Benford Behavior of Dependent Random Variables

Taylor Corcoran - University of Arizona
taylorc3@email.arizona.edu
Jaclyn Porfilio - Williams College
Jaclyn.D.Porfilio@williams.edu

Co-Authors: Joseph Iafrate, Jirapat Samranvedhya
Advisor: Steven J. Miller
Williams College

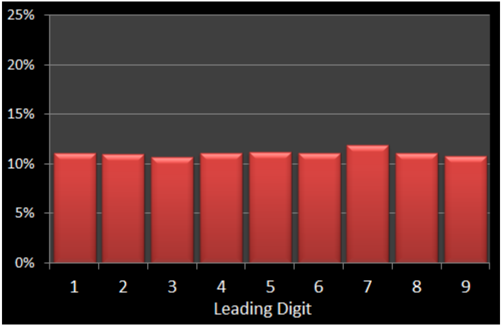
YMC - Ohio State
Friday, August 9th 2013

Summary

- Benford's Law
- Stick Decomposition
- Conjectures and Other Dependent Systems

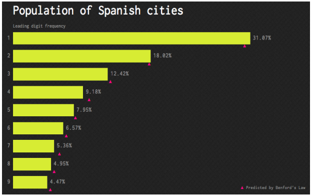
Consider the dataset of the populations of Spanish cities.
How often do you expect a leading digit of 1 to occur?

Consider the dataset of the populations of Spanish cities.
How often do you expect a leading digit of 1 to occur?



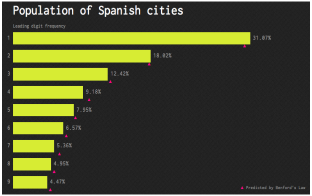
First Digit Bias

Population of Spanish Cities

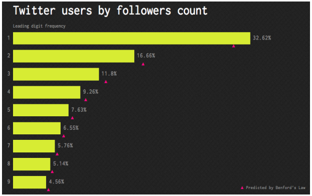


First Digit Bias

Population of Spanish Cities

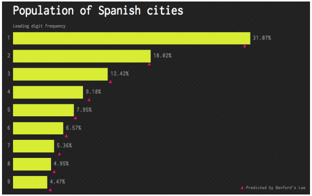


Twitter Followers per User

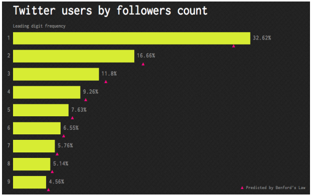


First Digit Bias

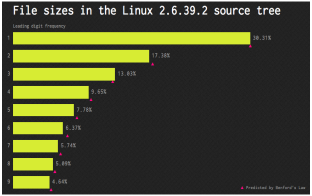
Population of Spanish Cities



Twitter Followers per User

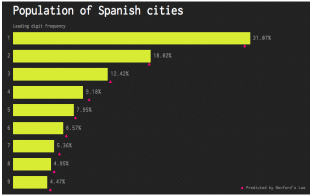


File Sizes in Linux Source Tree

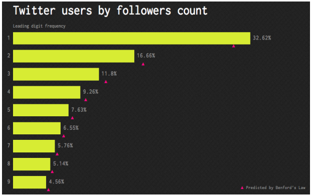


First Digit Bias

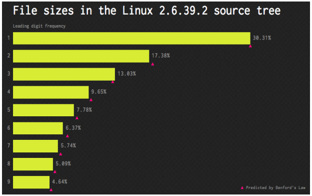
Population of Spanish Cities



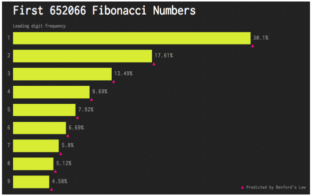
Twitter Followers per User



File Sizes in Linux Source Tree



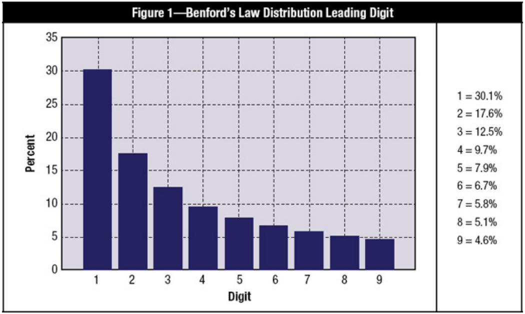
Fibonacci numbers



Benford's Law

Definition

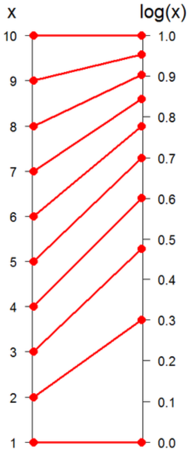
A dataset is said to follow **Benford's Law** (base b) if the probability of observing a first digit of d is $\log_b \frac{1+d}{d}$.



Logarithms and Benford's Law

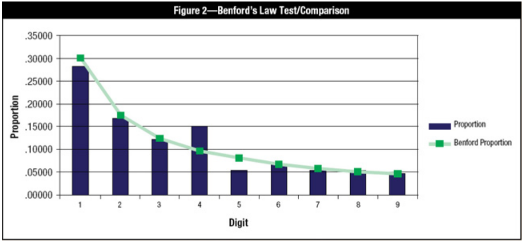
P(leading digit d) $=$ $\log_{10}(d+1) - \log_{10}(d)$

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



Applications of Benford's Law

- Fraud detection
- Data integrity
- Analyzing round-off errors



Previous Work

- Arithmetic operations on random variables.
- Reliance on independence of random variables.
- Becker, Greaves-Tunnell, Miller, Ronan, Strauch: techniques to deal with dependencies.
- Lemons: process of fragmenting a conserved quantity.

Fixed Proportion Decomposition

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

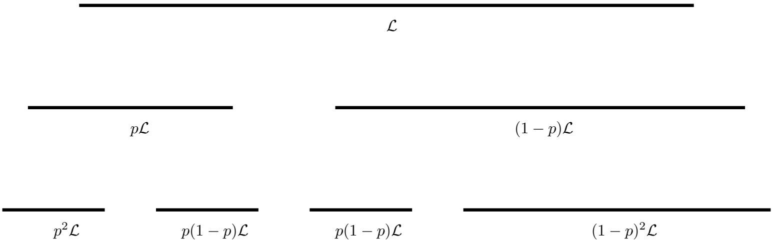
1. Consider a stick of length $\mathcal{L}$.
2. Uniformly choose a proportion $p \in (0, 1)$.
3. Break the stick into two pieces: lengths $p\mathcal{L}$ and $(1 - p)\mathcal{L}$.

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).

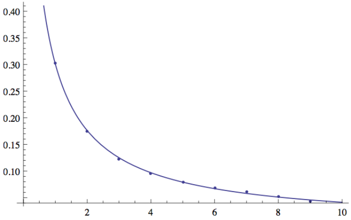
Fixed Proportion Decomposition Process



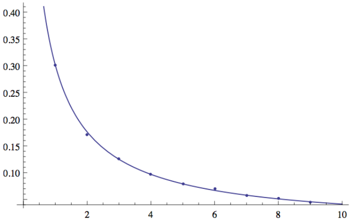
Fixed Proportion Conjecture

Joy Jing's Conjecture

The above decomposition process results in stick lengths that obey Benford's Law as $N \to \infty$ for any $p \in (0, 1)$, $p \neq \frac{1}{2}$.

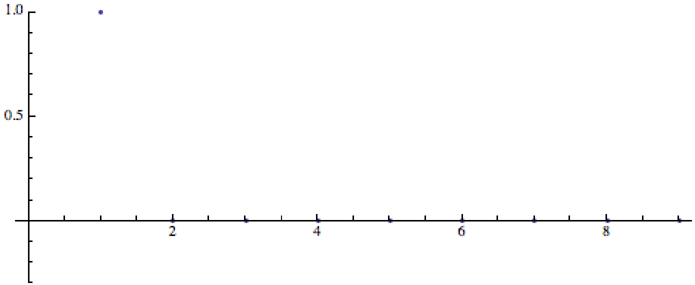


(B) $p = 0.51$ and $N = 10000$.



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

Counterexample: $p = \frac{1}{11},\ 1 - p = \frac{10}{11}$.



Benford Analysis

After Nth interation,

- 2^N sticks
- $N+1$ distinct lengths.

Distinct lengths are given by

$$x_{j+1} \;=\; \left(\frac{1-p}{p}\right) x_j,\;\; x_0 = p^N.$$

Frequency of $x_j = \binom{N}{j}$

Benford Analysis

After Nth interation,

- 2^N sticks
- $N+1$ distinct lengths.

Distinct lengths are given by

$$x_{j+1} = \left(\frac{1-p}{p}\right) x_j, \; x_0 = p^N.$$

Frequency of $x_j = \binom{N}{j}$

Let $\frac{1-p}{p} = 10^y$.

$\frac{1-p}{p} = 10^y$, **y** $\in \mathbb{Q}$

Theorem

Let $\frac{1-p}{p} = 10^y$. If $y \in \mathbb{Q}$, the described decomposition process results in stick lengths that do not obey Benford's Law.

$\frac{1-p}{p} = 10^y$, y $\in \mathbb{Q}$

Theorem

Let $\frac{1-p}{p} = 10^y$. If $y \in \mathbb{Q}$, the described decomposition process results in stick lengths that do not obey Benford's Law.

Let $y = \frac{r}{q}$.

Leading digit of x_j repeats every q indices. Thus,

$$\sum_k P(x_{j+kq}) = \sum_k \binom{N}{j+kq}.$$

Series Multisection

Multisection Formula

If $f(x) = \sum_{n=-\infty}^{\infty} a_n x^n$,

$$\sum_{k=-\infty}^{\infty} a_{kq+j} x^{kq+j} = \frac{1}{q} \sum_{p=0}^{q-1} \omega^{-jp} f(\omega^p x)$$

where ω is the primitive qth root of unity $e^{2\pi i/q}$.

Multisection of Binomial Coefficients

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

$\frac{1-p}{p} = 10^y,\ y \in \mathbb{Q}$

$$\sum_k P(x_{j+kq}) = \frac{1}{q}\left(1 + \mathscr{E}\left[(q-1)\left(\cos\frac{\pi}{q}\right)^N\right]\right)$$

$\frac{1-p}{p} = 10^y$, $y \in \mathbb{Q}$

$$\sum_k P(x_{j+kq}) = \frac{1}{q}\left(1 + \mathscr{E}\left[(q-1)\left(\cos\frac{\pi}{q}\right)^N\right]\right)$$

Digit frequencies are multiples of $\frac{1}{q}$.

Benford frequencies are irrational, so *not* perfect Benford.

$\frac{1-p}{p} = 10^y$, y $\notin \mathbb{Q}$: Outline

Theorem

Let $\frac{1-p}{p} = 10^y$. If $y \notin \mathbb{Q}$, the described decomposition process results in stick lengths that obey Benford's Law.

$\frac{1-p}{p} = 10^y$, y $\notin \mathbb{Q}$: Outline

Theorem

Let $\frac{1-p}{p} = 10^y$. If $y \notin \mathbb{Q}$, the described decomposition process results in stick lengths that obey Benford's Law.

$\{x_j\} \sim Bin(N, \frac{1}{2})$

mean: $\frac{N}{2}$

standard deviation: $\frac{\sqrt{N}}{2}$

Outline of proof strategy:

1. Truncation
2. Break into intervals
   - Roughly equal probability
   - Equidistribution

$\frac{1-p}{p} = 10^y$, $y \notin \mathbb{Q}$: Truncation

For $\epsilon > 0$, Chebyshev's Inequality gives

$$\begin{aligned} P\left(\left|x - \frac{N}{2}\right| \geq N^{\frac{1}{2}+\epsilon}\right) &= P\left(\left|x - \frac{N}{2}\right| \geq N^{\epsilon} N^{\frac{1}{2}}\right) \\ &\leq \frac{1}{N^{2\epsilon}}. \end{aligned}$$

$\frac{1-p}{p} = 10^y$, $y \notin \mathbb{Q}$: Truncation

For $\epsilon > 0$, Chebyshev's Inequality gives

$$\begin{aligned} P\left(\left|x - \frac{N}{2}\right| \geq N^{\frac{1}{2}+\epsilon}\right) &= P\left(\left|x - \frac{N}{2}\right| \geq N^{\epsilon} N^{\frac{1}{2}}\right) \\ &\leq \frac{1}{N^{2\epsilon}}. \end{aligned}$$

So we can limit our analysis to

- N^{ϵ} standard deviations
- Right half of binomial

$\frac{1-p}{p} = 10^y$, y $\notin \mathbb{Q}$: Intervals and Roughly Equal Probability

$\mathcal{I}_\ell = \{x_\ell, x_\ell + 1, \dots, x_\ell + N^\delta - 1\}$.

Let $x_0 = N/2$. It follows that $x_\ell = N/2 + \ell N^\delta$.

$$\left| \binom{N}{x_\ell} - \binom{N}{x_{\ell+1}} \right| \leq \binom{N}{x_\ell} N^{-\frac{1}{2}+\delta+\epsilon},$$

when $\delta < 1/2 - \epsilon$ and $\ell \leq N^{1/2-\delta+\epsilon}$.

$\frac{1-p}{p} = 10^y$, y $\notin \mathbb{Q}$: Equidistribution

Definition

$\{x_n\}_{n=1}^{\infty}$ is equidistributed modulo 1 if for any $[a, b] \subset [0, 1]$, $P\left(x_n \mod 1 \in [a, b]\right) \to b - a$:

$$\lim_{N\to\infty} \frac{\#\{n \leq N : x_n \mod 1 \in [a, b]\}}{N} = b - a.$$

Recall: Leading digits of stick lengths are Benford if their logarithms are equidistributed modulo 1.

$\frac{1-p}{p} = 10^y$, y $\notin \mathbb{Q}$: Equidistribution

Consider an interval I_ℓ where

$$I_\ell = \{x_\ell + i : i \in \{0, 1, \dots, N^\delta - 1\}\}$$

$$J_\ell \subset \{0, 1, \dots, N^\delta - 1\} = \{i : \log(x_\ell + i) \bmod 1 \in [a, b]\}.$$

$\frac{1-p}{p} = 10^y$, y $\notin \mathbb{Q}$: Equidistribution

Consider an interval I_ℓ where

$I_\ell = \{x_\ell + i : i \in \{0, 1, \ldots, N^\delta - 1\}\}$

$J_\ell \subset \{0, 1, \ldots, N^\delta - 1\} = \{i : \log(x_\ell + i) \bmod 1 \in [a, b]\}$.

If the irrationality exponent κ of y is finite,

$$|J_\ell| \;=\; (b-a)N^\delta + O(N^{\delta(1-\frac{1}{\kappa}+\epsilon)})$$

$\frac{1-p}{p} = 10^y$, y $\notin \mathbb{Q}$: Equidistribution

Using

- equidistribution within intervals
- roughly equal probability

we have

$$\sum_{\ell} \sum_{i \in J_\ell} f(x_\ell + i) \quad = \quad (b-a) + O(N^{\delta(-\frac{1}{\kappa}+\epsilon)} + N^{-\frac{1}{2}+\delta+\epsilon}).$$

where the irrationality exponent κ of y is finite.

Additive Decomposition

Additive Decomposition Process

Decomposition Process

1. Consider a stick of length $\mathcal{L}$.
2. Break the stick into two pieces, both of integer length.
3. Freeze a piece if its length exists in the specified stopping sequence.
4. Repeat decomposition with pieces that have not been frozen.

Additive Decomposition Processes: Conjectures

Benford

- Stop at evens (proved)
- Stop at primes

Non-Benford

- Stop at squares
- Stop at powers of two
- Stop at powers of three
- Stop at Fibonnaci numbers

Continuous Model

Continuous Model

Approach:

- Draw cut proportions from uniform distributon on $(0, 1)$.

Continuous Model

Approach:

- Draw cut proportions from uniform distributon on $(0, 1)$.
- Label sticks according to number of proportions in their product.

Continuous Model Stick Labeling

Iteration 0: $\mathcal{L}$

Continuous Model Stick Labeling

Iteration 0: $\mathcal{L}$

Iteration 1: $p_1\mathcal{L}$ $x_1 = (1 - p_1)\mathcal{L}$

Continuous Model Stick Labeling

Iteration 0: $\mathcal{L}$

Iteration 1: $p_1\mathcal{L}$ $x_1 = (1 - p_1)\mathcal{L}$

Iteration 2: $p_2p_1\mathcal{L}$ $x_2 = (1 - p_2)p_1\mathcal{L}$

Continuous Model Stick Labeling

Iteration 0: $\mathcal{L}$

Iteration 1: $p_1\mathcal{L}$ $\qquad x_1 = (1 - p_1)\mathcal{L}$

Iteration 2: $p_2 p_1\mathcal{L}$ $\qquad x_2 = (1 - p_2)p_1\mathcal{L}$

Iteration N: $p_N p_{N-1} \cdots p_1\mathcal{L}$ $\qquad x_N = (1 - p_N)p_{N-1} \cdots p_1\mathcal{L}$

Continuous Model

Approach:

- Draw cut proportions from uniform distributon on $(0, 1)$.
- Label sticks according to number of proportions in their product.

Continuous Model

Approach:

- Draw cut proportions from uniform distributon on $(0, 1)$.
- Label sticks according to number of proportions in their product.
- Do not consider first $\log N$ sticks as well as pairs of sticks x_i, x_j where i and j differ by less than $\log N$.

Continuous Model

Approach:

- Draw cut proportions from uniform distributon on $(0, 1)$.
- Label sticks according to number of proportions in their product.
- Do not consider first $\log N$ sticks as well as pairs of sticks x_i, x_j where i and j differ by less than $\log N$.
- Show $\mathbb{E}[P_N(s)] \to \log s$ and $\mathrm{Var}[P_N(s)] \to 0$.

Other Work and Conclusions

Determinant Expansions

Theorem

Let A be an $n \times n$ matrix with i.i.d. entries $a_{ij} \sim X$ with density f. The $n!$ terms in the determinant expansion of A are Benford if

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

Determinant Expansions

Theorem

Let A be an $n \times n$ matrix with i.i.d. entries $a_{ij} \sim X$ with density f. The $n!$ terms in the determinant expansion of A are Benford if

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

Important Technique: Quantify dependencies among terms.

Conclusions

- Dependent vs Independent Random Variables
- Stick Decomposition
  - Fixed Proportion Decomposition $\frac{1-p}{p} = 10^y$

    - $y \in \mathbb{Q}$, not Benford

    - $y \notin \mathbb{Q}$, Benford (finite vs infinite κ)
  - Additive Stick Decomposition

    - Conjectures

    - Continuous Model
- Determinant Expansion

Acknowledgements

We would like to thank our advisor, Steven J. Miller, our co-authors Joseph R. Iafrate, Jirapat Samranvedhya, B. G. Opher as well as Frederick W. Strauch and Joy Jing.

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

This research was funded by NSF grant DMS0850577.

Irrationality Exponent

Let $y \in \mathbb{R}$. Denote by $\mathcal{A}$ the set of positive numbers κ for which

$$0 \leq \left| y - \frac{p}{q} \right| \leq \frac{1}{q^n}$$

has at most finitely many solutions for $p, q \in \mathbb{Z}$.
The irrationality measure of y, denoted $\kappa(y)$, is $\inf_{\kappa \in \mathcal{A}} \kappa$.

If $\mathcal{A}$ is empty, $\kappa(y) = \infty$
For nonempty $\mathcal{A}$,

$$\kappa(y) = \begin{cases} 1 \text{ if } y \text{ is rational} \\ 2 \text{ if } y \text{ is algebraic of degree } > 1 \\ \geq 2 \text{ if } y \text{ is transcendental} \end{cases}$$

Proof of Multisection Formula I

As $f(x) = \sum_{n=0}^{\infty} a_n x^n$,

$$\sum_{j=0}^{k-1} \omega^{-jm} f(\omega^j x) = \sum_{j=0}^{k-1} \omega^{-jm} \sum_{n=0}^{\infty} a_n (\omega^j x)^n$$

$$= \sum_{n=0}^{\infty} a_n \left(\sum_{j=0}^{k-1} \omega^(n-m)j \right) x^n$$

Proof of Multisection Formula II

If $n - m = kl$ for some $l \in \mathbb{Z}$, then using the fact that $\omega^k = 1$ gives

$$\omega^{(n-m)j} = \omega^{klj} = 1$$

which gives $\sum_{j=0}^{k-1} \omega^{(n-m)j} = k$ If $n - m \neq lk$,

$$\sum_{j=0}^{k-1} \omega^{(n-m)j} = \frac{1 - \omega^{(n-m)k}}{1 - \omega^{n-m}} = 0$$

κ Infinite

Let $\alpha \notin \mathbb{Q}$. It is well known that $n\alpha$ mod 1 is equidistributed. For all $[a, b] \subset [0, 1]$, given $\epsilon > 0$, there exists $M(\epsilon, a, b, \alpha)$ such that

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

for all $N \geq M(\epsilon, a, b, \alpha)$.
Now let N be sufficiently large so that N^δ is greater than $M(\epsilon, a, b, \alpha)$.

Quantifying Dependencies

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

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

Let $K_{i,j}$ be the number of matrix entries shared by $X_{i,n}$ and $X_{j,n}$. Fixing $X_{i,n}$,

$$\begin{aligned} P(K_{i,j}=k) &= \frac{1}{k!}\sum_{\alpha=0}^{n-k}\frac{(-1)^{\alpha}}{\alpha!} \\ &= \frac{1}{ek!}+O\left(\binom{n}{k}\frac{1}{(n-k)n!}\right) \end{aligned}$$

Quantifying Dependencies

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

$$\begin{aligned}\mathrm{Var}\left(K_{i,j}\right) &= \sum_{k=0}^{n-2} \frac{k^2}{ek!} + \sum_{k=0}^{n-2} O\left(\binom{n}{k}\frac{k^2}{n!(n-k)}\right) - 1 \\ &\to 1\end{aligned}$$

Using Chebyshev's Inequality,

$$P(|K_{i,j} - 1| \geq \gamma) \leq 1/\gamma^2$$

References

BGMRS T. Becker, A. Greaves-Tunnell, S. Miller, R. Ronan, and F. Strauch, *Benford's Law and Continuous Dependent Random Variables*. http://arxiv.org/abs/1111.0568.

BLD *Benford's Law Distribution Leading Digit*. Digital image. ISACA. N.p., n.d. Web. 15 June 2013. http://www.isaca.org/Journal/Past-Issues/2011/Volume-3/PublishingImages/11v3-understanding-and.jpg.

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

DFD *Distribution of First Digit.* Digital image. Benford's Law. DataGenetics, n.d. Web. 15 July 2013. http://www.datagenetics.com/blog/march52012/.

JJ Jing, Joy. *Benford's Law and Stick Decomposition.* Undergraduate Mathematics Thesis, Williams College. Advisor: Steven Miller.

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.* http://arxiv.org/pdf/math/0412003v2.pdf

MT Miller, Steven J. and Ramin Takloo-Bighash. "Needed Gaussian Integral." *An Invitation to Modern Number Theory.* Princeton: Princeton UP, 2006. 222. Print. http://arxiv.org/pdf/1111.0568v1.pdf

S Singleton, Tommie W., Ph.D. *Benford's Law Test/Comparison.* Digital image. ISACA. N.p., 2011. Web. 16 July 2013. http://www.isaca.org/Journal/Past-Issues/2011/Volume-3/Pages/Understanding-and-Applying-Benfords-Law.aspx.