Benford's Law and Dependent Random Variables

Thealexa Becker[1], Alexander Greaves-Tunnell[2], Steven J. Miller[2], and Ryan Ronan[3]

[1]Smith College
[2]Williams College
[3]Cooper Union

Joint Meetings of the AMS / MAA, Boston
January 5, 2011

Benford's Law

Benford's Law: Newcomb (1881), Benford (1938)

For many data sets, probability of observing a first digit of d base B is $\log_B\left(\frac{d+1}{d}\right)$.

Benford's Law

Benford's Law: Newcomb (1881), Benford (1938)

For many data sets, probability of observing a first digit of d base B is $\log_B\left(\frac{d+1}{d}\right)$.

Examples:

- iterates of 3x+1, along with power, exponential and rational maps
- products of independent random variables
- particle decay, hydrology and financial data

Benford's Law

Benford's Law: Newcomb (1881), Benford (1938)

For many data sets, probability of observing a first digit of d base B is $\log_B\left(\frac{d+1}{d}\right)$.

Examples:

- iterates of 3x+1, along with power, exponential and rational maps
- products of independent random variables
- particle decay, hydrology and financial data

Question: Which systems lead to Benford behavior?

Dependence Considerations

Most efforts to identify Benford processes assume random variables are *independent*.

Dependence Considerations

Most efforts to identify Benford processes assume random variables are *independent*.

Motivated by "real world" processes shown to be Benford, we want to explore dependent systems.

Dependence Considerations

Most efforts to identify Benford processes assume random variables are *independent*.

Motivated by "real world" processes shown to be Benford, we want to explore dependent systems.

Our motivating example (Lemons 1986):
Decomposition of a conserved quantity as a model for particle decay.

A Model for Decomposition of Conserved Quantities

A discrete-time, continuous-division model of decay:

- “Stick” of length L divided in two at each stage.

A Model for Decomposition of Conserved Quantities

A discrete-time, continuous-division model of decay:

- "Stick" of length L divided in two at each stage.
- Cuts from continuous random variable K on $(0, 1)$.

A Model for Decomposition of Conserved Quantities

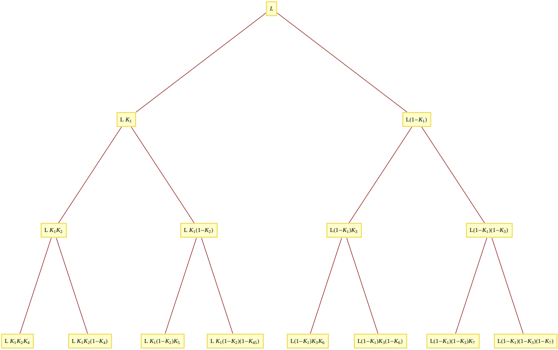
A discrete-time, continuous-division model of decay:

- “Stick” of length L divided in two at each stage.
- Cuts from continuous random variable K on $(0, 1)$.
- Pieces $\{X_i\}$ not independent, as must sum to L.

A Model for Decomposition of Conserved Quantities

A discrete-time, continuous-division model of decay:

- "Stick" of length L divided in two at each stage.
- Cuts from continuous random variable K on $(0, 1)$.
- Pieces $\{X_i\}$ not independent, as must sum to L.



Equidistribution and Benford's Law

Significand: for positive x, write $x = S_{10}(x) \cdot 10^{k(x)}$.

Equidistribution and Benford's Law

Significand: for positive x, write $x = S_{10}(x) \cdot 10^{k(x)}$.

Equidistribution: $\{y_i\}$ equidistributed mod 1 if for any $[a,b] \subset [0,1]$ have $\lim_{N\to\infty} \frac{\#\{n \le N : y_n \in [a,b]\}}{N} = b - a$.

For $s \in [1, 10)$, let

$$\varphi_s(u) \ := \ \begin{cases} 1 & \text{if } u\text{'s significand is at most } s \\ 0 & \text{otherwise.} \end{cases}$$

Equidistribution and Benford's Law

Significand: for positive x, write $x = S_{10}(x) \cdot 10^{k(x)}$.

Equidistribution: $\{y_i\}$ equidistributed mod 1 if for any $[a, b] \subset [0, 1]$ have $\lim_{N\to\infty} \frac{\#\{n \le N : y_n \in [a,b]\}}{N} = b - a$.

For $s \in [1, 10)$, let

$$\varphi_s(u) := \begin{cases} 1 & \text{if } u\text{'s significand is at most } s \\ 0 & \text{otherwise.} \end{cases}$$

Fundamental Equivalence

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

The Mellin Transform

Let $f(x)$ be a continuous real-valued function on $[0, \infty)$. Define its Mellin transform, $(\mathcal{M}f)(s)$, by

$$(\mathcal{M}f)(s) := \int_0^\infty f(x) x^s \frac{dx}{x}.$$

The Mellin Transform

Let $f(x)$ be a continuous real-valued function on $[0, \infty)$. Define its Mellin transform, $(\mathcal{M}f)(s)$, by

$$(\mathcal{M}f)(s) \; := \; \int_0^\infty f(x) x^s \frac{dx}{x}.$$

- Note that $(\mathcal{M}f)(s) = \mathbb{E}[x^{s-1}]$, thus results concerning expected values translate to results on Mellin transforms.

The Mellin Transform

Let $f(x)$ be a continuous real-valued function on $[0, \infty)$. Define its Mellin transform, $(\mathcal{M}f)(s)$, by

$$(\mathcal{M}f)(s) \;:=\; \int_0^\infty f(x)x^s \frac{dx}{x}.$$

- Note that $(\mathcal{M}f)(s) = \mathbb{E}[x^{s-1}]$, thus results concerning expected values translate to results on Mellin transforms.
- With a logarithmic change of variables, we can translate between Mellin and Fourier transforms.

Products of Independent Random Variables

Jang, Kang, Kruckman, Kudo and Miller (2009)

$\Xi_1, \dots, \Xi_N$ independent variables with densities f_{Ξ_m} and

$$\lim_{N\to\infty} \sum_{\substack{\ell=-\infty \\ \ell\neq 0}}^{\infty} \prod_{m=1}^{N} (\mathcal{M}f_{\Xi_m})\left(1 - \frac{2\pi i\ell}{\log B}\right) \;=\; 0 \;\; (C1).$$

Then as $\Xi_1 \cdots \Xi_N$ converges to Benford's law.

Products of Independent Random Variables

Jang, Kang, Kruckman, Kudo and Miller (2009)

$\Xi_1, \dots, \Xi_N$ independent variables with densities f_{Ξ_m} and

$$\lim_{N\to\infty} \sum_{\substack{\ell=-\infty \\ \ell\neq 0}}^{\infty} \prod_{m=1}^{N} (\mathcal{M} f_{\Xi_m}) \left(1 - \frac{2\pi i \ell}{\log B}\right) = 0 \quad (C1).$$

Then as $\Xi_1 \cdots \Xi_N$ converges to Benford's law.

- (C1) is quite weak, met by most distributions.

Products of Independent Random Variables

Jang, Kang, Kruckman, Kudo and Miller (2009)

$\Xi_1, \dots, \Xi_N$ independent variables with densities f_{Ξ_m} and

$$\lim_{N\to\infty} \sum_{\substack{\ell=-\infty \\ \ell\neq 0}}^{\infty} \prod_{m=1}^{N} (\mathcal{M}f_{\Xi_m}) \left(1 - \frac{2\pi i \ell}{\log B}\right) = 0 \quad (C1).$$

Then as $\Xi_1 \cdots \Xi_N$ converges to Benford's law.

- (C1) is quite weak, met by most distributions.
- **Corollary:** For $\Xi_1, \dots, \Xi_N$ uniformly distributed on $(0,1)$ and $N \geq 4$,
  $$|\varphi_s(u) - \log_{10} s| \leq \left(\frac{1}{2.9^N} + \frac{\zeta(N)-1}{2.7^N}\right) 2\log_{10} s.$$

Limiting Behavior of Decompostions

Theorem (Decomposition model as above)

Fix a continuous density f on $[0,1]$ st $f(x)$, $f(1-x)$ satisfy (C1), set

$$P_N(s) \ := \ \frac{\sum_{i=1}^{2^N} \varphi_s(X_i)}{2^N},$$

the fraction of pieces with significand less than or equal to s. Then

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

Limiting Behavior of Decompostions

Theorem (Decomposition model as above)

Fix a continuous density f on $[0,1]$ st $f(x)$, $f(1-x)$ satisfy (C1), set

$$P_N(s) \;:=\; \frac{\sum_{i=1}^{2^N} \varphi_s(X_i)}{2^N},$$

the fraction of pieces with significand less than or equal to s. Then

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

- Amalgamation of processes converges to Benford.

Limiting Behavior of Decompostions

Theorem (Decomposition model as above)

Fix a continuous density f on $[0,1]$ st $f(x)$, $f(1-x)$ satisfy (C1), set

$$P_N(s) \ := \ \frac{\sum_{i=1}^{2^N} \varphi_s(X_i)}{2^N},$$

the fraction of pieces with significand less than or equal to s. Then

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

- Amalgamation of processes converges to Benford.
- May consider a single process (if number stages tend to infinity).

Convergence of Amalgamated Lengths to Benford

$$\mathbb{E}[P_N(s)] = \mathbb{E}\left[\frac{\sum_{i=1}^{2^N} \varphi_s(X_i)}{2^N}\right] = \frac{1}{2^N}\sum_{i=1}^{2^N} \mathbb{E}[\varphi_s(X_i)].$$

Convergence of Amalgamated Lengths to Benford

$$\mathbb{E}[P_N(s)] \;=\; \mathbb{E}\left[\frac{\sum_{i=1}^{2^N} \varphi_s(X_i)}{2^N}\right] \;=\; \frac{1}{2^N}\sum_{i=1}^{2^N} \mathbb{E}[\varphi_s(X_i)].$$

Note: Dependencies exist in the pieces $\{X_i\}$, not the random variables K_i.

Convergence of Amalgamated Lengths to Benford

$$\mathbb{E}[P_N(s)] \;=\; \mathbb{E}\left[\frac{\sum_{i=1}^{2^N} \varphi_s(X_i)}{2^N}\right] \;=\; \frac{1}{2^N}\sum_{i=1}^{2^N} \mathbb{E}[\varphi_s(X_i)].$$

Note: Dependencies exist in the pieces $\{X_i\}$, not the random variables K_i.

We can apply JKKKM's theorem, as the Mellin transform at $1 - \frac{2\pi i \ell}{\log 10}$ is strictly less than 1.

For specific choices of f, can obtain precise bounds on the error. Ex: $\mathbb{E}[\varphi_s(X_i)] - \log_{10} s \ll \frac{1}{2.9^N}$.

Analysis of a Single Process

Dependence greatly complicates analysis here.

Problem: Evaluating cross terms $\mathbb{E}[\varphi_s(X_i)\varphi_s(X_j)]$ for $i \neq j$.

Analysis of a Single Process

Dependence greatly complicates analysis here.

Problem: Evaluating cross terms $\mathbb{E}[\varphi_s(X_i)\varphi_s(X_j)]$ for $i \neq j$.

Solution:

- Pair (X_i, X_j) shares M factors, split at $M + 1$. Remaining elements in the product are independent, so we can simplify our integral.

Analysis of a Single Process

Dependence greatly complicates analysis here.

Problem: Evaluating cross terms $\mathbb{E}[\varphi_s(X_i)\varphi_s(X_j)]$ for $i \neq j$.

Solution:

- Pair (X_i, X_j) shares M factors, split at $M + 1$. Remaining elements in the product are independent, so we can simplify our integral.
- Study integral as a function of the significand of first $M + 1$ variables, which we can bound by JKKKM.

Analysis of a Single Process

Dependence greatly complicates analysis here.

Problem: Evaluating cross terms $\mathbb{E}[\varphi_s(X_i)\varphi_s(X_j)]$ for $i \neq j$.

Solution:

- Pair (X_i, X_j) shares M factors, split at $M+1$. Remaining elements in the product are independent, so we can simplify our integral.
- Study integral as a function of the significand of first $M+1$ variables, which we can bound by JKKKM.
- Remaining task is to count for each of the 2^N choices of i and for $1 \leq n \leq N$, how many choices of X_j have n factors not in common with X_i.

Analysis of a Single Process

Dependence greatly complicates analysis here.

Problem: Evaluating cross terms $\mathbb{E}[\varphi_s(X_i)\varphi_s(X_j)]$ for $i \neq j$.

Solution:

- Pair (X_i, X_j) shares M factors, split at $M+1$. Remaining elements in the product are independent, so we can simplify our integral.
- Study integral as a function of the significand of first $M+1$ variables, which we can bound by JKKKM.
- Remaining task is to count for each of the 2^N choices of i and for $1 \leq n \leq N$, how many choices of X_j have n factors not in common with X_i.
- Resulting sum bounds variance above and goes to 0, thus $\lim_{N\to\infty} \mathrm{Var}\left(P_N(s)\right) = 0$.

Summary of Results

- Decay model: Stick decomposes in discrete stages, cut determined by continuous density function.
  - Pieces must sum to original stick length, thus they are dependent.
  - Allowable densities obey weak condition C1.

Summary of Results

- Decay model: Stick decomposes in discrete stages, cut determined by continuous density function.
    - Pieces must sum to original stick length, thus they are dependent.
    - Allowable densities obey weak condition C1.
- Expected lengths of pieces from amalgamation of processes converges to Benford distribution.
    - Key observation: dependencies exist among piece lengths, not densities.

Summary of Results

- Decay model: Stick decomposes in discrete stages, cut determined by continuous density function.
  - Pieces must sum to original stick length, thus they are dependent.
  - Allowable densities obey weak condition C1.
- Expected lengths of pieces from amalgamation of processes converges to Benford distribution.
  - Key observation: dependencies exist among piece lengths, not densities.
- Variance in piece length distribution goes to zero for a single process. Calculation complicated by dependencies:
  - Study expectation cross terms as integrals over their independent factors.

Thanks to:

- AMS / MAA
- Williams College SMALL 2011
- National Science Foundation
- Thealexa Becker, Professor Steven J. Miller, Ryan Ronan, Professor Frederick W. Strauch