Benford Behavior of Dependent Random Variables

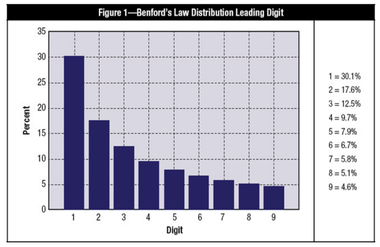
Taylor Corcoran, taylorc3@email.arizona.edu **Joseph Iafrate,** jri1@williams.edu **Jaclyn Porfilio,** jdp2@williams.edu **Jirapat Samranvedhya;** js4@williams.edu **Advisor: Steven J. Miller** sjm1@williams.edu

Number Theory and Probability Group — SMALL 2013 — Williams College

1. Background

Definition: Benford's Law of Leading Digit Bias (base B) states that in many real-life data sets, the proportion of values beginning with digit d is $\log_B(1+\frac{1}{d})$.

The Benford distribution of leading digits base 10 is:

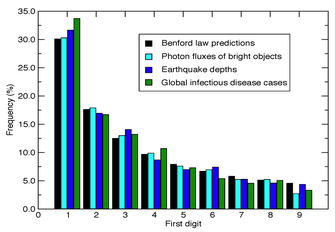


Abstract

Analyzing which datasets adhere to Benford's Law and how quickly Benford behavior sets in are the two most important problems in the field. Most previous analyses required the independence of the random variables in question. We study the case of dependent random variables by building on the work of Becker, Greaves-Tunnell, Miller, Ronan, Strauch, and Lemons to further develop techniques that allow us to analyze fragmentation models with correlated values and the determinant expansion of $n \times n$ matrices with entries drawn from 'nice' distributions.

2. Applications of Benford's Law

Benford's Law emerges in man-made, natural, and mathematical datasets and can be applied to a variety of fields, from economics to geoscience to computer science to psychology.



Examples of Benford's Law

Image obtained from [A]

When is Benford's Law used?

- ◇ Fraud detection and data integrity
  - – Accounting fraud
  - – Election fraud
- ◇ Errors in rounding or data collection methods
- ◇ With mathematical sequences that are Benford
  - – Iterates of $3x+1$ map
  - – Fibbonacci numbers

3. Research Questions

1. Is the distribution of leading digits of the sticks generated after the Nth iteration of a given cutting process Benford?
2. Is the distribution of leading digits of the $n!$ terms in the determinant expansion of an $n \times n$ matrix with independent, identically distributed, positive values Benford?

4. Stick Decomposition: Fixed Proportion

Fragmentation Process: In Stage 1, cut a given stick into two pieces whose lengths have proportion p to each other. In Stage 2, cut each resulting piece into two pieces with the lengths again in proportion p to each other.

Results: Cut at proportion p, and consider $\frac{1-p}{p} = 10^x$.
We consider two main cases:

1. $x \in \mathbb{Q}$: not Benford
2. $x \notin \mathbb{Q}$:Benford
   (a) $x \notin \mathbb{Q}$, x is of infinite irrationality type: Benford
   (b) x is of finite irrationality type: Benford with quantifiable convergence

1. $x \in \mathbb{Q}$
We used the multisection formula of binomial coefficients

$$\sum_{l=0}^{\infty} \binom{N}{lq+j} = \frac{2^N}{q} \sum_{s=0}^{q-1} \left(\cos\frac{\pi s}{q}\right)^N \cos\frac{\pi(N-2j)s}{q}. \quad (1)$$

and proved that the probability of observing a particular leading digit must be a multiple of $1/q$, which is a rational number. On the other hand, the probability from the Benford distribution is $\log(1+1/d)$ which is an irrational number. Therefore, we cannot get perfect Benford behavior.

2. $x \notin \mathbb{Q}$
In order to show that the leading digits of the 2^N stick lengths are Benford, we showed that the logarithms of the piece lengths are equidistributed. We broke the binomial distribution into intervals of width N^δ and showed both that the probability does not change much in each interval and that each interval is equidistributed.

(a). x does not have finite irrationality exponent
As a special case of these numbers, we applied our techniques to Liouville numbers. A **Liouville number** is a transcendental number that can be closely approximated by rationals [W]. Using the fact that $n\alpha \mod 1$ is equidistributed gives that, for all $[a,b] \subset [0,1]$, given $\epsilon > 0$, there exists $M(\epsilon,a,b,\alpha)$ such that

$$\#\{n \le N : n\alpha \bmod 1 \in [a,b]\} = (b-a)N + \mathscr{E}(\epsilon N) \quad (2)$$

for all $N \ge M(\epsilon,a,b,\alpha)$.
Letting N, the number of iterations of this cutting process, be sufficiently large so that N^δ is greater than $M(\epsilon,a,b,\alpha)$ causes this process to result in Benford behavior.

(b). x has finite irrationality exponent
To show that Benford behavior follows in this case, we used the same general approach as in part 2a. However, the fact that x was of finite irrationality type allowed us to explicitly determine the rate of convergence.

5. Stick Decomposition: Additive Model

Fragmentation Process: Consider a stick of length L where L is odd. Uniformly at random cut the stick into two pieces each of integer length: one even and one odd. Pieces of even length do not decompose. Recursively repeat the process on the piece of odd length until left with pieces only of even length or length 1.

Continuous Model: Let $p_1 \sim U(0,1)$ and decompose the initial stick into two pieces: one of length p_1 and one of length 1 - p_1. Without loss of generality, treat the stick of length $1-p_1$ as if it were of even length and do not decompose it further. Break the other stick at proportion $p_2 \sim U(0,1)$. Recursively repeat N - 1 times, leaving N sticks.

Results: The distribution of leading digits of stick lengths after the above decomposition process is Benford. The proof techniques used were very similar to the fixed proportion case. We do not consider the first $\log N$ pieces and all pairs of i,j such that X_i and X_j do not differ by at least $\log N$ terms. Removing these terms does not affect whether or not the distribution is Benford, but does remove dependencies that greatly complicate our analysis.

6. Stick Decomposition: Conjectures

We are numerically simulating the fragmentation of a stick into integer lengths subject to different conditions and exploring the connection between stopping sequence density and Benford behavior. We believe that Benford behavior occurs if a stick stops decomposing if its length is a prime or a 1 and does not occur if a stick stops decomposing if its length is a perfect square.

7. Determinant Expansion

Results: The terms in the determinant expansion are Benford provided the following condition on the density $f(x)$ is met:

$$\lim_{n\to\infty} \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \prod_{m=1}^{n} M_f\left(1 - \frac{2\pi i l}{\log 10}\right) = 0 \quad (3)$$

where $M_f(s) = \int_0^\infty x^{s-1} f(x)dx$ denotes the Mellin transform of f. Let

$$P_n(s) = \frac{\sum_{i=1}^{n!} \varphi_s(X_{i,n})}{n!} \quad (4)$$

To show that the first digits of $\{X_{i,n}\}_{i=1}^{n!}$ follow a Benford distribution, it suffices to show that

1. $\lim_{N\to\infty} \mathbb{E}[P_n(s)] = \log_{10}(s)$
2. $\lim_{N\to\infty} Var[P_n(s)] = 0$.

Given a fixed $X_{i,n}$, the number of terms in the determinant expansion that share k elements is

$$\binom{n}{k} \sum_{\alpha=0}^{n-k} \binom{n-k}{\alpha} (n-k-\alpha)!(-1)^\alpha \quad (5)$$

Let $K_{i,j}$ be the number of matrix entries that a given $X_{i,n}$ and $X_{j,n}$ share. Fix $X_{i,n}$, and it follows from (5) that

$$\begin{aligned} P(K_{i,j}=k) &= \frac{1}{k!} \sum_{\alpha=0}^{n-k} \frac{(-1)^\alpha}{\alpha!} \\ &\to 1. \quad (6) \\ \mathbb{E}[K_{i,j}] &= \sum_{k=0}^{n-2} \frac{k}{ek!} + \sum_{k=0}^{n-2} O\left(\binom{n}{k} \frac{k}{(n-k)n!}\right) \\ &\to 1. \quad (7) \\ \mathrm{Var}\left(K_{i;j}\right) &= \sum_{k=0}^{n-2} \frac{k^2}{ek!} + \sum_{k=0}^{n-2} O\left(\binom{n}{k} \frac{k^2}{n!(n-k)}\right) - 1 \\ &\to 1 \quad (8) \end{aligned}$$

These calculations allow us to treat the terms in the determinant expansion as relatively independent, which greatly simplifies our analysis.

8. Acknowledgements

We wish to thank Williams College and the NSF, whose generous support made this research possible. The authors and the advisor are funded by NSF DMS0850577.

References

[A] Examples of Benford's Law. Digital image. Australian National University Research School of Earth Sciences, n.d. Web. 15 July 2013.

[BGMRS] Becker T., A. Greaves-Tunnell, S. Miller, R. Ronan, and F. Strauch, *Benford's Law and Continuous Dependent Random Variables.*

[C] Chen, Hongwei (2010). *On the Summation of First Subseries in Closed Form.* International Journal of Mathematical Education in Science and Technology 41:4, 538-547.

[JKKKM] D. Jang, J. U. Kang, A. Kruckman, J. Kudo and S. J. Miller, *Chains of distributions, hierarchical Bayesian models and Benford's Law*, to appear in the Journal of Algebra, Number Theory: Advances and Applications.

[KM] Kontorovich, Alex V. and Steven J. Miller. *Benford's Law, Values of L-functions and the $3x+1$ Problem.*

[MT] Miller, Steven J. and Ramin Takloo-Bighash. "Needed Gaussian Integral." *An Invitation to Modern Number Theory.* Princeton: Princeton UP, 2006. 222. Print.

[W] Weisstein, Eric W. "Liouville Number." From MathWorld–A Wolfram Web Resource.