

A Survey on Sequences and Distribution Functions satisfying the First-Digit-Law.

Draft-Version: 25th October 2004

Peter N. Posch

Department of Finance, University of Ulm, Germany.

(e-mail: posch@mathematik.uni-ulm.de)

Abstract

====Draft Version. Only cite with permission from the author====

This paper provides a broad overview of both distribution function and sequences following the Newcomb-Benford Law. For the so-called Benford-Sequences strict mathematical proofs are given, which enable the reader to find close connections to other datasets and empirical problems.

Key words: First digit law, goodness-of-fit, quality of datasets

JEL Classification: xxx

1 Introduction

Almost every month a new empirical application of the Newcomb-Benford-Phenomenom¹ is discovered. It seems that the practical possibilities are as wide as the fancifulness of committed researchers. This paper gives a broad overview on the recent development. Furthermore I classify several distribution functions with specific parameter-classes on their fit on the

¹Other names for this empirical phenomenom include 'First-Digits-Law'', just 'Digit Law'' or, which despites the origin of Simon Newcomb, 'Benfords Law''.

Newcomb-Benford-Law (short: NBL). For some mathematical series a close fit can be shown analytically.

2 Formalities

First we introduce some notation. Any number $x \in \mathbb{R}$ can be written as $x = m_b \cdot b^n$ with $m_b \in [1, b - 1[$, for some $n \in \mathbb{N}$, where $b \in \{2, 3, \dots\}$ represents the basis of the number-system and m_b is called mantissae of x .² $D_k^{(b)}(x)$ denotes the k -th significant digit of $x \in \mathbb{R}^+$ in base b , i.e. $D_2^{(10)}(\pi) = 1$ and $\log_b(x)$ is the b -logarithm of $x \in \mathbb{R}^+$. When in the decimal system ($b = 10$) the superscript is suppressed. A more general version of the First-digit-Law can now be given by the following equation:

$$P(M_b(x) \leq t) = \log_b(t), \forall x \in \mathbb{R}^+ \quad (1)$$

using this equation one can easily (see e.g. Hill [1996]) derive the probability of a leading significant digit $d \in \{0, 1, \dots, b - 1\}$:

$$P(D_1(x) = d) = \log_b \left(1 + \frac{1}{d} \right) \quad (2)$$

and more general for any position $i > 2$:

$$P(D_n^{(b)} = d_n) = \sum_{k=b^{n-2}}^{b^{(n-1)}-1} \log_b \left(1 + (k \cdot b + d_n)^{-1} \right) . \quad (3)$$

Following Hill [1996] the so-called mantissae- σ -algebra will be denoted by \mathcal{M}_b and a probability measure, which assigns the above quoted probabilities to the significant digits and is countable-additive will be called 'benford probability measure (b.p.m.)'. Hill [1996] shows that a b.p.m. is scale- and base-invariant, which means that a change of the base (the b above) or a multiplication by any positive real number will not change the frequency of expected numbers.³

From now on let p_i^e be the expected probability of digit i , i.e. the Benford-probability in equation 1 or equation 3 respectively, and let p_i be the analogous observed frequency in a given dataset. Furthermore let $k \in \{0, 1, 2, \dots\}$ denote the order of the digit-class looked at, e.g. $k = 1$ would be the class of the first-significant digits, whereas $k = 3$ would refer to the digits from 100 to 999.

²Note that some authors use $m_b \in]0, b[$ instead, which is just a rescaling by the factor $\frac{1}{b}$.

³The restriction to positive numbers can be easily released, for our purposes taking the absolute value should be enough.

2.1 Generation of Benford-Random Variables

Let $UNI(a, b)$ be a uniform distribution on $[a, b]$. Using direct inversion of the definition (1) a random variable (RV) X can be derived, which digits exactly follow the NBL.

$$M_b \leftarrow b^\eta \text{ with } \eta \sim UNI(0, 1)$$

A RV D_1 in base 10 with exact fit on the NBL is thus generated by $D_1 \leftarrow \lfloor 10^\eta \rfloor$.

This method explains the early note of Newcomb [1881] stating that the mantissae of logarithm of the observed digits are distributed uniformly. (Note that Newcomb uses the term 'mantissae' in the sense of 'logarithmic mantissae', i.e. the fractional part of the logarithm of a real number.):

Let X be a RV, which realisations are the digits of which the mantissa is to be computed. X can be written as $X = M_b \cdot b^K$. Thus $M_b = X \cdot b^{-K} = b^{\log_b(X) - K}$. If $\log_b(X) - K \sim UNI(0, 1)$, then M_b is distributed according to the logarithmic mantissa-distribution and X is a Benford-RV.

With simple arithmetic it is obvious that $(\log_b(X) - K)$ is just the fractional part – equivalent the logarithm modulo one – $\log_b(X) - \lfloor \log_b(X) \rfloor$. This number is given in the logarithm-tables used by Newcomb and is called 'logarithmic Mantissae'. The equivalent proposition $\log_b(M_b) \sim UNI(0, 1)$ can be illustrated very intuitively.

Consider a circular slide rule. This slide rule multiplies numbers (or mantissae) via addition of their logarithms. E.g. $2 \cdot 9 = 10^{0.301} \cdot 10^{0.954} = 10^{1.255} \approx 18$.

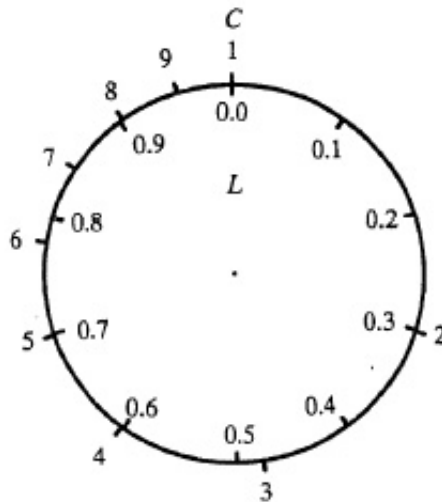


Figure 1: c-slide rule [Boyle, 1994, p. 880].

The numbers themselves are not uniformly distributed on the c -scale, but their logarithms on the L -scale indeed are.

The following interpretation by Boyle [1994], p. 881 shows which processes lead to a Benford RV. Consider the numbers which are to be multiplied as Mantissae and the slide-rule as a Wheel of Fortune. The successive multiplication of the RV is then given by a sequence of wheel turns. Assume every turn to be independent. Despite how 'unfair' one is trying to turn the wheel, after a while the RV 'Position of the Wheel of Fortune' will be approx. uniformly distributed. This corresponds to a uniform distribution of the logarithm of the Mantissae and (see above) to the NBL.

Remark. To test on the NBL one could test whether $(\log_b(X) - \lfloor \log_b(X) \rfloor) \sim UNI(0, 1)$ or $\log_b(M_b) \sim UNI(0, 1)$.

The direct inversion method can be generalized as follows:

Theorem 2.1. Let $X := b^\eta$ with $\eta \sim UNI(\alpha, \beta)$ and $\alpha < \beta$; $\alpha, \beta \in \mathbb{Z}$. The distribution of first significant digits, base b , follows the NBL, i.e. $P(D_1^{(b)}(X) = d_1) = \log_b(1 + d_1^{-1})$.

For the proof of this theorem the following transformation theorem is needed.

Theorem 2.2 (Transformation-Theorem). Let X be a RV with density f_X . Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be a continuous function, differentiable on $]a, b[$ ($a < b$), which covers the support of X . Furthermore assume $\frac{dg}{dx}(x) \neq 0, \forall x \in]a, b[$ and assume the existence of an inverse function g^{-1} of g with the support $D(Y) = g(D(X))$. Then $Y = g(X)$ is a RV with density function:

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{dg^{-1}}{dx}(x) \right| & , \text{ with } y \in g(D(X)) \\ 0 & , \text{ else.} \end{cases}$$

Proof of Theorem 2.1.

Let $\eta \sim UNI(\alpha, \beta)$ with $\alpha < \beta$, $g(x) := b^x$ and the RV X be given by $X := g(\eta) := b^\eta$. The inverse function of $g(x)$ is $g^{-1}(x) = \log_b(x)$.

The density function of η is given by:

$$f_\eta(x) = \begin{cases} \left(\frac{1}{\beta-\alpha}\right) & , \text{ with } \alpha \leq x \leq \beta \\ 0 & , \text{ else.} \end{cases}$$

With Theorem 2.2 the density function of X is given by:

$$f_X(x) = \begin{cases} \frac{1}{(\beta-\alpha)x \ln(b)} & , \text{ with } b^\alpha \leq x \leq b^\beta \\ 0 & , \text{ else.} \end{cases}$$

The distribution function of X is then:

$$F_X(x) = \begin{cases} 0 & , \text{ with } x < b^\alpha \\ \frac{\log_b(x) - \alpha}{\beta - \alpha} & , \text{ with } b^\alpha \leq x \leq b^\beta \\ 1 & , \text{ with } x > b^\beta . \end{cases}$$

and for the distribution of the first digits $D_1^{(b)} = d_1$:

$$\begin{aligned} P(D_1^{(b)} = d_1) &= \sum_{k=\alpha}^{\beta-1} P(d_1 \cdot b^k \leq X < (d_1 + 1) \cdot b^k) \\ &= \sum_{k=\alpha}^{\beta-1} \int_{d_1 \cdot b^k}^{(d_1+1)b^k} f_X(x) dx \\ &= \sum_{k=\alpha}^{\beta-1} \int_{d_1 \cdot b^k}^{(d_1+1)b^k} ((\beta - \alpha)x \ln(b))^{-1} dx \\ &= \sum_{k=\alpha}^{\beta-1} [(\beta - \alpha)^{-1} \cdot \log_b(x)]_{d_1 \cdot b^k}^{(d_1+1) \cdot b^k} \\ &= (\beta - \alpha)^{-1} \sum_{k=\alpha}^{\beta-1} \left(\log_b \left(\frac{(d_1 + 1) \cdot b^k}{d_1 \cdot b^k} \right) \right) \\ &\quad \text{(because the sum has } \beta - 1 - \alpha + 1 \text{ addends)} \\ &= (\beta - \alpha)^{-1} \cdot (\beta - 1 - \alpha + 1) \cdot \log_b \left(\frac{d_1 + 1}{d_1} \right) \\ &= \log_b(1 + d_1^{-1}) \end{aligned}$$

□

Example. $b = 10, \alpha = 2, \beta = 4, d = 2$. Then $X \in [100, 10.000]$ and:

$$\begin{aligned} P(D_1 = d) &= \sum_{k=\alpha}^{\beta-1} P(d \cdot b^k \leq X < (d + 1) \cdot b^k) \\ &= P(d \cdot 10^2 \leq X < (d + 1) \cdot 10^2) \\ &\quad + P(d \cdot 10^3 \leq X < (d + 1) \cdot 10^3) \\ &= P(200 \leq X < 300) \\ &\quad + P(2000 \leq X < 3000) \end{aligned}$$

Remark. Leemis et al. [2000], p. 238 remark that a generalization for $\alpha, \beta \in \mathbb{R}$ is possible, if $(\beta - \alpha) \in \mathbb{N}$ is assumed.

It is easy to show that the reciprocal of a Benford RV is still a Benford RV:

Proof . For all $t \in [1, b]$:

$$\begin{aligned}
P\left(M_b\left(\frac{1}{X}\right) \in [1, t[\right) &= P\left(\frac{1}{X} \in \bigcup_{n=-\infty}^{\infty} [1, t \cdot b^n \right) \\
&= P\left(X \in \bigcup_{n=-\infty}^{\infty}]\frac{1}{t}, 1] \cdot b^n \right) \\
&= P\left(M_b(X) \in \left] \frac{b}{t}, b \right] \right) \\
&= \log_b(b) - \log_b\left(\frac{b}{t}\right) \\
&= \log_b(t) .
\end{aligned}$$

□

3 Conditions for Benford-Sequences

From the theoretical examination of the properties of the NBL the question arises whether there are conditions under which a given sequence of real numbers would fulfill equation 1. Diaconis [1977] shows a close connection between sequences whose frequency of significant digits approaches $\log(1 + d^{-1})$, and the theory of uniformly modulo 1 (uni mod 1) distributed sequences. The latter theory is founded by Weyl [1916]. The following results are 'classic' in the sense, that we do not define a proper probability space. The logarithmic digit-distribution is rather obtained as a limit value of the observed significant digit frequency.

Following Berger et al. [2000] we define any real sequence in base b , which fulfills the NBL as follows:

Definition 3.1 (b -Benford-Sequence). A real sequence $(x_n)_{n \in \mathbb{N}}$ is called *b-Benford-Sequence*, if

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n I(M_b(|x_i|) \leq t) = \log_b(t)$$

holds for all $t \in [1, b[$.

Diaconis [1977] examines only decimal bases and uses the same definition 3.1 for what he calls 'strong Benford-Sequences' (see also Schatte [1989], p. 254). A sequence for which the above given definition holds for all bases $b > 1$ is called 'strict Benford-Sequence' (Berger et al. [2000]):

Definition 3.2 (Strict Benford Sequence). A real sequence $(x_n)_{n \in \mathbb{N}}$ is called (*strict*) *Benford-Sequence*, if for all $b \in \mathbb{N} \setminus \{1\}$: $(x_n)_{n \in \mathbb{N}}$ is a b -Benford-Sequence.

The above given definitions are propositions on the limit-frequency of a given dataset and therefore there are different definitions supposable. From now on we use the term 'Benford-Sequence' (BS) as given by definition 3.1 with $b = 10$, but any extension on arbitrary $b > 1$ is easily possible.

The connection between a BS and the sequences uniformly distributed modulo 1 is given by the following theorem (also consider the remark on p. 4):

Theorem 3.3 (Diaconis [1977], p. 74). A real sequence $(x_n)_{n \in \mathbb{N}}$ is a BS, if $(\log(|x_n|))_{n \in \mathbb{N}}$ is distributed uniformly mod 1.

Proof . For a more detailed (and formal) proof see Diaconis [1977], p. 74. Here a more intuitiv consideration (following an idea of Jamain [2001]) should be enough for our purposes: If $(\log(a_n))_{n \in \mathbb{N}}$ is distributed uniformly mod 1, then the sequence $(a_n)_{n \in \mathbb{N}}$ can be considered as a sub-sample of a sample which fulfills the NBL. (Compare the variate-generating method in chapter 2.1). Since the definition 3.1 deals solely with frequencies or rather limit values of frequencies, (a_n) is a BS. The reverse implication is proven analogously.

As a direct consequence, a transformation of a given BS $(x_n)_{n \in \mathbb{N}}$ by multiplying $\alpha \in \mathbb{R}$ or raising to any integer power $k \in \mathbb{Z}$, does not harm the BS, i.e. $(\alpha x_n^k)_{n \in \mathbb{N}}$ is still a BS. (Berger et al. [2000])

In addition to Theorem 3.3, the following corollary to a Theorem of Fejer (Theorem A.3, S. 15 in the appendix) gives a necessary condition for a given sequence beeing a BS:⁴

Corollary 3.4 (Diaconis [1977], p. 74).

For any BS $(x_n)_{n \in \mathbb{N}}$: $\limsup_{n \rightarrow \infty} n \left| \log \frac{|x_{n+1}|}{|x_n|} \right| = \infty$.

Proof . Directly, using Theorem A.3 with $f(n) := \log(x_n)$ and Theorem 3.3.

3.1 Examples for BS

We now list some real sequences which follow and some which do not fulfill the NBL. Some of the below given classes of sequences explain a lot of the empirical findings.

3.1.1 Geometric Sequence

From the Weyl-Criterium (see Theorem A.2 in the appendix) and Theorem 3.3 the following condition for geometric sequences building a BS can be derived:

⁴Note that Diaconis does not use the absolute function in his paper, which leads to wrong results.

Corollary 3.5. Let $x > 0$. A geometric Sequence $(x^n)_{n \in \mathbb{N}}$ is a BS, if $\log(x)$ is irrational.

Proof . Using Theorem A.2 it follows for any irrational θ : The sequence $(n\theta)_{n \in \mathbb{N}}$ is distributed uniformly mod 1. Now Theorem 3.3 completes the Proof. \square

Solely this results explains a variety of sequences, such as (2^n) used in Benford [1938]. Moreover the appearance of the Firs-Digit-Phenomenon in manifold natural processes can be explained.

3.1.2 Synthetic Benford-Sequences

Nigrini [2000], p. 10-12 proposes the use of geometric sequences to artificially generate BS, which he calls *synthetic Benford-Sequences*. Such a sequence $(S_n)_{n \in \mathbb{N}}$ on the real interval $[a, b]$ with N elements is generated by $S_n = k \cdot r^{n-1}$ using $r := 10^{d/N}$ and d given by $d := \log(\frac{b}{a})$. Note that the upper bound b is reached only approximately for sufficient large N .

3.1.3 Some other Sequences

It can be easily shown, that sequences analysed by Benford [1938], such as $(n^n)_{n \in \mathbb{N}}$ and $(n!)_{n \in \mathbb{N}}$ are BS. The proofs are given in the appendix, p. 16.

Furthermore Diaconis [1977], p. 75f. shows that the binomial array $\left(\binom{n}{k}\right)_{n \in \mathbb{N}}$ and a variety of recursive defined sequences approach the First-Digit-Law. For a more detailed treatment see also Schatte [1988] and Berger et al. [2000].

Using the corollary 3.5, it can be shown that neither $(n^b)_{n \in \mathbb{N}}$, nor $(bn)_{n \in \mathbb{N}}$ nor $(\log_b(n))_{n \in \mathbb{N}}$ are BS for arbitrary bases b . The proofs are given on page 17 in the appendix and provide a methodical insight on how a given sequence can be checked on a fit to the NBL.

Finally, neither the sequence of primes $(P_n)_{n \in \mathbb{N}}$ nor their logarithms $(\log P_n)_{n \in \mathbb{N}}$, where P_n denotes the n -th prime, are BS in the sense of definition 3.1. A proof is given by Diaconis [1977], p. 74. Following Whitney [1972] $(P_n)_{n \in \mathbb{N}}$ can be considered as a 'weak BS', see Goto [1992] for details.

It has been shown empirically that the Fibonacci- and Lucas-Numbers build a BS. Table 3.1.3 provides the details on this sequences, but recall that neither number set is independent and thus the χ^2 -Test does not provide clear critical values. (On the problem how to test on the NBL, see Posch [2004]).

Wlodarski [1971] observes the first 100 Fibonacci numbers ⁵ and Lucas numbers, while

⁵Recall that the Fibonacci sequence $(F_n)_{n \in \mathbb{N}}$ is defined recursively by $F_1 = F_2 = 1, F_{n+2} = F_n + F_{n+1}$.

Sentance [1973] extended the observation to the first 1.000 numbers. The results given below were calulated with the first two million Fibonacci numbers.

Test	χ^2	KS	MAD
Value	(0.0002496)	$1.2499 \cdot 10^{-6}$	$7.9144 \cdot 10^{-7}$

Table 1: Test Statistic for the first 2 Mio. Fibonacci numbers.

One reason for the excellent fit to the NBL could be the fact, that both sequences can be approximated by ϕ , where $\phi = \frac{1}{2}(\sqrt{5} - 1)$ is the so called 'golden cut' (see [Bronstein et al., 1999, p. 840]).

Digit	Abs.	Rel.	Abs.Dev.	Z-Stat
1	602 060	0.3010	0	$1 \cdot 10^{-5}$
2	352 185	0.1760	0	0.00461
3	249 877	0.1249	0	0.00101
4	193 817	0.0969	0	0.00723
5	158 366	0.0791	0	0.00919
6	133 890	0.0669	0	0.01013
7	115 985	0.0579	0	0.00335
8	102 305	0.0511	0	0.00014
9	91 515	0.0457	0	6E-05

Table 2: Digit Frequency for the first digits of the first 2 million Fibonacci numbers.

3.2 Distribution Functions

Beneath the theoretical deviation of conditions on which sequences follow the NBL, a closer look on distribution functions is required to make statements on the conditions of their possible convergence to the NBL. A distribution function is said to fulfill the NBL, if its underlying random variable is a Benford-RV.

Leemis et al. [2000] examine – using computer-based methods by Glen et al. [2001] – parametrical survival functions and some other common distributions on their goodness-of-fit to the NBL. As measures for the goodness of fit a χ^2 -test and the maximum absolute deviation of the empirical first significant digits d_1 to the expected digit frequency is used. Let Y be a Benford-RV and X the first significant digit in the survival time of the function

$S(t) = P(T \geq t)$ with $P(X = d_1) = \sum_{i=-\infty}^{\infty} [S(d_1 \cdot 10^i) - S((d_1 + 1) \cdot 10^i)]$ with $d_1 = 1, 2, \dots, 9$.

The Chi-Squared test-statistic is given by

$$c := \sum_{d_1=1}^9 \frac{[P(X = d_1) - P(Y = d_1)]^2}{P(Y = d_1)} \text{ und}$$

and the maximum absolute deviation:

$$m := \max_{d_1} \{|P(X = d_1) - P(Y = d_1)|\} .$$

Table 3 shows some of the results for different parameter values. The scale-parameter is given by λ , while the shape parameter is denoted as κ . The authors remark that dealing with distributions with shape-parameter κ the goodness-of-fit raises with increasing κ . Distributions with two parameters (κ, λ) are more sensitive on changes of κ and respond less on $\Delta\lambda$.

Distribution	λ	κ	c	m
Exponential	1	-	$0.61 \cdot 10^{-2}$	$0.29 \cdot 10^{-1}$
Exponential	5	-	$0.54 \cdot 10^{-2}$	$0.18 \cdot 10^{-1}$
Muth	-	0.1	$0.13 \cdot 10^{-1}$	$0.41 \cdot 10^{-1}$
Gompertz	5	1.1	$0.62 \cdot 10^{-2}$	$0.20 \cdot 10^{-1}$
Weibull	1	0.3	$0.37 \cdot 10^{-10}$	$0.16 \cdot 10^{-5}$
Weibull	1	2	0.19	0.11
Gamma	1	0.3	$0.15 \cdot 10^{-3}$	$0.29 \cdot 10^{-2}$
Gamma	1	2	$0.48 \cdot 10^{-1}$	$0.50 \cdot 10^{-1}$
Log Logistic	1	0.3	$0.86 \cdot 10^{-21}$	$0.67 \cdot 10^{-11}$
Log Logistic	1	2	$0.24 \cdot 10^{-1}$	$0.35 \cdot 10^{-1}$

Table 3: Reproduced of Leemis et al. [2000].

3.2.1 Construction of Benford-Sequences

Following chapter 2.1 every RV $X = b^\eta$ with $\eta \sim UNI(\alpha, \beta)$ and $\alpha, \beta \in \mathbb{N}$, $\alpha < \beta$ follows Benfords Law exactly, i.e. for the whole mantissae distribution. This arises the question how further distributions with exact fit in the digits frequency can be constructed. Leemis et al. [2000] show that these distributions are of the form $X = b^\eta$ where the support of η is an interval whose endpoints are integers.

Leemis et al. [2000], p. 238. Let η_1 be triangular distributed on $[0, 2]$ and η_2 non-symmetric on $[-1, 1]$ with the density function $f_{\eta_2}(x) = 1 - x^2$, for $-1 \leq x \leq 0$ and $f_{\eta_2}(x) = (x - 1)^2$, for $0 < x \leq 1$. Now both $X_1 = 10^{\eta_1}$ and $X_2 = 10^{\eta_2}$ are Benford RV

Before the proof is given, recall that the density function of a triangular distribution with parameters a, b, c is given by

$$f_{\eta}(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & , a < x < b \\ \frac{2(b-x)}{(c-a)(c-b)} & , b \leq x < c \end{cases}$$

Proof . Thus the triangular distribution on $[0, 2]$ has the density

$$f_{\eta}(x) = \begin{cases} x & , \text{für } 0 \leq x \leq 1 \\ 2 - x & , \text{für } 1 \leq x \leq 2 . \end{cases}$$

Let $Z := \eta - \lfloor \eta \rfloor$. The distribution function of Z is given by conditioning on $\lfloor \eta \rfloor$ for all $z \in [0, 1]$:

$$\begin{aligned} F_Z(z) &= P(\lfloor \eta \rfloor = 0) \cdot P(\eta \leq z \mid \lfloor \eta \rfloor = 0) \\ &\quad + P(\lfloor \eta \rfloor = 1) \cdot P(\eta - 1 \leq z \mid \lfloor \eta \rfloor = 1) \\ &= P(\eta \leq z \cap 0 \leq \eta < 1) \\ &\quad + P(\eta \leq z + 1 \cap 1 \leq \eta < 2) \\ &= \int_0^z x dx + \int_1^{z+1} (2 - x) dx \\ &= \frac{z^2}{2} + \left(z - \frac{z^2}{2} \right) \\ &= z , \end{aligned}$$

i.e. $Z \sim UNI(0, 1)$. Using the remark on page 4 completes the proof, i.e. $X = b^{\eta}$ is Benford-distributed.

Remark. This result can be generalized for $\eta \sim Triangular(a, b, c)$ with $a, b, c \in \mathbb{N}^*$ ($a < b < c$). See [Leemis et al., 2000, p. 238].

Use the same notation and procedure to prove the case of the asymmetric density:

$$f_{\eta}(x) = \begin{cases} 1 - x^2 & , \text{with } -1 \leq x \leq 0 \\ (x - 1)^2 & , \text{with } 0 \leq x \leq 1 \end{cases}$$

which yields the distribution function of Z :

$$\begin{aligned}
F_Z(z) &= P(\lfloor \eta \rfloor = -1) \cdot P(\eta + 1 \leq z \mid -1 \leq \eta \leq 0) \\
&\quad + P(\lfloor \eta \rfloor = 0) \cdot P(\eta \leq z \mid 0 \leq \eta \leq 1) \\
&= P(\eta \leq z - 1) + P(0 \leq \eta \leq z) \\
&= \int_{-1}^{z-1} (1 - x^2) dx + \int_0^z (x - 1)^2 dx \\
&= (z - 1) - \frac{(z - 1)^3}{3} + 1 - \frac{1}{3} + \frac{(z - 1)^3}{3} + \frac{1}{3} = z .
\end{aligned}$$

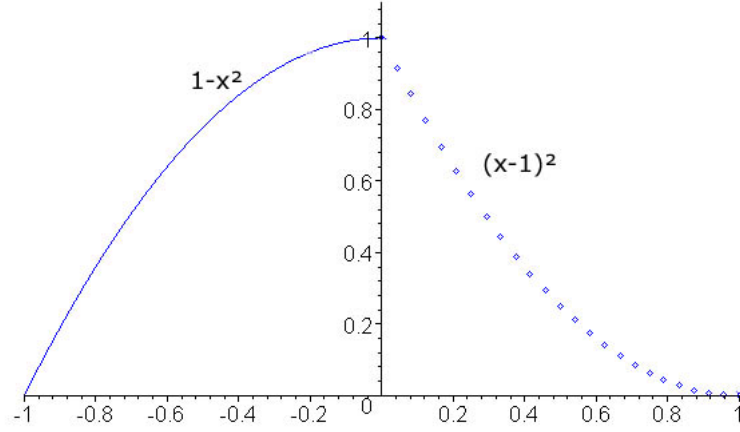


Figure 2: Asymmetric Benford distribution on $[-1, 1]$.

Another interesting result by Leemis et al. [2000], p. 240 is the fact that a mixture of Benford-RV is still a Benford-RV:

Theorem 3.6. Let X_1 and X_2 be non-negative Benford-RV. Let X be a RV with density $f_X(x) = p \cdot f_{X_1}(x) + (1 - p) \cdot f_{X_2}(x)$ with $x > 0$ and $0 \leq p \leq 1 \Leftrightarrow X$ is a Benford RV.

Proof . The proof is based on an idea by Leemis et al. [2000].

Let $Z_1 = \log_b X_1 - \lfloor \log_b X_1 \rfloor$, $Z_2 = \log_b X_2 - \lfloor \log_b X_2 \rfloor$ and $Z = \log_b X - \lfloor \log_b X \rfloor$. Because X_i ($i = 1, 2$) are Benford-RV: $F_{Z_i}(z) = z$. It is left to show, that the mixture fulfills $F_Z(z) = z$:

$$\begin{aligned}
F_X(x) &= p \cdot F_{X_1}(x) + (1 - p) \cdot F_{X_2}(x) \\
&= pz + (1 - p)z \\
&= z .
\end{aligned}$$

□

The following results hold for RV with the form $X = 10^{|\eta|}$ (see again Leemis et al. [2000]).

- (i) Distributions with only one modal value, which is the upper bound of the (open) support, do not fulfill the NBL for any parameter value (e.g. $\eta \sim EXP$).
- (ii) Some limiting distributions follow the NBL.
- (iii) Some other distributions (e.g. Weibull) approximate the NBL for some parameter values, but those values belong to specific parameter families.

3.2.2 Lognormal-Distribution

Leemis et al. [2000], p. 239 remark (but do not show explicitly), that $X = 10^\eta$ mit $\eta \sim N(\mu, \sigma^2)$ for sufficiently large σ converge to the NBL. Scott and Fasli [2001] show empirically , that a logarithmic normal distributed RV ⁶ is, at least for some parameter values, Benford distributed. The authors use random number generators like $\mu \cdot e^{\sigma \cdot N(0,1)}$ to generate artificial Benford sets.

The authors test several parameter values and find, that values of $\sigma \geq 1.2$ show a good goodness-of-fit to the NBL. They notice that the scale-parameter μ does not (as expected for a scale-invariant distribution, see Posch [2004]) influence the goodness-of-fit (see Table 4 for the details).

⁶A random variable X is said to be lognormal distributed ($X \sim \Lambda(\mu, \sigma^2)$), if the logarithm of X is normal distributed: $\ln X =: Y \sim N(\mu, \sigma^2)$. Recall that the position parameter μ changes to the scale parameter while the latter σ^2 will be the shape parameter of the lognormal distribution.

Parameter	χ^2
$\Lambda(1, 0.2^2)$	19160.89
$\Lambda(1, 1.2^2)$	14.06
$\Lambda(1, 10^2)$	10.66
$\Lambda(2, 0.2^2)$	21409.13
$\Lambda(2, 1.2^2)$	13.33
$\Lambda(2, 10^2)$	5.14
$\Lambda(2, 100^2)$	4.08
$\chi^2_{8;0.95}$	15.51
# Observations	10 000

Table 4: χ^2 -Goodness-of-fit for lognormal RV.

The Lognormal distribution can explain the NBL in a lot of natural phenomenons, such as economic values, the growth rate and natural occurrence of different species or the concentration of atmospheric particels. Furthermore the use of Lognormal distribution when dealing with the estimation of Options and Futures in financial markets (e.g. the Binomialmodel of Cox/Rox/Rubinstein, see eg. [Schürger, 1998, p. 461-480]) would indicate a more theoretical examination of the linkage between this distribution family and the NBL. (see Sandmann and Sondermann [1997] and Crow and Shimizu [1988] for more applications of lognormal distributed values)

4 Conclusion

The paper gives a comprehensive lists of distribution and sequences following the Newcomb-Benford Law. The selection given here is not complete, but provides a broad overview of this growing subfield of digital analysis.

A Sequences and Distributions with Benford Digits

The following definitions and theorems are needed for some of the above given proofs. For a more detailed treatment (and the proof of the below given theorems) see Kuipers and Niederreiter [1975].

A.1 Conditions for Benford Sequences

Definition A.1 (Uniform distribution mod 1). A real sequence $(x_n)_{n \in \mathbb{N}}$ is uniformly distributed modulo one (UNI mod 1), if for every pair a, b ($a, b \in \mathbb{R}$) with $0 \leq a < b \leq 1$:

$$\lim_{N \rightarrow \infty} \frac{A([a, b[; N; x)}{N} = b - a ,$$

where $A(E; N; x)$ is the number of terms of type x_n ($1 \leq n \leq N$) which $(x_n - \lfloor x_n \rfloor) \in E$

Theorem A.2 (Weyl-Criterion). The sequence $(x_n)_{n \in \mathbb{N}}$ is UNI mod 1, if

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N e^{2\pi i h x_n} = 0$$

for all $h \in \mathbb{N}$.

A.1.1 (n^n) and $(n!)$

The following proofs require some further theorems, see Kuipers and Niederreiter [1975] for proofs on these theorems.

Theorem A.3 (Fejer-Theorem). If $(f(n))_{n \in \mathbb{N}}$ is UNI mod 1, then:

$$\limsup_{n \rightarrow \infty} n |f(n+1) - f(n)| = \infty.$$

Theorem A.4 (Kuipers and Niederreiter [1975], Th. 1.2). If $(x_n)_{n \in \mathbb{N}}$ is UNI mod 1 and $(y_n)_{n \in \mathbb{N}}$ such, that $\lim_{n \rightarrow \infty} (x_n - y_n) = \alpha \in \mathbb{R}$, then $(y_n) \sim UNImod1$.

Theorem A.5 (Stirling Formula). For $n \rightarrow \infty$ the following approximation holds (see Schürger [1998] pp. 232 or Feller [1966] pp. 52):

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n .$$

Theorem A.6 (Van der Corput - Theorem). Let $a, b \in \mathbb{Z}$ with $a < b$ and f be twice differentiable on $[a, b]$ with $\frac{\partial^2 f}{\partial x^2} =: f''(x) \geq \rho > 0$ or $f''(x) \leq -\rho < 0$ with $x \in [a, b]$, then ⁷

$$\left| \sum_{n=a}^b e^{2\pi i f(n)} \right| \leq (|f'(a) - f'(b)| + 2) \left(\frac{4}{\sqrt{\rho}} + 3 \right) .$$

The following proofs follow ideas by Diaconis [1977] and Jamain [2001]. We denote the Landau - Symbol by \mathcal{O} , i.e. $f(x) = \mathcal{O}(g(x))$ with $x \rightarrow \infty$ iff, \exists constants N, C with $|f(x)| \leq C|g(x)| \forall x > N$.

⁷See Kuipers and Niederreiter [1975] Theorem 2.7, p. 17

Proof nⁿ.

Let $h \in \mathbb{Z} \setminus \{0\}$, $a = 1$ and $b = N \in \mathbb{N}$. Furthermore define $f(x) := hx \log(x)$, $x > 0$. Then:

$$f'(x) = h \log(x) + h$$

and

$$f''(x) = \frac{h}{x} \Rightarrow \rho = \frac{|h|}{N}$$

Using Theorem A.6:

$$\begin{aligned} \frac{1}{N} \left| \sum_{n=1}^N e^{2i\pi hn \log(n)} \right| &\leq \frac{1}{N} (|h| \log(N) + 2) \left(4\sqrt{\frac{N}{|h|}} + 3 \right) \\ &= \mathcal{O}\left(\frac{\log(N)}{\sqrt{N}}\right) \rightarrow 0 \text{ for } N \rightarrow \infty. \end{aligned}$$

With Theorem A.2: $(n \log(n)) \sim UNI \text{ mod } 1$ and using Theorem 3.3 completes the proof. \square

Proof on n! Considering Theorem 3.3 it is sufficient to show, that $\log(n!) \sim UNImod1$.

Using Stirlings Formula:

$\log(n!) = o(1) + (n + \frac{1}{2}) \log(n) - \frac{n}{\ln(10)} + \frac{1}{2} \log(2\pi)$ (recall: $\log(x) = \frac{\ln(x)}{\ln(10)}$). For that reason the sequence

$$\log(n!) - \left((n + \frac{1}{2}) \log(n) - \frac{n}{\ln(10)} \right)$$

with $n \rightarrow \infty$ tends to a constant. Using Theorem A.4 it is left to show, that $((n + \frac{1}{2}) \log(n) + kn) \sim UNImod1$.

Let $f(x) = h(x + \frac{1}{2}) \log(x) + h k x$, $x > 0$, with $h \in \mathbb{Z} \setminus \{0\}$ and $k \in \mathbb{R}$ constant. Furthermore let $a = 1$ and $b = N \in \mathbb{N}$. Then:

$$f'(x) = h \log(x) + \frac{h(x + \frac{1}{2})}{x \ln(10)} + h k$$

and

$$f''(x) = \frac{h}{x \ln(10)} - \frac{h}{2x^2 \ln(10)} \Rightarrow \rho = \frac{|h|}{\ln(10)} \left(\frac{1}{N} - \frac{1}{2N^2} \right)$$

With analogy to the proof given above the use of the Van der Corput Theorem yields:

$$\begin{aligned} \frac{1}{N} \left| \sum_{n=1}^N e^{2\pi i h f(n)} \right| &\leq \frac{1}{N \ln(10)} \left(|h| \log(N) + \left| \frac{h}{2} \right| \left(\frac{1}{N} - 1 \right) + 2 \right) \left(\frac{4\sqrt{2 \ln(10)} N}{\sqrt{|h|(2N-1)}} + 3 \right) \\ &= \mathcal{O} \left(\frac{\log(N)}{\sqrt{N}} \right) \rightarrow 0 \text{ für } N \rightarrow \infty \end{aligned}$$

The proof is completed with the Weyl-Criterium and Theorem 3.3. □

A.1.2 Other Sequences

It is left to show that neither (n^b) , nor (bn) or $(\log_b n)$ for arbitrary bases b build a BS (see Jamain [2001], p. 44): The proofs given below show a methodology for the use of the corollary 3.4, S. 7.

Proof .

- For (n^b) : $n \cdot \ln\left(\frac{(n+1)^b}{n^b}\right) = nb \ln(1 + n^{-1}) \rightarrow b$ ($n \rightarrow \infty$).
- For (bn) : $n \cdot \ln\left(\frac{b(n+1)}{bn}\right) = n \ln(1 + n^{-1}) \rightarrow 1$ ($n \rightarrow \infty$).
- For $(\log_b n)$: $n \ln\left(\frac{\log_b(n+1)}{\log_b n}\right) = n \ln\left(1 + \frac{\ln(1+n^{-1})}{\ln n}\right) \sim (\ln n)^{-1} \rightarrow 0$ ($n \rightarrow \infty$).

But e.g. (x^n) : $n \cdot \ln\left(\frac{x^{n+1}}{x^n}\right) = n \ln(x) \rightarrow \infty$ ($n \rightarrow \infty$) for $x > 1$. Note that this result means solely, that one cannot negate, that (x^n) with $x > 1$ is a Benford Sequence (see also section 3.1.1). □

A.2 Distribution Function

Exponential $X \sim EXP(\lambda)$ $F_X(x) = \lambda e^{-\lambda x}$

Gamma $X \sim \Gamma(\lambda, \kappa)$ $F_X(x) = \frac{\lambda(\lambda x)^{\kappa-1} e^{-\lambda x}}{\Gamma(\kappa)}$

Gompertz $X \sim Gompert(\delta, \kappa)$ $F_X(x) = \delta \kappa^x e^{-\frac{\delta(\kappa^x - 1)}{\ln(\kappa)}}$

LogLogistic $X \sim LogLogistic(\lambda, \kappa)$ $F_X(x) = \frac{\lambda \kappa (\lambda \kappa)^{\kappa-1}}{(1 + (\lambda x)^\kappa)^2}$

Muth $X \sim Muth(\kappa)$ $F_X(x) = (e^{\kappa x} - \kappa) e^{-\frac{e^{\kappa x}}{\kappa} + \kappa x + \frac{1}{\kappa}}$

Weibull $X \sim Weibull(\lambda, \kappa)$ $F_X(x) = \kappa \lambda^\kappa x^{\kappa-1} e^{-(\lambda x)^\kappa}$

References

- Frank Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572, 1938.
- A. Berger, L.A. Bunimovich, and T.P. Hill. One-dimensional dynamical systems and benford’s law. Submitted to Transactions of the American Mathematical Society, 2000.
- Jeff Boyle. An application of fourier series to the most significant digit problem. *American Mathematical Monthly*, 101:879–886, 1994.
- I.N. Bronstein, K.A. Semendjajew, G. Musiol, and H. Mühlig. *Taschenbuch der Mathematik*. Harri Deutsch, 4., überarb. und erw. auflage edition, 1999.
- E.L. Crow and K. Shimizu. *Lognormal Distributions: Theory and Applications*. Dekker, 1988.
- Persi Diaconis. The distribution of leading digits and uniform distribution mod 1. *The Annals of Probability*, 5(1):72–81, 1977.
- William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley & Sons, 3. auflage edition, 1966.
- Andrew G. Glen, Diane L. Evans, and Lawrence M. Leemis. Appl: A probability programming language. *The American Statistician*, 55(2):156–166, May 2001.
- Kazuo Goto. Some examples of benford sequences. *Math. Journal Okayama Univ.*, 34: 225–232, 1992.
- Theodore P. Hill. A statistical derivation of the significant-digit law. *Statistical Science*, 10 (4):354–363, 1996.
- Adrien Jamain. Benford’s law. Unpublished Dissertation Report, Department of Mathematics, Imperial College, London., 2001.
- L. Kuipers and H. Niederreiter. *Uniform Distribution of Sequences*. Wiley, 1975.
- Lawrence M. Leemis, Bruce W. Schmeiser, and Diane L. Evans. Survival distributions satisfying benford’s law. *The American Statistician*, 54(3):236–241, 2000.
- Simon Newcomb. Note on the frequency of use of the different digits in natural numbers. *Amer. J. Math.*, 4:39–40, 1881.

- Mark J. Nigrini. *Digital Analysis using Benford's Law*. Global Audit, 2000.
- Peter N. Posch. Benford or not-benford? how to test for the first digit law. Working Paper, 2004.
- Klaus Sandmann and Dieter Sondermann. Log-normal interest rate models: Stability and methodology. University Bonn, Discussion Paper B-398, January 1997.
- Peter Schatte. On the almost sure convergence of floating-point mantissas and benford's law. *Mathematische Nachrichten*, 135:79–83, 1988.
- Peter Schatte. On measures of uniformly distributed sequences and benford's law. *Monatshefte Mathematik*, 107:245–256, 1989.
- Klaus Schürger. *Wahrscheinlichkeitstheorie*. Oldenburg, 1998.
- P. Scott and M. Fasli. Benford's law: An empirical investigation and a novel explanation. 2001.
- W.A. Sentance. A further analysis of benford's law. *Fibonacci Quarterly*, 11:490–494, 1973.
- Hermann Weyl. Über die gleichverteilung von zahlen modulo eins. *Mathematische Annalen*, 77:313–352, 1916.
- R.E. Whitney. Initial digits for the sequence of primes. *American Mathematical Monthly*, 79: 150–152, 1972.
- J. Wlodarski. Fibonacci and lucas numbers tend to obey benford's law. *Fibonacci Quarterly*, 9:87–88, 1971.