Date: 2025-09-23

Conformity of Time Series to Benford's Law: Finite Range Formulation and Stochastic Evidence

Ben Hull https://bluehood.github.io/

1 Formulation of Benford's Law for Time Series Analysis

The classical form of Benford's law assumes that the Benford set spans many orders of magnitude[1]. In fact, this is often one of the factors that makes naturally occurring datasets likely to conform with BL [2].

1.1 Classical Benford's Law

Benford's law, in essence, is a property of distributions with densities of the form $f(x) \propto 1/x$ – that is, the distribution is scale-invariant. Statistical applications of BL are used to determine whether a set was sampled from such a distribution. A derivation of this result, therefore, assumes the underlying probability density function (PDF) describing the distribution is of the form:

$$f(x) \propto \frac{1}{x}$$

Theorem 1 (Benford's Law) *Consider a set L which is log-uniformly distributed in the range $[0, \infty)$. Then the probability of the integer $d_1 \in \{1, 2, ..., 9\}$ to occur in the first index D_1 is,*

$$P[D_1 = d_1] = \log\left(1 + \frac{1}{d_1}\right). \tag{1}$$

More generally, let D be an n-digit integer $D = \sum_{i=1}^{n} d_i \times 10^{n-i} = d_1 \times 10^{n-1} + d_2 \times 10^{n-2} + ... + d_n$. Then the probability of D to appear in the first n digits of a number is,

$$P_D(n) = \log\left(1 + \frac{1}{D}\right). \tag{2}$$

According to BL, if a set is Benford, the first digit of numbers is far more likely to be 1 (about 30%) than 2 (18%), continuing in this pattern to 9 (5%). This is contrary to one's

1. Mathematically, this is not a requirement but a general rule of thumb for practical applications of BL. For base-10 representations, any range $[1 \times 10^n, 1 \times 10^m)$ with $n, m \in \mathbb{N}, n < m$ where the underlying distribution is proportional to a log-uniform will suffice. Indeed, a union of such intervals will be Benford.
2. More generally, [1] suggests the key condition for BL to hold is that the PDF should be Riemann integrable on the positive real numbers and span many orders of magnitude, with the error from replacing discrete sums by integrals being small. However, for this work, we will consider PDF's proportional to a log-uniform distribution.

intuition – we might expect all digits to have the same probability of appearing. However, for such Benford distributions, this is not the case.

Proof See [1] for a full proof of the result. An intuitive sketch of the proof is given below for the first digit case, where we assume the PDF is scale-invariant[3] (as is the case for $1/x$).

The logarithm of any number x can be written as $\log_{10}(x) = n+\alpha$, where n is an integer and α is the mantissa or fractional part, with $0 \leq \alpha < 1$. The first digit of x is d_1 if and only if:

$$\log_{10}(d_1) \leq \alpha < \log_{10}(d_1 + 1)$$

Since the distribution of the mantissas is uniform in a scale-invariant distribution, the probability of the mantissa falling in this interval is simply the length of the interval.

$$P[D_1 = d_1] = \int_{\log_{10}(d_1)}^{\log_{10}(d_1+1)} 1 \cdot \mathrm{d}x = \log_{10}(d_1 + 1) - \log_{10}(d_1) = \log_{10}\left(\frac{d_1 + 1}{d_1}\right)$$

■

Remark 2 *A natural question arises: what is the most general analytic form of a scale-invariant distribution? By definition, a distribution is said to be scale-invariant if under the transformation $x \mapsto cx$ for $c > 0$, the probability law is preserved up to normalisation. Formally, this requires*

$$f(x)\,dx \quad \propto \quad f(cx)\,d(cx).$$

Suppose, more generally, that the density has the power-law form $f(x) \propto x^{-\alpha}$ for some $\alpha > 0$. Then

$$f(cx) = \frac{1}{c^\alpha} f(x).$$

For exact scale invariance, the distribution must reproduce itself under arbitrary rescalings c, which is only possible when $\alpha = 1$. Thus the unique analytic form of a scale-invariant density is

$$f(x) \propto \frac{1}{x}.$$

Of course, this is not normalizable over the full range $(0, \infty)$. Instead, one considers a restricted support $x \in [a, b]$, where the normalized form is the log-uniform distribution*:*

$$f(x) = \frac{1}{x \log(b/a)}, \quad a < x < b.$$

Philosophically, one could take Benford's law as emerging not from any mysterious numerological principle, but simply from the assumption of such scale-invariant laws of nature. In the limit where the interval $[a, b]$ spans many orders of magnitude, the mantissas of $\log_{10}(x)$ are uniformly distributed, and Benford's law follows inevitably.

3. More generally, we would need to consider another condition – that is the distribution should be invariant under a change of base

Remark 3 *Benford's Law often applies to datasets that are not themselves exactly log-uniform distributions of the form $1/x$ but still exhibit scale invariance. These distributions tend to span several orders of magnitude and often arise from multiplicative processes or combinations of different data sources. For example:*

- *Log-normal Distributions: The log-normal distribution is a common example. If a variable is log-normally distributed, its logarithm is normally distributed. As a result, when the set spans a wide range, the distribution of their logarithms becomes approximately uniform, leading to the Benford pattern.*
- *Power-Law Distributions: Many natural and social phenomena follow a power-law distribution, where the frequency of an event is proportional to a power of its size. Examples include city populations and word frequencies. These distributions are also scale-invariant and typically adhere to Benford's Law.*
- *Exponential Distributions: The exponential distribution, often used to model waiting times, can also conform to Benford's Law for certain parameters.*
- *Combinations of Distributions: Datasets created by combining numbers from various unrelated sources, such as a collection of different statistical tables, often obey Benford's Law. This is due to a central limit theorem for products of random variables, where the distribution of the logarithms of the products tends toward uniformity*

BL is a fascinating result in and of itself, and has many direct applications in the classical form. However, for the analysis of time-series data, the series may not span many orders of magnitude and may instead be constrained to a finite range. In such cases, our assumption that the underlying set we are analysing spans many orders of magnitude is not met, and we should not assume that BL is applicable. In such cases, we need to reformulate BL to account for such a finite range, aptly referred to as the Finite Range Benford Law.

1.2 Finite Range Benford Law

A general requirement of a set to be Benford is that the distribution should span many orders of magnitude. This is required for some distributions, such as log-normal distributions, but not for others, such as log-uniform distributions. We would therefore like to construct the mathematical framework for the application of Benford's law to finite ranges. This will allow us to determine how well a distribution that does not span a large range of values conforms with BL.

The sampling density is proportional to $1/x$ and the observables are taken from a finite multiplicative range (as before). Below $n \in \mathbb{N}$ denotes the number of leading (base-10) digits we inspect and D denotes an n-digit integer, i.e. $D = \sum_{i=1}^{n} d_i \times 10^{n-i} = d_1 \times 10^{n-1} + d_2 \times 10^{n-2} + ... + d_n$.

We define a set of numbers sampled from the density function[4]:

$$U \;=\; [\, a \times 10^{\alpha},\; b \times 10^{\beta}\,), \qquad 1 \le a < 10,\; 1 \le b < 10,\; \alpha, \beta \in \mathbb{Z},\; \alpha \le \beta.$$

4. When $\alpha = \beta$ the support lies inside a single decade $[a \times 10^{\alpha}, b \times 10^{\alpha})$; when $\alpha < \beta$ the support spans multiple decades.

The set of positive real numbers whose first n digits equal D is

$$\mathcal{I}_D \;=\; \bigcup_{m\in\mathbb{Z}} [\, D\cdot 10^{\,m-n+1},\ (D+1)\cdot 10^{\,m-n+1}\,).$$

Define the probability of observing leading digits D (for n digits) by

$$P_D(n) \;=\; \frac{\displaystyle\int_{U\cap\mathcal{I}_D} \frac{\mathrm{d}x}{x}}{\displaystyle\int_{U} \frac{\mathrm{d}x}{x}}.$$

Which is just the integral over the set of positive real numbers whose first n digits equal D divided by the integral over the entire range.

We now state the two useful closed-form results (single-decade $\alpha=\beta$, and multi-decade $\alpha<\beta$) in a compact form[5].

Theorem 4 (Single-decade ($\alpha=\beta$) finite-range formula) *Let $U=[a\times 10^{\alpha},\ b\times 10^{\alpha})$ with $1\le a<b<10$ and fix $n\in\mathbb{N}$ and D an n-digit integer. Define*

$$A := a\times 10^{\,n-1}, \qquad B := b\times 10^{\,n-1}.$$

Then the probability that a random $x\in U$ (with density $\propto 1/x$) has leading n digits equal to D is

$$P_D(n) \;=\; \frac{\log_{10}\!\left(\dfrac{\min(D+1,\ B)}{\max(D,\ A)}\right)}{\log_{10}\!\left(\dfrac{b}{a}\right)} \tag{3}$$

with the convention that the numerator is zero when the interval $[D,D+1)$ and $[A,B)$ do not overlap (so $P_D(n)=0$ in that case).

Proof Start from the integral definition

$$P_D(n) = \frac{\displaystyle\int_{\max(a\times 10^{\alpha},\, D10^{\alpha-n+1})}^{\min(b\times 10^{\alpha},\,(D+1)10^{\alpha-n+1})} \frac{\mathrm{d}x}{x}}{\displaystyle\int_{a10^{\alpha}}^{b10^{\alpha}} \frac{\mathrm{d}x}{x}}.$$

If the intersection in the numerator is empty the numerator is zero and the formula is trivial. Otherwise perform the substitution

$$x = 10^{\alpha-n+1}\,x', \qquad \mathrm{d}x/x = \mathrm{d}x'/x'.$$

Under this rescaling the numerator integrates $1/x'$ over $[\,\max(A,D),\ \min(B,D+1)\,)$ where $A=a\times 10^{\,n-1}$ and $B=b\times 10^{\,n-1}$. Using $\int_u^v \frac{\mathrm{d}x'}{x'} = \ln(v)-\ln(u) = \ln(10)\big(\log_{10} v-\log_{10} u\big)$,

5. Please also see [2] for a piecewise treatment of the case $\alpha<\beta$

the factor $\ln(10)$ cancels between numerator and denominator and we obtain

$$P_D(n) = \frac{\log_{10}\Big(\dfrac{\min(D+1, B)}{\max(D, A)}\Big)}{\log_{10}\Big(\dfrac{b}{a}\Big)},$$

which is (3). ■

Remark 5 (Reduction to classical Benford) *If $a = 1$ and $b = 10$ (i.e. U is a full decade $[10^\alpha, 10^{\alpha+1})$) then $A = 10^{n-1}$ and $B = 10^n$, the numerator equals $\log_{10}((D+1)/D) = \log_{10}(1 + 1/D)$ and the denominator equals $\log_{10}(10) = 1$, hence*

$$P_D(n) = \log_{10}\Big(1 + \frac{1}{D}\Big),$$

the standard Benford multi-digit law.

Theorem 6 (Multi-decade ($\alpha \le \beta$) finite-range formula) *Let $U = [a \times 10^\alpha,\ b \times 10^\beta)$ with $1 \le a, b < 10$ and integers $\alpha \le \beta$. Fix $n \in \mathbb{N}$ and an n-digit integer D. Define the rescaled endpoints*

$$A := a \cdot 10^{n-1}, \qquad B := b \cdot 10^{\beta-\alpha+n-1}.$$

Then

$$P_D(n) \;=\; \frac{\displaystyle\sum_{k=0}^{\beta-\alpha} \max\Big\{0,\ \log_{10}\Big(\frac{\min\big((D+1)10^k,\ B\big)}{\max\big(D\,10^k,\ A\big)}\Big)\Big\}}{\beta - \alpha + \log_{10}\Big(\dfrac{b}{a}\Big)} \tag{4}$$

where each summand contributes only when the k-th shifted D-block overlaps the rescaled support $[A, B)$.

Proof Begin with the definition

$$P_D(n) = \frac{\displaystyle\int_{U \cap \mathcal{I}_D} \frac{\mathrm{d}x}{x}}{\displaystyle\int_U \frac{\mathrm{d}x}{x}}.$$

Decompose $\mathcal{I}_D$ into disjoint decade-shifted blocks

$$\mathcal{I}_D = \bigcup_{m \in \mathbb{Z}} [\, D \cdot 10^{m-n+1},\ (D+1) \cdot 10^{m-n+1}\,).$$

Only those m for which $[D\cdot 10^{m-n+1},(D+1)\cdot 10^{m-n+1})$ intersects $U=[a\times 10^{\alpha},b\times 10^{\beta})$ can contribute. Perform the rescaling

$$x=10^{\alpha-n+1}\,x',\qquad \mathrm{d}x/x=\mathrm{d}x'/x'.$$

Under this map the data-support becomes $[A,B)$ with $A=a\times 10^{\,n-1}$ and $B=b\times 10^{\,\beta-\alpha+n-1}$, and each block $[D\cdot 10^{m-n+1},(D+1)\cdot 10^{m-n+1})$ becomes $[D\cdot 10^{m-\alpha},(D+1)\cdot 10^{m-\alpha})$. Letting $k=m-\alpha$ we see the only relevant k are $k=0,1,\dots,\beta-\alpha$. Thus the numerator (logarithmic length of the overlap) equals the finite sum

$$\sum_{k=0}^{\beta-\alpha}\int_{[A,B)\cap[D10^{k},(D+1)10^{k})}\frac{\mathrm{d}x'}{x'}\;=\;\sum_{k=0}^{\beta-\alpha}\max\Big\{0,\;\log_{10}\Big(\frac{\min\big((D+1)10^{k},\;B\big)}{\max\big(D\,10^{k},\;A\big)}\Big)\Big\},$$

where we again used $\int_u^v\frac{\mathrm{d}x'}{x'}=\ln(10)(\log_{10}v-\log_{10}u)$ and dropped the common factor $\ln(10)$. The denominator equals

$$\int_{a10^{\alpha}}^{b10^{\beta}}\frac{\mathrm{d}x}{x}=\ln(10)\Big(\beta-\alpha+\log_{10}\frac{b}{a}\Big),$$

so $\ln(10)$ cancels and we obtain (4). ■

Proposition 7 (Normalisation) *For fixed n and the support U above, the probabilities $\{P_D(n)\}$ sum to 1 over all n-digit D.*

Proof [Sketch of proof] This follows because the sets $\mathcal{I}_D$ partition the positive reals according to initial n digits, so $\sum_D \mathbf{1}_{\mathcal{I}_D}(x)=1$ for all $x>0$. Integrating $1/x$ over U and summing over D allows interchange of sum and integral (finite or absolutely convergent sum in our finite-range case) and yields $\sum_D$ numerator = denominator. After cancellation this gives $\sum_D P_D(n)=1$. ■

Corollary 8 (Recovery of classical Benford in the infinite-range limit) *Let the finite-range support be*

$$U_{(\alpha,\beta)}\;=\;[\,a\cdot 10^{\alpha},\;b\cdot 10^{\beta}\,),\qquad 1\le a,b<10,\quad \alpha<\beta,$$

and let $P_D^{(\alpha,\beta)}(n)$ denote the probability from (4) associated with $U_{(\alpha,\beta)}$ for fixed $n\in\mathbb{N}$ and fixed n-digit integer D. Put

$$N\;:=\;\beta-\alpha.$$

Then, keeping a,b,n,D fixed and letting $N\to\infty$ (equivalently letting the support expand multiplicatively to cover all positive reals), we have

$$\lim_{N\to\infty}P_D^{(\alpha,\beta)}(n)\;=\;\log_{10}\Big(1+\frac{1}{D}\Big),$$

i.e. the finite-range probabilities converge to the ordinary (classical) Benford n-digit law.

Proof Recall the multi-decade formula (with the rescaled endpoints as in the theorem)

$$P_D^{(\alpha,\beta)}(n) = \frac{\sum_{k=0}^{N} T_k(N)}{N + \log_{10}\Big(\frac{b}{a}\Big)},$$

where for brevity we put $N = \beta - \alpha$ and each summand is

$$T_k(N) \;=\; \max\Big\{0,\; \log_{10}\Big(\frac{\min\big((D+1)10^k,\; B(N)\big)}{\max\big(D\,10^k,\; A\big)}\Big)\Big\},$$

with $A = a \times 10^{n-1}$ (fixed) and $B(N) = b \times 10^{N+n-1}$ (grows with N).

(1) Each summand is uniformly bounded. For every k and N,

$$0 \le T_k(N) \le \log_{10}\Big(\frac{D+1}{D}\Big) \equiv L_D,$$

because the overlap of any block $[D10^k, (D+1)10^k)$ with $[A, B(N))$ cannot exceed the full block length $\log_{10}((D+1)/D)$.

(2) There are $N - O(1)$ full-block contributions. A summand $T_k(N)$ equals the full-block length L_D precisely when

$$D \cdot 10^k \ge A \qquad \text{and} \qquad (D+1)\cdot 10^k \le B(N).$$

Taking base-10 logarithms these inequalities are

$$k \ge \log_{10}\Big(\frac{A}{D}\Big) \quad \text{and} \quad k \le \log_{10}\Big(\frac{B(N)}{D+1}\Big).$$

Since $\log_{10}(B(N)/(D+1)) = N + C_1$ for a constant C_1 depending only on b, n, D (but *not* on N), and $\log_{10}(A/D) = C_0$ is a constant independent of N, it follows that the range of k for which both inequalities hold has length

$$\big(N + C_1\big) - C_0 + O(1) = N - O(1).$$

Hence the number m_N of indices $k \in \{0, \dots, N\}$ with $T_k(N) = L_D$ satisfies

$$m_N = N - O(1),$$

where the implicit $O(1)$ is a constant independent of N (at worst depending on a, b, n, D).

Decompose the numerator. Write

$$\sum_{k=0}^{N} T_k(N) \;=\; m_N\, L_D \;+\; R_N,$$

where the remainder R_N sums the at most $O(1)$ boundary/partial terms. By (1) and (2) we have the bounds

$$0 \le R_N \le C_2,$$

for some constant C_2 independent of N (indeed R_N is bounded by at most two partial-block contributions plus a bounded number of zero terms).

Take the limit. Now

$$\begin{aligned} P_D^{(\alpha,\beta)}(n) &= \frac{m_N\, L_D + R_N}{N + \log_{10}(b/a)} \\ &= \frac{\big(N - O(1)\big)\, L_D + O(1)}{N + \log_{10}(b/a)}. \end{aligned}$$

Dividing numerator and denominator by N and letting $N \to \infty$ gives

$$\lim_{N\to\infty} P_D^{(\alpha,\beta)}(n) \;=\; L_D \;=\; \log_{10}\Big(1 + \frac{1}{D}\Big),$$

which is the classical BL. ■

Remark 9 *The basic result here is that the first and last decade in the set can only contribute partially to BL – that is, all other ranges conform to BL fully according to (5) as they all span a full decade. As we take $N \to \infty$ the contribution of these decades at the ends of the range becomes smaller and smaller and eventually vanish.*

Remark 10 *It is also clear that in the case that $\alpha = \beta$ we recover the single decade result from (4). That is $\beta - \alpha = 0$ and the sum $\sum_{k=0}^{\beta-\alpha}$ disappears, reducing to the signle term $k = 0$. Therefore*

$$T_k(N) \;=\; \max\Big\{0,\; \log_{10}\Big(\frac{\min\big((D+1)10^k,\; B(N)\big)}{\max\big(D\,10^k,\; A\big)}\Big)\Big\},$$

becomes,

$$T(N) = \; \log_{10}\Big(\frac{\min\big((D+1),\; B\big)}{\max\big(D,\; A\big)}\Big),$$

and the denominator

$$\beta - \alpha + \log_{10}\Big(\frac{b}{a}\Big)$$

simply becomes

$$\log_{10}\Big(\frac{b}{a}\Big)$$

giving the expected format predicted by (3).

1.3 Application of the FRBL

We have developed the mathematical framework for analysing the Benford distributions over finite ranges. We now need to understand how to apply this formulation to datasets and discuss statistical tests to measure conformity BL. We will visualise both BL and the more general FRBL.

1.3.1 Statistical Test: d^*

We introduce the d^* metric, a test statistic that is based on the Euclidean distance between observed and expected digit occurrences in a set. Morrow [3] introduces a definition for the d^* measure and computes the associated p values for the first digit test (which are dicussed here [4])

$$d^* = \sqrt{N \sum_{i=1}^{9} (p_i - b_i)^2}, \tag{5}$$

where p_i is the observed proportion of digits in the first index having a value $i \in \{1, 2, ..., 9\}$ and b_i the corresponding expected probability according to BL, and N is the size of the set being analysed for conformity with BL. We can extend the definition of d^* to other digit tests as follows,

$$d^* := \sqrt{N \sum_{i} (p_i - b_i)^2}, \tag{6}$$

where p_i is the proportion of observations having $i = (d_n, ..., d_m, ..., d_2, d_1)$ as the $(D_n, ..., D_m, ..., D_2, D_1)$ indexes and b_i the corresponding expected proportion according to BL. The sum runs over all possible digit sequences in the indexes $(D_n, ..., D_m, ..., D_2, D_1)$.

We will use Morrow's definition of the d^* statistic to gauge conformity with BL and the FRBL. Rather than analysing the p-values directly, we will primarily be interested in changes in d^* between different sets of data and, in particular, different time series. However, losely speaking, a lower value of d^* corresponds to better conformity with a Benford distribution.

Please see [4] for a discussion of the benefits of using the d^* for Benford-specific applications and a more general discussion of the metric.

1.3.2 Visualising BL and FRBL

In order to analyse BL we need to apply it to a dataset. There are many examples of datasets that are Benford, as we have seen in this paper, that we could choose from. However, geometric series are known to be Benford and we can generate a geometric series in any given range with an arbitrary number of terms[4]. We will therefore use these synthetically generated geometric series as a basis for applying BL and FRBL to more general datasets. This will also give an intuitive understanding of how BL compares with the FRBL.

Fig 1 shows the application of BL and FRBL to a geometric series in the range $[1, 10000)$ with a thousand elements. We observe conformity with a Benford distribution in both cases – as expected for a geometric series. In particular, since the range of data perfectly spans 4 decades of data, the BL and FRBL converge to the same distribution of first digits, as expected.

We are not limited to only considering sets with ranges perfectly spanning decades of data – we can consider any range[6]. Fig 2 shows the FRBL applied to several ranges of data.

6. Indeed, we could also consider finite unions of such ranges as well. These should be finite as we actually need to compute the probabilities

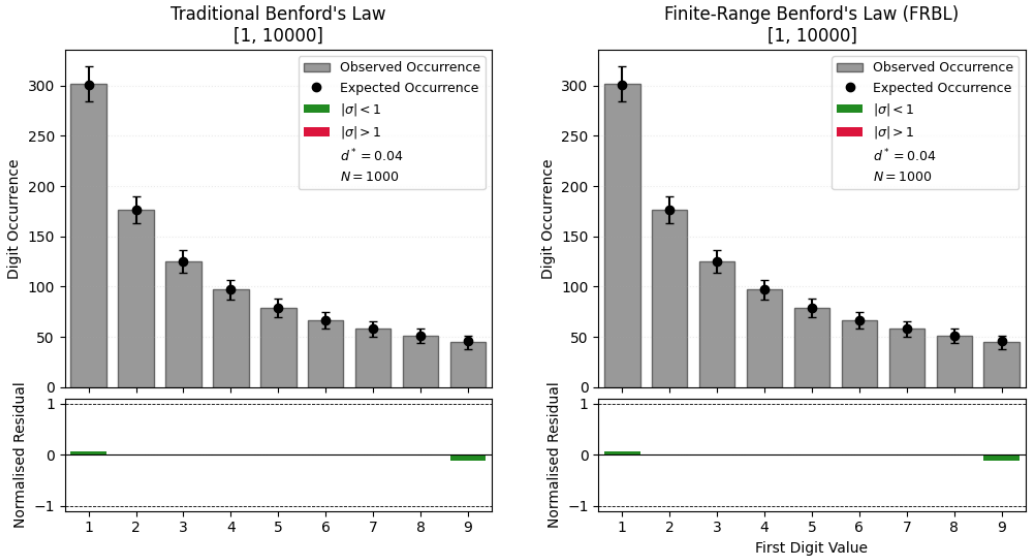

Figure 1: Histograms plotting the BL (left) and FRBL (right) to the same geometric series with a thousand terms in the range $[1, 10000)$. The expected value of the occurrence, E_i, is calculated using the formulation of BL and FRBL in section 1.2, with error bars given by the Poisson noise, $\sqrt{E_i}$. The actual observed occurrences, O_i, are shown for each of the first digits $\{1, 2, ..., 9\}$, as a bar. The normalised residuals are plotted under the graphs and show whether the observed data is within one standard error (shown in green) or outside one standard error (shown as red). In the case where we have 4 full decades in the range, the classical BL and FRBL plots are identical. The d^* metric, measuring the conformity with the underlying Benford distribution, demonstrates excellent conformity with both BL and the FRBL.

We can see that the distribution of first digits varies significantly from BL depending on the range. For example, for the range $[1, 5)$, we see the distribution only expects first digits in the set $\{1, 2, 3, 4\}$, which is correct. This distribution is clearly different from the classical BL, however, the general observation that the lower digits have a larger expectation still holds in this case. Again, all values of d^* suggest conformity with a Benford distribution.

However, these finite ranges of data do not necessarily conform with the classical BL, even though the underlying distribution is Benford. This is precisely because of the assumptions made when deriving the classical BL, that the tails of the distribution do not contribute significantly. This is equivalent to the datasets spanning several orders of magnitude.

Fig 3 illustrates the deviation from BL and the FRBL for geometric sets, as measured by the d^* metric. The x-axis represents the coefficient b, which determines the upper bound of the data range $[1, b \times 10^3]$, while the y-axis shows the d^* metric, a measure of conformity to the respective laws. The blue data points correspond to BL, and the red data points to FRBL.

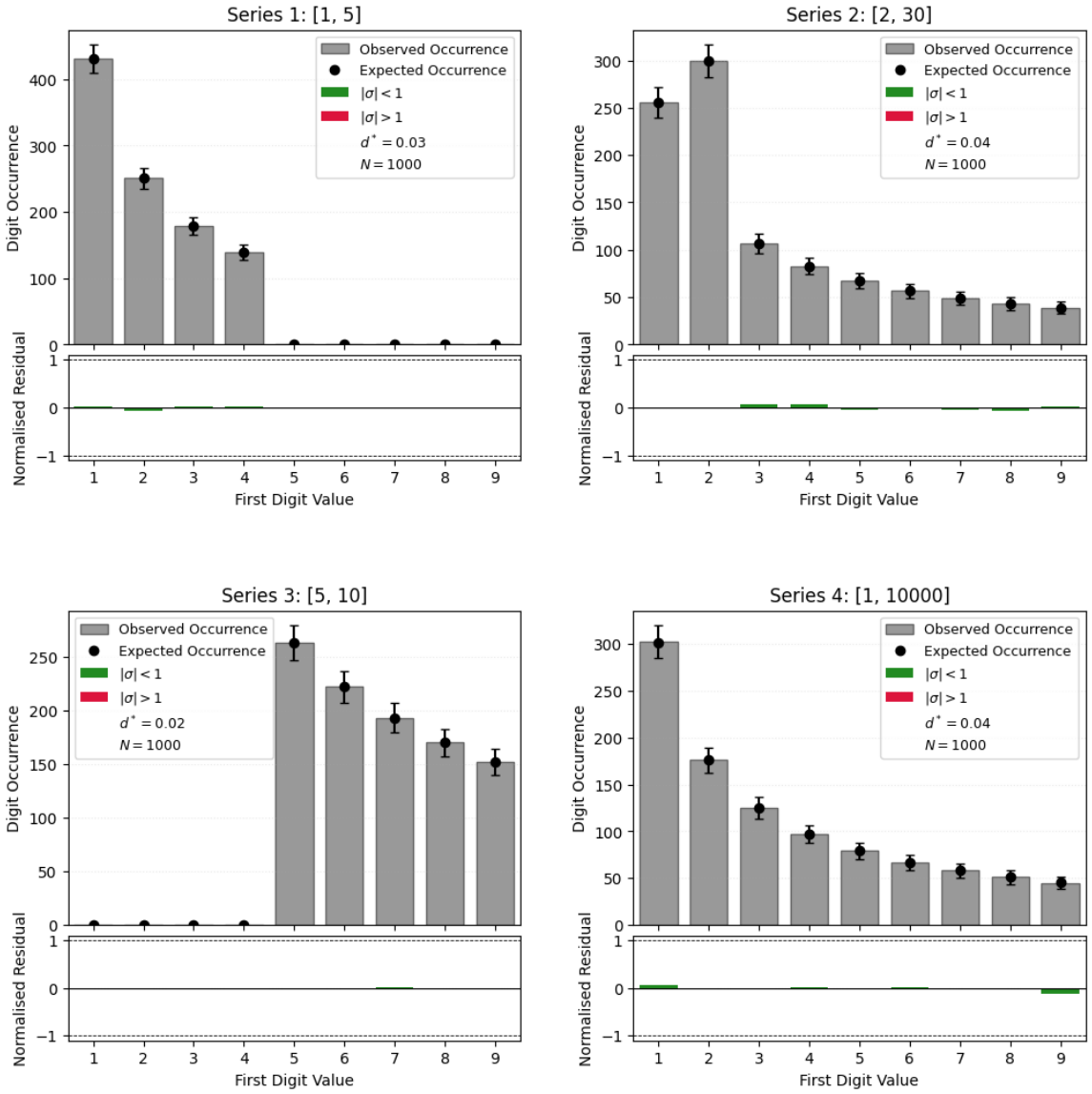

Figure 2: Comparison of observed first-digit counts with FRBL for four finite multiplicative ranges: Series 1, $[1, 5)$, only digits 1–4 occur; (b) Series 2, $[2, 30)$, digits 1–9 appear but 1 and 2 dominate; (c) Series 3, $[5, 10)$, only digits 5–9 occur; (d) Series 4, $[1, 10000)$, the broad multi-decade range shows the familiar Benford-like decay from 1 to 9. In all plots, the residuals are small, demonstrating close agreement between the data and the FRBL.

The blue data points show the variation of the d^* metric for Benford's Law across the range of b.

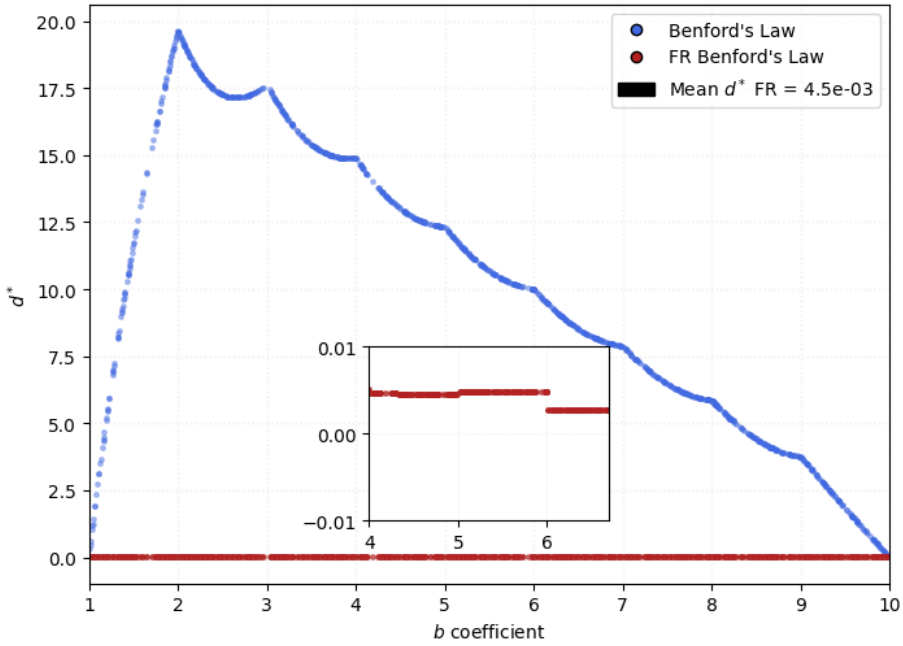

Figure 3: The variation of the d^* metric when the first digit test is applied to geometric Benford sets spanning different dynamical ranges, $[1, b \times 10^3]$, as a function of the coefficient $b \in [1, 10]$. BL is shown in blue and the FRBL is shown in red for the first index over this finite range. Each plot has one thousand data points. We measure good conformity with BL at the endpoints $b = 1, 10$ with weaker conformity away from these values for both metrics. There are local maxima at $b = 2$ for d^*. All elements of the base sets in the range $[10^3, b \times 10^3]$ have one in the first index when $b < 2$. This introduces a disproportionate number of ones into the distribution of first digits, resulting in a poor fit with BL. When $b > 2$, other digits appear in the first index $(2, 3, 4, ..., 9)$, which improves conformity with BL. For d^* we observe a monotone increase between $b = 1$ and $b = 2$ followed by a monotone decrease between $b = 2$ and $b = 10$. This shows that the d^* metric is sensitive to the distribution's underlying shape rather than deviations in individual bins. Note, we observe excellent conformity with the FRBL.

- Endpoints Conformity: The figure shows that the d^* metric is close to zero at the endpoints $b = 1$ and $b = 10$. This indicates good conformity with BL for these specific ranges, where the data sets are $[1, 10^3]$ and $[1, 10^4]$, respectively. This is expected, as over these ranges we are analysing full decades of data.

- Monotonic Behaviour: We observe a distinct pattern in the d^* metric for BL. There is a sharp, nearly monotonic increase from $b = 1$ to a local maximum around $b = 2$, followed by a general monotonic decrease from $b = 2$ to $b = 10$. This suggests that the fit to BL is worse around $b = 2$ and improves as b moves away from this value towards either endpoint.

- Impact of First Digit Distribution: For $b < 2$, all numbers in the range $[10^3, b \times 10^3]$ have a first digit of 1. This is because any number N in this range satisfies $1000 \leq N < 2000$, thus having a first digit of 1. This disproportionately large number of 1s skews the first-digit distribution, leading to a poor fit with BL's logarithmic distribution. As b increases beyond 2, other digits (2, 3, 4, etc.) begin to appear in the first position, which normalises the distribution and improves the fit with BL, leading to the observed decrease in d^*. The non-smooth, step-like nature of the decay for $b > 2$ is likely due to the successive appearance of new leading digits as the range expands.

The d^* metric for FRBL remains consistently near zero, providing strong empirical evidence that FRBL is the correct and more accurate model for predicting the first-digit distributions of Benford sets.

We now have a general framework for applying BL to constrained datasets. This will be invaluable when applying this technique to time series, which may not necessarily span a large range or may not span more than a single decade. In particular, we will begin by analysing time series of synthetically generated stock prices. We will provide the mathematical framework for doing so and show that time series generated under this framework are Benford in the finite range case. We will introduce methods for analysing conformity with BL and whether this may yield any useful results for quantitative analysis of price movements.

2 Geometric Brownian Motion and BL

Geometric Brownian motion (GBM) is the stochastic process most commonly used in the classical Black–Scholes framework to model the evolution of a non-dividend-paying stock price S_t. The GBM model assumes continuous paths, proportional (multiplicative) noise, constant drift and constant volatility. The model was central to the Black–Scholes option pricing theory and related developments in continuous-time finance [5; 6].

GBM offers a relatively simplistic model of stock price movements that we will use as a basis for an initial Benford analysis of stock price time series.

2.1 Geometric Brownian Motion

Theorem 11 (Time-Homogeneous Geometric Brownian motion) *A (time-homogeneous) geometric Brownian motion $\{S_t : t \geq 0\}$ with drift $\mu \in \mathbb{R}$ and volatility $\sigma > 0$ is given d as the (strong) solution of the stochastic differential equation*

$$dS_t = \mu S_t \, dt + \sigma S_t \, dW_t, \qquad S_0 > 0, \tag{7}$$

with the solution

$$S_t = S_0 \exp\Big(\big(\mu - \tfrac{1}{2}\sigma^2\big)t + \sigma W_t \Big). \tag{8}$$

Probability space and filtration. We work on a filtered probability space

$$(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$$

where the filtration $(\mathcal{F}_t)_{t\geq 0}$ satisfies the usual conditions (right-continuous and complete). The filtration models the flow of information available through time; requiring S_t to be adapted to $(\mathcal{F}_t)$ means the value of S_t at time t is determined by information available up to time t.

Standard Brownian motion. The process $(W_t)_{t\geq 0}$ is a standard Brownian motion (also called a Wiener process) relative to $(\mathcal{F}_t)$ and $\mathbb{P}$. Concretely:

- $W_0 = 0$ almost surely,
- W_t has continuous paths,
- increments are independent and Gaussian: for $0 \leq s < t$, $W_t - W_s \sim N(0, t-s)$ and is independent of $\mathcal{F}_s$.

The increment $\mathrm{d}W_t$ in the differential form is an informal notation for the infinitesimal increment of this continuous martingale and must be understood in the sense of Itô integration (see below). Standard references cover these facts in detail [7; 8].

Integral form and Itô interpretation. The differential (7) is shorthand for the integral equation

$$S_t \;=\; S_0 + \int_0^t \mu S_s \, \mathrm{d}s \;+\; \int_0^t \sigma S_s \, \mathrm{d}W_s, \tag{9}$$

where the first integral is a classical (Lebesgue) integral and the second is an Itô stochastic integral. The Itô integral

$$\int_0^t \sigma S_s \, \mathrm{d}W_s$$

is defined for adapted processes satisfying suitable integrability conditions and has the crucial isometry property

$$\mathbb{E}\Big[\Big(\int_0^t \sigma S_s \, \mathrm{d}W_s\Big)^2\Big] = \mathbb{E}\Big[\int_0^t \sigma^2 S_s^2 \, \mathrm{d}s\Big].$$

Thus statements about existence, uniqueness and moments are proved in the framework of Itô calculus [7; 8].

Meaning of the terms. Writing the SDE in relative form

$$\frac{\mathrm{d}S_t}{S_t} = \mu \, \mathrm{d}t + \sigma \, \mathrm{d}W_t, \tag{10}$$

makes the economic interpretation transparent:

- μ is the instantaneous (expected) *rate of return* per unit time (units: 1/time). Over a short time Δt, the deterministic contribution to the expected proportional change is approximately $\mu \, \Delta t$.
- $\sigma > 0$ is the instantaneous *volatility* (units: $1/\sqrt{\text{time}}$); over a short time interval of length Δt the random component has standard deviation approximately $\sigma\sqrt{\Delta t}$. The product $\sigma \, \mathrm{d}W_t$ therefore models random fluctuations in the *proportional* change of S_t.
- $S_0 > 0$ is the (non-random or $\mathcal{F}_0$–measurable) initial price.

Because both drift and diffusion coefficients are proportional to S_t, the noise is *multiplicative*: the magnitude of absolute fluctuations scales with the level of the process. This contrasts with additive noise models where fluctuations are independent of the level.

Proof To solve (7) apply Itô's lemma [9] to $X_t = \ln S_t$. If $S_t > 0$ and $f(s) = \ln s$, then

$$df(S_t) = \frac{1}{S_t} dS_t - \frac{1}{2} \frac{1}{S_t^2} (dS_t)^2.$$

Using (7) and the Itô calculus rule $(\mathrm{d}W_t)^2 = \mathrm{d}t$, we obtain

$$d(\ln S_t) = \left(\mu - \frac{\sigma^2}{2}\right) \mathrm{d}t + \sigma \, \mathrm{d}W_t.$$

Integrating from 0 to t gives

$$\ln S_t = \ln S_0 + \left(\mu - \frac{\sigma^2}{2}\right) t + \sigma W_t.$$

Exponentiating yields the explicit solution

$$S_t = S_0 \exp\Big(\big(\mu - \tfrac{1}{2}\sigma^2\big)t + \sigma W_t\Big). \tag{11}$$

Thus S_t is log-normally distributed for each fixed $t > 0$. ■

Remark 12 (Positivity construction) *One important modelling requirement for stock prices is non-negativity. For GBM the positivity is immediate from the explicit solution:*

$$S_t \;=\; S_0 \exp\Big((\mu - \tfrac{1}{2}\sigma^2)t + \sigma W_t \Big). \tag{12}$$

Since the exponential is strictly positive, $S_t > 0$ almost surely whenever $S_0 > 0$. Equivalently, S is the stochastic (Doléans–Dade) exponential of the continuous semimartingale $(\mu - \frac{1}{2}\sigma^2)t + \sigma W_t$; this perspective explains positivity more abstractly (see [7]).

Remark 13 (Path properties and Markov property) *The sample paths of S_t are continuous (as a continuous function of W_t) but almost surely nowhere differentiable — a consequence of the roughness of Brownian motion. The process (S_t) is Markovian: its future evolution depends only on the current value S_t (and time), not the past history. Furthermore, the process of log-prices $\log S_t$ has independent Gaussian increments, which implies that $\log S_t$ is an arithmetic Brownian motion and that S_t is log-normal at fixed times.*

Remark 14 (Why the assumptions are required / modelling remarks) *The assumptions of the model yield some useful properties:*

- *The condition $\sigma > 0$ ensures genuine randomness in returns; if $\sigma = 0$ the model reduces to deterministic exponential growth $S_t = S_0 e^{\mu t}$.*
- *The multiplicative form $\sigma S_t\, dW_t$ is chosen in finance because it yields stationary (time-homogeneous) distributional properties for log-returns and ensures positivity, both desirable for modelling asset prices.*

GBM's main simplifying assumptions are constant volatility and continuous paths. Empirically, asset returns show volatility clustering, leverage effects, and jumps. For these reasons, models with stochastic volatility or jump-diffusions are commonly used as extensions [10].

Additionally, given that solution S_t is log-normally distributed and the noise added to the price path is multiplicative, this suggests that GBM may be Benford. We will analyse this proposition in the next section.

2.2 Is GBM Benford?

The question of whether GBM obeys Benford's Law has been addressed in the literature from both rigorous and heuristic perspectives.

The key mechanism used in proofs is the *equidistribution modulo one* of the logarithm: if $\log_B S_t \bmod 1$ becomes (or is) uniformly distributed, then the leading digits of S_t follow Benford's law. For GBM one observes that $\log S_t$ is Gaussian with variance growing linearly in t. Such a distribution whose logarithms spread out (variance $\to \infty$) and satisfy mild regularity conditions produces Benford behaviour. This viewpoint is used explicitly in Kontorovich and Miller's structural approach (see their discussion of "spreading Gaussians" and how Poisson summation is applied to limiting densities) and is described in more expository form in Miller's and Berger–Hill's work [11; 12; 13; 14].

To be precise about *the sense* in which GBM is Benford, authors differ in formulation, and those distinctions matter for applications:

- **Ensemble / distributional sense.** Fix t (or let t vary over a diverging sequence); consider the law of S_t (or the law of S_{t_n} as $t_n \to \infty$). Because $\log S_t$ is Gaussian with variance $\sigma^2 t$, as $t \to \infty$ the density of $\log S_t$ "spreads out" and under the hypotheses used in Kontorovich–Miller this yields Benford behaviour for the distributions in the large-time limit (see [11; 12]).

- **Time-series / trajectory sense.** One may sample a single realisation $(S_t)_{t\geq 0}$ at times $t_1 < t_2 < \cdots$ and ask whether the empirical distribution of leading digits of the sampled sequence approaches the Benford proportions. Results here typically rely on ergodicity/mixing or on ensemble-to-time transfer arguments; the survey and primer by Berger and Hill summarise conditions under which continuous-time stochastic processes (including GBM as a primary example) produce Benford digit distributions along time or along typical trajectories [13; 14].

- **Practical / finite-sample caveats.** Even though mathematical limits (e.g. $t \to \infty$) indicate Benford behaviour, finite-time samples, strong deterministic trends, choice of sampling scheme, discretisation, and economic constraints can produce pronounced deviations. For example, empirical studies of real financial time series (e.g. S&P500) report mixed or negative conformity to Benford's law for raw price series, while log-returns sometimes conform more closely, demonstrating that model assumptions and sampling matter in practice [15].

If BL were to be used in quantitative analysis of time series, for example, as a signal to identify anomalies, we would need to be able to apply a finite range analysis. Locally at least, the time series element will be constrained to a finite range; the FRBL offers a promising candidate for local time series analysis.

Theorem 15 (Finite Range Benford Convergence for GBM) *Let X_t be geometric Brownian motion:*

$$X_t = X_0 \exp\big((\mu - \tfrac{1}{2}\sigma^2)t + \sigma W_t\big),$$

with $\sigma > 0$. Put

$$Y_t = \log_{10} X_t = \log_{10} X_0 + \alpha t + \beta W_t, \qquad \alpha = \frac{\mu - \frac{1}{2}\sigma^2}{\ln 10},\ \beta = \frac{\sigma}{\ln 10}.$$

Fix any interval $I = [a, b) \subset [1, 10)$. Then

$$\lim_{t\to\infty} \Pr\big(M_{10}(X_t) \in [a, b)\big) = \log_{10} \frac{b}{a},$$

or rather the time series is Benford when $t \to \infty$.

Proof For each integer k define

$$\varphi_t(k) := \mathbb{E}\big[e^{2\pi i k Y_t}\big].$$

Using $\mathbb{E}[e^{iuW_t}] = \exp(-u^2 t/2)$ we obtain, for $k \neq 0$,

$$\varphi_t(k) = e^{2\pi i k(\log_{10} X_0 + \alpha t)} \exp\Big(-\tfrac{1}{2}(2\pi k\beta)^2 t\Big).$$

Since $\beta > 0$, the Gaussian damping factor tends to zero as $t \to \infty$, hence

$$\lim_{t\to\infty} \varphi_t(k) = 0 \quad (k \neq 0), \qquad \varphi_t(0) \equiv 1.$$

Let μ_t be the distribution of the fractional part $\{Y_t\}$ on $\mathbb{R}/\mathbb{Z}$. The numbers $\varphi_t(k)$ are the Fourier coefficients of μ_t, and their pointwise limit (1 at $k = 0$, 0 elsewhere) are the Fourier coefficients of Lebesgue measure on the circle. By uniqueness of measures determined by Fourier coefficients,

$$\mu_t \xrightarrow{w} \text{Uniform}[0, 1).$$

Now let $I = [a, b) \subset [1, 10)$ be fixed and put

$$J = \{\log_{10} s : \ s \in I\} = [\log_{10} a, \log_{10} b) \subset [0, 1).$$

Then $\Pr(M_{10}(X_t) \in I) = \mu_t(J)$. Since J is an interval its boundary has at most two points; the Uniform$[0, 1)$ limit measure assigns those points zero mass.

Theorem 16 (Portmanteau theorem, special case) *If a sequence of probability distributions μ_t converges weakly to a probability distribution μ, then for any set A whose boundary has μ-measure zero, one has*

$$\lim_{t\to\infty} \mu_t(A) = \mu(A).$$

The Portmanteau theorem therefore gives

$$\lim_{t\to\infty} \mu_t(J) = \text{Leb}(J) = \log_{10} \frac{b}{a}.$$

■

Discussion

- The finite-range (first digit) version is just asking: pick a fixed interval $I \subset [1, 10)$, e.g. "first digit = 7" means $I = [7, 8)$. We show that the chance X_t falls in that interval converges to what Benford's law predicts.
- In showing the Fourier coefficients vanish, we rely on the Gaussian decay e^{-ct}. If the process had slower mixing or less randomness, one might not get such nice exponential damping, and then convergence could be slower or might fail.
- The same argument covers any finite union of fixed subintervals of $[1, 10)$. Edge cases involving an endpoint equal to 10 should be interpreted with the usual mantissa convention (values lie in $[1, 10)$), and do not change the conclusion because single points carry zero mass under the Uniform$[0, 1)$ limit.

We have shown that GBM follows the Benford distribution when t tends to infinity. Such a formalism may not be sufficient to analyse price paths over shorter periods. To do so, we'll need to quantify the expected error in our observations at smaller values of t.

2.2.1 Quantifying the Error with FRBL

We quantify the rate at which geometric Brownian motion (GBM) attains finite-range Benford behaviour as time $t \to \infty$. Starting from the Fourier representation of the law of the fractional part of the logarithm, we prove an exact Fourier-series expression for the Benford error and derive an exponentially small (Gaussian-in-t) bound with explicit constants.

Setup Let X_t be geometric Brownian motion (GBM)

$$X_t \;=\; X_0 \exp\big((\mu - \tfrac{1}{2}\sigma^2)t + \sigma W_t\big), \qquad \sigma > 0,$$

where $(W_t)_{t\geq 0}$ is a standard Wiener process. Put

$$Y_t := \log_{10} X_t = \log_{10} X_0 + \alpha t + \beta W_t, \qquad \alpha = \frac{\mu - \frac{1}{2}\sigma^2}{\ln 10}, \quad \beta = \frac{\sigma}{\ln 10}.$$

Fix an interval $I = [a, b) \subset [1, 10)$ and let

$$J := \{\log_{10} s : \; s \in I\} = [a', b') = [\log_{10} a, \log_{10} b) \subset [0, 1).$$

Denote by μ_t the distribution of the fractional part $\{Y_t\}$ on the unit circle $\mathbb{R}/\mathbb{Z}$. The quantity of primary interest is the *Benford error*

$$E_t(I) := \Pr\big(M_{10}(X_t) \in I\big) - \log_{10}\frac{b}{a} = \mu_t(J) - |J|, \qquad |J| := \mathrm{Leb}(J) = \log_{10}\frac{b}{a}.$$

We want to analyse the error term $E_t(I)$ for different values of $t \in [1, \infty)$. This will give us a more general understanding of the condition GBM should meet to be considered Benford.

Fourier Representation For each integer $k \in \mathbb{Z}$ define the characteristic (Fourier) coefficient

$$\varphi_t(k) := \mathbb{E}\big[e^{2\pi i k Y_t}\big].$$

Using the Gaussian characteristic function $\mathbb{E}[e^{iuW_t}] = \exp(-\frac{1}{2}u^2 t)$ one obtains, for $k \in \mathbb{Z}$,

$$\varphi_t(k) = e^{2\pi i k(\log_{10} X_0 + \alpha t)} \exp\Big(-\tfrac{1}{2}(2\pi k\beta)^2 t\Big) = e^{2\pi i k(\log_{10} X_0 + \alpha t)} e^{-2\pi^2 k^2 \beta^2 t}.$$

In particular $\varphi_t(0) \equiv 1$ and for $k \neq 0$ we have $\varphi_t(k) \to 0$ as $t \to \infty$.

The indicator 1_J of the interval $J \subset [0, 1)$ has Fourier coefficients

$$\widehat{1_J}(k) = \int_0^1 1_J(x) e^{-2\pi i k x}\, dx = e^{-2\pi i k \frac{a'+b'}{2}} \frac{\sin(\pi k |J|)}{\pi k}, \qquad k \in \mathbb{Z},\; k \neq 0,$$

and $\widehat{1_J}(0) = |J|$.

Since $\varphi_t(k)$ are the Fourier coefficients of μ_t, Parseval/Plancherel-type inversion for measures against L^1 test functions (or Fourier series representation on the circle) yields the following exact identity.

Proposition 17 (Exact Fourier representation of the Benford error) *For every $t > 0$ and every interval $I = [a, b) \subset [1, 10)$,*

$$E_t(I) \;=\; \sum_{k\in\mathbb{Z}\setminus\{0\}} \varphi_t(k)\,\widehat{1_J}(-k) \;=\; \sum_{k\neq 0} e^{2\pi i k(\log_{10} X_0+\alpha t)} e^{-2\pi^2 k^2\beta^2 t}\; e^{2\pi i k\frac{a'+b'}{2}}\,\frac{\sin(\pi k|J|)}{\pi k}. \tag{13}$$

In particular the following (uniform) bound holds:

$$|E_t(I)| \le \sum_{k\neq 0} e^{-2\pi^2 k^2\beta^2 t}\cdot \frac{|\sin(\pi k|J|)|}{\pi|k|}.$$

Proof The identity is a direct application of Fourier inversion on the torus: for any integrable function f on $[0,1)$ and any probability measure μ on the circle with Fourier coefficients $\widehat{\mu}(k)$ one has $\int f\,d\mu = \sum_{k\in\mathbb{Z}} \widehat{f}(-k)\widehat{\mu}(k)$ whenever the Fourier series converges in the appropriate sense. Here $f = 1_J$ and $\widehat{\mu_t}(k) = \varphi_t(k)$. The absolute bound follows by taking absolute values and using the explicit form of $\widehat{1_J}(k)$. ■

From the exact representation (13) we derive two useful explicit bounds.

Exponential (Gaussian-in-t) decay Set $c := 2\pi^2\beta^2 > 0$ (so $c = 2\pi^2\sigma^2/(\ln 10)^2$). Using $|\sin(\pi k|J|)| \le 1$ and pairing k with $-k$ we have

$$|E_t(I)| \le \frac{2}{\pi}\sum_{k=1}^{\infty} \frac{e^{-ck^2t}}{k}.$$

Split the sum into the $k = 1$ term and the tail:

$$\sum_{k=1}^{\infty} \frac{e^{-ck^2t}}{k} = e^{-ct} + \sum_{k=2}^{\infty} \frac{e^{-ck^2t}}{k} \le e^{-ct} + \int_1^{\infty} \frac{e^{-cx^2t}}{x}\,dx.$$

Substitute $u = \sqrt{ct}\,x$ to obtain

$$\int_1^{\infty} \frac{e^{-cx^2t}}{x}\,dx = \int_{\sqrt{ct}}^{\infty} \frac{e^{-u^2}}{u}\,du \le \frac{1}{2ct} e^{-ct},$$

where we used the standard Gaussian tail inequality $\int_y^{\infty} e^{-u^2}\,du \le \frac{1}{2y} e^{-y^2}$ and then integrated by parts. Combining the estimates yields the explicit bound

$$|E_t(I)| \le \frac{2}{\pi}\, e^{-ct}\Big(1 + \frac{1}{2ct}\Big) \;=\; \frac{2}{\pi}\, e^{-2\pi^2\beta^2 t}\Big(1 + \frac{1}{4\pi^2\beta^2 t}\Big).$$

Written in terms of the original volatility σ,

$$|E_t(I)| \le \frac{2}{\pi}\exp\Big(-\frac{2\pi^2\sigma^2}{(\ln 10)^2}t\Big)\Big(1 + \frac{(\ln 10)^2}{4\pi^2\sigma^2 t}\Big). \tag{14}$$

Remark 18 *This bound shows* exponential *decay of the Benford error as $t \to \infty$, with exact exponential rate $2\pi^2\beta^2$. The prefactor $2/\pi$ and the $1/t$ correction are explicit and small for moderate/large t. Indeed as $t \to \infty$ the $\exp(-t)$ terms tend to zero and the distribution converges to BL.*

2.2.2 Visualising the Error with FRBL

We have an upper bound for the expected error for GBM at some value of t in the time evolution. We can visualise how these error terms change as t increases via a Monte Carlo simulation[16; 17].

Monte Carlo methods are techniques which use repeated random sampling to obtain numerical results. They are generally used to model systems or phenomena that have a significant degree of uncertainty in their inputs or are too complex to be solved analytically. By aggregating the results of a set of simulations, the distribution of possible outcomes can be analysed, giving a statistically significant account of the modelled phenomena.

GBM falls within this category as a model of time series that has Gaussian noise added by construction. We can therefore simulate various GBM time series with different parameters to understand how errors might propagate as t increases. This will also confirm whether our analytic error bound given in (14) is correct. We compared the analytic error estimate with Monte Carlo simulations of GBM.

Monte Carlo Simulations: We fixed $X_0 = 1$, $\mu = 0.05$, $\sigma = 0.3$, and considered the interval $I = [1, 2)$ with theoretical mass $\log_{10} 2$. At each time t we sampled directly from the marginal distribution

$$Y_t \sim \mathcal{N}\big(\log_{10} X_0 + \alpha t,\ \beta^2 t\big), \qquad \alpha = \tfrac{\mu - \frac{1}{2}\sigma^2}{\ln 10}, \quad \beta = \tfrac{\sigma}{\ln 10},$$

and estimated the Benford error by

$$\widehat{E}_t(I) \ = \ \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{\{Y_t^{(j)}\} \in J\}} - |J|.$$

We generated ten independent batches with $n = 250{,}000$ samples each[7], and plotted the empirical absolute errors $|\widehat{E}_t(I)|$ against time, alongside the analytic Fourier–series value and the Gaussian–in–t upper bound. For each t we computed the groupwise signed errors and their mean.

Figure 4 shows the results of this simulation.

The dominant source of discrepancy between the analytic Benford error $E_t(I)$ and the Monte–Carlo curves is sampling variability of the binomial estimator. Let $p_t = \Pr(M_{10}(X_t) \in I)$ be the true probability and write the Monte–Carlo estimator $\widehat{p}_t = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{\{Y_t^{(j)}\} \in J\}}$. Then

$$(\widehat{p}_t) = \frac{p_t(1 - p_t)}{n}, \qquad \mathrm{sd}(\widehat{p}_t) = \sqrt{\frac{p_t(1 - p_t)}{n}}.$$

For our interval $I = [1, 2)$ one has $p_t \approx |J| = \log_{10} 2 \approx 0.30103$, hence $p_t(1 - p_t) \approx 0.30103 \cdot 0.69897 \approx 0.21$. Requiring a target pointwise standard error b leads to the sample–size rule

$$n \ \gtrsim \ \frac{p_t(1 - p_t)}{b^2} \approx \frac{0.21}{b^2}.$$

7. The standard error scales as $n^{-1/2}$, so further reductions require quadratically more samples. Hence $n = 250{,}000$ is a practical compromise between computational cost and the desired precision (it matches the natural target $b \approx 10^{-3}$ set by the analytic decay in our time window). If one needs reliable pointwise estimates below 10^{-4} one must increase n by another two orders of magnitude or use variance–reduction / pooled estimators.

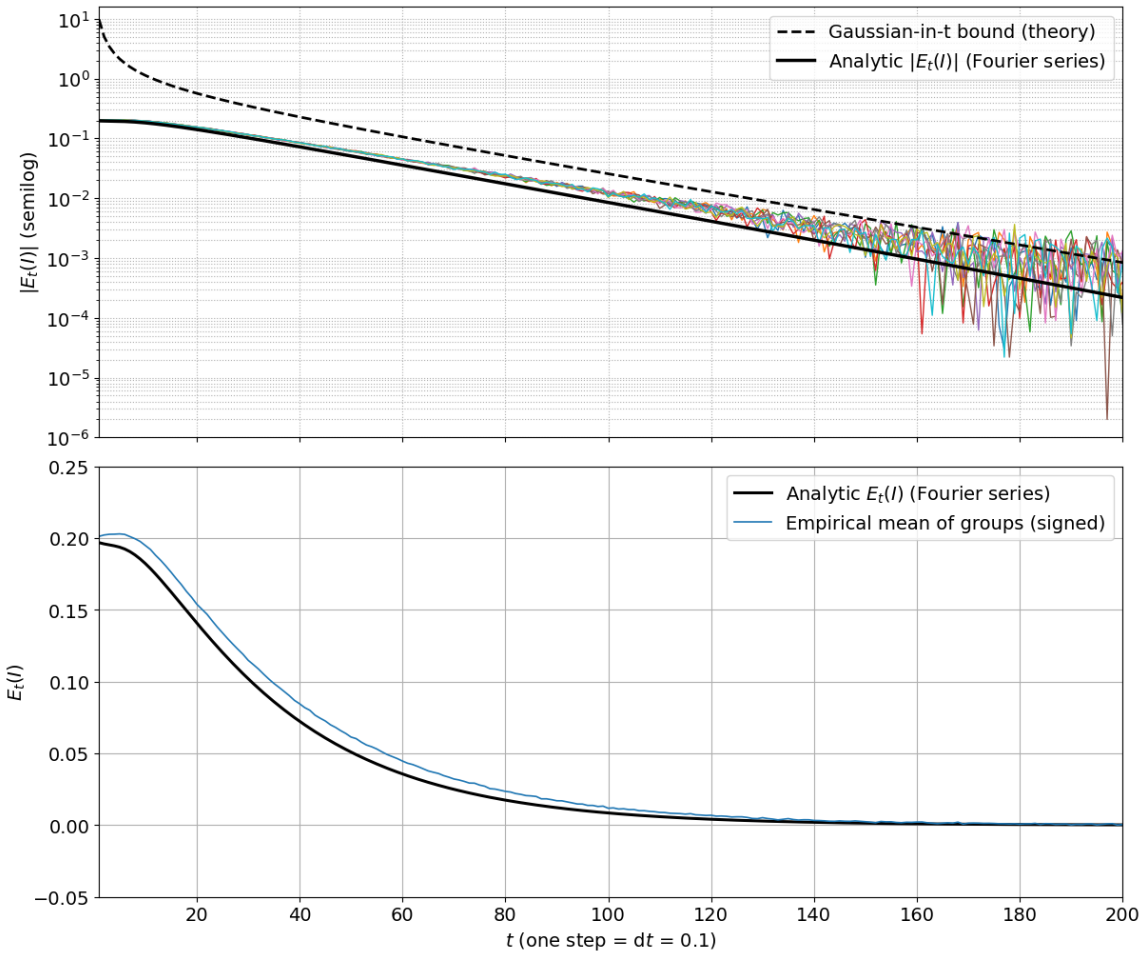

Figure 4: Comparison of analytic and Monte–Carlo estimates of the Benford error for GBM with $X_0 = 1$, $\mu = 0.05$, $\sigma = 0.3$, and $I = [1, 2)$. The top figure shows the absolute error $|E_t(I)|$ on a semilog scale for ten independent batches (coloured), the analytic Fourier–series value (black solid), and the Gaussian–in–t bound (black dashed). The analytic error decays exponentially in t, while the empirical curves saturate near the Monte–Carlo noise floor once the true error is $\ll n^{-1/2}$. The bottom figure shows the signed error $E_t(I)$, showing the analytic Fourier prediction (black) together with the empirical mean across the same ten batches of $n = 250{,}000$ samples each (blue). These results highlight that (i) the true Benford error decays exponentially fast with oscillatory corrections, and (ii) sufficiently large sample sizes are required to resolve the analytic behaviour below the natural Monte–Carlo variability.

In particular, to reduce the pointwise sampling noise to the level $b = 10^{-3}$ (i.e. to make the sampling sd comparable to analytic errors of order 10^{-3}) one needs

$$n \gtrsim \frac{0.21}{(10^{-3})^2} \approx 2.1 \times 10^5.$$

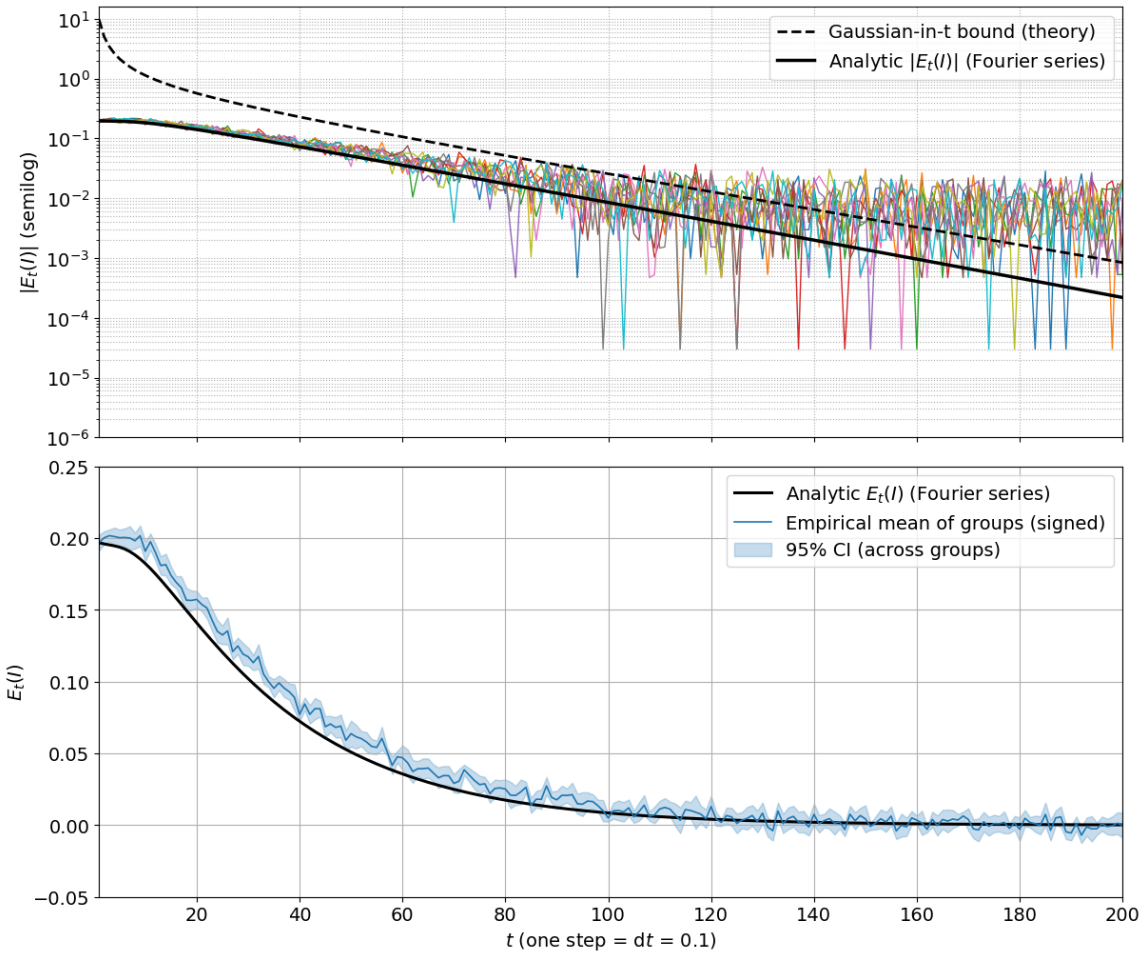

Figure 5: A near identical treatment as in Figure 4, except $n = 2000$. On the lower figure, we have shown the 95% confidence range bands for the Monte-Carlo noise, shown as a light blue shaded area, indicating that our error bound is within a reasonable confidence window. The Monte–Carlo noise floor becomes significant at lower values of t in this simulation, which is expected.

Thus $n = 250{,}000$ slightly exceeds this threshold and yields

$$\mathrm{sd}(\widehat{p}_t) \approx \sqrt{\frac{0.21}{250{,}000}} \approx 9.2 \times 10^{-4},$$

so typical Monte–Carlo fluctuations are $\mathcal{O}(10^{-3})$. At a lower value of n this becomes clear. Figure 5 shows a Monte–Carlo simulation for $n = 2000$, including the 95% Monte–Carlo confidence intervals for the signed error. At this lower sampling rate, we still see agreement between our analytical estimate and the observed error rate in the simulation, as evidenced by the confidence intervals.

This explains two important observations:

1. With $n = 2{,}000$ the sampling sd is $\approx 1.0 \times 10^{-2}$, much larger than analytic errors once t is moderate, hence the empirical $|\widehat{E}_t|$ routinely lies above the analytic curve.

2. With $n = 250{,}000$ the sampling sd is reduced to $\sim 10^{-3}$, so the empirical mean (or the pooled estimator using all samples) tracks the analytic signed error down to that noise floor; the analytic curve then lies inside the Monte–Carlo confidence intervals.

Appendix A. Test Appendix

This is a test appendix

References

[1] L. Wang and B.-Q. Ma, "A concise proof of benford's law," *Fundamental Research*, vol. 4, no. 4, pp. 841–844, 2023. [Online]. Available: https://doi.org/10.1016/j.fmre.2023.01.002

[2] M. Sambridge, H. Tkalčić, P. Arroucau *et al.*, "Benford's law of first digits: from mathematical curiosity to change detector," *Asia Pacific Mathematics Newsletter*, vol. 1, no. 4, pp. 1–6, 2011.

[3] J. Morrow, "Benford's law, families of distributions and a test basis," *Centre for Economic Performance, London School of Economics and Political Science*, 2014, url: http://eprints.lse.ac.uk/60364/1/dp1291.pdf.

[4] B. Hull, "Can benford's law be used to detect financial fraud?" 2021, level 4 Project, MSci Natural Sciences, Department of Physics, Durham University. [Online]. Available: https://bluehood.github.io/research/benh_benfords-law-financial-fraud_2021.pdf

[5] F. Black and M. S. Scholes, "The pricing of options and corporate liabilities," *Journal of Political Economy*, vol. 81, no. 3, pp. 637–654, 1973.

[6] R. C. Merton, "Theory of rational option pricing," *The Bell Journal of Economics and Management Science*, vol. 4, no. 1, pp. 141–183, 1973. [Online]. Available: https://www.jstor.org/stable/3003143

[7] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed., ser. Graduate Texts in Mathematics. Springer, 1991, vol. 113.

[8] B. Ø ksendal, *Stochastic Differential Equations: An Introduction with Applications*, 6th ed. Springer, 2003.

[9] K. Itô, "On stochastic differential equations," *Memoirs of the American Mathematical Society*, no. 4, pp. 1–51, 1951.

[10] S. L. Heston, "A closed-form solution for options with stochastic volatility," *The Review of Financial Studies*, vol. 6, no. 2, pp. 327–343, 1993.

[11] A. V. Kontorovich and S. J. Miller, "Benford's law, values of l-functions and the 3x+1 problem," *Acta Arithmetica*, vol. 120, no. 3, pp. 269–297, 2005. [Online]. Available: https://arxiv.org/abs/math/0412003

[12] S. J. Miller, "Benford's law: Theory and applications," in *Benford's Law: Theory and Applications*. Princeton University Press, 2015, edited volume; see in particular the chapters surveying stochastic/dynamical examples. [Online]. Available: https://press.princeton.edu/books/hardcover/9780691147611/benfords-law

[13] A. Berger and T. P. Hill, "A basic theory of benford's law," *Probability Surveys*, vol. 8, pp. 1–126, 2011. [Online]. Available: https://projecteuclid.org/euclid.ps/1315598749

[14] ——, "The mathematics of benford's law—a primer," *arXiv preprint*, 2019, accessible survey/primer; see arXiv version. [Online]. Available: https://arxiv.org/abs/1909.07527

[15] M. Ausloos, V. Ficcadenti, G. Dhesi, and M. Shakeel, "Benford's laws tests on s&p500 daily closing values and the corresponding daily log-returns both point to huge non-conformity," *Physica A: Statistical Mechanics and its Applications*, vol. 574, p. 125969, 2021. [Online]. Available: https://arxiv.org/abs/2104.07962

[16] P. Glassermann, *Monte Carlo Methods in Financial Engineering*, ser. Stochastic Modelling and Applied Probability. Springer, 2003, vol. 53, covers Monte Carlo simulation schemes, including Euler-Maruyama, Milstein, and applications to GBM.

[17] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed., ser. Springer Texts in Statistics. New York: Springer, 2004, comprehensive reference for Monte Carlo simulation concepts, theory, and methods. [Online]. Available: https://www.springer.com/gp/book/9781441915932