

Some Thoughts on Benford's Law

Steven J. Miller*

November 10, 2004

Abstract

For many systems, there is a bias in the distribution of the first digits. For example, if one looks at the first digit of 2^n in base 10, as n ranges over the positive integers, one observes 1 about 30% of the time (and not $\frac{1}{9} \approx .11\%$ of the time as one might expect). This bias is known as Benford's Law, and occurs in a variety of phenomena. In fact, the IRS uses Benford's Law to check the tax returns of large corporations!

We will show that if $y_n = \log_b x_n$ is equidistributed mod 1, then x_n is Benford base b . This is sufficient to prove that Recurrence Relations (with distinct roots $\lambda_1, \dots, \lambda_k$ such that $|\lambda_1| \geq \dots \geq |\lambda_k|$ and $|\lambda_1| \neq 1$) are Benford. In particular, this will imply that the Fibonacci numbers, which satisfy the Recurrence Relation $a_n = a_{n-1} + a_{n-2}$, are Benford.

In these notes we develop most of the techniques needed to prove these results. The only fact which we must assume is that if $\alpha \notin \mathbb{Q}$, then $n\alpha \bmod 1$ is equidistributed.

The first section introduces Benford's Law, and highlights the method of proof. The second section investigates Recurrence Relations. For a nice introduction to Benford's Law, the reader should see [Hi1]; for an exposition on Benford's Law and Recurrence Relations, see [BrDu, NS].

Contents

1	Benford's Law	2
1.1	Preliminaries	3
1.2	Equidistribution and Benford	3
2	Recurrence Relations and Benford's Law	4
2.1	Recurrence Preliminaries	4
2.2	Geometric Series are Benford	5
2.3	Recurrence Relations are Benford	5
2.4	Weakening of Recurrence Constraints (Sketch)	7

*E-mail: sjmiller@math.brown.edu

1 Benford's Law

While looking through tables of logarithms in the late 1800s, Newcomb noticed a surprising fact: certain pages were significantly more worn out than others. People were looking up numbers whose logarithm started with 1 significantly more frequently than other digits. In 1938, Benford observed the same digit bias in a variety of phenomenon. See [Hi1] for a description and history, [Hi2, BBH, KonMi] for recent results, and [Knu] for connections between Benford's law and rounding errors in computer calculations.

We say a sequence of positive numbers $\{x_n\}$ is **Benford (base b)** if the probability of observing the first digit of x_n (in base b) is j is $\log_b \left(1 + \frac{1}{j}\right)$.

More precisely, we would have

$$\lim_{N \rightarrow \infty} \frac{\#\{n \leq N : \text{first digit of } x_n \text{ is } j\}}{N} = \log_b \left(1 + \frac{1}{j}\right). \quad (1)$$

Note that $j \in \{1, \dots, b-1\}$. This is a division of probability, as one of the $b-1$ events must occur, and the total probability is

$$\begin{aligned} \sum_{j=1}^{b-1} \log_b \left(1 + \frac{1}{j}\right) &= \log_b \prod_{j=1}^{b-1} \left(1 + \frac{1}{j}\right) \\ &= \log_b \prod_{j=1}^{b-1} \frac{j+1}{j} = 1 \\ &= \log_b b = 1. \end{aligned} \quad (2)$$

Note it is possible to be Benford to some bases but not others. As $\log_{10} 2 \approx .3$, this means that about 30% of the time the first digit is a 1. This is a very strong digit bias; if all digits (1 through 9) were equally likely, than the probability of the first digit being 1 would be $\frac{1}{9} \approx .11$.

A common way to prove a sequence is Benford is to show its logarithms (modulo 1) are equidistributed. Recall

Definition 1.1 (Equidistributed). *A sequence $\{y_n\}_{n=1}^{\infty}$, $y_n \in [0, 1]$, is equidistributed in $[0, 1]$ if*

$$\lim_{N \rightarrow \infty} \frac{\#\{n : |n| \leq N, y_n \in [a, b]\}}{2N+1} = \lim_{N \rightarrow \infty} \frac{\sum_{n=-N}^N \chi_{(a,b)}(y_n)}{2N+1} = b-a \quad (3)$$

for all $(a, b) \subset [0, 1]$.

The following theorem will be central to our presentation, and will be proved in §1.2:

Theorem 1.2. *If $y_n = \log_b x_n$ equidistributed mod b , then x_n is Benford (base b).*

1.1 Preliminaries

We need the following simple fact:

Lemma 1.3. *If $u \equiv v \pmod{1}$, then the first digits of b^u and b^v are the same in base b .*

Proof. (of Lemma 1.3): As $u \equiv v \pmod{1}$, without loss of generality we may write $v = u + m$, $m \in \mathbb{Z}$. If

$$b^u = u_k b^k + u_{k-1} b^{k-1} + \cdots + u_0, \quad (4)$$

then

$$\begin{aligned} b^v &= b^{u+m} \\ &= b^u \cdot b^m \\ &= (u_k b^k + u_{k-1} b^{k-1} + \cdots + u_0) b^m \\ &= u_k b^{k+m} + \cdots + u_0 b^m. \end{aligned} \quad (5)$$

Thus, the first digits of each are u_0 , proving the claim. \square

The utility of the above lemma is that in order to study the first digit of b^y (in base b), it suffices to study $y \pmod{1}$.

1.2 Equidistribution and Benford

Proof (of Theorem 1.2): Assume $y_n = \log_b x_n$ is equidistributed mod 1. Consider the unit interval $[0, 1)$. For $j \in \{1, \dots, b\}$, define p_j by

$$b^{p_j} = j; \quad (6)$$

equivalently, we have

$$p_j = \log_b j. \quad (7)$$

For $j \in \{1, \dots, b-1\}$, let

$$I_j = [p_j, p_{j+1}) \subset [0, 1). \quad (8)$$

Claim 1.4. *If $y \pmod{1} \in I_j$, then b^y has first digit j .*

The proof is immediate. By Lemma 1.3, it is sufficient to prove this for $y \in I_j$, which we now assume. Then

$$y \in [p_j, p_{j+1}) \text{ implies that } b^{p_j} \leq b^y < b^{p_{j+1}}. \quad (9)$$

From the definitions of the p_j , it follows that

$$j \leq b^y < j+1, \quad (10)$$

proving the claim.

Thus, the measure of the subset of $[0, 1)$ which, when we exponentiate by b has first digit j , is simply the length of I_j . This is

$$|I_j| = p_{j+1} - p_j = \log_b \frac{j+1}{j} = \log_b \left(1 + \frac{1}{j}\right), \quad (11)$$

the Benford (base b) probabilities.

Returning to the proof of Theorem 1.2, we see that the intervals I_j have length $\log_b \left(1 + \frac{1}{j}\right)$. As y_n is equidistributed mod 1, in the limit the percent of time $y_n \in I_j$ is equal to $|I_j|$, ie, $\log_b \left(1 + \frac{1}{j}\right)$.

Now $x_n = b^{y_n}$. Each y_n is equivalent to some $\widetilde{y}_n \pmod 1$, and by Lemma 1.3, b^{y_n} and $b^{\widetilde{y}_n}$ have the same first digit.

Thus, in the limit, the probability that the first digit of x_n is j (base b) is just $\log_b \left(1 + \frac{1}{j}\right)$, proving the theorem. \square

2 Recurrence Relations and Benford's Law

2.1 Recurrence Preliminaries

We consider Recurrence Relations of the following form:

$$a_n = c_1 a_{n-1} + \cdots + c_k a_{n-k}, \quad (12)$$

where c_1, \dots, c_k, k are fixed integers. It is well known that we may explicitly write a_n in Binet form:

$$a_n = u_1 \lambda_1^n + \cdots + u_n \lambda_k^n, \quad (13)$$

where we have ordered the eigenvalues such that $|\lambda_1| \geq \cdots \geq |\lambda_k|$.

We give a quick sketch in a special case when $k = 2$; the reader can generalize the arguments. Assume $a_n = c_1 a_{n-1} + c_2 a_{n-2}$. Let us guess that $a_n = r^n$ for some r . If this were true, then

$$r^n = c_1 r^{n-1} + c_2 r^{n-2}. \quad (14)$$

After a little algebra, this leads us to the equation

$$r^2 - c_1 r - c_2 = 0. \quad (15)$$

There are two solutions to that, say r_1 and r_2 . A little algebra shows that any solution a_n is of the form

$$a_n = u_1 r_1^n + u_2 r_2^n, \quad (16)$$

for some $u_1, u_2 \in \mathbb{C}$. If we are given initial conditions (say the values of a_0 and a_1), we can then solve for α_1, α_2 ; if the two roots are the same.

Remark 2.1. We call the equation $r^2 - c_1r - c_2$ the characteristic polynomial. Technically, we need to assume its roots are distinct; if there are repeated roots, the solution must be modified. Below, we always assume we have Recurrence Relations where the roots are distinct.

For example, for the Fibonacci numbers $k = 2$, $c_1 = c_2 = 1$, $u_1 = -u_2 = \frac{1}{\sqrt{5}}$, and $\lambda_1 = \frac{1+\sqrt{5}}{2}$, $\lambda_2 = \frac{1-\sqrt{5}}{2}$.

If $|\lambda_1| = 1$, we do not expect the first digit of a_n to be Benford (base b). For example, if we consider

$$a_n = 2a_{n-1} - a_{n-2} \tag{17}$$

with initial values $a_0 = a_1 = 1$, every $a_n = n!$ If we instead take $a_0 = 0$, $a_1 = 1$, we get $a_n = n$.

2.2 Geometric Series are Benford

Let $\{x\} = x - [x]$ denote the fractional part of x , where $[x]$ as always is the greatest integer at most x . Recall the following:

Theorem 2.2. Let $\alpha \notin \mathbb{Q}$. Then the fractional parts of $n\alpha$ are equidistributed mod 1.

For a proof, see [HW].

From this and Theorem 1.2, it immediately follows that Geometric Series (series where $x_n = r^n$) are Benford (modulo a certain irrationality condition on r):

Theorem 2.3. Let $x_n = ar^n$, $\log_b r \notin \mathbb{Q}$. Then x_n is Benford (base b).

Proof: Let $y_n = \log_b x_n = n \log_b r + \log_b a$. As $\log_b r \notin \mathbb{Q}$, the fractional parts of y_n are equidistributed. Exponentiating by b , we obtain that x_n is Benford (base b) by Theorem 1.2.

2.3 Recurrence Relations are Benford

We first introduce some notation, and then show recurrence relations are Benford.

Definition 2.4 (Big-Oh, Little-Oh). If F and G are two real functions with $G(x) > 0$ for x large, we write

$$F(x) = O(G(x)) \tag{18}$$

if there exist $M, x_0 > 0$ such that $|F(x)| \leq MG(x)$ for all $x > x_0$. If

$$\lim_{x \rightarrow +\infty} \frac{F(x)}{G(x)} = 0, \tag{19}$$

we write $F(x) = o(G(x))$ and say F is little-oh of G .

An alternative notation for $F(x) = O(g(x))$ is $F(x) \ll G(x)$. If the constant depends on parameters α, β but not on parameters a, b , we sometimes write $F(x) \ll_{\alpha, \beta} G(x)$.

Exercise 2.5. Prove for any $r, \epsilon > 0$, as $x \rightarrow \infty$ we have $x^r = O(e^x)$ and $\log x = O(x^\epsilon)$.

Theorem 2.6. Let a_n be a Recurrence Relation as before, with $|\lambda_1| \neq 1$ (note $|\lambda_1|$ is the largest absolute value of the eigenvalues). If $\log_b |\lambda_1| \notin \mathbb{Q}$, then a_n is Benford (base b).

Proof: for notational simplicity, we assume $\lambda_1 > 0$, $\lambda_1 > |\lambda_2|$, and $u_1 > 0$. We will comment at the end on how to handle the more general case.

As always, let $y_n = \log_b x_n$. By Theorem 1.2, it is sufficient to show y_n is equidistributed mod 1. We have

$$\begin{aligned} x_n &= u_1 \lambda_1^n + \cdots + u_n \lambda_n^n \\ x_n &= u_1 \lambda_1^n \left[1 + O\left(\frac{k u \lambda_2^n}{\lambda_1^n}\right) \right], \end{aligned} \quad (20)$$

where $u = \max_i |u_i| + 1$ (so $ku > 1$ and the big-Oh constant is 1). Choose a small ϵ and an n_0 such that

1. $|\lambda_2| < \lambda_1^{1-\epsilon}$;
2. for all $n > n_0$, $\frac{(ku)^{\frac{1}{n}}}{\lambda_1^\epsilon} < 1$, and note $\frac{ku}{\lambda_1^\epsilon} = \left(\frac{(ku)^{\frac{1}{n}}}{\lambda_1^\epsilon}\right)^n$.

As $ku > 1$, $(ku)^{\frac{1}{n}}$ is monotonically decreasing to 1. Note $\epsilon > 0$ if $\lambda_1 > 1$ and $\epsilon < 0$ if $\lambda_1 < 1$. Letting

$$\beta = \frac{(ku)^{\frac{1}{n_0}} |\lambda_2|}{\lambda_1^\epsilon \lambda_1^{1-\epsilon}} < 1, \quad (21)$$

we find that the error term above is bounded by β^n for $n > n_0$, which tends to 0. Therefore

$$\begin{aligned} y_n &= \log_b x_n \\ &= \log_b(u_1 \lambda_1^n) + O(\log_b(1 + \beta^n)) \\ &= n \log_b \lambda_1 + \log_b u_1 + O(\beta^n), \end{aligned} \quad (22)$$

where the big-Oh constant is 1 (actually, the constant is slightly greater than 1, but for notational ease we will use 1 below). As $\log_b \lambda_1 \notin \mathbb{Q}$, the fractional parts of $n \log_b \lambda_1$ are equidistributed mod 1. Therefore, so are the shifts obtained by adding the fixed constant $\log_b u_1$.

We need only show that the error term $O(\beta^n)$ is negligible. It is possible for the error term to change the first digit; for example, if we had 9999999999999999 (or 100000000000), then if the error term contributes 2 (or -2), we would change the first digit (base 10).

However, for n sufficiently large, the error term will change a vanishingly small number of first digits.

Say $n \log_b \lambda_1 + \log_b u_1$ exponentiates (base b) to first digit j , $j \in \{1, \dots, b-1\}$. This means

$$n \log_b \lambda_1 + \log_b u_1 \in I_j = [p_{j-1}, p_j]. \quad (23)$$

The error term is at most β^n . Thus, y_n will have exponentiate to a different first digit than $n \log_b \lambda_1 + \log_b u_1$ only if one of the following holds:

1. $n \log_b \lambda_1 + \log_b u_1$ is within β^n of p_j , and adding the error term pushes us to or past p_j ;
2. $n \log_b \lambda_1 + \log_b u_1$ is within β^n of p_{j-1} , and adding the error term pushes us before p_{j-1} .

The first set is contained in $[p_j - \beta^n, p_j)$, of length β^n . The second is contained in $[p_{j-1}, p_{j-1} + \beta^n)$, also of length β^n .

Thus, the length of the interval where $n \log_b \lambda_1 + \log_b u_1$ and y_n could exponentiate (base b) to different first digits is of size $2\beta^n$. If we choose N sufficiently large, then for all $n > N$, we can make these lengths arbitrarily small.

Thus, as $n \log_b \lambda_1 + \log_b u_1$ is equidistributed mod 1, we can control the size of the subsets of $[0, 1)$ where $n \log_b \lambda_1 + \log_b u_1$ and y_n disagree. The Benford behavior (base b) of x_n now follows (in the limit, of course).

2.4 Weaking of Recurrence Constraints (Sketch)

We now show that we can weaken most of the Recurrence Relation assumptions, namely

1. $\lambda_1 > 0$,
2. $\lambda_1 > |\lambda_2|$,
3. $u_1 > 0$.

It is possible that $|\lambda_1| = |\lambda_2| = \dots = |\lambda_i|$. If so (including signs), we can combine these terms to give

$$u_1 \lambda_1^n + \dots + u_i \lambda_i^n = u_* \lambda_1^n + u_{\#} (-\lambda_1)^n. \quad (24)$$

Of course, if the different eigenvalues of modulus λ_1 range over more than $\pm \lambda_1$, one replaces the sum above with the obvious generalization.

The proof will proceed similarly if the $\lambda_1, \dots, \lambda_i$ are real-valued (simply split into even and odd powers of n , and $2 \log_b \lambda_1 \notin \mathbb{Q}$ (in the odd case, we get an extra translation by a multiple of $\log_b \lambda_1$). Note this shows how to handle the negative sign constraint (for we do not want to take logarithms of negative numbers, hence we break our sequence into two sequences). Similarly, if u_1 (or the net effect from eigenvalues of modulus $|\lambda_1|$) is negative, we consider $-x_n$, and show that satisfies Benford (base b).

References

- [BBH] A. Berger, Leonid A. Bunimovich and T. Hill, *One-dimensional dynamical systems and Benford's Law*, accepted for publication in Transactions of the American Mathematical Society.
- [BrDu] J. Brown and R. Duncan, *Modulo one uniform distribution of the sequence of logarithms of certain recursive sequences*, Fibonacci Quarterly **8**, 1970, 482-486.

- [Du] R. Durrett, *Probability: Theory and Examples*, Duxbury Press, second edition, 1996.
- [GKP] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics: A foundation for computer science*, Addison-Wesley Publishing Company, 1988.
- [HW] G. H. Hardy and E. Wright, *An Introduction to the Theory of Numbers*, fifth edition, Oxford Science Publications, Clarendon Press, Oxford, 1995.
- [Hi1] T. Hill, *The first-digit phenomenon*, *American Scientists* **86**, 1996, 358-363.
- [Hi2] T. Hill, *A statistical derivation of the significant-digit law*, *Statistical Science* **10**, 1996, 354-363.
- [Kel] D. Kelley, *Introduction to Probability*, Macmillian Publishing Company, 1994.
- [Knu] D. Knuth, *The Art of Computer Programming, Vol. 2*, Addison-Wesley, second edition, 1981.
- [KonMi] A. Kontorovich and S. J. Miller, *Poisson Summation, Benford's Law and values of L-functions*, preprint.
- [NS] K. Nagasaka and J. S. Shiue, *Benford's law for linear recurrence sequences*, *Tsukuba J. Math.* **11**, 1987, 341-351.
- [We] E. Weisstein, *MathWorld—A Wolfram Web Resource*, <http://mathworld.wolfram.com/>