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

(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

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

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

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

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*

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

Proof. Write (2.2) as a sum/difference of four terms $P_N(d)=\sum_{k=0}^{N}\{I_1(k)+I_2(k)-I_3(k)+I_4(k)\}$ with

$$\begin{aligned}
I_1(k)&=\tfrac{10^k}{9}\int_{10^k}^{(1+d)\cdot 10^k} b^{-1}dF(b)=\tfrac{10^k}{9}(\bar{G}(10^k)-\bar{G}((1+d)\cdot 10^k)),\\
I_2(k)&=\int_{d\cdot 10^k}^{(1+d)\cdot 10^k} dF(b)=\bar{F}(d\cdot 10^k)-\bar{F}((1+d)\cdot 10^k),\\
I_3(k)&=d\cdot 10^k\int_{d\cdot 10^k}^{(1+d)\cdot 10^k} b^{-1}dF(b)=d\cdot 10^k(\bar{G}(d\cdot 10^k)-\bar{G}((1+d)\cdot 10^k)),\\
I_4(k)&=\tfrac{10^{k+1}}{9}\int_{(1+d)\cdot 10^k}^{10^{k+1}} b^{-1}dF(b)=\tfrac{10^{k+1}}{9}(\bar{G}((1+d)\cdot 10^k)-\bar{G}(10^{k+1})).
\end{aligned}\tag{2.5}$$

Inserted into the preceding expression and simplifying one obtains immediately (2.4).□

Next, as in Rodriguez (2004), Section 2.2, we suppose that the random upper bound follows an arbitrary power law. However, instead of supporting it on a single interval $[10^k, 10^{k+1})$ we support it on the finite disjoint union of such intervals, i.e. on $S_N=[1, 10^N)$. Furthermore, we interpret the power law as an *extended truncated Pareto* law *ETPar*(α, N) with parameter $\alpha\in(-\infty,\infty)$ and probability density function

$$f_\alpha^N(b)=\begin{cases}\alpha\left(1-10^{-\alpha N}\right)^{-1}b^{-(\alpha+1)}, & b\in S_N,\quad \alpha\neq 0,\\ (bN\ \ln 10)^{-1}, & b\in S_N,\quad \alpha=0.\end{cases}\tag{2.6}$$

Proposition 2.2 (FSD of the *ETPar*(α,N)-URUB distribution). *The first significant digit distribution of the URUB family with extended truncated Pareto upper bound density $f_\alpha^N(b)$ supported on $S_N=[1, 10^N)$ is determined by*

$$P_N(d)=\begin{cases}\frac{\alpha}{\alpha+1}\left(\frac{1}{9}+\frac{d^{-\alpha}-(1+d)^{-\alpha}}{\alpha(1-10^{-\alpha})}\right), & \alpha\neq 0,-1,\\ \log(1+d^{-1}), & \alpha=0\quad(\text{Benford}),\\ \frac{1}{9}\left(1+\frac{10}{9}\ \ln 10+d\ln d-(1+d)\ln(1+d)\right), & \alpha=-1\quad(\text{Stigler}).\end{cases}\tag{2.7}$$

Proof. Three cases must be distinguished.

Case 1: $\alpha=0$ (Benford's law)

In the notations preceding Proposition 2.1 one has $\bar{F}(x)=1-\frac{1}{N}\log x$, $\bar{G}(x)=\frac{1}{N\ln 10}\left(\frac{1}{x}-10^{-N}\right)$, $S_{\bar{F},N}(x)=\frac{1}{2}(N+1)-\log x$, $S_{\bar{G},N}(x)=\frac{1}{N\ln 10}\left(\frac{N}{x}-\frac{1}{9}(1-10^{-N})\right)$. Inserted into (2.4) one gets Benford's law.

Case 2: $\alpha=-1$ (Stigler's law)

Similarly, one has $\bar{F}(x)=1-\frac{x}{10^N-1}$, $\bar{G}(x)=\frac{1}{10^N-1}(N\ln 10-\ln x)$, $S_{\bar{F},N}(x)=N-\frac{1}{9}x$, and $S_{\bar{G},N}(x)=A_N-\frac{1}{9}\ln x$ with $A_N=\frac{\ln 10}{10^N-1}\sum_{k=0}^{N-1}(N-k)\cdot 10^k$. Inserted into (2.4) one obtains after simplification Stigler's law.

Case 3: $\alpha\neq 0,-1$ (generic case)

From (2.6) one gets successively $\bar{F}(x)=\frac{x^{-\alpha}-10^{-\alpha N}}{1-10^{-\alpha N}}$, $\bar{G}(x)=\frac{\alpha}{\alpha+1}\frac{x^{-(\alpha+1)}-10^{-(\alpha+1)N}}{1-10^{-\alpha N}}$, $S_{\bar{F},N}(x)=\frac{x^{-\alpha}}{1-10^{-\alpha}}-A_N(\alpha)$, with $A_N(\alpha)=\frac{N10^{-\alpha N}}{1-10^{-\alpha N}}$, $S_{\bar{G},N}(x)=\frac{\alpha}{\alpha+1}\frac{x^{-(\alpha+1)}}{1-10^{-\alpha}}-B_N(\alpha)$, with $B_N(\alpha)=\frac{\alpha}{\alpha+1}\sum_{k=0}^{N-1}\frac{10^{-\alpha N-(N-k)}}{1-10^{-\alpha N}}$. Inserted into (2.4) one obtains after rearrangement (2.7).□

Besides the new simple general FSD formula (2.4) for the URUB family, the given proof of the one-parameter Rodriguez FSD (2.6) is shorter than the original one. Note that the parameter x of Rodriguez is related to the extended truncated Pareto index setting $\alpha=-(x+1)$. Since the FSD does not depend on the support one can ask whether it coincides for $\alpha>0$ with the FSD of the *Par*(α)-URUB distribution. The affirmative answer holds because in the limiting case as $N\to\infty$ one has $\lim_{N\to\infty} f_\alpha^N(b)=\alpha b^{-(\alpha+1)}$, $b\in[1,\infty)$, which is the Pareto density. This can also be verified very simply using (2.4) by noting that $S_{\bar{F},\infty}(x)=\frac{x^{-\alpha}}{1-10^{-\alpha}}$ and $S_{\bar{G},\infty}(x)=\frac{\alpha}{\alpha+1}\frac{x^{-(\alpha+1)}}{1-10^{-\alpha}}$.

Corollary 2.3 (FSD of the *Par*(α)-URUB distribution). *The first significant digit distribution of the URUB family with Pareto upper bound density coincides with (2.6) for $\alpha>0$.*

The last result has the following theoretical implication. For $\alpha>0$ the mixing *Par*(α) and the mixing *ETPar*(α, N) distributions in the URUB family are instances of probabilistic models for which it is not possible to distinguish between a Pareto and a truncated Pareto distribution. They lead both to the same FSD. In extreme value theory (EVT) the general problem of deciding between a Pareto-type and a truncated Pareto-type has been recently studied extensively in Beirlant, Fraga Alves, Gomes, and Meerschaert (2014). These authors have been motivated by typical applications, where it is not known a priori if a truncated distribution is more appropriate in tail fitting than an un-truncated one, as previously discussed in Aban, Merschaert, and Panorska (2006). For this reason, estimation methods should ideally apply in both settings. They develop a statistical EVT method that fulfils this purpose and illustrate it with some earthquake related data. Our example illustrates that this decision problem does not makes sense in all settings. This observation might be important for the study of FSD's from real-world data (cf. the geophysical data in Sambridge, Tkalčić, and Jackson, 2010). In our special situation it is not necessary to bother upon the truncation point of the Pareto as FSD's are equal being mixed from truncated or un-truncated Pareto distributions.

3. Other members of the URUB family and the generalized Benford law

Having obtained a closed-form formula for the FSD of the general URUB family, it is natural to look for alternatives to the Pareto URUB family introduced by Rodriguez. Among the many possibilities for the upper bound mixing distributions, we retain here truncated versions from below of the gamma, Erlang and Burr distributions. Their FSD fitting capabilities are compared in Section 4 with the generalized Benford (GB) law, which is a popular one-parameter extension of Benford's law. Based on the extended truncated Pareto distribution a new probabilistic derivation of it is proposed. A relationship with an exponential Benford (EB) law is also given.

3.1. The gamma URUB family

Consider the probability density and distribution functions of the *Gamma*(α, β) with shape and rate parameters $\alpha, \beta > 0$. For $x > 0$ they are given by $f(x) = \Gamma(\alpha)^{-1} x^{\alpha-1} e^{-\beta x}$, $F(x) = \Gamma(x; \alpha, \beta)$. The survival function is denoted by $\bar{F}(x) = \bar{\Gamma}(x; \alpha, \beta) = 1 - \Gamma(x; \alpha, \beta)$. Restricting the support to the interval $[1, \infty)$ the truncated gamma mixing survival function of the gamma URUB is replaced by

$$\bar{F}(x) = \frac{\bar{\Gamma}(x; \alpha, \beta)}{\bar{\Gamma}(1; \alpha, \beta)}, \quad x \in [1, \infty), \quad \alpha > 1, \quad \beta > 0 \tag{3.1}$$

Note that to simplify the numerical evaluation of formulas, the shape parameter is restricted to the range $\alpha > 1$. This is due to the fact that the survival like function $\bar{G}(x)$ in (2.3) takes the form

$$\bar{G}(x) = \int_x^{\infty} t^{-1} f(t) dt = \frac{\beta}{\alpha - 1} \cdot \frac{\bar{\Gamma}(x; \alpha - 1, \beta)}{\bar{\Gamma}(1; \alpha, \beta)}, \quad x \in [1, \infty), \quad \alpha > 1, \quad \beta > 0. \tag{3.2}$$

Together with (2.3) and (2.4) the gamma URUB FSD is herewith completely specified.

3.2. The Erlang URUB family

The *Erlang*(n, λ) distribution is the special case of the *Gamma*(α, β) setting $\lambda = \beta > 0$ and restricting the shape parameter $\alpha = n = 1, 2, 3, ...$ to a positive integer. The support of the Erlang URUB mixing distribution is again the infinite interval $[1, \infty)$. For numerical evaluation it is necessary to distinguish between two cases.

Case 1: $n = 2, 3, 4, \ldots$

It is well-known that the expressions (3.1) and (3.2) collapse to finite series as follows:

$$\bar{F}(x) = \frac{E(x; n, \lambda)}{E(1; n, \lambda)}, \quad \bar{G}(x) = \frac{\lambda}{n - 1} \cdot \frac{E(x; n - 1, \lambda)}{E(1; n, \lambda)}, \quad E(x; n, \lambda) = e^{-\lambda x} \cdot \sum_{j=0}^{n-1} \frac{(\lambda x)^j}{j!}. \tag{3.3}$$

Case 2: $n = 1$

The *Erlang*$(1, \lambda)$ coincides with the exponential *Exp*(λ) and the mixing survival like distributions are defined by

$$\bar{F}(x) = e^{\lambda} \cdot e^{-\lambda x}, \quad \bar{G}(x) = \lambda e^{\lambda} \cdot \int_x^{\infty} t^{-1} e^{-\lambda t} dt, \quad x \in [1, \infty), \quad \lambda > 0. \tag{3.4}$$

For the interested reader let us mention that an infinite series expansion of the exponential integral in (3.4) exists (e.g. Gradshteyn and Ryzhik (2007), p. 884).

3.3. The Burr URUB family

The probability density and the distribution functions of the *Burr*(c, γ) with parameters $c, \gamma > 0$ are given respectively by $f(x) = c\gamma \cdot x^{c-1}(1 + x^c)^{-(\gamma+1)}$ and $F(x) = 1 - (1 + x^c)^{-\gamma}$, with $x > 0$. It is an extension of the Pareto type II obtained in the special case $c = 1$. Restricting the support to the interval $[1, \infty)$ the truncated Burr mixing survival function of the Burr URUB is replaced by

$$\bar{F}(x) = 2^{\gamma} \cdot (1 + x^{c})^{-\gamma}, \quad x \in [1, \infty), \quad c > 0, \quad \gamma > 0, \tag{3.5}$$

and one has

$$\bar{G}(x) = 2^{\gamma} c\gamma \cdot \int_{x}^{\infty} t^{c-2} (1 + t^{c})^{-(\gamma+1)} dt, \quad x \in [1, \infty), \quad c > 0, \quad \gamma > 0. \tag{3.6}$$

3.4. The generalized Benford law

The generalized Benford law $GB(\alpha)$, with real parameter $\alpha \in (-\infty, \infty)$, is a popular one-parameter extension of Benford's law defined by

$$GB(d; \alpha) = \left\{ \begin{array}{ll} \dfrac{d^{-\alpha} - (1+d)^{-\alpha}}{1 - 10^{-\alpha}}, \quad \text{if} \quad \alpha \neq 0, & \\ & d = 1, \ldots, 9 \\ \log(1 + d^{-1}), \quad \text{if} \quad \alpha = 0, & \end{array} \right\}. \tag{3.7}$$

Up to the correct normalizing constant, which guarantees that (3.7) is a FSD, the GB has been derived in Pietronero, Tossati, Tossati, and Vespignani (2001), Eq. (3), from the power law solution to the functional equation for scale-invariance. The authors also derive a relationship between GB and Zipf's law. The latter has been used by Egghe and Guns (2012) (see also Egghe, 2011) to derive the GB from Zipf's law. A new probabilistic proof of it based on the extended truncated Pareto distribution follows. Though arbitrary bases can be considered the focus is here on decimal base.

Proposition 3.1 (Generalized Benford Law). *Suppose that the random variable $X(\alpha)$ has an extended truncated Pareto distribution with parameter $\alpha \in (-\infty, \infty)$, bounded support $[1, 10^{N}]$, $N \geq 1$, and probability density $f_{X(\alpha)}(x) = \left(1 - 10^{-N\alpha}\right)^{-1} \alpha \cdot x^{-(\alpha+1)}$ if $\alpha \neq 0$, and $f_{X(0)}(x) = (x \cdot N \ln 10)^{-1}$ if $\alpha = 0$. Then, the probability that $X(\alpha)$ has random first significant digit d in the decimal base is determined by (3.7).*

Proof. For simplicity set $X = X(\alpha)$, and let D denote the integer-valued random variable satisfying the inequality $10^{D} \leq X < 10^{D+1}$. Then, the first significant digit Y of X can be written as $Y = \lfloor X \cdot 10^{-D} \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function, and one has

$$P(Y = d) = \sum_{k=0}^{N-1} P(d \cdot 10^{k} \leq X < (1 + d) \cdot 10^{k}) = \sum_{k=0}^{N-1} \int_{d \cdot 10^{k}}^{(1+d) \cdot 10^{k}} f_{X}(x) dx. \tag{3.8}$$

Integrating (3.8) one obtains without difficulty the FSD (3.7).□

In the limiting case $\alpha \to 0$ the two formulas in (3.7) are consistent and generate Benford's law. One sees that the GB generates monotone decreasing probabilities if, and only if, one has $\alpha \geq -1$, where the limiting case $\alpha = -1$ is the discrete uniform distribution. Therefore, GB is defined as monotone decreasing FSD for all $\alpha \in [-1, \infty)$ and includes Benford's law as special case $\alpha = 0$. Moreover, this FSD is tilted towards a uniform distribution for $\alpha \in [-1, 0)$ and is more tilted than Benford's law for $\alpha \in (0, \infty)$. Next, we show and emphasize that an exponential random variable can also generate (3.7) for the special case $\alpha \in [0, \infty)$ only, however.

Proposition 3.2 (Exponential Benford Law). *For $b \in (0, 1)$ let $W(b)$ be a random variable with exponential density $f_{W(b)}(w) = -\ln(b) \cdot b^{w}$, $w \in [0, \infty)$, so that $X(b) = 10^{W(b)}$ has the probability density $f_{X(b)}(x) = -(x \cdot \ln 10)^{-1} \ln(b) \cdot b^{\log(x)}$, $x \in [1, \infty)$. Then, the probability that $X(b)$ has random first significant digit d in the decimal base is given by*

$$EB(d; b) = \frac{d^{\log(b)} - (1 + d)^{\log(b)}}{1 - b}, \quad d = 1, \ldots, 9. \tag{3.9}$$

Proof. Inserted into (3.8) with $N \to \infty$ one obtains the FSD (3.9) from the calculation

$$EB(d; b) = \sum_{k=0}^{\infty} b^{k} (b^{\log(d)} - b^{\log(1+d)}) = \frac{b^{\log(d)} - b^{\log(1+d)}}{1 - b} = \frac{d^{\log(b)} - (1 + d)^{\log(b)}}{1 - b}.$$

Furthermore, as $b \to 1$ in (3.9) one obtains from l'Hôpital's rule that

$$\lim_{b \to 1} EB(d; b) = \lim_{b \to 1} \frac{\log(d) \cdot \{b^{-1} d^{\log(b)}\} - \log(1 + d) \cdot \{b^{-1} (1 + d)^{\log(b)}\}}{-1} = \log(1 + d^{-1}), \tag{3.10}$$

which coincides with Benford's law. In view of these equations, the EB is justified for $b \in (0, 1]$. With the transformation of parameters $\alpha = -\log(b)$ the EB identifies with the GB for the restricted parameter range $\alpha \in [0, \infty)$. Without explicit mention the EB will always be identified with its GB counterpart.

Table 4.1
Real-world first digit data sets.

Real-world data sets	First digit								
	1	2	3	4	5	6	7	8	9
MEG	0.2580	0.2130	0.1610	0.1150	0.0820	0.0560	0.0450	0.0380	0.0320
Earth's gravity	0.3296	0.1660	0.1120	0.0850	0.0750	0.0670	0.0594	0.0557	0.0503
Geomagnetic field	0.2890	0.1770	0.1330	0.0940	0.0810	0.0690	0.0610	0.0510	0.0450
Seismic wavespeeds	0.3004	0.1760	0.1330	0.0980	0.0790	0.0640	0.0560	0.0489	0.0447
Star distances	0.3240	0.2130	0.1290	0.0890	0.0680	0.0520	0.0450	0.0410	0.0390
Dow Jones Index	0.2329	0.1888	0.1433	0.1168	0.0930	0.0746	0.0600	0.0500	0.0406
Population of countries	0.2741	0.1629	0.1230	0.1061	0.0934	0.0684	0.0653	0.0531	0.0537
Twitter followers	0.3262	0.1666	0.1181	0.0926	0.0763	0.0655	0.0577	0.0514	0.0456
Terrorism deaths	0.5193	0.1814	0.0990	0.0640	0.0469	0.0325	0.0247	0.0187	0.0135
Articles 10Y science	0.2811	0.1632	0.1273	0.1046	0.0863	0.0739	0.0624	0.0536	0.0476
Citations 10Y science	0.2994	0.1766	0.1244	0.0943	0.0789	0.0675	0.0608	0.0520	0.0461
Impact factors 10Y science	0.3118	0.1742	0.1172	0.0904	0.0796	0.0680	0.0597	0.0524	0.0467
Articles 5Y science	0.2767	0.1677	0.1326	0.1044	0.0877	0.0696	0.0611	0.0537	0.0465
Articles 5Y social science	0.2290	0.2650	0.1871	0.1177	0.0694	0.0506	0.0318	0.0260	0.0234

Table 4.2
MAD critical values and conformity to FSD.

MAD critical values	Conformity to FSD	Abbreviation
$\text{MAD} \le 6 \times 10^{-3}$	Close conformity	C
$6 \times 10^{-3} < \text{MAD} \le 12 \times 10^{-3}$	Acceptable conformity	AC
$12 \times 10^{-3} < \text{MAD} \le 15 \times 10^{-3}$	Marginal conformity	MC
$\text{MAD} > 15 \times 10^{-3}$	Nonconformity	NC

4. Application to real-world data sets: the fine structure index of Benford's law

To analyze the fitting capabilities of the considered FSD's a sample of fourteen data sets has been collected in Table 4.1 below. The first so-called MEG data (magneto-encephalograms from a healthy male) is from Cáceres et al. (2008) (cf. the RUBM discussion in Section 1). The next three real-world FSD's are taken from Sambridge et al. (2010) and concern the Geophysical and Earth Sciences. The fifth data set is about star distances discussed in Alexopoulos and Leontsinis (2014), and Fox and Hill (2014). The next one is the stock market data of Ley (1996) (Dow Jones Index) that has been extensively analyzed in Rodriguez (2004) and is an excellent illustration of the Pareto URUB family. The next three data sets are taken from the Social Sciences: the population data of countries around the world, the twitter users by followers count, and the people killed by terrorism (1970–2013), all found in Long (2014). The last five data sets illustrate the impact of the method for scientometrics, a subfield of informetrics. The first three are taken from Campanario and Coslado (2010), which is the first paper that analyzes Benford's law in scientometrics. They describe the first digit frequencies of the number of articles published, citations received and impact factors of all journals indexed in the Science Citation Index from 1998 to 2007 (aggregates over 10 years). Similarly, the last two describe the first digit frequencies of the number of articles published of journals indexed in the JCR® Sciences and Social Sciences Edition from 2007 to 2011 (5-year aggregates of the Tables 3 and 4 in Alves et al. (2014)).

Unfortunately, there exists at present no simple exact mathematical test to decide whether a given FSD conforms to Benford's law or another related parametric extension to it (for some recent efforts consult Morrow (2014), however). Despite this theoretical lack, a lot of experience has been accumulated to assess conformity to Benford's law, which might be extended to parametric FSD's. In this respect, the *mean absolute deviation* (MAD) test developed by Nigrini (2012), Table 7.1, suffices for our purpose (see Table 4.2). Recall the definition of the MAD statistics. Given two FSD's, which may depend on parameters or not, say $F_1(d)$ and $F_2(d)$, $d = 1, \ldots, 9$, the MAD measure is defined and denoted by

$$\text{MAD} = \frac{1}{9} \cdot \sum_{d=1}^{9} \left| F_1(d) - F_2(d) \right| . \tag{4.1}$$

The MAD statistics are calculated in Table 4.4 and re-used in Table 4.5 as follows. First, the minimum MAD estimators of the alternatives are computed and reported in Table 4.3, and their minimum values are listed in Table 4.4. Taking into account the (extended) critical values in Table 4.2 the conformity to the various FSD's is then assessed in Table 4.5.

Besides decision upon conformity, we use additionally the probability *weighted least squares* (WLS) measure used earlier by Leemis, Schmeier, and Evans (2000) (chi-square divided by sample size) to decide upon the preferred FSD choices. Again, this measure can be used for both theoretical FSD's or/and FSD's derived from sample data. Indeed, suppose $F_1(d)$ must be

Table 4.3
Minimum MAD estimators for the parametric URUB families.

Real-world FDD's	Minimum MAD estimators							
	Pareto α	Erl n	Erl λ	Gamma α	Gamma β	Burr c	Burr γ	GB α
MEG	−0.72618	5	0.00009703	4.700239	0.0000903	0.325688	0.303968	0.06539
Earth's gravity	0.519055	2	0.011313	2.208835	0.001243	0.372809	2.700676	−0.11603
Geomagnetic field	−0.217526	1	0.02022565	1.157985	0.021883	0.073306	0.680563	0.05005
Seismic wavespeeds	−0.011998	1	0.03122034	1.000001	0.030807	0.059993	0.505738	0.00203
Star distances	0.123828	3	0.008862	3.479352	0.0000099	0.248327	2.072275	−0.14144
Dow Jones Index	−0.969346	4	0.00005598	3.354999	0.0000471	0.134829	0.386285	0.26924
Population	−0.466144	2	0.00002294	2.154090	0.0000251	0.249429	0.279422	0.11326
Twitter followers	0.298598	1	0.00642082	1.352793	0.0076667	0.266579	2.118897	−0.10240
Terrorism deaths	0.565490	20	0.00000865	177.1143	0.0000764	2.945286	0.377731	−0.85339
Articles 10Y science	−0.351139	2	0.02274071	1.893601	0.0215797	0.059970	1.300629	0.08339
Citations 10Y science	0.021340	1	0.00775534	n.a.	n.a.	0.051090	2.632299	0.00674
Impact factors 10Y science	0.221617	1	0.00711865	1.0000001	0.0068387	0.093520	4.710478	0.00964
Articles 5Y science	−0.412188	2	0.02555946	1.873467	0.0239406	0.028151	1.381718	0.10212
Articles 5Y social science	−1.214343	11	0.00231574	10.76276	0.0022687	0.015990	0.867885	0.00000

chosen to approximate $F_2(d)$ and suppose both have been derived from a sample of same size N. Then, by definition of the WLS measure, one has

$$\mathrm{WLS} = \frac{1}{N} \cdot \sum_{d=1}^{9} \frac{(N \cdot F_1(d) - N \cdot F_2(d))^2}{N \cdot F_1(d)} = \sum_{d=1}^{9} \frac{(F_1(d) - F_2(d))^2}{F_1(d)}. \tag{4.2}$$

Theoretically, if the FSD's are known with certainty, the WLS measure does not depend on the sample size. It can therefore be used as a rule of thumb to choose the best fit to a given FSD among various alternatives.

Some interesting general observations can be borrowed from Table 4.5. Up to two exceptions the data sets conform to the Erlang- and Gamma-URUB families. The terrorism data only conforms to the GB. The 5Y social sciences data is acceptable conform to the Erlang/Gamma URUB family. For the terrorism data, a reason of the fit failure for the Erlang/Gamma URUB family might be the high proportion of numbers with first digit one, which presumably cannot be grasped this way. The 5Y social sciences data is not typically Benford like in the sense that the first digits frequencies are monotonically decreasing. Note that a complete breakdown of the Benford like pattern might not be unusual. For example, it has been observed by Ausloos, Herteliu, and Ileanu (2015) who analyze birth data. There exist data sets, for which the Pareto-URUB, the Burr-URUB and the GB are only acceptable or even marginally acceptable. The Stigler is sometimes marginally conform (e.g. MEG, geomagnetic field, population data, 5Y sciences data). It is conform to the Dow Jones data, as known to Rodriguez (2004). On the other hand, Benford's law is only conform to three data sets (geomagnetic field, seismic wavespeeds, twitter followers), acceptable conform to three of them (earth's gravity, population data and 5Y sciences data), marginally conform to star distances and not conform to the MEG data, the Dow Jones data and the 5Y social sciences data.

Table 4.4 helps differentiate and grasp better the Benford pattern, especially with regard to the selected families based on the WLS measure. First, in contrast to Cáceres et al. (2008), which claim that the RUBM/Stigler law "can likely explain the observed behaviour of MEG data", the WLS measure does not select Stigler's law to explain the MEG data. In fact, only the Erlang and Gamma-URUB models are selected and explain well the first digit behaviour. The next three data sets (earth's gravity, geomagnetic field, seismic wavespeeds), which are conform to all four URUB families, are also quite well fitted by them. For two of them (geomagnetic field, seismic wavespeeds) the gamma-URUB has the smallest min MAD measure but the first WLS choice is Erlang-URUB (however, both FSD's are quite close together). The latter property is also shared by the population data, the twitter followers, the 10Y citations data and the 5Y social sciences data. The quite recent data about star distances is very instructive. Benford's law is marginally acceptable but not selected by the WLS criterion. This might surprise in view of the two recent publications by Alexopoulos and Leontsinis (2014), and Fox and Hill (2014). The star data does not belong to the Pareto-URUB class of Rodriguez nor is it selected by Burr-URUB. It belongs to the GB but is best explained by the Erlang- and gamma-URUB (first choice being the gamma-URUB). The Dow Jones data is similar but even more special. As already known to Rodriguez (2004) Stigler's law is quite close to the optimal min MAD Pareto-URUB. But, the Erlang- and gamma-URUB improve the fit (first choice being again the gamma-URUB). While the population data and the twitter followers are respectively acceptable conform and conform to Benford's law, the best fit is for an Erlang-URUB. The three 10Y scientometric data sets are always selected by the Erlang-URUB class. The gamma-URUB minimum MAD estimator for the 10Y citations data could not be found with our computer program (probably due to a numerical instability). In this case the generalized Benford law is second choice. Without able to give any explanation, we note that the Burr-URUB is second choice for the 10Y impact factors data. The 5Y sciences data is similar to the population data but fits better the given FSD's. The fit to the Erlang/Gamma-URUB is excellent. Though it is not typically Benford the 5Y social sciences data is selected by the WLS measure for the Erlang/gamma-URUB. The terrorism data is clearly rejected by the URUB family, and we cannot explain further its excellent fit to the GB. Compared to the Erlang/Gamma-URUB family the GB never outperforms it when it is selected. With one more extra parameter the gamma-URUB has a smaller min MAD measure than the Erlang-URUB, as

Table 4.4
Goodness-of-fit to parametric FSD's (in bold the minimum MAD values).

Real-world FDD's	10^3 MAD							10^3 WLS						
	B	S	Par	Erl	Ga	Burr	GB	B	S	Par	Erl	Ga	Burr	GB
MEG	20.86	13.78	11.58	1.274	**0.811**	20.09	20.31	39.99	20.60	17.60	0.341	**0.227**	38.41	42.57
Earth's gravity	8.694	25.23	2.189	1.193	**0.769**	1.745	7.544	7.204	65.31	0.631	0.157	**0.156**	0.432	9.241
Geomagnetic field	3.522	14.10	2.469	2.102	**1.813**	3.473	2.507	1.367	21.98	1.045	**0.751**	0.794	1.385	1.098
Seismic wavespeeds	2.034	15.98	1.915	1.695	**1.653**	2.273	2.013	0.856	27.40	0.796	**0.411**	0.438	0.868	0.880
Star distances	14.21	26.06	14.11	2.303	**1.433**	13.71	8.313	21.11	62.64	22.27	0.798	**0.478**	23.80	8.620
Dow Jones Index	16.54	3.436	3.354	3.502	**3.156**	16.97	13.09	27.08	2.052	2.035	1.552	**1.513**	28.35	21.12
Population	9.347	14.14	7.744	3.450	**3.059**	10.04	4.119	9.257	24.57	10.88	**2.093**	2.202	12.10	2.382
Twitter followers	5.648	22.47	2.857	2.691	**2.597**	2.748	4.415	3.321	51.40	1.126	**0.766**	0.799	1.014	3.106
Terrorism deaths	49.68	61.77	47.65	26.08	**23.27**	44.02	**2.004**	268.3	459.2	234.0	106.9	121.4	202.2	**1.122**
Articles 10Y science	7.294	13.81	5.445	0.756	**0.377**	7.612	2.945	4.807	19.67	4.633	0.069	**0.025**	5.240	1.191
Citations 10Y science	1.124	16.83	1.106	**0.791**	n.a.	1.517	0.894	0.241	30.35	0.223	**0.131**	n.a.	0.271	0.204
Impact factors 10Y science	3.587	19.74	1.801	2.329	**1.416**	1.710	3.585	1.441	41.45	0.446	0.635	**0.314**	0.421	1.606
Articles 5Y science	7.271	12.32	4.083	1.440	**1.058**	7.429	3.366	4.742	15.78	3.202	0.318	**0.230**	5.028	1.503
Articles 5Y social science	38.19	27.50	27.20	7.754	**7.678**	38.28	38.19	137.8	90.75	94.66	**7.747**	7.828	138.1	137.8

Table 4.5
MAD test and WLS choices (WLS < 15×10^{-3}).

Real-world FDD's	Generalized MAD test							WLS criterion	
	B	S	Par	Erl	Ga	Burr	GB	Choice 1	Choice 2
MEG	NC	MC	AC	C	C	NC	NC	Gamma	Erlang(5)
Earth's gravity	AC	NC	C	C	C	C	AC	Gamma	Erlang(2)
Geomagnetic field	C	MC	C	C	C	C	C	Erlang(1)	Gamma
Seismic wavespeeds	C	NC	C	C	C	C	C	Erlang(1)	Gamma
Star distances	MC	NC	MC	C	C	MC	AC	Gamma	Erlang(3)
Dow Jones Index	NC	C	C	C	C	NC	MC	Gamma	Erlang(4)
Population	AC	MC	AC	C	C	AC	C	Erlang(2)	Gamma
Twitter followers	C	NC	C	C	C	C	C	Erlang(1)	Gamma
Terrorism deaths	NC	NC	NC	NC	NC	NC	C	GB	no choice
Articles 10Y science	AC	MC	C	C	C	AC	C	Gamma	Erlang(2)
Citations 10Y science	C	NC	C	C	C	C	C	Erlang(1)	GB
Impact factors 10Y science	C	NC	C	C	C	C	C	Erlang(1)	Burr
Articles 5Y science	AC	MC	C	C	C	AC	C	Gamma	Erlang(2)
Articles 5Y social science	NC	NC	NC	AC	AC	NC	NC	Erlang(11)	Gamma

expected. However, it is only selected in approximately half of the fitted cases. To us, it seems valuable to pursue further the idea that the Erlang exponent might be used as a *finite structure index* to distinguish between different patterns of Benford like data sets. Already six fitted data sets have here different indices. The index 1 occurs five times and the index 2 four times.

Let us conclude with a brief outlook. The considered few examples are close to the personal interests of the author. However, we believe the material to be of enough general interest to be applied in any scientific research. Moreover, the analyzed data fulfils the intended methodological goals. Perhaps, some interested and more specialized readers might analyze in future the first significant digits of further typical informetric related data sets.

Acknowledgment

The author is very grateful to an anonymous referee for pointing out the informetrics related paper by Alves et al. (2014).

References

Aban, I. B., Merschaert, M. M., & Panorska, A. K. (2006). Parameter estimation for the truncated Pareto distribution. *Journal of the American Statistical Association: Theory and Methods*, *101*(473), 270–277.
Alexopoulos, T., & Leontsinis, S. (2014). Benford's law in astronomy. *Journal of Astrophysics and Astronomy*, 1–10.
Alves, A. D., Yanasse, H. H., & Soma, N. Y. (2014). Benford's law and articles of scientific journals: comparison of JCR® and Scopus data. *Scientometrics*, *98*, 173–184.
Ausloos, M., Herteliu, C., & Ileanu, B. (2015). Breakdown of Benford's law for birth data. *Physica A: Statistical Mechanics and its Applications*, *419*, 736–745.
Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century: A review. *Journal of Informetrics*, *2*(1), 1–52.
Beirlant, J., Fraga Alves, M. I., Gomes, M. I., & Meerschaert, M. M. (2014). *Extreme value statistics for truncated Pareto-type distributions*. http://arxiv.org/pdf/1410.4097.pdf
Brookes, B. C. (1984). Ranking techniques and the empirical log law. *Information Processing and Management*, *20*(1–2), 37–46.
Brookes, B. C., & Griffiths, J. M. (1978). Frequency-rank distributions. *Journal of the American Society for Information Science*, *29*, 5–13.
Cáceres, H., García, J. L. P., Martínez Ortiz, C. M., & Dominguez, L. G. (2008). First digit distribution in some biological data sets. Possible explanations for departures from Benford's law. *Electronic Journal of Biomedicine*, *1*, 27–35.
Campanario, J. M., & Coslado, M. A. (2010). Benford's law and citations, articles and impact factors of scientific journals. *Scientometrics*, *88*(2), 421–432.
Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics*. Amsterdam: Elsevier Academic Press.
Egghe, L. (2011). Benford's law is a simple consequence of Zipf's law. *ISSI Newsletter*, *7*(3), 55–56.
Egghe, L., & Guns, R. (2012). Application of the generalized law of Benford to informetric data. *Journal of the American Society for Information Science and Technology*, *63*(8), 1662–1665.
Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics: Quantitative methods in library, documentation, and information science*. Amsterdam: Elsevier Science Publishers.
Fox, R. F., & Hill, T. P. (2014). *Hubble's law implies Benford's law for distances to stars*. arXiv:1412.1536.v2.[physics.data-an]
Gradshteyn, I. S., & Ryzhik, I. M. (2007). *Table of integrals, series, and products* (7th ed.). Burlington/San Diego/London: Academic Press/Elsevier.
Hood, W. W., & Wilson, C. S. (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, *52*(2), 291–314.
Lee, J., Cho, W. K. T., & Judge, G. G. (2010). Stigler's approach to recovering the distribution of first significant digits in natural data sets. *Statistics & Probability Letters*, *80*(2), 82–88.
Leemis, L. M., Schmeier, B. W., & Evans, D. L. (2000). Survival distributions satisfying Benford's law. *The American Statistician*, *54*(3), 1–6.
Ley, E. (1996). On the peculiar distribution of the US stock indexes' digits. *The American Statistician*, *50*(4), 311–313.
Long, J. (2014). *Testing Benford's law*. http://testingbenfordslaw.com/
Morrow, J. (2014). Benford's law, families of distributions and a test basis. In *CEP discussion paper no. 1291*. London: London School of Economics and Political Science.
Nacke, O. (1979). Informetrie: ein neuer Name für eine neue Disziplin. *Nachrichten für Dokumentation*, *30*(6), 219–226.
Nigrini, M. J. (2012). *Benford's law. Applications for forensic accounting, auditing, and fraud detection*. Hoboken, NJ: John Wiley & Sons.

Pietronero, L., Tossati, E., Tossati, V., & Vespignani, A. (2001). Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf. *Physica A*, *293*, 297–304.

Raimi, R. A. (1976). The first digit problem. *The American Mathematical Monthly*, *83*(7), 521–538.

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

Sambridge, M., Tkalčić, H., & Jackson, A. (2010). Benford's law in the natural sciences. *Geophysical Research Letters*, *37*, L22301.

Stigler, G. J. (1945). *The distribution of leading digits in statistical tables. Working paper*. University of Chicago, Regenstein Library Special Collections, George J. Stigler Archives.

Tague-Sutcliffe, J. (1992). An introduction to informetrics. *Information Processing and Management*, *28*(1), 1–3.