The Inverse Gamma Distribution and Benford's Law

R.F. Durst, C. Huynh, A. Lott, S.J. Miller,
E.A. Palsson, W. Touw, and G. Vriend*

Abstract - According to Benford's Law, many data sets have a bias towards lower leading digits (about 30% are 1's). The applications of Benford's Law vary: from detecting tax, voter and image fraud to determining the possibility of match-fixing in competitive sports. There are many common distributions that exhibit such bias, i.e. they are almost Benford. These include the exponential and the Weibull distributions. Motivated by these examples and the fact that the underlying distribution of factors in protein structure follows an inverse gamma distribution, we determine the closeness of this distribution to a Benford distribution as its parameters change.

Keywords : Benford's law; inverse gamma distribution; digit bias; Poisson summation

Mathematics Subject Classification (2020) **:** 60F05; 11K06; 60E10; 42A16; 62E15; 62P99

1 Introduction

1.1 Motivation

For a positive integer $B \geq 2$, any positive number x can be written uniquely in base B as $x = S_B(x) \cdot B^{k(x)}$ where $k(x)$ is an integer and $S_B(x) \in [1, B)$ is called the *significand* of x base B. Benford's Law describes the distribution of significands in many naturally occurring data sets and states that for any $1 \leq s < B$, the proportion of the set with significand at most s is $\log_B(s)$. In this paper, we examine the behavior of random variables, so we adopt the following definition.

Definition 1.1 (Benford's Law) *Let X be a random variable taking values in $(0, \infty)$ almost surely. We say that X follows Benford's Law in base B if, for any $s \in [1, B)$,*

$$\text{Prob}\left(S_B(X) \leq s\right) \;=\; \log_B(s). \tag{1}$$

In particular,

$$\text{Prob}\left(\textit{first digit of } X \textit{ is } d\right) \;=\; \log_B\left(\frac{d+1}{d}\right). \tag{2}$$

*This work was supported by NSF Grants DMS1265673, DMS1561945, and DMS1347804, Simons Foundation Grant #360560, Williams College, and the Williams Finnerty Fund.

Thus in base 10 about 30% of numbers have a leading digit of 1, as compared to only about 4.6% starting with a 9. For an introduction to the theory, as well as a detailed discussion of some of its applications in accounting, biology, economics, engineering, game theory, finance, mathematics, physics, psychology, statistics and voting see [6].

One of the most important applications of Benford's law is in fraud detection; it has successfully flagged voting irregularities, tax fraud, and embezzlement, to name just a few of its successes.[1] The motivation for this work was to see if a Benford analysis could have detected some fraud on protein structures, as well as serve as a protection against future unscrupulous researchers.

Proteins are the workhorses in all of biology; in plant, human, animal, bacterium, and slime mold, alike. They keep us together, digest our food, make us see, hear, taste, feel, and think, they defend us against pathogens, and they are the target of most existing medicines. Knowledge about the three-dimensional structure of proteins is a prerequisite for research in fields as diverse as drug design, bio-fuel engineering, food processing, or increasing the yield in agriculture.

These three-dimensional structures can be solved with X-ray crystallography, Nuclear Magnetic Resonance, or electron microscopy. Today, most structures are solved with X-ray crystallography. When structures are solved with this technique the experimentalist does not only obtain X, Y and Z coordinates for the atoms, but also a measure of their mobility, which is called the B factor.

After it was detected that 12 of the 14 structures deposited in the PDB protein data bank [1] by H. K. M. Murthy were not based on experimental data (see https://www.uab.edu/reporterarchive/71570-uab-statement-on-protein-data-bank-issues), two of the authors asked the question if their rather anomalous B-factor distributions could have been used to automatically detect the problems (see swift.cmbi.ru.nl/gv/Murthy/Murthy_4.html). In practice B-factor distributions are influenced by experiment conditions and human choices. For example, B factors may fit inverse Gamma distributions translated towards higher values [3, 7], or the inverse Gamma fit might be worse when upper and/or lower B factor limits are enforced by the experimentalist. The reported properties of each of the 14 structures were used to find in the PDB a legitimate protein structure of comparable experimental quality, deposition date, size, and B factor profile. In general, inverse Gamma parameters could be estimated well for both the Murthy structures and the legitimate structures by maximum likelihood estimation when accounting for the translation along the x-axis. This suggests the main question of this paper: how close is the inverse Gamma distribution, for various choices of its parameters, to Benford's law? While unfortunately a Benford analysis did not flag Murthy's structures from legitimate ones, the question of how close this special distribution is to Benford is still of independent interest, and we report on our findings below. This paper is a sequel to [2], where a similar analysis was done for the three parameter Weibull.

[1] See interestingly Section 6 of [5] for comments on rounding in Benford's original paper.

1.2 Results

In practice, it is easier to use the following equivalent condition for Benford behavior (see, for example, [4] or [6]), which we reprove here.

Definition 1.2 *We say that a random variable Y taking values in $[0,1]$ is equidistributed if, for any $[a,b] \subseteq [0,1]$,*

$$\text{Prob}\,(Y \in [a,b]) \;=\; b-a. \tag{3}$$

Theorem 1.3 *A random variable X follows Benford's Law in base B if and only if the random variable $Y := \log_B X \mod 1$ is equidistributed.*

Proof. We only prove the reverse direction here as that is all we need to prove our main result. Full details are given in [4]. Suppose $Y := \log_B X \mod 1$ is equidistributed. First note that

$$\begin{aligned} Y &= \log_B(X) \mod 1 \\ &= \log_B(S_B(X)\cdot B^{k(X)}) \mod 1 \\ &= \log_B(S_B(X)) + \log_B(B^{k(X)}) \mod 1 \\ &= \log_B(S_B(X)). \end{aligned} \tag{4}$$

Then, taking $a=0$, $b=\log_B(p)$ in the definition of equidistribution, we get

$$\text{Prob}\,(\log_B(S_B(X)) \in [0,\log_B(p)]) \;=\; \log_B(p). \tag{5}$$

Exponentiating gives

$$\text{Prob}\,(S_B(X) \in [1,p]) \;=\; \log_B(p), \tag{6}$$

which is exactly the statement of Benford's Law. □

In this paper, we examine the behavior of a random variable drawn from the inverse gamma distribution. For fixed parameters α, $\beta > 0$, this distribution has density defined by

$$f(x;\alpha,\beta) \;=\; \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(\frac{-\beta}{x}\right) \tag{7}$$

and cumulative distribution function

$$F(x;\alpha,\beta) \;=\; \frac{1}{\Gamma(\alpha)} \int_{\beta/x}^{\infty} t^{\alpha-1} e^{-t}\, dt \tag{8}$$

Let $X_{\alpha,\beta}$ be a random variable distributed according to (7) and let F_B be the cumulative distribution function of $\log_B(X_{\alpha,\beta}) \mod 1$. By Theorem 1.3, the assertion that $X_{\alpha,\beta}$ follows Benford's Law is equivalent to saying that $F_B(z) = z$ for all $z \in [0,1]$. In this paper, we investigate when the deviations of $F_B(z)$ from z are small, i.e., when $X_{\alpha,\beta}$ approximately follows Benford's Law. We do this by deriving a series expansion for $F_B'(z)$ of the form $1 +$ (error term), where the error term can be computed to great accuracy, and then integrating in order to return to the cumulative distribution function, $F_B(z)$.

In Section 2, we derive our series representation for $F'_B(z)$. In Section 3, we give bounds for the tail of the series, showing that the series can be computed to great accuracy by computing only the first few terms. This result is built upon in Appendix A. In Section 4, we use this result to generate some plots illustrating the Benfordness of the inverse gamma distribution as a function of α and β.

2 Series representation for $F'_B(z)$

Before beginning the analysis, we first note a useful invariant property of the Benfordness of this distribution.

Lemma 2.1 *For any $\alpha, \beta > 0$ and $z \in [0, 1]$,*

$$\text{Prob}\left(\log_B S_B(X_{\alpha,\beta}) \leq z\right) \;=\; \text{Prob}\left(\log_B S_B(X_{\alpha,B\cdot\beta}) \leq z\right). \tag{9}$$

In other words, the deviation from Benford's law of the inverse Gamma distribution doesn't change if we scale β by a factor of B.

Proof. Scaling β by a factor of B yields

$$\begin{aligned}\text{Prob}\left(\log_B S_B(X_{\alpha,B\cdot\beta}) \leq z\right) &= \sum_{k=-\infty}^{\infty} \text{Prob}\left(\log_B X_{\alpha,B\cdot\beta} \in [k, z+k]\right) \\ &= \sum_{k=-\infty}^{\infty} \text{Prob}\left(X_{\alpha,B\cdot\beta} \in [B^k, B^{z+k}]\right), \end{aligned} \tag{10}$$

which, by (8), is

$$\begin{aligned} &= \frac{1}{\Gamma(\alpha)} \sum_{k=-\infty}^{\infty} \left(\int_{B\cdot\beta/B^{z+k}}^{\infty} t^{\alpha-1}e^{-t}dt - \int_{B\cdot\beta/B^{k}}^{\infty} t^{\alpha-1}e^{-t}dt \right) \\ &= \frac{1}{\Gamma(\alpha)} \sum_{k=-\infty}^{\infty} \int_{B\cdot\beta/B^{z+k}}^{B\cdot\beta/B^{k}} t^{\alpha-1}e^{-t}dt \\ &= \frac{1}{\Gamma(\alpha)} \sum_{k=-\infty}^{\infty} \int_{\beta/B^{z+k-1}}^{\beta/B^{k-1}} t^{\alpha-1}e^{-t}dt \\ &= \text{Prob}\left(S_B(X_{\alpha,\beta}) \leq z\right). \end{aligned} \tag{11}$$

Thus, scaling β by a power of B only results in shifting k. Since we take an infinite sum over k, this shift does not change the final value of the probability. As a consequence of this, it is clear that scaling β by any power of B will yield the same result, shifting k by that power. □

Thus it suffices to study $1 \leq \beta < B$.

To show that the deviations of $F_B(z)$ from z are small, it is easier in practice to show that $F'_B(z)$ is close to 1, and then integrate. We derive a series representation for $F'_B(z)$, but first, we state a useful property of Fourier transforms (see, for example, [8]).

Throughout the course of this paper, we define the Fourier transform as follows.

Definition 2.2 (Fourier Transform) *Let $f \in L^1(\mathbb{R})$. Define the Fourier transform $\hat{f}$ of f by*

$$\hat{f}(\xi) \ := \ \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi} dx. \tag{12}$$

Furthermore, we will occasionally use the notation

$$\mathcal{F}(f(x))(\xi) \ := \ \hat{f}(\xi). \tag{13}$$

Our main tool is the Poisson summation formula, which we state here in a weak form (see Theorem 3.1 of [2] for a more detailed explanation).

Theorem 2.3 (Poisson Summation) *Let f be a function such that f, f', and f'' are all $O(x^{-(1+\eta)})$ as $x \to \infty$ for some $\eta > 0$. Then*

$$\sum_{k=-\infty}^{\infty} f(k) \ = \ \sum_{k=-\infty}^{\infty} \hat{f}(k). \tag{14}$$

Theorem 2.4 *Let α, $\beta > 0$ be fixed and let $B \geq 2$ be an integer. Let $X_{\alpha,\beta}$ be a random variable distributed according to equation (7). For $z \in [0,1]$, let $F_B(z)$ be the cumulative distribution function of $\log_B(X_{\alpha,\beta}) \mod 1$. Then $F_B'(z)$ is given by*

$$F_B'(z) \ = \ 1 + \frac{2}{\Gamma(\alpha)} \sum_{k=1}^{\infty} \Re \left(e^{2\pi i k (\log_B \beta - z)} \Gamma \left(\alpha - \frac{2\pi i k}{\log B} \right) \right). \tag{15}$$

Proof. By the argument leading to (11),

$$F_B(z) \ = \ \frac{1}{\Gamma(\alpha)} \sum_{k=-\infty}^{\infty} \int_{\frac{\beta}{B^{z+k}}}^{\frac{\beta}{B^k}} t^{\alpha-1} e^{-t} dt. \tag{16}$$

Taking the derivative yields

$$F_B'(z) \ = \ \frac{1}{\Gamma(\alpha)} \sum_{k=-\infty}^{\infty} \left(\frac{\beta}{B^{z+k}} \right)^{\alpha} \exp\left(\frac{-\beta}{B^{z+k}} \right) \ln B. \tag{17}$$

Applying Poisson summation to (17) gives

$$F_B'(z) \ = \ \frac{1}{\Gamma(\alpha)} \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\frac{\beta}{B^{z+t}} \right)^{\alpha} \exp\left(\frac{-\beta}{B^{z+t}} \right) \log B \ \exp(-2\pi i t k) \ dt. \tag{18}$$

We now let $x = \frac{\beta}{B^{z+t}}$ and $dx = \frac{-\beta}{B^{z+t}} \log B \; dt$ so that we have

$$\begin{aligned} F'_B(z) &= \frac{1}{\Gamma(\alpha)} \sum_{k=-\infty}^{\infty} \int_0^{\infty} x^{\alpha-1} \exp\left(-2\pi i k \left(\frac{\log \frac{\beta}{B^z x}}{\log B}\right)\right) e^{-x} dx \\ &= \frac{1}{\Gamma(\alpha)} \sum_{k=-\infty}^{\infty} \int_0^{\infty} x^{\alpha-1} \left(\frac{\beta}{B^z x}\right)^{\frac{-2\pi i k}{\log B}} e^{-x} dx \\ &= \frac{1}{\Gamma(\alpha)} \sum_{k=-\infty}^{\infty} \left(\frac{\beta}{B^z}\right)^{\frac{-2\pi i k}{\log B}} \int_0^{\infty} x^{\alpha-1+\frac{2\pi i k}{\log B}} e^{-x} dx \\ &= \frac{1}{\Gamma(\alpha)} \sum_{k=-\infty}^{\infty} \left(\frac{\beta}{B^z}\right)^{\frac{-2\pi i k}{\log B}} \Gamma\left(\alpha + \frac{2\pi i k}{\log B}\right). \end{aligned} \tag{19}$$

Note that $\left(\frac{\beta}{B^z}\right)^{2\pi i \theta} = \exp\left(2\pi i \theta \log \frac{\beta}{B^z}\right)$, so our sum becomes

$$F'_B(z) = \frac{1}{\Gamma(\alpha)} \sum_{k=-\infty}^{\infty} \exp\left(\frac{-2\pi i k \log \frac{\beta}{B^z}}{\log B}\right) \Gamma\left(\alpha + \frac{2\pi i k}{\log B}\right). \tag{20}$$

This form of our sum will become useful in a later proof, but for the purposes of this theorem, we further simplify our derivative and point out that the $k = 0$ term in (20) is equal to 1. Thus our equation becomes

$$\begin{aligned} F'_B(z) &= 1 + \frac{1}{\Gamma(\alpha)} \left[\sum_{k=1}^{\infty} \exp\left(\frac{2\pi i k \log \frac{\beta}{B^z}}{\log B}\right) \Gamma\left(\alpha - \frac{2\pi i k}{\log B}\right) \right. \\ &\qquad \left. + \sum_{k=1}^{\infty} \exp\left(\frac{-2\pi i k \log \frac{\beta}{B^z}}{\log B}\right) \Gamma\left(\alpha + \frac{2\pi i k}{\log B}\right) \right] \\ &= 1 + \frac{1}{\Gamma(\alpha)} \left[\sum_{k=1}^{\infty} \exp\left(2\pi i k (\log_B \beta - z)\right) \Gamma\left(\alpha - \frac{2\pi i k}{\log B}\right) \right. \\ &\qquad \left. + \exp\left(-2\pi i k (\log_B \beta - z)\right) \Gamma\left(\alpha + \frac{2\pi i k}{\log B}\right) \right]. \end{aligned} \tag{21}$$

Finally, using the identity that $\overline{\Gamma(a+ib)} = \Gamma(a-ib)$ for real numbers a and b, we have

$$F'_B(z) = 1 + \frac{2}{\Gamma(\alpha)} \sum_{k=1}^{\infty} \Re\left(e^{2\pi i k (\log_B \beta - z)} \Gamma\left(\alpha - \frac{2\pi i k}{\log B}\right)\right). \tag{22}$$

□

3 Bounding the truncation error

A key tool for the analysis in [2] is the identity

$$|\Gamma(1+ix)|^2 \;=\; \frac{\pi x}{\sinh(\pi x)} \tag{23}$$

for real x. Examining (22), it is clear that when $\alpha = 1$, our analysis of the truncation error is similar to that of [2]. Since the bound resulting from such analysis in the case of $\alpha = 1$ is tighter than the bound for an arbitrary α, we have included the proof in the appendix. However, when $\alpha \neq 1$, the identity (23) is no longer applicable, so a new approach is needed to bound the tails of the series expansion. We have the following bound on the truncation error.

Theorem 3.1 *Let $F'_B(z)$ be as in (20). Let $E_M(z)$ denote the two-sided tail of the series expansion, i.e.,*

$$E_M(z) \;:=\; \sum_{|k|\geq M} \exp\left(\frac{-2\pi ik \log\frac{\beta}{B^z}}{\log B}\right)\Gamma\left(\alpha + \frac{2\pi ik}{\log B}\right). \tag{24}$$

1. *We have*

$$|E_M(z)| \;\leq\; \frac{e^{\frac{\beta}{B^z}}\left(\frac{\beta}{B^z}\right)^{\alpha}}{\Gamma(\alpha)}\left(\frac{1}{\alpha}B^{-M\alpha} + \int_{B^M}^{\infty} e^{-x}x^{\alpha-1}dx\right). \tag{25}$$

2. *This is bounded uniformly on $z \in [0,1]$ by the constant*

$$|E_M(z)| \;\leq\; \frac{C(\alpha,\beta,B)}{\Gamma(\alpha)}\left(\frac{1}{\alpha}B^{-M\alpha} + \int_{B^M}^{\infty} e^{-x}x^{\alpha-1}dx\right) \tag{26}$$

   where $C(\alpha,\beta,B) = \max\left(e^{\beta}\beta^{\alpha},\; e^{\alpha}\alpha^{\alpha},\; e^{\frac{\beta}{B}}\left(\frac{\beta}{B}\right)^{\alpha}\right)$.

3. *Furthermore, for any $\epsilon > 0$, in order to have $|E_M(z)| < \epsilon$ in (26) it suffices to take*

$$M \;>\; \max\left(\alpha+1,\; -\log_B\left(\frac{\epsilon\cdot\Gamma(\alpha)}{2C(\alpha,\beta,B)}\right)\right) \tag{27}$$

   where $C(\alpha,\beta,B)$ is as above.

Proof.
Proof of part (1): locally bounding the truncation error

We begin with (20).

Let $\phi(z) = \log\frac{\beta}{B^z}$. We have

$$E(z) \;:=\; F'_B(z) - 1 \;=\; \frac{1}{\Gamma(\alpha)}\sum_{|k|\geq 1}\exp\left(-2\pi\frac{ik\phi(z)}{\log B}\right)\Gamma\left(\alpha + 2\pi\frac{ik}{\log B}\right). \tag{28}$$

Furthermore, given $\Gamma(a+2\pi ib) = \int_0^\infty e^{-x}x^{a+2\pi ib-1}dx$, we may perform a change of variables and let $x = e^{-u}$ so that we get

$$\Gamma(a+2\pi bi) \;=\; \int_{-\infty}^{\infty} e^{-e^{-u}}e^{-au}e^{-2\pi ibu}du \;=\; \mathcal{F}\left(e^{-e^{-u}}e^{-au}\right)(b), \tag{29}$$

where $\mathcal{F}(\cdot)$ denotes the Fourier transform, as stated in (13). This transforms our sum into the sum of terms of the form

$$\begin{aligned}&\exp\left(\frac{-2\pi ik\phi(z)}{\log B}\right)\Gamma\left(\alpha+2\pi\frac{ik}{\log B}\right)\\ &=\; \exp\left(\frac{-2\pi ik\phi(z)}{\log B}\right)\left[\mathcal{F}\left(e^{-e^{-u}}e^{-\alpha u}\right)\left(\frac{k}{\log B}\right)\right].\end{aligned} \tag{30}$$

Suppose $s \in L^1(\mathbb{R})$, $P > 0$, and $t \in \mathbb{R}$. Define

$$g(x) \;\equiv\; s(Px+t). \tag{31}$$

The scaling and frequency shift properties of Fourier transforms then yield

$$\hat{g}(\xi) \;=\; \frac{1}{P}\exp\frac{2\pi ikt}{P}\hat{s}\left(\frac{\xi}{P}\right). \tag{32}$$

Thus, if g meets the conditions required for Poisson summation, we have

$$P\sum_{n\in\mathbb{Z}} s(t+nP) \;=\; \sum_{k\in\mathbb{Z}}\exp\frac{2\pi ikt}{P}\mathcal{F}(s)\left(\frac{k}{P}\right). \tag{33}$$

Therefore, letting $s = e^{-e^{-u}}e^{-\alpha u}$, $P = \log B$, and $t = -\phi(z)$, we have

$$\begin{aligned}E(z) \;&=\; \sum_{|k|\geq 1}\mathcal{F}(s)\left(\frac{k}{P}\right)e^{2\pi i\frac{k}{P}t}\frac{1}{P}\\ &=\; \left(\sum_{k\in\mathbb{Z}}\mathcal{F}(s)\left(\frac{k}{P}\right)e^{2\pi i\frac{k}{P}t}\frac{1}{P}\right) \;-\; 1\\ &\leq P\sum_{k\in\mathbb{Z}} s(t+kP)\\ &=\; \frac{\log B\left(e^{-e^{\phi(z)}}e^{\alpha\phi(z)}\right)}{\Gamma(\alpha)}\sum_{k\in\mathbb{Z}} e^{-e^{-k\log B}}e^{-\alpha k\log B}.\end{aligned} \tag{34}$$

We now concentrate on the truncation error $E_M(z)$. We bound our sums by integrals and perform a change of variables, letting $x = e^{-k\log B}$:

$$E_M(z) \;\leq\; \frac{\log B\left(e^{-e^{\phi(z)}}e^{\alpha\phi(z)}\right)}{\Gamma(\alpha)} \sum_{|k|\leq M} e^{-e^{-k\log B}}e^{-\alpha k\log B}. \tag{35}$$

This may then be extended to give

$$\begin{aligned}
|E_M(z)| \;&\leq\; \frac{\left(e^{-e^{\phi(z)}}e^{\alpha\phi(z)}\right)}{\Gamma(\alpha)}\left(\int_{B^M}^{\infty} e^{-x}x^{\alpha-1}dx + \int_0^{B^{-M}} e^{-x}x^{\alpha-1}dx\right)\\
&\leq\; \frac{\left(e^{-e^{\phi(z)}}e^{\alpha\phi(z)}\right)}{\Gamma(\alpha)}\left(\int_{B^M}^{\infty} e^{-x}x^{\alpha-1}dx + \int_0^{B^{-M}} x^{\alpha-1}dx\right)\\
&\leq\; \frac{e^{-\beta/B^z}\left(\frac{\beta}{B^z}\right)^{\alpha}}{\Gamma(\alpha)}\left(\int_{B^M}^{\infty} e^{-x}x^{\alpha-1}dx + \frac{1}{\alpha}B^{-M\alpha}\right),
\end{aligned} \tag{36}$$

which is (25), thus proving (1).

Proof of part (2): uniformly bounding the truncation error for $z \in [0,1]$.To get (26), we simply maximize (25) with respect to z. Set

$$g(z) \;=\; e^{-\beta/B^z}\left(\frac{\beta}{B^z}\right)^{\alpha}, \tag{37}$$

set the derivative equal to 0 to get

$$g'(z) \;=\; e^{-\beta/B^z}(\beta/B^z)^{\alpha}\log B\left(\frac{\beta}{B^z}-\alpha\right) \;=\; 0, \tag{38}$$

and solve to get $z = \log_B\left(\frac{\beta}{\alpha}\right)$. Also note that $g'(z)$ is decreasing, so $g(z)$ has exactly one maximum at $z = \log_B\left(\frac{\beta}{\alpha}\right)$. Recalling that we only consider $|E_M(z)|$ on $z \in [0,1]$, we conclude that if $\log_B\left(\frac{\beta}{\alpha}\right) \leq 0$, $|E_M(z)|$ is maximized at $z = 0$, if $\log_B\left(\frac{\beta}{\alpha}\right) \in (0,1)$, $|E_M(z)|$ is maximized at $z = \log_B\left(\frac{\beta}{\alpha}\right)$, and if $\log_B\left(\frac{\beta}{\alpha}\right) \geq 1$, then $|E_M(z)|$ is maximized at $z = 1$. Calculating the value of (25) at these three points and letting $C(\alpha,\beta,B)$ be their maximum yields (26), so part (2) is proven.

Proof of part (3). Fix an $\epsilon > 0$ and suppose

$$M \;>\; \max\left(\alpha+1,\; -\log_B\left(\frac{\epsilon\cdot\Gamma(\alpha)}{2C(\alpha,\beta,B)}\right)\right). \tag{39}$$

In particular, this implies that $B^M > e^{\alpha+1}$, so for all $x \geq B^M$, $x/\log x > \alpha+1$, which implies that

$$e^{-x}x^{\alpha-1} \;\leq\; 1/x^2. \tag{40}$$

Equation (39) also implies that

$$\frac{1}{\alpha}B^{-M\alpha} + B^{-M} \;<\; 2B^{-M} \;<\; \frac{\epsilon \cdot \Gamma(\alpha)}{C(\alpha,\beta,B)}. \tag{41}$$

Combining (40) and (41) with (26), we have the bound

$$\begin{aligned} |E_M(z)| \;&<\; \frac{C(\alpha,\beta,B)}{\Gamma(\alpha)}\left(\frac{1}{\alpha}B^{-M\alpha} + \int_{B^M}^{\infty}\frac{1}{x^2}dx\right) \\ &<\; \frac{C(\alpha,\beta,B)}{\Gamma(\alpha)}\left(\frac{1}{\alpha}B^{-M\alpha} + B^{-M}\right) \\ &<\; \frac{C(\alpha,\beta,B)}{\Gamma(\alpha)}\frac{\epsilon \cdot \Gamma(\alpha)}{C(\alpha,\beta,B)} \;=\; \epsilon. \end{aligned} \tag{42}$$

□

4 Plots and analysis

Using Theorem 3.1 allows us to easily compare $F_B(z)$, the CDF of $\log X_{\alpha,\beta}$, with z, the Benford CDF. We simply integrate (22) from 0 to z, yielding

$$F_B(z) \;=\; z + \frac{1}{\Gamma(\alpha)}\sum_{|k|\geq 1}\Gamma\left(\alpha + \frac{2\pi ik}{\log B}\right)\frac{1}{2\pi ik}e^{-2\pi ik\log_B(\beta)}\left(e^{2\pi ikz} - 1\right). \tag{43}$$

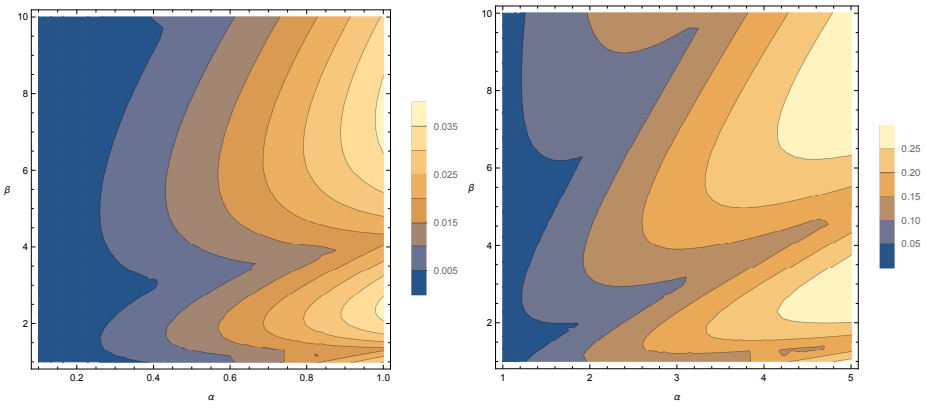
We now use Theorem 3.1 in the following way. Fix an $\epsilon > 0$. Then part (3) of Theorem 3.1 allows us to quickly compute the value of $|F_B'(z) - 1|$ to within ϵ of the true value. Thus, after integrating, since we are only working on $z \in [0,1]$, the mean value theorem guarantees that we now know $|F_B(z) - z|$ to within ϵ of the true value. In short, Theorem 3.1 allows us to obtain very good estimates for $|F_B(z) - z|$ by taking only the first few terms, which makes calculating the deviation more computationally feasible. To measure the closeness to Benford of the distribution, we use the quantity

$$\max_{z\in[0,1]} |F_B(z) - z|. \tag{44}$$

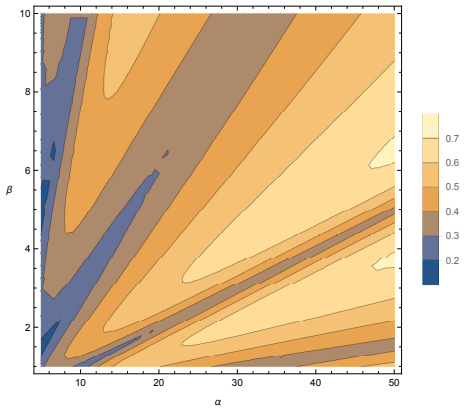
In Figure 1, we illustrate this quantity as a function of α and β with $B = 10$ fixed. The emergent trend is that as α increases, the distribution gets farther away from Benford, and the Benfordness is largely independent of β. This behavior is similar to that of the Weibull distribution exhibited in [2].

A Bounding the truncation error in the special case $\alpha = 1$

As mentioned above, when $\alpha = 1$ it is possible for us to achieve better bounds on the truncation error using methods similar to those in [2].



(a) $0.1 \leq \alpha \leq 1$ (b) $1 \leq \alpha \leq 5$



(c) $5 \leq \alpha \leq 50$

Figure 1: Contour plots of the quantity $\max_{z\in[0,1]} |F_B(z) - z|$ (see (43)) as a function of α and β with $B = 10$ fixed. Using part (3) of Theorem 3.1, we have made the displayed values accurate to within $\epsilon = 0.001$. Notice that the error is large for large α, meaning that the inverse gamma distribution only approximates Benford behavior for small α. Also notice that β has less of an effect on the error.

Theorem A.1 *Let $F'_B(z)$ be as in Theorem 2.4 with $\alpha = 1$.*

1. *For $M \geq \frac{\log 2 \log B}{4\pi^2}$, the contribution to $F'_B(z)$ from the tail of the expansion (from the terms with $k \geq M$ in (22)) is at most*

$$\frac{4(\pi^2 + \log B)}{\pi\sqrt{\log B}} M \exp\left(\frac{-\pi^2 M}{\log B}\right). \tag{45}$$

2. *For an error of at most ϵ from ignoring the terms with $k \geq M$ in (22), it suffices to take*

$$M = \frac{h + \log h + 1/2}{a} \tag{46}$$

where $a = \frac{\pi^2}{\log B}$, $h = \max\left(6, -\log \frac{a\epsilon}{C}\right)$, and $C = \frac{4(\pi^2+\log B)}{\pi \log B}$.

Proof.

1. As stated, we estimate the contribution to $F'_B(z)$ from the tail when $\alpha = 1$. Let

$$E_M(z) \;:=\; \frac{2}{\Gamma(1)} \sum_{k=M}^{\infty} \Re\left(e^{2\pi i k(\log_B \beta - z)} \Gamma\left(1 + \frac{-2\pi i k}{\log B}\right)\right) \tag{47}$$

where $\Gamma(1+iu) = \int_0^\infty e^{-x} x^{iu} dx$ with $u = \frac{-2\pi i k}{\log B}$ in our case. We note that as u increases, there is more oscillation, which means the integral would achieve a smaller value when u increases. Since $|e^{i\theta}| = 1$, when we take the absolute values inside the sum we get $|e^{2\pi i k(\log_B \beta - z)}| = 1$. Thus it is safe to ignore this term in computing the upper bound.

Using the fact that $|\Gamma(1+ix)|^2 = \frac{\pi x}{\sinh(\pi x)}$, we have from (47):

$$\begin{aligned}
|E_M(z)| \;&\leq\; \frac{2}{\Gamma(1)} \sum_{k=M}^{\infty} \left|e^{2\pi i k(\log_B \beta - z)}\right| \left|\Gamma\left(1 + \frac{-2\pi i k}{\log B}\right)\right| \\
&\leq\; \frac{2\sqrt{2}\pi}{\sqrt{\log B}} \sum_{k=M}^{\infty} \sqrt{\frac{k}{\sinh\left(\frac{2\pi^2 k}{\log B}\right)}} \\
&=\; \frac{2\sqrt{2}\pi}{\sqrt{\log B}} \sum_{k=M}^{\infty} \sqrt{\frac{2k^2}{\exp\left(\frac{2\pi^2 k}{\log B}\right) - \exp\left(\frac{-2\pi^2 k}{\log B}\right)}} \\
&\leq\; \frac{4\pi}{\sqrt{\log B}} \sum_{k=M}^{\infty} \sqrt{k^2 / \exp\left(\frac{2\pi^2 k}{\log B}\right)}.
\end{aligned} \tag{48}$$

Here we have overestimated the error by disregarding the difference in the denominator, which is very small when k is big. Let $u = \exp\left(\frac{2\pi^2 k}{\log B}\right)$. For $\frac{1}{u - 1/u} < \frac{2}{u}$, we must get $u \geq \sqrt{2}$, which means $\exp\left(\frac{2\pi^2 k}{\log B}\right) \geq \sqrt{2}$. Solving this gives us $k \geq \frac{\log 2 \log B}{4\pi^2}$, which will help us simplify the denominator as we can assume M exceeds this value and $k \geq M$. We can now substitute this bound into (48) to simplify further:

$$\begin{aligned}
|E_M(z)| \;&\leq\; \frac{4\pi}{\sqrt{\log B}} \sum_{k=M}^{\infty} \frac{\sqrt{2}k}{\exp\left(\frac{\pi^2 k}{\log B}\right)} \\
&\leq\; \frac{4\pi}{\sqrt{\log B}} \int_M^\infty m \exp\left(\frac{-\pi^2 m}{\log B}\right) dm.
\end{aligned} \tag{49}$$

We let $a = \frac{\pi^2}{\log B}$ and apply integration by parts to get

$$\begin{aligned} |E_M(z)| &\leq \frac{4\pi}{\sqrt{\log B}} \frac{1}{a^2} \left(aMe^{-aM} + e^{-aM} \right) \\ &\leq \frac{4\pi}{\sqrt{\log B}} \frac{a+1}{a} Me^{-aM} \\ &= \frac{4\pi(a+1)}{a\sqrt{\log B}} Me^{-aM}, \end{aligned} \tag{50}$$

which simplifies to

$$|E_M(z)| \leq \frac{4(\pi^2 + \log B)}{\pi\sqrt{\log B}} M \exp\left(\frac{-\pi^2 M}{\log B} \right), \tag{51}$$

proving part (1).

2. Let $C = \frac{4(\pi^2+\log B)}{\pi \log B}$ and $a = \frac{\pi^2}{\log B}$ as before. We want

$$CMe^{-aM} \leq \epsilon. \tag{52}$$

We will do this by iteratively expanding to improve the bounds. Let $v = aM$, then

$$\frac{C}{a} ve^{-v} \leq \epsilon \iff ve^{-v} \leq \frac{a\epsilon}{C}. \tag{53}$$

We carry out a change of variables one more time, letting $h = -\log \frac{a\epsilon}{C}$ and expanding v as $v = h + x$. This leads to

$$\begin{aligned} ve^{-v} &\leq e^{-h} \\ \longleftrightarrow \frac{h+x}{e^x} &\leq 1. \end{aligned} \tag{54}$$

Now we note that by expanding v in this way, solving for x is equivalent to solving for v , which is equivalent to solving for M. We guess $x = \log h + \frac{1}{2}$ then the left-hand-side of 54 becomes:

$$\frac{h + \log h + 1/2}{he^{1/2}} \leq 1 \leftrightarrow h + \log h + 1/2 \leq he^{1/2}. \tag{55}$$

Now what we want to do is to determine the value of h so that $\log h \leq h/2$ since this ensures the inequality above would hold. The aforementioned inequality gives $h \leq e^{h/2}$ or $h^2 \leq e^h$. Since for h positive, $e^h \geq \frac{h^3}{3!}$, it is sufficient to choose h such that $h^2 \leq h^3/6$ or $h \geq 6$. For $h \geq 6$,

$$h + \log h + \frac{1}{2} \leq h + \frac{h}{12} + \frac{h}{2} = \frac{19h}{12} \approx 1.5883h. \tag{56}$$

As $he^{1/2} \approx 1.64872h$, a sufficient cutoff for M in terms of h for an error of at most ϵ is

$$M = \frac{h + \log h + 1/2}{a} \tag{57}$$

with $a = \frac{\pi^2}{\log B}$, $h = \max\left(6, -\log \frac{a\epsilon}{C}\right)$.

□

Acknowledgments

We thank Peter Vijn for suggesting using Benford's law to study protein database submissions.

References

[1] H.M. Berman, K. Henrick, H. Nakamura, Announcing the worldwide Protein Data Bank, *Nat Struct Biol* **10** (2003), doi: 10.1038/nsb1203-980.

[2] V. Cuff, A. Lewis, S.J. Miller, The Weibull Distribution and Benford's Law, *Involve*, **8** (2015), 859–874.

[3] F. Dall'Antonia, J. Negroni, G.N. Murshudov, T.R. Schneider, Implementation of a B-factor validation protocol for macromolecular structures, *Acta Crystallographica Section A: Foundations of Crystallography*, **68** (2012), s81.

[4] P. Diaconis, The distribution of leading digits and uniform distribution mod 1, *Ann. Probab.*, **5** (1979), 72–81.

[5] P. Diaconis, D. Freedman, On Rounding Percentages, *J. Am. Stat. Assoc.*, **74** (1979), 359–364.

[6] S.J. Miller (editor), *Benford's Law: Theory and Application*, Princeton University Press, 2015.

[7] J. Negroni, Validation of Crystallographic B Factors and Analysis of Ribosomal Crystal Structures, Ph.D. Thesis, University of Heidelberg (2012), available online at the URL: `http://www.ub.uni-heidelberg.de/archiv/13142`.

[8] E. Stein, R. Shakarchi, *Fourier Analysis: An Introduction*, Princeton University Press, 2003.

[9] M. Pinsky, *Introduction to Fourier Analysis and Wavelets*, Brooks Cole, 2002.

Rebecca F. Durst
Department of Mathematics and Statistics
Williams College
Williamstown, MA 01267
E-mail: `rfd1@williams.edu`

Chi Huynh
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332
E-mail: `huynhngocyenchi@gmail.com`

Adam Lott
Department of Mathematics
University of Rochester
Rochester, NY 14627
E-mail: alott@u.rochester.edu

Steven J. Miller
Department of Mathematics and Statistics
Williams College
Williamstown, MA 01267
E-mail: sjm1@williams.edu

Eyvindur A. Palsson
Department of Mathematics
Virginia Tech
Blacksburg, VA 24061
E-mail: palsson@vt.edu

Wouter Touw
Department of Biochemistry
Netherlands Cancer Institute
Plesmanlaan 121
1066 CX Amsterdam
The Netherlands
E-mail: w.touw@cmbi.ru.nl

Gert Vriend
Radboud University Medical Centre
Centre for Molecular and Biomolecular Informatics
Geert Grooteplein Zuid 26-28 route 260
6525 GA Nijmegen
The Netherlands
E-mail: Gerrit.Vriend@radboudumc.nl

Received: June 19, 2020 **Accepted**: June 29, 2020
Communicated by Serban Raianu