Available online at www.sciencedirect.com

ScienceDirect

Physica A 380 (2007) 429–438

PHYSICA A

www.elsevier.com/locate/physa

An empirical non-parametric likelihood family of data-based Benford-like distributions

Marian Grendar[a,b,c], George Judge[d], Laura Schechter[e,*]

[a]Department of Mathematics, FPV UMB, Banska Bystrica, Slovakia
[b]Institute of Mathematics and CS of Slovak Academy of Sciences, Banska Bystrica, Slovakia
[c]Institute of Measurement Sciences SAS, Bratislava, Slovakia
[d]Graduate School, 207 Giannini Hall, UC Berkeley, Berkeley, CA 94720, USA
[e]Agricultural and Applied Economics, UW Madison, Madison, WI 53706, USA

Received 1 November 2006; received in revised form 4 January 2007
Available online 2 March 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 (FSDs). Newcomb [Note on the frequency of use of the different digits in natural numbers, Am. J. Math. 4 (1881) 39–40], and later Benford [The law of anomalous numbers, Proc. Am. Philos. Soc. 78(4) (1938) 551–572], conjectured that FSDs 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.

Keywords: Benford's law; First significant digit phenomenon; Relative frequencies; Information-theoretic method; Empirical likelihood; Minimum-divergence distance measure

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 (FSDs, or the first non-zero digit found when reading a number from left to right). This FSD phenomenon was first noticed by Newcomb [1] 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

*Corresponding author. Tel.: +1 608 2629482; fax: +1 608 2624376.
E-mail addresses: marian.grendar@savba.sk (M. Grendar), judge@are.berkeley.edu (G. Judge), lschechter@wisc.edu (L. Schechter).

0378-4371/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2007.02.062

created a formula reflecting the distribution of FSDs. Fifty years later, Benford [2] noted the same FSD characteristics in certain data sets and proposed that the digits, $d = 1, 2, \ldots, 9$, appear as FSDs with frequency proportional to the logarithmic rule

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

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 [3], Diaconis [4], Schatte [5], Hill [6], Scott and Fasli [7], Rodriguez [8], Hill and Schürger [9], Berger and Hill [10], and Miller and Nigrini [11]. As Rodriguez [8] notes, Raimi [3] 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 [13], Nigrini [14,15], de Marchi and Hamilton [16], Nigrini and Miller [17], and Judge and Schechter [18].

Benford's law postulates that lower digits are more likely to appear as FSDs 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 and Fasli [7] note, only about half of the data sets in Benford's original paper provide reasonably close matches. Leemis et al. [19] 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 Ref. [20]). 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 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 FSDs 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 (EL) methods [21] 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 EL 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

[1] Articles as early as Hamming [12] and as recent as Miller and Nigrini [11] 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.