Journal of Informetrics 9 (2015) 349–358

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

On the uniform random upper bound family of first significant digit distributions

Werner Hürlimann*

Feldstrasse 145, CH-8004 Zürich, Switzerland

ARTICLE INFO

Article history:
Received 16 December 2014
Received in revised form 23 January 2015
Accepted 20 February 2015
Available online 13 March 2015

MSC:
62E15
62P05
62P10

Keywords:
Benford's law
Stigler's law
Uniform distribution
Simulation algorithm
Extended truncated Pareto
Erlang distribution

ABSTRACT

The first significant digit patterns arising from a mixture of uniform distributions with a random upper bound are revisited. A closed-form formula for its first significant digit distribution (FSD) is obtained. The one-parameter model of Rodriguez is recovered for an extended truncated Pareto mixing distribution. Considering additionally the truncated Erlang, gamma and Burr mixing distributions, and the generalized Benford law, for which another probabilistic derivation is offered, we study the fitting capabilities of the FSD's for various Benford like data sets from scientific research. Based on the results, we propose the general use of a fine structure index for Benford's law in case the data is well fitted by the truncated Erlang member of the uniform random upper bound family of FSD's.

1. Introduction

Motivated by the first significant digit analysis of some biological data sets Cáceres, García, Martínez Ortiz, and Dominguez (2008) consider the following simulation model to generate a first significant digit distribution (FSD) and call it random upper bound model (RUBM):

"It seems plausible to explore whether the first digit law is a consequence of the finite nature of real data sets. The RUBM assumes that natural numbers span from 1 till an upper bound (for example 250). We call the number 250 "upper bound". For the case of uniform distribution of probability, number 1 will appear with a probability of 111/250 = 0.44, number 2 appears with 61/250 = 0.244, etc. RUBM assumes that the upper bound changes randomly. For each upper bound a number was randomly picked out and 10,000 simulations were performed for obtaining a frequency distribution histogram."

While such a model is a priori quite interesting its definition is incomplete because it does not specify the distribution according to which the upper bound changes randomly. Presumably, the authors mean "uniform distribution" but this notion cannot be grasped without precise mathematical modelling. From their pictured histogram one sees that the simulation comes very close or even coincides with the first digit law of Stigler (1945), which has been further discussed by Raimi

* Tel.: ++41 797188283.
E-mail address: whurlimann@bluewin.ch

http://dx.doi.org/10.1016/j.joi.2015.02.007
1751-1577/

(1976), Rodriguez (2004) and Lee, Cho, and Judge (2010). In particular, Rodriguez demonstrates how Benford's, Stigler's and the uniform FSD's can be embedded into a one-parameter extension by assuming a power law random behaviour for the upper bound. Using a modified more natural finite support for the random upper bound, we provide in Section 2 a new simpler proof that the Cáceres et al. RUBM coincides with the Stigler-RUBM first digit distribution as the number of simulations grows to infinity. In the context of statistical distributions, the random upper bound can be viewed as *extended truncated Pareto* distributed with arbitrary real index, i.e. as "analytical continuation" of the truncated Pareto with positive index. The obtained FSD is independent of the truncation point. In the special case of a positive Pareto index, we show that it coincides with the FSD from a RUBM with Pareto distributed upper bound. This is shown within the context of the uniform mixture model of Rodriguez (2004) with a general distribution of the random upper bound, called hereafter *uniform random upper bound* (URUB) family of FSD's.

Further specializations of the URUB family yield in Section 3 some new FSD's of independent interest. The upper bound mixing distribution is alternatively specified as a truncated version from below of the gamma, Erlang and Burr distributions respectively. As their FSD fitting capabilities are compared with the generalized Benford (GB) law, some brief information on it is included. In particular, based on the extended truncated Pareto distribution a new probabilistic derivation of the GB law is given. A relationship with an exponential Benford (EB) law is also given.

Applications to real-world data from various scientific disciplines (including some scientometric data) are presented in Section 4. Benford's law, which concerns a special but specific aspect of information, is an analytical tool of study in informetrics, a name coined by Nacke (1979) (see also Tague-Sutcliffe, 1992). Informetrics encompasses many subfields, in particular scientometrics, bibliometrics and webometrics. At the beginning of the 21st century one has a bibliography by Hood and Wilson (2001) and a review by Bar-Ilan (2008). Some books about informetrics include Egghe and Rousseau (1990) and Egghe (2005). The first digit phenomenon is mentioned in Brookes and Griffiths (1978) and Brookes (1984). Recent applications to scientometric data are due to Campanario and Coslado (2010) and Alves, Yanasse, and Soma (2014). Parts of their data will be used to illustrate the impact of the new method within informetrics. Based on the entire data analysis, we propose the use of a *fine structure index* for Benford's law in case Benford like data is well fitted by a truncated *Erlang* URUB FSD.

2. The uniform random upper bound family and the Stigler-RUBM FSD

Consider the uniform random upper bound (URUB) family of FSD's (to the decimal base) introduced in Rodriguez (2004). Given is a uniform random variable $U[0, b)$ with upper bound uniquely written as $b = m \cdot 10^k + c$, $m \in \{1, 2, \ldots, 9\}$, where m is an integer and $c \in [0, 10^k)$. Conditional on the value of b the probability that a random number drawn from $U[0, b)$ has a first significant digit $d \in \{1, 2, \ldots, 9\}$ is determined by (Rodriguez (2004), Eq. (1))

$$P(d/b) = \frac{10^{k(d)+1}}{9b} + \frac{b - m \cdot 10^k}{b} I(d),$$
$$I(d) = \begin{cases} 1, & d = m, \\ 0, & d \neq m, \end{cases} \qquad k(d) = \begin{cases} k, & d < m, \\ k-1, & d = m. \end{cases} \tag{2.1}$$

The general URUB family is defined to be equal to the FSD associated to the mixture of uniform random variables $U[0, b)$, where the random upper bound b has a distribution $F(b)$ with support S contained in the interval $[1, \infty)$. If the support is bounded we assume for simplicity that it is of the form $S_N = [1, 10^N)$ for some $N \geq 1$ and in case it is unbounded we set $S_\infty = \lim_{N\to\infty} [1, 10^N) = [1, \infty)$. Writing the support as disjoint union of intervals as $S_N = \bigcup_{k=0}^{N-1} [10^k, 10^{k+1})$ and using Eq. (2.1) one shows similarly to Rodriguez (2004), Eqs. (2) and (3), that the defined mixture of uniform random variables with support S_N has FSD

$$P_N(d) = \int_1^{10^N} P(d/b) dF(b) = \sum_{k=0}^{N} \left\{ \int_{10^k}^{d\cdot 10^k} \frac{10^k}{9b} dF(b) + \int_{d\cdot 10^k}^{(1+d)\cdot 10^k} \left(\frac{10^k}{9b} + \frac{b - d\cdot 10^k}{b} \right) dF(b) + \int_{(1+d)\cdot 10^k}^{10^{k+1}} \frac{10^{k+1}}{9b} dF(b) \right\}. \tag{2.2}$$

We show that (2.2) can be written in closed form in terms of the two finite series survival like functions

$$S_{\bar{F},N}(x) = \sum_{k=0}^{N-1} \bar{F}(x \cdot 10^k), \quad \bar{F}(x) = 1 - F(x),$$
$$S_{\bar{G},N}(x) = \sum_{k=0}^{N-1} 10^k \cdot \bar{G}(x \cdot 10^k), \quad \bar{G}(x) = \int_x^{10^N} b^{-1} dF(b) \tag{2.3}$$

Proposition 2.1 (FSD of the URUB family). *The first significant digit distribution of the URUB family with random upper bound distribution $F(b)$ supported on $S_N = [1, 10^N)$ is determined by*

$$P_N(d) = \tfrac{1}{9}(S_{\bar{G},N}(1) - 10 \cdot S_{\bar{G},N}(10)) + (S_{\bar{F},N}(d) - d \cdot S_{\bar{G},N}(d)) - (S_{\bar{F},N}(1+d) - (1+d) \cdot S_{\bar{G},N}(d)). \tag{2.4}$$