Generalized stick fragmentation and Benford's Law

Steven J Miller (sjm1@williams.edu), Williams College, with

*Xinyu Fang[1], *Maxwell Sun[2], Amanda Verga[3]

[1]fxinyu@umich.edu [2]mrsun@mit.edu [3]amanda.verga@trincoll.edu

MAA Southeastern Sectional Meeting, Special Session on the Theory of Integer Sequences, University of Tennessee

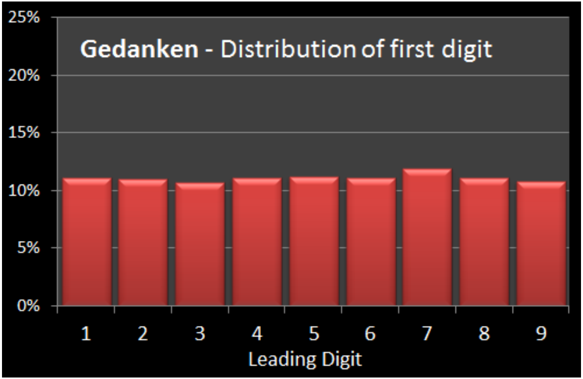
March 16, '24

Table of Contents

Table of Contents

Interesting Question

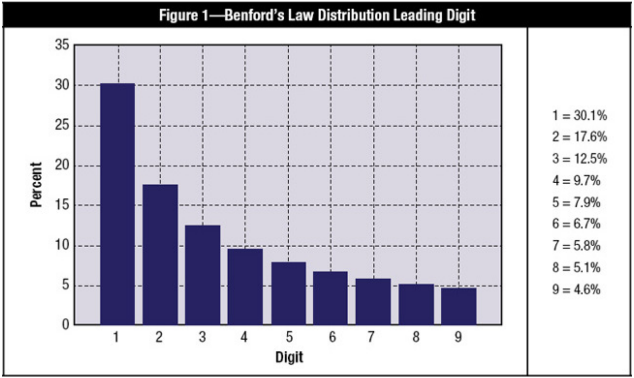
Motivating Question: For a nice data set, such as the Fibonacci numbers, stock prices, street addresses of college employees and students, ..., what percent of the leading digits are 1?



Natural guess: 10% (but immediately correct to 11%!).

Interesting Question

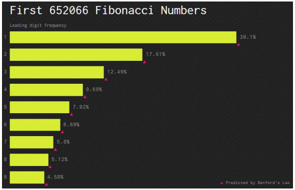
Motivating Question: For a nice data set, such as the Fibonacci numbers, stock prices, street addresses of college employees and students, ..., what percent of the leading digits are 1?



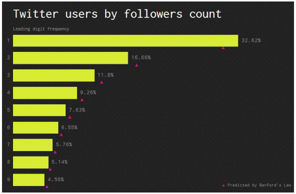
Answer: Benford's law!

Examples with First Digit Bias

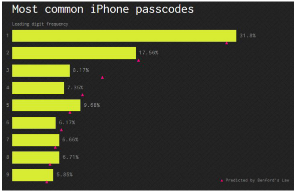
Fibonacci numbers



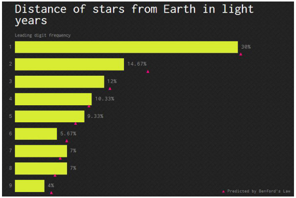
Twitter users by # followers



Most common iPhone passcodes



Distance of stars from Earth

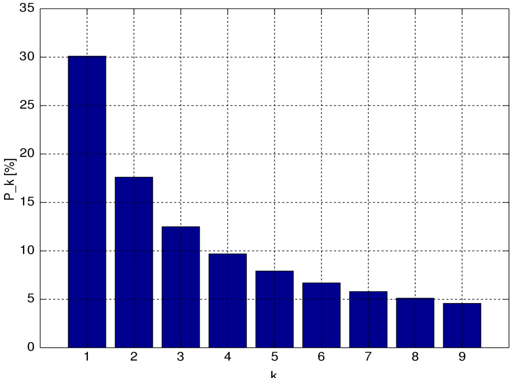


Benford's Law

Definition (Benford's Law)

A data set satisfies **Benford's Law base** B (where $B > 1$) if the probability of a first digit of d is $\log_B\left(\frac{d+1}{d}\right)$.

For example, when $B = 10$ (figure from Wikipedia):



Benford's Law

Definition (Benford's Law)

A data set is said to satisfy **Benford's Law base** B (where $B > 1$) if the probability of observing a value with first digit d is $\log_B\left(\frac{d+1}{d}\right)$.

Examples:

- special sequences and functions (e.g., $n!$ and the Fibonaccis)
- $3x + 1$ map (Kontorovich-Miller)
- financial data / fraud detection (Nigrini)
- **products of random variables** (Miller's REUs)

Background Material

- Modulo: $a = b \mod c$ if $a - b$ is an integer times c; thus $17 = 5 \mod 12$, and $4.5 = .5 \mod 1$.

- Significand: $x = S_{10}(x) \cdot 10^k$, k integer, $1 \leq S_{10}(x) < 10$. Thus $2024.1701 = 2.0241701 \cdot 10^3$.

- Mantissa: $M_{10}(x) = \log_{10} S_{10}(x)$.

- $S_{10}(a) = S_{10}(b)$ if and only if a and b have the same leading digits. Note $\log_{10} a = \log_{10} M_{10}(b) + k$.

- Key observation: $\log_{10}(x) = \log_{10}(\widetilde{x}) \bmod 1$ if and only if x and $\widetilde{x}$ have the same leading digits.

Thus often study $y = \log_{10} x \bmod 1$.
Advanced: $e^{2\pi i u} = e^{2\pi i (u \bmod 1)}$.

Strong Benford

Definition (Strong Benford)

A data set $\{x_n\}$ is **strong Benford base** B if $\{M_B(x_n)\}$ is distributed uniformly in $[0, 1]$. In other words, if

$$\mathbb{P}(M_B(x_n) \in [a, b]) \;=\; b - a$$

for all $[a, b] \subseteq [0, 1]$.

Equidistribution and Benford's Law

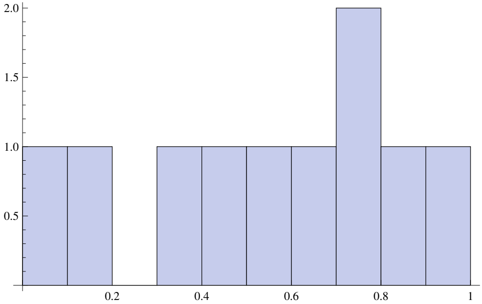
Equidistribution

$\{y_n\}_{n=1}^{\infty}$ is equidistributed modulo 1 if probability $y_n \bmod 1 \in [a, b]$ tends to $b - a$:

$$\frac{\#\{n \leq N : y_n \bmod 1 \in [a, b]\}}{N} \rightarrow b - a.$$

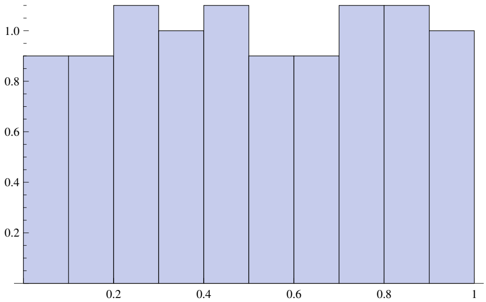
- Thm: $\beta \notin \mathbb{Q}$, $n\beta$ is equidistributed mod 1.
- Examples: $\log_{10} 2, \log_{10}\left(\frac{1+\sqrt{5}}{2}\right) \notin \mathbb{Q}$.

Example of Equidistribution: $n\sqrt{\pi} \bmod 1$



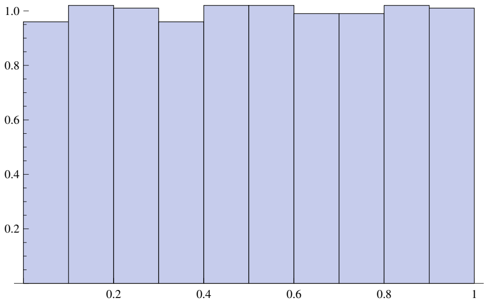
$n\sqrt{\pi} \bmod 1$ for $n \leq 10$

Example of Equidistribution: $n\sqrt{\pi}$ mod 1



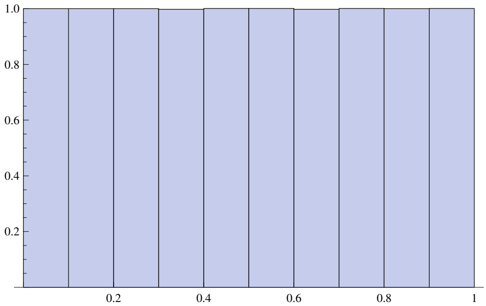
$n\sqrt{\pi}$ mod 1 for $n \leq 100$

Example of Equidistribution: $n\sqrt{\pi} \bmod 1$



$n\sqrt{\pi} \bmod 1$ for $n \leq 1000$

Example of Equidistribution: $n\sqrt{\pi} \bmod 1$



$n\sqrt{\pi} \bmod 1$ for $n \leq 10,000$

Logarithms and Benford's Law

Fundamental Equivalence

Data set $\{x_i\}$ is Benford base B if $\{y_i\}$ is equidistributed mod 1, where $y_i = \log_B x_i$.

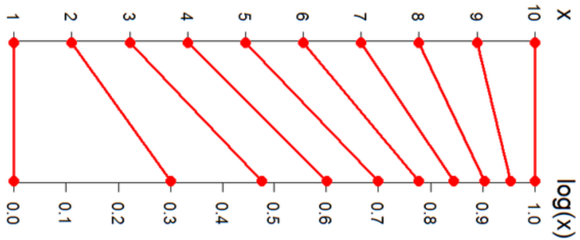
Logarithms and Benford's Law

Fundamental Equivalence

Data set $\{x_i\}$ is Benford base B if $\{y_i\}$ is equidistributed mod 1, where $y_i = \log_B x_i$.

$$x = S_{10}(x) \cdot 10^k \text{ then}$$

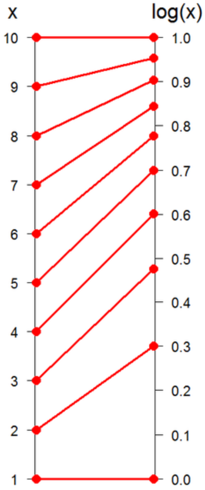
$$\log_{10} x = \log_{10} S_{10}(x) + k = \log_{10} S_{10} x \bmod 1.$$



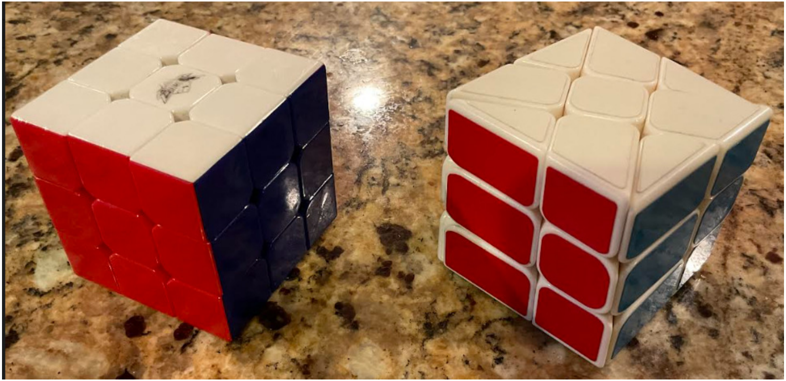
Logarithms and Benford's Law

$\mathrm{Prob}(\text{leading digit } d)$
$= \log_{10}(d+1) - \log_{10}(d)$
$= \log_{10}\left(\frac{d+1}{d}\right)$
$= \log_{10}\left(1+\frac{1}{d}\right).$

Have Benford's law $\leftrightarrow$ mantissa of logarithms of data are uniformly distributed



The Power of the Right Perspective



The Power of the Right Perspective

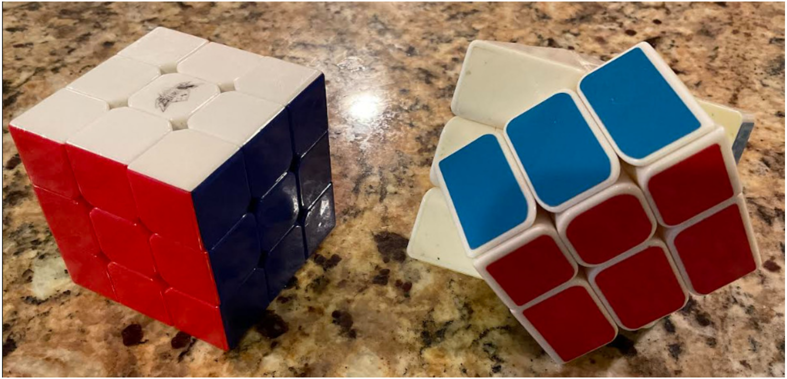


Table of Contents

Basic Stick Breaking Model

Start with a stick of length L. Choose a random point on the stick to break it in two, and repeat the process on each new stick obtained.

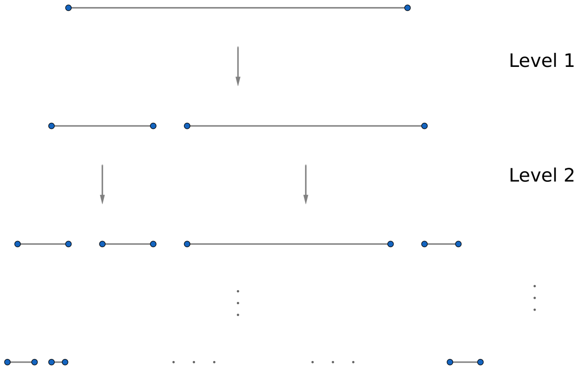


Figure 1: Illustration of stick breaking

Motivation from Physics

This process and its variations may be of interest to nuclear physicists for modelling particle decay (see Pain, *Benford's law and complex atomic spectra*, Physical Review E, 2008).

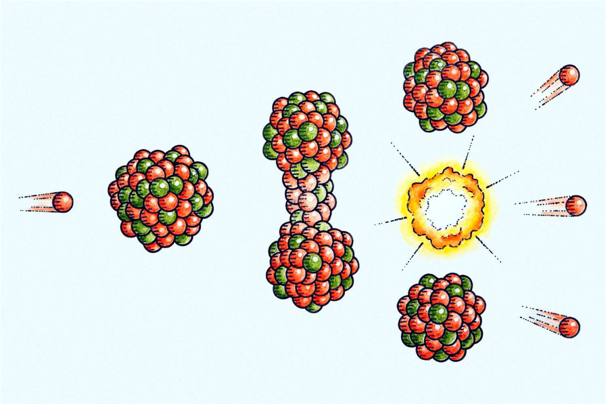


Figure 2: Random Stick Breaking similarities with Nuclear Fission.

Previous Results: Unrestricted Continuous Breaking

Theorem (Becker et. al.)

Fix some distribution $\mathcal{D}$ on $(0,1)$ satisfying a Mellin transform condition:

$$\lim_{n\to\infty} \sum_{\substack{\ell=-\infty \\ \ell\neq 0}}^{\infty} \prod_{m=1}^{n} \mathcal{M} f_{\mathcal{D}} \left(1 - \frac{2\pi i \ell}{\log B}\right) \ = \ 0.$$

Start with a stick of length L, and break in two with ratio sampled from $\mathcal{D}$. Repeat on all fragments for N levels, then the final collection of stick lengths converges to strong Benford as $N \to \infty$.

Benford's Law and Continuous Dependent Random Variables (Thealexa Becker, David Burt, Taylor C. Corcoran, Alec Greaves-Tunnell, Joseph R. Iafrate, Joy Jing, Steven J. Miller, Jaclyn D. Porfilio, Ryan Ronan, Jirapat Samranvedhya, Frederick W. Strauch and Blaine Talbut), Annals of Physics **388** (2018), 350–381.

https://arxiv.org/abs/1309.5603

Previous Results: Discrete One-Side Breaking

Theorem (Becker et. al.)

Start with a stick of integer length L. Choose an integer $X \in \{1, \cdots, L\}$ uniformly, and break off a fragment of length X. Repeat this process on the remaining stick $L - X$, until no more such breaking can be done. The final collection converges to strong Benford as $L \to \infty$.

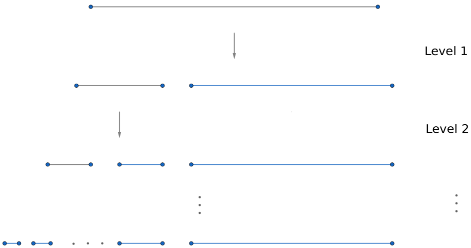


Figure 3: Illustration of discrete one-side breaking

Our Generalization: Discrete Breaking with Stopping Set

What if we break on both sides with extra *stopping conditions*?

- Fix $\mathfrak{S} \subseteq \mathbb{Z}_+$, *the stopping set*. Assume $1 \in \mathfrak{S}$.
- Declare a stick "dead" if its length falls into $\mathfrak{S}$ and do not break it further.
- Continue until all sticks are dead.

Our Generalization: Discrete Breaking with Stopping Set

What if we break on both sides with extra *stopping conditions*?

- Fix $\mathfrak{S} \subseteq \mathbb{Z}_+$, *the stopping set*. Assume $1 \in \mathfrak{S}$.
- Declare a stick "dead" if its length falls into $\mathfrak{S}$ and do not break it further.
- Continue until all sticks are dead.

Question

Which sets $\mathfrak{S}$ would lead to strong Benford behavior as $L \to \infty$?

Table of Contents

One-Sided Decomposition Conjecture

Theorem (Fang-Miller-Sun-Verga, 2023)

Start with an large odd integer length stick. Break it into two sticks, obtaining an even and an odd stick. Set the even stick aside and repeat the process on the resulting odd stick. As the initial length goes to ∞, the final empirical collection of sticks will converge to Benford behavior.

The above was conjectured by Becker et. al. We proved it and showed an even more general result.
Xinyu Fang, Steven J. Miller, Maxwell Sun, and Amanda Verga, *Generalized Continuous and Discrete Stick Fragmentation and Benford's Law*, preprint. https://arxiv.org/abs/2309.00766

Sharp Behavior Change

Theorem (Fang-Miller-Sun-Verga, 2023)

The process stops with probability 1 *and results in a collection of sticks that follows Benford's Law (in the limit) if and only if* $\mathfrak{S}$ *contains exactly* $n/2$ *residue classes.*

Sharp Behavior Change

Theorem (Fang-Miller-Sun-Verga, 2023)

The process stops with probability 1 *and results in a collection of sticks that follows Benford's Law (in the limit) if and only if* $\mathfrak{S}$ *contains exactly* $n/2$ *residue classes.*

With more residue classes, the mantissas are affected by the initial stick length. With less, we get lots of small sticks.

Idea of Proof

Idea used by Becker et. al.

1. Approximate the discrete process with a continuous analogue.

Idea of Proof

Idea used by Becker et. al.

1. Approximate the discrete process with a continuous analogue.
2. Show that the continuous analogue results in strong Benford behavior. (Easier!) [Key Input]

Idea of Proof

Idea used by Becker et. al.

1. Approximate the discrete process with a continuous analogue.
2. Show that the continuous analogue results in strong Benford behavior. (Easier!) [Key Input]
3. Deduce that the discrete process also results in strong Benford behavior by showing they are "close" enough. [Key Lemma]

Idea of Proof

Idea used by Becker et. al.

1. Approximate the discrete process with a continuous analogue.
2. Show that the continuous analogue results in strong Benford behavior. (Easier!) [Key Input]
3. Deduce that the discrete process also results in strong Benford behavior by showing they are “close” enough. [Key Lemma]

We also use this framework to prove our results.

Simulation Results: Stop At Odds, Many Trials

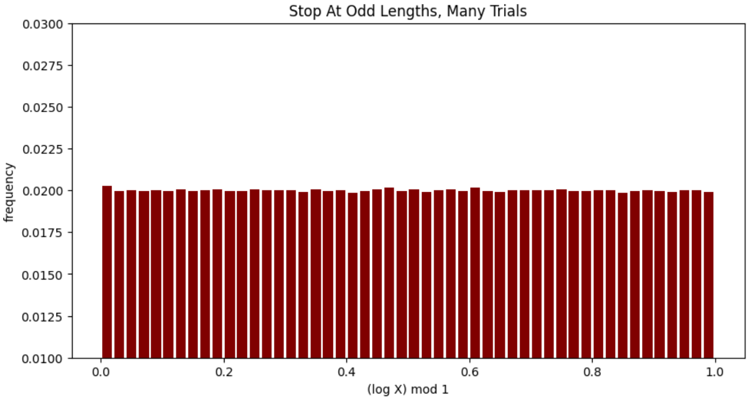


Figure 4: Histogram for $M_{10}(X)$, $L \approx 10^{1000}$, $R = 1000$ (R is the number of trials run with the same starting length L). The figure depicts the aggregated distribution of ending sticks from these trials.

Simulation Results: $n = 3$, stop at 1 residue class

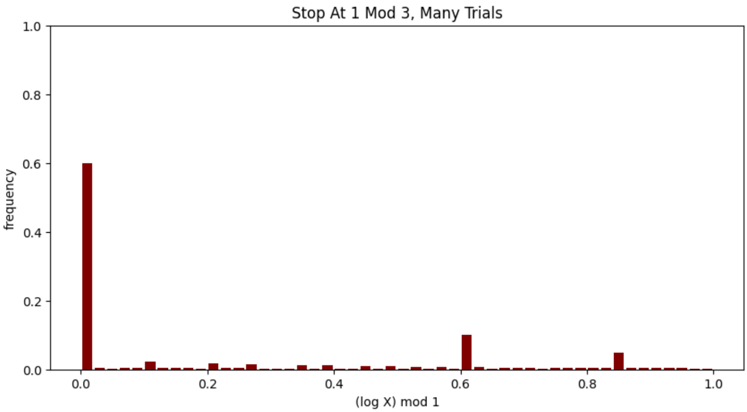


Figure 5: Histogram for $M_{10}(X)$, $L \approx 8 \cdot 10^{11}$, $R = 1000$.

Simulation Results: $n = 3$, stop at 2 residue classes

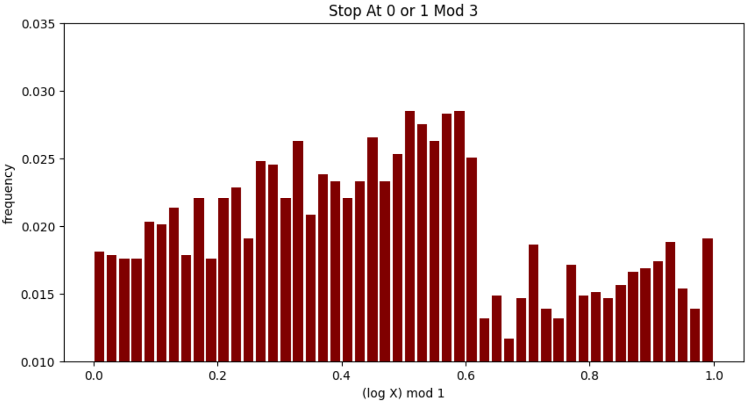


Figure 6: Histogram for $M_{10}(X)$, $L \approx 4 \cdot 10^{502}$, $R = 1000$.

Simulation Results: $n = 4$, stop at 2 residue classes

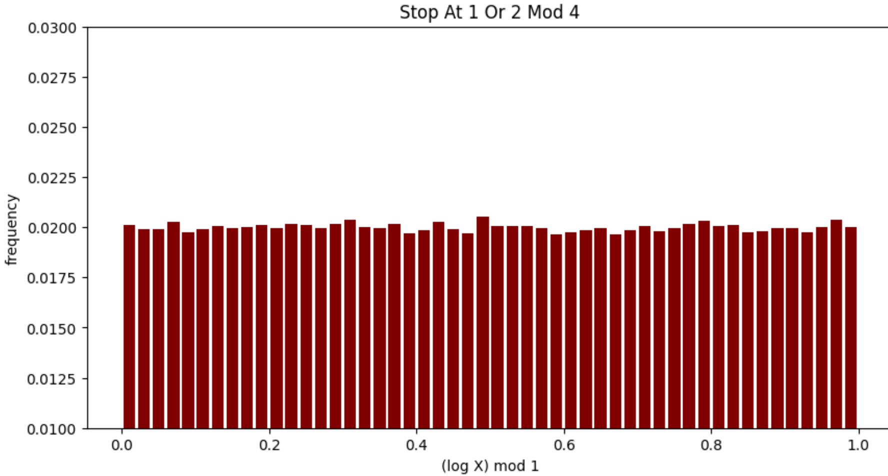


Figure 7: Histogram for $M_{10}(X)$, $L \approx 4 \cdot 10^{502}$, $R = 1000$.

When $|S| < n/2$: Non-Benford!

Theorem (Fang-Miller-Sun-Verga, 2023)

If $|S| < n/2$, then as $R \to \infty$ and $L \to \infty$, the collection of mantissas of ending stick lengths does not converge to any continuous distribution on $[0, 1]$. In particular, it does not converge to strong Benford behavior.

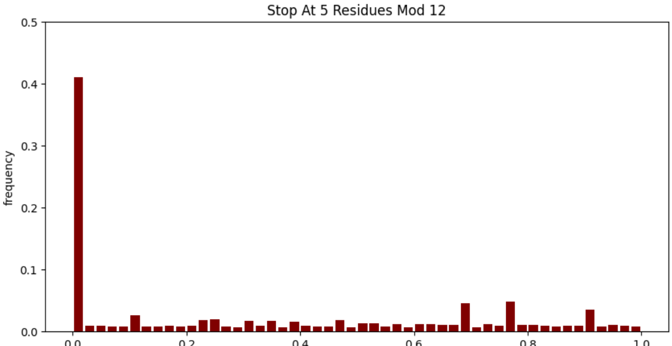


Table of Contents

References

References (Books)

- A. Berger and T. P. Hill, *An Introduction to Benford's Law*, Princeton University Press, Princeton, 2015. See also http://www.benfordonline.net/.

- A. E. Kossovsky, *Benford's Law: Theory, the General Law of Relative Quantities, and Forensic Fraud Detection Applications*, WSPC, 2014.

- S. J. Miller (editor), *Theory and Applications of Benford's Law*, Princeton University Press, 2015. https://web.williams.edu/Mathematics/sjmiller/public_html/benford/index.htm

- M. Nigrini, *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*, 1st Edition, Wiley, 2014.

Acknowledgements

Thank you!

Work supported by NSF Grants DMS2241623 and DMS1947438,
Williams College, and University of Michigan.

Table of Contents

Stick Decomposition

- T. Becker, D. Burt, T. C. Corcoran, A. Greaves-Tunnell, J. R. Iafrate, J. Jing, S. J. Miller, J. D. Porfilio, R. Ronan, J. Samranvedhya, F. W. Strauch and B. Talbut, *Benford's Law and Continuous Dependent Random Variables*, Annals of Physics **388** (2018), 350–381.
- J. Iafrate, S. J. Miller and F. W. Strauch, *Equipartitions and a distribution for numbers: A statistical model for Benford's law*, Physical Review E **91** (2015), no. 6, 062138 (6 pages).

Fixed Proportion Decomposition Process

Decomposition Process

1. Consider a stick of length $\mathcal{L}$.

Fixed Proportion Decomposition Process

Decomposition Process

1. Consider a stick of length $\mathcal{L}$.
2. Uniformly choose a proportion $p \in (0, 1)$.

Fixed Proportion Decomposition Process

Decomposition Process

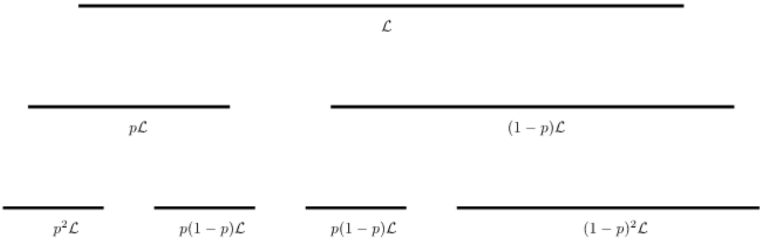
1. Consider a stick of length $\mathcal{L}$.
2. Uniformly choose a proportion $p \in (0, 1)$.
3. Break the stick into two pieces—lengths $p\mathcal{L}$ and $(1-p)\mathcal{L}$.

Fixed Proportion Decomposition Process

Decomposition Process

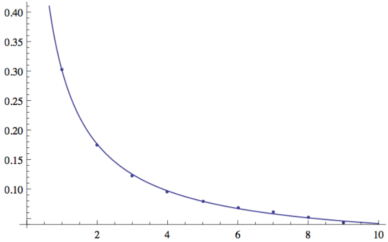
1. Consider a stick of length $\mathcal{L}$.
2. Uniformly choose a proportion $p \in (0, 1)$.
3. Break the stick into two pieces—lengths $p\mathcal{L}$ and $(1-p)\mathcal{L}$.
4. Repeat N times (using the same proportion).

Fixed Proportion Decomposition Process

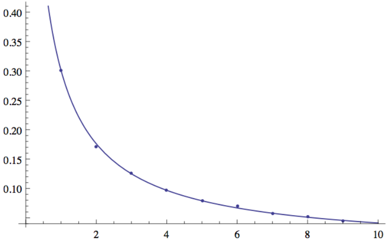


Fixed Proportion Conjecture (Joy Jing '13)

Conjecture: The above decomposition process is Benford as $N \to \infty$ for any $p \in (0, 1)$, $p \neq \frac{1}{2}$.



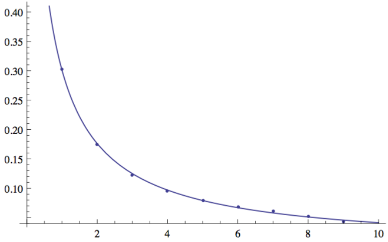
(B) $p = 0.51$ and $N = 10000$.



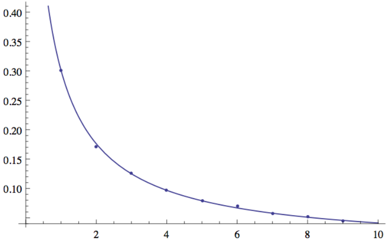
(B) $p = 0.99$ and $N = 50000$. Benford distribution overlaid.

Fixed Proportion Conjecture (Joy Jing '13)

Conjecture: The above decomposition process is Benford as $N \to \infty$ for any $p \in (0,1)$, $p \neq \frac{1}{2}$.



(B) $p = 0.51$ and $N = 10000$.



(B) $p = 0.99$ and $N = 50000$. Benford distribution overlaid.

Counterexample (SMALL REU '13): $p = \frac{1}{11}$, $1 - p = \frac{10}{11}$.

Benford Analysis

At N^{th} level,

- 2^N sticks
- $N+1$ distinct lengths: write $p^{N-j}(1-p)^j$ as

$$p^N\left(\frac{1-p}{p}\right)^j, \quad j \in \{0, \dots, N\}, \text{ have } \binom{N}{j} \text{ times.}$$

Benford Analysis

At N^{th} level,

- 2^N sticks
- $N+1$ distinct lengths: write $p^{N-j}(1-p)^j$ as

$$p^N\left(\frac{1-p}{p}\right)^j, \quad j\in\{0,\dots,N\}, \text{ have } \binom{N}{j} \text{ times.}$$

(Weighted) Geometric with ratio $\frac{1-p}{p}=10^y$; behavior depends on irrationality of y!

Benford Analysis

At N^{th} level,

- 2^N sticks
- $N+1$ distinct lengths: write $p^{N-j}(1-p)^j$ as

$$p^N\left(\frac{1-p}{p}\right)^j, \quad j\in\{0,\dots,N\}, \text{ have } \binom{N}{j} \text{ times.}$$

(Weighted) Geometric with ratio $\frac{1-p}{p} = 10^y$; behavior depends on irrationality of y!

Theorem: Benford if and only if y irrational.

Benford Analysis (cont)

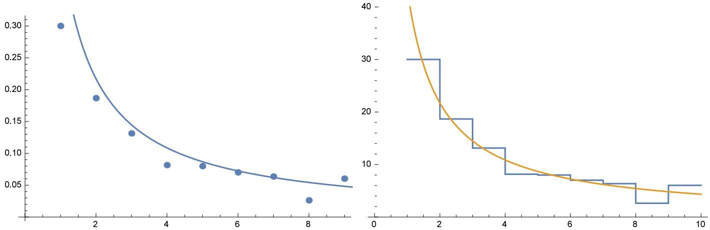
Say $\frac{1-p}{p} = 10^{r/q}$ for r, q integers.

All terms with index j mod q have same leading digit; probability index j mod q is

$$\begin{aligned}
\frac{1}{2^N}\left[\binom{N}{j}+\binom{N}{j+q}+\binom{N}{j+2q}+\cdots\right] &= \frac{1}{q}\sum_{s=0}^{q-1}\left(\cos\frac{\pi s}{q}\right)^N \cos\frac{\pi(N-2j)s}{q} \\
&= \frac{1}{q}\left(1+\sum_{s=1}^{q-1}\left(\cos\frac{\pi s}{q}\right)^N \cos\frac{\pi(N-2j)s}{q}\right) \\
&= \frac{1}{q}\left(1+\mathrm{Err}\left[(q-1)\left(\cos\frac{\pi}{q}\right)^N\right]\right),
\end{aligned}$$

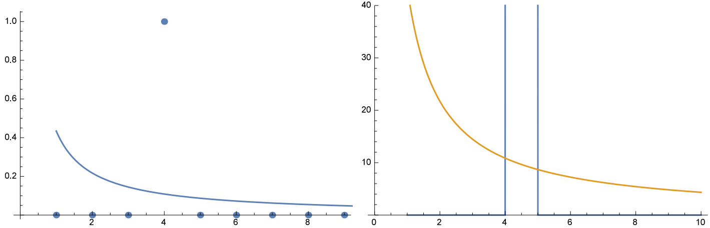
where $\mathrm{Err}[X]$ indicates an absolute error of size at most X

Examples



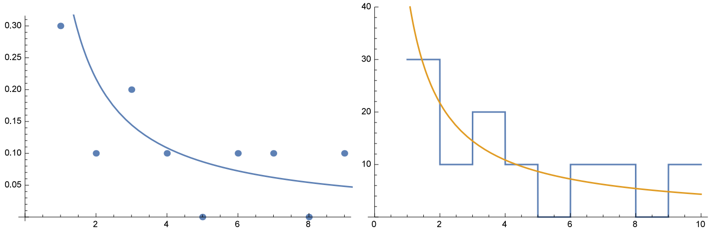
$p = 3/11$, 1000 levels; $y = \log_{10}(8/3) \notin \mathbb{Q}$ (irrational)

Examples



$p = 1/11$, 1000 levels; $y = 1 \in \mathbb{Q}$
(rational)

Examples



$p = 1/(1 + 10^{33/10})$, 1000 levels; $y = 33/10 \in \mathbb{Q}$ (rational)

Random Cuts

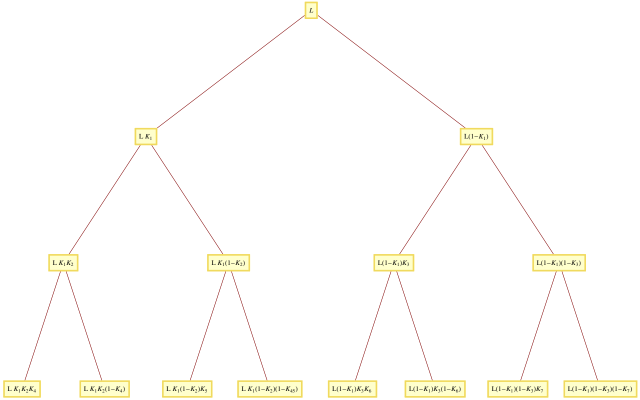


Figure 8: Unrestricted Decomposition: Breaking L into pieces, $N = 3$.