Benford's Law as a Logarithmic Transformation

J. M. Pimbley

Maxwell Consulting, LLC

Abstract

We review the history and description of Benford's Law - the observation that many naturally occurring numerical data collections exhibit a logarithmic distribution of digits. By creating numerous hypothetical examples, we identify several cases that satisfy Benford's Law and several that do not. Building on prior work, we create and demonstrate a "Benford Test" to determine from the functional form of a probability density function whether the resulting distribution of digits will find close agreement with Benford's Law. We also discover that one may generalize the first-digit distribution algorithm to include non-logarithmic transformations. This generalization shows that Benford's Law rests on the tendency of the transformed and shifted data to exhibit a uniform distribution.

Introduction

What the world calls "Benford's Law" is a marvelous mélange of mathematics, data, philosophy, empiricism, theory, mysticism, and fraud. Even with all these qualifiers, one easily describes this law and then asks the simple questions that have challenged investigators for more than a century.

Take a large collection of positive numbers which may be integers or real numbers or both. Focus only on the first (non-zero) digit of each number. Count how many numbers of the large collection have each of the nine possibilities (1-9) as the first digit. For typical

number collections – which we'll generally call "datasets" – the first digit is not equally distributed among the values 1-9. Instead, a first digit of "1" occurs roughly with frequency 30% while the frequency of first digit "9" is just 4.6%.

That's the easy description. The simple questions are: Does it really work? Why does it work? Why *should* it work? Does it ever fail?

In terms of written history, Simon Newcomb published the first known discussion of and solution to this "distribution of digits" problem in 1881.[1] The next development was Frank Benford's analysis of this same topic in 1938.[2] Benford had not been aware of the earlier Newcomb discovery. Newcomb and Benford both provided intriguing data, both reached the same conclusion regarding the mathematical form of the digit distribution, and both failed to convince subsequent investigators of the validity of their explanations. Ralph Raimi provided an excellent history and summary of Benford's Law research to the year 1976.[3] Theodore Hill published a quasi-proof of the Law in 1995.[4] Mark Nigrini has applied Benford's Law in numerous accounting contexts.[5] Rachel Fewster wrote an enjoyable history and some valuable thoughts on applicability of Benford in 2009.[6] Due to his unique insights, our favorite discussion is that of Steven Smith in 2007.[7]

Our primary interest in this topic is the alleged utility of Benford's Law in detecting fraud in financial and accounting contexts. We developed a straightforward understanding of Benford more generally and we share this approach in this article. Ultimately in a future

---

[1] S. Newcomb, "Note on the Frequency of Use of the Different Digits in Natural Numbers," *Am. J. Math.* **4**(1), 39-40, 1881. This is Newcomb (1881).
[2] F. Benford, "The Law of Anomalous Numbers," *Proc. Amer. Phil. Soc.* **78**, 551-72, 1938. This is Benford (1938).
[3] R. Raimi, "The First Digit Problem," *Amer. Math. Monthly* **83**(7), 521-38, 1976. This is Raimi (1976).
[4] T. P. Hill, "A Statistical Derivation of the Significant-Digit Law," *Stat. Sci.* **10**(4), 354-63, 1995. This is Hill (1995). In this paper, Hill described his Theorem 3 as "help[ing] explain and predict the appearance of the logarithmic distribution in significant digits of tabulated data."
[5] See, for example, M. J. Nigrini, "I've Got Your Number," *J. Accountancy*, May 1999. This is Nigrini (1999).
[6] R. M. Fewster, "A Simple Explanation of Benford's Law," *Amer. Stat.* **63**(1), 26-32, 2009. This is Fewster (2009).
[7] S. W. Smith, "Explaining Benford's Law," Chapter 34 in *The Scientist and Engineer's Guide to Digital Signal Processing*, available at http://www.dspguide.com/. This is Smith (2007).

article, we will re-focus on the fraud detection capability and find that practitioners and litigators must exercise caution. There do exist disputes in which Benford's Law is not nearly as appropriate or helpful as some litigants believe.

Benford Warmup

To re-state the Benford problem, first let's find or generate a large dataset of positive numbers. The first non-zero digit of each number will take one of the values k with $k = 1, 2, 3, \cdots, b-1$ where b is the base of the number system. (Conventionally, $b = 10$. We wish to retain flexibility to choose a different base. Retaining b also helps to maintain clarity on the logarithms in the mathematical expressions. All examples and discussion of this article pertain to base $b = 10$ unless we state otherwise. All logarithms are natural, *i.e.*, base *e*, unless we state otherwise as well.) If the dataset conforms with Benford's Law, the distribution of first digits p_k will be:[8]

$$p_k = \log\left(\frac{k+1}{k}\right)/\log b \quad k = 1, 2, 3, \cdots, b$$

(1)

This Benford Distribution is not uniform and, therefore, not intuitive. One's unthinking impression would be that a "9" is as likely to appear as a "1" as a leading digit for a collection of apparently random numbers. But the p_9 and p_1 values from equation (1) are 4.6% and 30.1%, respectively. Thus, a "1" is more than six times as likely to occur as a "9" in the leading digit of a Benford collection of numbers.

[8] As Newcomb (1881) demonstrated, it is little additional work to determine the distributions of the *trailing* digits as well as the leading digit.

Previous authors such as Benford (1938), Raimi (1976), Nigrini (1999), Fewster (2009), and Smith (2007) – and many others we do not cite[9] – provide datasets to show varying degrees of conformance with Benford's Law. There are dataset examples as well that do not comport with Benford. We provide our own "new contribution" here.

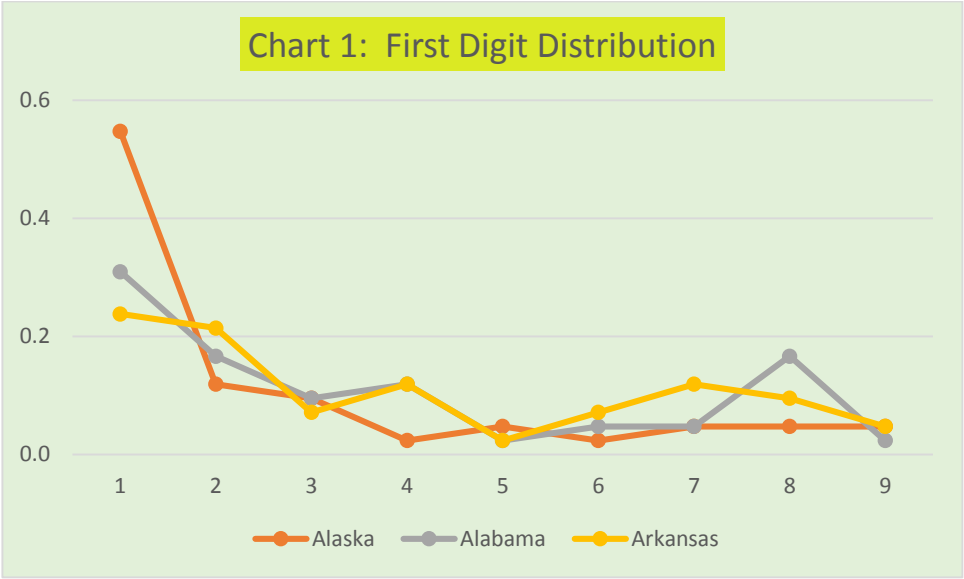
From the *Federal Housing Finance Agency* (FHFA) website we downloaded the average residential mortgage loan amount by state in the U.S. for every year from 1969 through 2010.[10] Given this span of years, there are 42 data points per state and more than 2,000 data points upon aggregation of all the states. Table I below shows the observed first-digit distribution p_k for all aggregated states and also for the first three states:

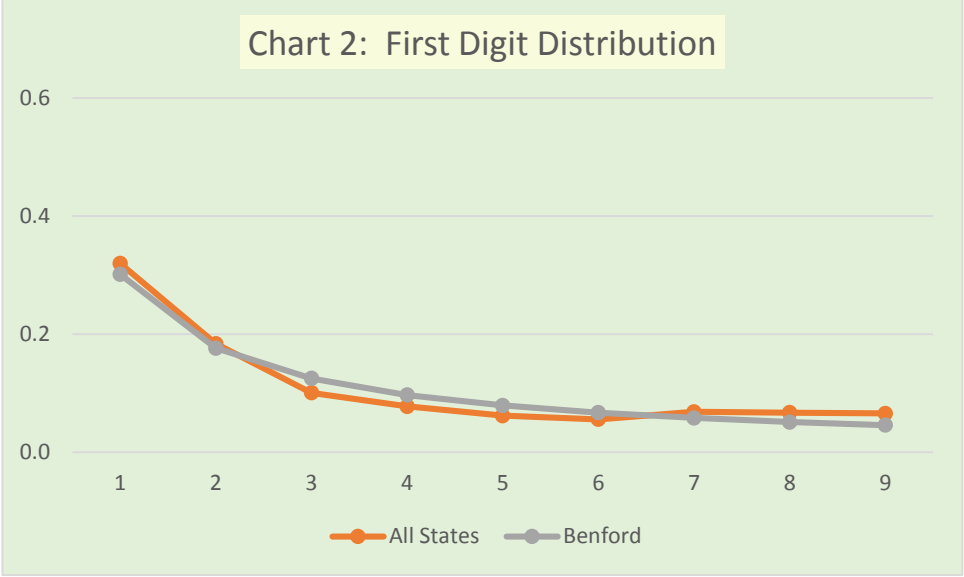
Digit	All States	Alaska	Alabama	Arkansas
1	0.320	0.548	0.310	0.238
2	0.184	0.119	0.167	0.214
3	0.101	0.095	0.095	0.071
4	0.078	0.024	0.119	0.119
5	0.062	0.048	0.024	0.024
6	0.056	0.024	0.048	0.071
7	0.068	0.048	0.048	0.119
8	0.067	0.048	0.167	0.095
9	0.066	0.048	0.024	0.048

Table I: First-Digit Distribution for Average Loan Amount by State for 1969-2010

Charts 1 and 2 below plot the distributions of Table I and the Benford Distribution.

[9] Raimi (1976), Hill (1995), and Fewster (2009) provide good citations to others' empirical data collections. As Nigrini (1999) notes, there exist more than 150 academic papers on Benford's Law published from the 1940's through the end of the twentieth century. An on-line record of more than 600 such papers exists at http://benfordonline.net/ .
[10] See the website http://www.fhfa.gov/Default.aspx?Page=158 .





The distributions for the individual states of Chart 1 differ from each other and do not approximate the Benford Distribution. Chart 2 shows, however, that the aggregated data for all U.S. states is much closer to Benford.[11] This observation that aggregating multiple datasets produces better agreement with Benford is unanimous among investigators and traces back to Benford (1938).

[11] Though "much closer to Benford," as we say, Chart 2 does not show a good fit. The Q statistic of the Pearson Chi-Square Test for the Chart 2 data is 71. Based on the chi-square distribution with eight degrees of freedom, one rejects the null hypothesis (agreement with Benford) with 99% confidence for this Q value since it is greater than 20.09. See, for example, M. H. DeGroot, *Probability and Statistics, 2nd Edition*, Addison-Wesley, 1989.

Benford in the Probability Density Function Perspective

We find it easiest to discuss, understand, and explain Benford's Law by reference to probability density functions (PDFs).[12] For positive real values *x*, the expression $f(x)dx$ with $0 < x < \infty$ is the probability that *x* resides in the domain $(x,\ x + dx)$. The PDF $f(x)$ is non-negative and must satisfy the normalization $\int_0^\infty f(x)dx = 1$. The first important aspect of the PDF is that it's not possible to *generate* example datasets without first specifying the distribution for the example. It is the PDF that provides this specification. To understand Benford, the natural question to study is which PDFs produce Benford distributions and which do not.

In fact, the very simplest PDF does not conform to Benford at all. Let the random variate *X* be uniformly distributed on the interval $x \in (0,1)$. The PDF $f(x) = 1$ when $0 < x < 1$ and is zero for all other values of *x*. The notation *U(0,1)* denotes the uniform distribution on $x \in (0,1)$. The derivation of first-digit probabilities p_k with $k = 1, 2, 3, \cdots, b-1$ is relatively straightforward to express (if not solve) for arbitrary $f(x)$:

$$p_k = \sum_{n=-\infty}^{+\infty} \int_{kb^n}^{(k+1)b^n} f(x)\, dx \qquad (2)$$

Plugging in the simple $f(x)$ for *U(0,1)* and evaluating the summation of equation (2) gives the *intuitive result* $p_k = 1/(b-1)$ for all $k = 1, 2, 3, \cdots, b-1$. That is, here's a case that gives precisely what our intuition suggests rather than Benford's Law. This intuitive case is the simple and ubiquitous uniform distribution *U(0,1)*.

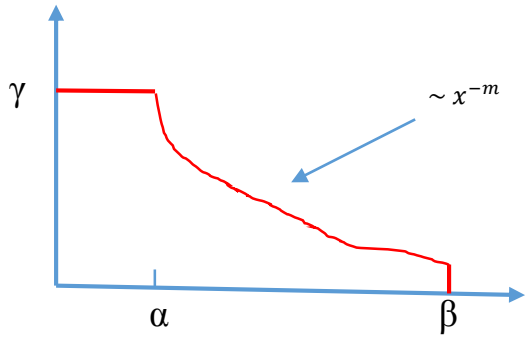
[12] Essentially all investigators touch on the PDF aspect implicitly. Smith (2007) gives much focus to PDFs. Few articles have attempted to compare numerous straightforward PDFs to Benford Law compliance. We are aware of A. K. Formann, "The Newcomb-Benford Law in its Relation to Some Common Distributions," *PLoS One.* **5**(5), May 2010. This is Formann (2010).

This *U(0,1)* case is itself rather special. Had we chosen *U(0,2)* – in which the random variate *X* is uniformly distributed within $x \in (0,2)$ - we would lose the result that all the p_k have equal value.

Instead of *U(0,1)* or *U(0,2)*, let's try *U(0, b^m)* in which the positive integer *m* may be arbitrarily large.[13] Our result of $p_k = 1/(b-1)$ for all $k = 1, 2, 3, \cdots, b-1$ remains unchanged. Thus, if we generate a dataset by selecting positive real values with equal probability up to the arbitrarily large b^m, we find all digits of the resulting dataset have equal probability. Again, this is the intuitive result. For a dataset to exhibit the Benford Distribution of digits (or any other non-uniform distribution), the original dataset must not be of the form *U(0, b^m)*. Hill noted that many mathematicians had attempted "to prove that the log law is a built-in characteristic of our number system – that is, to prove that the set of *all* numbers satisfies the log law ...."[14] Yet our result here is that the set of all numbers from zero to the arbitrarily large b^m will have uniform, rather than logarithmic, distribution of digits.

Since equation (2) enables the derivation of p_k from an assumed PDF $f(x)$, let's consider several more possible PDFs.

Flat-Top Algebraic



[13] The integer *m* may also be negative and be arbitrarily large and negative.
[14] T. P. Hill, "The First Digit Phenomenon," *Amer. Scientist*, 1998.

The PDF $f(x)$ above has constant value for $0 < x < \alpha$ and then declines algebraically as x^{-m} on the domain $\alpha < x < \beta$. The PDF is zero for $x > \beta$. The height of the flat-top is γ and we set this value such that the PDF satisfies the normalization condition. The parameters α and β are arbitrary with the exceptions that they are both integer powers of b and $\alpha \leq \beta$.

For $m = 1$, the equation (2) evaluation of the first-digit probabilities p_k gives:

$$p_k = \frac{\frac{1}{b-1} + \log(\beta/\alpha)\ \log\left(\frac{k+1}{k}\right)/\log b}{1 + \log(\beta/\alpha)}$$

(3)

For $m > 1$, the equation (2) evaluation of the first-digit probabilities p_k for the flat-top algebraic form gives:

$$p_k = \frac{\frac{m-1}{b-1} + \frac{b^{m-1}}{b^{m-1}-1}\,\frac{(k+1)^{m-1} - k^{m-1}}{k^{m-1}\,(k+1)^{m-1}}\left[1 - \left(\frac{\alpha}{\beta}\right)^{m-1}\right]}{m - \left(\frac{\alpha}{\beta}\right)^{m-1}}$$

(4)

Evaluating equations (3) and (4), we get

Digit	Benford	x^-1	x^-2	x^-3
1	0.301	0.244	0.322	0.325
2	0.176	0.156	0.146	0.121
3	0.125	0.121	0.102	0.091
4	0.097	0.101	0.085	0.082
5	0.079	0.089	0.076	0.078
6	0.067	0.080	0.071	0.077
7	0.058	0.074	0.068	0.076
8	0.051	0.069	0.066	0.075
9	0.046	0.066	0.064	0.075

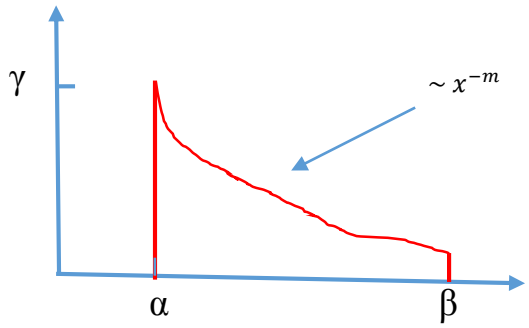
Table II: Flat-Top Algebraic Digit Distributions with $\alpha = 1$ and $\beta = 10$

Digit	Benford	x^-1	x^-2	x^-3
1	0.301	0.277	0.333	0.327
2	0.176	0.168	0.148	0.121
3	0.125	0.123	0.102	0.090
4	0.097	0.099	0.083	0.082
5	0.079	0.083	0.074	0.078
6	0.067	0.073	0.069	0.077
7	0.058	0.065	0.065	0.076
8	0.051	0.059	0.063	0.075
9	0.046	0.054	0.062	0.075

Table III: Flat-Top Algebraic Digit Distributions with $\alpha = 1$ and $\beta = 1000$

None of these cases of Tables II and III show agreement with the Benford Distribution. As is evident in equation (3), the flat-top algebraic form with x^{-1} will approach the Benford Distribution for very large values of β/α.

Pure Algebraic



The PDF $f(x)$ for this "pure algebraic" case above has zero value for $0 < x < \alpha$ and declines algebraically from a maximum value γ as x^{-m} on the domain $\alpha < x < \beta$. The PDF is zero for $x > \beta$. We determine the height γ as the value that normalizes the PDF. The parameters α and β are arbitrary with the exceptions that they are both integer powers of b and $\alpha < \beta$.

For $m = 1$, evaluation of the first-digit probabilities p_k for this pure algebraic form gives:

$$p_k = \log\left(\frac{k+1}{k}\right)/\log b \tag{5}$$

For $m > 1$, the equation (2) evaluation for the pure algebraic case of the first-digit probabilities p_k gives:

$$p_k = \frac{b^{m-1}}{b^{m-1}-1}\,\frac{(k+1)^{m-1}-k^{m-1}}{k^{m-1}(k+1)^{m-1}} \tag{6}$$

Note that the p_k of equation (5) are precisely those of Benford's Law of equation (1)! The "defining PDF" of Benford's Law, then, is $f(x) \sim x^{-1}$ over any interval $x \in (\alpha, \beta)$ such that α and β are integer powers of b and $\alpha < \beta$. Raimi (1976) recognizes that this $f(x) \sim x^{-1}$ is the only density function that satisfies Benford's Law exactly.[15]

Evaluating equation (6), we get

Digit	Benford	x^-2	x^-3	x^-4
1	0.301	0.556	0.758	0.876
2	0.176	0.185	0.140	0.088
3	0.125	0.093	0.049	0.021
4	0.097	0.056	0.023	0.008
5	0.079	0.037	0.012	0.003
6	0.067	0.026	0.007	0.002
7	0.058	0.020	0.005	0.001
8	0.051	0.015	0.003	0.001
9	0.046	0.012	0.002	0.000

Table IV: Pure Algebraic Digit Distributions from Equation (6)

Table IV shows that the behavior of the higher order pure algebraic forms is very far from the Benford Distribution.

[15] Raimi cites R. W. Hamming, "On the distribution of numbers," *Bell Sys. Tech. J.* **49**, 1609-25, 1970.

Positive Normal

The normal PDF has the mathematical form $(2\pi\sigma)^{-1/2}\, exp[-(x-\mu)^2/2\sigma^2]$ in which μ and σ are the mean and standard deviation, respectively. The random variate *X* is permitted to take negative values in this well-known normal distribution. We create the "positive normal" as the positive-*x* portion of the normal distribution with the negative-*x* portion "reflected back" to positive values. More clearly, with $\varphi(x;\mu,\sigma)$ as the normal PDF we stated above, the positive normal PDF is $f(x) = \varphi(x;\mu,\sigma) + \varphi(-x;\mu,\sigma)$ for $x \in (0,\infty)$. We cannot employ equation (2) to derive the p_k analytically, so we rely on Table V to help us determine whether a positive normal PDF comports with Benford's Law.

Digit	Benford	μ = 0	μ = 2	μ = 15
1	0.301	0.359	0.355	1.000
2	0.176	0.129	0.354	0.000
3	0.125	0.086	0.149	0.000
4	0.097	0.081	0.037	0.000
5	0.079	0.078	0.018	0.000
6	0.067	0.073	0.018	0.000
7	0.058	0.069	0.020	0.000
8	0.051	0.065	0.023	0.000
9	0.046	0.060	0.025	0.000

Table V: Positive Normal Digit Distributions from *Monte Carlo* integration with $\sigma = 1$

Table V clearly shows no agreement with the Benford Distribution. The case of $\mu = 15$ is "trivial" in the sense that one expects a normal distribution centered at $x = 15$ with $\sigma = 1$ to be almost entirely enclosed within the domain $x \in (10, 20)$. Thus, it's not surprising to see all the first digits are "1" in this case.

Positive Exponential

The exponential distribution is also well known with PDF $f(x) = \lambda\, exp(-\lambda x)$ for positive λ and $x \in (0,\infty)$. Though simple, this is also a distribution for which we cannot evaluate equation (2) analytically and rely on the *Monte Carlo* integration of Table VI.

Digit	Benford	λ = 1	λ = 2	λ = 5
1	0.301	0.330	0.287	0.297
2	0.176	0.175	0.158	0.194
3	0.125	0.113	0.123	0.135
4	0.097	0.086	0.102	0.098
5	0.079	0.073	0.087	0.076
6	0.067	0.064	0.075	0.061
7	0.058	0.059	0.064	0.052
8	0.051	0.053	0.055	0.046
9	0.046	0.049	0.049	0.041

Table VI: Positive Exponential Digit Distributions from *Monte Carlo* integration

The variance between the positive exponential and Benford digit distributions is not large for some digits and values of λ.

Log-Normal

The log-normal distribution, prevalent in the physical world and in financial markets, emerges from the exponentiation of a normal distribution. The domain of the log-normal PDF is $x \in (0, \infty)$. Though relatively simple, this is also a distribution for which we cannot evaluate equation (2) analytically.

Digit	Benford	σ = 0.5	σ = 1.0	σ = 1.2
1	0.301	0.251	0.298	0.301
2	0.176	0.309	0.184	0.177
3	0.125	0.202	0.130	0.126
4	0.097	0.108	0.098	0.097
5	0.079	0.056	0.078	0.079
6	0.067	0.029	0.064	0.067
7	0.058	0.018	0.055	0.057
8	0.051	0.013	0.049	0.050
9	0.046	0.013	0.044	0.046

Table VII: Log-Normal Digit Distributions from *Monte Carlo* integration with $\mu = 1$

Table VII does show log-normal agreement with Benford when the parameter $\sigma = 1.2$ with $\mu = 1$. The agreement improves further and appears to become numerically exact as σ increases beyond $\sigma = 1.2$. Smith (2007) first noted this conformance of the log-normal digit distribution with the Benford Distribution for specified parameters. Our results are entirely

consistent with Smith (2007), but we have large disagreement with the calculations of Formann (2010).[16]

When is the Data Benford?

In the preceding discussion, we found many example datasets that do not produce Benford Distributions. The "uniform distribution," from *U(0,1)* to *U(0,b^m)*, the "flat-top algebraic" distribution, the $m > 1$ "pure algebraic" distribution, and the "positive normal" distribution are all clearly non-Benford. The "positive exponential" distribution is somewhat closer to Benford in some cases. The "log-normal" distribution converges precisely to Benford as far as one can judge with numerical calculations. The x^{-1} "pure algebraic" distribution is exactly Benford.

We do not believe any previous author has considered all of these example datasets to show Benford versus non-Benford behavior.[17] But virtually all prior research addresses the topic of how to determine which datasets will conform to Benford and which will not. Two clear principles are that the dataset should span several orders of magnitude (*e.g.*, Fewster (2009)) and that *mixtures* of datasets are more likely to produce the Benford Distribution for digits (*e.g.*, Benford (1938), Raimi (1976), and Hill (1995)) than individual datasets.

The empirical requirement to span several orders of magnitude is understandable by reference to example. One of our earlier numerical cases was the normal distribution centered at $x = 15$ with standard deviation $\sigma = 1$. This distribution is almost entirely

[16] Formann (2010) found, as we do, that the log-normal distribution produces a near-Benford Distribution for some values of μ and σ but not for other values. It is in the specific values of these variables that we disagree. To take just one illustrative case, Formann (2010) writes "For μ=0, σ=1 and μ=.33, σ=1 the misfit [to Benford] is massive …" Yet we find that these two cases give agreement to the first-digit Benford Distribution to +/- 5%.

[17] Formann (2010) provided numerical calculations for the *U(0,1)*, positive exponential, and log-normal distributions and also provided some others we do not include.

contained between 10 and 20. The probability of the first digit being "1" is essentially 100%. To have any non-zero probability of finding a "9" as first digit, the distribution must extend upward to 90 and/or downward below 10. Thus, a dataset distribution must at least span one order of magnitude. An example of this sort provides intuition but not a solution.

Smith (2007) provided a creative and convincing condition that a PDF $f(x)$ must satisfy to be consistent with a Benford Distribution. Recall our equation (2) which we re-print here:

$$p_k = \sum_{n=-\infty}^{+\infty} \int_{kb^n}^{(k+1)b^n} f(x)\, dx$$

The limits of integration for each *n* specify the range in *x* pertaining to each digit *k*. Since the widths of these infinite number of "windows" are all equal under a logarithmic transformation, Smith (2007) applied the transformation $Y = \log X$ to work with the PDF $g(y)$ in this transformed variable *y*. Thus, we can write p_k as

$$p_k = \int_{-\infty}^{+\infty} dy\; g(y)\, W_k(y)$$

(7)

in which the $W_k(y)$ is an infinite sum of "window functions" equal to one in the intervals $n \log b + \log k < y < n \log b + \log(k+1)$ and equal to zero otherwise.

At this point, Smith (2007) adds a "scale variable" to measure how the p_k change when all "*x*-space" data values are multiplied by a common factor. In the logarithmic "*y*-space," this scaling is additive. This implied additional property of scale invariance has a long history in the scrutiny of Benford's Law (see Raimi (1976)). Denoting the arbitrary logarithmic shift as α,

$$p_k(\alpha) \;=\; \int_{-\infty}^{+\infty} dy\; g(y-\alpha)\, W_k(y)$$

(8)

Taking the Fourier transform of equation (8) gives the transformed $\tilde{p}_k(\lambda)$ $(= \int_{-\infty}^{+\infty} d\alpha\; p_k(\alpha) e^{-i\lambda\alpha})$ as a product (by the convolution theorem) of two individual transforms representing g and W_k. The transform for W_k gives an infinite sum of Dirac delta functions at evenly spaced frequencies $\lambda = 2n\pi/\log b$ for $n \in (-\infty, +\infty)$. Still following Smith (2007), the criterion for satisfaction of Benford's Law is that the transformed $\tilde{g}^*(\lambda)$ be small or zero at all non-zero "window frequencies" $2n\pi/\log b$ $(n \neq 0)$. Considering the magnitude of relevant terms, we write this "small" condition somewhat subjectively as

$$\left| \int_0^{\infty} dx\; f(x)\; e^{\pm i2n\pi \log x/\log b} \right| \;<\; 0.1$$

for all $n \neq 0$ (9)

after transforming back to the "*x*-space."

Inequality (9) is not the friendliest expression we've ever encountered, but it's tractable in several cases of interest. For example, recall the PDF for the uniform distribution *U(0,1)*: $f(x) = 1$ when $0 < x < 1$ and equals zero for all other values of *x*. Plugging this functional form into the left-hand side of (9), we get $|1 \pm i2n\pi/\log b|^{-1}$. This result clearly does not satisfy the inequality of (9) and, therefore, the distribution *U(0,1)* fails our "Benford Test." Had we employed *U(0, b^m)* rather than *U(0,1)*, we would have obtained precisely the same outcome. Failure of the Benfort Test, of course, is what we expect given our earlier direct calculation of p_k for the uniform distribution.

We can also evaluate (9) for the "pure algebraic" distribution function $f(x)\sim x^{-m}$. For $m = 1$, the left-hand side of (9) is identically zero (with our earlier stipulation that the limits of the domain are integer powers of b). This case passes the Benford Test consistent with our earlier identification of $f(x)\sim x^{-1}$ as providing an exact Benford Distribution. For $m > 1$, the left-hand side of (9) is $(m-1)|1 \pm i2n\pi/\log b|^{-1}$ – which fails the inequality for all $m > 1$. Again, this is what we expect given the numerical values of Table IV.

The last case for which we can evaluate inequality (9) analytically is the log-normal distribution. In terms of the parameters μ and σ of the associated normal distribution, the log-normal PDF $f(x)$ is

$$f(x) \;=\; \frac{1}{\sigma x\sqrt{2\pi}}\; exp\left[\frac{-(\log x - \mu)^2}{2\sigma^2}\right]$$

(10)

Evaluating inequality (9) with this PDF, we find the left-hand side to be[18]

$$\left|\int_0^\infty dx\, f(x)\, e^{\pm i2n\pi \log x/\log b}\right| \;=\; exp[-2n^2\pi^2\sigma^2/(\log b)^2]$$

(11)

Referring to the numerical results of Table VII, we see that the first-digit probabilities p_k move from "not Benford" to "Benford" over the values of σ from 0.5 to 1.0. In equation (11), the value of σ at which the right-hand side equals our (somewhat arbitrary) "0.1 demarcation value" is $\sigma = 0.8$ for $n = 1$. Hence, we conclude that our Benford Test of inequality (9) is supported by these examples we can calculate analytically. Note also that the log-normal Benford Test combining equation (11) and inequality (9) is independent of

[18] Smith (2007) provided a similar expression in the application of his Benford Test to the log-normal distribution.

the parameter μ. Our separate numerical calculations confirm this μ-independent behavior of the convergence of log-normal p_k to the Benford Distribution.

Why is the Data Benford?

There exists a relatively simple explanation for the appearance of the Benford Distribution of digits in sufficiently large and broad datasets. This explanation is entirely consistent with the Hill (1995) theorem regarding a limiting distribution (reminiscent of the Central Limit Theorem). Earlier authors – Newcomb (1881), Benford (1938), Raimi (1976), and others - had also recognized the importance of a mixture of distributions. Let's put this history aside momentarily and consider a different angle.

Reading the first digit of each number in a large dataset is a sequence of mathematical operations. Speaking now in base ten ($b = 10$), this sequence is:

(i) take the base-ten logarithm;

(ii) shift this step (i) value by the unique integer value that produces a result in the (logarithm) range *(0,1)*;

(iii) raise 10 to the power of this step (ii) value (which is the inverse transformation to step (i)); and

(iv) read the integer part of this step (iii) value which will be a number in the range 1, 2, …, 9.

As an example, consider the number 4,371.7. While we can see immediately that the first digit is "4," we must instead apply the mathematical sequence we just defined (or its equivalent) for a coded algorithm. The base-ten logarithm of step (i) is 3.6407 (to five significant digits). To do the translation to the range *(0,1)* of step (ii), we must subtract the integer "3" to get 0.6407. Raising 10 to the power of this value of step (iii) gives 4.3717.

Finally, reading the integer part of 4.3717 (the "int" operation in many computer languages) gives the result of "4" for this step (iv).

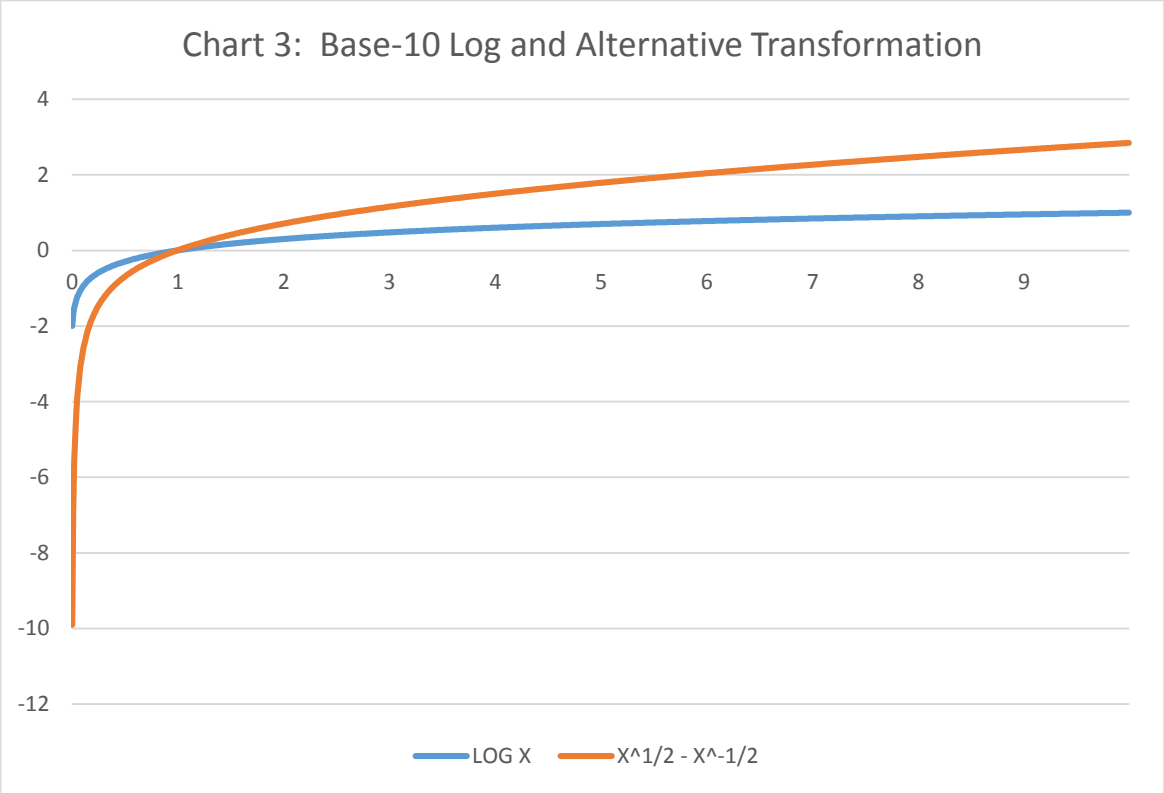
Of these four steps, it is the second "shifting step" that is critical to the Benford Distribution. The shifting step takes the values in every integer range such as *(3,4)*, *(11,12)*, *(-6,-5)*, *et cetera*, and combines them all into the range *(0,1)*. If the ultimate, aggregated collection of values shifted into *(0,1)* is uniform in this range, then the transformation of step (iii) to the *x*-range $x \in (1,10)$ will have a PDF $f(x) \sim x^{-1}$. As we determined earlier, it is this specific PDF that produces the Benford digit distribution.

We consider it reasonable and plausible that the "shifting step" will have the tendency to produce uniform distributions within the logarithm range *(0,1)* when the logarithm of the dataset has elements in many of the integer "bins." That is, when the original dataset spans several orders of magnitude, then several "bins" will be populated and the sum of contents of several "bins" may produce an approximation of a uniform distribution in *(0,1)*. These statements are mere conjecture.[19] Yet we propose the *ansatz* that naturally occurring datasets will tend to produce uniform distributions under the transformation and shifting of steps (i) and (ii) above when the number of "populated bins" is sufficiently large.

Given the *ansatz*, we then claim that the existence of the Benford Distribution is due only to the choice of transformation in step (i) (and its inverse in step (iii)). We get the Benford Distribution (for many but not all datasets) when the step (i) transformation is the base-ten logarithm. We will find an entirely different distribution of first digits when we choose a transformation other than the base-ten logarithm.

[19] We believe that Theorem 3 in Hill (1995) likely suffices to prove these statements for the special case of the base-ten logarithm transformation under Hill's stated assumptions and restrictions.

As one example, let's choose $Y = X^{1/2} - X^{-1/2}$ as the alternative transformation. Chart 3 below plots this function and the base-ten logarithm. Similar to the logarithm, the alternative transformation is defined for all positive *x* and has a range of all real values.



We now apply this alternative transformation and the four-step, first-digit algorithm to the FHFA average loan amount data presented earlier in Table I. Table VIII below shows the prior result in the fourth column (under "Log / Benford") next to the ideal Benford Distribution in the fifth column. The second and third columns show, respectively, the extracted and ideal first-digit probabilities for the alternative transformation. Roughly speaking, the agreement between the actual distribution and the ideal result is similar for the two transformations.

	X^1/2 - X^-1/2		Log / Benford	
Digit	**All States**	**Ideal**	**All States**	**Ideal**
1	0.232	0.248	0.320	0.301
2	0.151	0.157	0.184	0.176
3	0.112	0.121	0.101	0.125
4	0.119	0.101	0.078	0.097
5	0.088	0.089	0.062	0.079
6	0.082	0.080	0.056	0.067
7	0.078	0.073	0.068	0.058
8	0.077	0.067	0.067	0.051
9	0.060	0.063	0.066	0.046

Table VIII: First-Digit Distribution for Average Loan Amount by State for 1969-2012

To avoid confusion, we make a simple but important point here. The first digit distribution we extracted by this alternative transformation is *not* the first digit distribution of the actual raw data. Rather, it is the first digit distribution resulting from the four-step algorithm. For the alternative transformation, the algorithm becomes:

(i) determine $Y = X^{1/2} - X^{-1/2}$ for each raw data point;

(ii) shift this step (i) value by the unique integer multiple of η that produces a result in the (Y) range *(0,η)* – we choose η ($= 10^{1/2} - 10^{-1/2}$) such that the mapping to X is *(1,10)*;

(iii) determine $X = \left(Y + \sqrt{Y^2 + 4}\right)^2/4$ where Y is the step (ii) value while the functional form here is simply the inverse of the transformation in (i); and

(iv) read the integer part of this step (iii) value which will be a number in the range 1, 2, …, 9.

Using the earlier example of the raw number 4,371.7: step (i) produces 66.104 to five significant digits; step (ii) produces 0.64456; step (iii) produces 1.8849; and step (iv) gives "1." So the algorithm's "first digit" is not the first digit of the raw datum point 4,371.7. Only the base-ten logarithm as transformation will produce the *actual* first digit. Rather, we use "first digit" here to measure the cumulative distribution function of X on the domain

(1,10) at the completion of the four-step algorithm. It is this *algorithmic meaning* of "first digit" that explains the Benford Distribution.

Our point here with the algorithm and the *ansatz* and the example is that all transformations of this type will produce non-uniform "first digit" distributions.[20] The final distribution is merely a consequence of applying a mathematical transformation, shifting all the transformed data into a chosen interval to produce (often) a near-uniform distribution, and then performing the inverse transformation to a defined domain in the original variable *X*. The Benford Distribution takes the form of a base-ten logarithm only because this logarithm is the "natural transformation" of our base-ten number system. Observed first-digit distributions are "Benford" or "not Benford" depending on the *ansatz* of *Y*-space uniformity over a finite interval of the transformed and shifted data.

To show one more example for our alternative transformation $Y = X^{1/2} - X^{-1/2}$, we re-visit the log-normal example of Table VII.

	X^1/2 - X^-1/2		Log / Benford	
Digit	***Monte Carlo***	**Ideal**	***Monte Carlo***	**Ideal**
1	0.249	0.248	0.301	0.301
2	0.157	0.157	0.176	0.176
3	0.121	0.121	0.125	0.125
4	0.101	0.101	0.097	0.097
5	0.089	0.089	0.079	0.079
6	0.079	0.080	0.067	0.067
7	0.073	0.073	0.058	0.058
8	0.068	0.067	0.051	0.051
9	0.064	0.063	0.045	0.046

Table IX: Log-Normal Digit Distributions - *Monte Carlo* integration with $\mu = 1, \sigma = 4$

Results in the fourth and fifth columns of Table IX are similar to those of Table VII. The log-normal distribution gives the Benford Distribution of first digits for the chosen parameters under the base-ten logarithm transformation. The second and third columns

[20] As Newcomb (1881) showed, we actually can get marginal and joint distributions for *all* digits. Our repeated reference to "first digit" is simply to facilitate the discussion.

show that the log-normal also gives the ideal first-digit distribution for the alternative transformation. Similar to the earlier discussion, the actual and ideal results will differ for lower values of the standard deviation parameter σ. The convergence range is different for the alternative transformation than it is for the base-ten log transformation. This difference is not surprising since the inequality (9) that produces our Benford Test for equation (11) is dependent on the transformation. We would derive a different test for each transformation.

Benford's Law Implies Scale Invariance

As we stated earlier, many investigators have invoked "scale invariance" to uncover new insights into Benford's Law.[21] In this context, scale invariance means that a dataset with distribution of digits that approximates Benford's Law will also maintain this approximate agreement when an arbitrary scale factor (such as 1.002 or 39.7) multiplies all data points. Following Smith (2007), for example, we assumed scale invariance in writing equation (8) to derive inequality (9) – the Benford Test.

The four-step algorithm of the preceding section explains why observed compliance with Benford's Law implies scale invariance. After the (base-ten logarithm) transformation $Y = \log X$ of step (i) and the step (ii) integer shifting of all *y*-values to the range *(0,1)*, the aggregate distribution within *(0,1)* must be uniform to produce a Benford Distribution in the *x*-values (steps (iii) and (iv)). When this aggregated *y*-value distribution in *(0,1) is* uniform, then any additive shift to the *y*-values *before* the shifting and aggregation to *(0,1)* will *still* produce a uniform distribution on *(0,1)* upon shifting and aggregation. If the aggregated *y*-value distribution in *(0,1)* is *not* uniform, then there exist additive shifts to the *y*-values *before* the shifting and aggregation to *(0,1)* that will change the aggregated distribution on *(0,1)*

[21] See, for example, Raimi (1976), Hill (1995), and Smith (2007).

upon shifting and aggregation. Stated more simply, a Benford-compliant dataset will remain Benford upon a common additive shift to all *y*-values. A dataset that is not Benford-compliant will have a changing digit distribution upon some common additive shifts to all *y*-values.

Since an additive shift to the *y*-values denotes a multiplicative (or scale) shift to the *x*-values, we conclude that a dataset exhibiting a Benford Distribution of digits will continue to show this Benford Distribution under any scale shift. Thus, Benford does imply scale invariance. As an immediate consequence, it is logically valid to assume or require scale invariance for any derivation method or test of Benford properties. We do not need earlier and somewhat vague arguments for scale invariance.[22] Hill proved more rigorously the connection between scale invariance and the (Benford) logarithmic distribution.[23]

Conclusion

To gain a better understanding of the origin and meaning of Benford's Law, we created numerous hypothetical datasets with relatively simple probability density functions. We identified several examples that satisfy Benford's Law and several that do not. In particular, we showed that a uniform distribution spanning an unbounded set of positive real values does not comport with Benford's Law.

Building on the prior work of Smith (2007), we created a "Benford Test" to determine from the functional form of a probability density function whether the resulting distribution of digits would find close agreement with Benford's Law. Applying this test to several examples showed complete consistency.

[22] Raimi (1976) describes an early argument for scale invariance: "If the first digits of all the tables in the universe obey some fixed distribution law … that law must surely be independent of the system of units chosen." He cites R. S. Pinkham, "On the distribution of first significant digits," *Ann. Math. Statist.* **32**, 1223-30, 1961.
[23] T. P. Hill, "Base-Invariance Implies Benford's Law," *Proc. Amer. Math. Soc.* **123**(3), 887-95, 1995.

By deliberate specification of an algorithm to generate the first-digit distribution, we discovered that this algorithm is characterized by a mathematical transformation and its inverse. While the base-ten logarithm is the "natural transformation" of the base-ten number system, one may also specify alternative transformations and create "generalized Benford Laws" for the distribution of "algorithmic first digits." We found also that Benford's Law and the generalizations due to other transformations rest on an *ansatz* of uniform distribution in the transformed variable following the mathematical transformation and shifting. We provided several examples to support this explanation.

Consider, then, a naturally occurring dataset for which Benford's Law provides an accurate distribution of first (or other) digits. The "deeper meaning" of Benford is simply that a logarithmic transformation of the data coupled with the "shifting operation" that aggregates data over different orders of magnitude often produces a near-uniform distribution for the particular dataset in the reduced "y-space" range. This interpretation of Benford's Law does not diminish its effectiveness as a simple indicator for potentially fraudulent data.