An Empirical Non-Parametric Likelihood Family of Data-Based Benford-Like Distributions*

Marian Grendar George Judge Laura Schechter

January 4, 2007

Abstract

A mathematical expression known as Benford's law provides an example of an unexpected relationship among randomly selected sequences of first significant digits (FSD). Newcomb (1881), and later Benford (1938), conjectured that FSD's would exhibit a weakly monotonic decreasing distribution and proposed a frequency proportional to the logarithmic rule. Unfortunately, the Benford FSD function does not hold for a wide range of scale-invariant multiplicative data. To confront this problem we use information-theoretic methods to develop a data-based family of alternative Benford-like exponential distributions that provide null hypotheses for testing purposes. Two data sets are used to illustrate the performance of generalized Benford-like distributions.

*Marian Grendar is an assistant professor, Dept. of Mathematics, FPV UMB, Banska Bystrica; Inst. of Mathematics and CS of Slovak Academy of Sciences, Banska Bystrica; Inst. of Measurement Sciences SAS, Bratislava, Slovakia, e-mail: `marian.grendar@savba.sk`. George Judge is a professor in the Graduate School, 207 Giannini Hall, UC Berkeley, Berkeley CA 94720, e-mail: `judge@are.berkeley.edu`. Laura Schechter is an assistant professor, Agricultural and Applied Economics, UW Madison, Madison, WI 53706, e-mail: `lschechter@wisc.edu`. The order of the authors' names has only alphabetical significance. Laura Schechter is the corresponding author. Thanks to Wendy Cho and Maximilian Auffhammer for help with the computer code and to Joanne Lee, Lawrence Leemis, Douglas Miller, Steven J. Miller, and John Morrow for helpful comments. The first author received funding from VEGA grant 1/3016/06 and Australian Research Council grant DP0210999 while the third author received funding from USDA Hatch grant 142-1038.

Keywords: Benford's law, first significant digit phenomenon, relative frequencies, information-theoretic method, empirical likelihood, minimum-divergence distance measure.

AMS Classification: Primary 62E20.

JEL classification: C10, C24.

1 Introduction

Theoretical and applied-data outcomes involving unanticipated results have been important in the search for quantitative scientific knowledge. In this surprise-knowledge search context, a mathematical expression known as Benford's law provides a useful example of an unexpected relationship among randomly selected sequences of positive real numbers - first significant digits (FSD, or the first non-zero digit found when reading a number from left to right). This FSD phenomenon was first noticed by Newcomb (1881) who observed that the pages in logarithmic tables for numbers starting with 1 were significantly more worn than those starting with 9. Based on this discovery, he conjectured that FSD distributions over a variety of data sets would not be uniform and would exhibit a weakly monotonic decreasing distribution. From this conjecture he created a formula reflecting the distribution of FSD's. Fifty years later, Benford (1938) noted the same FSD characteristics in certain data sets and proposed that the digits, $d = 1, 2, ..., 9$, appear as FSD's with frequency proportional to the logarithmic rule

$$P(d = 1, 2, ..., 9) = \log_{10}(1 + d^{-1}) \tag{1.1}$$

that results in a uniform distribution in logarithmic space. Benford gave the resulting distribution (0.301, 0.176, 0.125, 0.097, 0.079, 0.067, 0.058, 0.051, 0.046) a theoretical basis by showing it could evolve from a mixture of uniform distributions.

Many others have attempted to rationalize Benford's logarithmic formula and provide a

stronger theoretical explanation for the empirically discovered FSD phenomenon. Overviews of the history and a sampling of the empirical and theoretical results include Raimi (1976), Diaconis (1977), Schatte (1988), Hill (1995), Scott & Fasli (2001), Rodriguez (2004), Hill & Schürger (2005), Berger & Hill (2006), and Miller & Nigrini (2006). As Rodriguez (2004) notes, Raimi (1976) contends that Benford's mixture scheme is rather arbitrary and suggests a wide variety of FSD distributions from mixtures of uniform distributions.[1] However, Benford's distribution continues to be the null hypothesis of choice for those tracking questions of human influence on or tampering with data. Papers using Benford's law to check the validity of purportedly scientific data in the social and physical sciences include Varian (1972), Nigrini (1996, 1999), de Marchi & Hamilton (2006), Nigrini & Miller (2006), and Judge & Schechter (2006).

Benford's law postulates that lower digits are more likely to appear as FSD's than higher ones and specifies a particular FSD distribution (1.1) that captures this phenomenon. Although Benford's logarithmic FSD function may be consistent with some data sets, it seems questionable that it holds for all sets of numerical data. As Scott & Fasli (2001) note, only about half of the data sets in Benford's original paper provide reasonably close matches. Leemis et al. (2000) and others have noted an elementary link between the underlying basic data and FSD distributions. Consequently, it seems reasonable that, in general, the scale-invariant multiplicative nature of the underlying distribution of the data induces the Benford-like FSD distribution (see Pietronero et al. 2001). Viewed in this context, the FSD distribution provides just another way to characterize the information in the underlying data distribution. Thus, in contrast to Benford's parametric distribution, using a family of FSD data-based distributions that incorporate the underlying characteristics of a data set may

[1] Articles as early as Hamming (1970) and as recent as Miller & Nigrini (2006) have noted that the product of two distributions is usually closer to Benford's law than either of the original distributions. As the number of terms increases, the resulting observation converges to Benford. The latter article reviews some of the literature related to this issue.

be a superior way to learn about and capture the data's unknown FSD distribution.

Within this context, the purpose of this article is to suggest, using information theoretic methods, a family of data-based Benford-like FSD distributions that are based on a first moment of the FSD data. The resulting family of distributions, based on a minimum-divergence distance measure and FSD moment conditions, exhibits weakly monotonically decreasing FSD probabilities and yields generalized Benford-like alternative exponential distributions as null hypotheses for use in confronting actual data probabilities. The same functional dependency between FSD's which we express in the form of an exponential or power law defines different functions depending on the first-moment domain of the observed data sample.

The organization of the paper is as follows. In Section 2 the identification of an FSD distribution is reformulated as an ill-posed inverse problem and information-theoretic solutions are suggested. In Section 3 empirical likelihood methods (Owen 2001) are demonstrated and investigated as a basis for developing data-adaptive FSD distributions. In Section 4, different data sets are used to illustrate the reach of the empirical likelihood information-theoretic method in recovering data-specific FSD distributions and the use of the data-based FSD distributions for checking tampering, behavioral, and human influence characteristics observed in data outcomes. In Section 5, methodological and applied implications are discussed.

2 Problem Reformulation and Solution

In identifying a unique FSD distribution to associate with sequences of positive real numbers, assume that on trial $i = 1, 2, \ldots, n$, one of nine digits $d_1, d_2, \ldots, d_9$ is observed with p_j as the probability that the jth digit is observed. Suppose after n trials we are given first-moment

information in the form of the average value of the FSD:

$$\sum_{j=1}^{9} d_j p_j = \bar{d}. \tag{2.1}$$

Given this first-moment information and the inverse problem of identifying an FSD distribution, we seek the best predictions of the unknown probabilities $p_1, p_2, \ldots, p_9$. It is readily apparent that there is one data point and nine unknowns so, from an information-recovery standpoint, the resulting inverse problem is ill-posed. Consequently, there exist an infinite number of possible discrete probability distributions with $\bar{d} \in [1, 9]$. For illustrative purposes, it might be useful to consider this problem within the context of a nine-sided die. The sample of realized values - sequences of positive real numbers - are then the result of rolling the die n times.

Based only on the information $\sum_{j=1}^{9} d_j p_j = \bar{d}$, $\sum_{j=1}^{9} p_j = 1$, and $0 \leq p_j \leq 1$, the problem cannot be solved for a unique solution. Consequently, a function must be inferred from insufficient information when only a feasible set of solutions is specified. In such a situation it would seem useful to have an approach that allows the investigator to use sample-based information recovery methods without having to choose, as in Equation (1.1), a parametric family of probability densities on which to base the FSD function. In other words, we seek a way to reduce the infinite dimensional nonparametric problem to a finite dimensional one.

2.1 An Information-Theoretic Approach

One way to solve this ill-posed inverse problem for the unknown p_j without making a large number of assumptions or introducing additional information is to formulate it as an extremum problem. This type of extremum problem is, in many ways, analogous to allocating probabilities in a contingency table where p_j and q_j are the observed and expected probabilities respectively of a given event. A solution is achieved by minimizing the divergence

between the two sets of probabilities, optimizing a goodness-of-fit (pseudo-distance measure) criterion subject to data-moment constraint(s). One possible set of divergence measures is the Cressie-Read (CR) power divergence family of statistics (Cressie & Read 1984, Read & Cressie 1988, Baggerly 1998):

$$I(\mathbf{p}, \mathbf{q}, \gamma) = \frac{1}{\gamma(1+\gamma)} \sum_{j=1}^{9} \left(p_j \left[\left(\frac{p_j}{q_j} \right)^{\gamma} - 1 \right] \right) , \tag{2.2}$$

where γ is an arbitrary unspecified parameter.

In the context of recovering the unknown FSD distribution, use of the CR criterion (2.2) suggests we seek, given $\mathbf{q}$, a solution to the following extremum problem:

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} \left[I(\mathbf{p}, \mathbf{q}, \gamma) \,\middle|\, \sum_{j=1}^{9} p_j d_j = \bar{d}, \sum_{j=1}^{9} p_j = 1, p_j \geq 0 \right] . \tag{2.3}$$

In the limit, as γ varies, a family of distance measures evolves. The variants $\gamma = -1$ and $\gamma = 0$ of $I(\mathbf{p}, \mathbf{q}, \gamma)$ have received explicit attention in the literature (see Mittelhammer et al. (2000)). Assuming for expository purposes that the reference distribution is discrete uniform, i.e. for all j, $q_j = 1/9$, then $I(\mathbf{p}, \mathbf{q}, \gamma)$ converges to an estimation criterion equivalent to Owen's (2001) empirical likelihood (EL) criterion $\sum_{j=1}^{9} \ln(p_j)$, when $\gamma \to -1$. The EL criterion assigns discrete mass across the nine possible FSD outcomes. In the sense of objective function analogies, the Owen EL is closest to the classical maximum-likelihood approach and in fact results in a maximum non-parametric likelihood alternative. Another prominent case for the CR statistic corresponds to letting $\gamma \to 0$ and leads to the criterion $-\sum_{j=1}^{9} p_j \ln(p_j)$, which is the maximum entropy (ME) function (Shannon 1948, Jaynes 1957*a*,*b*). Inserting the $\gamma = 0$ criterion in (2.3) leads to a maximum entropy formulation for the problem. Solutions for these distance measures cannot be written in a closed form and require a computer optimization algorithm.

3 Empirical Likelihood (EL) Formulation and Application

Given the two information-theoretic variants of the CR $I(\mathbf{p}, \mathbf{q}, \gamma)$ discrepancy-distance measures prominent in the literature, we demonstrate, in the case of the CR-EL criterion, $\gamma \to -1$, a uniform reference distribution $\mathbf{q}$ (for all j, $q_j = 1/9$), and first-moment information, a basis for recovering discrete FSD probability distributions such that the probabilities $\mathbf{p} > \mathbf{0}$ and $\sum_j p_j = 1$. Under this specification, when $\gamma \to -1$, the CR $I(\mathbf{p}, \mathbf{q}, \gamma)$ converges to an estimation criterion equivalent to Owen's (2001) empirical likelihood metric $9^{-1} \sum_{j=1}^{9} \ln(p_j)$. Our extremum problem likelihood function can then be formulated as

$$\max_{\mathbf{p}} \left[9^{-1} \sum_{j=1}^{9} \ln p_j \;\middle|\; \sum_{j=1}^{9} p_j d_j = \bar{d}, \sum_{j=1}^{9} p_j = 1 \right]. \tag{3.1}$$

The corresponding Lagrange function is

$$L(\mathbf{p}, \eta, \lambda) \equiv 9^{-1} \sum_{j=1}^{9} \ln p_j - \eta \left(\sum_{j=1}^{9} p_j - 1 \right) - \lambda \left(\sum_{j=1}^{9} p_j d_j - \bar{d} \right) \tag{3.2}$$

where $\mathbf{p} > \mathbf{0}$ is implicit in the structure of the problem. Solving the corresponding first order condition with respect to p_j leads to the solution

$$\hat{p_j}(\bar{d}, \hat{\lambda}) = 9^{-1} \left(1 + \hat{\lambda} \left(d_j - \bar{d} \right) \right)^{-1} \tag{3.3}$$

for the jth outcome where $\hat{\lambda}$ is such that $\hat{\mathbf{p}}_B(\overline{d}, \hat{\lambda})$ satisfies the mean constraint (2.1). This solution implies that, as the mean of the FSD varies over a range of actual data sets, an exponential family of distributions will result. In equation (4.2), $\hat{p}_j$ is a function of $\hat{\lambda}$, the Lagrange multiplier for constraint (2.1) and the information used as a basis for modifying the distribution of FSD probabilities. The CR-EL criterion, also specified as $\prod_{j=1}^{9} p_j$, provides

an empirical representation of the joint PDF of independent random variables. Maximizing $\prod_{j=1}^{9} p_j$, subject to the moment condition and the adding up restriction, the p_j are chosen to assign the maximum joint probability among all of the possible probability assignments.

3.1 Some Mean-Related EL Distributions

Given information about FSD means of data sets and the CR-EL formulation (3.1)-(4.2), some corresponding FSD distributions are presented in Appendix Table A-3 and illustrated in Figure 1. As expected, uniform $\hat{p}_j$ result when a uniform reference distribution and a FSD mean of 5 are used in (3.1). For mean FSD values less than 5, the resulting estimated FSD distribution is tilted toward the lower digits and reflects the monotonic decreasing FSD probabilities exhibited by the Benford distribution. For FSD means between 3 and 4, the correlation between Benford and the EL FSD proportions are high, approaching 1 as the FSD mean approaches the Benford mean of 3.44. For this FSD mean, the EL and FSD proportions are approximately equivalent. Because many empirical data sets have FSD means between 3 and 4, this explains why many seemingly unrelated data sets have been associated with Benford-like FSD distributions.[2] Note, however, in the rare event of an FSD mean greater than 5.0, the distribution is increasing (see Table A-3).

In this data-adaptive context, as a data set's FSD mean changes, alternative null hypotheses regarding the digit proportions are suggested. Thus, a basis is provided for realizing an exponential family of FSD distributions and relating it to a particular underlying data set. Consequently, data-based Benford-like alternative null hypotheses result and present an alternative basis for testing for human influence and/or errors of measurement in data sets.

[2]Another explanation would be that these data sets involve products of independent observations.

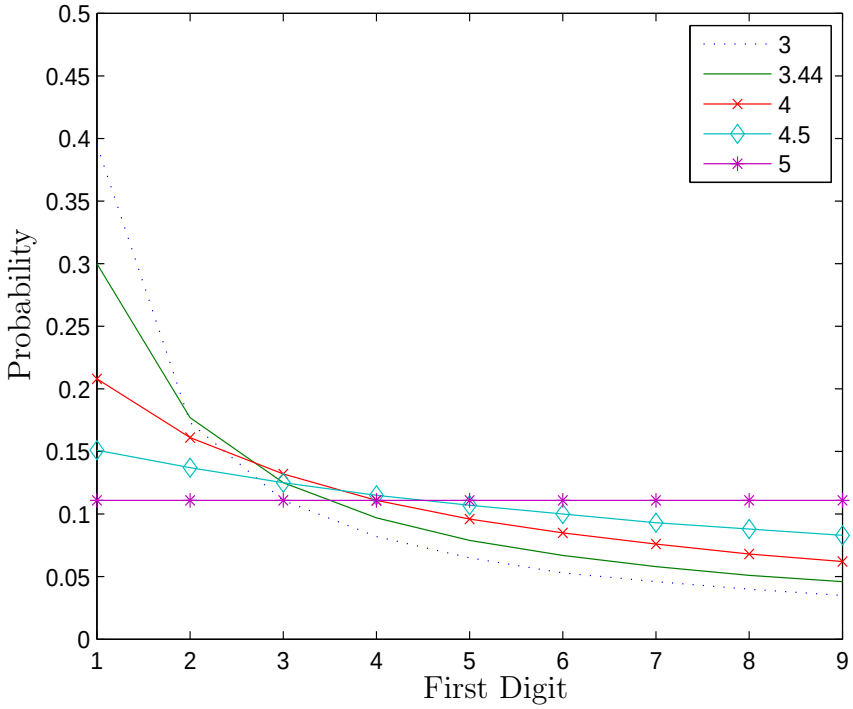

Figure 1: Empirical Likelihood (EL) Distribution (with uniform reference distribution)

4 Illustrations of EL Estimator's Performance

4.1 The Rodriguez Data

As one basis to illustrate the performance of the EL estimator in recovering FSD data-based distributions, we make use of data analyzed by Ley (1996) and Rodriguez (2004). These data on sales, total assets, net income, and stock prices are from the Disclosure Global Researcher SEC database. Ley (1996) originally analyzed DJ Returns 1, 2, and 3 which consist of the daily rates of return of the Dow Jones Industrial Average when their absolute values are below .1, greater than or equal to .1 but less than 1, and greater than or equal to 1 respectively. Rodriguez (2004) analyzed these variables as well as the daily closing values of the Dow Jones Industrial Average (DJ Value) which he took from the internet, recording all index values lower than 10,000 from January 2, 1930 to December 29, 2000. We analyze

this data since it has already been analyzed in the context of Benford's law in two important papers, is related to an interesting data set for economists, and involves FSD first moments that vary over a wide range, thus illustrating the reach of the EL criterion.

In Appendix A the frequencies for the actual data are presented in Table A-1 and the corresponding EL FSD frequencies are presented in Table A-2. Note the FSD means for these data range from 1.389 to 5.034 and three of the data sets have FSD means virtually equivalent to the Benford mean of 3.44. Consequently, these three data sets have an almost perfect correlation with the Benford and resulting EL distributions. Correlations and goodness-of-fit tests between the EL, Benford, and actual distributions are presented in Table 1.

The χ^2 goodness-of-fit test is the test most commonly used when comparing actual data with Benford's law. The χ^2 test has high power for large samples so even quite small deviations from Benford's law will be statistically significant. Giles (2006) has suggested using Kuiper's modified Kolmogorov-Smirnov goodness-of-fit test (V_N) instead. This test is less sensitive to sample size and also recognizes circularity of data. Critical values for a modified Kuiper test (V_N^*) have been given by Stephens (1970). Both the original and modified Kuiper tests were designed for use with continuous distributions, so the critical values given by Stephens (1970) are not accurate in the case of the discrete Benford distributions. Monte Carlo exercises suggest a 5% critical value of 1.34 for the Benford distribution (rather than 1.75 in the continuous case). Morrow (2006) shows that there are general properties under which we should expect both Benford's law and scale invariance to hold, however he also shows that the suitability of tests found in the literature is dependent on underlying distributional assumptions.

The Rodriguez data sets involving sales, total assets, and net income have FSD means consistent with the Benford mean. Therefore, in terms of goodness-of-fit with the actual data, both Benford and EL perform well. On the other hand, the DJ1 data has a mean of 5.03 and thus exhibits a non-decreasing monotonic FSD property more compatible with the

Table 1: Correlations (r), χ^2 Tests, and Kuiper V_N^* Tests, between the Empirical Distribution from the Rodriguez (2004) Data Sets and both the EL Estimated Distribution (with a uniform reference distribution) and Benford's Distribution

			EL-Emp			Ben-Emp		
Variable	Obs	Mean	r	χ^2	V_N^*	r	χ^2	V_N^*
DJ Return 1	6162	5.03	0.21	21.4	1.44	-0.48	2605.8	21.33
DJ Return 2	22598	3.61	0.97	304.9	8.26	0.96	612.1	11.21
DJ Return 3	5044	1.39	0.99	1216.2	8.42	0.94	5676.3	33.62
DJ Value	18392	4.17	0.65	5799.3	34.25	0.81	6320.8	25.79
Sales	11566	3.45	1.00	74.3	3.73	1.00	9.3	1.08
Total Assets	11565	3.44	1.00	5.2	0.55	1.00	5.1	0.58
Net Income	11566	3.44	1.00	12.3	0.76	1.00	12.2	0.79
Stock Prices	8584	3.26	0.99	67.2	3.69	1.00	72.3	3.53

The 10%, 5%, and 1% critical values for χ^2 with 8 degrees of freedom are 13.36, 15.51, and 20.09, and for V_N^* they are approximately 1.21, 1.34, and 1.61.

EL estimated distribution. This is likely the reason that the correlation between the empirical distribution and Benford's law is negative, while the correlation between the empirical distribution and the estimated EL distribution is positive. The DJ3 data set has an FSD mean of 1.39 and thus a highly tilted empirical distribution much different from Benford but consistent with the estimated EL distribution. The DJ Value data set has a mean of 4.17 and is close to the uniform and EL distributions. Generally, although Benford's distribution tends to be more highly correlated with the empirical data, the EL distribution yields superior goodness-of-fit for data sets with FSD means further from 3.44. For the Rodriguez data sets with FSD means close to 3.44, Benford and EL both appear to provide decent goodness-of-fit. Thus, the FSD sample mean appears to be a good predictor of goodness-of-fit with Benford and the resulting EL FSD distribution.

4.2 The Paraguay Data

We also use survey data from households in rural Paraguay to examine whether the information-theoretic EL methods can be used with survey data to assess its agreement or disagreement

with Benford's law. Correlation and goodness-of-fit tests are used to check the agreement, or disagreement, among the EL, Benford, and empirical distributions and the results are presented in Table 2. The self-reported survey data from rural Paraguayans exhibits a large number of outcomes with an FSD of 5, perhaps due to guesses by the respondents. A similar phenomenon is found by de Marchi & Hamilton (2006), who used Benford's law to test for tampering in self-reported toxic emissions by chemical plants.

In general, both Benford and EL do a good job of tracking the observed proportions. Again, the empirical data is more highly correlated with Benford than it is with the EL estimated distribution. The three variables for which the EL FSD distribution appears superior as seen in Table 2, with better goodness-of-fit according to both the χ^2 and V_N^* tests, are income, land owned, and the performance of the third enumerator in 2002. In previous work, we considered the fact that data on church donations do not conform with Benford's law to be suggestive evidence that people may not be reporting their donations correctly (Judge & Schechter 2006). This variable continues to perform poorly under the data-based EL. Again, departures of data sets' FSD means from 3.44 appear to be good predictors of relative goodness-of-fit with the Benford and EL FSD distributions.

4.3 Estimator Performance under a Non-uniform Reference Distribution

Thus far we have analyzed the CR distance measures using the assumption of a uniform reference distribution. We have noted the general monotonic decreasing nature of FSD distributions. We have also shown that the Benford FSD distribution offers good performance over a large number of data sets, as well as good performance relative to conventional EL. This suggests that there are data sets that induce the empirical Benford distribution. Consequently, it would seem that the Benford distribution is a natural choice as a reference

Table 2: Correlations (r), χ^2 Tests, and Kuiper V_N^* Tests, between the Empirical Distribution from the Paraguay Data Set and both the EL Estimated Distribution (with a uniform reference distribution) and Benford's Distribution

			EL-Emp			Ben-Emp		
Variable	Obs	Mean	r	χ^2	V_N^*	r	χ^2	V_N^*
Income	222	3.62	0.95	9.05	0.90	0.95	10.55	1.23
All Products	1632	3.16	0.96	115.25	3.17	0.97	101.34	2.69
Land Owned	223	3.61	0.98	7.19	0.64	0.98	8.49	0.72
Donations	197	2.94	0.88	49.47	2.28	0.92	38.93	1.64
Enu1 in 2002	177	3.45	0.98	4.37	0.74	0.98	4.23	0.72
Enu2 in 2002	184	3.37	0.96	10.75	0.72	0.96	10.38	0.64
Enu3 in 2002	198	3.24	0.98	4.76	0.43	0.98	5.96	0.83
Enu1 in 1999	85	2.93	0.83	26.89	1.68	0.87	21.95	1.20
Enu2 in 1999	94	2.79	0.81	30.83	1.88	0.96	7.86	1.20

The 10%, 5%, and 1% critical values for χ^2 with 8 degrees of freedom are 13.36, 15.51, and 20.09, and for V_N^* they are approximately 1.21, 1.34, and 1.61.

distribution ($\mathbf{q}_B$) in the CR-EL context. We pursue this idea in the next subsections.

4.3.1 EL Formulation with Benford Reference

To acknowledge the decreasing monotonic nature of FSD's, instead of a uniform distribution we now make use of the Benford distribution, $\mathbf{q}_B$, as the reference distribution in (2.2). Thus, in the Cressie-Read formulation (2.2), $\gamma = -1$ and Benford probabilities $\mathbf{q}_B$ replace the uniform reference distribution of Section 3. This leads to the BEL, or Benford Empirical Likelihood, criterion

$$\lim_{\gamma \to -1} I(\mathbf{p}, \mathbf{q}_B, \gamma) = \sum_{j=1}^{9} q_{jB} \ln(p_j / q_{jB}) = \sum_{j=1}^{9} q_{jB} \ln(p_j) - \sum_{j=1}^{9} q_{jB} \ln(q_{jB}) \tag{4.1}$$

where $\sum_{j=1}^{9} q_{jB} \ln q_{jB}$ is an added constant. Using this revised criterion and the data constraint (2.1), the adding-up condition, and selected FSD means over the range 2.0-5.5, results in

$$\hat{p_{jB}}(\bar{d}, \hat{\lambda}) = q_{jB} \left(1 + \hat{\lambda} \left(d_j - \bar{d}\right)\right)^{-1} \text{ for } j = 1, ..., 9 \tag{4.2}$$

where $\hat{\lambda}$ is such that $\hat{\mathbf{p}}_B(\overline{d}, \hat{\lambda})$ satisfies the mean constraint (2.1).

The BEL recovered FSD distributions, $\hat{\mathbf{p}}_B$, for the range of mean values 2.0–5.5 are presented in Table A-4. To see the impact of using a Benford reference distribution and a BEL criterion function, compare the estimates in Table A-4 with the conventional EL estimates in Table A-3. For clarity, Table A-5 shows the difference between the El and the BEL estimates. One interesting fact is that the Benford distribution and the BEL distribution are absolutely identical when the FSD mean is 3.44. In this case the Benford reference distribution is the minimum distance solution since it satisfies the constraints. With FSD means above 3.44, the BEL estimates tend to put higher probabilities on both higher and lower digits than do the EL estimates, which put higher probability on digits in the middle. For FSD means below 3.44, the El puts higher probability on a first digit of 1, whereas BEL puts higher probability on low digits greater than 1. Note also that the correlation between the EL estimates and Benford's distribution for sample FSD means between 3 and 4 is quite high.

Comparisons of the BEL distributions to the Rodriguez and Paraguay data sets suggest that the BEL distributions almost always lead to a closer fit with the respective empirical distributions than does Benford's distribution.[3] The list of variables for which we could not reject that the data were naturally occurring hardly changes when using the BEL rather than the Benford and conventional EL distributions. On the other hand, the use of BEL does not allow us to fail to reject that many more of the data sets are naturally occurring and free of human influence.

One of the referees raised the possibility that, in the event that the value of $\bar{d}$ for DJ in 2004 were available, one could use the corresponding EL-estimated FSD distribution as the reference distribution for DJ in 2005. This data-based reference distribution may, in many situations, be superior to both the fixed uniform and Benford distributions.

[3]These results are omitted to save space but are available from the authors upon request.

5 Summary and Implications

Benford's law and the corresponding logarithmic FSD distribution appear to capture the weakly monotonic nature of a range of data sets. Recognizing that the Benford FSD distribution does not hold in general for scale invariant distributions, we have suggested a family of data-based Benford-like distributions that are based on information-theoretic methods and a first moment of an FSD data distribution. This resulting family of distributions exhibits weakly monotonic Benford-like FSD probabilities and yields exponential distributions that may serve as null hypotheses when confronting empirical FSD proportions.

If a "natural" FSD distribution is to be used as a reference distribution to evaluate the impact of human influence or tampering on real data sets, it seems important that the reference distribution incorporate the characteristics defining that data set. Based on an empirical likelihood distance measure, a range of information-theoretic FSD distributions having different FSD means were analyzed and compared to Benford's FSD distribution under the assumptions of both uniform and Benford reference distributions. Two data sets were used to illustrate the reach of these data-based FSD methods. In both cases the information-theoretic FSD distributions performed well in assessing agreement or disagreement with Benford's law. Why some sequences of positive real numbers naturally exhibit the scale invariance multiplicative property is a question we, and many others, continue to ponder.

A Appendix: Extra Tables

Table A-1: Empirical Rodriguez Data Sets (Table 3 in Rodriguez (2004))

Data Set	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
DJ Return 1 (6,162)	0.101	0.110	0.121	0.103	0.116	0.119	0.116	0.107	0.109
DJ Return 2 (22,598)	0.233	0.189	0.143	0.117	0.093	0.075	0.060	0.050	0.041
DJ Return 3 (5,044)	0.773	0.145	0.040	0.021	0.011	0.005	0.002	0.002	0.001
DJ Value (18,392)	0.327	0.140	0.059	0.053	0.042	0.061	0.064	0.140	0.115
Sales (11,566)	0.306	0.174	0.121	0.097	0.077	0.065	0.058	0.052	0.051
Total Assets (11,565)	0.301	0.177	0.121	0.099	0.083	0.065	0.057	0.051	0.047
Net Income (11,566)	0.301	0.176	0.123	0.099	0.085	0.061	0.059	0.051	0.046
Stock Prices (8,584)	0.306	0.197	0.137	0.093	0.076	0.059	0.051	0.041	0.040

Table A-2: Estimated Empirical Likelihood (EL) Distributions (with uniform reference distribution) for the Rodriguez (2004) Data

Data Set	# of Obs	FSD Mean	$\hat{p_1}$	$\hat{p_2}$	$\hat{p_3}$	$\hat{p_4}$	$\hat{p_5}$	$\hat{p_6}$	$\hat{p_7}$	$\hat{p_8}$	$\hat{p_9}$
DJ Return 1	6,162	5.0336	0.109	0.109	0.110	0.111	0.111	0.112	0.112	0.113	0.113
DJ Return 2	22,598	3.6055	0.270	0.174	0.129	0.102	0.085	0.072	0.063	0.056	0.050
DJ Return 3	5,044	1.3898	0.869	0.047	0.024	0.016	0.012	0.010	0.008	0.007	0.006
DJ Value	18,392	4.1734	0.186	0.154	0.131	0.114	0.101	0.090	0.082	0.075	0.069
Sales	11,566	3.4523	0.298	0.177	0.126	0.097	0.080	0.067	0.058	0.051	0.046
Total Assets	11,565	3.4416	0.300	0.177	0.125	0.097	0.079	0.067	0.058	0.051	0.046
Net Income	11,566	3.4409	0.300	0.177	0.125	0.097	0.079	0.067	0.058	0.051	0.046
Stock Prices	8,584	3.2632	0.336	0.177	0.120	0.091	0.073	0.061	0.053	0.046	0.041

Table A-3: Estimated Empirical Likelihood (EL) Distributions (with uniform reference distribution) for the FSD Problem and their Correlation (r) with Benford's Distribution

FSD Mean	$\hat{p_1}$	$\hat{p_2}$	$\hat{p_3}$	$\hat{p_4}$	$\hat{p_5}$	$\hat{p_6}$	$\hat{p_7}$	$\hat{p_8}$	$\hat{p_9}$	r
2.0	0.673	0.111	0.061	0.042	0.032	0.026	0.021	0.018	0.016	0.925
3.0	0.395	0.173	0.111	0.082	0.065	0.053	0.046	0.040	0.035	0.990
3.44	0.300	0.177	0.125	0.097	0.079	0.067	0.058	0.051	0.046	1.000
4.0	0.208	0.161	0.132	0.111	0.096	0.085	0.076	0.068	0.062	0.980
4.5	0.151	0.137	0.125	0.115	0.107	0.100	0.093	0.088	0.083	0.932
5.0	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.000
5.5	0.083	0.088	0.093	0.100	0.107	0.115	0.125	0.137	0.151	-0.782

Table A-4: Estimated Empirical Likelihood (BEL) Distributions (with a Benford FSD reference distribution) for the FSD Problem and their Correlation (r) with Benford's Distribution

FSD Mean	$\hat{p_1}$	$\hat{p_2}$	$\hat{p_3}$	$\hat{p_4}$	$\hat{p_5}$	$\hat{p_6}$	$\hat{p_7}$	$\hat{p_8}$	$\hat{p_9}$	r
2.0	0.598	0.176	0.083	0.049	0.032	0.022	0.017	0.013	0.010	0.967
3.0	0.363	0.193	0.125	0.089	0.068	0.053	0.043	0.036	0.030	1.000
3.44	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	1.000
4.0	0.243	0.152	0.116	0.097	0.086	0.080	0.076	0.075	0.076	1.000
4.5	0.205	0.132	0.104	0.091	0.085	0.084	0.087	0.097	0.116	0.912
5.0	0.173	0.114	0.091	0.082	0.079	0.082	0.092	0.114	0.173	0.457
5.5	0.147	0.097	0.079	0.072	0.071	0.076	0.089	0.122	0.247	0.013

Table A-5: Difference Between the Estimated Empirical Likelihood EL (with a uniform FSD reference distribution) and BEL (with a Benford FSD reference distribution) Distributions for the FSD Problem

FSD Mean	$\hat{p_1}$	$\hat{p_2}$	$\hat{p_3}$	$\hat{p_4}$	$\hat{p_5}$	$\hat{p_6}$	$\hat{p_7}$	$\hat{p_8}$	$\hat{p_9}$
2.0	0.075	-0.065	-0.022	-0.007	0.000	0.004	0.004	0.005	0.006
3.0	0.032	-0.020	-0.014	-0.007	-0.003	0.000	0.003	0.004	0.005
3.44	-0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4.0	-0.035	0.009	0.016	0.014	0.010	0.005	0.000	-0.007	-0.014
4.5	-0.054	0.005	0.021	0.024	0.022	0.016	0.006	-0.009	-0.033
5.0	-0.062	-0.003	0.020	0.029	0.032	0.029	0.019	-0.003	-0.062
5.5	-0.064	-0.009	0.014	0.028	0.036	0.039	0.036	0.015	-0.096

References

Baggerly, K. (1998), 'Empirical likelihood as a goodness of fit measure', *Biometrika* **85**(3), 535–547.

Benford, F. (1938), 'The law of anomalous numbers', *Proceedings of the American Philosophical Society* **78**(4), 551–572.

Berger, A. & Hill, T. P. (2006), 'Newton's method obeys Benford's law', *American Mathematical Monthly* Forthcoming.

Cressie, N. & Read, T. R. C. (1984), 'Multinomial goodness of fit tests', *Journal of the Royal Statistical Society, Series B* **46**, 440–464.

de Marchi, S. & Hamilton, J. T. (2006), 'Assessing the accuracy of self-reported data: An evaluation of the toxics release inventory', *Journal of Risk and Uncertainty* **32**, 57–76.

Diaconis, P. (1977), 'The distribution of leading digits and uniform distribution mod 1', *The Annals of Probability* **5**(1), 72–81.

Giles, D. E. (2006), 'Benford's law and naturally occurring prices in certain ebaY auctions', *Applied Economics Letters* Forthcoming.

Hamming, R. W. (1970), 'On the distribution of numbers', *Bell System Technical Journal* **49**, 1609–1625.

Hill, T. P. (1995), 'A statistical derivation of the significant-digit law', *Statistical Science* **10**(4), 354–363.

Hill, T. P. & Schürger, K. (2005), 'Regularity of digits and significant digits of random variables', *Journal of Stochastic Processes and Their Applications* **115**, 1723–1743.

Jaynes, E. T. (1957*a*), 'Information theory and statistical mechanics', *Physical Review* **106**(4), 620–630.

Jaynes, E. T. (1957*b*), 'Information theory and statistical mechanics II', *Physical Review* **108**(4), 171–190.

Judge, G. & Schechter, L. (2006), 'Detecting problems in survey data using Benford's law'. Unpublished Manuscript.

Leemis, L. M., Schmeiser, B. W. & Evans, D. L. (2000), 'Survival distributions satisfying Benford's law', *The American Statistician* **54**(4), 236–241.

Ley, E. (1996), 'On the peculiar distribution of the U.S. stock indexes' digits', *The American Statistician* **50**(4), 311–313.

Miller, S. J. & Nigrini, M. J. (2006), Order statistics and shifted almost Benford behavior. Unpublished Manuscript.

Mittelhammer, R., Judge, G. G. & Miller, D. J. (2000), *Econometric Foundations*, New York: Cambridge University Press.

Morrow, J. (2006), Benford's law and families of distributions. Unpublished Manuscript.

Newcomb, S. (1881), 'Note on the frequency of use of the different digits in natural numbers', *American Journal of Mathematics* **4**, 39–40.

Nigrini, M. J. (1996), 'A taxpayer compliance application of Benford's law', *Journal of the American Taxation Association* **18**(1), 72–91.

Nigrini, M. J. (1999), 'Adding value with digital analysis', *The Internal Auditor* **56**(1), 21–23.

Nigrini, M. J. & Miller, S. J. (2006), Benford's law applied to hydrology data: Results and relevance to other geophysical data. Unpublished Manuscript.

Owen, A. B. (2001), *Empirical Likelihood*, Florida: Chapman & Hall/CRC.

Pietronero, L., Tosatti, E., Tosatti, V. & Vespignani, A. (2001), 'Explaining the uneven distribution of numbers in nature: The laws of Benford and Zipf', *Physica A: Statistical Methods and its Applications* **293**(1-2), 297–304.

Raimi, R. (1976), 'The first digit problem', *American Mathematical Monthly* **83**, 521–538.

Read, T. R. C. & Cressie, N. A. C. (1988), *Goodness-of-Fit Statistics for Discrete Multivariate Data*, New York: Springer-Verlag.

Rodriguez, R. J. (2004), 'First significant digit patterns from mixtures of uniform distributions', *The American Statistician* **58**(1), 64–71.

Schatte, P. (1988), 'On mantissa distributions in computing and Benford's law', *Journal of Information Processing and Cybernetics* **24**(10), 443–455.

Scott, P. D. & Fasli, M. (2001), Benford's law: An empirical investigation and a novel explanation. Unpublished Manuscript.

Shannon, C. E. (1948), 'A mathematical theory of communication', *Bell System Technical Journal* **27**, 379–423.

Stephens, M. A. (1970), 'Use of the Kolmogorov-Smirnov, Cramer-Von Mises and related statistics without extensive tables', *Journal of the Royal Statistical Society, Series B* **32**(1), 115–122.

Varian, H. (1972), 'Benford's law', *The American Statistician* **26**, 65.