BENFORD'S LAW: HAMMERING A SQUARE PEG INTO A ROUND HOLE?

Félix Balado and Guénolé C. Silvestre

School of Computer Science
University College Dublin, Ireland

ABSTRACT

Many authors have discussed the reasons why Benford's distribution for the most significant digits is seemingly so widespread. However the discussion is not settled because there is no theorem explaining its prevalence, in particular for naturally occurring scale-invariant data. Here we review Benford's distribution for continuous random variables under scale invariance. The implausibility of strict scale invariance leads us to a generalisation of Benford's distribution based on Pareto variables. This new model is more realistic, because real datasets are more prone to complying with a relaxed, rather than strict, definition of scale invariance. We also argue against forensic detection tests based on the distribution of the most significant digit. To show the arbitrariness of these tests, we give discrete distributions of the first coefficient of a continued fraction which hold in the exact same conditions as Benford's distribution and its generalisation.

***Index Terms*—** Benford's law, Pareto distribution, forensics, continued fractions.

1. INTRODUCTION

Benford's distribution for the most significant digits [1], first observed by Newcomb [2], is commonly called a "law". However this name is not justified —it should better be called a distribution, the criterion that we will adopt here— because there is no general theorem explaining its emergence and apparent ubiquitousness in datasets of diverse character, in particular in naturally occurring scale-invariant data. The purpose of this article is twofold. First of all, we review some basic facts about Benford's distribution in a general and simple way, restricting our presentation to continuous random variables. Although some of the results that we give are well known, we feel that the notation previously used to deal with more than one significant digit can be improved upon to increase clarity. In the process we fill some small gaps left by previous research, and we also propose a generalisation of Benford's distribution that emerges naturally from a realistic definition of scale invariance. To conclude, we argue against the use of forensic detection tests based on the distribution of the most significant digit. These testing approaches not only discard information that may prove valuable to a forensic examiner, but can also be replaced by other potentially better tests based on continued fraction expansions, which hold in the exact same conditions as Benford's distribution or its generalisation.

Notation. Calligraphic letters are sets, and $|\mathcal{V}|$ is the cardinality of $\mathcal{V}$. Random variables (r.v.'s) are denoted by capital letters. The cumulative distribution function (cdf) of r.v. X is $F_X(x) = \Pr(X \leq x)$, where $x \in \mathbb{R}$. If X is continuous with support $\mathcal{X}$, its probability density function (pdf) is $f_X(x)$, where $x \in \mathcal{X}$. If X is uniformly distributed between a and b, we write $X \sim U(a, b)$. The probability mass function (pmf) of a discrete r.v. X is denoted by $\Pr(X = x)$. We use two versions of the unit-step function: $u(x) = 1$ if $x \geq 0$, and $u(x) = 0$ otherwise, and $u^*(x) = 1$ if $x > 0$, and $u^*(x) = 0$ otherwise. The fractional part of $x \in \mathbb{R}$ is $\{x\} = x - \lfloor x \rfloor$. To keep notation standard, curly braces are also used to list the elements of a discrete set; the meaning of $\{x\}$ is clear from the context. The rising factorial of x is $x^{\overline{m}} = x(x+1)\cdots(x+m-1) = \Gamma(x+m)/\Gamma(x)$.

2. GENERAL PROBABILITY DISTRIBUTION OF THE k MOST SIGNIFICANT b-ARY DIGITS

We start by obtaining general expressions for the distribution of the k most significant digits of numbers written in a positional base b number system, where $b \in \mathbb{N}\backslash\{1\}$. The b-ary representation of $x \in \mathbb{R}^+$ is $x = \sum_{i\in\mathbb{Z}} a_i\, b^i$, where $a_i \in \{0, 1, \ldots, b-1\}$ are the b-ary digits of x. Letting $n = \lfloor \log_b x \rfloor$, the most significant b-ary digit of x is a_n because $b^n \leq x < b^{n+1}$. Using n, the k most significant b-ary digits of x can be inferred as follows:

$$a = \lfloor x\, b^{-n+k-1} \rfloor = \lfloor b^{\{\log_b x\}+k-1} \rfloor. \tag{1}$$

Defining the set of integers

$$\mathcal{A}_{(k)} = \{b^{k-1}, \ldots, b^k - 1\}$$

whose cardinality is $|\mathcal{A}_{(k)}| = b^k - b^{k-1}$, we can see that $a \in \mathcal{A}_{(k)}$ simply by using the inequalities $0 \leq \{\log_b x\} < 1$ in (1). To give an example, say that $b = 10$ and $x = 0.00456678$. In this case $n = \lfloor \log_{10} x \rfloor = -3$, so if we for instance choose $k = 2$ then $a = \lfloor x\, 10^4 \rfloor = 45 \in \mathcal{A}_{(2)} = \{10, 11, 12, \ldots, 98, 99\}$.

It follows from (1) that $a \leq b^{\{\log_b x\}+k-1} < a+1$, or, equivalently,

$$\log_b a - k + 1 \leq \{\log_b x\} < \log_b(a+1) - k + 1. \tag{2}$$

Assume next that x is drawn from a positive continuous random variable X, and that $A_{(k)}$ is the r.v. that models the k most significant b-ary digits of X, i.e. $A_{(k)} = \lfloor b^{\{\log_b X\}+k-1} \rfloor$. Then from (2) we have that the pmf of $A_{(k)}$ is $\Pr(A_{(k)} = a) = \Pr\big(\log_b a - k + 1 \leq \{\log_b X\} < \log_b(a+1) - k + 1\big)$, where $a \in \mathcal{A}_{(k)}$. If we define

$$Y = \log_b X, \tag{3}$$

then we can write the pmf of $A_{(k)}$ in terms of the cdf of $\{Y\}$ as

$$\Pr(A_{(k)} = a) = F_{\{Y\}}\big(\log_b(a+1)-k+1\big) - F_{\{Y\}}\big(\log_b a - k + 1\big), \tag{4}$$

where $a \in \mathcal{A}_{(k)}$ and $F_{\{Y\}}(y) = \sum_{i\in\mathbb{Z}} F_Y(y+i) - F_Y(i)$.

Finally, let us denote by $A_{[j]}$ the r.v. that models the j-th most significant b-ary digit of X, i.e. $A_{[j]} = A_{(j)} \pmod b$. Obviously, $A_{[1]} = A_{(1)}$. From the definition of $A_{[j]}$, its pmf for $j \geq 2$ is

$$\Pr(A_{[j]} = a) = \sum_{r\in\mathcal{A}_{(j-1)}} \Pr(A_{(j)} = rb + a), \tag{5}$$

where $a \in \{0, 1, \ldots, b-1\}$. For large j, if $\{Y\}$ has a well-behaved pdf then it holds that $\Pr(A_{(j)} = rb + a) \approx \Pr(A_{(j)} = rb)$ because $rb + a \approx rb$ (since $rb \geq b^{j-1} \gg a$). In such case (5) is approximately constant over the range of a, and so $\Pr(A_{[j]} = a) \approx b^{-1}$.

3. BENFORD'S DISTRIBUTION AND MORE

Let us first apply the general expressions in the previous section to a r.v. X for which $\{Y\} = \{\log_b X\}$ is uniformly distributed in $[0, 1)$. We call such a r.v. a *Benford variable*. In this case $F_{\{Y\}}(y) = y$ for $y \in [0, 1)$, and so (4) becomes

$$\Pr(A_{(k)} = a) = \log_b \left(1 + \frac{1}{a}\right), \quad \text{where } a \in \mathcal{A}_{(k)}, \tag{6}$$

which is Benford's distribution for the k most significant b-ary digits. As for $A_{[j]}$ when $j \geq 2$, we have from (5) and (6) that

$$\begin{aligned}\Pr(A_{[j]} = a) &= \sum_{r \in \mathcal{A}_{(j-1)}} \log_b \left(1 + \frac{1}{rb + a}\right) \\ &= \log_b \left(\prod_{r \in \mathcal{A}_{(j-1)}} \frac{(a+1)b^{-1} + r}{ab^{-1} + r} \right) \qquad (7) \\ &= \log_b \left(\frac{\Gamma\left((a+1)b^{-1} + b^{j-1}\right) \Gamma\left(ab^{-1} + b^{j-2}\right)}{\Gamma\left((a+1)b^{-1} + b^{j-2}\right) \Gamma\left(ab^{-1} + b^{j-1}\right)} \right), \qquad (8)\end{aligned}$$

where $a \in \{0, 1, \ldots, b-1\}$. The reason for the last equality is that the numerator and the denominator of the fraction in (7) are the rising factorials $((a+1)b^{-1} + b^{j-2})^{\overline{|\mathcal{A}_{(j-1)}|}}$ and $(ab^{-1} + b^{j-2})^{\overline{|\mathcal{A}_{(j-1)}|}}$, respectively. It appears that the analytic expression (8) for the distribution of $A_{[j]}$ was never given in previous studies of Benford's distribution. We can now tighten our informal asymptotic reasoning for (5) in the particular case (8), as, using $\lim_{x\to\infty} x^{w-v}\Gamma(x + v)/\Gamma(x + w) = 1$ [3] and letting $x \propto b^j$, we formally have

$$\lim_{j\to\infty} \Pr(A_{[j]} = a) = \lim_{j\to\infty} \log_b \frac{b^{(j-1)b^{-1}}}{b^{(j-2)b^{-1}}} = b^{-1}.$$

This asymptotic uniformity of $A_{[j]}$ is evinced in (8) when $j \gtrapprox 4$ (using $b = 10$). Because convergence speed is exponential on b, the distribution of $A_{[j]}$ tends to a uniform more slowly when b is smaller.

3.1. Benford Variables and Scaling

If X is a Benford variable, then it retains its character after being scaled by $\alpha > 0$, because if $\{\log_b X\} \sim U(0, 1)$ then $\{\log_b \alpha X\} = \{\log_b X + \delta\} \sim U(0, 1)$, with $\delta = \log_b \alpha$. For the same reason, one can see using conditioning that for any positive r.v. Z independent of X the ZX product is a Benford variable. These observations about scaling give clues about the persistence of Benford variables, but not about their emergence. Still, scaling is a topic that needs careful attention, because of the claim that scale-invariant variables —which are widespread— give rise to Benford's distribution, a claim that can be traced back to Pinkham [4]. A continuous r.v. X is strictly scale invariant whenever $\Pr\left(X \in (x', x)\right) = \Pr\left(X \in (\alpha x', \alpha x)\right)$ for $\alpha > 0$, assuming that both intervals are in the support of X. Using the cdf of X, an equivalent statement of strict scale invariance is $F_X(x) - F_X(x') = F_X(\alpha x) - F_X(\alpha x')$. Differentiating this equation with respect to x, we find that the pdf of a strictly scale-invariant r.v. X must have the following property:

$$f_X(x) = \alpha f_X(\alpha x). \tag{9}$$

Let us see the implications of this property for the distribution of the r.v. Y in (3). Since the cdf of Y is $F_Y(y) = \Pr(Y \leq y) = \Pr(X \leq b^y) = F_X(b^y)$, we have that its pdf is $f_Y(y) = (\ln b)\, b^y f_X(b^y)$, or, equivalently, $f_Y(\log_b x) = (\ln b)\, x f_X(x)$. Replacing x by αx in this expression and invoking property (9), one can see that $f_Y(\log_b \alpha x) = (\ln b)\, x\, \alpha\, f_X(\alpha x) = (\ln b)\, x f_X(x) = f_Y(\log_b x)$. This equation can also be put as

$$f_Y(y + \delta) = f_Y(y),$$

which means that Y must be uniform. However Y can only be uniform over a finite support $\mathcal{Y}$, and therefore X must also have a finite support $\mathcal{X}$. Since $f_{\{Y\}}(y) = \sum_{i\in\mathbb{Z}} f_Y(y + i)$ for $y \in [0, 1)$, if $\{Y\} \sim U(0, 1)$ then the difference between the endpoints of $\mathcal{Y}$ must be integer. This in turn implies that the logarithm base b of the ratio of the endpoints of $\mathcal{X}$ must be integer for X to be Benford. Therefore, a strictly scale-invariant r.v. cannot be exactly Benford for an arbitrary choice of b, in spite of claims to the contrary that are sometimes found in the literature.

A pdf that fulfils (9) must rely on the inverse law. Thus, excluding r.v.'s with piece-wise support, the only choice for a strictly scale-invariant Benford variable is the prize-competition distribution [5]:

$$f_X(x) = \left(x \ln(x_{\mathrm{M}}/x_{\mathrm{m}})\right)^{-1}, \quad 0 < x_{\mathrm{m}} \leq x \leq x_{\mathrm{M}}.$$

Proceeding as above, this pdf leads to $Y \sim U(\log_b x_{\mathrm{m}}, \log_b x_{\mathrm{M}})$. Letting $c = \log_b x_{\mathrm{m}}$ and $d = \log_b x_{\mathrm{M}}$, and assuming $\lfloor d \rfloor > \lfloor c \rfloor$ (as if $\lfloor d \rfloor = \lfloor c \rfloor$ then $\{Y\} \sim U(0, 1)$ is not possible) we have that

$$f_{\{Y\}}(y) = \frac{1}{d - c}\Big(\lfloor d \rfloor - \lfloor c \rfloor - 1 + u\big(y - \{c\}\big) + u^*\big(\{d\} - y\big)\Big)$$

for $y \in [0, 1)$. If $d - c = \lfloor d - c \rfloor$ then $\{c\} = \{d\}$ and therefore one, and only one, of the two unit-step functions above is always 1. Also $\lfloor d - c \rfloor = \lfloor \lfloor d \rfloor + \{d\} - \lfloor c \rfloor - \{c\} \rfloor = \lfloor \lfloor d \rfloor - \lfloor c \rfloor \rfloor = \lfloor d \rfloor - \lfloor c \rfloor$. Combining these facts one can verify that $f_{\{Y\}}(y) = 1$ for $y \in [0, 1)$ only when $d - c = \log_b(x_{\mathrm{M}}/x_{\mathrm{m}})$ is integer, as discussed above.

To conclude, in spite of these considerations many recurrence relations which exhibit scale invariance in the form of geometric growth do exactly comply with Benford's distribution [6], but this is beyond our focus on continuous r.v.'s. Also, given b one may artificially craft a Benford variable X which is not scale invariant (see [7] for two examples). However the discussion above shows that it is unlikely that Benford variables originate in strictly scale-invariant natural phenomena—even for one single value of b, since the prize-competition distribution is uncommon.

3.2. Beyond Benford Variables

So one must concede that the apparent prevalence of Benford's distribution among scale-invariant variables can only be explained because it approximates some other distribution. This fact is also hinted at by the following *ad-hoc* line of reasoning: Benford's distribution is widespread because, for many r.v.'s X, the distribution of $\{Y\}$ will approximately be $U(0, 1)$ when the variance of Y is large —an argument also first given by Pinkham [4] but popularised by Feller [8], and then rightly criticised by Berger and Hill [6].

In order to delve deeper into the observation that starts this section, it is worth pointing out that scale invariance has been defined in the literature in ways more relaxed than (9), but which bear much more relevance to natural processes. In particular, consider the following criterion for scale invariance of X [9] alternative to (9):

$$f_X(x) = \alpha^{\nu} f_X(\alpha\, x). \quad (10)$$

If X is scale invariant under this definition and $\nu > 1$, then the same analysis as in the previous section will show that Y is not, in general, uniformly distributed, and so X is not, in general, Benford. It turns out that the Pareto distribution, given by

$$f_X(x) = s\, x_{\mathrm{m}}^s\, x^{-(s+1)}, \quad 0 < x_{\mathrm{m}} \leq x,\ s > 0, \quad (11)$$

is the only one that exactly conforms to (10). Because of the positivity of s, the Pareto distribution does not obey the strict scale-invariance criterion (9). However, as $s \to 0$ it becomes increasingly closer to complying with it, and thus the corresponding Pareto variable becomes increasingly closer to being Benford.

Let us apply next the general expressions in Section 2 to a Pareto r.v. X. The pdf of Y is in this case $f_Y(y) = (\ln b)\, s\, x_{\mathrm{m}}^s\, b^{-sy}$, with $y \geq \log_b x_{\mathrm{m}}$. Since $F_Y(y) = 1 - x_{\mathrm{m}}^s b^{-sy}$, the cdf of $\{Y\}$ is

$$F_{\{Y\}}(y) = b^{s(\{\log_b x_{\mathrm{m}}\}-1)}\, \frac{1-b^{-sy}}{1-b^{-s}} + u\big(y - \{\log_b x_{\mathrm{m}}\}\big)\big(1 - b^{s(\{\log_b x_{\mathrm{m}}\}-y)}\big) \quad (12)$$

for $y \in [0,1)$, where $u(\cdot)$ is the unit-step function. Combining (4) and (12) and letting $\xi = \{\log_b x_{\mathrm{m}}\} + k - 1$, the distribution of the k most significant b-ary digits for a Pareto variable is

$$\begin{aligned}\Pr(A_{(k)} = a) = {} & \frac{b^{s(\xi-1)}}{1-b^{-s}}\, \big(a^{-s} - (a+1)^{-s}\big) \\ & + u\big(a + 1 - b^{\xi}\big)\big(1 - b^{s\,\xi}(a+1)^{-s}\big) \\ & - u\big(a - b^{\xi}\big)\big(1 - b^{s\,\xi}\, a^{-s}\big),\end{aligned} \quad (13)$$

where $a \in \mathcal{A}_{(k)}$. By application of l'Hôpital's rule, it can be verified that (13) tends to (6) as $s \to 0$ —as expected from our previous discussion about strict scale invariance— and so (13) generalises Benford's distribution. Importantly, the Pareto distribution can model a wealth of processes which cannot conform exactly to the stringent condition (9) leading to scale-invariant Benford variables. According to Nair et al. [9] "(…) the Gaussian distribution is common; however, heavy-tailed distributions such as the Pareto distribution are just as (if not more) prominent. The Pareto distribution, in particular, has been observed in hundreds of applications in physics, biology, computer science, the social sciences, and beyond over the past century". Therefore, it is entirely feasible that many scale-invariant datasets traditionally associated to Benford's distribution only follow it to a first approximation, and actually follow (13) more closely.

When $\{\log_b x_{\mathrm{m}}\} = 0$ we have that $\xi = k-1$, and (13) becomes

$$\Pr(A_{(k)} = a) = \frac{a^{-s} - (a+1)^{-s}}{b^{-s(k-1)} - b^{-sk}}. \quad (14)$$

In this case it is relatively simple to generalise (8) using (5) and (14) —this generalisation is also possible using (13), but the resulting expressions are more daunting. If $s \neq 1$, the distribution of $A_{[j]}$ for $j \geq 2$ can be written in semi-analytical form as

$$\Pr(A_{[j]} = a) = \frac{\sum_{l,m=0}^{1} (-1)^{l+m} \zeta\big(s, (a+l)\, b^{-1} + b^{j-m-1}\big)}{b^{-s(j-1)} - b^{-s(j-2)}}, \quad (15)$$

where $\zeta(s, v)$ is Hurwitz's zeta function, while if $s = 1$ then

$$\Pr(A_{[j]} = a) = \frac{\sum_{l,m=0}^{1} (-1)^{l+m+1} \psi\big((a+l)\, b^{-1} + b^{j-m-1}\big)}{b^{-(j-1)} - b^{-(j-2)}}, \quad (16)$$

where $\psi(v)$ is the digamma function. In both expressions above $a \in \{0, 1, \ldots, b-1\}$. Although we have not been able to prove it formally, (15) and (16) tend to b^{-1} as j increases for the reason discussed at the end of Section 2.

Previous work and the discrete truncated Pareto distribution. In the course of preparing this paper we became aware that (14) was recently found by Barabesi and Pratelli [10] —and much earlier for the case $k = 1$ by Pietronero et al. [11] The authors of [10] however missed the fact that (14) was first identified as a new distribution only a few years ago by Kozubowski et al. [12], who called it the discrete truncated Pareto (DTP) distribution. Kozubowski et al. also noticed that the DTP generalises Benford's distribution, but, unlike Barabesi and Pratelli or ourselves, they landed on this fact solely because of the mathematical form of (14). In fact, the practical motivation in [12] was far removed from the distribution of most significant digits: it was a biological problem involving the distribution of diet breadth in Lepidoptera. Strikingly, Kozubowski et al. found the DTP through the quantisation of a *truncated* Pareto variable, instead of through the discretisation of the fractional part of the logarithm of a *standard* Pareto variable (i.e. the procedure above), which is the ultimate reason why the DTP is connected with Benford's distribution. A Pareto variable must surely be the only choice for which two such remarkably different procedures yield the very same outcome.

4. CONTINUED FRACTIONS

The term "numerology" refers to the belief in the existence of underlying numerical patterns that carry some kind of hidden significance. Whereas there is no mysticism to be found in Benford's distribution, there is at least some numerological flavour to its use in forensic detection tests (see [13] for a survey). Why pay attention to the leading digits of numbers in an arbitrary base number system (e.g. base 10) rather than paying attention to those numbers themselves? Playing devil's advocate, we could argue that a human counterfeiter may be unwittingly biased towards altering decimal digits according to some detectable pattern. Nevertheless, it is even truer that any forensic detection test based on significant digits necessarily discards precious information by focussing on a discrete subspace of the data —an applied version of Plato's allegory of the cave.

Furthermore, we will see next that there exist alternatives to using significant digits in forensic tests which are equally valid (and equally numerological), and which hold in the exact same conditions as Benford's distribution (6) or its generalisation (13). Representing numbers using a base b number system was the central theme in the previous sections. Alternatively, numbers may be represented using continued fraction expansions. Take some $w_0 \in [0,1)$, and consider the following iterative procedure starting with $j = 1$: let $z_j = w_{j-1}^{-1}$, and if $z_j < \infty$ then let $w_j = \{z_j\}$, $a_j = \lfloor z_j \rfloor$, increment j and repeat; otherwise stop. It always holds that $z_j > 1$ because $0 \leq w_{j-1} < 1$, and thus $a_j \in \mathbb{N}$. Letting $a_0 = 0$, we can express w_0 as the following continued fraction (CF):

$$w_0 = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots}}$$

To model the a_j coefficients probabilistically, we will assume that w_0 is drawn from a continuous r.v. W_0 with support $[0,1)$. Because $\mathbb{Q}$ is of measure zero then $\Pr(W_0 \in \mathbb{Q}) = 0$, and so we may assume that the CF of W_0 is infinite. We can now define the r.v.'s $Z_j = W_{j-1}^{-1}$, $W_j = \{Z_j\}$ and $A_j = \lfloor Z_j \rfloor$ associated to the CF of W_0. The distribution of coefficient A_1 in this random setting is given by $\Pr(A_1 = a) = F_{Z_1}(a+1) - F_{Z_1}(a)$, where $F_{Z_1}(z) = \Pr(Z_1 \le z) = \Pr(W_0 \ge z^{-1}) = 1 - F_{W_0}(z^{-1})$.

Let us now assume that $W_0 = \{Y\} = \{\log_b X\}$, and that X is one of the two variables considered in Section 3. Firstly, if X is a Benford variable then $F_{Z_1}(z) = 1 - z^{-1}$, and so

$$\Pr(A_1 = a) = a^{-1} - (a+1)^{-1}, \quad (17)$$

where $a \in \mathbb{N}$. Incidentally, observe the parallelism of (17) with the DTP in (14) when $s = 1$, even though (14) is for a Pareto r.v. instead; the only difference is that the support in (14) is $\mathcal{A}_{(k)} \subset \mathbb{N}$. If X is actually a Pareto r.v. (assuming $\{\log_b x_{\mathrm{m}}\} = 0$ for simplicity) then from (12) we have that $F_{Z_1}(z) = b^{-sz^{-1}}/(1 - b^{-s})$, and therefore

$$\Pr(A_1 = a) = \frac{b^{-\frac{s}{a+1}} - b^{-\frac{s}{a}}}{1 - b^{-s}}, \quad (18)$$

where $a \in \mathbb{N}$. Again, (17) is the limit of (18) when $s \to 0$.

4.1. First CF Coefficient (A_1) vs First Significant Digit ($A_{[1]}$)

Assuming that X is a Benford variable, which of the following two discrete r.v.'s should be more effective in order to undertake forensic analysis of data that is hypothesised to originate from X?

a) $A_1 = \lfloor \{\log_b X\}^{-1} \rfloor$, which follows (17).

b) $A_{[1]} = \lfloor b^{\{\log_b X\}} \rfloor$, which follows (6) —Benford's pmf.

For the reasons stated at the start of Section 4, the right answer is: neither. Examining either a) the first coefficient of the CF of $\{\log_b X\}$ or b) the most significant b-ary digit of X only has a psychological justification: pretty *discrete* theoretical distributions are available in both cases. But it will always be better to directly test the uniformity of $\{Y\} = \{\log_b X\}$, because information is lost when producing either A_1 or $A_{[1]}$ from $\{Y\}$. Similar considerations would apply if X were Pareto instead, i.e. using distributions (18) and (14) with $k = 1$ for A_1 and $A_{[1]}$, respectively.

In any case, let us try to settle the question at the start of this section just for the sake of argument. As we have talked about information loss, why not have a look at the mutual information between the continuous r.v. $\{Y\} = \{\log_b X\}$ and each of the discrete r.v.'s A_1 and $A_{[1]}$? The first amount is $I(\{Y\}; A_1) = H(A_1) - H(A_1|\{Y\})$, where $H(\cdot)$ stands for discrete entropy. Since A_1 is a function of $\{Y\}$ then $I(\{Y\}; A_1) = H(A_1)$, and for the same reason $I(\{Y\}; A_{[1]}) = H(A_{[1]})$. When $b = 10$, $H(A_1) = 2.046$ nats whilst $H(A_{[1]}) = 1.993$ nats, and so A_1 furnishes more information about $\{Y\}$ than $A_{[1]}$ does. Thus, when X is Benford, a forensic test based on the the first CF coefficient should be more accurate than one based on the first significant digit —not that we advocate it! This result can be overturned using more significant digits, but no discretisation approach can be better than relying on $\{\log_b X\}$.

It is interesting to observe that the comparison above is reminiscent of Lochs' theorem [14], which says that, as $k \to \infty$, we need slightly less than k CF coefficients of w_0 to determine its k most significant decimal digits. The main difference is that here we consider a nonasymptotic random setting with $k = 1$, and we look instead at the most significant decimal digit of x when $w_0 = \{\log_b x\}$.

Finally, one may also wish to model the CF coefficients A_j for $j \ge 2$, which may be done using conditioning. In any case, the Gauss-Kuz'min theorem [15] states that $\lim_{j\to\infty} \Pr(A_j = a) = \log_2\left(1 + (a(a+2))^{-1}\right)$ independently of W_0, which echoes the fact that $\lim_{j\to\infty} \Pr(A_{[j]} = a) = b^{-1}$ independently of X. Therefore, CF coefficients —just like significant digits— provide diminishing returns as j increases in terms of their forensic value.

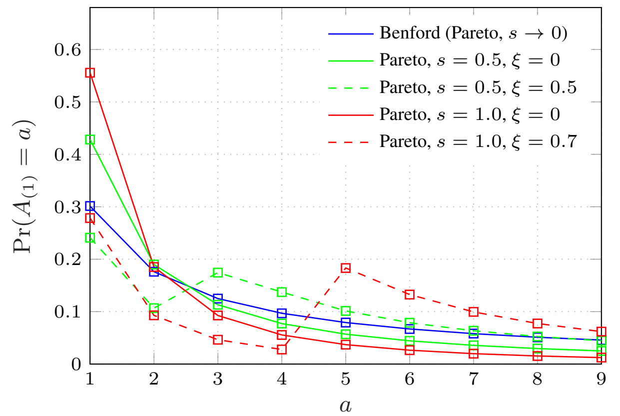

Fig. 1. Distributions of the most significant decimal digit. The theoretical pmf's are (6) and (13), and the empirical frequencies correspond to 10^7 pseudorandom outcomes.

5. EXPERIMENTS

In all figures in this section, solid or dashed lines connect theoretical probability masses, whereas symbols (sometimes connected by dots for clarity) show empirical frequencies. For validation purposes, Figs. 1 and 2 compare the distributions in Section 3 to empirical results from pseudorandom outcomes. Next, Figs. 3 and 4 show results for real datasets. Given a dataset $\{x_1, x_2, \dots, x_p\}$, where $x_i > 0$, we use the maximum likelihood (ML) estimators $\hat{x}_{\mathrm{m}} = \min_i x_i$ and $\hat{s} = \left(\frac{1}{p}\sum_i \ln(x_i/\hat{x}_{\mathrm{m}})\right)^{-1}$ to drive (13). As it can be seen, datasets that would be usually assumed to follow Benford's distribution (on the basis of scale invariance considerations) are much better modeled by (13). The DTP (14) is rarely enough for a good fit, i.e. assuming $\{\log_b x_{\mathrm{m}}\} = 0$ does not generally work. Many other datasets similarly support the validity of (13), but cannot be shown here due to lack of space: length of rivers, elevation of mountains, wealth of richest people, etc. Also, estimation approaches better than ML are possible when the data is not exactly Paretian and/or the sample size p is small [16]. This may be useful when applying (13) to datasets following heavy-tailed distributions which are also asymptotically scale invariant according to (10), in the sense that their tails follow the same power law as the Pareto distribution (such as the Fréchet, Burr, or Lévy distributions, among others).

To conclude, Fig. 5 empirically validates (17) and (18) using two datasets: one that follows Benford's distribution and a Paretian one. We have also plotted the asymptotic Gauss-Kuz'min law, which, intriguingly, turns out to be fairly close to (17).

6. REFERENCES

[1] F. Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572, 1938.

[2] S. Newcomb. Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, 4(1):39–40, 1881.

[3] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. U.S. Government Printing Office, New York, 1972.

[4] R. Pinkham. On the distribution of first significant digits. *Ann. Math. Statist.*, 32(4):1223–1230, December 1961.

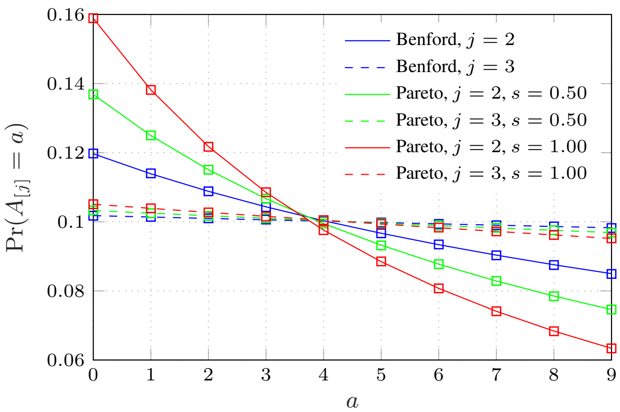

Fig. 2. Distributions of the j-th most significant decimal digit, for $j = 2, 3$. The theoretical pmf's are (8), (15) and (16), and the empirical frequencies correspond to 5×10^7 pseudorandom outcomes, with $\{\log_{10} x_{\mathrm{m}}\} = 0$ for the Pareto samples.

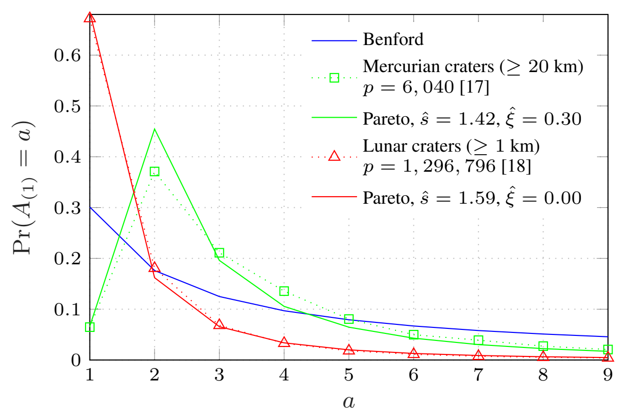

Fig. 3. Distributions of the most significant decimal digit for the diameter of impact craters (in km).

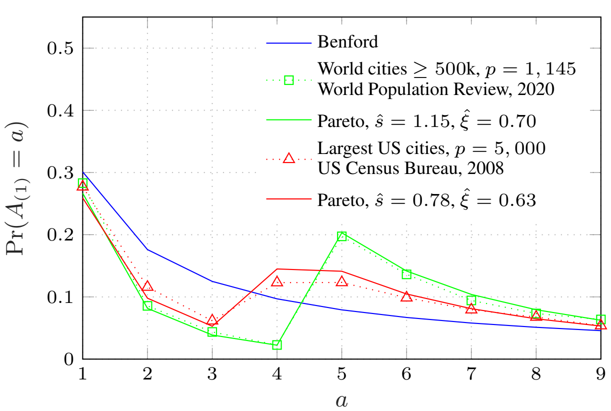

Fig. 4. Distributions of the most significant decimal digit for the population of cities.

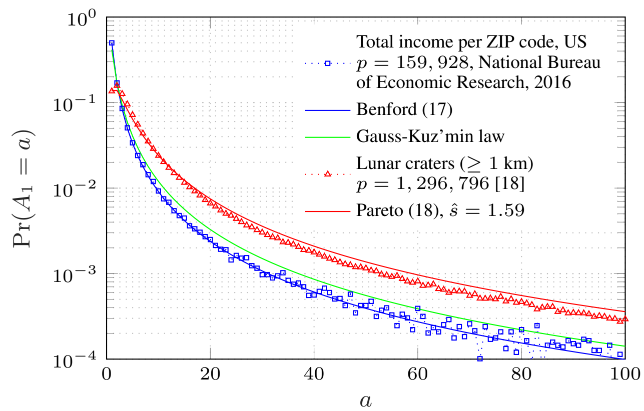

Fig. 5. Distributions of the first CF coefficient of $\{\log_{10} x_i\}$ for Benford and Pareto (with $\{\log_{10} \hat{x}_{\mathrm{m}}\} \approx 0$) datasets.

[5] C. Kleiber and S. Kotz. *Statistical Size Distributions in Economics and Actuarial Sciences*. John Wiley & Sons, 2003.

[6] A. Berger and T. Hill. Benford's law strikes back: No simple explanation in sight for mathematical gem. *Math Intelligencer*, 33:85–91, 2011.

[7] L. Leemis, B. Schmeiser, and D. Evans. Survival distributions satisfying Benford's law. *The American Statistician*, 54(4):236–241, 2000.

[8] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley & Sons, 2nd edition, 1965.

[9] J. Nair, A. Wierman, and B. Zwart. The fundamentals of heavy tails: Properties, emergence, and estimation. To appear, 2021.

[10] L. Barabesi and L. Pratelli. On the generalized Benford law. *Statistics & Probability Letters*, 160:108702, 2020.

[11] L. Pietronero, E. Tosatti, V. Tosatti, and A. Vespignani. Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf. *Physica A: Statistical Mechanics and its Applications*, 293(1):297–304, 2001.

[12] T.J. Kozubowski, A.K. Panorska, and M.L. Forister. A discrete truncated Pareto distribution. *Statistical Methodology*, 26:135–150, 2015.

[13] M.J. Nigrini. *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*. Wiley & Sons, 2012.

[14] G. Lochs. Vergleich der Genauigkeit von Dezimalbruch und Kettenbruch. *Abh. Math. Sem. Hamburg*, 27:142–144, 1964.

[15] A.Y. Khinchin. *Continued Fractions*. The University of Chicago Press, 3rd edition, 1961.

[16] J.H.T. Kim, S. Ahn, and S. Ahn. Parameter estimation of the Pareto distribution using a pivotal quantity. *Journal of the Korean Statistical Society*, 46:438–450, 2017.

[17] C.I. Fassett, S.J. Kadish, J.W. Head, S.C. Solomon, and R.G. Strom. The global population of large craters on Mercury and comparison with the Moon. *Geophys. Res. Lett.*, 28, 2011.

[18] S.J. Robbins. A new global database of Lunar impact craters $>$ 1–2 km: 1. Crater locations and sizes, comparisons with published databases, and global analysis. *Journal of Geophysical Research: Planets*, 124(4):871–892, 2018.