

BENFORD'S LAW, FAMILIES OF DISTRIBUTIONS AND A TEST BASIS

JOHN MORROW†

This Draft: January 22, 2009

First Draft: August 6, 2006

ABSTRACT. This paper presents improved, asymptotically valid test values for Benford's Law, a particular distribution of First Significant Digits. Also derived are new test values for heuristic tests used in the literature and small sample properties of the tests are investigated. Since not all data should be expected to satisfy Benford's Law, a simple method is presented by which all continuous distributions may be transformed to satisfy Benford with arbitrary precision and induce *scale invariance*, one of the properties underlying Benford's Law in the literature. This allows application of Benford tests to arbitrary samples, a hurdle to current empirical work. The results yield improved tests for Benford's law applicable to a broader class of data.

JEL Codes: C10, C24, C46

AMS Classification: 62E20, 62F12

Acknowledgement. I thank George Judge, Thomas Kurtz and Laura Schechter for helpful comments, guidance and encouragement. This paper has also benefited from discussions with William Brock, Swati Dhingra, Ching-Yang Lin and Mian Zhu.

† UW-Madison, Contact: morrow1@wisc.edu. A utility for testing Benford's Law is available from www.checkyourdata.com. Comments on the paper or utility are much appreciated.

1. INTRODUCTION

Benford’s Law states that for commonly observed empirical data, regularities should occur in the First Significant Digits (FSD)s of the data. The FSD of a number x is the leading digit of x in the base 10 numbering example, for instance

$$\text{FSD of } \pi = 3 \text{ since } \pi = \underbrace{3}_{\text{FSD}}.14159\dots$$

In its strong form, Benford’s law says for the FSDs $\{1, \dots, 9\}$, the frequency observed of each digit $d \in \{1, \dots, 9\}$ should be approximately $\log_{10} \left(1 + \frac{1}{d}\right)$. Many papers have detailed occurrences of Benford’s Law, and the Law has also been used to test for fraud and error present in a variety of contexts¹. A few papers have also categorized properties that characterize distributions satisfying Benford’s Law, or found distribution families which satisfy it for particular parameter values². Unfortunately, no *general principle* has been found to explain the Benford phenomenon in data, nor provide general criteria as to when to expect Benford’s Law to hold. This paper contributes towards efforts providing such a general principle and focuses on the testing issues that arise when assessing conformance with Benford’s Law.

Testing for Benford’s Law has recently been performed on a variety of data sets, in the broad context of detecting fraud. This paper focuses on two testing issues. The first is the suitability of existing tests which have been used in the literature. Such tests are too conservative and consequently Section 2 derives new asymptotically valid test values which allow for more powerful tests and evaluates small sample values of the tests. Measures of fit have also been used as “rules of thumb” to check concordance with Benford’s Law. Section 2 also provides a new interpretation for such measures and derives critical values for hypothesis testing. The second testing issue is the application of tests on data which inherently does not satisfy the law.³ Clearly, rejection of tests for Benford on data which inherently fails the law will not help uncover fraud or error. Section 3 develops results which allow the transformation of data sets to satisfy Benford within arbitrary precision thereby allowing application of the above tests to any sample. Section 4 provides a discussion of the main results and applies them to distribution families of interest and concludes.

¹For occurrences see Benford (1938); Giles (2007); Berger and Hill (2007). Examples using Benford’s law for fraud and error detection include tax fraud Nigrini (1996), reliability of survey data Judge and Schechter (2007), environmental law compliance Marchi and Hamilton (2006) and campaign finance Cho and Gaines (2007).

²For characterizations of general properties see Hill (1995b); Boyle (1994); Allaart (1997) and for distribution families see Scott and Fasli (2001); Leemis et al. (2000).

³For a discussion regarding when testing the Law is appropriate, see Durtschi et al. (2004).

2. TESTING AND BENFORD'S LAW

One of the most popular applications of Benford's Law is fraud detection and testing of data quality. A few tests have been constructed, and new tests recently proposed, but at present it appears that properties of the estimators themselves are not well understood. In fact, asymptotic results indicate that the test values used in some recently published papers are too conservative for the significance levels reported (e.g. Giles (2007); Cho and Gaines (2007)). More importantly, such tests appear rather *ad hoc* and the power of such tests appears to be almost wholly unexamined. I now discuss the four tests in use, provide asymptotically valid test values, and explore their small sample properties.

2.1. Popular Tests in Use. Pearson's χ^2 test is a natural candidate for testing whether an observed sample satisfies Benford's Law, however, due to its low power for even moderately small sample sizes it is often unsuitable. Consequently, other tests have been devised, and commonly used tests for conformance with Benford's Law include the Kolmogorov-Smirnov test and the Kuiper test. More recently Leemis et al. (2000) have introduced the statistic m (max).

$$m \equiv \max_{d \in \{1, \dots, 9\}} \left| \Pr(X \text{ has FSD} = d) - \log_{10} \left(1 + \frac{1}{d} \right) \right|$$

Similarly, Cho and Gaines (2007) propose the d (distance) statistic.

$$d \equiv \left[\sum_{d \in \{1, \dots, 9\}} [\Pr(X \text{ has FSD} = d) - \log_{10} \left(1 + \frac{1}{d} \right)]^2 \right]^{1/2}$$

2.2. Issues with current tests in use: Kolmogorov-Smirnov and Kuiper. The χ^2 , Kolmogorov-Smirnov (D_N) and Kuiper (V_N) tests for a sample of size N appear to be the most common tests in use. In fact, latter two have a "correction factor" introduced by Stephens (1970) which when applied to such tests produce fairly accurate test statistics regardless of sample size. Denote these tests with the correction factor applied as D_N^* and V_N^* , respectively. For instance, for the modified Kuiper test V_N^* presented in Stephens, a 99% confidence set is produced by all samples $\{X_i\}$ such that $V_N^* < 2.001$. However, such tests are based on the null hypothesis of a continuous distribution, and are generally conservative for testing discrete distributions as discussed by Noether (1963). A simple example in the appendix shows that such critical values derived for continuous distributions can be *extremely* conservative. The example involves the critical region for the V_N^* test at 99%

significance, which generates in fact a .99994% critical region. The example illustrates that tests based on the assumption of a continuous distribution for Benford’s law (which is discrete) can have substantially lower power than test values analytically derived.

The Stephens (1970) test values for the modified Kuiper (D_N^*) and Kolmogorov-Smirnov (V_N^*) tests at commonly used significance levels are reported in the first column of Table 1. New asymptotically valid test values under the specific null hypothesis that Benford’s Law holds are in the second column of Table 1. These test values are derived from an application of the CLT to a multivariate Bernoulli variable which corresponds to a random variable which exactly satisfies Benford’s Law. Inspection shows that in fact the test values based on the assumption of a continuous underlying distribution are too high, and thus too conservative.⁴ Furthermore, the test statistics as in Table 1 allow easy computation of the relevant test as well as evaluate existing published literature.

TABLE 1. Continuous vs Benford Specific Test Values

Test Statistic	Continuous			Benford Specific		
	$\alpha = .10$	$\alpha = .05$	$\alpha = .01$	$\alpha = .10$	$\alpha = .05$	$\alpha = .01$
Kuiper Test (V_N^*)	1.620	1.747	2.001	1.191	1.321	1.579
KS Test (D_N^*)	1.224	1.358	1.628	1.012	1.148	1.420

2.3. The m and d tests and critical values. As far as the m and d tests are concerned, no test values have yet to be reported for use addressing the above issues. In order to derive asymptotic test statistics, define the modified test statistics m_N^* and d_N^* given in Equation (2.1), where N is the number of observations.

$$(2.1) \quad m_N^* \equiv \sqrt{N} \cdot m \quad d_N^* \equiv \sqrt{N} \cdot d$$

The reason for the appearance of the \sqrt{N} term is as follows. The true FSD frequencies $\Pr(X \text{ has FSD} = k)$ correspond to Bernoulli parameters as do the Benford $\log_{10}(1 + \frac{1}{k})$ terms. Letting $\mathbf{1}_{FSD=k}(X)$ be the indicator that X has a FSD equal to k , the random vector

$$T_N \equiv \left[\overline{\mathbf{1}_{FSD=1}(X)} - \log_{10}\left(1 + \frac{1}{1}\right) \quad \dots \quad \overline{\mathbf{1}_{FSD=8}(X)} - \log_{10}\left(1 + \frac{1}{8}\right) \right]$$

is iid and by the CLT, $\sqrt{N}T_N$ converges in distribution to a $N(0, \Sigma)$ random variable. Both m_N^* and d_N^* can be formed as continuous mappings of $\sqrt{N}T_N$ in which the \sqrt{N} term can be slipped

⁴In fact, Conover (1972) shows how one may construct appropriate tests based on the Kolmogorov-Smirnov test for discrete distributions, but this approach suffers from the joint problems of being (1) rather involved and (2) computationally expensive relative to all the other suggested tests.

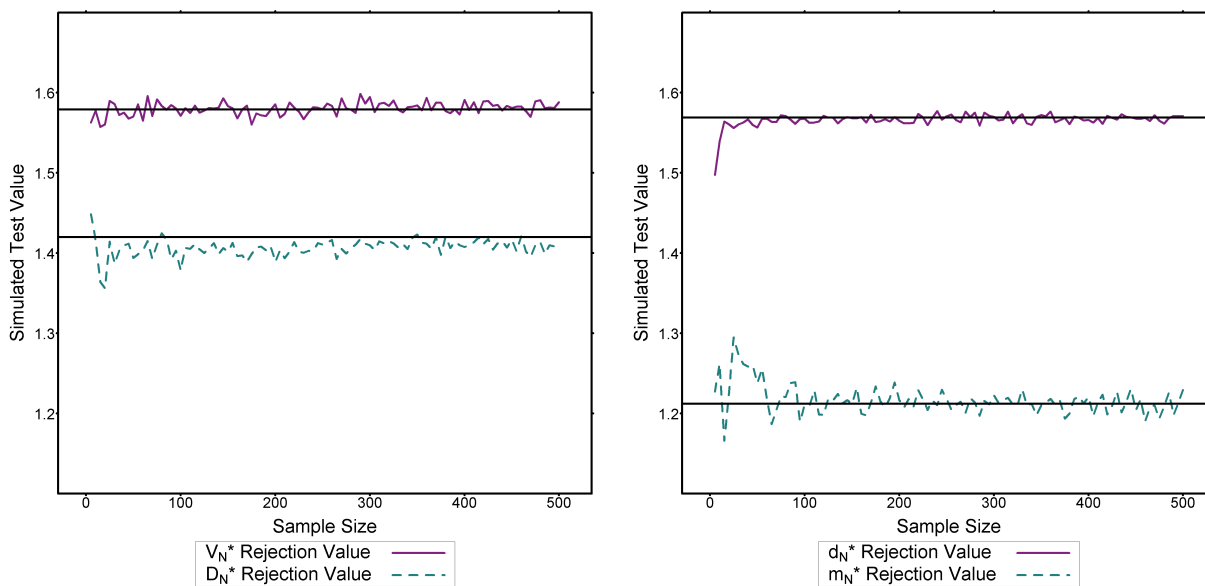
outside of the m and d terms since these functions are homogeneous. The end result in both cases is convergence in distribution to a continuous function of a $N(0, \Sigma)$ variable. Rejecting the null that Benford's Law holds when m_N^* and d_N^* are large provides a consistent test statistic (e.g. van der Vaart (2000, Lemma 14.15)). Rejection regions for common test levels are provided in Table 2.

TABLE 2. m^* and d^* Test Values

Test Statistic	Asymptotic Test Level		
	$\alpha = .10$	$\alpha = .05$	$\alpha = .01$
Max Test (m_N^*)	0.851	0.967	1.212
Distance Test (d_N^*)	1.212	1.330	1.569

Naturally, the question arise of how good the critical values reported in Tables 1 and 2 are in practice for small sample sizes. For sample sizes $N \leq 500$ I have numerically computed the appropriate test values for a level $\alpha = .01$ test for all four statistics as shown in Figure 1, based on 10^5 simulated draws for each sample size. The Figure contains numerically obtained test values in sample size increments of $N = 5$, and horizontally superimposed are the asymptotic test values for each test. The small N performance is fairly good in that the simulated test statistics are very close to the asymptotic values. This shows that the critical regions in Table 2 are reasonable for small as well as large N .

FIGURE 1. m_N^* and d_N^* Test Values for Small Samples



(a) Kuiper and KS Tests

(b) Max and Distance Tests

3. ASSURING CONFORMITY TO BENFORD’S LAW

The general approach of using Benford’s Law for fraud detection is to compare FSD frequencies in sample data with the Law. Of course, whether Benford’s Law holds for a particular sample depends upon the underlying distribution. Therefore testing for Benford is restricted by the underlying properties of data. One of the major obstacles in using this approach is that often the distribution one would like to test does not remotely satisfy Benford’s Law, regardless of data quality (see Table 3). The results in this section ameliorate this issue by developing a transformation (Theorem 1) that may be applied to data that induce compliance with Benford’s Law. The implications of Theorem 1 are further developed in the next Section.

Before applying tests based on Benford’s Law to a random variable X , one should first expect that X is approximately satisfies Benford. This idea is formalized in the following Definition.

Definition. A random variable X ϵ -satisfies Benford’s Law if for all FSDs d

$$|\Pr(X \text{ has FSD} = d) - \log_{10}(1 + \frac{1}{d})| < \epsilon$$

Before applying the tests in Section 2 it is necessary to ensure that the sample ϵ -satisfies Benford’s Law. This is best illustrated with an example. Consider a sample S composed of two sub-samples, S_H and S_C and hypothesize S_H comes from an “Honest” data source while S_C comes from “Cheaters”. The underlying assumption for fraud detection is that S_H is closer to satisfying Benford than S_C . But to apply the tests of Section 2, a minimum requirement is that S_H is approximately Benford, one option being that X ϵ -satisfies Benford’s Law. If the sample S could be transformed in some way to satisfy the Law so that S_H satisfies the Law while S_C fails, the transformation would be a basis for detecting anomalies in S_C . The main result of this Section, Theorem 1, provides such a means of transforming S .⁵

Theorem 1 (Exponential-Scale Families). *Let X be a random variable with continuous pdf and fix $\epsilon > 0$. There is an α^* such that for all $\alpha \geq \alpha^*$:*

$$(X/\sigma)^\alpha \quad \epsilon - \text{satisfies Benford's Law for all } \sigma$$

In light of the above discussion if one is fairly confident about the distribution of X (say, using a Kernel Density Estimate), one strategy is to apply Theorem 1 to transform X to ϵ -satisfy Benford’s

⁵For another line of thought, that of generalizing the class of FSD distributions considered, see for instance Rodriguez (2004); Grendar et al. (2007); Hurlimann (2006).

Law and then perform tests. Methods for computing sufficiently large α follow from the intermediate results in this Section. To be concrete, suppose we have a random sample $\{X_i\}$ and we feel confident that $\frac{X-\mu}{\sigma} \sim N(0, 1)$, perhaps by estimating μ and σ from the sample. There are several values of μ and σ where we should not expect that the sample will obey Benford's Law. However, fix any $\epsilon > 0$ and from Theorem 1 we know there is an $\alpha(\epsilon)$ such that for $Y \sim (\frac{X-\mu}{\sigma})^{\alpha(\epsilon)}$, the FSD frequencies observed in Y should be within ϵ of Benford's Law. A sufficiently large $\alpha(\epsilon)$ may be calculated from the distribution of X using the techniques below. Accordingly, m_N^* and d_N^* calculated with Y in place of X should be close to zero, allowing for detection of anomalous observations. This Section proceeds with intermediate steps leading up to a proof of Theorem 1.

3.1. Approximation by step functions. The following definition has an important relationship with Benford's Law, as will be shown shortly.

Definition. Let Y be a random variable with pdf $f(y)$. Fix $\epsilon > 0$ then Y can be ϵ -approximated by integer step functions, denoted $Y \in I(\epsilon)$ if there exist $\{c_i\}$ s.t. for every measurable set $A \subset \mathbb{R}$

$$|\int_A f(y)dy - \int_A \sum c_i \mathbf{1}_{[i, i+1)}(y)dy| \leq \epsilon$$

For example, by taking $c_i \equiv 0$ for any Y , $Y \in I(1)$. Although the definition of $I(\epsilon)$ is simple, any continuous random variable X for which $\log_{10} X \in I(\epsilon)$ "approximately" satisfies Benford's Law. The formal statement of this fact is Lemma 1.

Lemma 1. *Suppose X is a random variable on $(0, \infty)$ with continuous pdf and let $Y \sim \log_{10} X$. If $Y \in I(\epsilon)$ then X ϵ -satisfies Benford's Law.*

Proof. See Appendix. □

This lemma provides a check of whether a random variable X ϵ -satisfies Benford's law by checking whether $\log_{10} X \in I(\epsilon)$. Since Lemma 1 will be the workhorse throughout the rest of the paper, some remarks on its hypotheses are in order. First, the assumption of a continuous pdf is fairly mild and examination of the proofs shows it can be weakened, but this assumption will be maintained for brevity. Second, the restriction that the random variable in question has support on $(0, \infty)$ rather than all of \mathbb{R} is really not an imposition since with respect to FSD frequencies, the FSDs of X are identical to the FSDs of $|X|$.

3.2. Characterization of $I(\epsilon)$. The simplicity of the definition of $I(\epsilon)$ allows for a precise characterization of the least ϵ s.t. $X \in I(\epsilon)$. By the definition of $I(\epsilon)$, $X \in I(\epsilon)$ requires that

$$(3.1) \quad \sup_{A \text{ measurable}} \left| \int_A f(y)dy - \int_A \sum c_i \mathbf{1}_{[i,i+1)}(y)dy \right| \leq \epsilon$$

In solving for the best choice of $\{c_i\}$ it suffices to consider each interval $[i, i + 1]$ individually. Surprisingly, the solution to these individual problems is quite simple in that the optimal c_i turn out to be the gross estimates $c_i \equiv \int_{[i,i+1]} f(x)dx$. These c_i are optimal because of the “maximin” nature of Equation (3.1): the best c_i must minimize integrals of the form $|\int_A [f(y) - c_i]_- dy|$ and $|\int_A [f(y) - c_i]_+ dy|$. Following this idea leads to a proof of Lemma 2.

Lemma 2. *Let f be an L^1 function. Then*

$$\arg \min_c \sup_{A \text{ measurable}} \left| \int_{[0,1] \cap A} [f(x) - c] dx \right| = \int_{[0,1]} f(x) dx$$

and for $c^* \equiv \int_{[0,1]} f(y)dy$, the minimum attained is $\frac{1}{2} \int_{[0,1]} |f(x) - c^*| dx$.

Proof. See Appendix. □

A first consequence of Lemma 2 is that for random variables X_k with pdfs of the form $f(x) = k \mathbf{1}_{[0, \frac{1}{k}]}$, $X_k \in I(1 - \frac{1}{k})$ so considering large k , nothing can be said about $X \in I(\epsilon)$ for $\epsilon < 1$ without more information about the distribution of X . A second consequence of Lemma 2 is that choosing the optimal $\{c_i\}$ allows computation of the least ϵ such that $X \in I(\epsilon)$ directly. This characterizes the sets $I(\epsilon)$ completely, a consequence stated as Theorem 2.

Theorem 2. *Let X be a random variable with pdf f . Then the least ϵ s.t. $X \in I(\epsilon)$ is given by*

$$(3.2) \quad \epsilon = \sum_i \int_{[i,i+1]} \left| f(x) - \int_{[i,i+1]} f(t) dt \right| dx$$

Proof. Application of Lemma 2 on each interval $[i, i + 1]$. □

Paired with Lemma 1 this forms a method to test for conformance with Benford’s Law within a parametric family using analytic methods: take any random variable X with parameters θ , find the pdf of $\log_{10} X$, say g , and solve Equation (3.2) for g . Intuitively, for parameters θ where g is fairly “flat,” $\int_{[i,i+1]} |g(x) - \int_{[i,i+1]} g(t) dt| dx$ is fairly small. Lemma 1 implies that X will ϵ -satisfy Benford’s Law for such θ . These results provide precise analytical tools to find parameters θ for X which will induce Benford’s Law.

3.3. Transformations and $I(\epsilon)$. By virtue of the fact $Y \in I(\epsilon)$ means Y can be approximated by integer step functions, integer shifts and scaling of Y preserve the ability to approximate Y by

integer step functions. In particular for $a, b \in \mathbb{Z}$, let $Z \equiv aY + b$ and then Z can be approximated by translating the $\{c_i\}$ used to approximate Y . The details are routine and left to the appendix, but the new approximation will guarantee $Z = aY + b \in I(\epsilon)$. Since this holds for all $a, b \in \mathbb{Z}$, $I(\epsilon)$ is invariant under such transformations as summarized in Lemma 3.

Lemma 3. $Y \in I(\epsilon)$ iff $aY + b \in I(\epsilon)$ for all $a, b \in \mathbb{Z}$ with $a \neq 0$.

Proof. See Appendix. □

The last step towards proving Theorem 1 is a method of transforming any random variable within its mean-scale family so that the transformed variable is in $I(\epsilon)$ for arbitrary ϵ . This result is given in Theorem 3 and is followed by a sketch of the proof.

Theorem 3 (Mean-Scale Approximation). *For any random variable Y with continuous pdf, and $\epsilon > 0$ there exists a $s \in \mathbb{R}$ s.t. $\sigma \leq s$ implies*

$$Y/\sigma \in I(\epsilon)$$

Additionally $\forall \mu \in \mathbb{R}$,

$$2(Y - \mu)/\sigma \in I(\epsilon)$$

Proof. See Appendix. □

The basic idea of the proof is as follows. To show that $Y/\sigma \in I(\epsilon)$ consider σ as a transformation that flattens out the pdf of Y/σ as $\sigma \rightarrow 0$. Once Y/σ is sufficiently flattened out, approximate its pdf via constants $\{c_i\}$ which correspond to appropriately chosen elements of a Riemann sum, giving an ϵ approximation to the pdf. In order to show $2(Y - \mu)/\sigma = 2Y/\sigma - 2\mu/\sigma \in I(\epsilon)$ appeal to Lemma 3 to argue that without loss of generality $2\mu/\sigma \in [0, 1]$. Finally, show that smoothing Y/σ further to $2Y/\sigma$ is enough that the improved approximation absorbs the $2\mu/\sigma$ term.

3.4. Proof of Theorem 1. With the above results, it is a simple step to get to the main result of the section, Theorem 1. Let X and Y be as in the hypotheses of Theorem 1 and first assume X has positive support. Then

$$\log_{10} |X/\sigma|^\alpha = [\log_{10} X - \log_{10} |\sigma|] / [1/\alpha]$$

so from Theorem 3 for sufficiently large α^* , for all $\alpha \geq \alpha^*$, $\log_{10} (X/|\sigma|)^\alpha \in I(\epsilon)$ for all σ . The result then follows from an application of Lemma 1. If X has positive and negative support a similar argument applies to $|X|$.

4. DISCUSSION: EXPONENTIAL-SCALE FAMILIES

This section discusses additional implications of Theorem 1. For ease of reference the theorem is restated here:

Theorem. *Let X be a random variable with continuous pdf and fix $\epsilon > 0$. There is an α^* such that for all $\alpha \geq \alpha^*$ $(X/\sigma)^\alpha$ ϵ -satisfies Benford's Law for all σ .*

Another way of stating this result is that the exponential transformation $g(x) = x^\alpha$ induces conformity to Benford's Law for all sufficiently large α . More surprising is that this transformation simultaneously induces approximate *scale invariance*, in that $(X/\sigma)^\alpha$ satisfies Benford's Law for any scaling parameter σ . Scale invariance is one of the fundamental properties that distributions satisfying Benford's Law should have.⁶ Earlier work has detailed experimental evidence of high exponents of random variables to tend to conform to Benford's Law independent of scale (see for instance (Scott and Fasli, 2001) who find that the Log-Normal distribution satisfies the Law for $\sigma \gtrsim 1.2$).

Raising a random variable Y to the power α has the effect of leveling out the pdf of $\log_{10} Y^\alpha$. Looking back to Theorem 2, this has the effect of scaling the $\int_{[i,i+1]} |f(x) - \int_{[i,i+1]} f(t)dt|dx$ terms in Equation (3.2) to $\int_{[i,i+1]} |f(x/\alpha)/\alpha - \int_{[i,i+1]} f(t/\alpha)/\alpha dt|dx$ thereby improving the approximation. More generally, any transformation g which has this effect on $\log_{10} Y$ will eventually make $g(Y)$ ϵ -satisfy Benford's Law. However, the particular transformation $g(x) = x^\alpha$ is of interest due to its simplicity and relevance for commonly modelled distributions. FSD frequencies of common distributions are contrasted the with the same distributions raised to the tenth power in Table 3.

TABLE 3. FSD Frequencies (Sample Size= 10^7)

	<i>First Significant Digit</i>									Max Dev.	Upper Bound
	1	2	3	4	5	6	7	8	9		
Benford's Law	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	0.000	0.000
Normal(0,1)	0.359	0.129	0.087	0.081	0.077	0.073	0.069	0.064	0.060	0.058	0.673
Uniform(0,1)	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.190	0.538
Log-Normal(0,1)	0.308	0.170	0.119	0.094	0.079	0.068	0.060	0.053	0.048	0.007	0.547
Exponential(1)	0.330	0.174	0.113	0.086	0.072	0.064	0.058	0.053	0.049	0.029	0.520
Pareto(1,1)	0.556	0.185	0.093	0.056	0.037	0.026	0.020	0.015	0.012	0.255	0.538
Normal(0,1) ¹⁰	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	0.000	0.056
Uniform(0,1) ¹⁰	0.277	0.171	0.126	0.100	0.084	0.072	0.063	0.056	0.051	0.024	0.058
Log-Normal(0,1) ¹⁰	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	0.000	0.046
Exponential(1) ¹⁰	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	0.000	0.042
Pareto(1,1) ¹⁰	0.326	0.180	0.123	0.093	0.075	0.062	0.053	0.046	0.041	0.025	0.058

⁶For a more formal definition of *scale invariance* and results relating to it see (Raimi, 1976; Hill, 1995a).

Table 3 shows a striking convergence of FSDs to Benford’s Law following the transformation of being raised to the tenth power. The Max Dev. column lists the maximum FSD frequency deviation from the Benford prediction for each row, showing that even the Uniform(0,1)¹⁰ distribution obeys Benford’s Law reasonably well. The Upper Bound column lists the Upper Bound on deviation from Benford’s Law given by Theorem 2. Although this bound is not terribly good for the distributions in the first five rows of Table 3, they become reasonable for the second five rows after the transformation x^{10} is applied.

4.1. Particular Families. As illustrated by the Log-Normal case, it is a natural question to ask which families of distributions will satisfy Benford’s law for particular parameter values. From Theorem 1, a natural way to start looking is to find families of a variable X where X^s is again within the family. Three such common families are the Log-Normal, Weibull, and Pareto distributions. The effect of a transformation of $X \rightarrow (X/\nu)^s$ within these families are summarized in Table 4. Theorem 1 shows it is no coincidence that the Log-Normal and Pareto families appear in the Table and the literature on scaling laws. If such distributions commonly occur in data, since for particular parameter values Theorem 1 applies, Benford’s Law will be commonly observed in samples drawn from these distributions as well.

TABLE 4. Families Closed under Powers

Distribution	Functional Form	X Dist.	$(X/\nu)^s$ Dist.	$\text{Var}(X)$
		Parameters	Parameters	
Log-Normal	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2}$	(μ, σ)	$(s\mu - \ln \nu, s\sigma)$	$(e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$
Weibull	$\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k}$	(k, λ)	$(k/s, \lambda^s/\nu)$	$\lambda^2[\Gamma(1 + \frac{2}{k}) - \Gamma(1 + \frac{1}{k})^2]$
Pareto	$kb^k x^{-(k+1)}\mathbf{1}_{[b,\infty)}(x)$	(k, b)	$(k/s, b^2/\nu)$	$b^2k/[(k-1)^2(k-2)]$

For example, according to Table 4, if X is distributed Log-Normal(μ, σ^2) then $(X/\nu)^s$ is distributed Log-Normal($s\mu - \ln \nu, s^2\sigma^2$). Appealing to Theorem 1, $(X/\nu)^s$ ϵ -satisfies Benford’s Law for sufficiently large s , or equivalently X ϵ -satisfies Benford’s Law for sufficiently large σ^2 . Consequently, for each distribution in Table 4 and $\epsilon > 0$ there is a region in the parameter space where the distribution will ϵ -satisfy Benford’s Law. Referring to the Variance column in Table 4 this is roughly when the variance or shape parameter is sufficiently large. Theorem 1 implies that the transformed variables $(X/\nu)^s$ will ϵ -satisfy Benford’s Law for sufficiently large s and any ν . This formally confirms observations by Leemis et al. (2000, pg. 237) that increases in the shape parameter increase compliance with Benford’s Law.

4.2. Conclusion. This paper derives new test values and improves upon existing tests for evaluating compliance with Benford’s Law. Also provided new results which broaden the range of data to which such tests can be applied through a simple transformation. This transformation also induces scale invariance with respect to compliance with Benford’s Law which frees tests from dependence of choice of measurement units. Methods in this paper may also be used to characterize precisely which particular members of a family of distributions satisfy Benford’s Law, and have particularly clean implications for the Log-Normal, Weibull, and Pareto families.

REFERENCES

- Allaart, Pieter C.**, “An Invariant-Sum Characterization of Benford’s Law,” *Journal of Applied Probability*, 1997, *34* (1), 288–291.
- Benford, Frank**, “The Law of Anomalous Numbers,” *Proceedings of the American Philosophical Society*, 1938, *78* (4), 551–572.
- Berger, A. and T. P. Hill**, “Newton’s Method Obeys Benford’s Law,” *American Mathematical Monthly*, 2007, *114* (7), 588–601.
- Boyle, Jeff**, “An Application of Fourier Series to the Most Significant Digit Problem,” *The American Mathematical Monthly*, 1994, *101* (9), 879–886.
- Cho, W. K. T. and B. J. Gaines**, “Breaking the (Benford) law: Statistical fraud detection in campaign finance,” *The American statistician*, 2007, *61* (3), 218–223.
- Conover, W. J.**, “A Kolmogorov Goodness-of-Fit Test for Discontinuous Distributions,” *Journal of the American Statistical Association*, 1972, *67* (339), 591–596.
- Durtschi, C., W. Hillison, and C. Pacini**, “The Effective Use of Benford’s Law to Assist in Detecting Fraud in Accounting Data,” *Journal of Forensic Accounting*, 2004, *5*, 17–34.
- Giles, D. E.**, “Benford’s law and naturally occurring prices in certain ebaY auctions,” *Applied Economics Letters*, 2007, *14* (3), 157–161.
- Grendar, M., G. Judge, and L. Schechter**, “An empirical non-parametric likelihood family of data-based Benford-like distributions,” *Physica A: Statistical Mechanics and its Applications*, 2007, *380*, 429–438.
- Hill, T. P.**, “A Statistical Derivation of the Significant-Digit Law,” *Statistical Science*, 1995, *10* (4), 354–363.
- Hill, Theodore P.**, “Base-Invariance Implies Benford’s Law,” *Proceedings of the American Mathematical Society*, 1995, *123* (3), 887–895.
- Hurlimann, W.**, “Generalizing Benford’s law using power laws: application to integer sequences,” *Arxiv preprint math.ST/0607166*, 2006.
- Judge, George and Laura Schechter**, “Detecting Problems in Survey Data using Benford’s Law,” *Forthcoming in Journal of Human Resources*, November 2007.
- Leemis, Lawrence M., Bruce W. Schmeiser, and Diane L. Evans**, “Survival Distributions Satisfying Benford’s Law,” *The American Statistician*, 2000, *54* (4), 236–241.
- Marchi, S. and J. T. Hamilton**, “Assessing the Accuracy of Self-Reported Data: an Evaluation of the Toxics Release Inventory,” *Journal of Risk and Uncertainty*, 2006, *32* (1), 57–76.
- Nigrini, M.**, “A taxpayer compliance application of Benford’s law,” *Journal of the American Taxation Association*, 1996, *18* (1), 72–91.
- Noether, G. E.**, “Note on the Kolmogorov statistic in the discrete case,” *Metrika*, 1963, *7*, 115–116.
- Raimi, R. A.**, “The First Digit Problem,” *The American Mathematical Monthly*, 1976, *83* (7), 521–538.
- Rodriguez, R. J.**, “First Significant Digit Patterns from Mixtures of Uniform Distributions.,” *The*

- American Statistician*, 2004, 58 (1), 64–72.
- Scott, P. D. and M. Fasli**, “Benford’s Law: An Empirical Investigation and a Novel Explanation,” Technical Report, CSM Technical Report 349, Department of Computer Science, University Essex 2001.
- Stephens, M. A.**, “Use of the Kolmogorov-Smirnov, Cramer-Von Mises and Related Statistics Without Extensive Tables,” *Journal of the Royal Statistical Society. Series B (Methodological)*, 1970, 32 (1), 115–122.
- van der Vaart, A. W.**, *Asymptotic Statistics*, Cambridge University Press, 2000.