leaves open the question of why it is reasonable to assume that the *logarithm* of the spread, as opposed to the spread itself—or, say, the log log spread—should be large. As seen above, those assumptions contain subtle differences, and lead to very different conclusions about the distributions of significant digits. Moreover, via (1) and (3), assuming large spread on a logarithmic scale is equivalent to assuming an approximate conformance with BL. Quite likely, Feller realized this, and in (2) specifically did *not* hypothesize that the log of the range was large.

A related and apparently widespread misconception is that claim (2) or claim (3)—notwithstanding the incorrectness of the former—implies that a larger spread or log spread automatically means better conformance with BL. For example, [W] concludes that "datasets with large logarithmic spread will naturally follow the law, while datasets with small spread will not", and the Conclusion of the study [AP2, p.12] states,

> On a small stage (18 data-sets) we have checked a theoretical prediction. Not just the literal assertion of Benford's law — that in a data-set with large spread on a logarithmic scale, the relative frequencies of leading digits will approximately follow the Benford distribution — but the rather more specific prediction that distance from Benford should decrease as that spread increases. In one sense it's not surprising this works out.

But it doesn't. Distance from BL does not always decrease as the spread increases, regardless of whether the spread is measured on the original scale or on the logarithmic scale. A simple way to see this is as follows: Again, for simplicity, let Y be a random variable uniformly distributed on (0,1), and let $X = 10^Y$ and $Z = 10^{3Y/2}$. Then by (1), X follows BL exactly, since $\log_{10}X = Y$, while Z is not close to BL, for $3Y/2$ mod 1 is not close to uniform on (0,1). Yet for any reasonable definition of spread, including all those mentioned earlier, the spread of Z is larger than the spread of X, and the spread of $\log_{10}Z = 3Y/2$ is larger than the spread of $\log_{10}X = Y$. Another way to see that the distance from BL does not decrease as the spread increases is contained in the proof of Proposition 1: For X_T a random variable uniformly distributed on $(0,T)$, it is shown there that the Kolmogorov-Smirnov distance between $\log_{10} X_T$ mod 1 and $U(0,1)$ is a continuous 1-periodic function of $\log_{10}T$.

Moreover, when employing a logarithmic scale it is important to keep in mind that what is considered large generally depends on the base of the logarithm. For example, as noted earlier, if Y is uniformly distributed on (0,1) then $X = 10^Y$ is exactly Benford base 10, yet it is not Benford base 2 even though its spread on the $\log_2$-scale is $\log_2 10 \approx 3.322$ times as large.

Conclusion

Classroom experiments based on Feller's derivation or on an assumption of large spread on a logarithmic scale (e.g., [AP1, AP2, Few, W]) should be used with caution. As alternative, a supplement or teachers might also ask students to compare the significant digits in the first 20-30 articles in tomorrow's *New York Times* against BL, thereby testing real-life data against the explanation given in the main theorem in [H2], which, without any assumptions on magnitude of spread, shows that mixing data from different distributions in an unbiased manner leads to exact conformance with BL.

Although some experts may still feel that "like the birthday paradox, there is a simple and standard explanation" for BL [AP2, p.6] and that this explanation "occurs quickly to those with appropriate mathematical background", there does not appear to be a simple derivation of BL that both offers a "correct explanation" [AP2, p.7] and satisfies Speed's goal to provide insight. A broad and often ill-understood phenomenon need not always be reduced to a few theorems. Although many facets of BL now rest on solid ground, there is currently no unified approach that simultaneously explains its appearance in dynamical systems, number theory, statistics, and real-world data. In that sense, most experts seem to agree with [Few] that the ubiquity of BL, especially in real-life data, remains mysterious.

ACKNOWLEDGMENT

The authors are grateful to Rachel Fewster, Kent Morrison, and Stan Wagon for excellent suggestions that helped to improve the exposition.

REFERENCES

[AP1] Aldous, D., and Phan, T. (2009), ''When Can One Test an Explanation? Compare and Contrast Benford's Law and the Fuzzy CLT'', Class project report dated May 11, 2009, Statistics Department, UC Berkeley; accessed on May 14, 2010, at [BH].

[AP2] Aldous, D., and Phan, T. (2010), ''When Can One Test an Explanation? Compare and Contrast Benford's Law and the Fuzzy CLT'', Preprint dated Jan. 3, 2010, Statistics Department, UC Berkeley; accessed on May 14, 2010, at [BH].

[Ben] Benford, F. (1938), ''The Law of Anomalous Numbers'', *Proc. Amer. Philosophical Soc*. 78, 551–572.

[Ber] Berger, A. (2010), ''Large Spread Does Not Imply Benford's Law'', Preprint; accessed on May 14, 2010, at http://www.math.ualberta.ca/~aberger/Publications.html.

[BBH] Berger, A., Bunimovich, L., and Hill, T.P. (2005), ''One-dimensional Dynamical Systems and Benford's Law'', *Trans. Amer. Math. Soc*. 357, 197–219.

[BH] Berger, A., and Hill, T.P. (2009), *Benford Online Bibliography*; accessed May 14, 2010, at http://www.benfordonline.net.

[Fel] Feller, W. (1966), *An Introduction to Probability Theory and Its Applications* vol. 2, 2nd ed., J. Wiley, New York.

[Few] Fewster, R. (2009), ''A Simple Explanation of Benford's Law'', *American Statistician* 63(1), 20–25.

[Ga] Gardner, M. (1959), ''Mathematical Games: Problems involving questions of probability and ambiguity'', *Scientific American* 201, 174–182.

[GD] Gauvrit, N., and Delahaye, J.P. (2009), ''Loi de Benford générale'', *Mathématiques et sciences humaines* 186, 5–15; accessed May 14, 2010, at http://msh.revues.org/document11034.html.

[H1] Hill, T.P. (1995), ''Base-Invariance Implies Benford's Law'', *Proc. Amer. Math. Soc*. 123(3), 887–895.

[H2] Hill, T.P. (1995), ''A Statistical Derivation of the Significant-Digit Law'', *Statistical Science* 10(4), 354–363.

[K] Knuth, D. (1997), *The Art of Computer Programming*, pp. 253-264, vol. 2, 3rd ed, Addison-Wesley, Reading, MA.

[N] Newcomb, S. (1881), ''Note on the Frequency of Use of the Different Digits in Natural Numbers'', *Amer. J. Math*. 4(1), 39–40.

[P] Pinkham, R. (1961), ''On the Distribution of First Significant Digits'', *Annals of Mathematical Statistics* 32(4), 1223–1230.

[R1] Raimi, R. (1976), ''The First Digit Problem'', *Amer. Mathematical Monthly* 83(7), 521–538.

[R2] Raimi, R. (1985), ''The First Digit Phenomenon Again'', *Proc. Amer. Philosophical Soc*. 129, 211–219.

[S] Speed, T. (2009), ''You Want Proof?'', *Bull. Inst. Math. Statistics* 38, p 11.

[W] Wagon, S. (2010), ''Benford's Law and Data Spread''; accessed May 14, 2010, at http://demonstrations.wolfram.com/BenfordsLawAndDataSpread.