Benford's Law and Multinomial Distributions in Stick Fragmentation

Ella Lippelman
Colorado College
e_lippelman@coloradocollege.edu

Co-Author: Ava Irons
Advisor: Steven J. Miller
Williams College

YMC – Ohio State
August 14, 2024

Summary

- Benford's Law

Summary

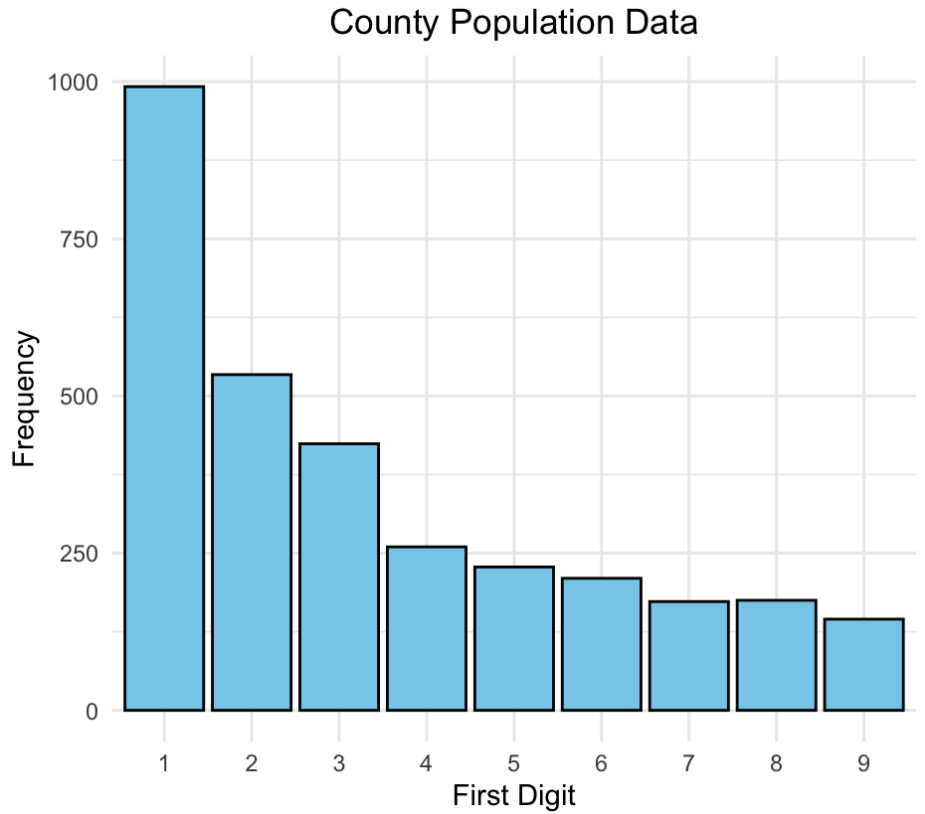
- Benford's Law
- Fixed Proportion Decomposition Model

Summary

- Benford's Law
- Fixed Proportion Decomposition Model
- Conjecture and Future Work

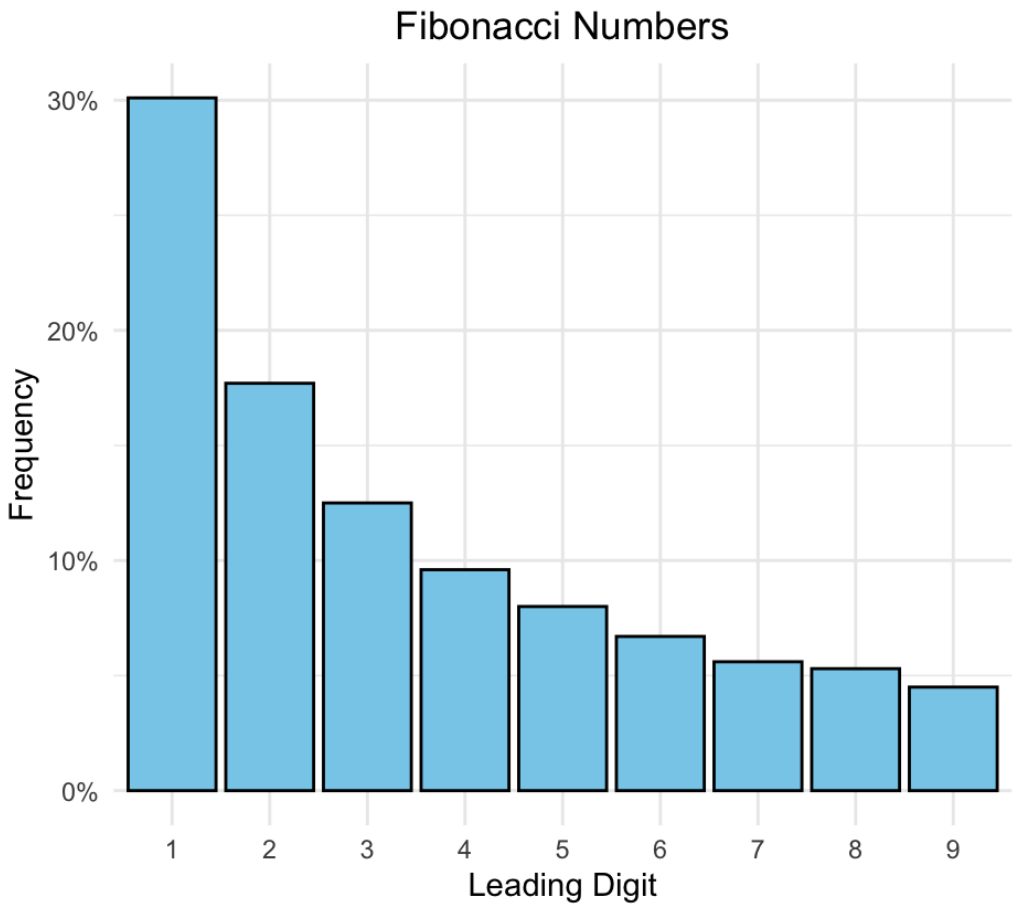
Consider the population of 3141 counties in the U.S. How often do you expect to see the leading digit 1 occur?

Consider the population of 3141 counties in the U.S. How often do you expect to see the leading digit 1 occur?



Consider the Fibonacci numbers. How often do you expect to see the leading digit 1 occur?

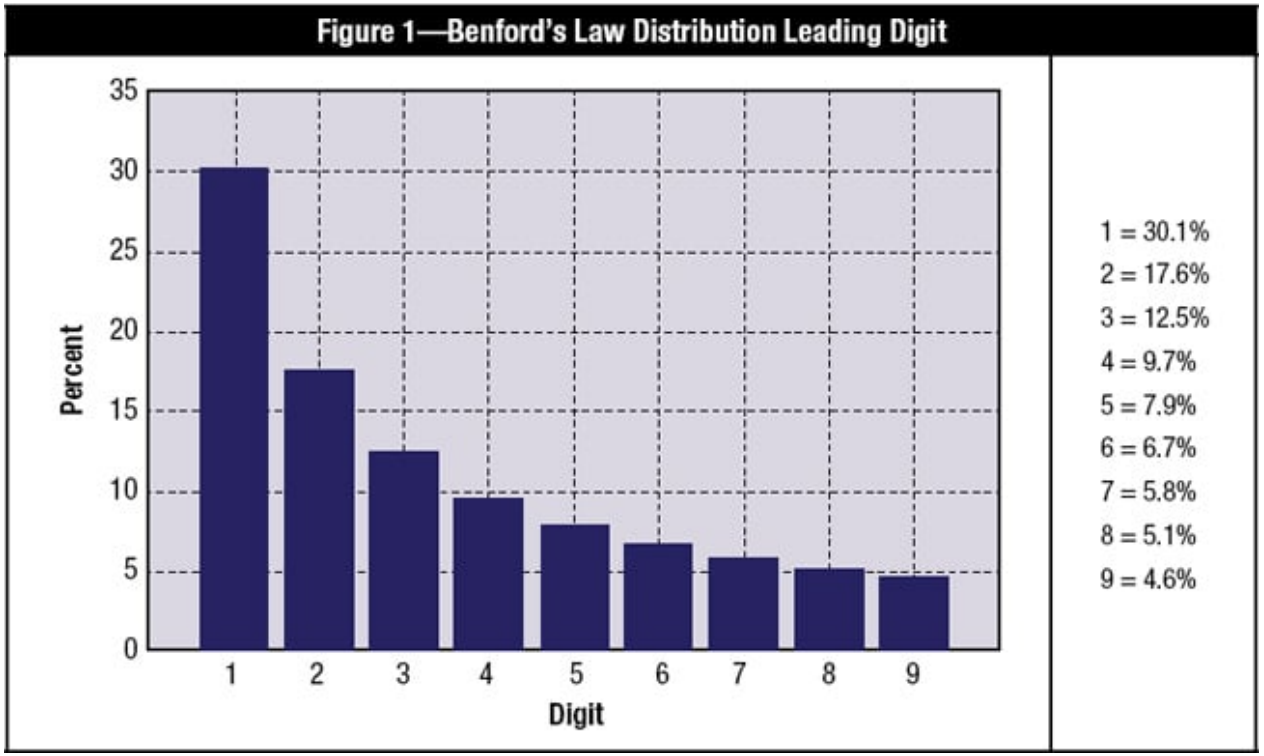
Consider the Fibonacci numbers. How often do you expect to see the leading digit 1 occur?



Benford's Law

Definition

A dataset is said to follow Benford's Law (base b) if the probability of observing a first digit of d is $\log_b \frac{1+d}{d}$.



Benford's Law and Logarithms Mod 1

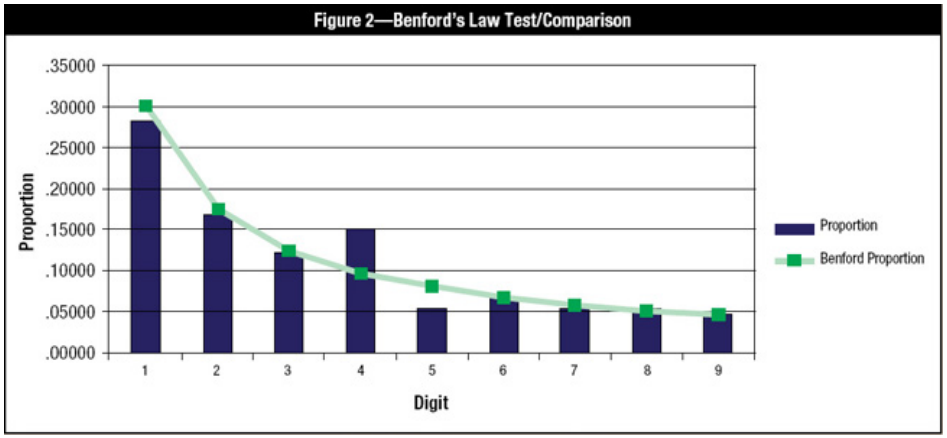
Uniform Distribution Modulo 1

A sequence (x_n) of real numbers is uniformly distributed modulo 1 if $\lim_{N\to\infty} \frac{\#\{1 \le n \le N : \langle x_n \rangle \le s\}}{N} = s$ for all $s \in [0, 1)$.

Uniform Characterization Theorem

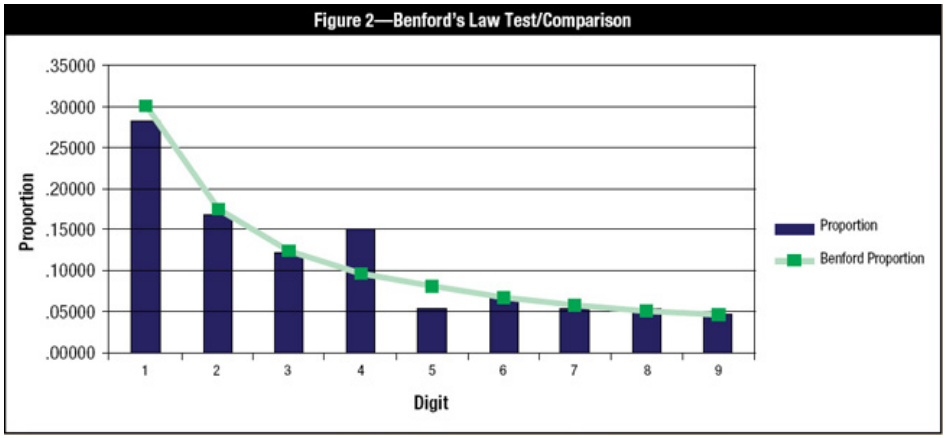
A sequence of real numbers is Benford base 10 if and only if the base 10 logarithms of its absolute values is uniformly distributed modulo 1.

Applications of Benford's Law



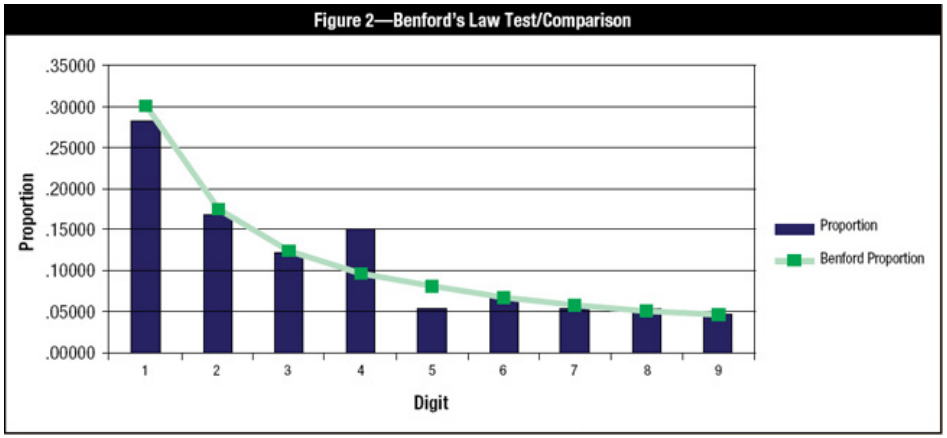
- Data integrity

Applications of Benford's Law



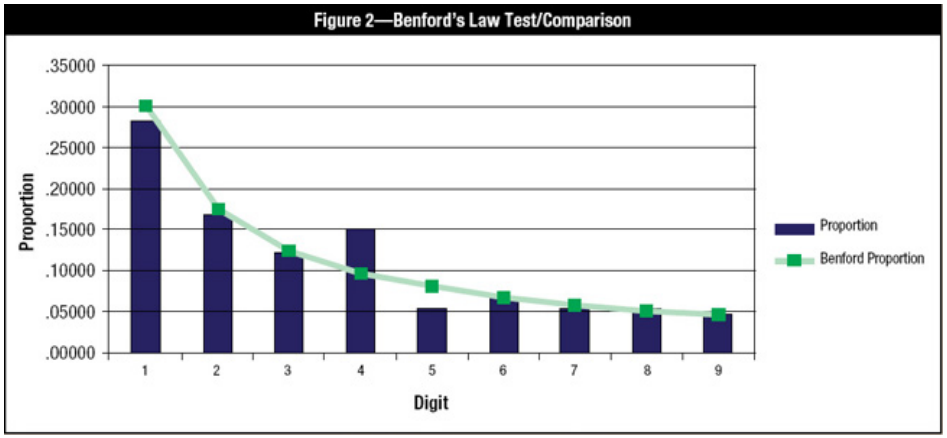
- Data integrity
- Fraud detection

Applications of Benford's Law



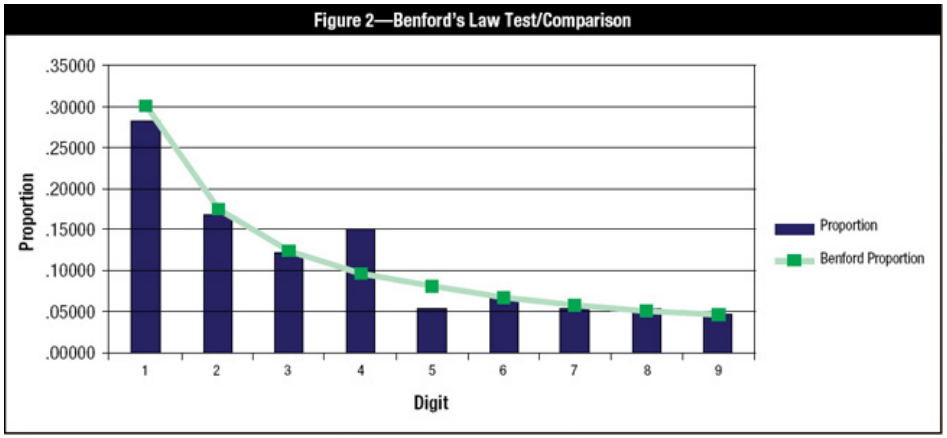
- Data integrity
- Fraud detection
  - taxes

Applications of Benford's Law



- Data integrity
- Fraud detection
  - taxes
  - insurance

Applications of Benford's Law



- Data integrity
- Fraud detection
  - taxes
  - insurance
  - expenses

Prior Work

- Focus on independent random variables.

Prior Work

- Focus on independent random variables.
- Janson and Neininger: number of pieces at the end of a tree fragmentation problem.

Prior Work

- Focus on independent random variables.
- Janson and Neininger: number of pieces at the end of a tree fragmentation problem.
- Lemons: fragmentation of conserved quantities.

Prior Work

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, B. Talbut and F. W. Strauch (Becker et. al)

Prior Work

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, B. Talbut and F. W. Strauch (Becker et. al)

- analyzes tree fragmentation

Prior Work

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, B. Talbut and F. W. Strauch (Becker et. al)

- analyzes tree fragmentation

Fixed proportion decomposition model

Prior Work

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, B. Talbut and F. W. Strauch (Becker et. al)

- analyzes tree fragmentation

Fixed proportion decomposition model

- cut a stick L at a fixed proportion $p \neq 0.5$

Prior Work

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, B. Talbut and F. W. Strauch (Becker et. al)

- analyzes tree fragmentation

Fixed proportion decomposition model

- cut a stick L at a fixed proportion $p \neq 0.5$
- 2 sticks after first cut

Prior Work

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, B. Talbut and F. W. Strauch (Becker et. al)

- analyzes tree fragmentation

Fixed proportion decomposition model

- cut a stick L at a fixed proportion $p \neq 0.5$
- 2 sticks after first cut
- N iterations, results in 2^N sticks with $N + 1$ distinct lengths

Fixed Proportion Decomposition

Fixed Proportion Decomposition Process

1. Consider a stick of length L.

Fixed Proportion Decomposition Process

1. Consider a stick of length L.
2. Choose finitely many fixed proportions $p_1, p_2, \ldots, p_{m-1}$ in $(0, 1)$.

Fixed Proportion Decomposition Process

1. Consider a stick of length L.
2. Choose finitely many fixed proportions $p_1, p_2, \ldots, p_{m-1}$ in $(0, 1)$.
3. In Stage 1, cut the stick at $p_1, p_2, \ldots, p_{m-1}$ to create m pieces.

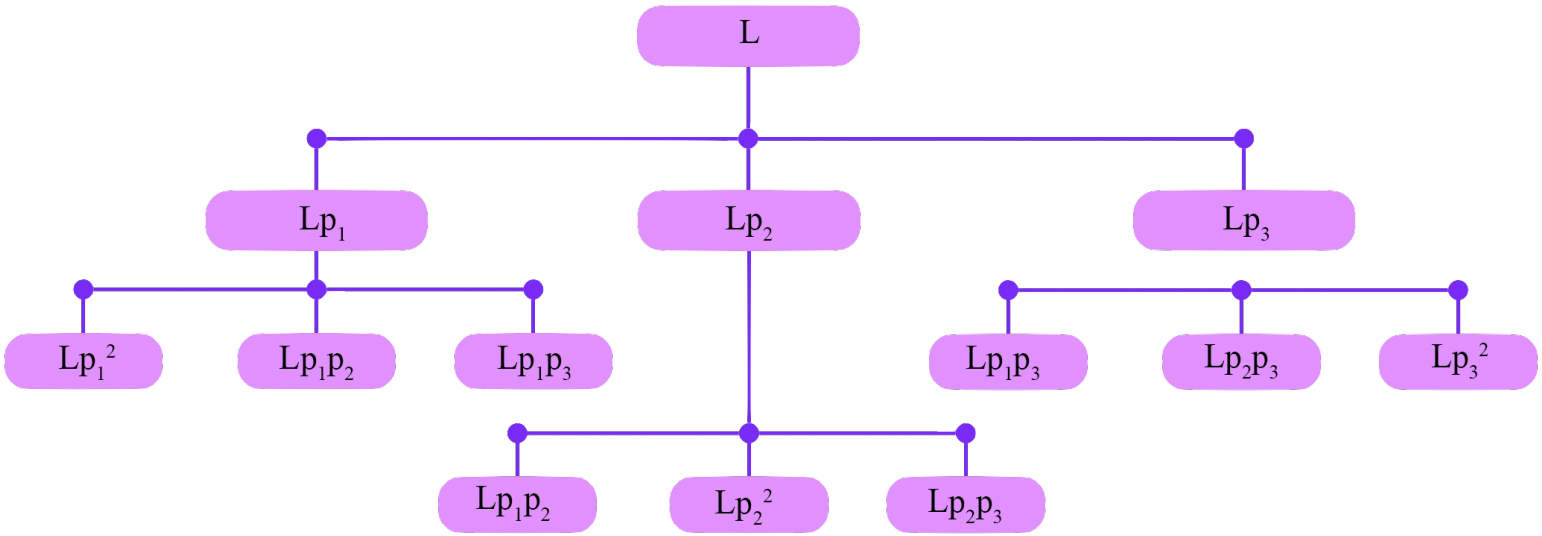
Fixed Proportion Decomposition Process

1. Consider a stick of length L.
2. Choose finitely many fixed proportions $p_1, p_2, \ldots, p_{m-1}$ in $(0, 1)$.
3. In Stage 1, cut the stick at $p_1, p_2, \ldots, p_{m-1}$ to create m pieces.
4. Repeat the process for N iterations, which results in

Fixed Proportion Decomposition Process

1. Consider a stick of length L.
2. Choose finitely many fixed proportions $p_1, p_2, \ldots, p_{m-1}$ in $(0, 1)$.
3. In Stage 1, cut the stick at $p_1, p_2, \ldots, p_{m-1}$ to create m pieces.
4. Repeat the process for N iterations, which results in
   - m^N sticks
   - $\binom{m+N-1}{N}$ distinct lengths for generic p_i

Fixed Proportion Decomposition Process Example



Example: Decomposition for $m = 3$ and $N = 2$.

Our model follows a multinomial distribution.

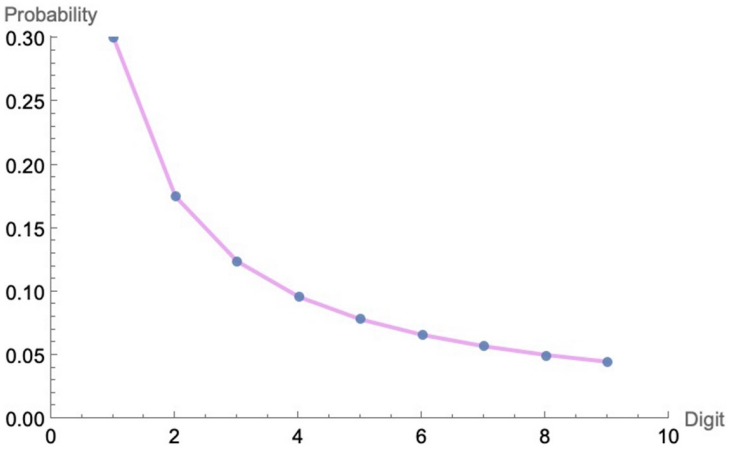
Our model follows a multinomial distribution.

Multinomial Coefficient

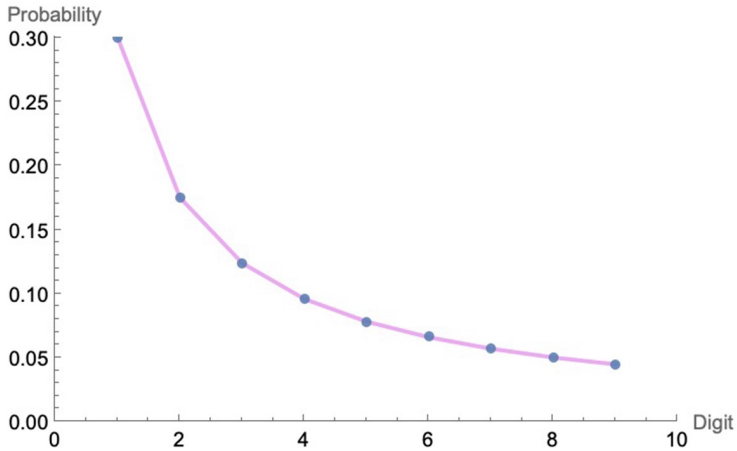
For any nonnegative integer N and positive integer m, we define the multinomial coefficient, for non-negative k summing to N, by $\binom{N}{k_1,k_2,\ldots,k_m} = \frac{N!}{k_1!k_2!\cdots k_m!}$.

Multinomial Theorem

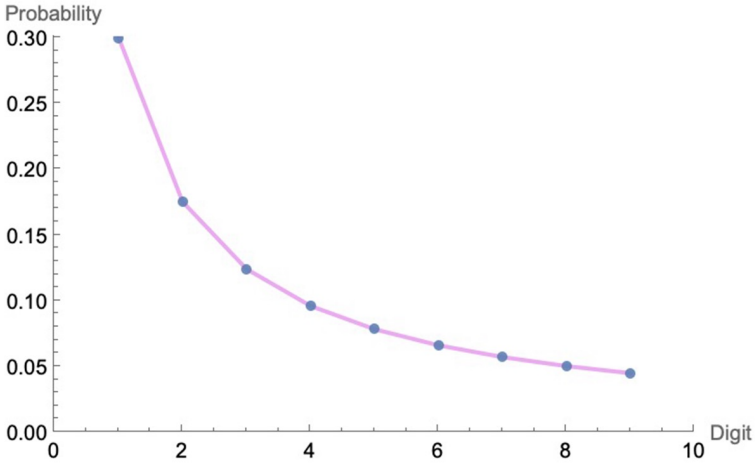
Let N be any non-negative integer and $p_1, p_2, \ldots, p_m$ real numbers. Then $(p_1 + p_2 + \cdots + p_m)^N = \sum_{k_1+k_2+\cdots+k_m\geq 0} \binom{N}{k_1,k_2,\ldots,k_m} p_1^{k_1} p_2^{k_2} \cdots p_m^{k_m}$, where the k_j's are non-negative integers and 0^0 is interpreted as unity.



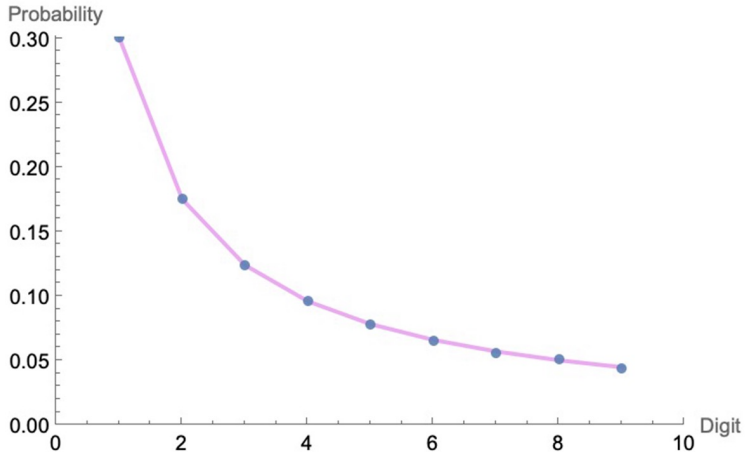
$p_1 = 0.05, p_2 = 0.16, N = 1000$



$p_1 = 0.07, p_2 = 0.34, N = 1000$



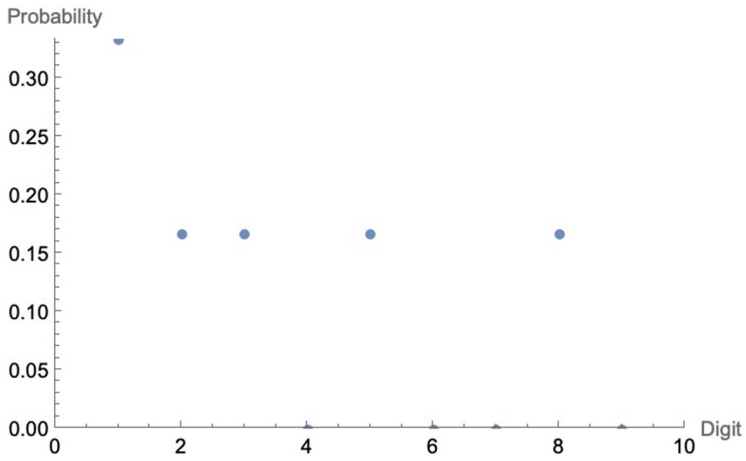
$p_1 = 0.11, p_2 = 0.23, N = 1000$



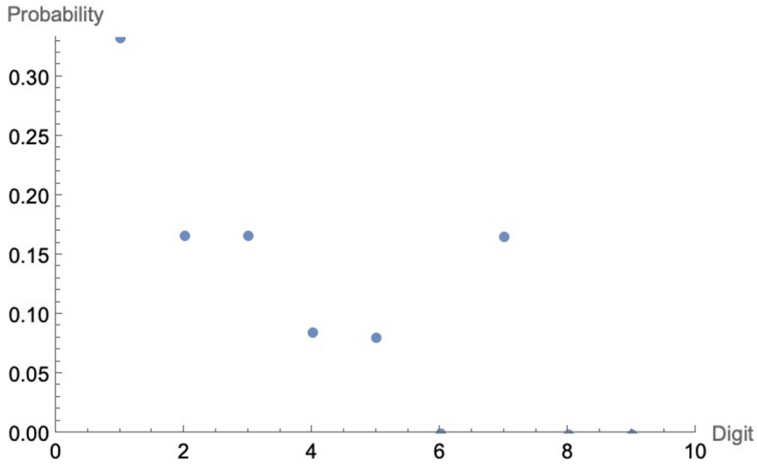
$p_1 = 0.13, p_2 = 0.3, N = 1000$

Theorem (Irons-Lippelman-Miller)

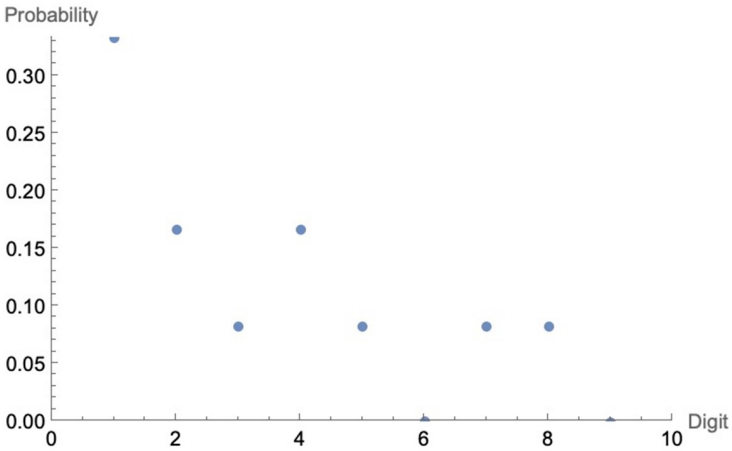
Consider the above described decomposition process. Choose $y_1, y_2, \ldots, y_{m-1}$ such that $10^{y_1} = p_1/p_2$, $10^{y_2} = p_2/p_3, \ldots, 10^{y_{m-1}} = p_{m-1}/p_m$, which are the ratios of adjacent lengths. The decomposition process results in a distribution of stick lengths that do not follow Benford's Law if all $y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$.



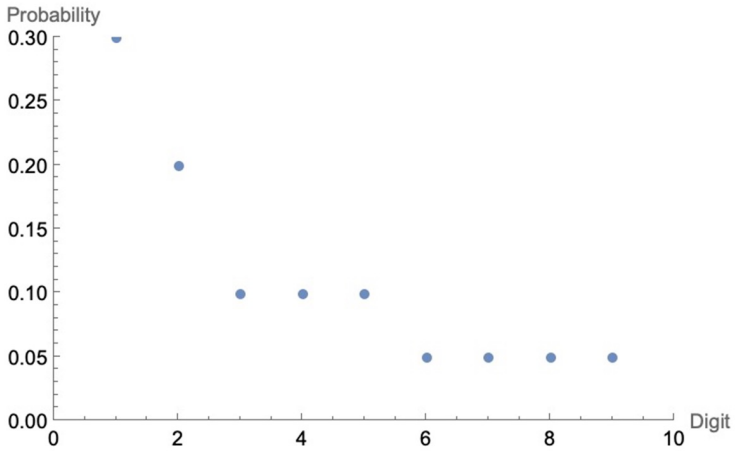
$p_1 \approx 0.1003, p_2 \approx 0.2161, y_1, y_2 \in \mathbb{Q}, N = 1000$



$p_1 \approx 0.0911, p_2 \approx 0.2881, y_1, y_2 \in \mathbb{Q}, N = 500$



$p_1 \approx 0.1856, p_2 \approx 0.33, y_1, y_2 \in \mathbb{Q}, N = 1000$



$p_1 \approx 0.1851, p_2 \approx 0.2933, y_1, y_2 \in \mathbb{Q}, N = 500$

$y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$

Break a stick $m - 1$ times at distinct proportions $p_1, p_2, \ldots, p_{m-1}$.

$y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$

Break a stick $m-1$ times at distinct proportions $p_1, p_2, \ldots, p_{m-1}$.

The frequency of lengths follows the m-multinomial distribution, as the lengths are given by

$$p_1^{k_1} \cdot p_2^{k_2} \cdots p_m^{k_m},$$

$y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$

Break a stick $m-1$ times at distinct proportions $p_1, p_2, \ldots, p_{m-1}$.

The frequency of lengths follows the m-multinomial distribution, as the lengths are given by

$$p_1^{k_1} \cdot p_2^{k_2} \cdots p_m^{k_m},$$

where k_j is the number of times the stick was cut at proportion p_i,

$$0 \leq k_1 \leq N, 0 \leq k_2 \leq N - k_1, \ldots, 0 \leq k_m \leq N - \sum_{j=1}^{m-1} k_j.$$

$y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$

Multiplying $p_1^{k_1} \cdot p_2^{k_2} \cdots p_m^{k_m}$ by one yields

$y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$

Multiplying $p_1^{k_1} \cdot p_2^{k_2} \cdots p_m^{k_m}$ by one yields

$$= \mathrm{p}_1^{k_1} \cdot \left(\frac{p_2}{p_2}\right)^{k_1} \cdot p_2^{k_2} \cdots p_m^{k_m}$$

$y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$

Multiplying $p_1^{k_1} \cdot p_2^{k_2} \cdots p_m^{k_m}$ by one yields

$$= \mathrm{p}_1^{k_1} \cdot \left(\frac{p_2}{p_2}\right)^{k_1} \cdot p_2^{k_2} \cdots p_m^{k_m}$$

$$= \left(\frac{p_1}{p_2}\right)^{k_1} \cdot p_2^{k_1+k_2} \cdots p_m^{k_m}$$

$y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$

Multiplying $p_1^{k_1} \cdot p_2^{k_2} \cdots p_m^{k_m}$ by one yields

$$
\begin{aligned}
&= p_1^{k_1} \cdot \left(\frac{p_2}{p_2}\right)^{k_1} \cdot p_2^{k_2} \cdots p_m^{k_m} \\
&= \left(\frac{p_1}{p_2}\right)^{k_1} \cdot p_2^{k_1+k_2} \cdots p_m^{k_m} \\
&= \left(\frac{p_1}{p_2}\right)^{k_1} \cdot \left(\frac{p_2}{p_3}\right)^{k_1+k_2} \cdot \left(\frac{p_3}{p_4}\right)^{k_1+k_2+k_3} \cdots \left(\frac{p_{m-1}}{p_m}\right)^{\sum_{j=1}^{m-1} k_j} \cdot (p_m)^N .
\end{aligned}
$$

$y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$

Multiplying $p_1^{k_1} \cdot p_2^{k_2} \cdots p_m^{k_m}$ by one yields

$$
\begin{aligned}
&= p_1^{k_1} \cdot \left(\frac{p_2}{p_2}\right)^{k_1} \cdot p_2^{k_2} \cdots p_m^{k_m} \\
&= \left(\frac{p_1}{p_2}\right)^{k_1} \cdot p_2^{k_1+k_2} \cdots p_m^{k_m} \\
&= \left(\frac{p_1}{p_2}\right)^{k_1} \cdot \left(\frac{p_2}{p_3}\right)^{k_1+k_2} \cdot \left(\frac{p_3}{p_4}\right)^{k_1+k_2+k_3} \cdots \left(\frac{p_{m-1}}{p_m}\right)^{\sum_{j=1}^{m-1} k_j} \cdot (p_m)^N .
\end{aligned}
$$

Let $q_i = \sum_{j=1}^{i} k_j$. The above equation becomes

$$\left(\frac{p_1}{p_2}\right)^{q_1} \cdot \left(\frac{p_2}{p_3}\right)^{q_2} \cdot \left(\frac{p_3}{p_4}\right)^{q_3} \cdots \left(\frac{p_{m-1}}{p_m}\right)^{q_{m-1}} \cdot (p_m)^N .$$

$y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$

Let $10^{y_i} = p_i/p_{i+1}$, where $y_i \in \mathbb{Q}$. Then there must exist $a_i \in \mathbb{Z}$ and $b_i \in \mathbb{N}$ where $\gcd(a_i, b_i) = 1$ for all i where $y_i = a_i/b_i$.

$y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$

Let $10^{y_i} = p_i/p_{i+1}$, where $y_i \in \mathbb{Q}$. Then there must exist $a_i \in \mathbb{Z}$ and $b_i \in \mathbb{N}$ where $\gcd(a_i, b_i) = 1$ for all i where $y_i = a_i/b_i$.

Substituting in we obtain

$$(p_m)^N \cdot \prod_{i=1}^{m-1} (10^{y_i})^{q_i} = (p_m)^N \cdot \prod_{i=1}^{m-1} \left(10^{\frac{a_i}{b_i}}\right)^{q_i}.$$

$y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$

Let $10^{y_i} = p_i/p_{i+1}$, where $y_i \in \mathbb{Q}$. Then there must exist $a_i \in \mathbb{Z}$ and $b_i \in \mathbb{N}$ where $\gcd(a_i, b_i) = 1$ for all i where $y_i = a_i/b_i$.

Substituting in we obtain

$$(p_m)^N \cdot \prod_{i=1}^{m-1} (10^{y_i})^{q_i} = (p_m)^N \cdot \prod_{i=1}^{m-1} \left(10^{\frac{a_i}{b_i}}\right)^{q_i}.$$

Thus, for $1 \leq i \leq m-1$, the significands of $10^{a_i/b_i}$ repeat every b_i indices.

$y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$

Thus, the significands repeat periodically and are scaled by p_m^N.

$y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$

Thus, the significands repeat periodically and are scaled by p_m^N.

Therefore we have at most $\prod_{i=1}^{m-1} b_i$ distinct significands.

$y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$

Thus, the significands repeat periodically and are scaled by p_m^N.

Therefore we have at most $\prod_{i=1}^{m-1} b_i$ distinct significands.

- significands take on only finitely many values modulo one, thus $(0, 1)$ not covered entirely

$y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$

Thus, the significands repeat periodically and are scaled by p_m^N.

Therefore we have at most $\prod_{i=1}^{m-1} b_i$ distinct significands.

- significands take on only finitely many values modulo one, thus $(0, 1)$ not covered entirely
- cannot be uniformly distributed modulo one

$y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$

Thus, the significands repeat periodically and are scaled by p_m^N.

Therefore we have at most $\prod_{i=1}^{m-1} b_i$ distinct significands.

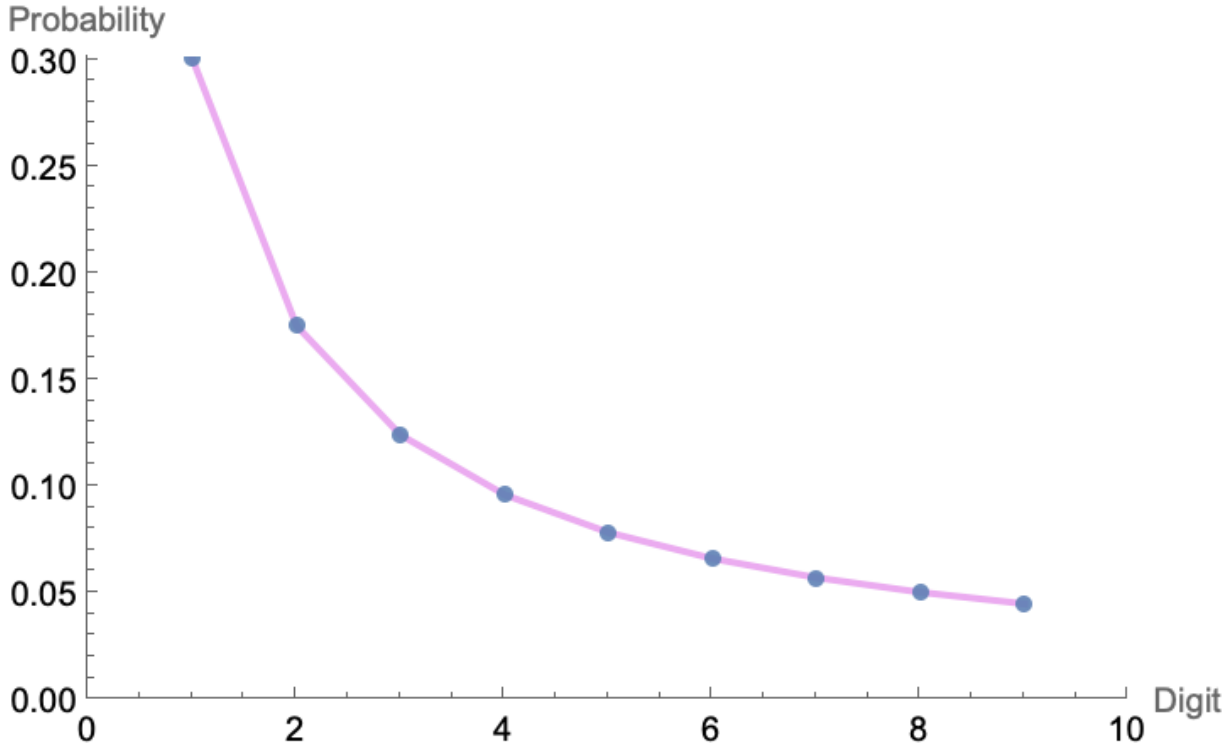
- significands take on only finitely many values modulo one, thus $(0, 1)$ not covered entirely
- cannot be uniformly distributed modulo one

Thus, by the Uniform Characterization Theorem, the distribution of lengths cannot follow Benford's Law.

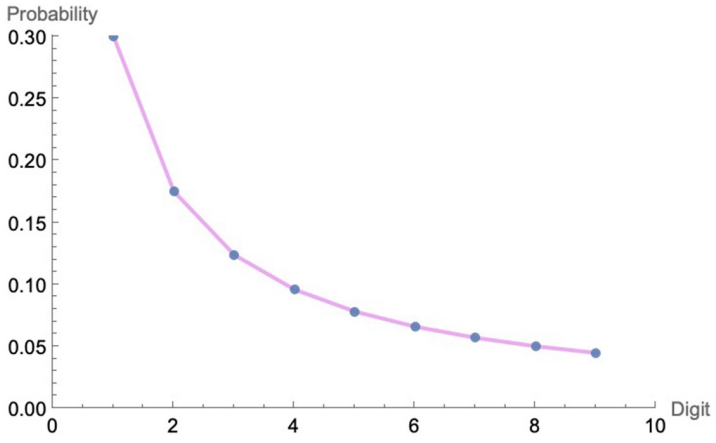
Conjecture, Future Work, and Conclusions

Conjecture

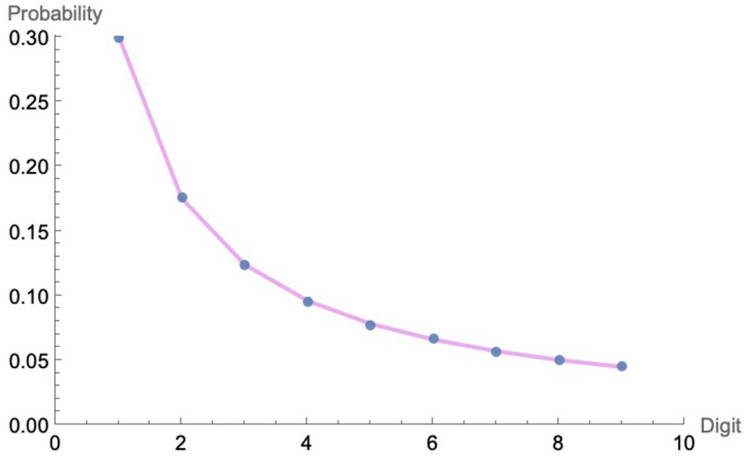
Choose $y_1, y_2, \ldots, y_{m-1}$ such that $10^{y_1} = p_1/p_2$, $10^{y_2} = p_2/p_3, \ldots, 10^{y_{m-1}} = p_{m-1}/p_m$, which is the ratio of adjacent lengths. If $y_1, y_2, \ldots, y_{m-1}$ are all irrational, then the distribution of stick lengths follow Benford's Law.



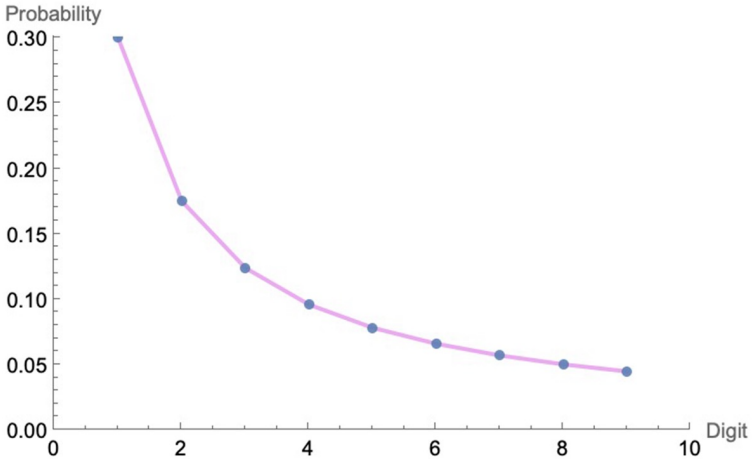
$p_1 = 0.05, p_2 = 0.2,$ and $N = 1000$



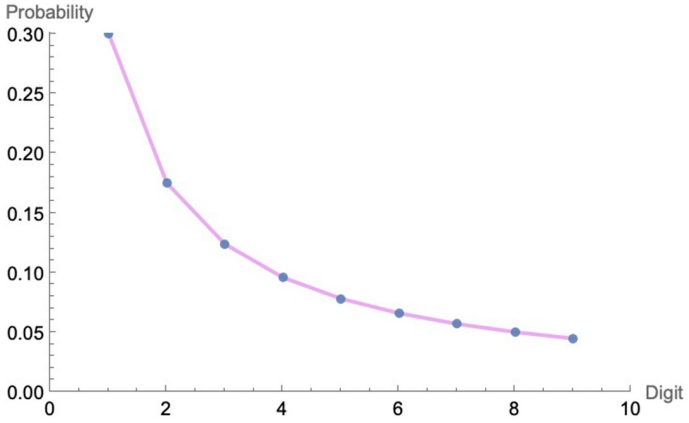
$p_1 = 0.1, p_2 = 0.3, y_1, y_2 \notin \mathbb{Q}, N = 1000$



$p_1 = 0.1, p_2 = 0.25, y_1, y_2 \notin \mathbb{Q}, N = 1000$

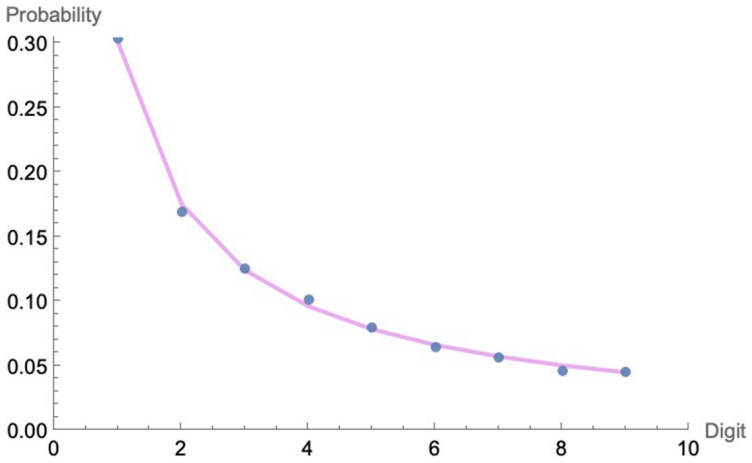


$p_1 = 0.09, p_2 = 0.39, y_1, y_2 \notin \mathbb{Q}, N = 1000$

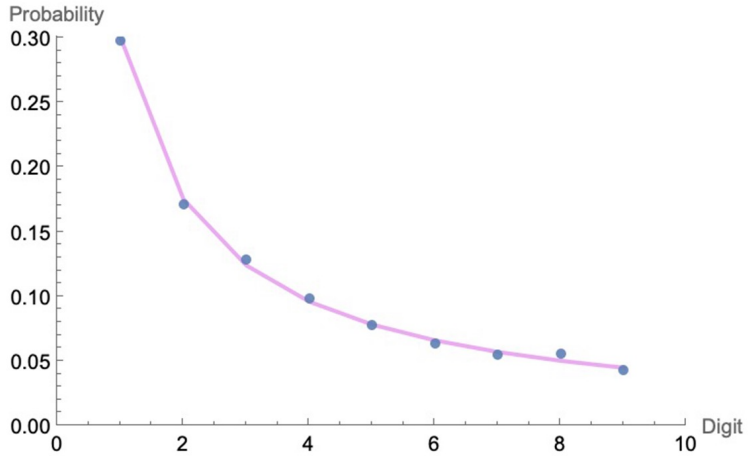


$p_1 = 0.05, p_2 = 0.25, y_1, y_2 \notin \mathbb{Q}, N = 1000$

Exploring the mixed case where some $y_i \in \mathbb{Q}$ and some y_i are irrational.



$p_1 \approx 0.0058, p_2 \approx 0.3152, y_1 \notin \mathbb{Q}, y_2 \in \mathbb{Q}, N = 1000$



$p_1 \approx 0.1160, p_2 \approx 0.3667, y_1 \in \mathbb{Q}, y_2 \notin \mathbb{Q}, N = 1000$

Future Work

- Prove Conjecture.

Future Work

- Prove Conjecture.
- Handling mixed cases where some y_i are irrational and some $y_i \in \mathbb{Q}$.

Conclusions

- Independent and Dependent Random Variables

Conclusions

- Independent and Dependent Random Variables
- Fixed Proportion Decomposition
  - $y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$, not Benford
- Future Work

Conclusions

- Independent and Dependent Random Variables
- Fixed Proportion Decomposition
  - $y_1, y_2, \ldots, y_{m-1} \in \mathbb{Q}$, not Benford
- Future Work
  - $y_1, y_2, \ldots, y_{m-1}$ are all irrational
  - some y_i are irrational and some $y_i \in \mathbb{Q}$
  - other models

Acknowledgements

I would like to thank my advisor, Steven J. Miller, and my co-author Ava Irons.

I would also like to thank the Williams College SMALL summer research program.

This research was funded by the Williams College Summer Science Program Research Fellowship, Finnerty Fund, and NSF Grant DMS2241623.

References

- 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, B. Talbut and F. W. Strauch, *Benford's Law and Continuous Dependent Random Variables*, Annals of Physics, **388** (2018), 350–381.
- *Benford's Law Distribution Leading Digit.* Digital Image. ISACA. N.p., Web. 06 August 2024. https://www.isaca.org/resources/isaca-journal/past-issues/2011/understanding-and-applying-benfords-law.
- P. Diaconis, *The distribution of leading digits and uniform distribution mod 1*, Ann. Probab. **5** (1979), 72–81.
- S. Janson and R. Neininger, *The size of random fragmentation trees*, Probab. Theory Related Fields **142** (2008), no. 3-4, 399–442.

References

- K. K. Kataria, *A Probabilistic Proof of the Multinomial Theorem*, The American Mathematical Monthly, **123** (2016), 94–96.
- D. S. Lemons, *On the number of things and the distribution of first digits*, Am. J. Phys. **54** (1986), no. 9, 816–817.
- S. J. Miller (editor), *Theory and Applications of Benford's Law*, Princeton University Press, 2015.
- Singleton, Tommie W., Ph.D. Benfordâs Law Test/Comparison. Digital image. ISACA. N.p., 2011. Web. 16 July 2013, https://www.isaca.org/resources/isaca-journal/past-issues/2011/understanding-and-applying-benfords-law.