Title

Does china income FSDs follow Benford? A comparison between Chinese income first significant digit distribution with Benford distribution

Journal

China Economic Journal, 12(1)

ISSN

1753-8963

Authors

Fu, Qiuzi
Villas-Boas, Sofia B
Judge, George

Publication Date

2019-01-02

DOI

10.1080/17538963.2018.1477418

Peer reviewed

China Economic Journal

ISSN: 1753-8963 (Print) 1753-8971 (Online) Journal homepage: https://www.tandfonline.com/loi/rcej20

Does china income FSDs follow Benford? A comparison between Chinese income first significant digit distribution with Benford distribution

Qiuzi Fu, Sofia B. Villas-Boas & George Judge

To cite this article: Qiuzi Fu, Sofia B. Villas-Boas & George Judge (2019) Does china income FSDs follow Benford? A comparison between Chinese income first significant digit distribution with Benford distribution, China Economic Journal, 12:1, 68-76, DOI: 10.1080/17538963.2018.1477418

To link to this article: https://doi.org/10.1080/17538963.2018.1477418

Published online: 05 Jun 2018.

CHINA ECONOMIC JOURNAL
2019, VOL. 12, NO. 1, 68–76
https://doi.org/10.1080/17538963.2018.1477418

Routledge
Taylor & Francis Group

Does china income FSDs follow Benford? A comparison between Chinese income first significant digit distribution with Benford distribution

Qiuzi Fu[a], Sofia B. Villas-Boas[b] and George Judge[b]

[a]National School of Development, Peking University, Beijing, 100871, China; [b]Department of Agricultural and Resource Economics, University of California, CA, Berkeley, USA

ABSTRACT

Since Benford's law is an empirical phenomenon that occurs in a range of data sets, this raises the question as to whether or not the same thing might be true in terms of the Chinese income distribution data. We focus on the first significant digit (FSD) distribution of Chinese micro income data from the 2005 Inter-Census sample, which corresponds to 1% of Chinese population and other micro income data from the China family panel studies (CFPS) and Chinese General Social Survey (CGSS). We use information theoretic-entropy based methods to investigate the degree to which Benford's FSD law is consistent with the FSD of Chinese income data and our findings suggest consistency between the Chinese FSD income distribution and Benford's distribution. The close connection between the two distributions has implications for the quality of the sample of Chinese micro data.

ARTICLE HISTORY
Received 21 November 2017
Accepted 14 May 2018

KEYWORDS
First significant digits; Benford's law; information theoretic methods; empirical likelihood; minimum divergence measure; adaptive behavior; causal entropy maximization

JEL CLASSIFICATION
C1; C10; C24

1. Introduction

Simon Newcomb (1881), an astronomer and mathematician, in 1881 conjectured that in natural data sets, the first digits did not occur with equal frequency, but rather he suggested that the occurrence of numbers is such that all mantissa of their logarithms are equally probable. Fifty-seven years later, Benford (1938) empirically began to test Newcomb's hypothesis and demonstrated that a large number of seemingly unrelated sets of numbers provided a good fit to the FSD exponential distribution and gave it law status. Since then, others have published studies showing that 'Benford's Law' applies to a surprisingly large number of natural-behavioral data sets and has the nice properties of being scale and base invariant (see Varian (1972) and Miller (2015)). This FSD phenomenon was then named 'Benford's Law', after its popularizer rather than its discoverer.

It was another 57 years before Hill (1995), using a base-invariance argument, became the first to rigorously prove Benford's Law. Prior to Hill, others had only suggested possible explanations for the phenomenon. For instance, Benford suggested that the law held when data came from a mixture of uniform distributions that were more likely to

CONTACT Sofia B. Villas-Boas sberto@berkeley.edu ARE, University of California, Berkeley, CA 94720, USA

have relatively small upper bounds. More recently, power-law and information-theoretic methods have been proposed as being more intuitively appealing and general ways of determining similar FSD distributions (Grendar, Judge, and Schechter 2007). Pietronero et al. (2001) suggest that Benford's Law is a special case of Zipf's Law, which claims that all rankings of natural processes by size follow power laws. For example, word frequencies have been shown to have such a distribution where the most frequent word occurs approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word (Zipf 1949). In this fashion, probability of occurrence is inversely proportional to its rank. The argument that Zipf's Law is a generalization of Benford's Law is based on the scale-invariant nature of both laws' respective applications, but since Zipf's Law is simply an empirical observation of a family of distributions, the claim that Zipf's Law justifies Benford's Law is intriguing but not rigorous.

Moving in a rigorous direction, we have noted the role of scale invariance underlying data outcomes in Hill's 1995 proof of Benford's Law. Scale invariance occurs when multiplying either the underlying data distribution P(D) or its FSD counterpart P(d) by a constant s yields an identical outcome. Following Pietronero et al. (2001), we note that scale invariance leads to the functional relation

$$\mathrm{P(sD)} = \mathrm{PD}^{*}) = \mathrm{K(p)P(D)} \tag{1.1}$$

and that the general solution to (1.1) has the power law nature

$$\mathrm{P(D}^{*}) = \mathrm{P(D}^{*-\alpha}) = \mathrm{s}^{-\alpha}\mathrm{D}^{-\alpha} \tag{1.2}$$

For these types of distributions, we can compute the probability of the first digit by noting that we have the same (uniform) relative probability for the integers $d = 1, 2, \ldots, 9$, for each cycle. Following Pietronero et al. (2001), we can write for $P(d)$ that, for $\alpha \neq 1$,

$$P(D^{*}) = \int_{\alpha}^{\alpha+1} D^{-\alpha}\, dD = \frac{1}{1-\alpha}\left[(d+1)^{1-\alpha} - d^{1-\alpha}\right] \tag{1.3}$$

Then, for $d = 1$:

$$P(D^{*}) = \int_{d}^{d+1} D\, dD = \int_{d}^{d+1} d(logD) = log\left(\frac{d+1}{d}\right) \tag{1.4}$$

This expresses Benford's law as determined from the underlying data distribution. Consequently, in a power law context when $\alpha = 1$, we have a uniform FSD in logarithmic space. For values of $\alpha > 1$, the FSD distribution is more tilted than Benford. For values of $\alpha < 1$, the FSD distribution is tilted toward a uniform FSD distribution. Pietronero et al. (2001) calls this family of power laws a generalized Benford law.

Zipf's Law, as mentioned earlier, is an instance of a rank order statistic that is scale invariant and applicable to a large range of phenomena, including income distributions, city sizes, and linguistics (Pietronero et al. 2001; Raimi 1976; Zipf 1949). Of particular interest is the connection between Benford's Law and Zipf's law. Following Pietronero et al. (2001), in analyzing the rank-order properties of a set of numbers extracted from a general distribution, $P(N) \sim N^{\alpha}$, if a maximum number Nmax corresponds to the rank $k = 1$ and the rank Nk is given by all the numbers between Nk and $N_{\max}$, then

$$k = \int_{N(k)}^{N_{max}} P(N)\ dN \sim N(k)^{1-\alpha} \tag{1.5}$$

Inverting (1.5) gives us

$$N(k) \sim k^{\frac{1}{1-\alpha}}, \tag{1.6}$$

which highlights a link between the Benford and Zipf's Law.

Overviews of the history and theoretical explanations include Raimi (1976), Diaconis (1977), Hill (1995), Berger and Hill (2006), Miller and Nigrini (2009), Judge and Schechter (2009). Even when FSD data sets deviate from the Benford pattern, the lower digits are favored and decline monotonically. Furthermore, in the physical sciences area, Shao and Ma (2010a and 2010b) demonstrate empirically that in physical statistics, the Boltmann–Gibbs and Fermi–Derac distributions with respect to the temperature of the system, fluctuate around the Benford distribution and that the Bose-Einstein distribution exactly conforms to it.

Since Benford's law is an empirical phenomenon that occurs in a range of data sets, this raises the question as to whether or not the same thing might be true in terms of the Chinese income distribution data. To pursue this question, we focus on the first significant digit (FSD) distribution of Chinese micro income data from the 2005 Inter-Census sample, which corresponds to 1% of Chinese population. We use information theoretic-entropy based methods to investigate the degree to which Benford's FSD law is consistent with the FSD of Chinese income data.

The rest of the paper proceeds as follows. In section 2, we introduce the information theoretic methods used to recover the FSD distributions from the micro income data. In section 3, we present in graph form the relationship between Benford and the data-based FSD income distributions. In section 4, we note the agreement between Benford and the empirical income FSD distributions and speculate on the implications of the results.

2. The conceptual framework

Pre analysis knowledge suggests that the FSD distribution of a sequence of positive real numbers from scale-independent multiplicative data should vary with the phenomena in question. In this context entropy based information theoretic methods offer a natural way to establish a data-based link that captures the varying monotonically decreasing nature of the FSD.

To use information theoretic methods to recover the FSD distribution from a sequence of positive real numbers, we assume for the discrete random variable d_i (for $i = 1,\ 2, \ldots,\ 9$), that at each trial, one of nine digits is observed with probability p_i. Suppose after n trials, we have first-moment information in the form of the average value of the FSD:

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

Based on *sample 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 nine digit FSD ill-posed inverse recovery problem cannot be solved for a unique solution. In such a situation it seems useful to have an approach that permits the investigator to use sample based information recovery methods without having to choose a parametric family of probability densities on which to base the FSD probability density function.

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-optimization problem. In this context a solution is achieved by minimizing the divergence between the two sets of probabilities and an optimizing goodness-of-fit criterion, subject to data-moment constraints. One attractive set of divergence measures is the Cressie-Read (CR) power divergence family of statistics (Cressie and Read (1984), Read and Cressie (1988), and Judge and Mittelhammer (2011, 2012)):

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

where γ is an arbitrary unspecified parameter. All well known entropy divergences belong to the class of CR functions. In the context of recovering the unknown sample information FSD distribution, we make use of the CR criterion (2.2) and seek a solution to the following extremum problem:

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

When $\gamma \to -1$ and $I(\boldsymbol{p}, \boldsymbol{q}, \gamma)$ converges to an estimation criterion equivalent to the empirical likelihood (EL) criterion $\sum_{j=1}^{N} \ln(p_j)$. As γ varies, power law like behavior is efficiently described and the resulting estimators that minimize power divergence exhibit qualitatively different sampling behavior. Over defined ranges of the divergence measures, the CR and entropy families are equivalent.

In terms of the information-theoretic variants of the CR $I(\boldsymbol{p}, \boldsymbol{q}, \gamma)$ we demonstrate for the Benford recovery problem the case of the Maximum Entropy Empirical Likelihood (MEEL)-CR $\gamma \to -1$, anda uniform reference distribution $\boldsymbol{q}$ ($q_j = 1/9, \forall j$). First moment information $\bar{d}$ is used as a basis for recovering discrete FSD probability distributions. As noted above, under the criterion CR $\gamma \to -1$, the CR $I(\boldsymbol{p}, \boldsymbol{q}, \gamma)$ converges to the empirical likelihood criterion metric $9^{-1} \sum_{j=1}^{9} (ln p_j)$ and the extremum likelihood function

$$\max_{\boldsymbol{p}} \left[9^{-1} \sum_{j=1}^{9} \ln p_j \mid \sum_{j=1}^{9} p_j d_j = \bar{d}, \sum_{j=1}^{9} p_j = 1 \right]. \tag{2.4}$$

The corresponding Lagrange function is

$$L(\boldsymbol{p},\eta,\lambda)\equiv 9^{-1}\sum_{j=1}^{9}\ln p_j-\eta\left(\sum_{j=1}^{9}p_j-1\right)-\lambda\left(\sum_{j=1}^{9}p_jd_j-\bar{d}\right) \tag{2.5}$$

with the solution

$$\widehat{p_j}(\bar{d},\lambda)=\left[9^{-1}\left(1+\hat{\lambda}\left(d_j-\bar{d}\right)\right)\right]^{-1}, \tag{2.6}$$

for the jth FSD outcome. As the mean of the significant first digits varies a family of probability density functions-distributions result In Equation (2.6), $\hat{p}_j$ is a function of $\hat{\lambda}$, the Lagrange multiplier for constraint (2.3). This information may be used as a basis for modifying the distribution of FSD probabilities.

For mean FSD values less than 5, the resulting estimated FSD distribution reflect the monotonic decreasing FSD probabilities exhibited by the Benford distribution. As the FSD mean approaches the Benford mean 3.44, the CR-EL and FSD distributions are approximately equal. If we use the$_9$CR in the limit $\gamma\rightarrow -1$ criterion and a Benford reference distribution $I(\boldsymbol{p},\boldsymbol{q_B},\gamma)=\sum_{j=1}^{9}(\ln p_j/q_{jB})$, then with the first moment condtion of 3.44, the Benford FSD distribution is exactly reproduced.

3. Benford and the Chinese micro income data

The data used in this section originate from the China's inter-census survey from 2005. The inter-census survey represents 1% of the population and is conducted every 10 years for years ending in 5. The sample contains over one million observations on personal characteristics and income data among the population of current residence. The survey covers all the 2861 counties of China and is representative of the 333 prefectures.

In Figure 1, we display for the Chinese 2005 micro data, the FSD probability density function and the Benford distribution. The fit of the Chinese FSDs and Benford FSDs is very good with a correlation of 0.964 and Chi-square of 0.031.

If in the above CR formulation, $\gamma\rightarrow -1$, *the*Benford reference distribution probabilities q_B replaces the uniform reference distribution, this leads to the BEL or Benford EL, criterion

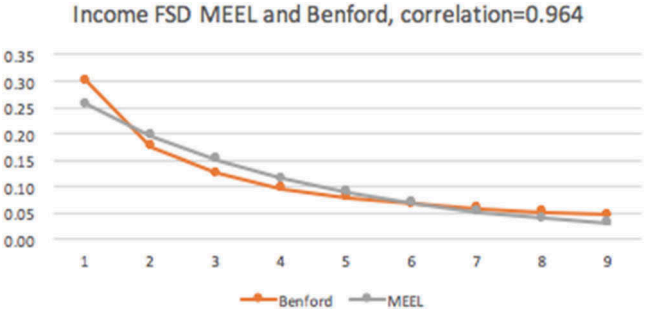

Figure 1. Benford and Chinese 2005 data FSD distribution.

$$\lim_{\gamma \to -1} I(p, q_B, \gamma) = \sum_{j=1}^{9} q_{jB} \ln(p_j / q_{jB}) = \sum_{j=1}^{9} q_{jB} \ln(p_j) - \sum_{j=1}^{9} q_{jB} \ln(q_{jB}). \tag{3.1}$$

where $\sum_{j=1}^{9} q_{jB} \ln(q_{jB})$ is an added constant(see Grendar, Judge, and Schechter 2007). Using the criterion (3.1), the data constraint $\sum_{j=1}^{9} d_j p_j = \bar{d}$, andthe probabilities adding-up condition,resultsin

$$\widehat{p_{jB}}\left(\bar{d}, \hat{\lambda}\right) = q_{jB}(1 + \hat{\lambda}(d_j - \bar{d}))^{-1}, \tag{3.2}$$

where $\hat{\lambda}$ is such that $\widehat{p_B}$ (d, $\hat{\lambda}$) satisfies the mean FSD constraint. When the Benford distribution is used as a reference distribution, the Benford distribution is exactly reproduced.

4. The China Family Panel Studies (CFPS) income data case

As an additional example of the association between the two distributions, we make use of micro data from the China Family Panel Studies (CFPS). The CFPS is a nationally representative, annual longitudinal survey of Chinese communities, families, and individuals launched in 2010 by the Institute of Social Science Survey (ISSS) of Peking University, China. The CFPS is designed to collect individual-, family-, and community-level longitudinal data in China. The studies focus on the economic, as well as the non-economic, wellbeing of the Chinese population, with a wealth of information such topics as economic activities, education outcomes, family dynamics and relationships, migration, and health. As a follow-up to the 2010, 2012, and 2014 data, the CFPS 2016 data also contains the five parts: community, family roster, family, adult and child. We use household income from family roster part. The data is monthly salary per capita for household and is in 'Yuan'. A graph of the two income FSDs and Benford is given in Figure 3.3. As in the comparisons in Figure 2, there is a degree of association between the two distributions as noted by a correlation of 0.97 and a Chi-square of 0.999.

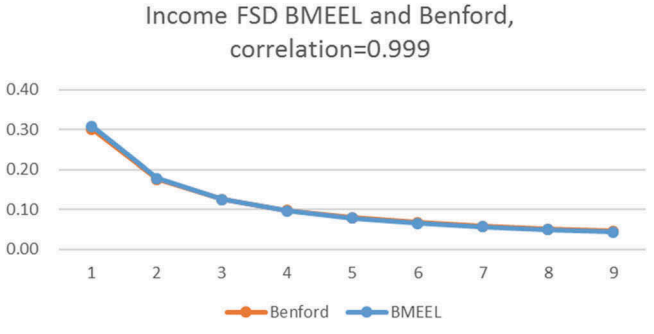

Figure 2. Benford and Chinese 2005 FSD distributions with Benford prior.

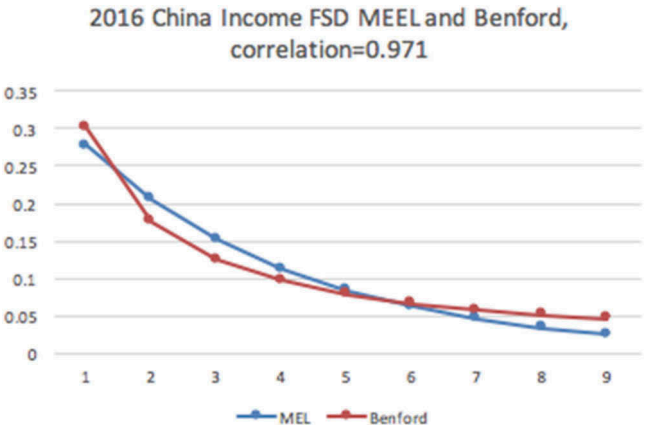

Figure 3. Benford and the Chinese 2016 CFPS FSD income data.

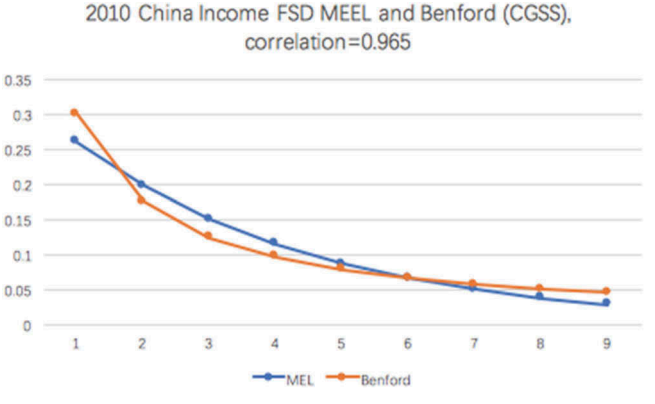

Figure 4. Benford and the CGSS 2010 FSD income data.

5. The Chinese General Social Survey (CGSS) income data case

The General Social Survey (CGSS) survey is a national representative continuous survey project launched in 2003. Conducted by Renmin University of China and Hong Kong University of Science and Technology, CGSS is one of the earliest and most comprehensive dataset in China. CGSS contains social and quality-of-life variables of individuals in both urban and rural China. The released data are from 2003 to 2015, so the latest data that we can find are in 2014. The quality-of-life part of CGSS includes individual salary of the past year. So the Chinese transformed the data into monthly salary (unit: Yuan) and try to analyze the first digit distribution pattern. As shown in Figures 4 and 5, both 2010 and 2014 CGSS income data show correlated income FSD with Benford. The Chi-squares are both 0.999 as well.

6. Summary and conclusions

In this paper, we have used a sample of Chinese micro income data and an entropy based member of the Cressie–Read family to present evidence that the associated micro

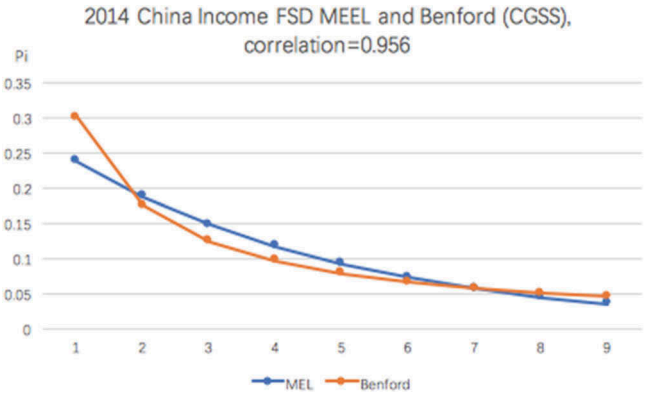

Figure 5. Benford and the CGSS 2014 FSD income data.

income FSD data and the Benford distribution are closely linked. This result along with a similar FSD results for Australian income data (see Villas-Boas, Fu, and Judge 2015) and physical systems indicate that the data from these behavioral worlds are closely linked to the Benford distribution. From a methodological view point we have demonstrated how entropy based information theoretic methods may be used in identifying and making distributional comparisons. Benford's law can also be used to test for errors in micro sampling data (see Shechter and Judge 2009; Cho and Gaines 2007; Cho and Judge 2015). The close association between Benford's law and the Chinese income FSDs bodes well for the quality of the 2005 micro data and the CFPS and CGSS data used in this study. Looking ahead, we plan to use the 2005 Chinese micro income data to develop income probability density functions-distributions and entropy inequalities measures for China and its 31 provinces.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Benford, F. 1938. "The Law of Anomalous Numbers." In *Proceedings of the American Philosophical Society* 78 (4): 551–572.

Berger, A., and T. P. Hill. 2006. "Newton's Method Obeys Benford's Law." *American Mathematical Monthly* 14: 588–601.

Cho, L. J. W., and G. Judge. 2015. "Stigler's Approach to Recovering the Distribution of First Significant Digits in Natural Data Sets." In *Of Benford's Law: Theory and Applications*, edited by S. Miller, 304–315, Princeton University Press.

Cho, W., and B. Gaines. 2007. "Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance." *The American Statistician* 61: 1–6.

Cressie, N., and T. Read. 1984. "Multinomial Goodness-Of-Fit Tests." *Journal of the Royal Statistical Society. Series B (Methodological)* 46: 440–464.

Diaconis, P. 1977. "The Distribution of Leading Digits and Uniform Distribution Mod 1." *The Annals of Probability* 5: :72–81. doi:10.1214/aop/1176995891.

Grendar, M., G. Judge, and L. Schechter. 2007. "An Empirical Non-Parametric Likelihood Family of Data-Based Benford-Like Distributions." *Physica A: Statistical Mechanics and Its Applications* 380: 429–438. doi:10.1016/j.physa.2007.02.062.

Hill, T. 1995. "A Statistical Derivation of the Significant Digit Law." *Statistical Science* 10 (4): 354–363. doi:10.1214/ss/1177009869.

Judge, G., and L. Schechter. 2009. "Detecting Problems in Survey Data Using Benford's Law." *Journal of Human Resources* 40 (1): 1–24.

Judge, G., and R. Mittelhammer. 2011. *An Information Theoretic Approach to Econometrics*. Cambridge: Cambridge University Press.

Judge, G., and R. Mittelhammer. 2012. "Implications of the Cressie-Read Family of Additive Divergences for Information Recovery." *Entropy* 14 (12): 2427–2438. doi:10.3390/e14122427.

Miller, S. ed. (2015). *Benford's Law: Theory and Applications*. Princeton, NJ: Princeton University Press.

Newcomb, S. 1881. "Note on the Frequency of Use of the Different Digits in Natural Numbers." *American Journal of Mathematics* 4 (1/4): 39–40. doi:10.2307/2369148.

Nigrini, M. U., and S. J. Miller. 2009. "Data Diagnostics Using Secondorder Tests of Benford's Law." *Auditing: a Journal of Practice & Theory* 28 (2): 305–324.

Pietronero, L., E. Tosatti, V. Tosatti, and A. Vespignani. 2001. "Explaining the Uneven Distribution of Numbers in Nature: The Laws of Benford and Zipf." *Physica A: Statistical Mechanics and Its Applications* 293: 297–304. doi:10.1016/S0378-4371(00)00633-6.

Raimi, R. 1976. "The First Digit Problem." *The American Mathematical Monthly* 83: 521–538. doi:10.1080/00029890.1976.11994162.

Read, T.R. and N.A. Cressie, 1988, Goodness of Fit Statistics for Discrete Multivariate Data, New York: Springer Verlag.

Shao, L., and B.-Q. Ma. 2010a. "The Significant Digit Law in Statistical Physics." *Physica A: Statistical Mechanics and Its Applications* 389: 3109–3116. doi:10.1016/j.physa.2010.04.021.

Shao, L., and B.-Q. Ma. 2010b. "First Digit Law in Nonextensive Statistics." *Physical Review* E 82: 04111014.

Shechter, L., and G. Judge. 2009. "Detecting Problems in Survey Data Using Benford's Law." *Journal of Human Resources* 44: 1–26.

Varian, H. 1972. "Benford's Law." *The American Statistician* 26: 65–66.

Villas-Boas, S., Q. Fu, and G. Judge. 2015. "Is Benford's Law a Universal Behavioral Theory?" *Econometrics* 3: 698–708. doi:10.3390/econometrics3040698.

Zipf, G. K. 1949. *Human Behavior and The Principle of Least Effort*. Oxford: Addison-Wesley Press.