Hindawi Publishing Corporation
International Journal of Mathematics and Mathematical Sciences
Volume 2008, Article ID 382948, 19 pages
doi:10.1155/2008/382948

Research Article

Order Statistics and Benford's Law

Steven J. Miller[1] and Mark J. Nigrini[2]

[1] *Department of Mathematics and Statistics, Williams College, Williamstown, MA 01267, USA*
[2] *Accounting and Information Systems, School of Business, The College of New Jersey, Ewing, NJ 08628, USA*

Correspondence should be addressed to Steven J. Miller, steven.j.miller@williams.edu

Received 2 June 2008; Revised 6 September 2008; Accepted 13 October 2008

Recommended by Jewgeni Dshalalow

Fix a base $B > 1$ and let ζ have the standard exponential distribution; the distribution of digits of ζ base B is known to be very close to Benford's law. If there exists a C such that the distribution of digits of C times the elements of some set is the same as that of ζ, we say that set exhibits shifted exponential behavior base B. Let $X_1, \ldots, X_N$ be i.i.d.r.v. If the X_i's are Unif, then as $N \to \infty$ the distribution of the digits of the differences between adjacent order statistics converges to shifted exponential behavior. If instead X_i's come from a compactly supported distribution with uniformly bounded first and second derivatives and a second-order Taylor series expansion at each point, then the distribution of digits of any N^δ consecutive differences *and* all $N - 1$ normalized differences of the order statistics exhibit shifted exponential behavior. We derive conditions on the probability density which determine whether or not the distribution of the digits of all the unnormalized differences converges to Benford's law, shifted exponential behavior, or oscillates between the two, and show that the Pareto distribution leads to oscillating behavior.

1. Introduction

Benford's law gives the expected frequencies of the digits in many tabulated data. It was first observed by Newcomb in the 1880s, who noticed that pages of numbers starting with a 1 in logarithm tables were significantly more worn than those starting with a 9. In 1938 Benford [1] observed the same digit bias in a variety of phenomena. From his observations, he postulated that in many datasets, more numbers began with a 1 than with a 9; his investigations (with 20,229 observations) supported his belief. See [2, 3] for a description and history, and [4] for an extensive bibliography.

For any base $B > 1$, we may uniquely write a positive $x \in \mathbb{R}$ as $x = M_B(x) \cdot B^k$, where $k \in \mathbb{Z}$ and $M_B(x)$ (called the mantissa) is in $[1, B)$. A sequence of positive numbers $\{a_n\}$ is *Benford base* B if the probability of observing a mantissa of a_n base B of at most s is $\log_B s$.

More precisely, for $s \in [1, B]$, we have

$$\lim_{N\to\infty} \frac{\#\{n \le N : 1 \le M_B(\alpha_n) \le s\}}{N} = \log_B s. \tag{1.1}$$

Benford behavior for continuous functions is defined analogously. (If the functions are not positive, we study the distribution of the digits of the absolute value of the function.) Thus, working base 10 we find the probability of observing a first the probability of observing a first digit of d is $\log_{10}(d+1) - \log_{10}(d)$, implying that about 30% of the time the first digit is a 1.

We can prove many mathematical systems follow Benford's law, ranging from recurrence relations [5] to $n!$ [6] to iterates of power, exponential and rational maps, as well as Newton's method [7–9]; to chains of random variables and hierarchical Bayesian models [10]; to values of L-functions near the critical line; to characteristic polynomials of random matrix ensembles and iterates of the $3x + 1$-Map [11, 12]; as well as to products of random variables [13]. We also see Benford's law in a variety of natural systems, such as atomic physics [14], biology [15], and geology [16]. Applications of Benford's law range from rounding errors in computer calculations (see [17, page 255]) to detecting tax (see [18, 19]) and voter fraud (see [20]).

This work is motivated by two observations (see Remark 1.9 for more details). First, since Benford's seminal paper, many investigations have shown that amalgamating data from different sources leads to Benford behavior; second, many standard probability distributions are close to Benford behavior. We investigate the distribution of digits of differences of adjacent ordered random variables. For any $\delta < 1$, if we study at most N^δ consecutive differences of a dataset of size N, the resulting distribution of leading digits depends very weakly on the underlying distribution of the data, and closely approximates Benford's law. We then investigate whether or not studying all the differences leads to Benford behavior; this question is inspired by the first observation above, and has led to new tests for data integrity (see [21]). These tests are quick and easy-to-apply, and have successfully detected problems with some datasets, thus providing a practical application of our main results.

Proving our results requires analyzing the distribution of digits of independent random variables drawn from the standard exponential, and quantifying how close the distribution of digits of a random variable with the standard exponential distribution is to Benford's law. Leemis et al. [22] have observed that the standard exponential is quite close to Benford's law; this was proved by Engel and Leuenberger [23], who showed that the maximum difference in the cumulative distribution function from Benford's law (base 10) is at least .029 and at most .03. We provide an alternate proof of this result in the appendix using a different technique, as well as showing that there is no base B such that the standard exponential distribution is Benford base B (Corollary A.2).

Both proofs apply Fourier analysis to periodic functions. In [23, equation (5)], the main step is interchanging an integration and a limit. Our proof is based on applying Poisson summation to the derivative of the cumulative distribution function of the logarithms modulo 1, F_B. Benford's law is equivalent to $F_B(b) = b$, which by calculus is the same as $F_B'(b) = 1$ and $F_B(0) = 0$. Thus, studying the deviation of $F_B'(b)$ from 1 is a natural way to investigate the deviations from Benford behavior. We hope the details of these calculations may be of use to others in investigating related problems (Poisson summation has been fruitfully used by Kontorovich and Miller [11] and Jang et al. [10] in proving many systems are Benford's; see also [24]).

1.1. Definitions

A sequence $\{a_n\}_{n=1}^{\infty} \subset [0,1]$ is equidistributed if

$$\lim_{N\to\infty} \frac{\#\{n : n \le N, a_n \in [a,b]\}}{N} = b - a \quad \forall [a,b] \subset [0,1]. \tag{1.2}$$

Similarly, a continuous random variable on $[0,\infty)$, whose probability density function is p, is equidistributed modulo 1 if

$$\lim_{T\to\infty} \frac{\int_0^T \chi_{a,b}(x)p(x)dx}{\int_0^T p(x)dx} = b - a, \tag{1.3}$$

for any $[a,b] \subset [0,1]$, where $\chi_{a,b}(x) = 1$ for $x \bmod 1 \in [a,b]$ and 0 otherwise.

A positive sequence (or values of a function) is Benford base B if and only if its base B logarithms are equidistributed modulo 1; this equivalence is at the heart of many investigations of Benford's law (see [6, 25] for a proof).

We use the following notations for the various error terms.

(1) Let $\mathcal{E}(x)$ denote an error of at most x in absolute value; thus $f(b) = g(b) + \mathcal{E}(x)$ means $|f(b) - g(b)| \le x$.

(2) Big-Oh notation: for $g(x)$ a nonnegative function, we say $f(x) = O(g(x))$ if there exist an x_0 and a $C > 0$ such that, for all $x \ge x_0$, $|f(x)| \le Cg(x)$.

The following theorem is the starting point for investigating the distribution of digits of order statistics.

Theorem 1.1. *Let ζ have the standard (unit) exponential distribution*

$$\text{Prob}(\zeta \in [\alpha,\beta]) = \int_\alpha^\beta e^{-t}dt, \quad [\alpha,\beta] \in [0,\infty). \tag{1.4}$$

For $b \in [0,1]$, let $F_B(b)$ be the cumulative distribution function of $\log_B \zeta \bmod 1$; thus $F_B(b) := \text{Prob}(\log_B \zeta \bmod 1 \in [0,b])$. Then, for all $M \ge 2$,

$$\begin{aligned} F_B'(b) &= 1 + 2\sum_{m=1}^{\infty} \text{Re}\left(e^{-2\pi i m b}\Gamma\left(1 + \frac{2\pi i m}{\log B}\right)\right) \\ &= 1 + 2\sum_{m=1}^{M-1} \text{Re}\left(e^{-2\pi i m b}\Gamma\left(1 + \frac{2\pi i m}{\log B}\right)\right) + \mathcal{E}\left(4\sqrt{2}\pi c_1(B)e^{-(\pi^2 - c_2(B))M/\log B}\right), \end{aligned} \tag{1.5}$$

where $c_1(B)$, $c_2(B)$ are constants such that for all $m \geq M \geq 2$, one has

$$\begin{aligned} e^{2\pi^2 m/\log B} - e^{-2\pi^2 m/\log B} &\geq \frac{e^{2\pi^2 m/\log B}}{c_1^2(B)}, \\ \frac{m}{\log B} &\leq e^{2c_2(B)m/\log B}, \\ 1 - e^{(\pi^2 - c_2(B))M/\log B} &\geq \frac{1}{\sqrt{2}}. \end{aligned} \tag{1.6}$$

For $B \in [e, 10]$, take $c_1(B) = \sqrt{2}$ and $c_2(B) = 1/5$, which give

$$\begin{aligned} &\operatorname{Prob}\left(\log \zeta \bmod 1 \in [a,b]\right) \\ &\quad = b - a + \frac{2r}{\pi} \cdot \sin\left(\pi(b+a) + \theta\right) \cdot \sin\left(\pi(b-a)\right) + \mathcal{E}\left(6.32 \cdot 10^{-7}\right), \end{aligned} \tag{1.7}$$

with $r \approx 0.000324986$, $\theta \approx 1.32427186$, and

$$\begin{aligned} \operatorname{Prob}\left(\log_{10} \zeta \bmod 1 \in [a,b]\right) &= b - a + \frac{2r_1}{\pi} \sin\left(\pi(b+a) - \theta_1\right) \cdot \sin\left(\pi(b-a)\right) \\ &\quad - \frac{r_2}{\pi} \sin\left(2\pi(b+a) + \theta_2\right) \cdot \sin\left(2\pi(b-a)\right) + \mathcal{E}\left(8.5 \cdot 10^{-5}\right) \end{aligned} \tag{1.8}$$

with

$$\begin{aligned} r_1 &\approx 0.0569573, \qquad \theta_1 \approx 0.8055888, \\ r_2 &\approx 0.0011080, \qquad \theta_2 \approx 0.1384410. \end{aligned} \tag{1.9}$$

The above theorem was proved in [23]; we provide an alternate proof in Appendix A. As remarked earlier, our technique consists of applying Poisson summation to the derivative of the cumulative distribution function of the logarithms modulo 1; it is then very natural and easy to compare deviations from the resulting distribution and the uniform distribution (if a dataset satisfies Benford's law, then the distribution of its logarithms is uniform). Our series expansions are obtained by applying properties of the Gamma function.

Definition 1.2 (Definition exponential behavior, shifted exponential behavior). Let ζ have the standard exponential distribution, and fix a base B. If the distribution of the digits of a set is the same as the distribution of the digits of ζ, then one says that the set exhibits exponential behavior (base B). If there is a constant $C > 0$ such that the distribution of digits of all elements multiplied by C is exponential behavior, then one says that the system exhibits shifted exponential behavior (with shift of $\log_B C \bmod 1$).

We briefly describe the reasons behind this notation. One important property of Benford's law is that it is invariant under rescaling; many authors have used this property to characterize Benford behavior. Thus, if a dataset is Benford base B, and we fix a positive number C, so is the dataset obtained by multiplying each element by C. This is clear if, instead of looking at the distribution of the digits, we study the distribution of the base B

logarithms modulo 1. Benford's law is equivalent to the logarithms modulo 1 being uniformly distributed (see, e.g., [6, 25]); the effect of multiplying all entries by a fixed constant simply translates the uniform distribution modulo 1, which is again the uniform distribution.

The situation is different for exponential behavior. Multiplying all elements by a fixed constant C (where $C \neq B^k$ for some $k \in \mathbb{Z}$) does not preserve exponential behavior; however, the effect is easy-to-describe. Again looking at the logarithms, exponential behavior is equivalent to the base B logarithms modulo 1 having a specific distribution which is almost equal to the uniform distribution (at least if the base B is not too large). Multiplying by a fixed constant $C \neq B^k$ shifts the logarithm distribution by $\log_B C$ mod 1.

1.2. Results for differences of orders statistics

We consider a simple case first, and show how the more general case follows. Let $X_1, \ldots, X_N$ be independent identically distributed from the uniform distribution on $[0, L]$. We consider L fixed and study the limit as $N \to \infty$. Let $X_{1:N}, \ldots, X_{N:N}$ be the X_i's in increasing order. The $X_{i:N}$ are called the order statistics, and satisfy $0 \leq X_{1:N} \leq X_{2:N} \leq \cdots \leq X_{N:N} \leq L$. We investigate the distribution of the leading digits of the differences between adjacent $X_{i:N}$'s, $X_{i+1:N} - X_{i:N}$. For convenience, we periodically continue the data and set $X_{i+N:N} = X_{i:N} + L$. As we have N differences in an interval of size L, the average value of $X_{i+1:N} - X_{i:N}$ is of size L/N, and it is sometimes easier to study the normalized differences

$$Z_{i;N} = \frac{X_{i+1:N} - X_{i:N}}{L/N}. \tag{1.10}$$

As the X_i's are drawn from a uniform distribution, it is a standard result that as $N \to \infty$, the $Z_{i;N}$'s are independent random variables, each having the standard exponential distribution. Thus, as $N \to \infty$, the probability that $Z_{i;N} \in [a, b]$ tends to $\int_a^b e^{-t} dt$ (see [26, 27] for proofs).

For uniformly distributed random variables, if we know the distribution of $\log_B Z_{i;N}$ mod 1, then we can immediately determine the distribution of the digits of the $X_{i+1:N} - X_{i:N}$ base B because

$$\log_B Z_{i;N} = \log_B \left(\frac{X_{i+1:N} - X_{i:N}}{L/N} \right) = \log_B (X_{i+1:N} - X_{i:N}) - \log_B \left(\frac{L}{N} \right). \tag{1.11}$$

As $Z_{i;N}$ are independent with the standard exponential distribution as $N \to \infty$; if X_i are independent uniformly distributed, the behavior of the digits of the differences $X_{i+1:N} - X_{i:N}$ is an immediate consequence of Theorem 1.1.

Theorem 1.3 (Shifted exponential behavior of differences of independent uniformly distributed random variables). *Let $X_1, \ldots, X_N$ be independently distributed from the uniform distribution on $[0, L]$, and let $X_{1:N}, \ldots, X_{N:N}$ be X_i's in an increasing order. As $N \to \infty$, the distribution of the digits (base B) of the differences $X_{i+1:N} - X_{i:N}$ converges to shifted exponential behavior, with a shift of $\log_B(L/N)$ mod 1.*

A similar result holds for other distributions.

Theorem 1.4 (Shifted exponential behavior of subsets of differences of independent random variables). *Let $X_1, \dots, X_N$ be independent, identically distributed random variables whose density $f(x)$ has a second-order Taylor series at each point with first and second derivatives uniformly bounded, and let the $X_{i:N}$'s be the X_i's in increasing order. Fix a $\delta \in (0,1)$. Then, as $N \to \infty$ the distribution of the digits (base B) of N^δ consecutive differences $X_{i+1:N} - X_{i:N}$ converges to shifted exponential behavior, provided that $X_{i:N}$'s are from a region where $f(x)$ is nonzero.*

The key ingredient in this generalization is that the techniques, which show that the differences between uniformly distributed random variables become independent exponentially distributed random variables, can be modified to handle more general distributions.

We restricted ourselves to a subset of all consecutive spacings because the normalization factor changes throughout the domain. The shift in the shifted exponential behavior depends on which set of N^δ differences we study, coming from the variations in the normalizing factors. Within a bin of N^δ differences, the normalization factor is basically constant, and we may approximate our density with a uniform distribution. It is possible for these variations to cancel and yield Benford behavior for the digits of all the unnormalized differences. Such a result is consistent with the belief that amalgamation of data from many different distributions becomes Benford; however, this is not always the case (see Remark 1.6). From Theorems 1.1 and 1.4, we obtain the following theorem.

Theorem 1.5 (Benford behavior for all the differences of independent random variables). *Let $X_1, \dots, X_N$ be independent identically distributed random variables whose density $f(x)$ is compactly supported and has a second-order Taylor series at each point with first and second derivatives uniformly bounded. Let the $X_{i:N}$'s be the X_i's in an increasing order $F(x)$ be the cumulative distribution function for $f(x)$, and fix a $\delta \in (0,1)$. Let $I(\epsilon, \delta, N) = [\epsilon N^{1-\delta}, N^{1-\delta} - \epsilon N^{1-\delta}]$. For each fixed $\epsilon \in (0, 1/2)$, assume that*

(i) *$f(F^{-1}(kN^{\delta-1})$ is not too small for $k \in I(\epsilon, \delta, N)$*

$$\lim_{N\to\infty} \max_{k \in I(\epsilon,\delta,N)} \frac{\min(N^{-(\epsilon+\delta/2)}, N^{\delta-1})}{f(F^{-1}(kN^{\delta-1}))} = 0, \tag{1.12}$$

(ii) $\log_B f(F^{-1}(kN^{\delta-1}) \bmod 1$ *is equidistributed: for all $[\alpha, \beta] \subset [0,1]$*

$$\lim_{N\to\infty} \frac{\#\{k \in I(\epsilon,\delta,N) : \log_B f(F^{-1}(kN^{\delta-1})) \bmod 1 \in [\alpha,\beta]\}}{N^\delta} = \beta - \alpha. \tag{1.13}$$

Then, if $\epsilon > \max(0, 1/3 - \delta/2)$ and $\epsilon < \delta/2$, the distribution of the digits of the $N-1$ differences $X_{i+1:N} - X_{i:N}$ converges to Benford's law (base B) as $N \to \infty$.

Remark 1.6. The conditions of Theorem 1.5 are usually not satisfied. We are unaware of any situation where (1.13) holds; we have included Theorem 1.5 to give a sufficient condition of what is required to have Benford's law satisfied *exactly*, and not just approximately. In Lemma 3.3, we show with: Example 3.3 shows that the conditions fail for the Pareto distribution, and the limiting behavior oscillates between Benford and a sum of shifted exponential behavior. (If several datasets each exhibit shifted exponential behavior but with distinct shifts, then

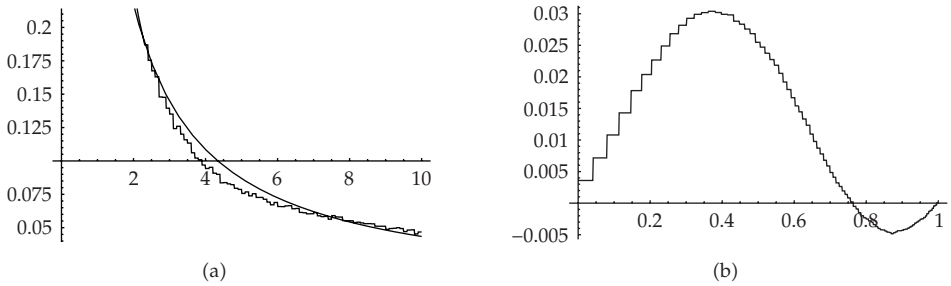

Figure 1: All 499 999 differences of adjacent order statistics from 500 000 independent random variables from the Pareto distribution with minimum value and variance 1. (a) Observed digits of scaled differences of adjacent random variables versus Benford's law; (b) scaled observed minus Benford's law (cumulative distribution of base 10 logarithms).

the amalgamated dataset is closer to Benford's law than any of the original datasets. This is apparent by studying the logarithms modulo 1. The differences between these densities and Benford's law will look like Figure 1(b) (except, of course, that different shifts will result in shifting the plot modulo 1). The key observation is that the unequal shifts mean that we do not have reinforcements from the peaks of modulo 1 densities being aligned, and thus the amalgamation will decrease the maximum deviations.) The arguments generalize to many densities whose cumulative distribution functions have tractable closed-form expressions (e.g., exponential, Weibull, or $f(x) = e^{-e^x}e^x$).

The situation is very different if instead we study normalized differences:

$$\widetilde{Z}_{i:N} = \frac{X_{i+1:N} - X_{i:N}}{1/Nf(X_{i:N})}, \tag{1.14}$$

note if $f(x) = 1/L$ is the uniform distribution on $[0, L]$, (1.14) reduces to (1.10).

Theorem 1.7 (Shifted exponential behavior for all the normalized differences of independent random variables). *Assume the probability distribution f satisfies the conditions of Theorem 1.5 and* (1.12) *and $\widetilde{Z}_{i;N}$ is as in* (1.14). *Then, as $N \to \infty$, the distribution of the digits of $\widetilde{Z}_{i:N}$ converges to shifted exponential behavior.*

Remark 1.8. Appropriately scaled, the distribution of the digits of the differences is universal, and is the exponential behavior of Theorem 1.1. Thus, Theorem 1.7 implies that the natural quantity to study is the normalized differences of the order statistics, not the differences (see also Remark 3.5). With additional work, we could study densities with unbounded support and show that, through truncation, we can get arbitrarily close to shifted exponential behavior.

Remark 1.9. The main motivation for this work is the need for improved ways of assessing the authenticity and integrity of scientific and corporate data. Benford's law has been successfully applied to detecting income tax, corporate, and voter fraud (see [18–20]); in [21], we use these results to derive new statistical tests to examine data authenticity and integrity. Early applications of these tests to financial data showed that it could detect errors in data downloads, rounded data, and inaccurate ordering of data. These attributes are not

easily observable from an analysis of descriptive statistics, and detecting these errors can help managers avoid costly decisions based on erroneous data.

The paper is organized as follows. We prove Theorem 1.1 in Appendix A by using Poisson summation to analyze $F'_B(b)$. Theorem 1.3 follows from the results of the order statistics of independent uniform variables. The proof of Theorem 1.4 is similar, and given in Section 2. In Section 3, we prove Theorems 1.5 and 1.7.

2. Proofs of Theorems 1.3 and 1.4

Theorem 1.3 is a consequence of the fact that the normalized differences between the order statistics drawn from the uniform distribution converge to being independent standard exponentials. The proof of Theorem 1.4 proceeds similarly. Specifically, over a short enough region, any distribution with a second-order Taylor series at each point with first and second derivatives uniformly bounded is well approximated by a uniform distribution.

To prove Theorem 1.4, it suffices to show that if $X_1, \dots, X_N$ are drawn from a sufficiently nice distribution, then for any fixed $\delta \in (0,1)$ the limiting behavior of the order statistics of N^δ adjacent X_i's becomes Poissonian (i.e., the $N^\delta - 1$ normalized differences converge to being independently distributed from the standard exponential). We prove this below for compactly supported distributions $f(x)$ that have a second-order Taylor series at each point with the first and second derivatives uniformly bounded, and when the N^δ adjacent X_i's are from a region where $f(x)$ is bounded away from zero.

For each N, consider intervals $[a_N, b_N]$ such that $\int_{a_N}^{b_N} f(x)dx = N^\delta/N$; thus, the proportion of the total mass in such intervals is $N^{\delta-1}$. We fix such an interval for our arguments. For each $i \in \{1, \dots, N\}$, let

$$w_i = \begin{cases} 1, & \text{if } X_i \in [a_N, b_N] \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

Note w_i is 1 with probability $N^{\delta-1}$ and 0 with probability $1 - N^{\delta-1}$; w_i is a binary indicator random variable, telling us whether or not $X_i \in [a_N, b_N]$. Thus,

$$\mathbb{E}\left[\sum_{i=1}^{N} w_i\right] = N^\delta, \qquad \mathrm{Var}\left(\sum_{i=1}^{N} w_i\right) = N^\delta \cdot \left(1 - N^{\delta-1}\right). \tag{2.2}$$

Let M_N be the number of X_i in $[a_N, b_N]$, and let β_N be *any* nondecreasing sequence tending to infinity (in the course of the proof, we will find that we may take any sequence with $\beta_N = O(N^{\delta/2})$). By (2.2) and the central limit theorem (which we may use as w_i's satisfy the Lyapunov condition), with probability tending to 1, we have

$$M_N = N^\delta + O\left(\beta_N N^{\delta/2}\right). \tag{2.3}$$

We assume that in the interval $[a_N, b_N]$ there exist constants c and C such that whenever $x \in [a_N, b_N]$, $0 < c < f(x) < C < \infty$; we assume that these constants hold for all regions investigated and for all N. (If our distribution has unbounded support, for any

$\epsilon > 0$, we can truncate it on both sides so that the omitted probability is at most ϵ. Our result is then trivially modified to be within ϵ of shifted exponential behavior.) Thus,

$$c \cdot (b_N - a_N) \le \int_{a_N}^{b_N} f(x)dx = N^{\delta-1} \le C(b_N - a_N), \tag{2.4}$$

implying that $b_N - a_N$ is of size $N^{\delta-1}$. If we assume that $f(x)$ has at least a second-order Taylor expansion, then

$$\begin{aligned} f(x) &= f(a_N) + f'(a_N)(x - a_N) + O((x - a_N)^2) \\ &= f(a_N) + f'(a_N)(x - a_N) + O(N^{2\delta-2}). \end{aligned} \tag{2.5}$$

As we assume that the first and second derivatives are uniformly bounded, as well as f being bounded away from zero in the intervals under consideration, all Big-Oh constants below are independent of N. Thus,

$$b_N - a_N = \frac{N^{\delta-1}}{f(a_N)} + O(N^{2\delta-2}). \tag{2.6}$$

We now investigate the order statistics of the M_N of the X_i's that lie in $[a_N, b_N]$. We know $\int_{a_N}^{b_N} f(x)dx = N^{\delta-1}$; by setting $g_N(x) = f(x)N^{1-\delta}$, then $g_N(x)$ is the conditional density function for X_i, given that $X_i \in [a_N, b_N]$. Thus, $g_N(x)$ integrates to 1, and for $x \in [a_N, b_N]$, we have

$$g_N(x) = f(a_N) \cdot N^{1-\delta} + f'(a_N)(x - a_N) \cdot N^{1-\delta} + O(N^{\delta-1}). \tag{2.7}$$

We have an interval of size $N^{\delta-1}/f(a_N) + O(N^{2\delta-2})$, and $M_N = N^\delta + O(\beta_N N^{\delta/2})$ of the X_i lying in the interval (remember that β_N are any nondecreasing sequence tending to infinity). Thus, with probability tending to 1, the average spacing between adjacent ordered X_i is

$$\frac{N^{\delta-1}/f(a_N) + O(N^{2\delta-2})}{M_N} = (f(a_N)N)^{-1} + N^{-1} \cdot O(\beta_N N^{-\delta/2} + N^{\delta-1}), \tag{2.8}$$

in particular, we see that we must choose $\beta_N = O(N^{\delta/2})$. As $\delta \in (0, 1)$, if we fix a k such that $X_k \in [a_N, b_N]$, then we expect the next X_i to the right of X_k to be about $t/Nf(a_N)$ units away, where t is of size 1. For a given X_k, we can compute the conditional probability that the next X_i is between $t/Nf(a_N)$ and $(t + \Delta t)/Nf(a_N)$ units to the right. It is simply the difference of the probability that all the other $M_N - 1$ of the X_i's in $[a_N, b_N]$ are not in the interval $[X_k, X_k + t/Nf(a_N)]$ and the probability that all other X_i in $[a_N, b_N]$ are not in the interval $[X_k, X_k + (t + \Delta t)/Nf(a_N)]$; note that we are using the wrapped interval $[a_N, b_N]$.

Some care is required in these calculations. We have a conditional probability as we assume that $X_k \in [a_N, b_N]$ and that exactly M_N of the X_i are in $[a_N, b_N]$. Thus, these

probabilities depend on two random variables, namely, X_k and M_N. This is not a problem in practice, however (e.g., M_N is tightly concentrated about its mean value).

Recalling our expansion for $g_N(x)$ (and that $b_N - a_N = N^{\delta-1}/f(a_N) + O(N^{2\delta-2})$ and t is of size 1), after simple algebra, we find that with probability tending to 1, for a given X_k and M_N, the first probability is

$$\left(1 - \int_{X_k}^{X_k + t/Nf(a_N)} g_N(x)dx\right)^{M_N - 1}. \tag{2.9}$$

The above integral equals $tN^{\delta} + O(N^{-1})$ (use the Taylor series expansion in (2.7) and note that the interval $[a_N, b_N]$ is of size $O(N^{\delta-1})$). Using (2.3), it is easy to see that this is a.s. equal to

$$\left(1 - \frac{t + O(N^{\delta-1} + \beta_N N^{-\delta/2})}{M_N}\right)^{M_N - 1}. \tag{2.10}$$

We, therefore, find that as $N \to \infty$, the probability that $M_N - 1$ of the X_i's ($i \neq k$) are in $[a_N, b_N] \setminus [X_k, X_k + t/Nf(a_N)]$, conditioned on X_k and M_N, converges to e^{-t}. (Some care is required, as the exceptional set in our a.s. statement can depend on t. This can be surmounted by taking expectations with respect to our conditional probabilities and applying the dominated convergence theorem.)

The calculation of the second probability, the conditional probability that the $M_N - 1$ other X_i' that are $[a_N, b_N]$ not in the interval $[X_k, X_k + (t + \Delta t)/Nf(a_N)]$, given X_k and M_N, follows analogously by replacing t with $t + \Delta t$ in the previous argument. We thus find that this probability is $e^{-(t+\Delta t)}$. As

$$\int_t^{t+\Delta t} e^{-u}du = e^{-t} - e^{-(t+\Delta t)}, \tag{2.11}$$

we find that the density of the difference between adjacent order statistics tends to the standard (unit) exponential density; thus, the proof of Theorem 1.4 now follows from Theorem 1.3.

3. Proofs of Theorems 1.5 and 1.7

We generalize the notation from Section 2. Let $f(x)$ be any distribution with a second-order Taylor series at each point with first and second derivatives uniformly bounded, and let $X_{1:N}, \ldots, X_{N:N}$ be the order statistics. We fix a $\delta \in (0, 1)$, and for $k \in \{1, \ldots, N^{1-\delta}\}$, we consider bins $[a_{k;N}, b_{k;N}]$ such that

$$\int_{a_{k;N}}^{b_{k;N}} f(x)dx = \frac{N^{\delta}}{N} = N^{\delta-1}, \tag{3.1}$$

there are $N^{1-\delta}$ such bins. By the central limit theorem (see (2.3)), if $M_{k;N}$ is the number of order statistics in $[a_{k;N}, b_{k;N}]$, then, provided that $\epsilon > \max(0, 1/3 - \delta/2)$ with probability tending to 1, we have

$$M_{k;N} = N^{\delta} + O(N^{\epsilon+\delta/2}), \tag{3.2}$$

of course we also require $\epsilon < \delta/2$, as, otherwise, the error term is larger than the main term.

Remark 3.1. Before we considered just one fixed interval; as we are studying $N^{1-\delta}$ intervals simultaneously, we need ϵ in the exponent so that with high probability, all intervals have to first order N^{δ} order statistics. For the arguments below, it would have sufficed to have an error of size $O(N^{\delta-\epsilon})$. We thank the referee for pointing out that $\epsilon > 1/3 - \delta/2$, and provide his argument in Appendix B.

Similar to (2.8), the average spacing between adjacent order statistics in $[a_{k;N}, b_{k;N}]$ is

$$\left(f(a_{k;N})N\right)^{-1} + N^{-1} \cdot O(N^{-(\epsilon+\delta/2)} + N^{\delta-1}). \tag{3.3}$$

Note that (3.3) is the generalization of (1.11); if f is the uniform distribution on $[0, L]$, then $f(a_{k;N}) = 1/L$. By Theorem 1.4, as $N \to \infty$, the distribution of digits of the differences in each bin converges to shifted exponential behavior; however, the variation in the average spacing between bins leads to bin-dependent shifts in the shifted exponential behavior.

Similar to (1.11), we can study the distribution of digits of the differences of the normalized order statistics. If $X_{i:N}$ and $X_{i+1:N}$ are in $[a_{k;N}, b_{k;N}]$, then

$$\begin{aligned} Z_{i;N} &= \frac{X_{i+1:N} - X_{i:N}}{\left(f(a_{k;N})N\right)^{-1} + N^{-1} \cdot O(N^{-(\epsilon+\delta/2)} + N^{\delta-1})}, \\ \log_B Z_{i;N} &= \log_B(X_{i+1:N} - X_{i:N}) + \log_B N - \log_B\left(f(a_{k;N})^{-1} + O(N^{-(\epsilon+\delta/2)} + N^{\delta-1})\right). \end{aligned} \tag{3.4}$$

Note we are using the *same* normalization factor for all differences between adjacent order statistics in a bin. Later, we show that we may replace $f(a_{k;N})$ with $f(X_{i:N})$. As we study all $X_{i+1:N} - X_{i:N}$ in the bin $[a_{k;N}, b_{k;N}]$, it is useful to rewrite the above as

$$\log_B(X_{i+1:N} - X_{i:N}) = \log_B Z_{i;N} - \log_B N + \log_B\left(f(a_{k;N})^{-1} + O(N^{-(\epsilon+\delta/2)} + N^{\delta-1})\right). \tag{3.5}$$

We have $N^{1-\delta}$ bins, so $k \in \{1, \ldots, N^{1-\delta}\}$. As we only care about the limiting behavior, we may safely ignore the first and last bins. We may, therefore, assume that each $a_{k;N}$ is finite, and $a_{k+1;N} = b_{k;N}$. (Of course, we know that both quantities are finite as we assumed that our distribution has compact support. We remove the last bins to simplify generalizations to noncompactly supported distributions.)

Let $F(x)$ be the cumulative distribution function for $f(x)$. Then,

$$F(a_{k;N}) = \frac{(k-1)N^{\delta}}{N} = (k-1)N^{\delta-1}. \tag{3.6}$$

For notational convenience, we relabel the bins so that $k \in \{0, \ldots, N^{1-\delta} - 1\}$; thus $F(a_{k;N}) = kN^{\delta-1}$.

We now prove our theorems which determine when these bin-dependent shifts cancel (yielding Benford behavior), or reinforce (yielding sums of shifted exponential behavior).

Proof of Theorem 1.5. There are approximately N^{δ} differences in each bin $[a_{k;N}, b_{k;N}]$. By Theorem 1.4, the distribution of the digits of the differences in each bin converges to shifted exponential behavior. As we assume that the first and second derivatives of f are uniformly bounded, the Big-Oh constants in Section 2 are independent of the bins. The shift in the shifted exponential behavior in each bin is controlled by the last two terms on the right-hand side of (3.5). The $\log_B N$ shifts the shifted exponential behavior in each bin equally. The bin-dependent shift is controlled by the final term

$$\log_B\left(f(a_{k;N})^{-1} + O\left(N^{-(\epsilon+\delta/2)} + N^{\delta-1}\right)\right) = -\log_B f(a_{k;N}) + \log_B\left(1 + \frac{\min(N^{-(\epsilon+\delta/2)}, N^{\delta-1})}{f(a_{k;N})}\right). \tag{3.7}$$

Thus, each of the $N^{1-\delta}$ bins exhibits shifted exponential behavior, with a bin-dependent shift composed of the two terms in (3.7). By (1.12), $f(a_{k;N})$ are not small compared to $\min(N^{-(\epsilon+\delta/2)}, N^{\delta-1})$, and hence the second term $\log_B(1 + (\min(N^{-(\epsilon+\delta/2)}, N^{\delta-1})/f(a_{k;N})))$ is negligible. In particular, this factor depends only very weakly on the bin, and tends to zero as $N \to \infty$.

Thus, the bin-dependent shift in the shifted exponential behavior is approximately $-\log_B f(a_{k;N}) = -\log_B f(F^{-1}(kN^{\delta-1}))$. If these shifts are equidistributed modulo 1, then the deviations from Benford behavior cancel, and the shifted exponential behavior of each bin becomes Benford behavior for *all* the differences. □

Remark 3.2. Consider the case when the density is a uniform distribution on some interval. Then, all $f(F^{-1}(kN^{\delta-1}))$ are equal, and each bin has the same shift in its shifted exponential behavior. These shifts, therefore, reinforce each other, and the distribution of all the differences is also shifted exponential behavior, with the same shift. This is observed in numerical experiments (see Theorem 1.3 for an alternate proof).

We analyze the assumptions of Theorem 1.5. The condition from (1.12) is easy-to-check, and is often satisfied. For example, if the probability density is a finite union of monotonic pieces and is zero only finitely often, then (1.12) holds. This is because for $k \in I(\epsilon, \delta, N)$, $F^{-1}(kN^{\delta-1}) \in [F^{-1}(\epsilon), F^{-1}(1-\epsilon)]$, and this is, therefore, independent of N (if f vanishes finitely often, we need to remove small subintervals from $I(\epsilon, \delta, N)$, but the analysis proceeds similarly). The only difficulty is basically a probability distribution with intervals of zero probability. Thus, (1.12) is a mild assumption.

If we choose any distribution *other than* a uniform distribution, then $f(x)$ is not constant; however, (1.13) does not need to hold (i.e., $\log_B f(a_{k;N})$ mod 1does not need to be equidistributed as $N \to \infty$). For example, consider a Pareto distribution with minimum value 1 and exponent $a > 0$. The density is

$$f(x) = \begin{cases} ax^{-a-1} & \text{if } x \geq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

The Pareto distribution is known to be useful in modelling natural phenomena, and for appropriate choices of exponents, it yields approximately Benford behavior (see [16]).

Example 3.3. If f is a Pareto distribution with minimum value 1 and exponent $a > 0$, then f does not satisfy the second condition of Theorem 1.5, (1.13).

To see this, note that the cumulative distribution function of f is $F(x) = 1 - x^{-a}$. As we only care about the limiting behavior, we need only to study $k \in I(\epsilon, \delta, N) = [\epsilon N^{1-\delta}, N^{1-\delta} - \epsilon N^{1-\delta}]$. Therefore, $F(a_{k;N}) = kN^{\delta-1}$ implies that

$$a_{k;N} = \left(1 - kN^{\delta-1}\right)^{-1/a}, \qquad f\left(a_{k;N}\right) = a\left(1 - kN^{\delta-1}\right)^{(a+1)/a}. \tag{3.9}$$

The condition from (1.12) is satisfied, namely,

$$\lim_{N\to\infty} \max_{k\in I(\epsilon,\delta,N)} \frac{\min(N^{-(\epsilon+\delta/2)}, N^{\delta-1})}{f(a_{k;N})} = \lim_{N\to\infty} \max_{k\in I(\epsilon,\delta,N)} \frac{\min(N^{-(\epsilon+\delta/2)}, N^{\delta-1})}{a(kN^{\delta-1})^{(a+1)/a}} = 0, \tag{3.10}$$

as k is of size $N^{1-\delta}$.

Let $j = N^{1-\delta} - k \in I(\epsilon, \delta, N)$. Then, the bin-dependent shifts are

$$\begin{aligned} \log_B f\left(a_{k;N}\right) &= \frac{a+1}{a}\log_B\left(1 - kN^{\delta-1}\right) + \log_B a \\ &= \frac{a+1}{a}\log_B\left(jN^{1-\delta}\right) + \log_B a \\ &= \log_B\left(j^{(a+1)/a}\right) + \log_B\left(aN^{(1-\delta)(a+1)/a}\right). \end{aligned} \tag{3.11}$$

Thus, for a Pareto distribution with exponent a, the distribution of *all* the differences becomes Benford if and only if $j^{(a+1)/a}$ is Benford. This follows from the fact that a sequence is Benford if and only if its logarithms are equidistributed. For fixed m, j^m is *not* Benford (e.g., [6]), and thus the condition from (1.13) fails.

Remark 3.4. We chose to study a Pareto distribution because the distribution of digits of a random variable drawn from a Pareto distribution converges to Benford behavior (base 10) as $a \to 1$; however, the digits of the differences do not tend to Benford (or shifted exponential) behavior. A similar analysis holds for many distributions with good closed-form expressions for the cumulative distribution function. In particular, if f is the density of an exponential or Weibull distribution (or $f(x) = e^{-e^x}e^x$), then f does not satisfy the second condition of Theorem 1.5, (1.13).

Modifying the proof of Theorem 1.5 yields our result on the distribution of digits of the normalized differences.

Proof of Theorem 1.7. If f is the uniform distribution, there is nothing to prove. For general f, rescaling the differences eliminates the bin-dependent shifts. Let

$$\widetilde{Z}_{i:N} = \frac{X_{i+1:N} - X_{i:N}}{1/Nf(X_{i:N})}. \tag{3.12}$$

In Theorem 1.5, we use the same scale factor for all differences in a bin (see (3.4)). As we assume the first and second derivatives of f are uniformly bounded, (2.5) and (2.6) imply that for $X_{i:N} \in [a_{k;N}, b_{k;N}]$,

$$\begin{aligned} f(X_{i:N}) &= f(a_{k;N}) + O(b_{k;N} - a_{k;N}) \\ &= f(a_{k;N}) + O\left(\frac{N^{\delta-1}}{f(a_{k;N})} + N^{2\delta-2}\right), \end{aligned} \tag{3.13}$$

and the Big-Oh constants are independent of k. As we assume that f satisfies (1.12), the error term is negligible.

Thus, our assumptions on f imply that f is basically constant on each bin, and we may replace the local rescaling factor $f(X_{i:N})$ with the bin rescaling factor $f(a_{k;N})$. Thus, each bin of normalized differences has *the same* shift in its shifted exponential behavior. Therefore all the shifts reinforce, and the digits of all the normalized differences exhibit shifted exponential behavior as $N \to \infty$. □

As an example of Theorem 1.7, in Figure 1 we consider 500,000 independent random variables drawn from the Pareto distribution with exponent

$$a = \frac{4 + \sqrt[3]{19 - 3\sqrt{33}} + \sqrt[3]{19 + 3\sqrt{33}}}{3} \tag{3.14}$$

(we chose a to make the variance equal 1). We study the distribution of the digits of the differences in base 10. The amplitude is about .018, which is the amplitude of the shifted exponential behavior of Theorem 1.1 (see the equation in [23, Theorem 2] or (1.5) of Theorem 1.1).

Remark 3.5. The universal behavior of Theorem 1.7 suggests that if we are interested in the behavior of the digits of all the differences, the natural quantity to study is the *normalized* differences. For any distribution with uniformly bounded first and second derivatives and a second-order Taylor series expansion at each point, we obtain shifted exponential behavior.

Appendices

A. Proof of Theorem 1.1

To prove Theorem 1.1, it suffices to study the distribution of $\log_B \zeta \bmod 1$ when ζ has the standard exponential distribution (see (1.4)). We have the following useful chain of equalities. Let $[a, b] \subset [0, 1]$. Then,

$$\begin{aligned} \text{Prob}(\log_B \zeta \bmod 1 \in [a, b]) &= \sum_{k=-\infty}^{\infty} \text{Prob}(\log_B \zeta \in [a + k, b + k]) \\ &= \sum_{k=-\infty}^{\infty} \text{Prob}(\zeta \in [B^{a+k}, B^{b+k}]) \\ &= \sum_{k=-\infty}^{\infty} (e^{-B^{a+k}} - e^{-B^{b+k}}). \end{aligned} \tag{A.1}$$

It suffices to investigate (A.1) in the special case when $a = 0$, as the probability of any interval $[\alpha, \beta]$ can always be found by subtracting the probability of $[0, \alpha]$ from $[0, \beta]$. We are, therefore, led to study, for $b \in [0, 1]$, the cumulative distribution function of $\log_B \zeta \bmod 1$,

$$F_B(b) := \text{Prob}\big(\log_B \zeta \bmod 1 \in [0, b]\big) = \sum_{k=-\infty}^{\infty} \big(e^{-B^k} - e^{-B^{b+k}}\big). \tag{A.2}$$

This series expansion converges rapidly, and Benford behavior for ζ is equivalent to the rapidly converging series in (A.2) equalling b for all b.

As Benford behavior is equivalent to $F_B(b)$ equals b for all $b \in [0, 1]$, it is natural to compare $F_B'(b)$ to 1. If the derivatives were identically 1, then $F_B(b)$ would equal b plus some constant. However, (A.2) is zero when $b = 0$, which implies that this constant would be zero. It is hard to analyze the infinite sum for $F_B(b)$ directly. By studying the derivative $F_B'(b)$, we find a function with an easier Fourier transform than the Fourier transform of $e^{-B^u} - e^{-B^{b+u}}$, which we then analyze by applying Poisson summation.

We use the fact that the derivative of the infinite sum $F_B(b)$ is the sum of the derivatives of the individual summands. This is justified by the rapid decay of the summands (see, e.g., [28, Corollary 7.3]). We find

$$F_B'(b) = \sum_{k=-\infty}^{\infty} e^{-B^{b+k}} B^{b+k} \log B = \sum_{k=-\infty}^{\infty} e^{-\beta B^k} \beta B^k \log B, \tag{A.3}$$

where for $b \in [0, 1]$, we set $\beta = B^b$.

Let $H(t) = e^{-\beta B^t} \beta B^t \log B$; note $\beta \geq 1$. As $H(t)$ is of rapid decay in t, we may apply Poisson summation (e.g., [29]). Thus,

$$\sum_{k=-\infty}^{\infty} H(k) = \sum_{k=-\infty}^{\infty} \widehat{H}(k), \tag{A.4}$$

where $\widehat{H}$ is the Fourier transform of H: $\widehat{H}(u) = \int_{-\infty}^{\infty} H(t) e^{-2\pi i t u} dt$. Therefore,

$$F_B'(b) = \sum_{k=-\infty}^{\infty} H(k) = \sum_{k=-\infty}^{\infty} \widehat{H}(k) = \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\beta B^t} \beta B^t \log B \cdot e^{-2\pi i t k} dt. \tag{A.5}$$

Let us change variables by taking $w = B^t$. Thus, $dw = B^t \log B \, dt$ or $dw/w = \log B \, dt$. As $e^{-2\pi i t k} = (B^{t/\log B})^{-2\pi i k} = w^{-2\pi i k / \log B}$, we have

$$\begin{aligned} F_B'(b) &= \sum_{k=-\infty}^{\infty} \int_0^{\infty} e^{-\beta w} \beta w \cdot w^{-2\pi i k / \log B} \frac{dw}{w} \\ &= \sum_{k=-\infty}^{\infty} \beta^{2\pi i k / \log B} \int_0^{\infty} e^{-u} u^{-2\pi i k / \log B} du \\ &= \sum_{k=-\infty}^{\infty} \beta^{2\pi i k / \log B} \Gamma\left(1 - \frac{2\pi i k}{\log B}\right), \end{aligned} \tag{A.6}$$

where we have used the definition of the Γ-function

$$\Gamma(s) = \int_0^\infty e^{-u} u^{s-1} du, \qquad \mathrm{Re}(s) > 0. \tag{A.7}$$

As $\Gamma(1) = 1$, we have

$$F_B'(b) = 1 + \sum_{m=1}^{\infty} \left[\beta^{2\pi i m / \log B} \Gamma\left(1 - \frac{2\pi i m}{\log B}\right) + \beta^{-2\pi i m / \log B} \Gamma\left(1 + \frac{2\pi i m}{\log B}\right) \right]. \tag{A.8}$$

Remark A.1. The above series expansion is rapidly convergent, and shows the deviations of $\log_B \zeta \bmod 1$ from being equidistributed as an infinite sum of special values of a standard function. As $\beta = B^b$, we have $\beta^{2\pi i m / \log B} = \cos(2\pi m b) + i \sin(2\pi m b)$, which gives a Fourier series expansion for $F'(b)$ with coefficients arising from special values of the Γ-function.

We can improve (A.8) by using additional properties of the Γ-function. If $y \in \mathbb{R}$, then from (A.7), we have $\Gamma(1 - iy) = \overline{\Gamma(1 + iy)}$ (where the bar denotes complex conjugation). Thus, the mth summand in (A.8) is the sum of a number and its complex conjugate, which is simply twice the real part. We have formulas for the absolute value of the Γ-function for large argument. We use (see [30, page 946, equation (8.332)]) that

$$\left|\Gamma(1 + ix)\right|^2 = \frac{\pi x}{\sinh(\pi x)} = \frac{2\pi x}{e^{\pi x} - e^{-\pi x}}. \tag{A.9}$$

Writing the summands in (A.8) as $2\mathrm{Re}(e^{-2\pi i m b} \Gamma(1 + 2\pi i m / \log B))$, (A.8) becomes

$$F_B'(b) = 1 + 2 \sum_{m=1}^{M-1} \mathrm{Re}\left(e^{-2\pi i m b} \Gamma\left(1 + \frac{2\pi i m}{\log B}\right)\right) + 2 \sum_{m=M}^{\infty} \mathrm{Re}\left(e^{-2\pi i m b} \Gamma\left(1 + \frac{2\pi i m}{\log B}\right)\right). \tag{A.10}$$

The rest of the claims of Theorem 1.1 follow from simple estimation, algebra, and trigonometry.

With constants as in the theorem, if we take $M = 1$ and $B = e$ (resp., $B = 10$) the error is at most .00499 (resp., .378), while if $M = 2$ and $B = e$ (resp., $B = 10$), the error is at most $3.16 \cdot 10^{-7}$ (resp., .006). Thus, just *one* term is enough to get approximately five digits of accuracy base e, and two terms give three digits of accuracy base 10. For many bases, we have reduced the problem to evaluate $\mathrm{Re}(e^{-2\pi i b} \Gamma(1 + 2\pi i / \log B))$. This example illustrates the power of Poisson summation, taking a slowly convergent series expansion and replacing it with a rapidly converging one.

Corollary A.2. *Let ζ have the standard exponential distribution. There is no base $B > 1$ such that ζ is Benford base B.*

Proof. Consider the infinite series expansion in (1.5). As $e^{-2\pi i m b}$ is a sum of a cosine and a sine term, (1.5) gives a rapidly convergent Fourier series expansion. If ζ were Benford base B, then $F_B'(b)$ must be identically 1; however, $\Gamma(1 + 2\pi i m / \log B)$ is never zero for m a positive integer because its modulus is nonzero (see (A.9)). As there is a unique rapidly convergent

Fourier series equal to 1 (namely, $g(b) = 1$; see [29] for a proof), our $F'_B(b)$ cannot identically equal 1. □

B. Analyzing $N^{1-\delta}$ intervals simultaneously

We show why in addition to $\epsilon > 0$ we also needed $\epsilon > 1/3 - \delta/2$ when we analyzed $N^{1-\delta}$ intervals simultaneously in (3.2); we thank one of the referees for providing this detailed argument.

Let $Y_1, \ldots, Y_N$ be i.i.d.r.v. with $\mathbb{E}[Y_i] = 0$, $\mathrm{Var}(Y_i) = \sigma^2$, $\mathbb{E}[|Y_i|^3] < \infty$, and set $S_N = (Y_1 + \cdots + Y_N)/\sqrt{N\sigma^2}$. Let $\Phi(x)$ denote the cumulative distribution function of the standard normal. Using a (nonuniform) sharpening of the Berry-Esséen estimate (e.g., [31]), we find that for some constant $c > 0$,

$$\left|\mathrm{Prob}(S_N \le x) - \Phi(x)\right| \le \frac{c\mathbb{E}[|Y_1|^3]}{\sigma^3\sqrt{N}(1+|x|)^3}, \quad x \in \mathbb{R},\ N \ge 1. \tag{B.1}$$

Taking $Y_i = w_i - N^{\delta-1}$, where w_i is defined by (2.1), yields

$$\begin{gathered} S_N = \frac{M_N - N^\delta}{\sqrt{N^\delta(1-N^{\delta-1})}}, \\ \sigma^2 N^{\delta-1}(1-N^{\delta-1}), \\ \mathbb{E}\left[|Y_i|^3\right] \le 2N^{\delta-1}. \end{gathered} \tag{B.2}$$

Thus, (B.1) becomes

$$\left|\mathrm{Prob}\left(\frac{M_N - N^\delta}{\sqrt{N^\delta(1-N^{\delta-1})}} \le x\right) - \Phi(x)\right| \le \frac{3cN^{-\delta/2}}{(1+|x|)^3}, \tag{B.3}$$

for all $N \ge N_0$ (for some N_0 sufficiently large, depending on δ).

For each N, k, and ϵ consider the event

$$A_{N,k,\epsilon} = \left\{\frac{M_{k;N} - N^\delta}{\sqrt{N^\delta(1-N^{\delta-1})}} \in \left[-N^\epsilon, N^\epsilon\right]\right\}. \tag{B.4}$$

Then, as $N \to \infty$, we have

$$\mathrm{Prob}\left(\bigcap_{k=1}^{N^{1-\delta}} A_{N,k,\epsilon}\right) \longrightarrow 1, \tag{B.5}$$

provided that

$$\sum_{k=1}^{N^{1-\delta}} \mathrm{Prob}\left(A^c_{N,k,\epsilon}\right) \longrightarrow 0, \tag{B.6}$$

as $N \to \infty$. Using (B.3) gives

$$\begin{aligned} \mathrm{Prob}\left(A^c_{N,k,\epsilon}\right) &\leq \frac{6cN^{-\delta/2}}{\left(1+N^{\epsilon}\right)^3} + 2\left(1-\Phi\left(N^{\epsilon}\right)\right) \\ &\leq 6cN^{-\delta/2-3\epsilon} + \sqrt{\frac{2}{\pi}}N^{-\epsilon}\exp\left(-\frac{N^{2\epsilon}}{2}\right) \end{aligned} \tag{B.7}$$

(e.g., [32]). Thus, the sum in (B.6) is at most

$$6cN^{1-3\delta/2-3\epsilon} + \sqrt{\frac{2}{\pi}}N^{1-\delta-\epsilon}\exp\left(-\frac{N^{2\epsilon}}{2}\right), \tag{B.8}$$

and this is $O(1)$ provided that $\epsilon > 0$ and $\epsilon > 1/3 - \delta/2$.

Acknowledgments

The authors would like to thank Ted Hill, Christoph Leuenberger, Daniel Stone, and the referees for numerous helpful comments. S. J. Miller was partially supported by NSF (Grant no. DMS-0600848).

References

[1] F. Benford, "The law of anomalous numbers," *Proceedings of the American Philosophical Society*, vol. 78, no. 4, pp. 551–572, 1938.
[2] T. Hill, "The first-digit phenomenon," *American Scientists*, vol. 86, pp. 358–363, 1996.
[3] R. A. Raimi, "The first digit problem," *The American Mathematical Monthly*, vol. 83, no. 7, pp. 521–538, 1976.
[4] W. Hurlimann, "Benford's law from 1881 to 2006," preprint, http://arxiv.org/abs/math/0607168.
[5] J. L. Brown Jr. and R. L. Duncan, "Modulo one uniform distribution of the sequence of logarithms of certain recursive sequences," *The Fibonacci Quarterly*, vol. 8, no. 5, pp. 482–486, 1970.
[6] P. Diaconis, "The distribution of leading digits and uniform distribution mod1," *The Annals of Probability*, vol. 5, no. 1, pp. 72–81, 1977.
[7] T. P. Hill, "A statistical derivation of the significant-digit law," *Statistical Science*, vol. 10, no. 4, pp. 354–363, 1995.
[8] A. Berger, L. A. Bunimovich, and T. P. Hill, "One-dimensional dynamical systems and Benford's law," *Transactions of the American Mathematical Society*, vol. 357, no. 1, pp. 197–219, 2005.
[9] A. Berger and T. P. Hill, "Newton's method obeys Benford's law," *American Mathematical Monthly*, vol. 114, no. 7, pp. 588–601, 2007.
[10] D. Jang, J. U. Kang, A. Kruckman, J. Kudo, and S. J. Miller, "Chains of distributions, hierarchical Bayesian models and Benford's law," preprint, http://arxiv.org/abs/0805.4226.
[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.
[12] J. C. Lagarias and K. Soundararajan, "Benford's law for the $3x+1$ function," *Journal of the London Mathematical Society*, vol. 74, no. 2, pp. 289–303, 2006.

[13] S. J. Miller and M. J. Nigrini, "The modulo 1 central limit theorem and Benford's law for products," *International Journal of Algebra*, vol. 2, no. 1–4, pp. 119–130, 2008.
[14] J.-C. Pain, "Benford's law and complex atomic spectra," *Physical Review E*, vol. 77, no. 1, Article ID 012102, 3 pages, 2008.
[15] E. Costas, V. López-Rodas, F. J. Toro, and A. Flores-Moya, "The number of cells in colonies of the cyanobacterium *Microcystis aeruginosa* satisfies Benford's law," *Aquatic Botany*, vol. 89, no. 3, pp. 341–343, 2008.
[16] M. Nigrini and S. J. Miller, "Benford's Law applied to hydrology data—results and relevance to other geophysical data," *Mathematical Geology*, vol. 39, no. 5, pp. 469–490, 2007.
[17] D. E. Knuth, *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*, Addison-Wesley, Reading, Mass, USA, 3rd edition, 1997.
[18] M. Nigrini, "Digital analysis and the reduction of auditor litigation risk," in *Proceedings of the Deloitte & Touche / University of Kansas Symposium on Auditing Problems*, M. Ettredge, Ed., pp. 69–81, University of Kansas, Lawrence, Kan, USA, 1996.
[19] M. Nigrini, "The use of Benford's law as an aid in analytical procedures," *Auditing: A Journal of Practice & Theory*, vol. 16, no. 2, pp. 52–67, 1997.
[20] W. R. Mebane Jr., "Election forensics: the second-digit Benford's law test and recent American presidential elections," in *Presented at the Election Fraud Conference*, Salt Lake City, Utah, USA, September 2006.
[21] M. Nigrini and S. J. Miller, "Data diagnostics using second order tests of Benford's law," preprint.
[22] L. M. Leemis, B. W. Schmeiser, and D. L. Evans, "Survival distributions satisfying Benford's law," *The American Statistician*, vol. 54, no. 4, pp. 236–241, 2000.
[23] H.-A. Engel and C. Leuenberger, "Benford's law for exponential random variables," *Statistics & Probability Letters*, vol. 63, no. 4, pp. 361–365, 2003.
[24] R. S. Pinkham, "On the distribution of first significant digits," *Annals of Mathematical Statistics*, vol. 32, pp. 1223–1230, 1961.
[25] S. J. Miller and R. Takloo-Bighash, *An Invitation to Modern Number Theory*, Princeton University Press, Princeton, NJ, USA, 2006.
[26] H. A. David and H. N. Nagaraja, *Order Statistics*, Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, NJ, USA, 3rd edition, 2003.
[27] R.-D. Reiss, *Approximate Distributions of Order Statistics. With Applications to Nonparametric Statistics*, Springer Series in Statistics, Springer, New York, NY, USA, 1989.
[28] S. Lang, *Undergraduate Analysis*, Undergraduate Texts in Mathematics, Springer, New York, NY, USA, 2nd edition, 1997.
[29] E. M. Stein and R. Shakarchi, *Fourier Analysis: An Introduction*, vol. 1 of *Princeton Lectures in Analysis*, Princeton University Press, Princeton, NJ, USA, 2003.
[30] I. Gradshteyn and I. Ryzhik, *Tables of Integrals, Series, and Products*, Academic Press, New York, NY, USA, 5th edition, 1965.
[31] V. V. Petrov, *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*, vol. 4 of *Oxford Studies in Probability*, The Clarendon Press, Oxford University Press, New York, NY, USA, 1995.
[32] W. Feller, *An Introduction to Probability Theory and Its Applications. Vol. I*, John Wiley & Sons, New York, NY, USA, 2nd edition, 1962.